Utilisation des méthodes de criblage virtuel dans un contexte de santé humaine et environnementale : application aux récepteurs nucléaires et aux perturbateurs endocriniens

Asma Sellami

▶ **To cite this version:**

Asma Sellami. Utilisation des méthodes de criblage virtuel dans un contexte de santé humaine et environnementale : application aux récepteurs nucléaires et aux perturbateurs endocriniens. Biotechnologie. HESAM Université, 2022. Français. NNT : 2022HESAC019 . tel-03975056

HAL Id: tel-03975056

https://theses.hal.science/tel-03975056v1

Submitted on 6 Feb 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

HESAM UNIVERSITÉ

le cnam

École doctorale Sciences et Métiers de l'Ingénieur
Génomique Bioinformatique et Chimie Moléculaire

THÈSE

présentée par : **Asma SELLAMI**

soutenue le : **05 octobre 2022**

pour obtenir le grade de : **Docteur d'HESAM Université**

préparée au : **Conservatoire national des arts et métiers**

Discipline/ Spécialité : **Bioinformatique structurale**

Utilisation des méthodes de criblage virtuel dans un contexte de santé humaine et environnementale :

application aux récepteurs nucléaires et aux perturbateurs endocriniens

THÈSE dirigée par :
Pr Matthieu MONTES

et co-encadrée par :
Dr Nathalie LAGARDE

Jury

Mme Esther KELLENBERGER, Professeur, Université de Strasbourg	Rapportrice
M. Sébastien FIORUCCI, Maître de conférences, Université Côte d'Azur	Rapporteur
Mme Karine AUDOUZE, Professeur, Université Paris Cité	Examinatrice
M. Gautier MOROY, Maître de conférences, Université Paris Cité	Examinateur

THÈSE

Affidavit

Je soussigné / soussignée, Asma SELLAMI, déclare par la présente que le travail présenté dans ce manuscrit est mon propre travail, réalisé sous la direction scientifique de Pr. Matthieu MONTES (directeur) et de Dr. Nathalie LAGARDE (co-directrice), dans le respect des principes d'honnêteté, d'intégrité et de responsabilité inhérents à la mission de recherche. Les travaux de recherche et la rédaction de ce manuscrit ont été réalisés dans le respect de la charte nationale de déontologie des métiers de la recherche.
Ce travail n'a pas été précédemment soumis en France ou à l'étranger dans une version identique ou similaire à un organisme examinateur.

Fait à Paris, le 29/07/2022
Signature
Asma SELLAMI

Affidavit

I, undersigned, Asma SELLAMI, hereby declare that the work presented in this manuscript is my own work, carried out under the scientific direction of Pr. Matthieu MONTES (thesis director) and of Dr. Nathalie LAGARDE (co-thesis director), in accordance with the principles of honesty, integrity and responsibility inherent to the research mission. The research work and the writing of this manuscript have been carried out in compliance with the French charter for Research Integrity.
This work has not been submitted previously either in France or abroad in the same or in a similar version to any other examination body.

Paris, 29/07/2022
Signature
Asma SELLAMI

«Le meilleur moyen de réussir, c'est toujours d'essayer encore une fois »

Thomas Edison

Remerciements

Je tiens tout d'abord à remercier le Professeur Jean-François Zagury de m'avoir accueillie au sein de son laboratoire, de sa confiance et son soutien tout au long de cette thèse.

J'adresse également tous mes remerciements au Professeur Matthieu Montes pour la confiance qu'il m'a accordée et pour avoir supervisé mes travaux de thèses. Je le remercie également pour sa bienveillance et tous les conseils prodigués qui ont certainement forgé la chercheuse que je suis devenue.

Je tiens à exprimer ma gratitude la plus infinie au Docteur Nathalie Lagarde pour avoir cru en moi depuis le début lorsque je n'étais qu'une stagiaire en recherche d'un stage, pour m'avoir prise sous son aile et pour m'avoir formée. Je la remercie également pour sa patience, sa bienveillance et sa générosité inégalables. Merci infiniment pour tous les *memes* et les post instagrams partagés, pour toutes les lettres de recommandations rédigées et pour tous les sangs d'encre que je t'ai fait subir.

J'exprime toute ma reconnaissance au Professeur Esther Kellenberger et au Docteur Sébastien Fiorucci d'avoir accepté de juger mon travail de thèse en tant que rapporteurs.

Je remercie également le Professeur Karine Audouze et le Docteur Gautier Moroy de m'avoir accompagné tout au long de ces trois années en tant que membres du comité de suivi de thèse d'abord et membre du jury pour clôturer cette belle aventure.

Lorsque j'ai commencé cette aventure il y a trois ans, j'étais loin de me douter de tous les rebondissements qu'il y aurait dans ma vie et dans le monde (clin d'œil au COVID oblige !). Une chose est sure, j'étais certaine du soutien et de la bienveillance dans laquelle j'allais évoluer. Si je le pouvais j'écrirais une page pour remercier individuellement chaque membre du laboratoire GBCM qui a croisé ma route pour quelques mois ou quelques années en particulier Sigrid pour les longues conversations philosophiques et existentielles qu'on a pu échanger mais surtout pour son « explosivité » et son accessibilité; Taoufik pour le calme et la bonne humeur contagieuse, Manon dont je suis les traces, merci pour tes conseils , ta disponibilité et ta bienveillance ; Myriam qui a partagé mon aventure au Cnam depuis les premiers jours du stage , pour ce goût de Tunisie au laboratoire, pour ta générosité sans faille et pour m'avoir remonté le moral quand il était au plus bas , je te souhaite le plus grand succès pour la thèse ; Josselin pour les parenthèses musicales partagées ; toute l'équipe UDock/VTX : Florent pour le co-bureau, Léa pour la complicité, Benjamin pour le calme absolu, Macha pour la bienveillance, Maxime pour les sessions « *Food testing* » inoubliables, Nico et Simon pour votre gentillesse et patience; Maïté pour tout le travail fait pour les thésards à l'école doctorale

et enfin Sofia pour ta disponibilité, ta réactivité et pour la maman poule que tu es pour chacun d'entre nous. Je remercie également tous les stagiaires en bioinformatique structurale qui ont croisé mon chemin : Ikram, Floriane, Gian Franco, Ella et bien sûr Mehdi que j'ai connu stagiaire et à qui je souhaite le meilleur pour les années de thèses devant lui ! Je remercie enfin toutes les personnes du Cnam qui ont jalonné mes 3 années ; Chloé pour sa bonne humeur contagieuse, Cécile pour son grain de folie, Laura pour les discussions interminables, Elyas pour ses petites attentions et son aide, Céline pour sa douceur (et les tablettes de chocolat !) et la dernière arrivée Raïssa pour le soleil qu'elle ramène au labo !

Je remercie tous mes amis qui m'ont soutenu tout au long de cette thèse. D'abord Fatma pour sa bonne humeur inégalée pour le confinement partagé et l'hospitalité démesurée ; Imon pour sa philosophie de vie ; Oumayma et Haifa : les sœurs Bouattour toujours là pour mettre l'ambiance ; Khadija ma cousine, mon amie et ma supportrice N°1 ; Kais mon frère-ami ! ; Enfin, Ines, Souhayel, Omar, Hella, Nour, Salma ; Selim ; Olfa ; Hamza, Ahmed, Baha, Faten, Zied, et tous ceux que j'ai bassinés avec ma thèse et qui ont été malgré cela d'un soutien infaillible.

Je remercie également mes parents et mon petit frère qui n'ont cessé de croire en moi et qui m'ont soutenu lorsque j'ai décidé d'entreprendre cette aventure en France. Je dédie cette thèse à la mémoire de mes grands-parents partis tôt pour savourer avec moi ce moment et surtout à mamie Zohra partie au tout début de la thèse. J'espère que là où vous êtes je vous rends fiers.

Je termine par la clé de la réussite de cette thèse, mon mari Amir sans qui je n'y serais jamais arrivé. Merci d'avoir cru en moi quand moi-même je n'y croyais plus, de m'avoir poussé à toujours être la meilleure version de moi-même. Merci pour tous les sacrifices que tu as faits et pour toutes les ~~petites~~ grandes attentions. Je te dédie aussi cette thèse car c'est quelque part la tienne également.

Résumé

Les perturbateurs endocriniens (PE) constituent un problème de santé publique. En effet, l'exposition humaine à ces composés est associée à un risque accru de développement de plusieurs pathologies. Les PE sont capables de pénétrer dans l'organisme et d'interférer avec les fonctions du système endocrinien par divers mécanismes, dont la liaison directe aux récepteurs nucléaires (NR). La détection précoce de potentiels PE est donc nécessaire pour garantir la sécurité de l'Homme et de l'environnement. Ceci est possible en utilisant des tests expérimentaux sur les composés suspects, mais cela reste une tâche difficile notamment en raison du nombre considérable de composés à évaluer. Ainsi, les méthodes *in silico* peuvent être utilisées en amont de ces derniers pour prioriser les composés à tester expérimentalement. Avec cet objectif, nous avons utilisé des méthodes basées sur la structure de la cible (SB) et des méthodes basées sur les ligands (LB) pour prédire la liaison de composés aux NR. Nous avons établi une preuve de concept sur le récepteur ER alpha, le plus étudié des NR. Nous avons ensuite proposé un protocole combinant des modèles de docking et de pharmacophores pour prédire les potentiels PE en se basant sur leur capacité à se lier à six récepteurs nucléaires : AR, ERα, ERβ, GR, PPARγ et TRα. Les modèles développés permettent ainsi de catégoriser la probabilité de composés requêtes de se lier aux NR et donc, en fonction du mécanisme direct, d'être des PE.

Mots clés : In silico, criblage virtuel, docking, pharmacophore, perturbateurs endocriniens, récepteurs nucléaires

Résumé en anglais

Endocrine disrupting chemicals (EDCs) are considered as a public health threat as human exposure to these compounds have been associated with increased risk of several diseases. EDCs are able to penetrate the body and to interfere with the functions of the endocrine system through various mechanisms including a direct binding to nuclear receptors (NR). Early detection of potential EDCs becomes an imperative to prevent human and environmental safety issues. This can be achieved using experimental tests, but it remains a challenging task in particular because of the considerable number of compounds to be evaluated. *In silico* methods can then be used in complement, to prioritize compounds for experimental testing and to help elucidating toxicity mechanisms. In this work, we used structure-based (SB) and ligand-based (LB) *in silico* methods to predict compounds binding to NR. We established a first proof of concept on ER alpha, the most studied NR. We then proposed a pipeline combining docking and pharmacophore models to predict potential EDCs based on their ability to bind six nuclear receptors: AR, ERα, ERβ, GR, PPARγ and TRα. The pipeline output enables to categorize query compounds according to their probability of being NR binders and thus, accordingly to the direct mechanism, potential EDCs.

Key words: In silico, Virtual screening, docking, pharmacophore, endocrine disrupting chemicals, nuclear receptors

Table des matières

Liste des tableaux

Liste des figures

Liste des annexes

Première partie

Introduction

1. Introduction

La toxicologie, comme les autres sciences, s'est développée par phases. Les toxicologues affirment toutefois que la phase initiale de cette discipline a précédé celle de la plupart des autres sciences biologiques, puisqu'elle impliquait la reconnaissance par l'homme primitif des agents sûrs et dangereux présents dans son environnement [1]. La phase suivante (l'Antiquité et le Moyen-Âge) a été caractérisée par l'utilisation de ces informations pour le bien (thérapeutique) et le mal (empoisonnement). C'est à la Renaissance que Paracelse a reconnu l'importance du paradigme dose-réponse, ce qui a marqué le début de la toxicologie moderne. Aujourd'hui, la toxicologie se concentre sur les mécanismes moléculaires, et l'utilisation des nouvelles technologies pour stocker et diffuser ces informations [1].

L'un des objectifs de la toxicologie est donc l'évaluation du risque d'exposition à un produit chimique, que ce soit un composé présent dans l'environnement ou un médicament par exemple. Ces effets néfastes sont déterminés grâce aux tests expérimentaux. L'expérimentation humaine étant à exclure (sauf dans le cas de substances thérapeutiques) pour des raisons morales, éthiques et légales, la principale méthode utilisée est l'expérimentation sur les animaux (modèles ***in vivo***), cependant, deux problématiques majeures y sont associées. La première est que la plupart des modèles d'extrapolation des risques, qui utilisent les résultats des tests sur les animaux, ne peuvent pas tenir compte du fait que l'espèce testée et l'espèce cible sont biologiquement différentes. Certains effets peuvent en effet apparaitre chez l'animal sans pour autant se manifester chez l'Homme et vice versa [2,3]. Cela a incité à reconsidérer les tests animaux et leurs légitimités à être utilisés seuls pour déceler le caractère toxique des composés. La seconde problématique est encore une fois liée à une raison éthique. Les exigences en matière d'expérimentation animale et de bien-être des animaux ont ainsi évolué et le principe des « 3R » (replacement, reduction and refinement) [4,5] a émergé. Ce principe prône de privilégier les approches mécanistiques basées sur des données empiriques et a conduit au développement de méthodes alternatives en toxicologie, en particulier le renforcement des essais ***in vitro*** qui permettent de classer par ordre de priorité les substances chimiques pour lesquelles le mécanisme d'action devra être étudié [4].

De nos jours, les méthodes de criblage computationnelles ou méthodes de criblage ***in silico*** ont aussi pris de l'importance en toxicologie grâce aux développements de l'informatique et des capacités de calculs combinées au développement en chimie et biologie moléculaire [2,3,6].

2. Evaluation des risques en toxicologie

L'évaluation des risques de toxicité est cruciale pour les différentes industries (pharmaceutiques, cosmétiques et chimiques) ainsi que pour les organismes réglementaires. Les objectifs de ces évaluations sont toutefois très divers. Dans l'industrie pharmaceutique, l'évaluation des risques est menée tout au long du processus de découverte et de développement des médicaments. Elle commence avant même sa synthèse, et se poursuit jusqu'aux essais cliniques où l'évaluation de la balance bénéfices/risques est un point crucial pris en compte dans la demande de mise sur le marché (AMM) [7]. Malgré une évaluation rigoureuse dès les premières phases de développement, le taux d'attrition reste élevé, souvent en raison de problèmes de sécurité [8]. Dans les industries chimiques, les objectifs principaux de l'évaluation de risque de toxicité sont orientés vers la sécurité des travailleurs et l'évaluation des risques environnementaux. Dans l'industrie cosmétique, les évaluations des risques portent essentiellement sur la sécurité des consommateurs. Le développement des connaissances scientifiques ainsi que les progrès technologiques ont permis l'évolution de l'évaluation des risques en toxicologie en passant d'une toxicologie descriptive basée sur les tests *in vivo* sur animaux mais aussi *in vitro* à une toxicologie mécanistique intégrant les connaissances sur les voies de toxicologie responsables des effets indésirables (*in vitro* et *in silico)* [9]. Cette évolution est de plus en plus intégrée aux processus règlementaires [4,10] laissant entrevoir l'espoir de réduire voire de s'affranchir des tests sur les animaux [5].

2.1. Etudes *in vivo*

Classiquement, ces études sont réalisées sur des animaux sains et adultes. Mais des variations peuvent exister pour étudier un phénomène toxicologique chez une population particulière (par exemple des animaux génétiquement modifiés pour étudier l'impact d'un toxique sur un processus pathogénique donné) [9]. Au cours de ces tests, un composé est administré à différentes doses et à des fréquences variées. Généralement, les principales voies d'exposition sont étudiées (voie orale, voie respiratoire et voie dermique). L'apparition ou non d'un effet est analysée pour (1) déterminer la toxicité aiguë d'un composé (exposition unique et courte de moins de 24h) et pour (2) déterminer les effets de l'exposition répétée c'est-à-dire évaluer la toxicité subaiguë (exposition répétée inférieure à un mois), sub-chronique (exposition répétée entre 1 et 3 mois)

et chronique (exposition répétée pour plus de 3 mois allant jusqu'au deux tiers de la durée de vie de l'animal) [9].

Les tests de toxicité aiguë sont généralement utilisés pour déterminer les doses utiles : la dose effective médiane (DE50) soit la dose qui affecte la moitié d'un groupe étudié et la dose létale médiane (DL50) c'est-à-dire la dose entrainant la mort de la moitié de la population animale étudiée [11]. Des tests de toxicité sur organes peuvent être pratiqués (toxicité locale) [5]. Les études à doses répétées permettent généralement de déterminer les effets indésirables résultants d'une exposition longue durée à une substance toxique. Ainsi, les valeurs critiques d'exposition peuvent être déterminées ainsi que les différents organes impactés et les effets de cette exposition sur ces organes (par exemple défaillance d'un organe). La reprotoxicité, la tératotoxicité ainsi que la carcinogénicité sont évaluées aux moyens de protocoles spécifiques [9,12].

Bien que les tests *in vivo* soient considérés comme le « *gold standard* » de l'évaluation toxicologique, ces derniers présentent diverses limites. En effet, les études *in vivo* ne sont pas toujours capables de modéliser la toxicité chez l'Homme en raison des différences métaboliques ou toxicocinétiques entre l'Homme et l'animal entre autres. La catastrophe du thalidomide en reste le plus terrible témoin [13]. De plus, ces études étant longues et coûteuses, elles ne permettent pas de tester beaucoup de composés ce qui constitue une véritable limite dans un contexte industrialisé où le nombre de nouvelles molécules synthétisées ne cesse de croitre. Enfin, la dernière limite et sans doute la plus importante est l'enjeu éthique que représentent ces tests sur les animaux. Dès 1959, Russel et Bursh[14] ont introduit le concept des 3R qui incite à Réduire, Raffiner et Remplacer ces expérimentations.

2.2. Etudes in vitro

Dans ce contexte, les tests *in vitro* et plus spécifiquement les tests sur les cultures cellulaires présentent certains avantages. En effet, ils permettent d'étudier l'effet des substances sur des cellules humaines et de s'émanciper des différences inter-espèces en plus de pouvoir étudier l'effet direct d'une substance sans interférence hormonale et dans les conditions qui assurent l'imputabilité directe de la toxicité à l'exposition (contrôle de l'humidité, pH, oxygénation etc..). De plus, les tests *in vitro* permettent de tester un grand nombre de produits avec un large panel de doses et de temps d'expositions. Additionnellement, la miniaturisation des dispositifs d'expérimentation permet de réaliser les tests avec de très faibles quantités du produit [3]. Enfin, grâce à l'automatisation, la parallélisation et la robotisation des processus (*high throughput*

screening ou HTS), des données haut débit peuvent être obtenues ce qui permet de pouvoir trier rapidement les composés sur la base d'un point de vue toxicologique particulier (toxicité cardiaque, homéostasie calcique, mort cellulaire etc.) [15,16]. Originairement développées pour la découverte de médicaments, les méthodes HTS sont devenues de plus en plus populaires dans les études toxicologiques, car rapides et permettant de réduire considérablement le coût des tests expérimentaux [3,17] (Figure 1).

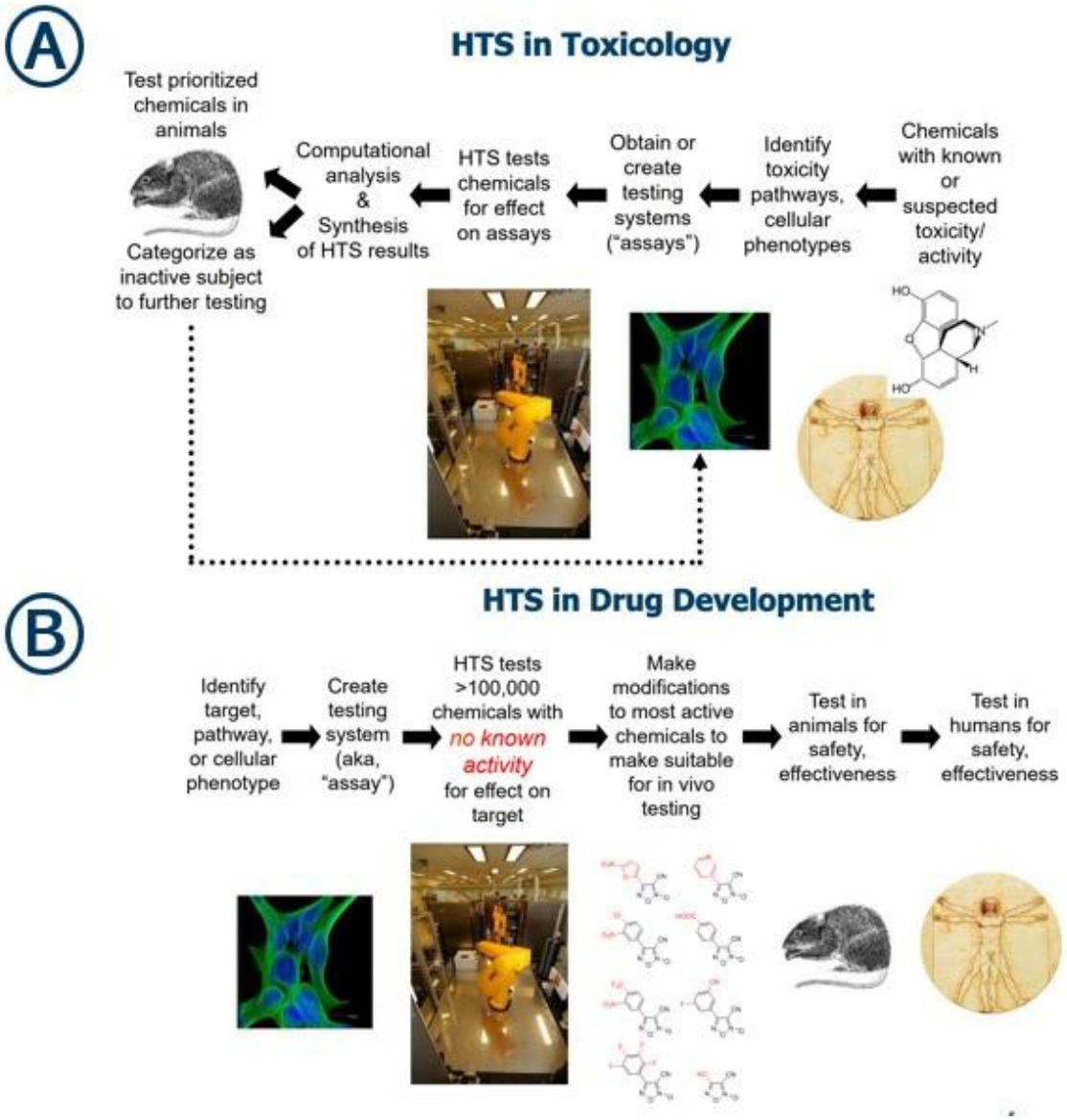


Figure 1 : Paradigme du criblage à haut débit (HTS) en toxicologie (A) Par rapport à sa place en *drug design* (B) d'après [18]

Il existe une multitude de paramètres toxicologiques pouvant être évalués *in vitro* et différents mécanismes de toxicités peuvent être explorés via différents test *in vitro* dédiés. En effet, les

tests mis au point permettent de détecter et mesurer l'activité biologique comme la liaison à des cibles toxicologiques connues, le changement de biomarqueurs ou la dégradation cellulaire [15]. Aujourd'hui, les efforts visent à acquérir de nouveaux tests biologiques permettant d'élargir les données mesurables relatives à des manifestations toxicologiques. Les scientifiques cherchent à identifier de nouveaux biomarqueurs de toxicité par exemple [18].

Toutefois, les études *in vitro* présentent aussi quelques limites. La première est liée aux conditions expérimentales comme la non-spécificité des tests due à la réactivité des composés avec les supports expérimentaux (plastique) ou encore la dégénérescence des cellules monoclonales utilisées pour les tests. De plus, la mise en place de certains tests *in vitro* et leur validation restent un processus assez long ce qui fait que ces derniers sont peu utilisés dans le cadre d'une validation règlementaire contrairement aux tests animaux [9].

2.3. Etudes *in silico*

Au vu de la quantité de produits chimiques qui sont utilisés aujourd'hui et des nombreux autres synthétisés et potentiellement synthétisables, il devient vital de disposer de méthodes efficaces pour évaluer l'effet de ces composés sur l'environnement et la santé humaine. Bien que les tests expérimentaux *in vivo* et *in vitro* restent indispensables pour confirmer la toxicité d'un composé, ces derniers sont à la fois longs, coûteux et parfois difficiles à mettre en place. Ainsi, des méthodes *in silico* précises permettant de réaliser un premier criblage conduisant à l'identification des risques et à la hiérarchisation des composés [19] représentent un atout complémentaire. Par ailleurs, ces méthodes combinées avec les tests *in vitro* permettront probablement un jour de remplacer les tests *in vivo* [12]. En effet, les méthodes HTS permettent de fournir une grande quantité de données et les approches informatiques et statistiques permettent de les analyser et les interpréter correctement. Ainsi, les résultats sont susceptibles d'être utilisés pour construire des modèles prédisant le potentiel de toxicité des nouveaux produits chimiques sur la base de leur comportement dans les essais *in vitro*. Les résultats du criblage intégrés dans des modèles devraient permettre de mieux comprendre aussi les mécanismes d'action de toxicité, ce qui sera très utile pour l'évaluation des risques [12].

La toxicologie *in silico* appelée aussi la toxicologie computationnelle est une sous-discipline de la toxicologie. Elle est définie comme « l'application de modèles mathématiques et méthodes informatiques pour prédire les effets néfastes et mieux comprendre les mécanismes uniques ou multiples par lesquels un produit chimique donné induit des dommages et *in fine* pouvoir

prédire les effets néfastes des composés toxiques pour la santé humaine et/ou l'environnement » [12,20,21]. La toxicologie computationnelle est un domaine multidisciplinaire qui combine des connaissances sur les voies de toxicité avec des données chimiques et biologiques pertinentes. Elle a pour but d'assurer le développement, la vérification et l'évaluation de modèles informatiques multi-échelles utilisés pour mieux comprendre les mécanismes par lesquels un produit chimique donné induit des dommages. Pour cela, de grands ensembles de données biologiques et chimiques sont analysés pour extraire les données à fort contenu informatif, en s'appuyant pour cela sur de nouvelles méthodes biostatistiques et sur la puissance de calcul [20]. La toxicologie *in silico* peut s'organiser en 3 axes distincts mais intimement reliés au sein desquels différentes méthodes sont regroupées (Figure 2). Des méthodes comme le QSAR ou le *Read-across* peuvent ainsi être utilisées pour différents cas et ce même lorsque les données sont manquantes. Le schéma de mode d'action ou AOP (Adverse Outcome Pathways) quant à lui permet de décrire qualitativement (schématiquement) les mécanismes de toxicité entre un évènement initiateur et un effet indésirable. Ainsi les méthodes de criblage virtuel habituellement utilisées dans un contexte de *drug design* font partie intégrante des outils *in silico* pour la toxicologie.

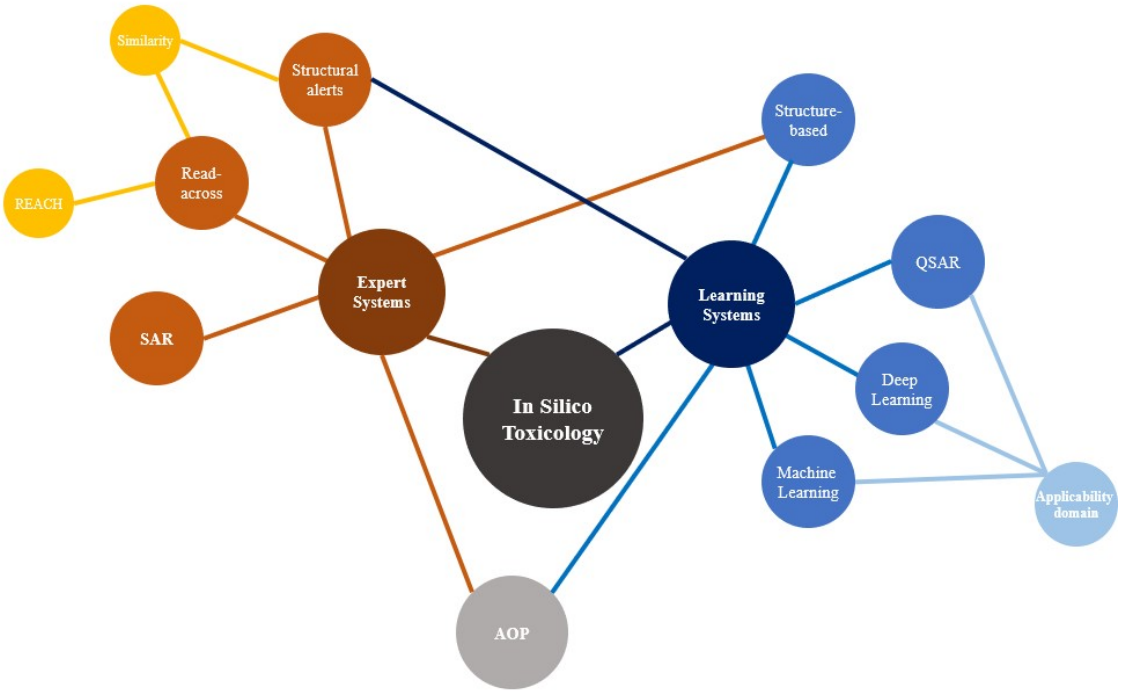


Figure 2 : Prise de décision en toxicologie *in silico* d'après [12]

2.3.1. Les voies impliquées dans les effets indésirables : AOP

Le concept d'AOP (*adverse outcome pathway*) permet de décrire la relation causale existant entre un évènement moléculaire déclencheur provoqué par un toxique, des évènements cellulaires, moléculaires et physiologiques intermédiaires clefs ainsi que l'effet néfaste sur un organisme ou une population [9]. La construction d'une AOP nécessite donc des données de type mécanistique pour identifier les voies biologiques sous-jacentes à la toxicité étudiée (*données in vivo*, *in vitro* et même *in silico*) mais aussi de définir la stratégie de tests intégrés qui permettent de mesurer la tolérance de ces voies à la perturbation sans induire une voie spécifique de toxicité.

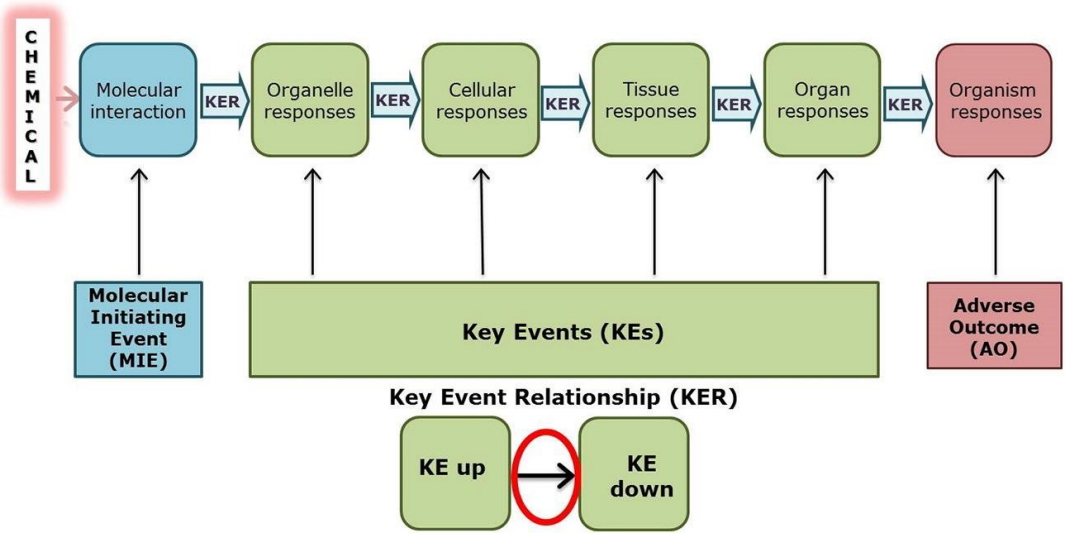


Figure 3 : Représentation du schéma de mode d'action (AOP) d'après [22]

2.3.2. Les systèmes d'experts

Les systèmes ou méthodes d'experts se basent sur les connaissances (des experts) pour déduire, prédire ou expliquer les mécanismes de toxicité pour un seul composé ou pour un groupe de composés pour lesquels le potentiel toxique est inconnu. Ainsi, les systèmes d'experts peuvent utiliser pour leurs prédictions aussi bien les données mécanistiques (AOP) que les données de criblage. Parmi les méthodes, nous retrouvons le « Read-Across », les alertes structurales et les études de relation structure-activité. La méthode « Read-across » est très populaire et très utilisée en toxicologie pour l'identification des composés dangereux. Elle permet d'extrapoler les résultats d'un test *in vivo* pour un ou plusieurs composés sources à des composés cibles sur la base d'une similarité entre la source et la cible [23] (c.f. paragraphe 3.1 criblage basé sur la structure du ligand).

2.3.3. Les méthodes d'apprentissage

Les méthodes d'apprentissage regroupent les méthodes de QSAR (c.f. paragraphe 3.1.3 méthodes QSAR)- très populaires en toxicologie-, les méthodes de *machine learning* et *de deep learning* ainsi que les méthodes basées sur la structure des protéines (c.f. paragraphe 3.2. criblage basé sur la structure des cibles).

3. Le criblage virtuel à haut débit (VS) en toxicologie

Différentes méthodes de toxicologie *in silico* existent au sein desquelles les méthodes de criblage virtuel occupent une place importante comme présenté plus haut.

Les méthodes de criblage virtuel utilisées dans un **contexte thérapeutique**, dans le cadre d'un projet de *drug design*, permettent de filtrer un grand nombre de molécules afin d'en extraire un ou plusieurs touches (*hits)*. Ces *hits* sont ensuite testés expérimentalement et optimisés chimiquement pour espérer obtenir un candidat thérapeutique pour entrer en phase clinique de développement de médicaments. La base de données de molécules (ou chimiothèque) criblée virtuellement peut contenir des molécules déjà synthétisées ou des molécules virtuelles. En effet, il existe un grand nombre de chimiothèques qui peuvent servir de point de départ pour une campagne de criblage. Lors d'une campagne de criblage dans un **contexte toxicologique**, le but est de déterminer avec le plus de certitude possible les composés susceptibles d'être dangereux. L'effet toxicologique ne pouvant être évalué qu'expérimentalement, il serait donc utile d'établir une liste prioritaire des composés à tester en urgence. Les méthodes de criblage virtuel peuvent être employées pour un large spectre d'applications d'intérêt toxicologique dont le (1) le développement de médicaments pour déceler de potentiels effets indésirables et (2) la toxicologie environnementale (éco-toxicologie) afin de prédire le potentiel toxique des composés chimiques issues de l'industrie agro-alimentaire ou cosmétique par exemple[24].

En fonction de l'information à disposition, ces méthodes peuvent être classées en méthodes basées sur la structure de la cible appelées méthodes *Structure Based* (SB) et méthodes basées sur la structure du ligand (endogène ou connu pour être actif sur la cible) appelées méthodes *Ligand Based* (LB). Dans cette partie nous aborderons les deux catégories de méthodes en mettant à chaque fois l'accent sur les aspects à prendre en considération pour un contexte toxicologique.

3.1. Criblage virtuel basé sur la structure du ligand

Les méthodes LB s'appuient sur un même postulat de départ qui est le **principe de similarité**. Autrement dit, deux molécules structurellement similaires ont une forte probabilité d'avoir le même profil d'activité par rapport à une cible donnée [25]. Ainsi en partant de molécules actives et en recherchant celles qui leur sont similaires, nous pouvons prétendre à retrouver de nouveaux composés actifs. Pour ce faire, il existe différentes approches et algorithmes pour

calculer et opérer une **recherche de similarité**[26]. Ce principe de similarité est aussi utilisé pour définir les propriétés communes entre ces différentes molécules actives et de les exprimer sous forme de **modèles pharmacophoriques** ou en d'équations mathématiques (**méthodes QSAR**).

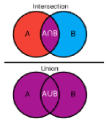
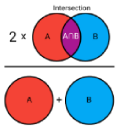
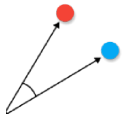
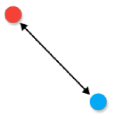
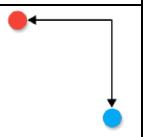
3.1.1. Recherche de similarité

La recherche de similarité est une méthode très populaire dans les différentes branches de la chémoinformatique car elle permet de prédire le comportement moléculaire ainsi que le devenir de composés structurellement apparentés. En toxicologie computationnelle, la recherche de similarité est typiquement utilisée pour les approches de « Read-across » où une estimation de l'activité d'un composé chimique est déterminée en utilisant des données expérimentales disponibles pour d'autres composés qui lui sont hautement similaires[27]. La recherche de similarité requiert de choisir une méthode permettant **la description de la structure** et **une mesure de similarité** (la mesure quantitative de la similarité entre deux représentations numériques)[27].

3.1.1.1. Similarité basée sur les descripteurs moléculaires

Un descripteur est défini comme le résultat final d'expérimentations standardisées ou d'une procédure mathématique et logique qui transforme l'information chimique (une molécule encodée grâce à une représentation symbolique) en un nombre ou vecteur utile[28]. Il existe donc différentes méthodes pour calculer les descripteurs moléculaires[29–36]. Ainsi, chaque molécule étudiée, décrite par un jeu de n descripteurs, aura des coordonnées définies dans l'espace de référence à n dimensions (n -D). La mesure de la similarité se fera à partir de ces coordonnées pour calculer les distances comme présenté dans la première colonne du **Tableau 1** où x_i correspond à la valeur du descripteur pour une coordonnée x. Plus la distance entre deux molécules est petite, plus celles-ci seront similaires dans l'espace x n -D étudié et inversement. Il existe différentes méthodes pour mesurer les distances entre les molécules et la plus étudiée est la distance euclidienne. Enfin Au-delà d'une mesure de similarité, ces **descripteurs numériques** sont utilisés afin de construire des modèles QSAR (c.f. paragraphe 3.1.3 les méthodes QSAR).

Tableau 1 : Mesures de similarité pour le cas des variables continues et binaires d'après[37]

Nom	Formule pour des variables continues (Rang de valeurs)	Formule pour des variables binaires	Schéma
Coefficient de Tanimoto (ou de Jaccard)	$S_{AB} = \frac{\sum_{i=1}^{N} x_{iA}x_{iB}}{\sum_{i=1}^{N}(x_{iA})^2 + \sum_{i=1}^{N}(x_{iB})^2 - \sum_{i=1}^{N} x_{iA}x_{iB}}$ $(-0{,}333 \text{ à } 1)$	$S_{AB} = \frac{c}{a+b-c}$ (0 à 1)	
Coefficient de Dice	$S_{AB} = \frac{2\sum_{i=1}^{N} x_{iA}x_{iB}}{\sum_{i=1}^{N}(x_{iA})^2 + \sum_{i=1}^{N}(x_{iB})^2}$ $(-1 \text{ à } 1)$	$S_{AB} = \frac{2c}{a+b}$ (0 à 1)	
Similarité cosinus	$S_{AB} = \frac{\sum_{i=1}^{N} x_{iA}x_{iB}}{\sqrt{\sum_{i=1}^{N}(x_{iA})^2 \sum_{i=1}^{N}(x_{iB})^2}}$ $(-1 \text{ à } 1)$	$S_{AB} = \frac{c}{\sqrt{ab}}$ (0 à 1)	
Distance euclidienne	$D_{AB} = \sqrt{\sum_{i=1}^{N}(x_{iA} - x_{iB})^2}$ $(0 \text{ à } \infty)$	$D_{AB} = \sqrt{a+b-2c}$ $(0 \text{ à } N)$	
Distance de Manhattan	$D_{AB} = \sum_{i=1}^{N} \lvert x_{iA} - x_{iB} \rvert$ $(0 \text{ à } \infty)$	$D_{AB} = a+b-2c$ $(0 \text{ à } N)$	

3.1.1.2. Similarité basée sur les *fingerprints*

La manière la plus populaire pour décrire une structure et calculer sa similarité avec d'autres composés reste l'utilisation **d'empreintes moléculaires** (ou *fingerprints* en anglais) [26]. Ces derniers prennent la forme d'une chaine de *bits* (0 et 1) décrivant les groupements chimiques, les sous-structures et les types de liaison au sein d'une molécule. Il existe des fingerprints dits (*keyed*) définis à partir d'une liste de fragments. Ces derniers encodent la présence (par 1) ou l'absence (par 0) de fragments particuliers d'un dictionnaire de référence [38]. Ainsi la longueur du fingerprint est égale au nombre de fragments (sous-structures) recherchés. Bien qu'intuitifs, les *keyed fingerprints* s'accompagnent d'une perte de l'information chimique. Un autre type d'empreintes appelées fragmentaux (*hashed*) [39], répond à ce déficit en permettant d'encoder l'occurrence des fragments d'une molécule sous forme d'un ou plusieurs bits à l'aide d'un algorithme appelé algorithme de hachage ou « *hashing algorithms* ». Un célèbre exemple de ce type d'empreinte sont les daylight *fingerprints*[40] qui listent tous les motifs formés par les différentes combinaisons d'atomes de la molécule de référence (les motifs de 1 atome, puis les motifs de 2 atomes, de 3, de 4 etc..). Chaque motif active un certain nombre de positions (bits)

dans le *fingerprint*. L'algorithme est fait de manière à ce qu'il soit toujours possible d'assigner des bits à un motif. Ce type de fingerprints ne nécessite donc pas de dictionnaire de fragments de référence mais a besoin de définir certains paramètres à l'avance comme la longueur du fingerprint, la taille des motifs à hacher et le nombre de bits à activer pour chaque motif[40]. Les fingerprints étant des descripteurs dichotomiques, il est possible de calculer un coefficient de similarité. Il existe plusieurs coefficients de similarité (colonne 2 du **Tableau 1**) [41]. Pour calculer un coefficient de similarité entre deux molécules, il faudra déterminer les variables *a* et *b* constituant le nombre de *bits* propres à A et B respectivement et la variable *c* constituant le nombre de propriétés en communs entre les deux.

3.1.1.3. Similarité basée sur la forme

Enfin, il est possible de décrire la molécule à partir de sa forme et sa géométrie 3D et il existe différentes stratégies pour la calculer [42–45]. Le programme ROCS[42] est considérée comme la référence pour la similarité de forme. Cette méthode permet de comparer deux molécules en superposant leur volume représenté par un ensemble de sphères centrées autour des atomes. Ces sphères sont construites à partir de fonction gaussienne permettant d'approximer le volume des atomes[46]. Par ailleurs, ROCS permet aussi de préciser des types d'atome (cycles, H-donneur etc..). Ainsi, pour déterminer la similarité de deux molécules, on définit le chevauchement maximal entre les volumes de deux molécules en utilisant le coefficient de Tanimoto (volume de chevauchement total divisé par somme des volumes individuels) et la distance de Tversky (volume de chevauchement total divisé par le volume de chevauchement auquel s'additionnent les volumes non chevauchés de chaque molécule).

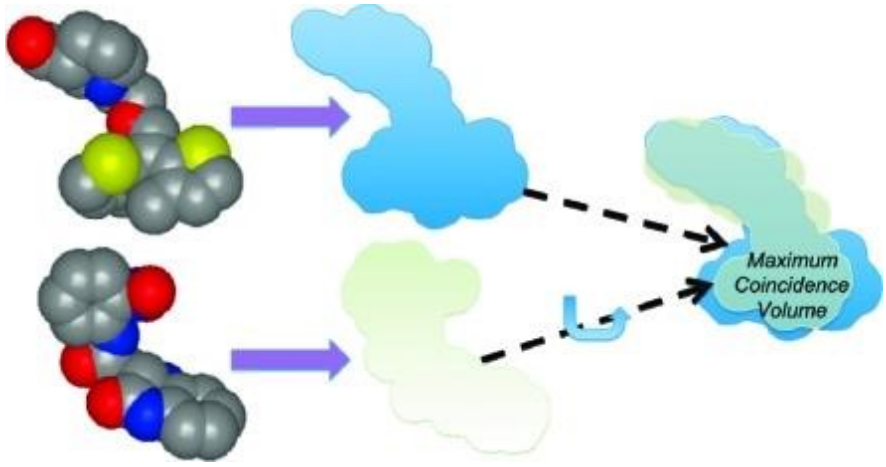


Figure 4 : Similarité de forme entre deux molécules par la détermination du chevauchement de leurs volumes d'après [47]

3.1.2. Modèle de pharmacophores LB

3.1.2.1. Définition

Au-delà du principe de similarité, la définition de pharmacophore est une extension du concept de bioisostérisme qui dit que certains groupements chimiques possèdent des propriétés chimiques et physiques similaires [48]. L'IUPAC définit un pharmacophore comme l'« ensemble des éléments stériques et électroniques d'une molécule nécessaire pour assurer une interaction supramoléculaire avec une cible biologique et pour déclencher ou bloquer une réponse biologique » [49]. Ainsi, cette définition s'émancipe des groupements chimiques et de leurs connectivités. Un pharmacophore peut donc être considéré comme une représentation abstraite de la structure chimique qui met l'accent sur la capacité de composés présentant des effets biologiques comparables (même site actif, même cible biologique) à établir des interactions stériques et électroniques communes. Ce modèle abstrait peut alors servir de filtre de recherche efficace pour le criblage virtuel [48,50]. Il existe deux catégories de pharmacophores dépendants de l'information utilisée : 1) les pharmacophores LB construits sur la base de molécules actives de références et 2) les pharmacophores SB générés à partir de la structure protéique de la cible thérapeutique ou le plus souvent de son complexe avec un ligand. Dans ce paragraphe, nous aborderons les différentes étapes de construction de modèles pharmacophoriques LB. Les modèles SB seront abordés dans une autre partie (c.f. 3.2 Criblage basé sur la structure des cibles).

3.1.2.2. Etapes préliminaires

Deux étapes préliminaires sont nécessaires avant l'étape de génération des modèles à proprement dit : la division des données initiales et l'analyse conformationnelle de ligands.

3.1.2.2.1. Division des données

Tout d'abord, les données initiales brutes sont divisées en 2 jeux : un jeu d'entrainement (training set), un jeu de test (test set). (1) Le jeu d'entrainement est formé d'un groupe de molécules pour lesquelles l'activité est connue. Il peut donc être composé uniquement d'actifs ou contenir à la fois des actifs et des inactifs. Dans les deux cas, un minimum de deux molécules actives est requis au sein du jeu et généralement seules les molécules actives de ce jeu serviront pour la génération des modèles pharmacophoriques. (2) Le jeu de test sert à valider les performances du modèle et est, par conséquent, composé de molécules actives et inactives.

3.1.2.2.2. Analyse conformationnelle

Comme indiqué par la définition d'un pharmacophore, les molécules de références utilisées pour générer le modèle doivent agir au niveau du même site d'action et avoir la même activité biologique. Les molécules utilisées doivent de préférence être dans leurs conformations bioactives faute de quoi il faudra générer les conformations les plus probables. Notons que certains logiciels permettent de partir des structure 2D des molécules et de générer les conformations pendant la création du modèle pharmacophorique lui-même [50,51]. La structure 3D du ligand peut ainsi être extraite de données expérimentales de la PDB ou bien générée computationnellement à partir de la structure 2D au moyen d'outils précis ayant déjà fait leurs preuves [52]. L'un des outils les plus populaire est OMEGA[53] utilisant un algorithme systématique en trois grandes phases : (1) assemblage de la structure 3D à partir d'une bibliothèque de fragments, (2) énumération exhaustive de toutes les torsions rotatives à partir d'une liste d'angle préétablie, (3) échantillonnage du grand nombre de conformations générées en se basant sur des critères de géométrie et d'énergie [54,55].

3.1.2.3. Génération de pharmacophores

Cette étape s'articule en plusieurs sous étapes comme le montre la **Figure 5** :

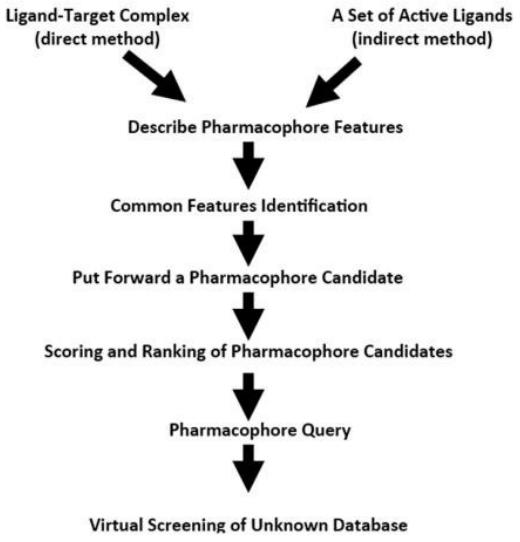


Figure 5 : Etapes de génération du modèle pharmacophorique

3.1.2.3.1. Description des points pharmacophoriques pour chaque ligand

Un pharmacophore 3D est généralement défini comme un ensemble de points pharmacophoriques caractérisés par leurs types, leurs orientations spatiales (coordonnées xyz) et leurs poids (importance). Les points pharmacophoriques typiquement utilisés sont les

groupements donneurs et accepteurs de liaisons hydrogène, les régions hydrophobes, les noyaux aromatiques et les groupements chargés positivement et négativement [48,50]. Par ailleurs, ces points possèdent un centre et une sphère de tolérance qui sont importants pour déterminer l'efficacité du modèle [50]. Il existe différents logiciels destinés à générer des modèles pharmacophoriques comme LigandScout[56], Discovery Studio[57], MOE[58] ou encore PHASE[59] et chacun représente les points pharmacophorique d'une manière différente (**Figure 6**).

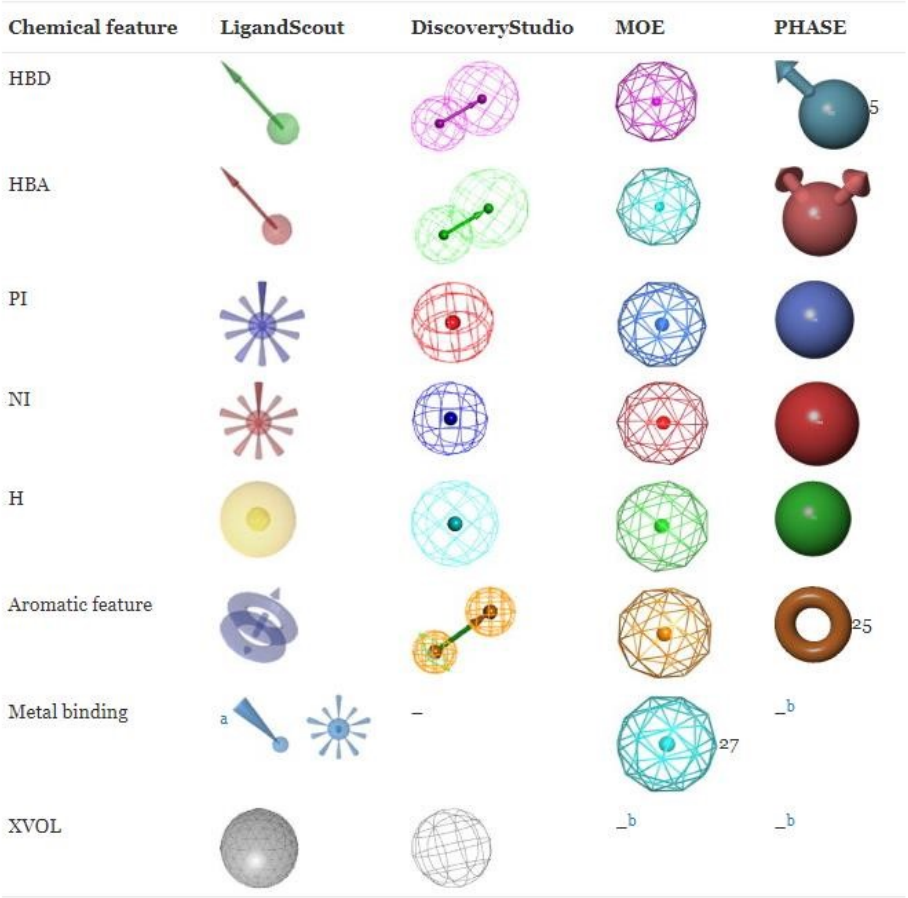


Figure 6 : Description des différents points pharmacophoriques et leurs implémentations au niveau des logiciels LigandScout, Discovery Studio, MOE et PHASE d'après [60]

3.1.2.3.2. Identification des points pharmacophoriques communs

Cette étape consiste à retrouver les points pharmacophoriques en commun entre tous ceux définis pour chaque ligand pour générer des modèles de pharmacophores. Ce procédé s'appelle le *pharmacophore mapping.* Les ligands doivent être alignés pour faire correspondre le

maximum de sous-structures en commun et en générer un modèle. De nombreuses techniques implémentées dans les différents logiciels permettent de le faire[50,61,62]. Parmi celles-ci, il est possible de citer l'utilisation de l'algorithme de détection de « CLIQUE » [63] utilisé dans le logiciel DISCO [64] ou encore l'algorithme génétique utilisé dans GASP[65]. Cette étape conduit à la génération de plusieurs modèles. Un score est attribué à chaque modèle pour les classer entre eux et choisir le meilleur. Ce score permet à chaque fois d'apprécier la qualité de l'alignement réalisé, le volume de recouvrement et de suivre l'évolution de l'étape d'élucidation du pharmacophore. Certaines fonctions de score comme celles associées à l'algorithme HipHop[66] ou au logiciel PHASE[59] prennent en considération la rareté du modèle pharmacophorique de manière à attribuer un score faible à des modèles retrouvés dans des molécules à activité biologiques différentes. Cette récurrence signifie que ces modèles ne sont pas spécifiques d'une interaction en particulier.

3.1.2.4. Validation des modèles pharmacophoriques

Une fois les modèles de pharmacophores générés, une dernière étape de sélection des plus pertinents est réalisée en utilisant les données du jeu test. Il existe différentes manières de déterminer la pertinence des modèles choisis [50]. Les plus couramment utilisées sont l'analyse statistique, les méthodes d'enrichissement, la courbe de ROC (c.f. 3.3. évaluation de méthodes de criblage) et la validation expérimentale (activité biologique) [62]. Cette étape de validation peut mener à recommencer à nouveau l'étape de génération de pharmacophores jusqu'à obtenir des performances satisfaisantes [50].

3.1.2.5. Modèles pharmacophoriques et toxicité

En raison de leur universalité, les modèles pharmacophoriques sont assez efficaces mais surtout utilisables dans différentes applications notamment en toxicologie. En appliquant la même procédure que celle utilisée lors de la recherche de molécules capables d'interagir avec une cible thérapeutique, les pharmacophores peuvent être utilisés en toxicologie pour retrouver les propriétés pharmacophoriques nécessaires pour agir avec une cible non désirée : une anti-cible [67]. Par exemple, ce principe d'anti-cible peut être utilisé dans le cadre de prédiction de toxicité hépatique en ciblant les composés capables d'interagir avec certains récepteurs comme le récepteur CAR ou encore PPARγ [68], ou encore dans le cadre de cardiotoxicité en ciblant les canaux ioniques hERG [69]. Les modèles de pharmacophores peuvent aussi être utilisés pour identifier de potentiels perturbateurs endocriniens ayant la capacité d'interagir avec certains récepteurs aux hormones [70]. Enfin, les modèles pharmacophoriques peuvent être utilisés pour

des études ADMET (absorption, distribution, métabolisme, élimination et toxicité) [50,67]. Les principales cibles sont alors différents cytochromes (CYP 1A2, 2C9, 2C19, 2D6 et 3A4) [71,72] en plus des pompes à efflux comme la P-glycoprotéine. Dans le cas des études ADMET, la prédiction de potentiels inhibiteurs et substrats sont tous les deux intéressants.

3.1.3. Les méthodes QSAR

3.1.3.1. Définition

Les méthodes QSAR (*quantitative structure activity relationship*) permettent de prédire l'activité biologique d'un ensemble de molécules grâce à l'analyse d'une équation reliant la structure chimique et les activités expérimentales d'un autre groupe de molécules analogues (appelées groupe d'entrainement) [73,74]. L'équation de la forme $Ai = \hat{k}(D1, D2, \dots Dn)$, permet de mettre en relation ($\hat{k}$: transformation mathématique empirique) un ensemble pondéré de descripteurs représentant la structure chimique (D_1, D_2, …, D_n) avec l'activité biologique (Ai). L'équation mathématique modèle sera appliquée pour déterminer l'activité d'autres composés non testés expérimentalement.

Il existe différentes dimensions de QSAR en fonction des variables utilisées pour la construction des modèles : 1D pour les modèles basés sur des descripteurs constitutionnels, 2D pour ceux basés sur les indices de connectivités, 3D met en évidence les interactions non liées en utilisant des descripteurs de champs, 4D en rapport avec la corrélation entre les configurations des ligands, 5D en rapport avec la corrélation des modèles *induced-fit* en 4D QSAR et enfin 6D qui permet de décrire la variation de solvatation au niveau des 5D QSAR [75].

3.1.3.2. Construction d'un modèle QSAR

La construction d'un modèle QSAR suit 3 étapes : (1) Préparation et analyse des données, (2) développement du modèle et enfin (3) sa validation.

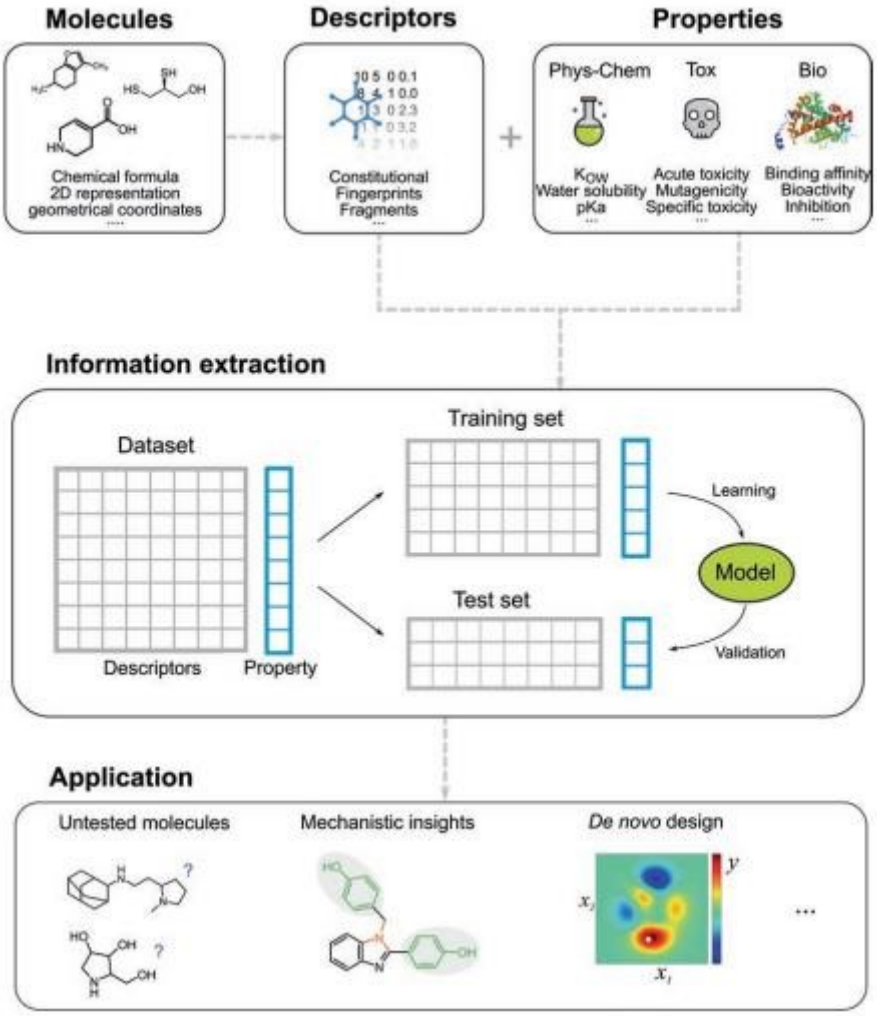


Figure 7 : Principales étapes de construction d'un modèle QSAR d'après [76]

3.1.3.2.1. Les descripteurs

Comme précédemment mentionné, un modèle QSAR met en relation un ensemble de descripteurs moléculaires et une activité ou une toxicité sous forme d'équation mathématique [7]. Une étape critique dans la construction d'un modèle QSAR est le choix et la sélection des bons descripteurs [77]. En effet, il existe de nombreux critères pour qu'un descripteur soit adapté. En particulier, il faut qu'il permette de corréler les caractéristiques moléculaires avec les propriétés physicochimiques décrivant au mieux l'activité biologique tout en ayant le minimum de corrélation avec les autres descripteurs [73,77].

Il existe différentes méthodes pour calculer les descripteurs moléculaires[29–36] et différentes façons de les catégoriser. Nous pouvons les classer en descripteurs calculés et expérimentaux, en descripteurs globaux, locaux et de champs mais la manière usuelle est de se baser sur la

dimensionnalité de la représentation structurale[78]. Il existe de nombreux logiciels permettant de générer des descripteurs comme CDK et Dragon [79,80].

Descripteurs 1D

Ces descripteurs aussi appelés descripteurs constitutionnels sont calculés à partir d'informations 1D de la molécule comme le poids moléculaire, le nombre d'atomes ou le nombre de groupes fonctionnels. [81]

Descripteurs 2D

Ces descripteurs sont dérivés à partir de la représentation 2D de la structure chimique aussi appelée graphe moléculaire. On trouve dans cette catégorie par exemple les empreintes moléculaires ou *fingerprints* (c.f. 3.1.1. recherche de similarité) ainsi que les descripteurs topologiques [38,82,83]. Ces derniers sont basés sur la représentation du graphe moléculaire permettant d'intégrer plusieurs informations relatives à la taille, au degré de ramifications ou même à la flexibilité de la molécule. Ainsi, le graphe moléculaire est encodé en représentation matricielle à partir de laquelle sont calculés plusieurs descripteurs. Par exemples les indices topologiques de Zagreb[83] sont calculés à partir des matrices d'adjacence (matrices indiquant les différentes connectivités existantes entre les atomes d'une même molécule).

Descripteurs 3D

Cette catégorie comprend par exemple les descripteurs géométriques et les descripteurs de surface. Les descripteurs géométriques incluent par exemple l'ovalité ou le volume moléculaire. Les descripteurs de surfaces décrivent aussi bien l'aire de la surface moléculaire que la distribution des charges comme l'aire de la surface polaire (PSA, *polar surface area*), la surface de van der Waals, la surface moléculaire accessible au solvant ou encore le moment dipolaire ou la polarisabilité.

Il est à noter que les descripteurs 1D sont le plus souvent utilisés pour filtrer la base de données et pour construire des modèles QSAR plutôt que pour mesurer les similarités entre molécules. A contrario, les descripteurs 2D et 3D peuvent être utilisés pour une recherche de similarité ainsi que pour les méthodes QSAR [81].

3.1.3.2.2. Préparation et analyse des données

Tout d'abord, les données doivent être préparées en suivant les règles de bonnes pratiques relatives à la standardisation des structures [84]. Par la suite, une étude statistique descriptive doit être menée pour vérifier que la distribution des données dans l'espace chimique est homogène mais surtout pour sélectionner les descripteurs les plus pertinents.

Une fois les données prêtes et les descripteurs générés et sélectionnés, les données d'entrée vont être séparées en 2 parties : une partie regroupée sous le terme jeu d'entrainement (*training set*), est utilisée pour construire le modèle et la deuxième partie est utilisée pour former le jeu de test (*test set*) qui permet de tester le pouvoir prédictif du modèle. Cette division en jeux d'entrainement et de test vise à éviter un phénomène récurrent en modélisation : le surentrainement ou l*'overfitting*. Ainsi, cela permet de prouver qu'un modèle est capable de prédire aussi bien les données utilisées pour le construire que des données qu'il n'aura jamais « vu ». Un modèle qui est validé pourra donc être utilisé pour prédire l'activité de molécules pour lesquelles nous ne possédons pas de données expérimentales.

3.1.3.2.3. Développement du modèle

Traditionnellement les premiers modèles de QSAR suivaient un modèle linéaire (de régression linéaire) [75]. Grâce aux progrès de l'apprentissage automatique (*Machine Learning*), des modèles plus sophistiqués ont été mis au point. Nous pouvons distinguer les modèles linéaires comprenant la régression linéaire (*Linear Regression*), l'analyse en composantes principales (PCA), ou la régression des moindres carrés partiels (*Partial Least Square regression*) et les modèles non linéaires comme les k plus proches voisins (*k-Nearest Neighbours*), les réseaux de neurones artificiels (*Artifical Neural Networks*), les arbres de décisions (*Decision Trees*), les forêts d'arbres de décision (*Random Forest*) ou les machines à vecteurs de support (*support vector machines SVMs)*. En fonction de la nature du problème, ces méthodes peuvent être classées en méthodes de classification (prédire deux classes : active et inactive) et méthodes de régression (prédire les valeurs d'activités biologique).

3.1.3.2.4. Validation

Il existe deux types de validation : (1) validation interne qui implique l'utilisation des données d'entrainements et (2) une validation externe qui se base sur les données de test. Pour la validation interne, différentes méthodes de cross-validation peuvent être employées. Ainsi, le *leave one ou*t (LOO) ou *leave many out* (LMO) consistent à retirer respectivement une ou plusieurs molécules du jeu d'entrainement. Le *Boostrap* consiste à rééchantillonner le jeu d'entrainement au moyen de tirages avec remise. Le but de ces méthodes est de générer des sous-ensembles et de contrôler la performance "dessus".

L'étape de validation se fait au moyen de métriques statistiques permettant d'apprécier les performances du modèle. Pour les méthodes de classification, la sensibilité, la spécificité, l'exactitude et la précision peuvent être calculées. Pour les méthodes de régression, des

métriques comme l'erreur absolue moyenne (MAE), l'erreur quadratique moyenne (RMSE) et le coefficient de détermination (R^2) seront calculés. Notons aussi que pour que le jeu de test soit valide, il faut qu'il soit compris dans le domaine d'applicabilité du modèle. Cette notion centrale pour les modèles QSAR permet de déterminer la région de l'espace chimique pour laquelle les prédictions réalisées sont jugées fiables. Il existe différentes méthodes pour calculer le domaine d'applicabilité [85–88] et ce dernier devra être déterminé pour n'importe quel jeu de donné à prédire.

3.1.3.3. Place des méthodes QSAR en toxicologie

Les méthodes QSAR occupent une place importante en toxicologie computationnelle prospective.

Au-delà de l'industrie pharmaceutique, les méthodes QSAR sont largement employées dans d'autres industries ainsi que par des organismes réglementaires. En effet, ces derniers ont commencé à créer leurs propres bases de données de structures moléculaires ainsi que leurs effets pour de nombreux paramètres physiques et biologiques permettant ainsi la création de modèles sur des bases d'apprentissage plus importantes [74]. Ainsi, depuis l'introduction de mesures de protection contre les risques possibles des produits chimiques (réglementés aux États-Unis par l'EPA et dans l'Union européenne par la règlementation REACH) [10,89], les modèles de QSAR sont devenus un outil de routine en parallèle d'autres outils instrumentaux et de recherche. Toutefois, chaque modèle QSAR est limité dans sa prédictivité à une famille structurale unique et pour un seul effet toxique étudié (effet biologique) [9,12].

3.2. Criblage virtuel basé sur la structure des cibles

Avec le nombre croissant des structures de protéines enregistrées au niveau de la P*rotein Data Bank* (PDB) [90], les méthodes SB se positionnent aussi comme un outil valide en toxicité *in silico*. Les coordonnées 3D des cibles peuvent être résolues expérimentalement ou modélisées à l'aide d'algorithmes dédiés et deux cas de figures existent : 1) le site actif de la protéine est connu et la méthode SB choisie peut donc être appliquée directement ou 2) le site actif n'a pas encore été caractérisé et il faudra le prédire en amont pour pouvoir appliquer la majorité des méthodes SB.

3.2.1. Obtention de la structure 3D

La structure 3D de la protéine cible étudiée peut être résolue expérimentalement et stockée dans la base de données de référence, la *Protein Data Bank* (PDB) [90], ou bien obtenue par des méthodes de prédiction *in silico.*

3.2.1.1. Résolution expérimentale de la structure 3D

Afin d'obtenir les coordonnées 3D d'une protéine, 3 méthodes expérimentales peuvent être employées, à savoir la cristallographie aux rayons X, la résonnance magnétique nucléaire (RMN) et la cryomicroscopie électronique (Cryo-EM).

3.2.1.1.1. Cristallographie aux rayons X

La cristallographie aux rayons X est la méthode la plus populaire et la plus utilisée pour élucider la structure 3D des protéines [91]. Le principe de la méthode est d'obtenir une carte représentant la distribution de densité des électrons (liée à la position atomique) à partir des données de diffraction de rayons X par un cristal de la protéine. La première étape, une étape clé et facteur limitante pour la réussite de cette méthode, est donc de réussir à isoler un cristal à partir d'une solution sursaturée dans des conditions expérimentales optimales (réactif, pH, température et additifs). Une fois le cristal isolé, un faisceau de rayons X est dirigé vers le cristal et diffracté par les électrons ordonnés de la molécule. Les motifs de diffraction pour des orientations différentes du cristal seront ainsi enregistrés au moyen de capteurs pour obtenir une carte de densité électronique. Enfin, une sonde protéique, i.e. un enchainement d'acides aminés, sera insérée au niveau des maillages de la carte (étape de *fitting*) et son positionnement sera optimisé (étape de *refinement*).

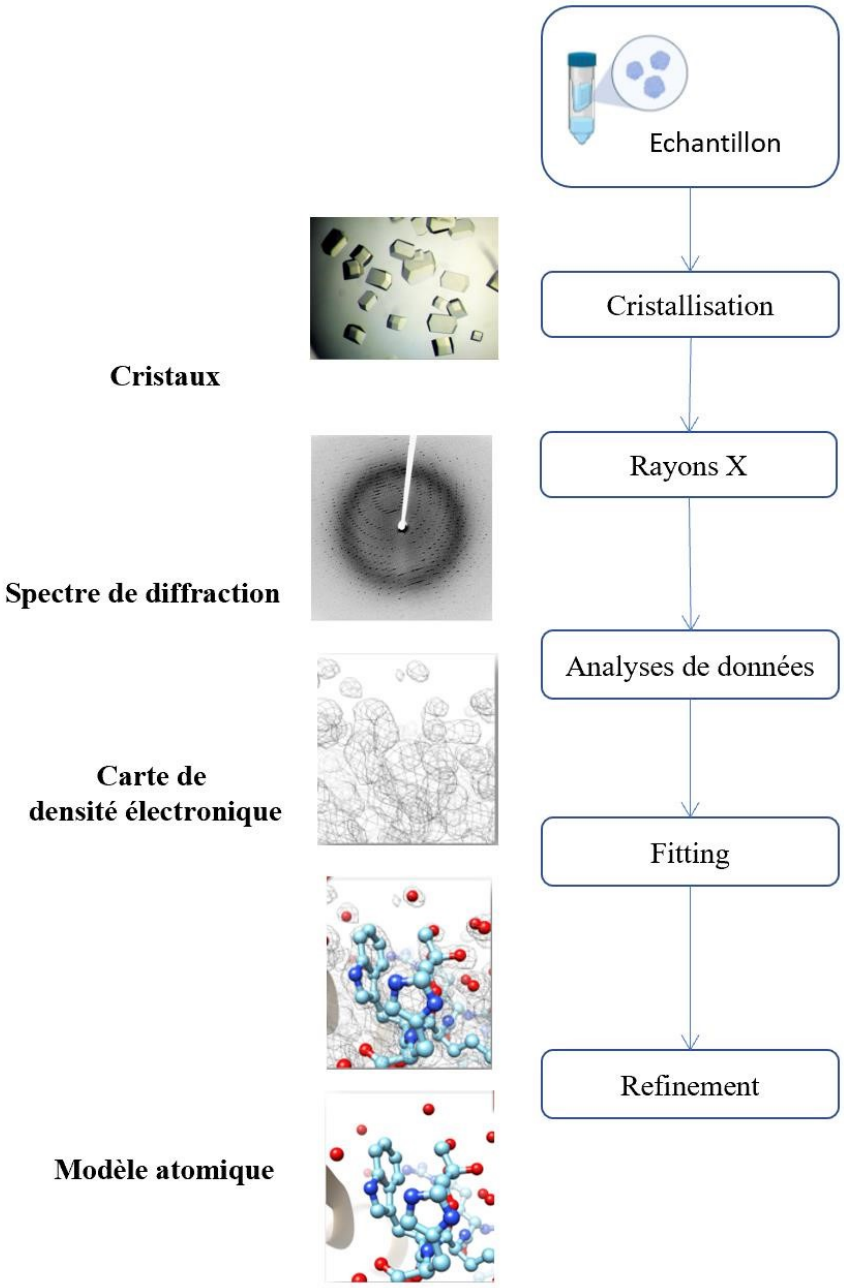


Figure 8 : Etapes de résolution de la structure 3D des protéines par la méthode de cristallographie aux rayons X

3.2.1.1.2. Spectroscopie à résonance magnétique nucléaire (RMN)

La RMN permet de mesurer de manière très fine les champs magnétiques régnants dans la matière. Cette technique se base sur les états de spins particuliers que possèdent certains noyaux atomiques comme l'hydrogène ou le carbone 13 ou encore le fluor. Ces derniers peuvent ainsi absorber l'énergie émise par un rayonnement électromagnétique et l'émettre à une fréquence très précise dépendante du champ magnétique appliqué [92]. Ainsi, pour résoudre une structure, une protéine purifiée est diluée dans une solution aqueuse qui sera mise dans une sonde RMN.

La sonde est exposée à un champ magnétique fort entrainant la résonance de certains noyaux (les ^{1}H, ^{13}C et ^{15}N)[93]. Ainsi les fréquences de résonnance sont mesurées et enregistrées sous forme d'un spectre. C'est l'environnement de chaque atome qui détermine si ces fréquences seront hautes ou faibles. En utilisant des méthodes computationnelles, les mesures faites seront converties en graphiques représentant les fréquences comme des pics correspondant à une localisation spécifique d'un groupe d'atomes appelés contraintes de distances et de géométrie. Cette information est optimisée et combinée avec d'autres expérimentations RMN pour déterminer la structure 3D répondant au mieux aux contraintes collectées. La RMN possède l'avantage majeur de s'affranchir de l'étape de cristallisation permettant à la protéine d'adopter plusieurs conformations qui seront utilisées pour générer différents spectres conduisant à plusieurs structures plausibles.

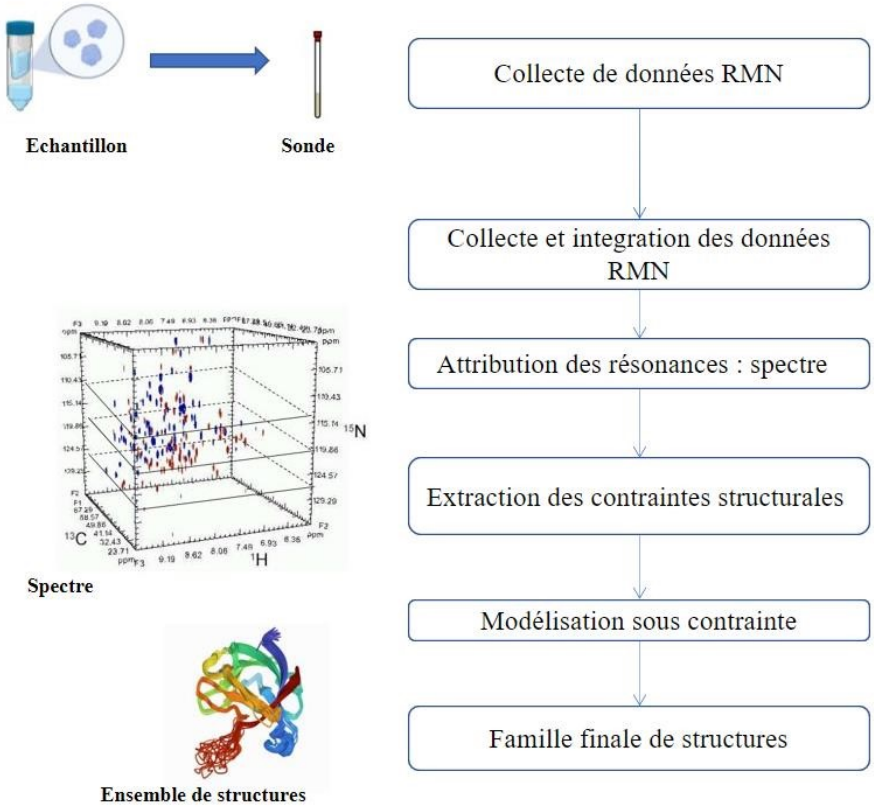


Figure 9 : Etapes générales de la résolution de structure de protéines par spectroscopie RMN d'après [93,94]

3.2.1.1.3. Cryomicroscopie électronique

En 2017, l'avancée majeure apportée par la cryomicroscopie électronique dans la détermination des structures de protéine a été saluée par l'attribution d'un prix Nobel de chimie [95]. En effet, cette technique d'imagerie permet d'obtenir de façon directe des images agrandies de gros

complexes de protéines à une très haute résolution par rapport aux méthodes de cristallographie aux rayons X et RMN. Pour cette méthode, une petite quantité de protéine purifiée est étalée sur une grille de cuivre. Ensuite, l'échantillon est plongé rapidement dans de l'éthane liquide pour immobiliser les molécules (on utilise l'éthane pour éviter la formation de cristaux d'eau congelée qui ajouterait du bruit au signal). En effet, sans préparation préalable, l'échantillon est sensible au faisceau d'électrons et au vide qui entrainent la dénaturation des protéines et sa dégradation [96]. Enfin, l'échantillon est chargé dans un microscope électronique. Dans le microscope, l'échantillon est exposé à un faisceau d'électrons accélérés et les images issues du microscope sont capturées. Dans l'échantillon les protéines adoptent différentes orientations et l'ensemble des images générées par le microscope sont combinées par orientation et intégrées pour en déduire une carte de densité électronique. Une étape de *fitting* vient finaliser le protocole, au cours de laquelle les atomes sont raccordés dans la carte obtenue pour en déduire la structure 3D [95,97].

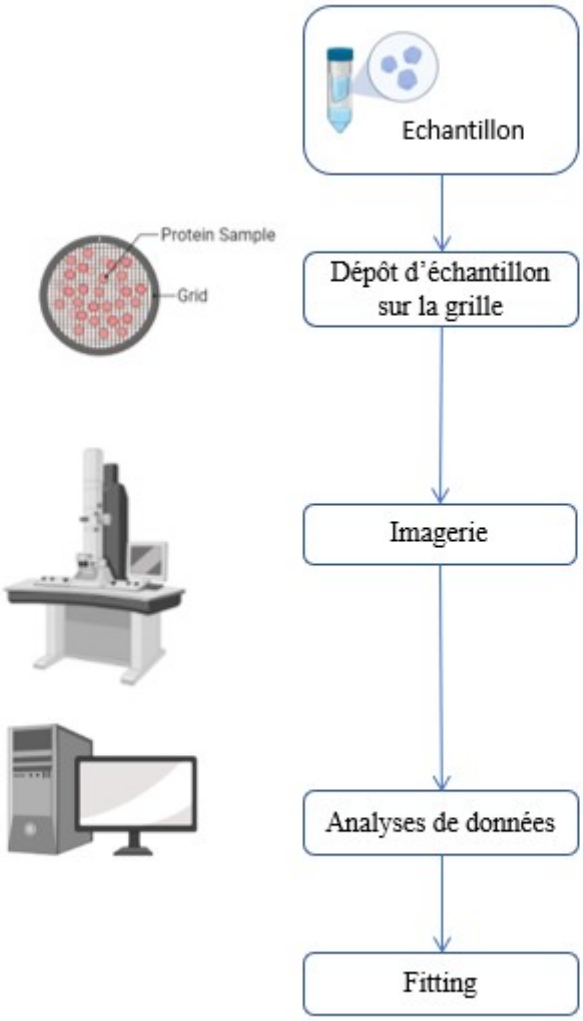


Figure 10 : Etapes générales de la résolution de structure de protéines par Cryo-EM

3.2.1.1.4. Prédiction de la structure 3D *in silico*

La structure d'une protéine peut être prédite computationnellement soit 1) en partant de structures de protéines homologues en appliquant la théorie de l'évolution (modélisation par homologie ou par enfilage) soit 2) grâce à des méthodes de modélisation en appliquant les lois de la physique (modélisation *Ab Initio*).

3.2.1.1.4.1. Modélisation par homologie

La modélisation part homologie par du principe qu'au cours de l'évolution, les structures sont plus conservées que les séquences elles-mêmes. Par conséquence, deux protéines ayant une forte identité de séquence ont de fortes chances d'avoir des structures similaires. Cette méthode

de modélisation s'articule en trois étapes. D'abord, les séquences similaires sont identifiées grâce à des approches séquence-séquence (BLAST) ou profile-séquence (PSI-BLAST) [98] ou profil-profil (HHsearch [99] ou ORION [100]). Ensuite leurs séquences seront alignées à celles de la protéine d'intérêt dans le but de trouver l'alignement optimal ce qui permet de transférer les coordonnées atomiques du modèle à la séquence protéique étudiée. Enfin, les boucles et les chaines latérales sont modélisées et le modèle final est optimisé pour éliminer les encombrements stériques et minimiser l'énergie globale. De nombreux outils permettent de réaliser la modélisation par homologie comme Modeller [101] ou encore Rosetta qui combine les méthodes d'homologie et les méthodes *Ab Initio* [102] .

3.2.1.1.4.2. Modélisation par enfilage

Connue sous le nom de threading, cette méthode consiste à enfiler la séquence de la protéine à étudier dans des structures protéiques connues. Plus concrètement, il s'agit d'aligner des portions de la protéine sur des structures provenant des bases de données comme la PDB et d'en calculer un score basé sur la compatibilité des structures secondaires et notamment le potentiel de mutation. Ainsi, les structures pour lesquelles l'enfilage va donner le meilleur score pourront être identifiées. De la même manière que pour la modélisation par homologie, la structure obtenue sera optimisée en termes d'énergie (minimisation d'énergie).

3.2.1.1.4.3. Modélisation Ab Initio

Cette méthode consiste à échantillonner les repliements d'une protéine en s'affranchissant de tout a priori structural. En effet, le mécanisme de repliement a été au centre de nombreuses études (Levinthal [103] et Afinsen [104]) et les méthodes Ab Initio s'appuient aujourd'hui sur le principe que l'information sur le repliement d'une protéine est contenue dans sa structure. Le modèle 3D de la protéine sera ainsi construit sur la base de la séquence, par simulation des forces qui gouvernent le repliement, pour trouver la structure de plus basse énergie. Cette méthode de modélisation est utilisée plutôt pour des petites portions de protéines car il est difficile d'atteindre des minima globaux sur une protéine entière. Aujourd'hui, grâce à l'avènement des super-ordinateurs ainsi que des méthodes d'échantillonnages, des structures de plus de 100 acides aminés ont pu être résolues avec succès [105–107]. Par ailleurs, des algorithmes de Deep Learning ont permis de créer une véritable révolution grâce l'essor d'AlphaFold1 en 2018 et sa version améliorée Alphafold2 en 2020 ce qui a permis d'accélérer la prédiction des structures 3D de protéines à partir de leurs séquences [108,109]. Les deux versions d'AlphaFold suivent la même logique et s'articulent en deux étapes. La première étape implique

l'utilisation d'algorithmes de *deep learning* pour prédire une matrice de distances et de distribution des angles de torsions entre les différentes paires d'acides aminés de la protéine. Pour ce faire, une multitude de données est fournie à l'algorithme pour l'enrichir et pour qu'il puisse apprendre au mieux. Les informations comprennent notamment des données physicochimiques sur les relations entre les AA ainsi que des données issues de protéines déjà connues et sont fournies sous forme d'images à une centaine de canaux. La deuxième étape permet à partir de la matrice de distance de déterminer les positions dans l'espace des différents résidus puis d'optimiser la structure de la protéine en minimisant son énergie en employant une méthode de descente de gradient. La différence entre les deux versions réside dans la première étape et l'algorithme utilisé. En effet AlphaFold1 s'appuie sur des réseaux de convolution [108] et AlphaFold 2 utilise plutôt des réseaux avec des mécanismes d'attention [109,110] .

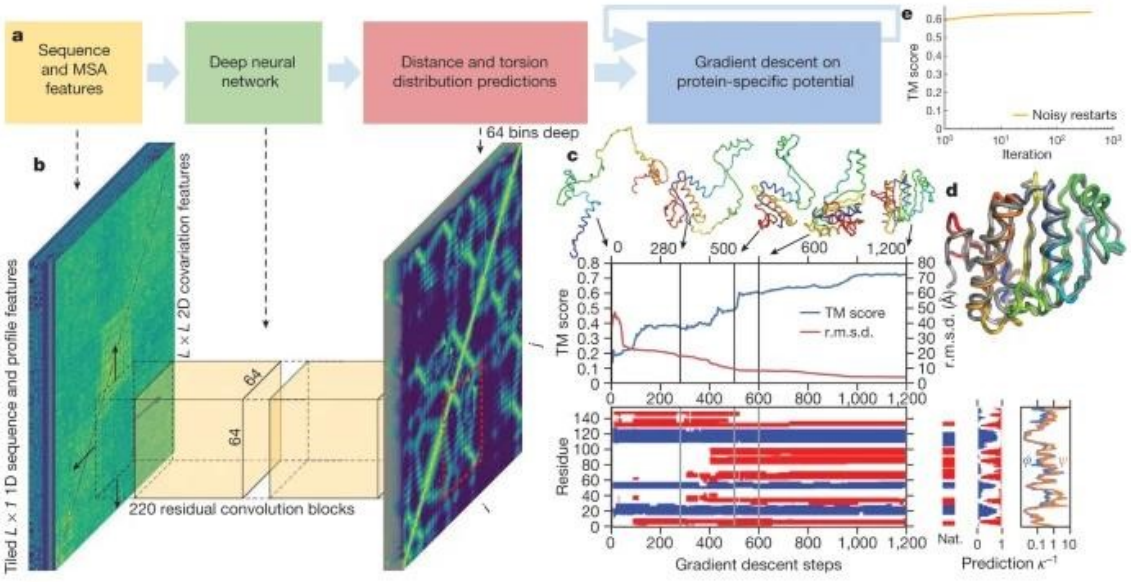


Figure 11 : Prédiction du repliement des protéines tel que présenté par AlphaFold1d'après [108]

Ainsi, AlphaFold2 a permis de prédire de nombreuses structures jusqu'alors non résolues ce qui a conduit à la création de la base de données libre du même nom : AlphaFoldDB [111]. Cette dernière a permis une expansion sans précédent de la couverture structurale de l'espace des séquences protéiques connues (plus de 360 000 structures prédites pour les protéomes de 21organismes modèles) [111].

3.2.2. Outils de prédiction du site de liaison

Une fois la structure de la protéine résolue, deux cas de figures sont possibles ; 1) la structure a été résolue en complexe avec un ligand (structure *holo*). Ce dernier sera ainsi utilisé comme

référence/sonde pour définir un site de liaison tout autour. 2) la structure isolée est *apo* i.e. isolée sans ligand. Dans ce cas-là, il est possible d'utiliser des méthodes de prédictions de site de liaison. Notons que ces méthodes peuvent aussi être employées lorsqu'on cherche à identifier un autre site de liaison différent de celui associé au ligand de référence. Trois catégories de méthodes existent pour prédire le site de liaison : les méthodes basées sur les connaissances (*knowledge based*), celles sur la géométrie et celles sur l'énergie d'interaction. Notons que le site de liaison peut aussi être déterminé à partir de données expérimentales qui identifieent les résidus impliqués dans l'interaction d'un ligand avec la protéine.

3.2.2.1. Méthodes basées sur les connaissances

Ces méthodes se basent sur des données biochimiques, de mutagenèse dirigée et de similarité de séquence ou de structure. La similarité de séquence se base sur l'idée que deux protéines avec des séquences similaires possèdent une forte probabilité d'avoir des sites de liaisons similaires [112]. Ainsi, en partant de bases de données comme PROSITE [113] qui regroupe des données sur les sites de liaisons et leurs séquences associées, des études d'homologie peuvent être réalisées pour définir le site de liaison de la protéine étudiée (exemple ConSurf [114]). En ce qui concerne les méthodes de similarité de structures, ces dernières se basent sur deux fondements. Le premier est que les sites de liaisons peuvent adopter un nombre réduit de formes [115]. Le deuxième fondement est que deux protéines exerçant la même fonction biologique présentent généralement une similarité structurale. Ainsi, à partir de banques de données contenant des informations structurales sur les sites de liaison expérimentalement déterminé [116–120], un alignement structural peut être réalisé (par exemple avec l'outil TM-align [121]) avec la protéine étudiée pour en déduire un site de liaison potentiel. Enfin notons qu'il existe des outils comme COFACTOR [122] qui allient à la fois les méthodes de prédictions de structure et basés sur la similarité.

3.2.2.2. Méthodes basées sur la géométrie

Le postulat de départ de ces méthodes est que les cavités et les poches présentes au niveau d'une structure protéique sont généralement associées à un site de liaison [123,124]. Afin d'identifier ces poches, de nombreux outils et approches existent. Un exemple est l'outil POCKET[125] permet de générer la surface protéique et de la placer dans une grille cartésienne. Les informations extraites de cette grille seront utilisées pour générer la topologie de la protéine et identifier des poches. D'autres approches permettent de s'émanciper de l'utilisation des grilles en se basant sur l'utilisation des sondes sphériques « déroulées » à la surface de la protéine pour identifier

des zones concaves correspondant aux poches (exemple SURFNET [126]) ou encore les méthodes utilisant le diagramme de Voronoï et les triangulations de Delaunay conduisant à une représentation α-*Shape* de la protéine. A partir de cette représentation, les arrêtes qui se trouvent à l'extérieur de la protéine sont soustraites correspondant ainsi aux poches [127].

3.2.2.3. Méthodes basées sur les énergies

Ces méthodes permettent d'estimer les énergies d'interaction entre une sonde (groupe méthyl, hydroxyle ou amine) et un point donné de la protéine. Les sondes en interagissant avec des points de la molécule permettent d'imiter les propriétés communes des petites molécules comme les propriétés hydrophobes ou donneur/accepteur d'hydrogènes. Ainsi, cela permet de définir une zone favorable d'interactions par génération de cluster contenant les zones où les sondes se lient avec des énergies favorables [128,129]. Il existe différents outils se basant sur l'énergie pour la prédiction des poches comme VolSite [120] qui utilise 7 types de sondes (donneur/accepteur de liaison H, donneur de liaison H, accepteur de liaison H, chargé (+), chargé (-), aromatique et hydrophobe. Pour cet outil, la protéine est placée dans une grille dont chaque point est sondé. Cela permet de définir des zones « IN » i.e à l'intérieur de la protéine et des zones « OUT » i.e. hors de la protéine. Enfin, des propriétés sont attribuées aux zones « IN » correspondant aux images négatives des types de sondes associées.

3.2.3. Outils de criblage

3.2.3.1. Modélisation de pharmacophores

Un pharmacophore décrit l'arrangement de l'ensembles des propriétés physicochimiques et électroniques nécessaires pour établir une interaction entre un ligand et une protéine [130]. Dans le cas de l'approche SB, un pharmacophore pourra être généré à partir d'une structure protéique en complexe avec un ligand (structure *holo*) ou bien à partir d'une structure non complexée avec un ligand (*apo*). Ainsi les modèles de pharmacophore SB sont riches en information relative aux interactions potentielles entre un ligand et la cible étudiée. Par conséquent, et au-delà d'être un outil de criblage prisé, une application directe de ces modèles serait de les utiliser pour la prédiction et/ou pour le classement (*ranking*) des poses de docking [131].

3.2.3.1.1. Modèle pharmacophorique à partir d'un complexe ligand-récepteur

Dans ce cas de figure, les informations/profils d'interaction entre le ligand et sa protéine récepteur sont utilisées pour définir les points pharmacophoriques constituants le modèle et ces

derniers sont généralement dérivés à partir de complexes présents dans la PDB. Enfin, pour s'assurer de la robustesse et de la fiabilité d'un pharmacophore généré, il est nécessaire de regrouper les informations provenant de plusieurs complexes pour la même protéine et de vérifier la pérennité des points pharmacophoriques entre les différents pharmacophores issus de ces complexes. Ces derniers peuvent 1) être des structures 3D d'une même protéine au niveau de la PDB (soit complexé avec le même ligand ou pas) ou 2) provenir de conformations issues d'une simulation de dynamique moléculaire. Au cours des travaux relatifs à cette thèse, le logiciel LigandScout a été utilisé pour générer des pharmacophores SB et la procédure est détaillée dans la partie résultats de cette thèse.

3.2.3.1.2. Modèle pharmacophorique à partir de récepteur non complexé avec un ligand

Les modèles de pharmacophores SB dérivés à partir d'une structure en complexe avec un ligand sont limités en nombre [130]. Les conséquences de cette situation sont que (i) un nombre restreint d'options est disponible pour construire des modèles de pharmacophores pour des cibles lorsqu'aucun ligand de liaison n'est connu, et (ii) lorsque seule une petite quantité de ligands est connue, les pharmacophores qui en résultent sont limités aux interactions que ces molécules forment ne permettant pas d'échantillonner la totalité de l'espace pharmacophorique [132]. Ainsi, les méthodes pharmacophoriques se basant sur les structures Apo s'imposent surtout au vu du nombre croissant de structures découvertes régulièrement parfois même sans informations préalables sur leurs fonctions [132]. Pour ce faire, différentes approches peuvent être utilisées. Il existe par exemple des méthodes basées sur **l'alignement des séquences**. Le pharmacophore est généré à partir de résidus clés identifiés par alignements localisés au niveau du site d'interaction [133]. Ces résidus seraient conservés tout au long de l'évolution pour 2 raisons possibles : ils sont soit nécessaires à l'interaction avec le ligand, soit apportent de la spécificité. Une autre approche consiste à simuler le **comportement dynamique** de sondes chimiques (eau et solvants organiques) sur une surface moléculaire flexible[134,135]. Les molécules sondes minimisées permettent ainsi de révéler des sites d'interaction favorables à la surface de la protéine, qui peuvent être convertis en points pharmacophoriques par la suite. Bien que plus longue, cette approche peut être efficace lorsque l'eau, le solvant naturel des protéines, est utilisée comme sonde. Cependant, les sondes organiques utilisées pour détecter les régions hydrophobes d'une macromolécule créent un environnement non naturel pour la plupart des protéines, induisant *in silico* des conformations qui ont peu de chances d'être observées *in vivo*. La méthode la plus populaire pour obtenir des pharmacophores *apo* reste celle basée sur les

grilles tridimensionnelles [132,135]. Il s'agit de placer la zone d'interaction dans une grille (boîte) et d'y calculer et identifier les points d'énergies les plus favorables et de les grouper pour en déduire les différents points pharmacophoriques associés aux régions d'interactions. Il est à noter que la génération de la grille peut soit être définie par l'utilisateur comme le cas du logiciel Pocket v2 [136], soit générée automatiquement au moyen d'algorithmes de détection dédiés comme le cas des outils Ph4Dock[137] ou LigandScout [56]. Une fois la grille placée et les énergies calculés, les points d'énergies sont examinés afin d'éliminer ceux qui sont éloignés du site et ceux ayant une énergie insuffisante. Enfin, les différents points sont regroupés en clusters définissants une région favorable pour l'interaction : point pharmacophorique.

3.2.3.1.3. Apport de la dynamique moléculaire

3.2.3.1.3.1. Pharmacophore dynamique

Le concept de pharmacophore dynamique (dynophore) consiste à coupler la génération de pharmacophore SB à la dynamique moléculaire. Cela permet de prendre en considération la flexibilité du site de liaison- en plus de celle du ligand- et d'échantillonner ainsi l'espace conformationnelle de la protéine à la recherche de motifs d'interactions encore non décelées par les approches classiques i.e. protéine rigide [50,138]. D'un point de vue pratique, il s'agit d'analyser les différentes interactions ligand-protéine et extraire les points pharmacophoriques correspondants à partir d'une trajectoire de dynamique moléculaire. Les différents points extraits sont ainsi statistiquement analysés et classés entre eux sur la base des différents motifs d'interactions et afin de générer le ou les meilleurs modèles. Différentes approches ont été proposées pour optimiser les différents modèles extraits à partir d'une trajectoire dont l'approches des hits communs ou « *common hits approach* » (CHA)[139] ou l'approche MYSPACE[140].

3.2.3.1.3.2. Hydration-site resctricted pharmacophore

Ce type de pharmacophore est extrait à partir de la trajectoire de dynamique moléculaire d'une structure *apo.* L'idée est que lorsqu'un ligand se lie à la protéine, les groupes fonctionnels remplaçant les molécules d'eau contribuent fortement à l'affinité de liaison du ligand à cause du gain en énergie libre. Ainsi, au cours d'une simulation de dynamique moléculaire, chaque site d'hydratation est soumis à une analyse thermodynamique pour en calculer l'énergie libre[141]. En parallèle, des modèles pharmacophoriques sont générés et ceux qui sont co-localisés avec les sites d'hydratation à impact favorable pour l'énergie libre de liaison sont sélectionnés pour

former un modèle réduit. Cette approche permet de cribler des hits plus pertinents puisqu'entropiquement favorables ce qui se traduit par un bon enrichissement et une forte efficacité [50]. De plus, cette approche conduit à une réduction du temps de criblage d'un facteur 200-500 comparé aux approches classiques parcourant la totalité de la protéine.

3.2.3.2. Docking protéine-ligand

Le docking ligand-protéine est une méthode utilisée pour prédire la capacité d'un composé à interagir avec une cible protéique [142]. Le premier modèle d'interaction substart-enzyme a été introduit par Fisher [143,144]. Il s'agit du modèle clé-serrure qui suggère une reconnaissance et une interaction entre deux corps rigides et qui a inspiré les méthodes de docking dites rigides. Koshland introduit par la suite la notion d'ajustement ou *induced fit* qui décrit l'adaptation conformationnelle du ligand et de la protéine lors de leurs interactions [144]. Ainsi, pour tenter de prendre en compte cette notion, les méthodes de docking ont évolué et à ce jour les méthodes les plus efficaces sont celles réalisées avec un ligand dit flexible et une protéine rigide. Néanmoins, il est possible de prendre en compte la flexibilité de la protéine d'une façon explicite ou implicite, totale ou partielle [145] (c.f. 3.2.3.2.3. Prise en compte de la flexibilité).

Les méthodes de docking sont généralement constituées de deux étapes clés. La première appelée **échantillonnage** consiste à prédire les différentes poses que le ligand peut adopter au niveau d'un site de la protéine (sa conformation et sa position). La deuxième étape consiste à calculer un **score** pour chaque pose générée, les poses associées aux meilleurs scores sont celles qui seraient les plus probables pour le ligand étudié. Ainsi le docking peut être utilisé pour formuler des hypothèses sur le mode d'interaction dominant d'un composé avec une protéine et pour cribler une chimiothèque et classer les composés selon leur score de docking associé [146].

3.2.3.2.1. Etape d'échantillonnage

Une petite molécule est par définition flexible car elle possède 3N degrés de liberté (**Figure 12**) : 3 mouvements de translation, 3 de rotations et 3 de vibration/torsion.

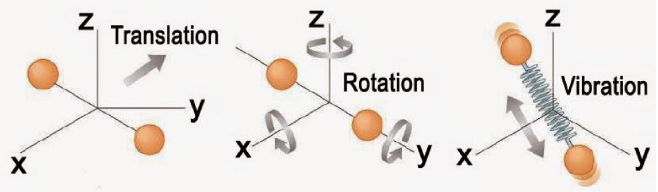


Figure 12 : Les 3N degrés de liberté d'une molécule d'après [147]

3.2.3.2.1.1. Docking rigide

Les premiers algorithmes de docking ne prenaient en considération que 6 degrés de libertés (3 translations et 3 rotations) pour la recherche conformationnelle. Cette dernière se faisait de façon indépendante du récepteur puis les conformations générées étaient insérées au niveau du site de liaison en éliminant celles qui ne pouvaient pas le complémenter. C'est le cas de certains outils comme DOCK [148], FLOG [149] et FRED [150].

3.2.3.2.1.2. Méthodes d'échantillonnage

Comme décrit plus haut, en plus des 6 degrés de liberté considérés par les algorithmes de docking rigide, une molécule possède d'autres degrés de libertés translationnels liés aux liaisons rotatives (torsions) (c.f.**Figure 12**). Ainsi, obtenir toutes les conformations possibles d'une molécule et en déduire les potentielles interactions avec le récepteur devient une tache assez coûteuse tant en temps qu'en ressources computationnelles à mesure que la taille de la molécule augmente. Il est possible d'utiliser des programmes de docking protéine-ligand qui se basent sur des algorithmes de recherche conformationnelles divisés en 2 groupes : les algorithmes stochastiques, les algorithmes de recherche systématiques.

3.2.3.2.1.2.1. Méthodes déterministes ou de simulation

Dans cette catégorie de méthodes, on compte les méthodes dites de simulations et les méthodes de minimisation. **Les méthodes de dynamique moléculaire** permettent de calculer la trajectoire d'un système par l'application de la mécanique newtonienne. Les forces sont calculées sur chaque atome à partir du faible changement d'énergie potentielle entre la position actuelle et la nouvelle position. Ces forces atomiques et les masses correspondantes des atomes sont utilisées pour déterminer les positions atomiques sur une série de petits pas de temps en intégrant la deuxième loi du mouvement de Newton [151]. On obtient ainsi une trajectoire des changements de positions atomiques dans le temps. Le problème de la simulation de la dynamique moléculaire est qu'elle est généralement incapable de franchir des barrières à haute énergie dans une période de simulation praticable. Par conséquent, la simulation de dynamique moléculaire peut localiser les ligands dans des minima locaux. Le complément d'autres méthodes suivies (comme le recuit simulé) par la simulation de dynamique moléculaire peut fournir de meilleurs résultats. En 1999, Mangoni et ses collègues [152] ont décrit un protocole de dynamique moléculaire pour simuler l'interaction de petits ligands flexibles à des cibles flexibles dans l'eau. Le mouvement du centre de masse du ligand et ses mouvements internes et rotationnels ont été séparés et couplés à différents bains de température. Plus tard, on a introduit l'approche du complexe relaxé qui tourne autour des conformations de liaison qui ne

peuvent se produire que rarement dans les protéines cibles non liées. Dans cette approche, la simulation de dynamique moléculaire de la cible sans ligand est effectuée pendant 2 ns, puis le docking du ligand est réalisé. Outre la dynamique molécules, **les méthodes de minimisation** peuvent aussi être employées. Les techniques de minimisation de l'énergie conduisent à des minima énergétiques locaux, c'est pourquoi ces méthodes sont souvent utilisées en complément d'autres méthodes de recherche. Par exemple, le programme DOCK procède à une étape de minimisation successive à chaque ajout de fragment, suivie d'une minimisation finale avant l'étape de scoring.

3.2.3.2.1.2.2. Méthodes de recherche conformationnelle

Les méthodes de recherche conformationnelles permettent comme leur nom l'indique de faire une recherche de l'espace conformationnel du ligand. Nous distinguons deux types d'algorithmes : les algorithmes de recherche systématique ou exhaustifs et les algorithmes stochastiques ou aléatoires.

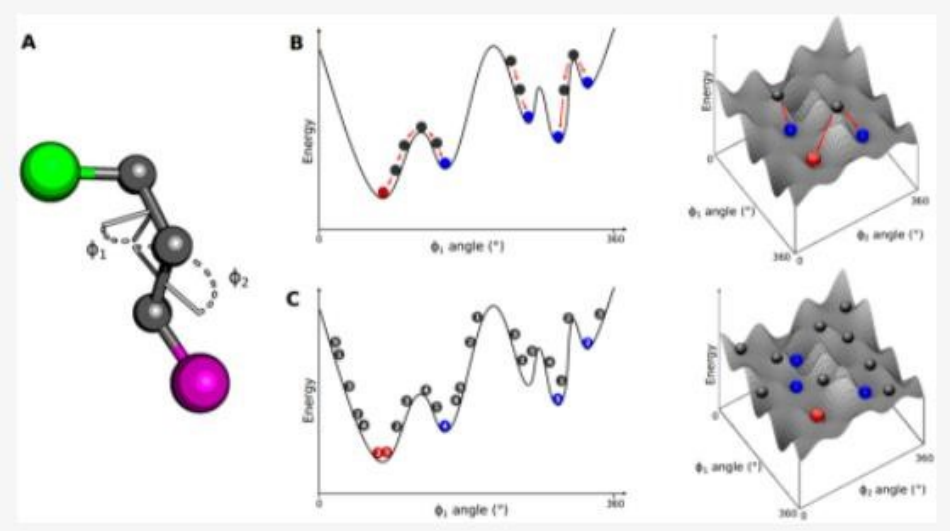


Figure 13 : Méthodes de recherche conformationnelle d'une petite molécule (A). Les sphères grises correspondent aux énergies initiales des structures, les rouges aux minima globaux et en bleu aux minima locaux. La courbe en (B) illustre les méthodes de recherches systématiques et en (C) les méthodes stochastiques qui augmentent les chances de tomber sur un minimum global. D'après [153]

Les méthodes de recherche systématique

Il existe des **algorithmes exhaustifs** qui sondent le profil énergétique de l'espace conformationnel et, après de nombreux cycles de recherche et d'évaluation, converge vers la solution associée au minima énergétique correspondant au mode de liaison le plus probable.

Cependant, cette méthode peut converger vers un minimum local plutôt que global (**Figure 13** B) [154]. Toutefois, cette recherche exhaustive fait que le nombre de combinaisons possibles croît de manière exponentielle à mesure que les degrés de liberté associés au ligand augmentent, ce qui entraîne un phénomène connu sous le nom d'explosion combinatoire. Afin de résoudre ce problème, différentes approches peuvent être utilisées. Par exemple, les programmes de docking tels que FRED, Surflex-dock et DOCK appliquent un **algorithme de construction incrémentale** dans lequel le ligand est progressivement construit dans le site de liaison [155–157]. Pour cela, la structure chimique est initialement décomposée en plusieurs fragments. Ensuite, l'un de ces fragments est sélectionné comme fragment d'ancrage et est inséré au niveau de la région complémentaire du site de liaison. Les fragments restants sont ajoutés séquentiellement. Le processus se poursuit jusqu'à ce que le ligand entier ait été reformé. L'algorithme effectue la recherche conformationnelle uniquement pour les fragments ajoutés, ce qui réduit les degrés de liberté à explorer et évite ainsi l'explosion combinatoire [158].

Les méthodes stochastiques

Les méthodes stochastiques effectuent la recherche conformationnelle en modifiant de manière aléatoire les paramètres structuraux des ligands [159]. Pour cela, l'algorithme génère des ensembles de conformations moléculaires correspondant à différents points au niveau de l'espace conformationnel (**Figure 13**C). Ces dernières seront acceptées ou rejetées en se basant sur une fonction de probabilité. Cette stratégie permet de prendre en considération la totalité de l'espace conformationnel et ainsi d'augmenter la probabilité de trouver un minimum global plutôt que de piéger la solution finale à un minimum énergétique local [153]. Malheureusement, elle implique aussi un grand cout computationnel. Plusieurs algorithmes stochastiques existent et les plus utilisées sont les **algorithmes de Monte Carlo** [160]et l'**algorithme génétique**.

L' algorithme génétique par exemple permet de résoudre le problème du coût de calcul élevé associé aux méthodes stochastiques en appliquant les concepts de la théorie de l'évolution et de la sélection naturelle. Dans un premier temps, l'algorithme code tous les paramètres structurels de la structure initiale dans un chromosome, qui est représenté par un vecteur. À partir de ce chromosome, l'algorithme de recherche aléatoire génère une population initiale de chromosomes couvrant une large zone du paysage énergétique. Cette population est évaluée et les chromosomes les plus adaptés (c'est-à-dire ceux qui présentent les valeurs d'énergie les plus faibles) sont sélectionnés comme modèles pour la génération de la population suivante. Cette procédure diminue l'énergie moyenne de l'ensemble des chromosomes en transmettant les caractéristiques structurales favorables d'une population à descendance, réduisant ainsi

l'espace conformationnel à explorer. La recherche par algorithme génétique est exécutée de manière récursive et, après un nombre défini de cycles de recherche et d'évaluation des conformations, elle converge vers une conformation (chromosome) correspondant au minimum énergétique global. Cet algorithme est intégré par exemple dans les programmes AutoDock [161] et Gold[162].

3.2.3.2.2. Etape de scoring

Une fonction de score permet d'évaluer l'énergie de liaison d'un complexe ligand-récepteur prédit. La prédiction de l'énergie de liaison est réalisée en évaluant les phénomènes physico-chimiques les plus importants impliqués dans la liaison ligand-récepteur, notamment les interactions intermoléculaires, la désolvatation et les effets entropiques [163]. Cependant, le coût de calcul augmente proportionnellement à ce nombre et afin de conserver des temps de calculs compatibles avec le criblage de larges bases de données, les fonctions de score doivent permettre d'assurer un équilibre entre la précision des résultats et la vitesse d'exécution. Les fonctions de score sont classées en trois groupes : les fonctions de score basées sur les champs de force, les fonctions de score empiriques et les fonctions de score basées sur les connaissances [163]. Chaque fonction de score possède des avantages et des limites. C'est pourquoi l'utilisation simultanée de différentes méthodes de scoring combinée pour obtenir un score consensus [164], est une approche intéressante pour combiner les avantages et atténuer simultanément les inconvénients de chaque méthode [165]. Ce score consensus peut être obtenu en combinant différentes fonctions de score choisies par l'utilisateur mais il existe aussi des fonctions de score consensus comme MultiScore, X-Cscore, Gfscore, SCS, SeleX-CS et CONSENSUS-DOCK [166–171].

3.2.3.2.2.1. Les fonctions de score basées sur le champ de force

Ce groupe de fonctions de score permet d'estimer l'énergie de liaison en additionnant les contributions des termes de liaisons reliés aux interactions covalentes dites liées (étirement de la liaison, flexion de l'angle et variation du dièdre) et les termes non liés (interactions électrostatiques et de van der Waals). Ce type de fonction de score applique une méthode ab initio pour calculer l'énergie associée à chaque terme de la fonction en utilisant les équations issues de la mécanique classique [172].

$$E = \sum_{Liaisons} K_L(r - r_0)^2 + \sum_{Valences} K_V(\theta - \theta_0)^2 + \sum_{Torsions} K_T[1 + \cos(n\phi - \phi_0)] + \sum_{VDW} 4\varepsilon_{ij}\left(\frac{\sigma_{ij}^{12}}{r_{ij}^{12}} - \frac{\sigma_{ij}^{6}}{r_{ij}^{6}}\right) + \sum_{Coulomb} \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}}$$

Tel que :

- K_L, K_V et K_T sont respectivement les écarts de distances de liaison, les écarts d'angles de valences et les écarts d'angles de torsions par rapport à la valeur idéale.
- Les couples r, θ et ϕ représentent respectivement les distances, les angle de valences et les angles de torsions mesurés
- Les couples r_0, θ_0 et ϕ_0 représentent respectivement les distances, les angle de valences et les angles de torsions de référence
- ε_{ij} : profondeur du puit du potentiel de Lennard Jones
- $q_i q_j$: charge coulombienne
- ε_0 : Constante diélectrique
- σ_{ij} : Le rayon de van der Waals

Une limitation majeure des méthodes basées sur les champs de force est leur imprécision dans l'estimation des contributions entropiques. Cette lacune est due à l'absence d'un modèle physique correct pour décrire le phénomène. De plus, le solvant n'est pas explicitement pris en compte, ce qui ne permet pas l'estimation des énergies de désolvatation [142]. En l'absence d'effets de désolvatation, la fonction de score est d'avantage orientée vers les interactions coulombiennes favorisant ainsi les ligands fortement chargés. Une manière d'introduire l'effet des termes de solvatation est de traiter les molécules d'eau de façon explicite. Cependant, ces méthodes sont gourmandes en termes de calculs. Il est possible alors de remédier à cela en traitant l'eau comme un milieu diélectrique continu (de façon implicite). Ces modèles comprennent les méthodes de Poisson-Boltzmann (PB/SA) et de Generalized Born (GB/SA) souvent utilisées pour re-scorer les poses choisies avec une autre fonction de score moins gourmande. Ainsi, dans les méthodes GB/SA, les interactions électrostatiques et les coûts de désolvatation électrostatique sont calculés avec le modèle GB, tandis que les contributions hydrophobes pour les atomes non polaires sont estimées à l'aide des surfaces accessibles aux solvants (SA) des atomes. Le potentiel de Lennard-Jones est utilisé pour l'estimation des énergies de van der Waals. Les paramètres des contributions de van der Waals, hydrophobes et électrostatiques sont optimisés en accord avec les données expérimentales d'affinité.
Diverses fonctions de scores basées sur les champs de force existent comme Gscore qui est basé sur le champ de force Tripos et AutoDock sur le champ de force AMBER [173–175].

3.2.3.2.2.2. Les fonctions de score empiriques

Cette catégorie de fonctions de score est pensée de telle manière à ce qu'elles soient capables de reproduire les données expérimentales comme des données de binding, d'énergie et de conformations [142]. Une série de complexes protéine-ligand (données cristallographiques) avec des affinités de liaison connues est utilisée pour effectuer une analyse de régression linéaire multiple. Ensuite, les constantes (ou poids) générées par le modèle statistique sont utilisées comme coefficients pour ajuster les termes de l'équation. Chaque terme de la fonction décrit un type d'évènement physique impliqué dans l'interaction ligand-récepteur tel que les liaisons hydrogènes, les interactions polaires ou encore les effets désolvatation [176]. La fonction de score empirique ChemScore de Gold [177] par exemple, comprend les termes suivants : liaisons hydrogène, effets lipophiles des atomes et nombre effectif de liaisons à rotation libre dans le ligand. Un inconvénient des fonctions empiriques est leur dépendance aux données utilisées pour développer le modèle [178]. Toutefois, en raison de la simplicité des termes énergétiques employés, les fonctions empiriques sont plus rapides que celles basées sur les champs de force [155,179].

3.2.3.2.2.3. Les fonctions de score basées sur la connaissance.

Cette méthode aussi appelée "potentiel statistique" utilise des potentiels d'énergie calculés à partir des paires ligand-récepteur issues de la PDB [180]. Ces potentiels sont construits en tenant compte de la fréquence à laquelle une paire d'atomes ligand/protéine se trouve à une distance donnée dans l'ensemble de données structurales. Les différents types d'interaction observés dans l'ensemble de données sont classés et pondérés en fonction de leur fréquence d'apparition. Ainsi, le score final est donné comme la somme de ces interactions individuelles. Contrairement aux méthodes précédentes, les fonctions basées sur la connaissance ne reposent pas sur la reproduction des affinités de liaison (méthodes empiriques) ni sur des calculs ab initio (méthodes de champ de force). Elles constituent ainsi un bon équilibre entre précision et rapidité. Elles demeurent toutefois dépendantes de la qualité de résolution des complexe ligand-récepteur [181].

3.2.3.2.3. Prise en compte de la flexibilité de la protéine

La majorité des méthodes de docking couramment utilisées ne prennent en considération que la flexibilité du ligand. Or les protéines sont des entités flexibles et dynamiques. Cette flexibilité est par exemple visible en comparant les conformations de structures non liées (*apo*) et liées (*holo*) d'une protéine, et peut prendre la forme de simples réarrangements locaux de chaînes latérales [182–184] et aller jusqu'à des mouvements de domaines entiers. Il existe différentes

manières d'appréhender la flexibilité du docking (**Figure 14**) et toutes ont un impact direct sur les performances, pouvant conduire à des résultats différents [185,186]

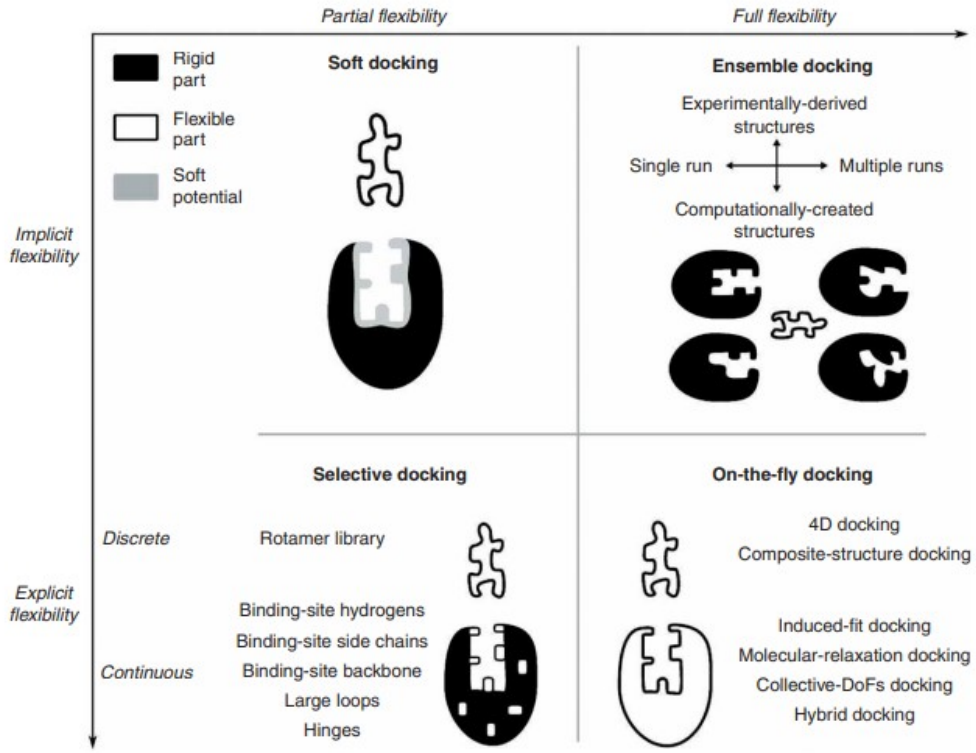


Figure 14 : Classification des méthodes de docking prenant en compte la flexibilité du récepteur d'après [145]

La prise en compte de la flexibilité d'une protéine peut être classifiée en deux grands axes (les deux colonnes de la **Figure 14**) : 1) Flexibilité partielle et 2) Flexibilité totale. Dans les deux cas, un deuxième niveau de complexité est rajouté pour diviser les méthodes selon le degré d'explicité de la flexibilité. Ainsi les 4 grandes catégories de docking flexible sont : Le *Soft* docking (flexibilité implicite partielle), le docking sélectif (flexibilité explicite partielle), le docking d'ensemble (flexibilité implicite totale) et enfin le *On the fly* docking ou à la volée (flexibilité explicite totale).

3.2.3.2.3.1. Le *Soft* docking (flexibilité implicite partielle)

L'idée est de permettre de petits chevauchements entre le ligand et le récepteur en « lissant » les potentiels de van der Wals et de tolérer de petites collisions stériques entre le ligand et la protéine en ajustant la fonction énergétique du potentiel de Lennard-Jones (utilisé pour calculer les interactions de VDW) [187,188]. Cela entraine un site de liaison légèrement plus grand tout en gardant quelques contraintes de plasticité conformationnelle. L'outil ADAM [189] par exemple utilise une variante de cette méthode appelée « VDW-offset grid ». Le plus grand avantage de

ce type de docking est qu'il n'entraine pas de coût supplémentaire de calcul. Malheureusement, il reste limité aux réarrangements à petite échelle associés à la flexibilité des chaînes latérales ne permettant pas de mouvements de grande envergure comme les mouvements du squelette [145]. De ce fait le soft docking est rarement utilisé seul mais plutôt en amont d'autres méthodes plus complexes [187].

3.2.3.2.3.2. Le docking sélectif (flexibilité explicite partielle)

Le docking sélectif peut s'avérer utile lorsqu'il existe une bonne connaissance structurale du récepteur. Ce dernier consiste à sélectionner quelques degrés de liberté « critiques » de la protéine, en plus de ceux du ligand, et à explorer explicitement leur variabilité [190,191]. Malheureusement cette approche est aussi associée à un coût computationnel important et de ce fait les approches actuelles se limitent à explorer les mouvements des chaines latérales à l'étape d'échantillonnage. Cela se fait 1) de façon discrète à l'aide d'une bibliothèque de rotamères [192] (contenant les rotamères des chaînes latérales d'acides aminés préférentiellement observés expérimentalement) ou bien via 2) une approche continue pour laquelle les liaisons à rotation libre des chaînes latérales sélectionnées sont entièrement explorées [193]. De nombreux outils incorporent cette approche comme l'outil SPECITOPE [194] qui se base sur l'approche continue en incorporant la flexibilité des chaines latérales ainsi que le squelette, AutoDockFR-AutoDock [195] qui permet de sélectionner et d'explorer la flexibilité des chaines latérales choisies. Plus récemment l'outil Smina, une branche de Vina, permet aussi de sélectionner les chaines latérales à considérer comme flexible en plus du fait qu'elle permette une suggestion de résidus autour du ligand [196].

3.2.3.2.3.3. Le docking d'ensemble (flexibilité implicite totale)

La simulation de la flexibilité de la protéine s'effectue en réalisant les docking sur différentes conformations d'une même structure. Cela est proposé par certains logiciels comme par exemple, DOCK [123] qui génère un potentiel d'énergie moyenné sur l'ensemble des conformations considérées ou FlexE [197] qui permet de fusionner les parties similaires de la protéine et considère les parties dissimilaires comme des alternatives différentes pour en déduire les interactions potentielles avec le ligand. Il est aussi possible de réaliser un docking d'ensemble en utilisant des logiciels de docking "classiques". D'abord, chaque ligand sera séquentiellement docké dans plusieurs structures protéiques. Ensuite, les résultats sont traités pour ne conserver, pour chaque ligand, que le meilleur score ou la moyenne des scores parmi toutes les structures [198,199]. Tous les ligands sont ensuite classés en fonction de ces nouveaux

scores. Les différentes structures utilisées peuvent provenir 1) de la PDB et correspondent donc à différentes structures résolues expérimentalement ou bien 2) provenir de différents *frames* issus d'une trajectoire de dynamique moléculaire ou d'analyse des modes normaux [185]. Dans les deux cas, il est clair que la qualité et la diversité des données structurales des protéines (l'espace conformationnel) influencent les résultats issus de ces méthodes [145].

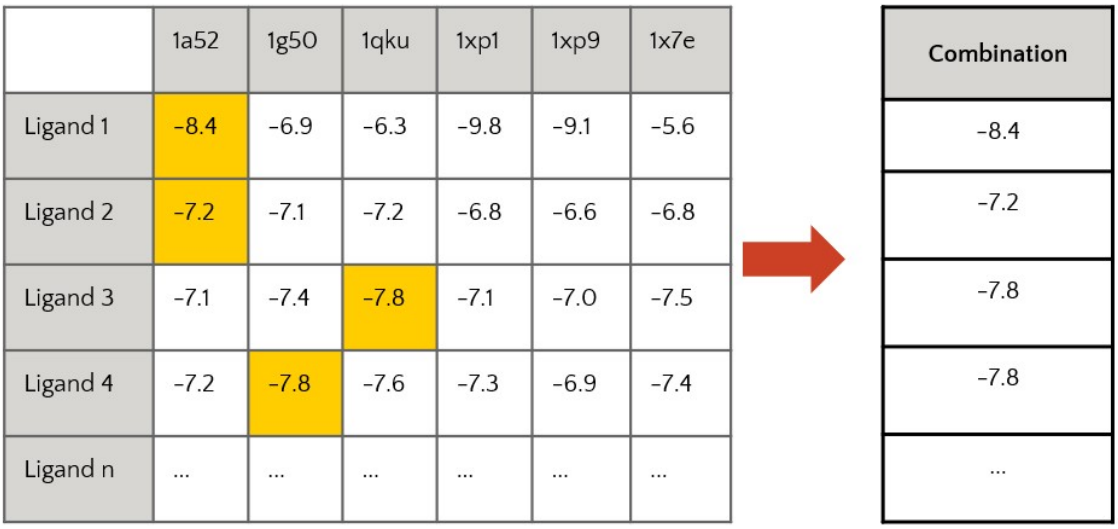


Figure 15 : Schématisation de l'étape d'intégration des données du docking d'ensemble

3.2.3.2.3.4. *On the fly* docking ou docking à la volée (flexibilité explicite totale)

Cette approche traite la flexibilité des protéines de manière explicite en générant de nouvelles conformations de protéines « à la volée » durant le docking. Pour ce faire, diverses techniques d'échantillonnage conformationnel et/ou des algorithmes d'optimisation heuristiques sont utilisés pour que le problème reste gérable sur le plan du cout computationnel [145]. Par exemple le 4D Docking [200] permet d'explorer un ensemble de conformations d'une même structure et ce de façon simultanée (contrairement à la manière successive avec laquelle on procède lors du docking d'ensemble). La notion de 4D n'implique pas l'ajout d'une quatrième dimension à proprement dit mais fait référence aux 4 variables à traiter lors de l'étape d'échantillonnage. En effet, en plus des 3 degrés de libertés explorées, une quatrième variable discrète relative à l'index de la conformation de la protéine au sein de l'ensemble est explorée [145].

3.2.3.2.4. Aspects pratiques de la méthode de docking

3.2.3.2.4.1. Préparation de la structure protéique

La qualité des modèles de docking est reliée à la qualité de la structure protéique. De plus à cause des nombreuses simplifications qui existent au niveau des deux étapes d'échantillonnage

et de scoring, il est important d'utiliser des structures cohérentes. L'étape de préparation du récepteur comprend aussi bien des étapes simples telles que l'ajout de résidus manquants, que des étapes un peu plus complexes dont la **détermination de l'état de protonation**, **l'identification des molécules d'eau inhérentes à l'interaction ligand-protéine**. La détermination de l'état de protonation constitue un facteur déterminant affectant la pertinence de la prédiction de l'affinité de liaison ligand-protéine. Un cas concret illustrant l'importance de la protonation des protéines est l'histidine, un AA dont la charge peut varier en fonction du pH. Cette dernière peut être impliquée dans les interactions ligand-protéine puisqu'elle peut être donneur ou accepteur de liaison hydrogène en fonction de son état électronique. Il est donc important de déterminer son état de protonation. Pour ce faire il faut d'abord identifier le pH physiologique de la protéine, se baser sur la conformation liée de cette dernière et analyser les prédictions de protonation en concordance avec les observations expérimentales. Il est à noter que les techniques actuellement utilisées pour la protonation dépendent en grande partie de la nature de la protéine étudiée [201].

3.2.3.2.4.2. Solvant

Un autre facteur important à prendre en compte est le solvant. Les interactions ligand-protéine se déroulent dans un milieu physiologique et impliquent donc des interactions avec l'eau.
Lorsque le point de départ est une structure cristallographique, il est important de vérifier la présence de molécules d'eau stables ou non. En effet, certaines molécules d'eau peuvent jouer un rôle dans la médiation de la liaison du ligand avec la protéine en initiant les liaisons hydrogènes et contribuant aux termes entropiques et enthalpiques [201]. Toutefois, l'inclusion des molécules d'eau dans le site actif diminue considérablement le volume de la poche et, par conséquent, les conformations possibles que le ligand peut adopter [202,203]. Il est donc important de déterminer lesquelles conserver lors d'un criblage virtuel et lesquelles exclure. Il existe quelques recommandations générales pour le faire, toutefois la meilleure manière de procéder reste de comparer le profil d'interactions avec et sans les molécules d'eaux « suspectées » afin de leur imputer ou non un rôle dans l'interaction ligand-protéine [201].

De plus, avant que l'interaction entre un ligand et une protéine ne se produise, chaque élément, i.e le ligand et la protéine, est entouré de molécules d'eau avec lesquelles elles interagissent. Au moment de l'interaction ligand-protéine, ces interactions sont rompues entrainant une contribution enthalpique qu'il faudra compenser par l'interaction ligand-protéine. Ainsi, plusieurs logiciels incluent dans leurs fonctions de scores un terme de solvatation-désolvatation

[204,205]. Par ailleurs, il est possible d'avoir recours aux calculs d'énergies libres de liaison en employant des approches prenant en compte l'interaction du ligand avec le solvant de manière implicite. Les méthodes MM-PBSA et MM-GBSA sont ainsi utilisées en amont de l'étape classique de docking dans une optique d'affiner les résultats de docking [206].

3.2.3.2.4.3. Fonctions de score et machine Learning

L'efficacité d'une méthode de docking est tributaire de la fonction de score utilisée qui permet d'évaluer la prédiction de l'affinité de liaison. Ainsi, les fonctions de score basées sur les champs de force sont moins susceptibles aux erreurs de protonations comparées aux méthodes basées sur les connaissances [201]. Toutefois, les fonctions de score classiques se basent sur les coordonnées atomiques du couple ligand-récepteur et ne permettent donc pas d'être toujours très pertinentes. Ainsi, les méthodes de *machine learning* et de *deep learning* ont été proposées pour améliorer ces fonctions de scores. Ces méthodes permettent de prédire l'affinité de liaison ou de classer la pose du ligand (en actif ou inactif) à partir d'un large jeu d'entrainement contenant des données expérimentales soit sur l'affinité de binding ou sur la pose expérimentale du ligand. Plusieurs algorithmes supervisés peuvent être utilisés dans ce contexte. La première application de ce type était l'utilisation de l'algorithme de *random forest* qui a permis d'améliorer la prédiction des affinités de liaisons [207]. Les réseaux neuronaux convolutifs (CNN) qui ont déjà fait leurs preuves en reconnaissance d'images ont été utilisés afin de scorer les poses [208]. Par ailleurs, notons qu'il est possible d'utiliser les méthodes de machine learning pour une étape supplémentaire de re-scoring également [209].

3.2.3.2.5. Evaluation des performances de l'algorithme de recherche des méthodes de docking

3.2.3.2.5.1. Self docking et cross-docking

Lorsque qu'une structure de la protéine en complexe avec un ligand est résolue, ce dernier (sa position et sa conformation) pourra être utilisé pour valider l'algorithme de recherche de la méthode de docking employée. En effet, la plupart des programmes de docking sont capables de prédire avec succès la conformation du ligand dans le site de liaison. Ceci peut être confirmé en réalisant une comparaison entre des complexes prédits et les structures résolues expérimentalement correspondantes à l'aide de mesures d'évaluation des poses (c.f paragraphe 3.2.3.2.5.2 ci-dessous). Deux méthodes communément utilisées pour valider la méthode de docking sont 1) le self-docking et le 2) cross-docking. Un self docking consiste à extraire le

ligand lié à la protéine et de réaliser une étude de docking avec ce ligand au niveau du site actif de la protéine. Les poses prédites par l'algorithme de recherche sont comparées à la pose de référence [210]. L'algorithme de recherche de la méthode de docking sera validé lorsque les poses prédites sont proches des poses observées dans les structures résolues expérimentalement. La méthode de cross-docking peut se faire lorsque plusieurs structures résolues expérimentalement avec différents ligands pour une même protéine sont disponibles. Les structures sont alignées et les ligands sont extraits. Ensuite, chaque ligand est successivement docké au niveau de chacune des structures et les poses sont comparées à celle du ligand de référence.

3.2.3.2.5.2. Mesures d'évaluation des poses

Afin d'apprécier la fidélité de la pose générée par le docking par rapport à la conformation observée dans la structure expérimentalement résolue, plusieurs mesures peuvent être effectuées. La plus populaire est la mesure de l'écart quadratique moyen (*Root Mean Square Deviation, RMSD*) en raison de sa simplicité, son objectivité et sa sensibilité [211]. Ainsi, le RMSD évalue la distance entre la position des atomes (généralement atomes lourds) entre la structure expérimentale et la structure prédite [211]. Plus la valeur du RMSD est faible, meilleure sera jugée la qualité de la prédiction des poses de la méthode de docking. La formule du RMSD est la suivante

$$RMSD\ (v,w) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}((v_{ix} - w_{ix})^2 + \left(v_{iy} - w_{iy}\right)^2 + (v_{iz} - w_{iz})^2}$$

Tel que :

v_i et w_i : les atomes identiques entre la structures expérimentale et la structure prédite

x, y et z : les coordonnées cartésiennes

n : Le nombre d'atomes total

Il est possible de calculer un RMSD moyen pour évaluer la qualité de toutes les poses générées par le docking. Dans ce cas-là, il est possible d'obtenir des divergences entre la valeur moyenne et les valeurs individuelles de RMSD. Il est donc préférable de calculer l'erreur relative au déplacement (RDE) qui permet de réduire l'impact de ces grandes divergences [212]. La formule est la suivante

$$RDE = 100(1 - \frac{L}{N}(\sum_{i=1,N} \frac{1}{L + D_{ii'}}))$$

Tel que

N : le nombre d'atomes

$D_{ii'}$: la déviation de l'atome i par rapport à l'atome i'

3.3. Evaluation des méthodes de criblage virtuel

L'évaluation des méthodes de criblage se fait généralement sur la base de métriques génériques et des mesures d'enrichissement. Afin de pouvoir calculer ces dernières, nous avons besoin de chimiothèques dites d'évaluation.

3.3.1. Les banques d'entraînement / d'évaluation

Les banques d'évaluation contiennent des composés dont l'activité pour une ou plusieurs cibles a été mesurée. Généralement, des composés dits actifs, i.e. qui ont démontré une activité biologique contre une cible particulière, y sont rassemblés avec un nombre d'autres molécules qualifiées d'inactives. Idéalement, ces inactifs ont aussi été testés expérimentalement pour la même activité biologique et n'ont pas démontré d'effet. Malheureusement, ce n'est pas toujours le cas car les résultats négatifs dans les tests expérimentaux sont rarement renseignés. Par conséquent, des leurres ou « *decoys* » peuvent être utilisés en tant qu'inactifs dans les banques d'évaluation [211]. L'étape de sélection des *decoys* est très importante lors de la préparation de la base de données. Lors de cette étape, des biais peuvent survenir notamment un biais d'analogie (le manque de diversité au sein des *decoys*), ou un biais de complexité (les *decoys* sélectionnés sont moins complexes chimiquement et donc facilement éliminés) qui peuvent entraîner une surestimation des performances des méthodes explorées. A l'inverse, l'utilisation de *decoys* peut être associée à une sous-estimation des performances des méthodes de criblage virtuel à cause de l'incorporation de faux positifs. En effet, les *decoy*s étant piochés aléatoirement, des molécules actives peuvent être intégrées par erreur et donc impacter les performances. Il existe plusieurs stratégies pour surmonter ces biais comme imposer la diversité structurale aux actifs par rapport aux *decoys* pour éviter le biais d'analogie [213] ou encore en imposant une similarité physico chimique entre les actifs et les *decoys* pour éviter le biais de complexité [214]. D'autres efforts sont menés dans le sens de rationaliser la sélection des decoys comme notamment la

diversification des cibles intégrées aux bases de données d'évaluation. Néanmoins, la manière la plus efficace reste l'intégration de données négatives expérimentales [215,216].

3.3.2. Les métriques d'évaluation

3.3.2.1. Métriques génériques

Le but des outils de criblage virtuel est de prédire le plus fidèlement possible deux catégories de molécules : les actifs et les inactifs. Il s'agit donc d'un problème de classification pour lequel 4 descripteurs sont définis : les vrais positifs (VP) , les vrais négatifs (VN), les faux positifs (FP) et les faux négatifs (FN). Les VP et les VN représentent la fraction des molécules dont l'activité a été correctement prédite et vice versa pour les FP et FN.

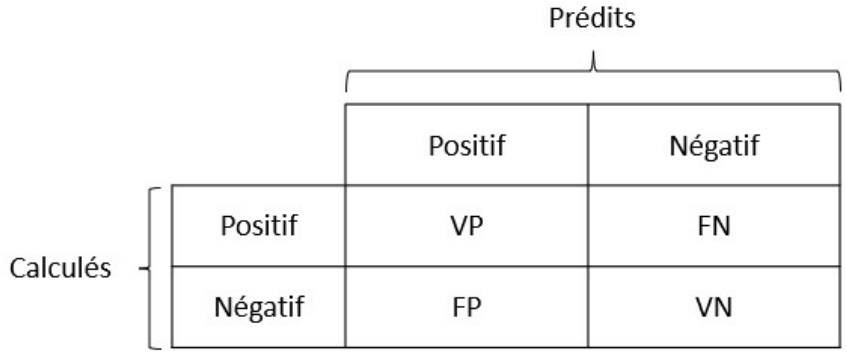


Figure 16 : Matrice de confusion ou tableau croisé

Ainsi à partir de ces 4 descripteurs, il est possible de calculer un ensemble de métriques utiles pour l'évaluation de la qualité du criblage. Les mesures les plus populaires sont **la sensibilité (Se)** et **la spécificité (Sp)**. Cependant, aucune de ces deux métriques ne rend compte de toutes les informations de la matrice de confusion. À moins que le nombre d'actifs et d'inactifs soient égaux, les valeurs de la Se et Sp peuvent être biaisées, voire dénuées de sens [217]. En réalité, les banques de données sont généralement très asymétriques avec une proportion d'inactifs supérieure à celle des actifs. Différentes approches existent pour pallier ce problème de métriques. La plus simple consiste à réduire la taille de la classe de données la plus importante pour qu'elle corresponde à celle de la classe la plus petite. Toutefois, ceci intègrerait un biais de représentativité car la réalité de l'espace chimique est que le déséquilibre inactifs/ actifs est toujours présent pour une cible donnée [87]. Il faudrait donc avoir recours à des métriques qui prennent compte de ce déséquilibre. L'***Accuracy***, le coefficient de corrélation de Matthews

(**MCC)** et le score **F1** -moins populaire- , contrairement à la Se et la Sp, utilisent toute l'information fournie par la matrice (c.f **Figure 16**) et seraient donc mieux adaptés dans le cas d'un déséquilibre de classes.

Tableau 2 : Les différentes métriques utilisées pour valider les méthodes de criblage. Pour toutes les métriques, la meilleure performance est la plus proche de 1. Pour le MCC -1 indique une parfaite corrélation négative, 0 une classification aléatoire et 1 une corrélation parfaite

Métrique	Formule	Intervalle
Sensibilité	$Se = \frac{TP}{TP + FN}$	[0 ;1]
Spécificité	$Spf = \frac{TN}{TN + FP}$	[0 ;1]
Accuracy	$Acc = \frac{TP + TN}{TP + TN + FP + FN}$	[0 ;1]
Matthews Correlation Coefficient	$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}}$	[-1 ;1]
F1-score	$F1 = \frac{2 * TP}{2 * TP + FP + FN}$	[0 ;1]

3.3.2.2. La courbe ROC

La *receiver operating characteristics curve* ou courbe ROC permet d'analyser visuellement les performances globales d'une méthode de criblage. En effet, elle représente la sensibilité en fonction de (1-spécificité) ou en d'autres termes le pourcentage d'actifs en fonction du pourcentage d'inactifs. Dans un cas idéal, tous les composés actifs sont classés avant les inactifs par la méthode de criblage virtuel étudiée. La courbe correspondante illustrera ce classement en montant jusqu'à (0,1) puis continue tout droit vers la droite avec tous les TN. Un classement aléatoire situerait la courbe sur la diagonale, c'est-à-dire un mélange de vraies et fausses prédictions. Plus la courbe monte rapidement, meilleure est la performance globale de la méthode. Il est à noter qu'on peut comparer différentes méthodes en utilisant l'analyse ROC uniquement lorsque le même ensemble de données est utilisé [217].

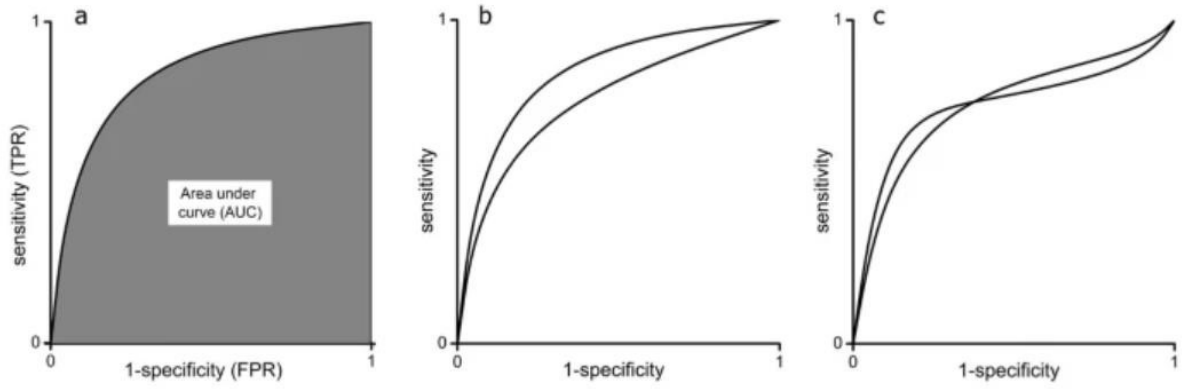


Figure 17 : Analyse de la courbe ROC. La courbe qui s'élève le plus correspond à une meilleure méthode. Si les courbes se croisent, cela signifie que la comparaison n'a plus de sens d'après [217]

Par ailleurs, l'aire sous la courbe ROC (AUC) peut être utilisée comme mesure de la performance des prédictions (formule). Elle représente approximativement la probabilité de classer un composé actif choisi au hasard plus haut qu'un inactif choisi au hasard. Les valeurs d'AUC sont comprises entre 0 et 1, une valeur de 0,5 indique que la performance de la méthode est assimilable à l'aléatoire tandis qu'une valeur de 1 est associée à une performance idéale.

$$AUC = \sum_{i} [Se_{i+1} \, (Sp_{i+1} - Sp_i)]$$

Il faut garder à l'esprit que la courbe ROC n'indique pas directement la performance d'une méthode. Elle montre le potentiel de classement de la méthode, qui est lié à la performance globale, ce qui renforce encore le fait qu'une seule mesure ne peut pas décrire entièrement la performance prédictive. Le meilleur moyen pour évaluer une méthode reste une analyse de toutes les métriques et surtout leur contextualisation [217].

3.3.2.3. Métrique d'enrichissement

Le facteur d'enrichissement (**EF**) est calculé généralement pour une fraction de la chimiothèque (généralement 10% et 25 %) [211]

$$EF_{100*(n/N)\%} = \frac{TP/n}{(TP + FN)/N}$$

Equation de mesure de l'enrichissement pour la fraction n/N % de la chimiothèque où n : nombre de composés dans la fraction de la chimiothèque étudiée, N : nombre total de composés dans la chimiothèque.

Cette métrique permet d'évaluer la capacité d'une méthode à retrouver les actifs dans une fraction donnée en comparaison à une sélection aléatoire. L'EF permet donc une mesure d'une performance sur une fraction d'intérêt contrairement à l'AUC qui est une mesure de performance globale. L'EF est donc particulièrement intéressant dans le cas du criblage virtuel dans lequel une fraction de composés les mieux classés après le criblage virtuel vont être étudiés expérimentalement. Tout comme les Se et la Sp, l'EF est tributaire du ratio entre actifs et inactifs. De plus, il présente une deuxième faiblesse qui est sa non-sensibilité à la distribution des actifs dans la fraction étudiée. Cela veut dire que pour deux méthodes capables de retrouver le même nombre d'actifs dans les premiers n%, le EF ne permet pas de savoir laquelle retrouve plus d'actifs plus tôt. La courbe de ROC permet de pallier ce défaut ce qui explique pourquoi elle est généralement préférée aux courbes d'enrichissements [217].

3.3.3. Importance des métriques en fonction du contexte

Selon le contexte dans lequel les méthodes de criblage virtuel sont appliqués mais aussi selon les capacités expérimentales associées pour vérifier les prédictions, une importance plus grande peut être donnée à un paramètre d'évaluation de ces méthodes. Ainsi, dans le contexte de la recherche de candidats médicaments assistée par ordinateur, l'objectif est de réduire le grand nombre de composés à tester expérimentalement à un plus petit nombre potentiellement optimisable. Dans ce contexte-là, les faux négatifs peuvent être tolérés mais ne sont pas critiques. En toxicologie, le but du criblage virtuel est de prioriser les molécules potentiellement toxiques à tester en urgence à l'aide de méthodes expérimentales afin d'identifier et préserver l'exposition humaine et environnementale à ces composés. De ce fait, il est important de correctement prédire le potentiel toxique de tous les composés chimiques et de réduire au maximum le nombre de faux négatifs[16].

4. Les perturbateurs endocriniens

L'évaluation des risques en toxicologie comme abordée plus tôt est cruciale à la fois pour les futurs candidats médicaments mais aussi pour garantir la sécurité des composés provenant d'industries chimiques ainsi que les composés environnementaux. Une catégorie particulière de composés a donné lieu à de nombreuses études toxicologiques. Il s'agit des Perturbateurs Endocriniens (PE). Les travaux menés pendant ces 3 années de thèses ont porté sur cette famille de composés. Dans cette partie, nous commencerons par définir ce groupe de molécules, leurs mécanismes d'action et les conséquences de l'exposition à ces PE. Ensuite, nous aborderons les tests biologiques utilisés pour les identifier ainsi que les caractéristiques chimiques et structurales propres à ce groupe. Enfin, nous tenterons de contextualiser le problème des perturbateurs endocriniens et le besoin de développer des modèles *in silico* pour les prédire.

4.1. Définition

Le sujet des perturbateurs endocriniens (PE) a été abordé pour la première fois en tant que problème de santé publique et environnementale mondial lors du « sommet de Rio » en 1992. Depuis, les PE sont réglementés par divers organismes internationaux comme l'organisation de coopération et de développement économique (OECD), la Food and drug administration (FDA), l'agence américaine de protection de l'environnement (EPA), les Nations Unies et l'organisation mondiale de la santé (OMS) [218]. Plusieurs définitions des PE ont été proposées mais c'est celle de l'OMS qui est la plus communément utilisée. Elle définit un PE comme une substance ou un mélange de substances exogènes qui altère la ou les fonctions du système endocrinien et provoque par conséquent des effets indésirables dans un organisme intact, ou sa progéniture, ou des (sous-)populations [219,220].

4.1.1. Points de discorde

La variété de définitions existantes et le fait qu'aucune harmonisation n'a été faite est causé par la divergence sur différentes questions. Un premier point de discorde concerne la question des doses d'expositions ainsi que des mélanges (c.f 4.3.3 Facteurs influençant les mécanismes d'action). Bien que cette dernière ait été soulignée dans le rapport de l'EDSTAC en 1998, il n'existe actuellement aucun mécanisme permettant d'intégrer cette question dans les stratégies de tests biologiques [218] . Mais le principal point de discorde reste l'utilisation du terme

« néfaste » pour définir un PE. En effet, pour qu'un produit chimique soit réglementé, il faut démontrer qu'il induit un effet néfaste, mais cette question reste complexe dans le domaine des perturbateurs endocriniens en raison de la double nature du phénomène. Par exemple, l'œstrogénicité en soi n'est pas un effet indésirable, c'est un mécanisme naturel d'action hormonale contrôlé par des mécanismes homéostatiques. Cependant, un produit chimique aux propriétés œstrogéniques agissant hors contexte dans le système endocrinien, ou à un moment critique du développement, peut potentiellement induire un effet indésirable. A cet égard, la société d'endocrinologie précise qu'il n'est pas nécessaire qu'un produit chimique provoque des effets indésirables pour être considéré comme un PE, dès lors qu'il interfère avec un aspect quelconque de l'action hormonale. En effet, les PE ne peuvent pas être traités comme des toxines ordinaires et leur définition doit être encadrée par des principes d'endocrinologie afin de mieux concevoir le criblage et les tests biologiques visant à détecter les substances chimiques susceptibles de perturber le système endocrinien [218,219,221].

4.1.2. Sources de contamination

Les PE sont des composés pour la plupart synthétiques d'origine industrielle, même si certains ont une origine naturelle [222]. Les PE sont présents dans de nombreux produits, notamment les bouteilles en plastique, les boîtes de conserve en métal, les détergents, les retardateurs de flamme, les aliments, les jouets, les cosmétiques et les pesticides qui constitue la catégorie de produits à susciter l'intérêt pour les PE en premier [219,223,224]. Les aliments constituent la principale voie d'exposition aux PE. De plus, l'augmentation des produits ménagers contenant des polluants et la diminution de la ventilation des bâtiments font de l'air intérieur une source importante d'exposition aux perturbateurs endocriniens. Plusieurs substances environnementales, y compris des métaux lourds, qui semblent agir comme des perturbateurs endocriniens, ont été signalées. De nombreuses études ont démontré que les tissus humains, notamment les reins, le foie et les testicules, sont sensibles à la toxicité des métaux lourds libérés dans l'environnement par les produits industriels et agricoles. L'exposition au tabac, en particulier est la principale source d'exposition humaine aux métaux lourds [219,224]. L'Homme est soumis à une exposition continue et simultanée à différents PE. Ainsi, la bioaccumulation devient inévitable dans de nombreux cas, ce qui pourrait causer des dommages permanents par défaut d'adaptation physiologique [225].

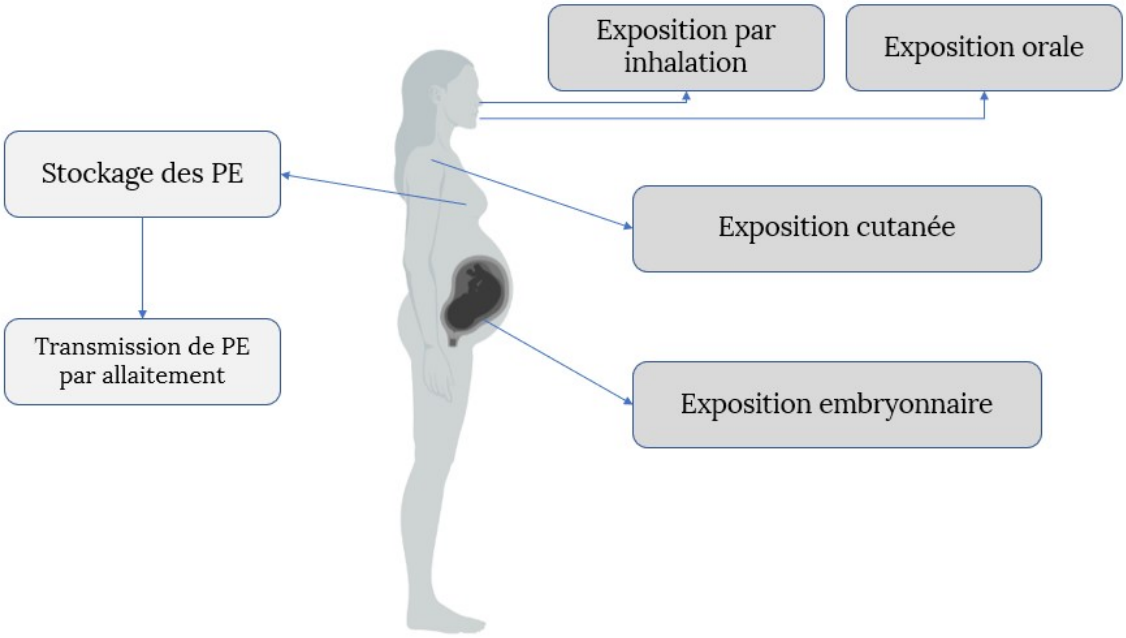


Figure 18 : Principales sources d'exposition aux perturbateurs endocriniens d'après [226]

4.2. Le système endocrinien

4.2.1. Généralités

Le système endocrinien est un réseau de glandes et de cellules capables de synthétiser des hormones et de les libérer dans le sang. Les hormones peuvent ainsi atteindre leurs tissus cibles pour exercer leur fonction biologique soit directement en se liant aux récepteurs appropriés soit indirectement après leur métabolisme. Le système endocrinien contrôle de nombreuses fonctions biologiques comme le métabolisme, la croissance, le développement des caractères sexuels secondaires et l'activité sexuelle [227]. La sécrétion hormonale peut être déclenchée par (1) une stimulation nerveuse (décharge d'adrénaline lors d'un stress par exemple), (2) une modification de l'homéostasie (une augmentation de la concentration en glucose dans le sang stimule la sécrétion d'insuline par le pancréas) ou (3) par d'autres hormones (les hormones de l'hypophyse FSH et LH qui stimulent la production d'œstrogènes et de progestérone par les ovaires).

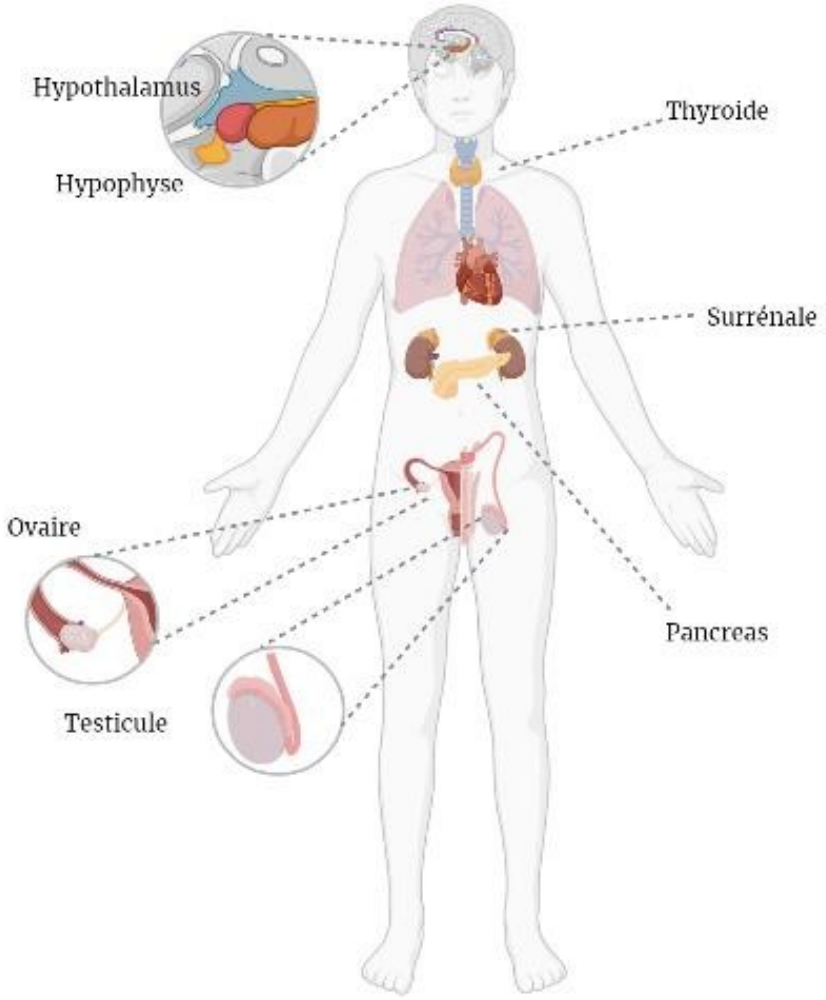


Figure 19 : Le système endocrinien

4.2.2. Les Hormones

Les hormones sont des molécules produites par les glandes endocrines en réponse à un certain stimulus et libérés dans la circulation sanguine [227]. L'effet d'une hormone est destiné à un organe cible présentant un groupe de cellules spécifiques. Si l'hormone agit sur d'autres cellules, des troubles pathologiques peuvent se développer.

Les Afin d'atteindre leurs organes cibles, les hormones circulent dans le sang soit de manière active via des transporteurs spécifiques soit de manière passive sans transporteurs. Pendant le transport, les hormones sont sujettes à une dégradation enzymatique. Ainsi, pour atteindre l'organe cible en concentrations suffisantes, de nombreuses hormones sont libérées par l'action coordonnée de nombreuses cellules et la libération se fait de manière pulsatile [228].

Il existe différentes catégories d'hormones comprenant, (1) les hormones stéroïdiennes, (2) les hormones dérivées d'acides aminés et (3) Les hormones de nature polypeptidique [227]. Les

différences résident généralement dans une différence structurale ce qui induit des différences en termes les cellules cibles et de mécanismes d'action.

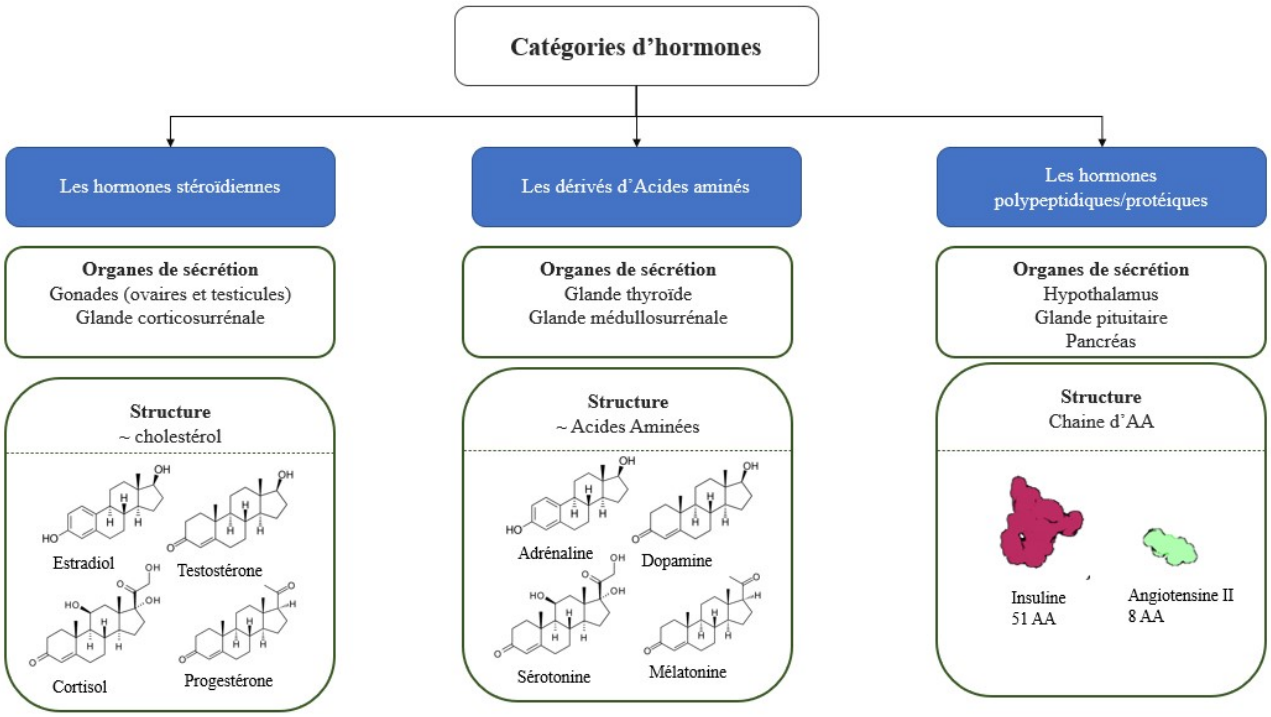


Figure 20 : Les différentes catégories d'hormones. Les différences structurales entre les différentes catégories fait que certaines sont plus aptes à passer la membrane cellulaire et atteindre les récepteurs au niveau du noyaux. C'est le cas des hormones stéroïdiennes et des dérivés d'acides aminés.

4.2.2.1. Les stéroïdes

Dans cette classe d'hormones, nous trouvons les estrogènes (dont l'estradiol) et les androgènes (dont la testostérone), les corticostéroïdes et les minéralocorticoïdes (le cortisol et l'aldostérone) et les hormones de gestation (progestérone). Tous ces stéroïdes sont dérivés du cholestérol et leur synthèse se fait en plusieurs étapes nécessitant plusieurs enzymes. Le produit stéroïdien final peut être stocké dans la cellule mais le plus souvent, il est libéré immédiatement après synthèse par diffusion dans l'environnement cellulaire. Ainsi, la libération d'une hormone stéroïdienne est régulée non pas par un signal de libération de molécules préexistantes, mais par l'activation des gènes régulant l'expression de la cascade de synthèse [228].

4.2.2.2. Les hormones polypeptidiques/protéiques

Cette première catégorie comprend les hormones régulant la reproduction (gonadotrophines), l'équilibre énergétique (insuline) ou la pression artérielle (angiotensines). De plus d'autres

hormones responsables régulant la synthèse de certaines hormones appartiennent aussi à cette catégorie comme l'hormone de libération des gonadotrophines (GnRH) et l'hormone de libération de l'hormone de croissance (GHRH), et les hormones contrôlant l'alimentation, comme la leptine et le neuropeptide Y [228]. Certaines hormones possèdent un effet neurotransmetteur en plus des effets endocriniens. Une distinction semble difficile à établir car on ne sait pas si, par exemple, la noradrénaline est libérée par le système nerveux autonome ou par la médullosurrénale. Cela pourrait refléter la connexion intime entre le système endocrinien et le système nerveux [228].

4.2.2.3. Les dérivés d'acides aminés

Cette classe comprend des hormones comme les hormones thyroïdiennes (la triiodothyronine et la thyroxine), les catécholamines dérivant de la tyrosine (dopamine, noradrénaline et adrénaline) et les indolamines (mélatonine et les molécules apparentées comme la sérotonine) fabriquées à partir du tryptophane [227,228].

4.2.3. Les récepteurs nucléaires

Les hormones endocrines engendrent la majorité de leurs effets via la voie des récepteurs nucléaires (NR). Ces derniers forment une superfamille de facteur de transcription impliqués non seulement dans d'importantes fonctions physiologiques (le contrôle du développement embryonnaire, de la physiologie des organes, de la différenciation cellulaire et de l'homéostasie [229–231] mais aussi identifiée dans leurs dérégulations associées à de nombreux processus pathologiques (le cancer, le diabète, la polyarthrite rhumatoïde, l'asthme ou les syndromes d'hormonorésistance) [232,233]. Outre les perturbations d'ordre physiopathologique, les NR constituent la cible des PE aussi bien dans le cas des mécanismes d'action directe qu'indirecte. Les NR suscitent donc un intérêt aussi bien thérapeutique que toxicologique intimement lié au sujet des PE. La place de cette famille de protéine dans ces 2 thématiques a été longuement traitée dans une revue réalisée durant ce travail de thèse. Cette revue, présentée dans la partie Résultats de ce mémoire, a pour but d'énumérer les projets *in silico* d'intérêt thérapeutique et toxicologique dédiés aux NR (c.f partie résultats). Dans cette partie du mémoire, nous présenterons les aspects structuraux et mécanistiques de ces protéines, nécessaires pour une meilleure compréhension du contexte des travaux portant sur les NR.

Tableau 3 : Les membres de la superfamille des récepteurs nucléaires humains

Subgroup	Characteristics	Members	Complete name	Endogenous ligands
Subgroup 0	Atypical NR presenting only a LBD with a co-activator motif able to interact with other NR LBD	DAX	dosage-sensitive sex reversal-adrenal hypoplasia congenital critical region on the X chromosome	-
		SHP	small heterodimer partner	-
Subgroup 1	Regulated by lipophilic signaling molecules	**THRA**	thyroid hormone receptors alpha	Thyroid hormones
		THRB	thyroid hormone receptors beta	Thyroid hormones
		RARA	retinoic acid receptors alpha	Retinoic Acid
		RARB	retinoic acid receptors beta	Retinoic Acid
		RARG	retinoic acid receptors gamma	Retinoic Acid
		PPARA	peroxisome proliferator activated receptors alpha	-
		PPARD	peroxisome proliferator activated receptors delta	Prostaglandin
		PPARG	peroxisome proliferator activated receptors gamma	Fatty acids
		PXR	pregnane X receptors	-
		REV-ERA	reverse-ERa receptors	Heme
		REV-ERB	reverse-ERb receptors	Heme
		RORA	retinoic acid related receptors	Cholesterol
		RORB	retinoic acid related receptors	Retinoic Acid
		RORC	retinoic acid related receptors	-
		FXR	farnesoid X receptors	Bile acids
		LXRA	liver X receptors alpha	Oxyterols

		LXRB	liver X receptors beta	Oxyterols
		CAR	Constitutive androstane receptor	-
		VDR	vitamin D receptors	vitamin D
Subgroup 2	Orphan receptors able to bind fatty acids with unclear effect on protein regulation mechanism	**RXRA**	retinoid X receptors alpha	Retinoic Acids
		RXRB	retinoid X receptors beta	Retinoic Acids
		RXRG	retinoid X receptors gamma	Retinoic Acids
		TR2	Testicular receptor 2	-
		TR4	Testicular receptor 4	-
		TLX	homologue of the Drosophila tailless gene	-
		PNR	photoreceptor cell-specific nuclear receptor	-
		COUP-TFII	chicken ovalbumin upstream promoter transcription factors I	-
		COUP-TFII	chicken ovalbumin upstream promoter transcription factors II	-
		HNF4 A	hepatocyte nuclear Factor 4	-
		HNF4 B	hepatocyte nuclear Factor 4	-
		NR2F6	V-erbA-related protein 2	-
Subgroup 3	Regulated by cholesterol-derived hormones	**ERA**	Estrogen receptor alpha	17 beta estadiol
		ERB	Estrogen receptor beta	17 beta estadiol
		ERRalpha	Estrogen related receptor aalpha	-
		ERRb	Estrogen related receptor beta	-
		ERRgamma	Estrogen related receptor gamma	-
		AR	Androgen receptor	Testosterone
		PR	progesterone receptor	Progesterone
		GR	Glucorticoid receptor	Cortisol

		MR	Mineralocorticoid receptor	Aldosterone, 11-deoxycorticosteron and cortisol
Subgroup 4	Orphan nuclear receptors	NGF1-B	nerve growth Factor 1B	-
		NURR1	nurr-related Factor-1	-
		NOR-1	neuron-derived orphan Receptor-1	-
Subgroup 5	Regulated by phospholipids	**SF-1**	steroidogenic Factor 1	-
		LRH-1	Liver receptor Homolog-1	-
Subgroup 6	Presenting critical difference in the LBD and no activator function (AF-H)	GCNF	germ cell nuclear factor	-

4.2.3.1. Structure de NR

4.2.3.1.1. Structure générale

La superfamille des NR est composée de 48 membres chez l'Homme répartis en 7 sous-groupes (**Tableau 3**). A l'exception de SHP et DAX, tous les NR partagent globalement la même architecture comprenant 6 domaines, nommés de A à F, dont chacun joue un rôle spécifique. Le domaine A/B est le domaine N-terminal (NTD), contenant la région Function-1 de l'activateur (AF-1), qui interagit avec une variété de protéines co-régulatrices de manière spécifique à la cellule et au promoteur. Le domaine C constitue le domaine de liaison à l'ADN (DBD), la région la plus conservée parmi tous les NR. Il contient deux sous-domaines qui coordonnent la formation du motif canonique du doigt de zinc se liant à l'ADN. Certains NR, comme LRH-1 et GCNF, contiennent une extension C-terminale (CTE) de ce DBD, permettant des contacts supplémentaires avec l'ADN. Le domaine D est la région charnière, un élément de liaison entre le DBD et le LBD. Il s'agit de la séquence la plus courte et la région la moins conservée parmi les NR. Enfin, le domaine E est une région de signalisation allostérique structurellement conservée formant le domaine de liaison au ligand (LBD). Et enfin le domaine F au niveau du C-terminal dont la fonction reste indéfinie [234]. Il s'agit de la région la moins conservée qui est d'ailleurs inexistante pour beaucoup de NR[235].

Les NR sont généralement présents sous forme de monomères en solution, mais la plupart d'entre eux se transforme en complexes d'ordre supérieur lorsqu'ils se lient à l'ADN (homodimères ou hétérodimères), ce qui permet d'augmenter leur surface de contact.

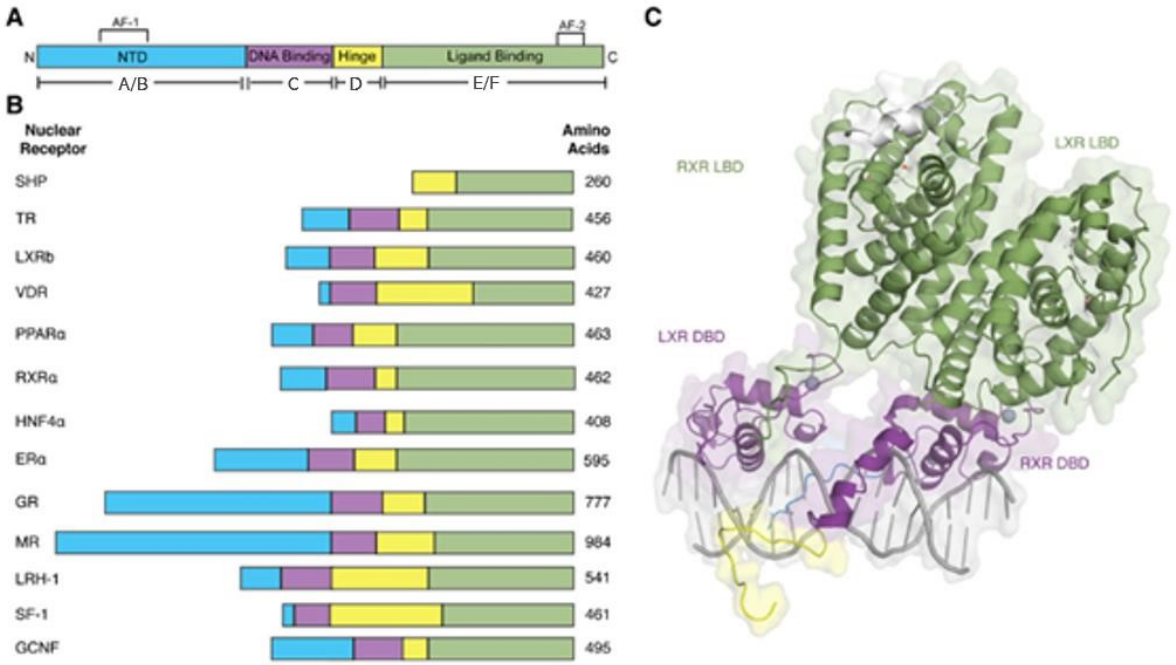


Figure 21 : Les différents domaines structuraux des récepteurs nucléaires. (A) la structure linéaire des différents domaines avec notamment la fonction d'activation AF-1, le doigt de zinc au niveau du DBD et la fonction AF-2 qui se lient au coactivateur/répresseur à la suite de la liaison du ligand au LDB. (B) Taille générale et longueur de la chaine pour chaque domaine pour différents NR. (C) Exemple d'une structure 3D d'un NR (l'hétérodimère LXR-RXR) montrant le DBD en violet, le domaine charnière en jaune et le LBD en vert d'après [236]

4.2.3.1.2. Structure du site de liaison du ligand

Le LBD est un domaine de signalisation allostérique complexe qui se lie non seulement aux ligands, mais interagit aussi directement avec les protéines co-régulatrices [237]. Ce domaine structualement conservé contient généralement 11 hélices α et 4 brins β qui se replient en trois couches parallèles pour former un « sandwich » hélicoïdal α créant une poche hydrophobe de liaison au ligand (*Ligand Binding Pocket* ou LBP) à la base du récepteur [237]. La partie supérieure du récepteur est la plus conservée alors que la partie basse contenant la LBP, est plus variable [238]. C'est cette variabilité entre les NR qui leur permet de reconnaître des ligands spécifiques. Le LBD contient la fonction d'activation (AF-2), qui est composée des hélices 3, 4 et 12. L'hélice H12, appelée hélice de la fonction d'activation (AF-H), change de conformation lors de la liaison au ligand. En effet, en absence d'un ligand, l'hélice est éloignée du LBD ce qui la rend apte à fixer un corépresseur. En présence du ligand, l'hélice 12 se referme à la manière d'un couvercle piégeant ainsi le ligand et permettant de fixer des co-activeurs [239]. Il est à noter

que sur certains NR comme le récepteur constitutif aux androstanes (CAR), l'hélice H12 peut être fermée constitutionnellement correspondant ainsi à la conformation liée d'autres NR (i.e la conformation active) du NR. La fixation d'un ligand au LBD entraine ainsi une stabilisation de cette conformation [240].

4.2.3.2. Mécanistique

Les récepteurs nucléaires sont des protéines solubles qui peuvent se lier à des éléments spécifiques de régulation de l'ADN et agir comme des régulateurs de type cellulaire [241]. Contrairement à d'autres facteurs de transcription, la plupart des NR sont régulés de manière endogène par des ligands apparentés. Cependant, pour 25 des NR appelés récepteurs orphelins, aucun ligand endogène n'a été caractérisé. La liaison du ligand à la poche hydrophobe du LBD entraîne un changement de conformation de la protéine. Le NR ainsi activé est capable de se lier à une séquence d'ADN spécifique et de recruter la machinerie appropriée pour la transactivation ou la transrepression des gènes ciblés [236,242]. Outre les ligands endogènes, d'autres composés synthétiques peuvent agir sur les NR avec un large spectre d'activité en tant qu'agoniste, antagoniste, agoniste ou antagoniste partiel, agoniste inverse ou encore modulateur spécifique (SARM) [216]. Des différences structurales au niveau des ligands ont été identifiés[199] et il est actuellement admis que l'action antagoniste est due à une perturbation du repliement de l'hélice 12. Différentes hypothèses concernant ce mécanisme existent mais ce dernier reste malgré tout encore non résolu [243,244].

4.3. Mécanismes d'action

Les PE peuvent agir sur deux familles de récepteurs à savoir les NR et les récepteurs membranaires. Les NR, étant des facteurs de transcription, l'effet induit par les PE est un effet à long terme. En revanche pour les récepteurs membranaires, l'effet sera plutôt à court terme [224]. Les PE peuvent agir selon différents mécanismes d'action classés en voies directe et indirecte. La voie directe implique la liaison directe des PE aux NR agissant en tant qu'agonistes ou antagonistes, mimant ou bloquant ainsi l'effet d'une hormone. Cependant, d'autres mécanismes d'action des PE ont été décrits et catégorisés comme indirect. Dans cette voie, les PE entrainent un déséquilibre de l'homéostasie. Ce mécanisme n'est pas directement lié à une action mimétique des hormones et implique des composés aussi bien similaires que différents structuralement des hormones [224,245,246]. Au-delà de ces mécanismes d'action directs et indirects, d'autres mécanismes méconnus confèrent aux PE d'autres types de toxicité telles que la cytotoxicité, la reprotoxicité, la tératoxicité ou encore la génotoxicité entrainant

l'apparition de cancer non hormonaux. Ces effets sont particulièrement inquiétants car les altérations de la programmation génétique au cours des premiers stades du développement peuvent avoir des effets profonds des années plus tard et peuvent également conduire à une transmission transgénérationnelle de la maladie [224,247]. Ces mécanismes dépassent le champ de nos compétences et ne seront pas abordés dans ce manuscrit. Seuls les mécanisme directes et indirects seront expliqués.

4.3.1. Mécanismes d'actions directs des PE

La voie directe comprend aussi bien l'activation que l'inhibition de la voie de signalisation d'une hormone en se fixant sur son récepteur [224]. En effet, grâce à leurs structures chimiques, leurs forme et taille similaires à celles des hormones, certains PE peuvent atteindre le site de liaison des hormones et interférer avec leurs mécanismes d'action. De plus, les PE sont présents en grande quantité au niveau du site d'action ce qui leur permet d'entrer en compétition avec les hormones au niveau du site de liaison et ce malgré une affinité réduite pour les récepteurs. Par ailleurs, ces composés sont dotés d'un temps de demi-vie systémique et cellulaire assez long qui fait qu'ils sont moins efficacement dégradés que les hormones et que par conséquence, leur action soit plus forte. Certains PE peuvent, en se liant au NR, entrainer leur blocage dans une seule conformation antagonisant ainsi l'action des hormones. Ce mécanisme d'action fait que certains PE ont une action quasi instantanée. Ainsi, les BPC (biphényles polychlorés) agissent en inhibant la liaison de T3 au THR entrainant la dissociation de l'hétérodimère actif TR/RXR de l'élément de réponse aux thyroïdes (TRE) [224].

4.3.2. Mécanisme d'actions indirects des PE

En parallèle de la voie directe, la voie indirecte comprend plusieurs mécanismes d'actions dont (1) la perturbation de la voie de signalisation des hormones, (2) l'effet sur la concentration des hormones, (3) l'activation ou l'inhibition de la synthèse ou de la dégradation des protéines de transport et (4) l'activation ou l'inhibition de l'expression du récepteur à l'hormone (*Turn-over*) [224,246].

(1) Dans le premier cas, les PE vont interagir avec des éléments impliqués dans la voie de signalisation des hormones en aval du récepteur. Les effets toxiques sont ainsi non endocriniens à proprement dit mais affectent d'autres phénomènes biologiques comme l'expression/inhibition de certains gènes. Par exemple, Caron et al [248] ont démontré que l'exposition des cellules Hs578t (lignée de cellules épithéliales isolées à partir de tissus cancéreux du sein) à des pesticides nicotinamides, entrainant une surexpression de l'aromatase

suite à un changement de promoteur. L'aromatase est une enzyme clé pour la synthèse des estrogènes et une augmentation de son taux entraine donc une augmentation des concentrations d'estrogènes dans l'environnement de la tumeur et induit au développement du cancer. Pour une cellule saine, l'expression du gène de l'aromatase est sous le contrôle du promoteur I.4 qui permet une expression à bas niveau. Dans une cellule cancéreuse, il y a inhibition du I.4 et activation de deux autres promoteurs I.3 et I.7 qui eux vont entrainer une augmentation de l'aromatase. Ce changement de promoteur est ainsi induit par l'exposition aux pesticides concernées par l'étude.

(2) Les PE exercent un effet sur la concentration des hormones en stimulant ou inhibant la synthèse ou la dégradation de ces dernières. Par exemple le triclosan entraine la sécrétion de VEGF (*Vascular Endothelial Growth Factor*) par les cellules cancéreuses de la prostate chez l'homme [249]. Les parabènes, conservateurs largement utilisés dans les cosmétiques inhibent la dégradation d'estrogène en inhibent la 17β-hydroxysteroïd déshydrogénase entrainant l'augmentation de concentration de l'œstrogène dans le sang [250].

(3) De plus certains PE peuvent entrer en compétition avec les stéroïdes au niveau des protéines de transport ou les éléments de liaison aux hormones circulants diminuant ainsi la concentration des hormones (en particulier les stéroïdes) dans le sang [251]. Certains PE peuvent aussi affecter la synthèse ou la dégradation de ces protéines de transport. Ainsi, plusieurs composés hépatotoxiques peuvent être classés comme PE car les protéines de transport sont généralement synthétisées et dégradées dans le foie. Par exemple, les PBDE agissent en inhibant les TTR (protéines de transport de transthyrétine) entrainant une diminution de la concentration de T4 sanguine [252].

(4) Enfin, la synthèse et la dégradation des récepteurs est sujette à un rétrocontrôle établi par les hormones elles-mêmes [218]. Ainsi, la surexpression ou le déficit des RN perturbe l'homéostasie endocrinienne. Par exemple, l'exposition au cadmium des cellules HUVEC *in vitro* entraine une augmentation de l'expression de ERβ ainsi que la diminution de celle de AR après plus de 24h d'exposition [253]. Un autre exemple, les BPA administrés à faible dose oralement, entrainent l'inhibition de l'expression des AR et ER chez les rats [254,255].

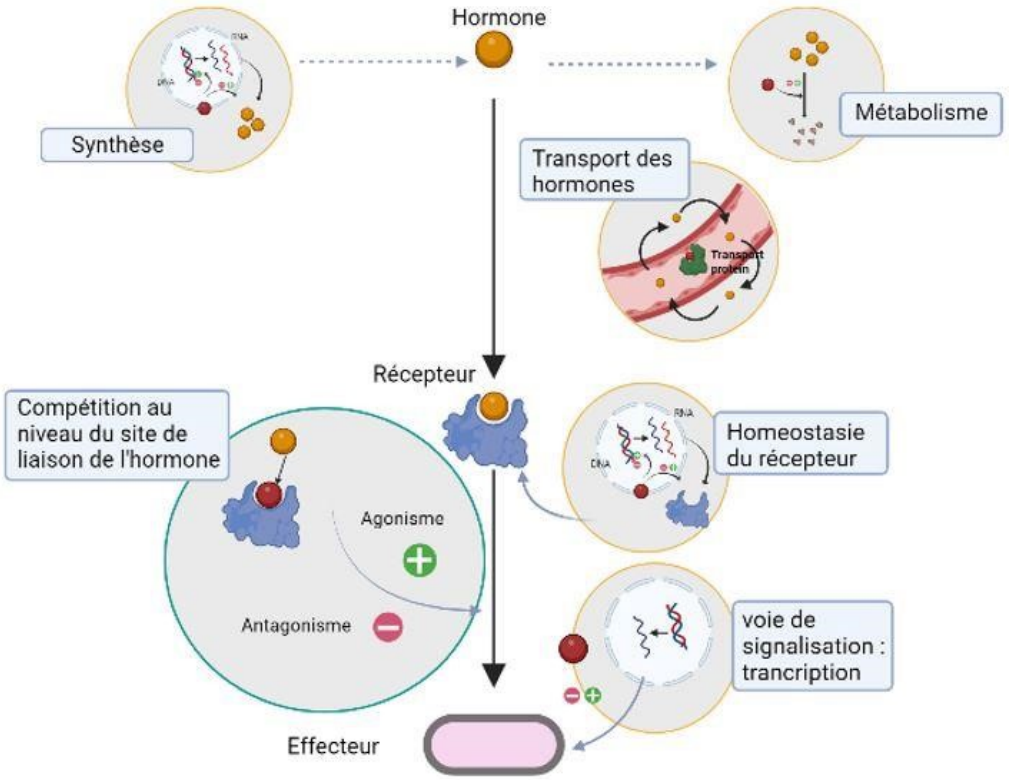


Figure 22 : Principaux mécanismes d'action des perturbateurs endocriniens. Les mécanismes directs sont encerclés en vert et le reste i.e mécanismes indirectes sont encerclées en jaune. Les flèches noires représentent la voie de signalisation des hormones (sphère jaune) via les récepteurs nucléaires. Les flèches grises sont en rapport avec l'effet des PE (sphère rouges) sur cette voie d'après [224,256]

4.3.3. Facteurs influençant les mécanismes d'actions

Certains facteurs doivent être pris en compte lors de la compréhension des mécanismes de perturbation endocrinienne comme (1) la période d'exposition, (2) l'effet cocktail, (3) la dynamique dose-réponse, (4) la variabilité inter-individuelle et (5) la complexité des mécanismes.

4.3.3.1. Période d'exposition et latence de réponse

Plus l'âge d'exposition est précoce et plus l'impact de l'exposition aux PE sera important. Un fœtus, un nourrisson ou un enfant seront plus sensibles notamment aux substances affectant leur développement et croissance. Notons qu'il peut exister une certaine latence entre la période d'exposition et l'apparition des effets. On parle de plusieurs années à plusieurs décennies voire l'apparition d'effets sur la descendance [257].

4.3.3.2. Effet « cocktail »

Le terme effet cocktail est utilisé pour décrire l'effet due à une poly-exposition à différents PE. L'importance de cet effet s'est considérablement accru au cours des dernières décennies en raison de leurs effets significatifs, même à faibles doses, par rapport aux effets des substances chimiques individuelles [258]. L'effet cocktail peut se manifester comme une addition ou une synergie d'actions.

4.3.3.3. Dynamique dose-réponse

Beaucoup de PE n'obéissent pas au paradigme de base de la toxicologie : le principe de dose réponse qui stipule que l'effet toxicologique observé est dépendant de la dose de produit assimilée. Cela peut clairement s'observer à travers les courbes doses-réponses de ces substances (courbe en U inversé). Pour certains PE, de faibles doses entrainent parfois un effet plus prononcé que des doses plus grandes [257].

4.3.3.4. Variabilité inter-individuelle

Ce paramètre découle de la variabilité génétique de métabolisme des xénobiotiques ce qui inclut par conséquent les PE. Ainsi la variabilité de métabolisme induit une variabilité de susceptibilité à ces substances entre les individus [257,259].

4.3.3.5. Complexité des mécanismes d'action

Un seul PE peut avoir plusieurs mécanismes d'actions dépendants de la cible. Ainsi un même PE peut être à la fois oestrogénique et anti-androgénique. De plus, les PE comme les autres xénobiotiques sont métabolisés conduisant à des métabolites pouvant eux aussi avoir des effets différents [257].

4.4. Physiopathologie des perturbateurs endocriniens

Bien que les premiers effets observés pour le PE soient liés à la reproduction [223], les PE exercent une pléthore d'activités sur le système endocrinien qui leur est conférée par la diversité de leurs mécanismes d'action ainsi que de leurs cibles. Dans cette partie nous explorerons quelques effets.

4.4.1. Effets sur la fonction de reproduction

Les manifestations engendrées par l'exposition aux PE sur le système reproductif touchent à la fois les fonctions de reproduction féminine et masculine. En effet, la fonction et la maturation

normales des glandes et de l'appareil reproducteur sexuel sont affectées par les PE au cours du processus de développement [245].

4.4.1.1. Système de reproduction féminin

Les effets des PE sur l'appareil reproducteur féminin commencent dès le plus jeune âge. En effet, chez les filles, ces composés peuvent causer des anomalies lors des divisons méiotiques avant ou après la puberté conduisant à des manifestations tardives comme l'aneuploïdie (nombre anormal de chromosomes), l'insuffisance ovarienne et l'augmentation de la fréquence des fausses couches. A l'âge adulte et plus particulièrement lors de la grossesse, les PE pourraient également être mis en cause dans de nombreux problèmes tels que l'avortement spontané, la grossesse extra-utérine, la mort fœtale, la mortinaissance, l'accouchement prématuré, le faible poids de naissance ou la modification du sex-ratio. De plus, les interférences dues aux PE engendrent des effets sur la fertilité féminine notamment ceux incorporant des métaux lourds. L'effet de ces derniers ainsi que les organochlorés sur le cycle menstruel a été démontré [260,261] comme affectant entre autres la fécondité [262–267].

Par ailleurs les PE sont suspectés d'être à l'origine d'autres pathologies touchant les glandes mammaires et l'endomètre tels que l'endométriose (maladie gynécologique liée à la présence de muqueuse utérine en dehors de la cavité utérine) ou encore l'apparition de fibromes utérins.

4.4.1.2. Système de reproduction masculin

Le système reproductif masculin peut également être perturbé par les effets d'une substance ou d'un mélange de substances chimiques perturbatrices du système endocrinien. On pense que les PE augmentent la fréquence d'anomalies graves telles que le cancer des cellules testiculaires [268], l'infertilité notamment par la réduction de la qualité du sperme et la manifestation pathologique de troubles urogénitaux, notamment la cryptorchidie (non-descente des testicules) et l'hypospadias (le positionnement anormal de l'ouverture de l'urètre) [269–271]

4.4.2. Effet sur le métabolisme

L'obésité est une maladie métabolique communément associée à une suralimentation couplée à un manque d'exercice physique. Récemment, l'accroissement accrue de ce phénomène a poussé les scientifiques à se pencher sur le sujet et le rôle des contaminants chimiques dans la prévalence de l'obésité a été soulevée parmi les 10 causes principales [272,273]. Les PE induisant l'obésité, appelés obésogènes, ciblent les régulateurs de transcription impliqués dans le contrôle de l'homéostasie des lipides intracellulaires ainsi que dans la prolifération et la différenciation

des adipocytes. Les principales cibles de ces PE sont les NR activés par les proliférateurs de peroxysomes (PPARα, δ et γ)[219]. Afin de fonctionner correctement, ces derniers doivent former un hétérodimère avec un autre récepteur : RXR lui-même obésogène mais avec un degré moindre [274]. De plus, PPARγ joue un rôle clé dans la biologie des adipocytes et est considéré comme le régulateur principal de l'adipogenèse [275]. Plusieurs perturbateurs endocriniens sont connus pour affecter l'activité de PPARγ et induire l'adipogenèse comme les organoétains tels que le tributylétain et le triphénylétain et certains phtalates [276,277]. D'autres perturbateurs endocriniens sont connus pour favoriser l'adipogenèse, mais n'agissent probablement pas par le biais de PPARγ. Il s'agit du BPA, des pesticides organophosphorés, du glutamate monosodique et des PBDE [278,279]. Bien que plusieurs perturbateurs endocriniens soient associés à l'adipogenèse et à l'obésité dans des modèles animaux, le tributylétain est le seul perturbateur endocrinien connu pour avoir des effets *in utero* sur les adipocytes via l'activation de PPARγ [280]. D'autres perturbateurs endocriniens sont susceptibles de favoriser l'adipogenèse *in utero*, bien que cela puisse être secondaire à des déséquilibres métaboliques plus larges. Par exemple, certains PCB et PBDE réduisent la fonction thyroïdienne, tout comme le triclosan, un composé antibactérien [281,282]. Les mécanismes d'action ne sont pas tout à fait certains, mais les modes possibles incluent une interférence avec la synthèse, le transport, le métabolisme ou la clairance des hormones thyroïdiennes [283].

4.4.3. Oncogenèse

La fréquence des cancers augmente considérablement dans les pays industrialisés. Ce groupe de maladies est causé par une dérégulation du cycle cellulaire et des changements dans les niveaux d'expression des gènes liés à ce dernier [245]. Les cancers restent des maladies multifactorielles mais la probabilité de développement de nombreux cancers est augmentée par l'exposition à certains PE comme les cancers du sein, les cancers testiculaires et les cancers de la prostate dont l'incidence a augmenté au cours des dernières années dans les pays industrialisés. En ce qui concerne les cancers du sein, les périodes de vulnérabilité particulièrement élevée d'exposition aux PE semblent être des périodes sensibles comme la puberté, la grossesse ou la ménopause [284,285]. Le cancer de la prostate est un cancer fréquent chez l'homme, mal diagnostiqué, et la signalisation hormonale stéroïdienne semble jouer un rôle essentiel dans sa formation et ses métastases [286]. La prostate exprime à la fois ER-α et ER-β, et la signalisation médiée par les hormones stéroïdes régule le développement des organes reproducteurs masculins et des caractéristiques sexuelles à l'âge adulte. Bien qu'il soit difficile

d'étudier l'association directe entre le risque de cancer de la prostate et l'exposition aux PE chez l'homme [287], la prolifération des cellules cancéreuses de la prostate est stimulée par l'exposition aux PE dans les modèles animaux [288].

4.4.4. Système nerveux

Les PE peuvent affecter le développement du système nerveux et même entrainer les dysfonctionnements neurologiques [289,290], l'apparition de maladies mentales [291], les modifications du comportement de reproduction (réceptivité sexuelle et instinct maternel) ainsi que les troubles de la vision, de la cognition et de la mémoire [245]. La majorité de ces effets sont dus à l'exposition aux xénœstrogènes survenant lors du développement intra-utérin [292] alors que d'autres sont causés par l'interférence avec certains neurotransmetteurs comme la dopamine et la noradrénaline [293]. En plus du système nerveux, le système neuroendocrinien [294] contrôle diverses fonctions importantes telles que la reproduction, les réponses au stress, la croissance, la lactation, le métabolisme, l'équilibre énergétique et d'autres processus permettant à l'organisme de réagir à son environnement [295–297]. La littérature documente surtout les effets sur l'axe hypothalamus-hypophyse-gonades ainsi que le système neuroendocrine thyroïdien [219].

4.5. Caractérisation *in vitro* des PE

Afin de démontrer qu'une substance est un perturbateur endocrinien, Il faut montrer 1) qu'elle engendre un effet indésirable et 2) qu'un lien existe entre cet effet indésirable et cette action endocrinienne puis déterminer la dose toxique.

La première étape critique de l'identification du danger doit être correctement conçue pour garantir une évaluation précise de la sensibilité des populations humaines et animales aux substances chimiques. Ainsi un PE doit être identifié à la fois en termes de mode d'action (capacité à interférer avec l'action des hormones) et de capacité à produire des effets indésirables/dangereux. La définition d'un PE doit se concentrer sur sa capacité à interférer avec l'action des hormones plutôt que de stimuler l'apparition d'effets indésirables. Une nuance importante à considérer est de ne pas caractériser comme PE tout produit interférant avec un aspect de l'action hormonale. Le risque dépend de l'exposition et de la puissance de la substance chimique. Quant à l'estimation de la capacité d'un produit à provoquer des effets néfastes (la puissance), elle est aussi compliquée que l'étude du rôle du système endocrinien dans le développement et la physiologie de l'adulte.

Par conséquent, le criblage et les tests biologiques de caractérisation des PE et l'estimation de leur toxicité nécessitent des connaissances dérivées des principes de l'endocrinologie, leurs effets et les conséquences du dérèglement hormonal et des maladies endocriniennes [219]. Les mécanismes d'action détaillés plus haut (c.f. paragraphe 4.3 Mécanismes d'action) peuvent facilement être investigués *in vitro* aux moyens de tests cellulaires et biochimiques [224]. Ainsi le criblage pour un effet donné peut se faire rapidement voire à haut débit et fournir des résultats aussi bien qualitatifs que quantitatifs. Les tests le plus communément utilisés pour évaluer l'action des PE sont les tests de liaison (*binding*), les tests de prolifération cellulaire et les tests de transactivation [298].

4.5.1. Tests de liaison ou tests de *binding*

Les mécanismes directs peuvent être mis en évidence grâce aux tests de *binding*. Il s'agit de déterminer la concentration à laquelle un potentiel PE est capable de déplacer un ligand de référence (généralement une hormone) au niveau de la protéine cible. Les tests de binding ont été largement utilisés [298,299] parce qu'ils sont faciles à réaliser, rapides et relativement bon marché, ce qui en fait un excellent choix pour le criblage à grande échelle. Néanmoins, certaines limites des tests de liaison aux récepteurs existent. La première limite de ce test c'est qu'ils ne permettent pas de déterminer le profil agoniste ou antagoniste du ligand [298,300]. De plus, les tests de *binding* ne sont pas valables pour la détection de composé nécessitant un métabolisme préalable à la liaison aux récepteurs comme les pro-œstrogènes [298]. Finalement, les tests de *binding* permettent d'identifier uniquement les PE agissant via le mécanisme direct. Il est donc utile de compléter les résultats obtenus avec d'autres tests.

4.5.2. Tests de prolifération cellulaire

Le principe de ces tests est de mesurer la croissance d'une lignée cellulaire hormonodépendante. Les tests de prolifération ont l'avantage de pouvoir être utilisés pour évaluer de nombreux processus biologiques relatifs au fonctionnement cellulaire y compris les processus passant par la voie génomique. Malheureusement, ces tests présentent quelques inconvénients dont le premier est leur lenteur (quelques jours) ainsi que la faible spécificité du mécanisme d'action. Enfin et malgré leurs succès, ces tests peuvent présenter une variabilité inter-laboratoires due à la différence des souches cellulaires utilisées et des conditions opératoires [298].

4.5.3. Tests de transactivation ou tests de gènes rapporteurs

Les tests de gènes rapporteurs sont utilisés pour analyser la capacité d'une substance à activer la transcription d'un promoteur sensible aux hormones au niveau de cellules eucaryotes (généralement des cellules de mammifère ou de levure). Les cellules sont transfectées avec un vecteur d'expression codant pour le récepteur d'intérêt et un vecteur de gène rapporteur (composé d'un promoteur sensible aux composés étudiés, lié à un gène rapporteur). Ce gène rapporteur code généralement pour une protéine qui peut être facilement détectée et quantifiée à l'aide d'un substrat approprié. L'activation du récepteur par une substance d'essai entraîne la stimulation de l'expression du gène rapporteur après activation de l'élément de réponse. Les cellules de mammifères sont les plus utilisées et le test le plus employé pour étudier les PE est celui lié au gène rapporteur de la luciférase où la luminescence produite est proportionnelle à l'activité transcriptionnelle. Un avantage des tests de transactivation c'est d'être plus spécifique que les tests de prolifération et de permettre d'identifier la nature du composé étudié : agoniste ou antagoniste. En revanche, la réponse transcriptionnelle provoquée n'est pas toujours due à la fixation du ligand sur le récepteur. En effet, elle peut être le résultat d'une cascade réactionnelle par exemple [298].

4.5.4. Limites aux évaluation *in vitro*

Malheureusement, les effets cytotoxiques de certaines substances peuvent fausser les résultats des tests *in vitro* entrainant une difficulté dans l'interprétation des résultats [301]. Les PE représentent aussi un défi car leurs effets dépendent à la fois du niveau et du moment de l'exposition, et sont particulièrement critiques lorsque l'exposition a lieu pendant la période de développement. Par conséquent, le moment de l'exposition et ses limites quantitatives acceptables sont d'un intérêt majeur pour évaluer le risque [302].

Néanmoins, l'existence de seuils de dose pour les perturbateurs endocriniens continue à être débattue [303–307] car les courbes de réponse à la dose non monotone (NMDR) sont souvent considérées comme une propriété intrinsèque des PE par le grand public. Il s'agit plutôt d'une propriété dérivée de la complexité des régulations endocriniennes [306]. Il n'en reste pas moins qu'il peut être difficile de distinguer un seuil réel valide d'un seuil apparent, qui découle simplement des limites de détection du système expérimental utilisé. Si une molécule présente une courbe dose-réponse en forme de U dans un système expérimental donné, la branche descendante en U de la courbe doit être utilisée comme base pour déterminer le seuil si la réponse enregistrée est liée à l'effet indésirable de la molécule. Cela peut conduire à des limites

excessivement basses mais, au moins, cela est plus satisfaisant que de refuser, par principe, toute limite. Bien entendu, la détermination de la valeur témoin en l'absence totale de la molécule testée est primordiale pour démontrer un effet significatif, positif ou négatif, à ces très faibles doses [219,224].

De plus, des études ont montré que plusieurs produits chimiques considérés séparément n'avaient pas d'effet observé mais que lorsqu'ils étaient présents simultanément en tant que mélange, ils étaient associés à un effet négatif. Cela attire l'attention sur les études de mélange [308,309]. En effet de nombreuses réponses biologiques sont la convergence de plusieurs voies de signalisation [220]. Bien que la question des mélanges ait été soulignée dans le rapport de l'EDSTAC [310](EDSTAC, 1998), il n'existe actuellement aucun mécanisme permettant d'intégrer cette question dans les stratégies de test *in vitro* [218].

4.6. Caractérisation *in silico* des PE

Outre les tests *in vitro*, de nombreux travaux ont été réalisés pour prédire le potentiel perturbateur endocrinien des composés chimiques via des méthodes *in silico*[311]. Ces méthodes sont ainsi capables de traiter des données diverses et de relier la chimie au niveau atomique à l'activité biologique aux niveaux de la cellule, de l'organe et de l'organisme. Comme abordé plus tôt (c.f paragraphe 2.2 Place de la toxicologie *in silico*), ces méthodes sont variables en fonction de la nature des données qu'elles utilisent et selon leur demande en ressources informatiques. Par ailleurs les PE n'étant pas à l'origine conçus pour interagir avec la même cible protéique et en l'occurrence avec les NR, une grande variabilité chimique en découle. Ainsi, les méthodes de criblage SB en plus des LB trouvent toute leur légitimité. Les méthodes LB permettent d'extraire les propriétés chimiques saillantes pour une grande quantité de données structurellement liées ou pas (série chimique ou simplement un ensemble de composés connues pour être actifs). Les méthodes SB quant à elle permettent de réaliser des études plus prospectives en définissant les propriétés nécessaires à un ligand pour pouvoir interagir avec la cible.

5. Objectif de la thèse

L'objectif de cette thèse et de mettre à profit les méthodes de criblage virtuel dans un cadre toxicologique. Nous nous intéressons à une catégorie de composés appelés perturbateurs endocriniens (PE) agissant via une pléthore de mécanismes d'action. Dans le cadre de nos travaux, nous nous intéressons au mécanisme direct d'interaction avec une famille de protéines appelées récepteurs nucléaires (NR). Il s'agit de facteurs de transcription dont l'activité est régie en majorité par des hormones et responsables du fonctionnement normal du système endocrinien. Parallèlement, les NR et plus spécifiquement l'interaction directe de composés chimiques avec ces protéines sont aussi étudiés dans un contexte thérapeutique afin de trouver des traitements aux pathologies endocriniennes.

Le travail mené au cours de cette thèse s'articule en trois parties. Dans un premier temps, une revue sur les NR a été réalisée. Il s'agit d'une étude de plus de 100 articles dans laquelle nous faisons l'état des lieux de l'utilisation des méthodes de criblage virtuel appliquée aux NR pour les usages thérapeutique et toxicologique sur la période 2010-2020. L'analyse critique de ces résultats nous a permis d'entamer la construction de nos modèles de prédictions. Une preuve de concept a été réalisée sur un membre de la famille des NR : le récepteur aux œstrogènes alpha (ERα). Après avoir validé nos méthodes et résultats, nous avons étendu l'étude à 5 autres NR : les récepteurs aux androgènes (AR), les récepteurs aux œstrogènes beta (ERβ), les récepteurs aux glucocorticoïdes (GR), les récepteurs activés par les proliférateurs des peroxysomes gamma (PPARγ) et les récepteurs aux hormones thyroïdiennes alpha (TRα).

Deuxième partie

Résultats

1. Etat de l'art :

1.1. Les récepteurs nucléaires, cible thérapeutique et toxicologique

1.1.1. Introduction

L'objectif de cette publication était de rassembler dans une même référence les études faites sur une thématique phare du laboratoire : les récepteurs nucléaires. La réalisation d'un état de l'art est en effet une étape importante à faire en amont de la création de nouveaux modèles de prédiction. Cette revue de l'état de l'art permet notamment de comprendre quelles *méthodes in silico*, aussi bien LB que SB, ont déjà été utilisées et avec quel succès, de relever les principales difficultés rencontrées et de tenter de prendre en considération les recommandations faites par les différents groupes de chercheurs qui s'y sont intéressés.

L'article présenté ci-dessous est une revue extensive des études réalisées entre 2010 et 2020 portants sur les NR aussi bien dans le contexte d'applications thérapeutiques que toxicologiques. Ainsi, la réalisation de la revue s'est faite en deux étapes ; une première étape de collecte des articles scientifiques répondants aux critères de cette revue et une deuxième étape de tri et de classification des articles. Pour chaque modèle décrit dans un article, différents aspects ont été identifiés : (1) le type de NR étudié, (2) la base de données utilisée pour l'entrainement, le test et la validation, (3) la ou les méthode(s) *in silico* employée(s), (4) le caractère prospectif ou rétrospectif de l'étude, (5) le niveau de reproductibilité de l'étude et (6) le contexte de cette dernière, i.e. thérapeutique ou toxicologique. La revue a été publié dans le journal *Frontiers in Endocrinology* le 13 septembre 2022.

1.1.2. Publication

frontiers | Frontiers in Endocrinology

TYPE Review
PUBLISHED 13 September 2022
DOI 10.3389/fendo.2022.986016

Check for updates

OPEN ACCESS

EDITED BY
Marica Cariello,
University of Bari Aldo Moro, Italy

REVIEWED BY
Elena Piccinin,
University of Bari Aldo Moro, Italy
Jean-Marc A. Lobaccaro,
Université Clermont Auvergne, France

*CORRESPONDENCE
Nathalie Lagarde
nathalie.lagarde@lecnam.net

SPECIALTY SECTION
This article was submitted to Molecular and Structural Endocrinology, a section of the journal Frontiers in Endocrinology

RECEIVED 04 July 2022
ACCEPTED 08 August 2022
PUBLISHED 13 September 2022

CITATION
Sellami A, Réau M, Montes M and Lagarde N (2022) Review of *in silico* studies dedicated to the nuclear receptor family: Therapeutic prospects and toxicological concerns. *Front. Endocrinol.* 13:986016. doi: 10.3389/fendo.2022.986016

COPYRIGHT
© 2022 Sellami, Réau, Montes and Lagarde. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Review of *in silico* studies dedicated to the nuclear receptor family: Therapeutic prospects and toxicological concerns

Asma Sellami, Manon Réau, Matthieu Montes and Nathalie Lagarde*

Laboratoire GBCM, EA 7528, Conservatoire National des Arts et Métiers, Hésam Université, Paris, France

Being in the center of both therapeutic and toxicological concerns, NRs are widely studied for drug discovery application but also to unravel the potential toxicity of environmental compounds such as pesticides, cosmetics or additives. High throughput screening campaigns (HTS) are largely used to detect compounds able to interact with this protein family for both therapeutic and toxicological purposes. These methods lead to a large amount of data requiring the use of computational approaches for a robust and correct analysis and interpretation. The output data can be used to build predictive models to forecast the behavior of new chemicals based on their *in vitro* activities. This atrticle is a review of the studies published in the last decade and dedicated to NR ligands *in silico* prediction for both therapeutic and toxicological purposes. Over 100 articles concerning 14 NR subfamilies were carefully read and analyzed in order to retrieve the most commonly used computational methods to develop predictive models, to retrieve the databases deployed in the model building process and to pinpoint some of the limitations they faced.

KEYWORDS

nuclear receptors, in silico, endocrine disrupting chemicals, docking, pharmacophore model, QSAR

1 Introduction

Nuclear receptors (NRs) are a large family of transcription factors. They are involved in a wide variety of biological and physiological processes such as growth, metabolism, reproduction, cell proliferation, differentiation, development and homeostasis (1, 2). The NRs superfamily is composed of 48 members in humans and is divided in 7 subgroups. Apart

from SHP and DAX, all the NRs share the overall same architecture comprising 5 domains, named A to E, each of which playing a specific role. The E domain is a structurally conserved allosteric signaling region forming the ligand binding domain LBD. This domain includes 12 helices and 4 β-strands that create a buried hydrophobic ligand-binding pocket (LBP) able to interact with small molecules ligands. The LBD and in particular the helix H12, called the activation function helix (AF-H), change conformation upon ligand-binding facilitating the interaction of the LBD with co-activators or co-repressors proteins. The DNA binding domain (C domain) of the activated NR can bind to a specific DNA sequence and recruit co-regulator proteins that either promote or repress the DNA transcription and therefore, specific gene expression (1).

Although the mechanism of action of NRs is very well-adjusted, dysfunctions of their signaling pathways have been linked with diseases such as cancers, diabetes or auto-immune disorders. NRs biology and ability to interact with small lipophilic molecules in the LBD have led to their emergence as a major class of therapeutic drug targets for these diseases, accounting for more than 10% of FDA-approved drugs (1–5).

However, exogenous ligand binding may also cause dysfunctions of NRs pathways. As such, some (candidate) drugs side effects are related to their unforeseen interaction with NRs like nifedipine and the antifungal clotrimazole who are able to activate PXR (6). More recently, a category of compounds called the endocrine disrupting chemicals (EDCs), have been associated with NRs, especially the sex hormones receptors such as Estrogen Receptors (ER) and Androgen Receptors (AR). Although the concept has been introduced since 1958, the scientific outline has evolved to settle in 2012 on a definition proposed by the Endocrine society (7). EDCs are environmental compounds that can be found in fumes, cosmetic additives or pesticides that can enter the human body due to their high lipophilicity (8). They can affect the endocrine system through several mechanisms (8) (9) including "direct" and "indirect" mechanisms. In the "direct" mechanism, EDCs bind to human NRs (hNRs) LBD and dysregulate the functioning of the endocrine system either by excessively activating or repressing the associated biological activity (7, 9). In the "indirect" mechanism, EDCs can alter not only the synthesis, metabolism, transportation and fate of hormones in the body but also the hormone-receptors expression, the signal transduction as well as the epigenic behind (8). Up to 2019, more than 1000 suspected EDCs have been reported (10). It is to note that EDCs can also affect other protein families than the NRs (11–13). For this review, we will focus on the NRs related "direct" mechanism of action.

Identifying hNRs binders, i.e. compounds able to bind to hNRs, is important for both therapeutic and toxicologic purposes: to discover new therapeutic compounds for several NRs-related diseases, to predict potential off-target effects, to guarantee the safety of novel synthesized molecules and to identify potential EDCs in the exposome. *In silico* methods emerge as an asset to achieve this goal. The construction of predictive models of NRs ligand binding using *in silico* methods (14–16) enable to prioritize compounds that should be biologically evaluated to reduce the time, cost and technical issues associated with the experimental tests of a large number of compounds.

Previous reviews mostly focused on specific NRs as potential drug discovery targets (17, 18) while other evaluated models dedicated to EDCs prediction (7, 19). Herein, we present an exhaustive review of 89 *in silico* initiatives carried out on NRs on both therapeutic and toxicologic levels and published between 2010 and 2020. We provide a summary in order to 1) enumerate them in a referential, 2) list the available NRs-related data that can be used in modeling approaches and 3) learn from previous experiences to enable the elaboration of more accurate predictions.

2 General overview

This review focuses on 14 NRs subfamily members for which publications falling within the scope of our review were retrieved. 89 articles were carefully selected, read and analyzed. For each collected model, several aspects were identified including 1) the NR for which the model was developed, 2) the computational method used to build the model, 3) the DB used for the model training, testing and validation, 4) the level of reproducibility 5) if the study was prospective or retrospective, and 6) the purpose of the study (toxicological or therapeutic). The result of our bibliographic search is presented in Table 1–4 and point 2) to 4) are detailed in a following paragraph in this review.

In total, in this review, we identified 38 projects dedicated to the identification of therapeutic compounds, 44 to the prediction of toxicological compounds and 3 projects addressing both purposes. Distribution of publications related to each studied NR are depicted in Figure 1. The in silico methods used to construct the models are divided into two approaches: Ligand-Based (LB) and Structure-Based (SB) (Figure 2). The reviewed studies are also classified according to their study design into prospective and retrospective studies. Prospective studies are based on compounds for which no biological data is available for the query target. These compounds are subjected to virtual screening protocols in order to assess if they can be considered as potential NRs modulators. The obtained predictions are then validated by experimental assays. In retrospective studies, models that are developed aim at correctly forecasting the already known data activities. The models can then be used to achieve predictions for compounds with unknown activities as long as they remain within the activity domain. The reviewed papers were perfectly balanced in terms of study design as 43 model were retrospective and 43 were prospective with the latter including a majority of LB methods.

3 Studied nuclear receptors

In addition to the IUPAC classification, NR can be distinguished according to structural differences, functions,

TABLE 1 Review of the different initiatives dedicated to Steroid Hormones nuclear receptors.

Receptor	methods	Approach	Database	Reproducibility	Prospective or Retrospective	Application	Year	Ref
AR	COMPARA	both	1,746 compounds from ToxCast/Tox21	Medium	retrospective	Toxicological	2020	(20)
AR	QSAR : Machine learning methods (kNN, lazy IB1, and ADTree methods)	LB	In house data (292 compounds) and collection from literature (231 compounds)	Medium	prospective	Toxicological	2010	(21)
AR	Docking and 3D QSAR (CoMSIA)	both	Collected from the literature (76 compounds)	Medium	retrospective	Toxicological	2013	(22)
AR	Docking, MD, and 3D QSAR (Comsia)	both	In house database of flavonoids (21 compounds)	Medium	retrospective	Toxicological	2016	(23)
AR	Docking	SB	Collected from the literature (20 bisphenols compounds)	Medium	retrospective	Toxicological	2016	(24)
AR	Docking	SB	EPA (1689 compounds)	Medium	prospective	Toxicological	2017	(25)
AR	Docking + molecular dynamics	SB	NR-List BDB (3233 compounds)	Medium	retrospective	Toxicological	2018	(26)
AR	QSAR : Machine learning methods (Bernoulli Naive Bayes, RF, NNN)	LB	COMPARA calibration set (1689 compounds from EPA) and external validation set (3882 compounds from EPA)	High	retrospective	Toxicological	2019	(27)
AR	QSAR : machine learning methods (ANNs,SVM,DT)	LB	CoMPARA dataset (1689 compounds from EPA), EDKB (202 compounds)	Medium	retrospective	Toxicological	2019	(28)
AR	Docking and LB Pharmacophore	both	NR-DBIND (812 compounds), Tox21 (5690 compounds)	High	retrospective	Both	2019	(29)
AR	QSAR : Machine learning (Bayesian models, RF, kNN, SVM, naïve Bayesian, AdaBoosted DT) and DL	LB	Toxcast (8645 compounds)	High	retrospective	Toxicological	2020	(30)
AR, ER	QSAR: KNN + Local method (lazy learning) + RF	LB	METI (900 compounds), EDKB (87 compounds)	Medium	prospective	Toxicological	2010	(31)
AR, ER	QSAR: ML (kNN, DT, NB SVM)	LB	Collected from the literature (1157 compounds in the training set and 121 compounds in the external validation set)	High	retrospective	Toxicological	2014	(14)
AR, ER	QSAR: ANN	LB	Collected from the literature (879 compounds for ER, 930 compounds for AR)	High	Retrospective	Toxicological	2015	(32)
AR, ER	3D QSAR and bayesian statistics	LB	Toxcast (1853 compounds) + 42 compounds	Medium	retrospective	Toxicological	2016	(33)
AR, ER alpha	Hierarchical charactarestic fragments, docking and MD simulations	both	ToxCast/Tox21 and ChEMBL (2458 compounds for ER, 2843 compounds for AR)	Medium	prospective	Toxicological	2020	(34)
AR, GR	Similarity	LB	Toxcast (7027 compounds for AR, 7329 compounds for GR)	High	retrospective	Toxicological	2020	(35)
ER	CERAPP	both	Collected from the literature (1677 compounds)	Medium	retrospective	Toxicological	2016	(36)
ER	QSAR: ANN	LB	Collected from the literature (174 compounds)	Medium	Both	Toxicological	2010	(37)
ER	QSAR (single task an multi task learning KNN) and docking	both	Collected from the literature including EDKB and ChEMBL (QSAR data sets: 546 compounds for ERa, 137 compounds for ERb; docking data sets: 106 binders/ 4018 decoys for ERa, 80 binders/ 2000 for ER b)	Medium	Both	Toxicological	2013	(38)
ER	QSAR:machine learning methods (LDA / CART/ SVM)	LB	Toxcast (1814 compounds) and Tox21 (8303 compounds)	Low	retrospective	Toxicological	2013	(39)
ER	Docking	SB	EPA (1677 compounds)	Medium	retrospective	Toxicological	2015	(40)
ER	Docking and QSAR :machine learning methods (LDA, decision tree, SVM)	both	Collected from the literature (440 compounds)	Medium	retrospective	Toxicological	2016	(41)
ER	QSAR: Machine learning (Bernoulli Naive Bayes,	LB	Collected from the literature (1677 compounds from the CERAPP data set, 7351 compounds from	High	retrospective	Toxicological	2018	(42)

(Continued)

TABLE 1 Continued

	AdaBoost Decision Tree, RF, SVM) and deep learning (DNN) methods		Tox21, 3474 compounds for ERa, 2775 compounds for ERb)					
ER	QSAR: Machine learning method (GkNN)	LB	ToxCast and CERAPP databases (1677 compounds)	Low	retrospective	Toxicological	2018	(43)
ER	QSAR: Machine learning method (Bayesian models)	LB	"Toxcast2019" and two publications	Medium	Both	Toxicological	2020	(44)
ER	QSAR: Machine-learning (BNB, kNN, RF, and SVM) and deep learning (DNN) methods	LB	ToxCast and Tox21 (7576 compounds)	Medium	retrospective	Toxicological	2020	(45)
ER alpha	3D QSAR + 2D QSAR : machine learning methods (PLS, SVR, LR)	LB	In house (68 raloxifene's derivatives)	Low	prospective	Therapeutic	2013	(46)
ER alpha	Docking	SB	Ligands extracted from cristallographic complexes (66 compounds) and DUD-E's set (106 binders, 4018 decoys)	Medium	retrospective	Therapeutic	2014	(47)
ER alpha	Docking and aggregated potential field similarity	both	NCTRER binding database, ChEMBL, DUD (1691 active and 4785 inactive/decoy compounds) and Tox21 for prospective screening	Medium	prospective	Toxicological	2014	(48)
ER alpha	Docking	SB	Drug-Bank Database and collection from literature (105 compounds)	Medium	prospective	Therapeutic	2019	(49)
ER alpha	QSAR : Machine learning (RF)	LB	EABD (3308 compounds) and Toxcast (1641 compounds)	Medium	retrospective	Toxicological	2015	(50)
ER alpha and ER beta	QSAR: ANN	LB	Collected from the literature (170 compounds)	Low	retrospective	Toxicological	2011	(51)
ER beta	LB Pharmacophore modeling and QSAR (MLR)	LB	Collected from the literature (119 compounds) and NCI list of compounds for prospective screening	Medium	prospective	Therapeutic	2010	(52)
ER beta	LB pharmacophore modeling and docking	both	Maybridge and Enamine	Low	prospective	Therapeutic	2014	(53)
ER beta	docking and MD simulations	SB	18 ligands from crystal structures, 40 compounds collected from the literature and 2570 DUD decoys, 400000 compounds from commercial databases for prospective screening	Medium	prospective	Therapeutic	2014	(54)
ER beta	QSAR: Machine learning methods (Naïve bayes, KNN, RF, SVM)	LB	CHEMBL20 (356 active compounds and 107 inactive compounds) + 249 DUD-E decoys	Medium	retrospective	Therapeutic	2016	(55)
ER beta	QSAR: Machine learning (RF)	LB	EADB (2492 compounds) and ToxCast (1805 compounds)	Medium	retrospective	Both	2017	(56)
PR	Docking, MD, Binding energy calculation	SB	Collected from the literature (12 compounds); ZINC db (20000 compounds) for prospective screening	High	prospective	Therapeutic	2018	(57)

tissue specificity, DNA binding motifs or the knowledge or not about an endogenous ligand. For this review, we divided the studied NR into three groups according to their structural similarities and function in three groups. 1) steroid hormones receptors, (Table 1) 2) RXR and its partners (Table 2) 3) monomeric orphan receptors (Table 3) (107, 108).

3.1 Steroid hormones receptors

The NR belonging to this class are all activated by an endogenous ligand presenting a steroid core. Upon binding to their native ligand, steroid hormones receptors undergo conformational changes leading to their homodimerization and subsequently to DNA binding. The reviewed articles dedicated to the steroid hormone receptors are listed in Table 1.

3.1.1 Estrogen receptors

Two isoforms of ER exist, namely ERα and ERβ. Both isoforms exhibit similar affinity for their native ligand, 17β Estradiol, but differential expression in the body and unique roles in estrogens action *in vivo* (55). Indeed, ERα's effects are prominent in the mammary gland, uterus and in the preservation of skeletal homeostasis and the regulation of metabolism, while

TABLE 2 Review of the different initiatives dedicated to RXR and its partners NR.

Receptor	methods	Approach	Database	Reproducibility	Prospective or Retro-spective	Application	Year	Ref
FXR	SB pharmacophores	SB	ChEMBL (221 compounds); NCI database (247041 compounds) for prospective screening	Medium	prospective	Therapeutic	2011	(58)
FXR	SB pharmacophores	SB	in-house Chinese Herbal Medicine database (10216 compounds) for prospective screening	Low	prospective	Therapeutic	2011	(59)
FXR	LB Pharmacophore and free energy calculations	LB	ChemBridge (~520000 compounds) for prospective screening	Low	prospective	Therapeutic	2015	(60)
FXR	QSAR: Machine learning methods (SVM, C4.5 DT, k-NN, RF, NV), MoSS and SARpy	LB	Tox21 (688 compounds), ChEMBL (460 compounds), D3R CG2 (76 compounds)	Low	retrospective	Toxicological	2018	(61)
FXR	QSAR: Machine Learning (counter-propagation artificial neural network, kNN)	LB	ChEMBL (896 compounds), Asinex (3383942 compounds) for prospective screening	Medium	prospective	Therapeutic	2018	(62)
LXR	SB pharmacophore and shape similarity	both	Collected from the literature 41 compounds + 67059 decoys from Derwent World Drug Index); NCI database (250761 compounds) for prospective screening	Medium	prospective	Therapeutic	2012	(63)
LXR	self-organizing maps (SOM)	LB	ChEMBL (458 compounds); DrugBank (1280 compounds) for prospective screening	Low	prospective	Therapeutic	2017	(64)
LXR beta	2D fragment-based HQSAR and HQSSR (structure selectivity) and Docking	both	Collected from the literature (62 quinolines and cinnolines)	Medium	prospective	Therapeutic	2012	(65)
LXR alpha and LXR beta	Docking and MD	SB	ChEMBL database + DecoyFinder (769 compounds for LXRa, 570 compounds for LXRb); MolMall subset of the ZINC (~20000 compounds) for prospective screening	High	prospective	Therapeutic	2018	(66)
LXR beta	QSAR (MLR) and Docking	both	Collected from the literature (53 compounds with dual activity LRα/β)	Medium	prospective	Therapeutic	2018	(67)
PPAR alpha and PPAR gamma	2D-, 3D-QSAR and docking	both	In-house library (22 compounds)	Medium	prospective	Therapeutic	2013	(68)
PPAR alpha and PPAR gamma	QSAR, SB pharmacophore modelling and docking	both	In-house library (46 phenylpropanoic acid derivatives)	Medium	prospective	Therapeutic	2016	(69)
PPAR alpha and PPAR gamma	docking and MD	SB	Asinex (292,724 compounds) for prospective screening	Medium	prospective	Therapeutic	2018	(70)
PPAR alpha and PPAR gamma	docking, binding energy calculations, MD	SB	ChemDiv database (7476 compounds) for prospective screening	Low	prospective	Therapeutic	2019	(71)
PPAR alpha and	Docking and MD	SB	Ligand Expo components database	Medium	prospective	Therapeutic	2020	(72)

(Continued)

TABLE 2 Continued

Receptor	methods	Approach	Database	Reproducibility	Prospective or Retro-spective	Application	Year	Ref
PPAR gamma								
PPAR gamma	LB Pharmacophores and 3D QSAR	LB	Collected from the library (88 compounds)	Medium	retrospective	Therapeutic	2010	(73)
PPAR gamma	QSAR :Machine learning methods (MLR, SVM and Bayes Network Toolbox (BNT)), docking and MD	both	Traditional Chinease Medicine (TCM) database (9,029 compounds)	High	prospective	Therapeutic	2014	(74)
PPAR alpha and PPAR gamma	Docking and MD	SB	Compounds collected from the literature (51 compounds + 3600 DUD decoys); "clean-leads" ZINC's subset for prospective screening (740000 compounds)	Low	prospective	Therapeutic	2015	(75)
PPAR gamma	SB and LB pharmacophore-, shape similarity and docking	both	Collected from the literature (51 partial agonists, 14 agonists + 812 inactives from ToxCast and literature); Maybridge database (52000 compounds) for prospective screening	Low	prospective	Therapeutic	2016	(76)
PPAR gamma	docking and MD	SB	Zbc subset of ZINC database (180313 compounds)	Low	prospective	Therapeutic	2018	(77)
PPAR gamma	Docking, binding energy calculations and MD simulations	SB	Seaweed Metabolite Database (1110 compounds)	Medium	prospective	Therapeutic	2021	(78)
CAR	docking,SB and LB pharmacophore, QSAR (SVM)	both	Collected from the literature (392 compounds)	Low	retrospective	Therapeutic	2017	(79)
CAR, PXR	Docking	SB	Collected from the literature (106 compounds)	Medium	retrospective	Toxicological	2017	(80)
PXR	Docking and QSAR (Bayesian classification)	both	Toxcast (308 compounds)	medium	prospective	Toxicological	2010	(81)
PXR	QSAR : C5.0	LB	In-house collection (202 compounds) and collection from the literature (434 compounds)	High	retrospective	Both	2012	(82)
PXR	QSAR: partial logistic regression(PLR)	LB	Collected from the literature (631 compounds)	medium	retrospective	Both* (PXR activation is an unwanted side effects of drugs)	2012	(83)
PXR	QSAR, similarity	LB	Prestwick Chemical Library (1120 compounds)	Low	prospective	Both* (PXR activation is an unwanted side effects of drugs)	2015	(84)
PXR	SB Pharmacophore and docking	SB	Binding DB (266 compounds); PubChem (820 herbs compounds) for prospective screening	Medium	prospective	Both* (PXR activation is an unwanted side effects of drugs)	2015	(85)
PXR	SB Pharmacophore	SB	Collected from the literature (18 compounds), Mitsubishi Tanabe Pharma Corporation (68 compounds), NPC (2816 compounds)	Low	retrospective	Both* (PXR activation is an unwanted side effects of drugs)	2017	(86)
TR	Docking and MD simulations	both	Collected from the literature (16 HO-PBDEs compounds)	Medium	retrospective	Toxicological	2016	(87)

(Continued)

TABLE 2 Continued

Receptor	methods	Approach	Database	Reproducibility	Prospective or Retro-spective	Application	Year	Ref
TR	QSAR (C4.5 ,SVM and Random Forest)	LB	Collected from the literature (258 compounds)	Medium	retrospective	Toxicological	2019	(88)
TR beta	Docking and MD	SB	DUD-E (7556 compounds), in-house indoor dust contaminant inventory (485 compounds)	Medium	retrospective	Toxicological	2016	(89)
TR beta	3D QSAR, Docking and MD	both	Collected from the literature (33 compounds)	Medium	retrospective	Therapeutic	2015	(90)
TR beta	Docking and QSAR (PLS)	both	Collected from the literature (18 HO-PBDEs compounds)	Medium	prospective	Toxicological	2010	(91)
VDR	de novo design, docking, MD, free energy calculation	both	Fragments extracted from 6 VDR agonists collected from the literature	Low	prospective	Therapeutic	2012	(92)
VDR	Docking, LB Pharmacophores, 3D QSAR, MD	both	ChEMBL (478 compounds)	Medium	retrospective	Therapeutic	2020	(93)
VDR	LB pharmacophore, molecular docking, binding free energy calculation, Density Functional Theory (DFT) study and MD	both	Binding database (31 compounds); for prospective screening: Life chemicals, Enamine, MayBridge, and TCM	Low	prospective	Therapeutic	2020	(94)

the activation of ERβ impacts the immune and central nervous systems. Moreover, ERβ exerts an anti-proliferative and pro-apoptotic activity that counteracts ERα's actions towards cell growth and proliferation (54, 55). The established link between impairment of these ER pathways and diseases such as breast and endometrial cancers, osteoporosis, metabolic and cardiovascular diseases (55, 109) explains the number of published *in silico* studies with a therapeutic scope (8 up to the 23 ER-related reviewed studies). Additionally, ER are a well-known target for EDCs, and it has been shown that exposure to these chemicals increase the risk of breast cancer and immune diseases development (18). Consequently, several projects focused on this hNR as a central component of toxicological pathways including EDCs. It is the case of CERAPP, a large-scale screening project initiated by the US Environmental Protection Agency (EPA) (36). The project gathers models from American as well as European research groups to develop *in silico* models to evaluate thousands of chemicals for ER-related activity and

TABLE 3 Review of the different initiatives dedicated to monomeric orphan receptors.

Receptor	methods	Approach	Database	Reproducibility	Prospective or Retro-spective	Application	Year	Ref
ERR	Combination of QSAR models	LB	Tox21 (5077 compounds for ERR agonism, 6526 compounds for ERR inhibition); HMDB (3092 compounds) and EU pesticides dataset (888 compounds) for prospective screening	high	prospective	Toxicological	2019	(95)
ERR	molecular similarity and docking	both	KEGG COMPOUND database (10739 compounds)	high	prospective	Therapeutic	2013	(96)
LRH-1	Docking	SB	ZINC database (5.2 million compounds) for prospective screening	Low	prospective	Therapeutic	2013	(97)
RORγt	docking and similarity	both	ChEMBL (502 compounds); Specs commercial database (116495 compounds) for prospective screening	Medium	prospective	Therapeutic	2018	(98)
RORγt	SB pharmacophore and Docking	SB	Asinex Gold–Platinum (289174 compounds)	Medium	prospective	Therapeutic	2020	(99)

TABLE 4 Review of the different initiatives dedicated to projects targeting several NR.

Receptor	methods	Approach	Database	Reproducibility	Prospective or Retro-spective	Application	Year	Ref
AhR, AR, CAR, ER, ERR, FXR, GR,PPARd, PPARg, PR, PXR, RAR, ROR, RXR, TR, VDR	QSAR : Deep learning method (molecular image-based method)	LB	Tox21 Data Challenge 2014 (~7000 compounds / NRs)	Low	prospective	Toxicological	2020	(100)
AhR, AR, ER, PPARg	QSAR: Machine learning (RF) and Deep Learning (Deep Neural Network) methods	LB	Tox21 (10255 compounds curated from the original 12707 compounds)	Medium	retrospective	Toxicological	2016	(101)
AhR, AR, ER, PPARg	QSAR :Deep learning method (DNN)	LB	Tox21 (8694 compounds curated from the original 12707 compounds)	Medium	retrospective	Toxicological	2016	(102)
AhR, AR, ER, PPARg	QSAR: Machine learning (RF and SVM)/ Molecular similarity/ SB Pharmacophore modeling	LB	Tox21 (~7000 compounds / NRs)	High	retrospective	Toxicological	2018	(103)
AR, ER alpha, ER beta, GR, PPAR alpha, PPAR beta, PPAR gamma, PR, RXR alpha, and TR alpha and TR Beta	Docking	SB	DUDE-E, ChEMBL	High	retrospective	Toxicological	2014	(104)
AR, ER,GR, PPAR gamma, TR	QSAR : Machine learning methods (SVM, RF)	LB	Tox21 (7248 compounds)	High	retrospective	Toxicological	2019	(19)
AhR, AR, ERα, ERβ, GR, LXR, MR, PPAR gamma, PR, TR alpha, TR beta	Virtual Tox Lab software (docking and mQSAR)	both	Collected from the literature (1016 compounds)	Medium	prospective	Toxicological	2012 and 2014	(12, 105)
AR, GR, MR, PPAR alpha, PPAR beta, PPAR gamma, PR, RAR alpha, RXR alpha, TR beta, VDR	Docking	SB	Collected from the literature (157 compounds)	High	retrospective	Therapeutic	2010	(106)

prioritize them for further testing. CERAPP models, like most of the ER focused models reported in the literature, do not differentiate the two ER isoforms during the development of the models. It is to note that this distinction is crucial for therapeutic projects as many ligands display different affinities for each isoform (110). Designing selective ERβ ligands could help reducing ER related side effects besides exerting the desired estrogenic activity. Niu et al. (55), succeeded in building machine leaning models with high accuracies (77.10 to 88.34%) and AUCs greater than 0.8 that performed equally on an external validation dataset composed of ERβ selective agonists. It is to note that the negative data used for this study consisted of generated decoys rather than selective ERα ligands. Another particularity we could notice for ER models, was the predominance of Quantitative Structure Activity Relationship (QSAR) models based on machine learning (ML) algorithms. The most used ML algorithms were random forest, SVM and Naïve Bayes. Other computational approaches comprise molecular docking and pharmacophore modelling.

3.1.2 Androgen receptors

Modulated by dihydrotestosterone and testosterone, AR is involved in several physiological processes such as the male sexual differentiation, the development and maintenance of musculoskeletal and cardiovascular systems, as well as functionality of female ovarian follicles and ovulation (29, 111). The malfunctioning of these receptors is linked to several diseases including prostate, testicular and ovarian cancers, impaired reproduction system development and neuromuscular diseases (29, 112). Despite this large therapeutic potential, the 23 models and projects identified for AR during this review process were for a large majority conducted in the EDCs risk assessment perspective. Among them, the CoMPARA project (20) is the AR counterpart of the CERAPP project and falls under the Endocrine Disruptor Screening Program of the U.S. Environmental Protection Agency (EPA). The 25 international research teams forming the CoMPARA consortium used a common training set of 1,746

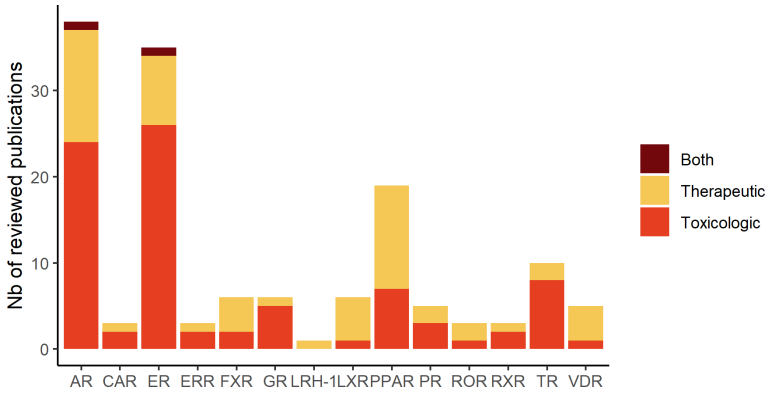


FIGURE 1
Number of publications related to each studied hNR subfamily described in the review.

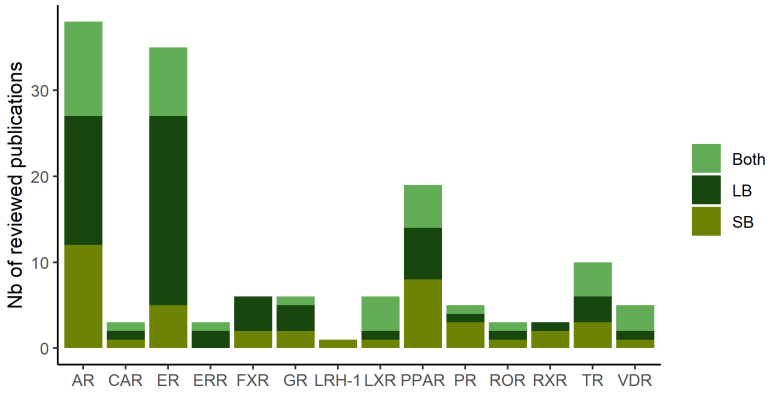


FIGURE 2
Distribution of the computational approach (SB: structure-based, LB: ligand-based and both: combination of SB and LB methods) in the reviewed publications for the different hNR.

chemicals compiled from a data set of 11 ToxCast/Tox21 HTS *in vitro* assays to generate 91 predictive QSAR models for AR binding, agonist, and antagonist activity predictions. The resulting models were evaluated using non-overlapping curated literature data. Finally, all predictions were combined into consensus models with an average accuracy of 80%. In complement to the CoMPARA project, several computational LB models, obtained using deep learning and QSAR methods, were described in the literature. SB methods such as docking and MD were also used, albeit to a lesser extend because of a lack of AR structural data. In fact, there is no AR antagonist-bound crystal structure yet available in the PDB which can impair the prediction of AR antagonist molecules and the development of AR SB models. However, some groups managed to overcome this issue and relied on molecular dynamics simulations to generate AR structures in antagonist conformations that led to a vast improvement for the docking of AR antagonists (26).

3.1.3 Progesterone receptor

PR is expressed primarily in female reproductive tissues and in the central nervous system as two isomers, PRα and PRβ. Both isomers present identical structures except for the A/B domain but they target distinct gene networks in progesterone-responsive cells (113). PR is activated by binding its endogenous ligand, the progesterone. PR regulates a wide range of biological function in a context- and cell-specific manner including the development and differentiation processes in normal and malignant female tissues, in particular in patients suffering from breast cancer (114). Selective progesterone modulators (SPRMs) have thus been developed as drug candidates for breast cancer therapies (57). Moreover, PR is suspected to be involved in endocrine disruptor toxicity pathway as bisphenols and some pesticides have been characterized with a PR activity (107). The reviewed models reflect the involvement of this NRs

in both aspects. Three *in silico* endocrine disruptors prediction tools (12, 100, 104, 105) include PR models. In particular, a prediction model of PR agonist compounds was constructed using a deep learning approach (100) and the Tox21 data. This model was associated with high prediction performance as measured by a Matthew correlation coefficient superior to 0.8 and an AUC value of 0.95. A structure-based prospective virtual screening to identify new PR inhibitors was also conducted (57) but no experimental validation of the results was provided.

3.1.4 Glucocorticoid receptors

Besides their well-known anti-inflammatory, anti-proliferative, pro-apoptotic and anti-angiogenic roles, glucocorticoids are involved in several other physiological processes affecting the nervous, cardiovascular, immune and respiratory systems (115, 116). Additionally, GR is incriminated in several toxicological pathways resulting in decreased male fertility and depression (35) and are a potential target for EDCs (19). GR is thus a relevant target for both therapeutic and toxicologic compounds. There are several isoforms of GR distributed through various tissues. GRα and GRβ present a similar sequence that differs only in the C-terminal region, whereas GRγ, GR-A, and GR-P are less characterized. It has been shown that GRβ negatively regulate the action of GRα as well as exerting its own. The remaining isoforms are associated to glucocorticoids insensitivity (115). In this review, only one therapeutic project was identified (106) whereas, 4 projects with toxicological applications for GR were found. Both SB, LB and their combination were applied to identify GR ligands. None of these projects did the distinction between all isoforms despite the difference in affinities towards glucocorticoids (115). It is to note that all the available GR structures in the PDB include the same mutation in helix 5 from a Phenylalanine residue to a Serine to overcome solubility problems during the crystallization experiments (117). However, this mutation is not located in the ligand binding site.

3.2 RXR and its partners

RXRα presents the particularity of forming heterodimers with one third of the 48 human NR (107) and is thus able to regulate a wide range of biological functions in a cell and tissue-specific manner (118). This class of NR can be subdivided in permissive heterodimers that can be activated by ligands of either RXR or its partner in the dimer (PPARs, LXRs, FXR, PXR and CAR) and non permissive heterodimers that generally require the RXR to be unbound in order to be activated by the native ligand of the dimer partner (TRs, RARs) (107). The reviewed articles focusing on RXR and its partners are presented in Table 2.

3.2.1 Retinoid X receptor alpha

As previously mentioned, RXRα can form heterodimers with several NR partners, but it can also be assembled and activated as an homodimer (108). A synthetical RXRα ligand, bexarotene, is notably marketed for the treatment of cutaneous T cell lymphoma (119). RXRα ligands have also demonstrated neuroprotective properties and are considered for the design of drug candidates for Alzheimer disease treatment (120). No study specifically dedicated to RXRα ligands prediction was identified during this review. However, 3 articles presenting models for several NRs including RXRα were available. Among them, two studies used SB methods. However, one of this study (106) is more focused on the comparison the performance of single and ensemble docking approaches with an active RXR data set limited to 11 compounds. In the other SB study, the RXR DUD-E data set (121) was used with a single docking approach to select the RXRα structure. Overall good performance in predicting RXR ligands was obtained with an AUC value close to 0.8 and an early enrichment factor value EF1% of 8.6. The third study presents a RXRα agonist ligands prediction model developed using a LB deep learning approach (100). The model was trained on the Tox21 database and was associated with an AUC value similar to those obtained in the previous study with the docking approach.

3.2.2 Peroxisome proliferator-activated receptors

Since the cloning of the first PPAR from the rat liver in 1990, the subfamily of PPAR has been enlarged and currently counts three members PPARα, PPARδ and PPARγ. Each isoform is differentially expressed in tissues and associated with different biological functions (122): PPARα is mainly expressed in the liver and regulates lipid metabolism (123); PPARγ is mainly expressed in adipose tissues and controls adipogenesis and carbohydrate metabolism (124); PPARδ is ubiquitously expressed and involved in atherosclerosis, pathologies of the nervous system, embryonic, organ and tissue development and metabolism of lipid and glucose (122). PPARδ is the less studied receptor of this group (75) while PPARα and PPARγ have been extensively investigated as therapeutic targets especially in metabolic diseases like type 2 diabetes (T2D) (68, 71, 78, 125, 126). Recently, PPARγ has also been proposed as a therapeutic target for ovarian cancer (77) and Alzheimer Disease (AD) (78). In this review, 11 models and projects dedicated to PPARs have been identified. The implication of this receptor in several pathologies explains the fact that all project specific to PPAR fall under the therapeutic application. Indeed, 8 studies described prospective virtual screening protocols combining several in *silico* methods for the prediction of dual PPARα/γ agonist ligands with therapeutic applications for T2D (70–72)

and of PPARγ ligands (74–78). However, experimental validation was achieved for only 3 of these studies (75–77), which limits the evaluation of the protocols performance. The protocol presented by Kaserer et al. for PPARγ partial agonist ligands prediction (76) is of particular interest since a retrospective study was first achieved to select optimal models for the prospective screening. This protocol combines 3 pharmacophore models (associated with EF and AUC values of 6.5 and 0.92 respectively in the retrospective evaluation), 5 shape-based models (associated with EF and AUC values of 11.0 and 0.83 respectively in the retrospective evaluation), and a docking protocol (associated with EF and AUC values of 2.2 and 0.65 respectively in the retrospective evaluation). The top-ten ranked compounds by the three methods (29 compounds) were biologically evaluated and 9 novel PPARγ ligands were identified. Retrospective studies were also conducted on the PPARα/γ using 2D, 3D-QSAR, and docking (68) and to understand the structural factors responsible for PPARγ agonistic activity using a combined pharmacophoric/3D-QSAR approach (75). As PPARγ data is available from the Tox21 program, several studies trying to model the Tox21 data for a large panel of NRs also present PPARγ dedicated models (19, 100–103). All of these models were associated with high predictive power with AUC values comprised between 0.8 and 0.9.

3.2.3 Human pregnane X receptor

Due to its flexibility and the relative bulkiness of its binding site in comparison with other NRs (80), PXR activity can be modulated by binding to a wide range of endogenous compounds, from bile acids to steroid hormones but also xenobiotics, such as drugs or environmental chemicals (pesticides, phenols, cosmetics, phytoestrogens) that can dysregulate normal physiological functions (127–129). PXR is responsible for the modulation of the expression of enzymes and transporters associated with the metabolism and transport of several drugs. Thus, PXR can mediate drug-drug interactions either by reducing the therapeutic efficacy or by increasing the concentration of reactive metabolites leading to toxicity (86). PXR may also lead to the so called "cocktail effect" that is the adverse effects caused by several chemicals present at the same time exhibiting low individual toxicities (83, 129). In this sense, it is important to identify compounds that activate PXR to avoid such effect and modify the drug design during early stages. Several projects focused on developing models able to identify PXR agonists: Torimoto-Katori et al. (86), designed a pharmacophore with high accuracies (over 0.7) yet with lower sensitivities suggesting that reinforcing the methods with other *in silico* methods could help achieve better performances. This was done by Cui et al. (85), who combined pharmacophore and docking method to detect ingredients from herbs able to activate PXR in order to avoid herb-drug effects. Pharmacophore models achieved similar performances as the model described before with sensibilities equal to 0.54 and specificities of 0.8. Interestingly, adding docking methods on top of pharmacophore models enhanced the performances to a detection rate of 0.6 i.e. the ratio of positive hits among all the database compounds. Three other groups built QSAR models to identify potential activators of PXR. The first one (84), achieved performances of $R^2 = 0.64$ and was applied to identify 16 novel activators of PXR. Dybdahl et al. (83), built a model associated with a sensitivity of 82% and a specificity of 85% that was used to identify potential PXR activators in a database of environmental chemicals. Interestingly, these molecules were also linked to cause adverse effects such genotoxicity or teratogenicity for examples.

3.2.4 Liver X receptor

LXR presents two isoforms (LXRα and LXRβ) that are both ubiquitously expressed (65, 67, 130–132). LXR serves primarily as a reverse cholesterol transporter within the lipid metabolism, protecting cells from cholesterol excess (66, 67) and is involved in many physiological processes. Thus, LXR represents a promising therapeutic target for cardiovascular diseases, dyslipidemia and cancer treatment (66). However, it has been shown that LXRα activation can lead to undesired lipogenic effects like increased hepatic lipogenesis or liver steatosis whereas LXRβ activation does not and can even reduce them (67). Designing LXRβ selective compounds to treat dyslipidemia appears as a promising strategy yet difficult to achieve due to the high similarity between both isoforms LBD. For this reason, only two projects among the reviewed ones were dedicated to the identification of LXRβ selective agonists whereas the remaining two others did not make any distinction. Both projects used a virtual screening workflow based on a combination of LB and SB methods. In the first article, Chen et al. (67) built a selective LB model using an association of Kohonen maps and stepwise multiple regression. The model relied on the structures of newly reported dual agonists and was associated with R^2 equal to 0.837 and 0.843 for train and test set respectively. This model was then used to perform predictions of potential selective ligands from the ZINC database falling within the applicability domain of the model. A promising compound was found with a predicted pEC50 = 7.0 for LXRβ and pEC50 = 6.095 for LXRα and was used as template to design potential inhibitors. The latter molecules were incorporated to a docking analysis to better understand the underlying mechanism of the selective activity. Similarly, the second article (65) was also a combination of LB (QSAR) and SB (docking) methods to firstly unravel new selective LXRβ ligands and then analyze the interaction mode using docking studies. No toxicological study related to LXR was collected during our bibliographic search.

3.2.5 Constitutive androstane receptor

CAR is involved in regulation of the transcription of genes encoding for the metabolism of xenobiotic and steroid (133). It is highly expressed in the liver and to less extend the small intestine. Although structurally similar to other NR, the LBD of CAR contains particular residues and motifs and thus present an original conformation. The CAR LBD in its unbound form present a similar conformation to those of other NR when bound to their endogenous ligands. Thus, in the absence of ligand CAR can recruit coactivators. CAR agonist compounds, referred to as "phenobarbital alike", can increase coactivator recruitment and thus promote the expression of cytochrome P450 enzyme and other proteins involved in metabolism of xenobiotic compounds. Besides the agonist compounds, inverse agonists, such as the androstane metabolites, are also able to bind to CAR LBD. Upon binding to CAR, these compounds inhibit the CAR constitutive activity through the release of coactivators (134, 135). CAR is considered to be a sensor to several xenobiotics including EDCs such as phthalates and triclocarban but it is also considered as an interesting therapeutic target for metabolic diseases such as type 2 diabetes (79, 80). Only few attempts to develop *in silico* models for CAR were found (79, 80, 100). One therapeutic study was conducted with the dual objective of predicting CAR agonist compounds and collecting knowledge about CAR/ligand interactions. To do so, Lee et al. (79) developed a machine-learning model based on pharmacophoric descriptors with good predictive power with accuracy values equal to 0.875 and 0.854 for the training and test sets, respectively and MCCs values equal to 0.744 and 0.701 for the training and test sets, respectively. Additionally, the group identified the critical elements involved in the binding affinity with CAR. Additionally, two toxicological studies that aim to understand the effect of EDCs on the CAR were also retrieved.

3.2.6 Farnesoid X receptor

FXR is expressed in the liver, the kidney, the intestine, and the adrenal glands. It regulates the metabolism of glucose and lipids and the maintenance of the bile acids homeostasis. It is thus a suitable therapeutic target for the prevention and treatment of metabolic syndrome, dyslipidemia, atherosclerosis, and type 2 diabetes (59). Additionally, some exogenous compounds able to bind to FXR can induce the dysfunction of the receptor and are suspected to be responsible for liver toxicity and hepato-biliary injuries (61). In this review, 4 *in silico* models dedicated to the identification of FXR modulators for the treatment of hepatic and metabolic disorders have been analyzed. As a limited number of diverse known FXR modulators was available, SB methods used to be elected to identify potential therapeutic compounds. 3 studies present the generation of FXR SB pharmacophore models and their use in prospective screening in combination with experimental testing to identify FXR agonist hits (57, 59, 60). For example, in the study of Schuster et al. (58), a set of SB pharmacophore models was generated and theoretically evaluated by calculating the enrichment factors using several data sets. The combination of all pharmacophore models was able to retrieve 87.8% of a list of FXR actives but the performance obtained varied according to the model and to the data set studied. EF values were used to produce pharmacophore models ranking for each data set and the 3 most suitable pharmacophore models were selected accordingly for prospective screening. In complement, 3 LB approaches were described in the literature. The Tox21 FXR agonism and antagonism assay data were used in 2 different studies to obtain respectively FXR disruptors model using machine learning methods (61) and separated FXR agonism and FXR antagonism models using a deep learning approach (100). Both studies present high predictive accuracy with area under the ROC curve values reaching 0.8. The third study focus on different machine learning methods (counter-propagation artificial neural network, similarity of 3D pharmacophore feature distributions method and k-nearest neighbor learner) that were optimized and combined to identify new FXR modulators molecular frameworks (62). This ensemble machine learning approach was used in a prospective screening of 3 million commercially available compounds and enable the discovery of 4 new experimentally validated FXR agonist and 2 FXR antagonist compounds with original molecular frameworks.

3.2.7 Thyroid hormones receptors

TR, are regulated endogenously by the thyroid hormones (TH) that play major role in metabolism and growth processes (136). The 2 isoforms, TRα and TRβ, are expressed differently in tissues and play different roles in the TH signaling (137). Despite the great potency of TR as therapeutic target in the field of dyslipidemia and liver diseases, the development of TR modulators has been limited by selectivity problems and associated undesirable side effects on heart and bone. TRα activation being associated with the cardiac effects of TH, TRβ selective compounds are currently investigated for the treatment of metabolic and brain disorders (138). Several environmental chemicals have also been documented as TR modulators, some of them being suspected to be EDCs. It is notably the case of Hydroxylated polybrominated diphenyl ethers that present structural similarity with endogenous TH and may interfere with the TH binding to TR (87). TR are thus largely studied in the EDCs context (139, 140). In this review, 10 models dedicated to TR are listed, including 3 models specifically developed for TRβ and 4 tools to predict potential EDCs that include SB (12, 104, 105) and LB (19, 100) models for a panel of NRs and not only TR. It is to note that these LB initiatives trained their model

using the TR Tox21 dataset and that the associated performance was among the lowest of all studied receptors. Among the TR focused studies, Zhang et al. (89) presented a virtual screening protocol combining ensemble docking and MM-GBSA rescoring to identify TRβ ligands. This protocol was developed and evaluated using the TRβ DUD-E data set and was associated with an AUC value of 0.865 and an EF10% value of 7.418. This protocol was then applied in a prospective virtual screening of an indoor dust contaminants inventory but as it was developed using the TRβ DUD-E dataset, it could also be applied to identify potential TRβ binders for therapeutic applications. Among the remaining studies, it is to note that one was dedicated to the classification of TR agonist and antagonist compounds using ML methods but was not applicable to the classification of TR binders and TR non binders (88), and 2 studies using QSAR models were associated with good predictive performance of TRβ agonist compounds (90) and TRβ binders (91) respectively, but both suffer from a limited (<30) number of compounds included in the training and test sets.

3.2.8 Vitamin D receptor

VDR is expressed in various tissues especially in the gastrointestinal tract and the kidneys. It plays a major role in the regulation of vitamin D thus controlling the calcium homeostasis, the bone mineralization and remodeling and immune pathways. Additionally, VDR is responsible for the detoxification of both endogenous ligands and xenobiotics. The natural ligand of VDR is the active form of vitamin D called D3. It has been reported in several studies that the lack of Vitamin D leads to hypocalcemia and hypophosphatemia resulting in chronic kidney disease (CKD) also found in patients under chronic hemodialysis, mineral-bone disorders, osteoporosis and cancer (92, 94, 141, 142). In this review, we collected 3 articles that aimed at identifying novel VDR modulators for therapeutic purpose. 2 articles both used a combination of LB (pharmacophore-based 3D-QSAR models) and SB (docking and molecular dynamics) methods to identify respectively VDR inhibitors (93) and VDR agonists (94). Both models are associated with good correlation value (R^2=0.8869 and R^2=0.8676 respectively) and predictive score on the training set (Q^2=0.8870 and Q^2=0.8523 respectively). However, the low diversity and limited number of compounds used to train the model in the first study (93) and the difficulty to understand and thus reproduce the protocol used in the second one (94) may limit the applicability of these 2 models. In the third study (92), *de novo* design, docking and molecular dynamics was used to design new potential VDR agonists. However, except a redocking evaluation of D3, the performance of the protocol was not assessed retrospectively and no experimental test was conducted to validate the predicted hits. Additionally, one article (100) used data from the Tox21 initiative to build deep learning toxicological models for 35 NRs including VDR.

3.3 Monomeric orphan receptors

The members of this category of NR are characterized by an incomplete knowledge about their endogenous ligands and their ability to be activated as a monomer (or homodimer) by opposition to the orphan receptor that are partners to RXR (108). 3 NR studied during this review process and acting as monomers belong to this category (Table 3).

3.3.1 Estrogen related receptors

ERR are orphan receptors, i.e. no endogenous ligands have yet been characterized for the ERR. They do not bind any natural hormones, not even estrogen despite their names (which has been chosen regarding the high sequence similarity in the DNA-binding domain between ERR and ER) but they do bind some synthetic estrogenic compounds (143). ERR are present in the human body under three main isoforms, ERRα, ERRβ and ERRγ that modulate cartilage development, mitochondrion organization and T-cell activation and differentiation. ERRα has particularly been investigated due to its tight similarity with ERα. Common DNA regions can be activated by both ERRα and ERα (144) and they shared common co-activators which may explain the fact that ERRα is involved in estrogen-related diseases. However, the lack of known endogenous ligand has impaired the establishment of protein-ligand interaction profile and the discovery of synthetic ligands. However, the Tox21 program has provide a huge amount of binding and activity data for ERR that could be used to build prediction models. Klimenko et al. (95), proposed a ligand-based (LB) approach to identify ERRα agonists by combining QSAR models. Each QSAR model is specific of a particular biological endpoint aiming at identifying potential ERRα agonists. The models were selected according to the balance between several statistical parameters i.e sensitivity, specificity, Accuracy, balanced accuracy, PPV and NPV. Other investigated ERRα ligands are inverse agonist and antagonist compounds that can be used for cancer patients with resistance to hormonal therapy. In order to identify such compounds, Chitrala et al. (96) used a library from KEGG COMPOUNDS, containing an ensemble of metabolic compounds, pharmaceutical and environmental compounds. This library was pre-filtered to select 8 compounds presenting a similarity score greater than 0.3 with an antagonist compound co-cristallized into the ERRα binding site. The ERRα structure in its antagonist conformation was then used to dock these compounds. Unfortunately, no biological evaluation of the proposed hit compound was achieved.

3.3.2 Liver receptor homolog 1

LRH-1, also called NR5A2, plays an essential role in the well-functioning of the liver, the pancreas and the intestines by controlling the level of cholesterol, bile acids and pancreatic enzymes. LRH-1 is also involved in cell differentiation and associated with key developmental pathways. However, dysregulated LRH-1 activity and unexpected re-activation of the previously mentioned developmental pathways has been linked with breast, endometrial, intestinal and pancreatic malignancies (97). Limited data is available for LRH-1. In particular, LRH-1 is still classified as an orphan receptor, but several studies pointed out phospholipid species as possible endogenous ligands. This has impacted the development of LRH-1 modulators and of *in silico* models dedicated to LRH-1. Only one article dedicated to LRH-1 was found during the review process (97). This article published in 2013 presents a prospective structure-based screening with the aim of identifying the first synthetic LRH-1 antagonist compound. Using a docking method, 2 new LRH-1 antagonists were discovered and proposed to be used as a probe to help decipher LRH-1 mechanism of action.

3.3.3 Retinoic acid-related orphan receptor

Three subtypes of ROR exist, namely RORα, RORβ and RORγ with its two isoforms, RORγ1 and RORγt (RORγ2). Each subtype presents a different pattern of expression, RORα being highly expressed in the brain, RORβ in the central nervous system, RORγ1 in the liver, the adipose tissue, the kidney, the small intestines, and the skeletal muscles and RORγt being exclusively expressed in cells of the immune system (145, 146). RORs are implicated in several key physiological functions. In particular, RORγt plays a major role in the differentiation of T-helper 17 (TH17) cells that produce the cytokine IL-17, itself involved in several diseases such as psoriasis, multiple sclerosis, rheumatoid arthritis and type 1 diabetes (98, 99). Additionally, the inhibition of RORγt stimulates the AR gene transcription and can be a strategy to follow for prostate cancer treatment (99). Identifying RORγt modulators emerges as a promising therapeutic strategy for all these conditions. However, until very recently, the known RORγt modulators were all sharing the same scaffold and efforts have been made towards the discovery of new RORγt modulators with original scaffolds. Two articles (98, 99) described SB *in silico* approaches to identify potent inverse agonists of RORγt that can be used in auto-immune diseases treatment. In the first article, docking and negative image-based methods (NIB) were both first evaluated retrospectively with RORγt experimental data extracted from the ChEMBL database to define docking score threshold and cutoff similarity value, respectively. These 2 methods were then used in parallel to screen a collection of more than 100000 molecules commercially available. Experimental tests validated as RORγt inverse agonists 11 up to the 34 predicted consensus hit compounds with an original scaffold.

3.4 Projects targeting several NR

Besides articles that were solely dedicated to a specific NR presented above, other projects focused on ensemble of NRs to provide more global NRs ligand identification models. Except one article (106), all these projects fall under a toxicological scope with the aim of identifying potential EDCs and understanding their mechanism of action. The most targeted NR combination is ER and AR due to extensive literature and data available. For example, Li et al. (31) studied the ability of some EDCs to present a dual activity by interacting with both AR and ER. To do so, they constructed ER binding models using a QSAR approach and these models were used to screen a database of AR antagonist compounds. It is to note that some of these projects are made available as webservers and software (12, 103–105). Open VirtualToxLab (105) is a software that estimates the toxicological potential of the screened compounds by relying on an automated combination of flexible docking and free energy calculations. Open VirtualToxLab focuses on 16 biological targets among which are 9 NR: AR, ERα, ERβ, GR, LXR, MR, PPARγ, PR, TRα, TRβ. The proof of concept of this software was established with a series of 2564 compounds yielding accurate predictions [C.f. Table 2 of (105)]. Endocrine disruptome (104) is a freely available open-source project that allows users to test the ED potential of their compounds against 14 NR involved in several biological processes: AR, ER α, ER β, GR, PPARα, PPARβ, PPARγ, PR, RXRα, and TRα and TRβ. This webserver is based on docking models elaborated for each NR individually. The major drawback of this tool is the use of DUD-E decoys to validate the protocol. These decoys are putative inactive compounds, and not experimentally validated inactive compounds, which may bias the evaluation of the performance of the models. ProTox-II (103) is a webserver predicting the potential toxicity of small molecules by using 33 models of different toxicity endpoints. Among these endpoints are AhR, AR, ER and PPARγ signaling pathways. The models for each one of these NRs performed with high accuracy on the Tox21 dataset [C.f Table S2 of (103)].

3.5 Related receptors: Aryl hydrocarbon receptors

Although AhR does not belong to the NR superfamily, this receptor displays functional and structural similarities with the members of this family, particularly the presence of a DBD and a LBD activated upon ligand binding (147). During the review

process, we noticed that AhR is often associated with other NR in toxicological studies and in the EDCs context and we decided to list also the *in silico* studies dedicated to this receptor (Table S1). AhR is expressed in various tissues such as the lung, liver, kidney, skin, sleen and placenta. It can be activated by several hazardous chemicals such as polycyclic aromatic hydrocarbons (PAHs) and persistent organic pollutants (POPs) especially dioxins and polychlorinated biphenyls (PCBs). These chemicals result from combustion and are massively found in the air (148). Due to its pivotal role as an environmental pollution sensor and mediator, AhR emerges as a major target for toxic compounds which is illustrated by the predominance of reported toxicological projects (8 toxicological among 10 in total). Most of the data used in these projects are collected from the literature and these studies are in their large majority dedicated to specific chemical series of environmental compounds that have been proved experimentally to be able to interact with AhR (71, 149, 150). The corresponding models are associated with high predictive power, but their applicability domain is relatively limited. Additionally, data from the Tox21 project was also used with both SB and LB methods. However, the application of SB methods is limited by the fact that the 3D structure of this receptor has not been solved yet. A preliminary homology modelling step is thus necessary in all the studies relying on SB methods to obtain a model of the AhR LBD. AhR belongs to the PAS-domain protein group (151) and usually the structure of the PAS-B of the Hypoxia Inducible Factor 2 α (HIF-2α) is used (152, 153) sharing 31% of sequence similarity and 52% identity with the target considered as the highest sequence identity and similarity. More recently, potential therapeutic applications of AhR modulation have emerged. AhR activation has been shown to be involved in hematopoiesis and inflammation process especially the production of inflammatory cytokines whereas AhR repression has been shown to be beneficial in anti-cancer therapies especially for glioblastomas and breast cancers (154). Two therapeutic projects were analyzed in this review, the first being a LB model aiming to discover new AhR antagonists. For this purpose, Parks et al. used successively the Rapid Overlay of Chemical Structures (ROCS) and the electro-static overlap to identify one compound that yield promising experimental results in three *in vitro* and *in vivo* AhR-dependent assays (154). The second therapeutic project applied a SB protocol to decipher the molecular mechanisms behind the AhR activation or inhibition (152). In this article, several entries for the template protein were used to simulate the flexibility and plasticity of the binding domain and homology models were selected and used to perform the docking of 10 representative AhR agonists from different chemical classes. The outcome of docking coupled with MD simulations allowed the analysis of the predominant poses and thus the description of ligand binding and the identification of the interacting residues within AhR.

4 Studied methods and associated data

4.1 Computational methods

Along this review process, we analyzed models dedicated to the identification of NRs ligands obtained with a very large spectrum of *in silico* methods categorized in two approaches i.e SB and LB as shown in Figure 3.

4.1.1 Molecular dynamics simulations

MD simulations are used to study the dynamic of a biological system as a function of time. The movements of the system are computed through the integration of Newton's equation of motion. MD simulations can thus be used to consider the conformational changes associated with the NRs LBD. Indeed, the NRs LBD is characterized by its hydrophobicity and flexibility allowing binding of ligands of different sizes and shapes. NRs structures are subject to various modifications occurring after several events such as the DNA transcription or the binding of the ligand (15). The helix H12 is known to modify its orientation according to the bound-ligand profile. MD can thus be used to span the NRs flexibility and deliver unbiased theoretical structures, but also to assert the stability of a ligand-protein complex and study the time of residence of the ligand. In the reviewed articles, MD has been used for several projects relevant to AR, ER, LXR, PPAR, TR and VDR always in complement of another computational method such as QSAR (74, 150) or pharmacophore models (93, 94) and frequently prior or after docking simulations (66, 75, 152).

4.1.2 Docking

Protein-ligand docking is a SB method used to predict the ability of a compound to interact with a target. Docking methods used a search algorithm that generate multiple potential conformations of the molecular complex and a scoring function to rank them. Docking can be used to propose structural hypotheses on the dominant binding mode of a compound, to screen large libraries of compounds and to rank them according to docking scores. In this way, docking is used to prioritize and reduce the number of compounds that should be experimentally tested (156). Docking methods can be applied to any target for which an experimental 3D structure is available. Consequently, docking methods were largely employed to develop predictive models of NR-ligand interactions and for each of the 14 reviewed NR subfamily, at least one project using a docking method was available. However, the docking protocol used must be finely tuned for each NR. A first important criterion is the rational selection of the initial docking structure(s) since the NRs LBDs undergo significant conformational changes upon ligand binding, dependent of the

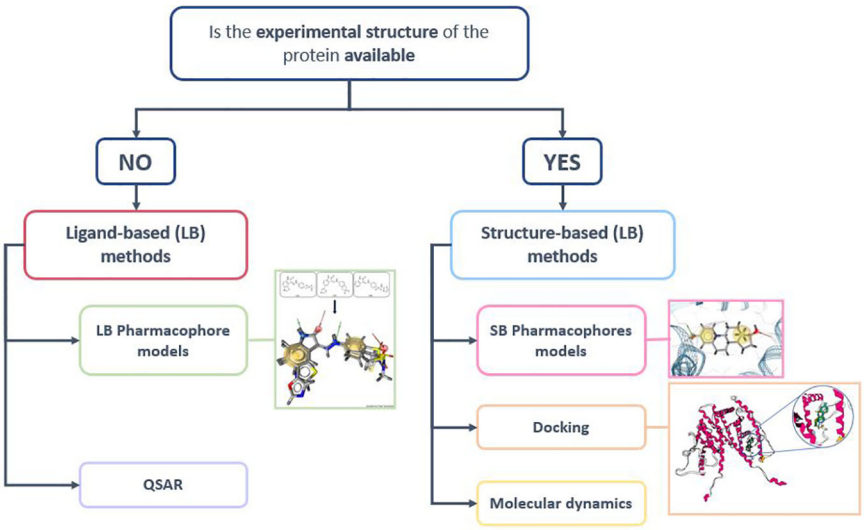


FIGURE 3
Structure-based (SB) and Ligand-based (LB) screening methods (155).

pharmacological profile of the ligand. NRs agonist compounds tend to stabilize the interactions between the residues of helix H12 and other helices (H3, H5 and H11) of the LBD, creating a lid shape of the H12 over the ligand. In contrast, other compounds (antagonists or partial agonists) are not able to stabilize the previous interactions and the observed repositioning of the H12 is different. The choice of an agonist-bound or antagonist-bound structure may thus impact the docking performances in predicting NRs ligands binding (15). This is particularly important for docking protocols aiming to discover therapeutic compounds targeting the NRs as a specific pharmacological profile is usually targeted and the corresponding agonist- or antagonist-bound structure should be selected. In contrast, EDCs can present different pharmacological profiles and it is thus important to evaluate both type of structures to ensure that the screening is able to retrieve all EDCs regardless of their pharmacological profile. In the reviewed articles, two main docking approaches have been listed: single and an ensemble structure docking. In the ensemble structure docking approach, a ligand is successively docked against several protein conformations (Multiple PDB entries or snapshots extracted from MD simulations or a gaussian transformation of the atom localization) and the results are post processed to only keep the best score among all the structures. This approach provides a better picture of the protein flexibility but is more computationally and time expensive and it is not always associated to enhanced docking performances (157).

4.1.3 Pharmacophore modeling

A pharmacophore model is defined by the IUPAC as "an ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target and to trigger (or block) its biological response" (158). Two types of pharmacophore models exist (1): LB pharmacophore models established upon the superimposition of known active molecules and the retrieval of the common chemical features that are necessary for the bioactivity (2); SB pharmacophore models based on the structure of a protein or more usually of a complex protein-ligand by probing the possible interactions in the macromolecule binding site (159).

In total, 17 reviewed articles used pharmacophore modelling. The generated pharmacophore models are in a large majority combined with other computational models (QSAR and docking models) and only two publications present a protocol combining SB and LB pharmacophores (76, 79). It is to note that both LB and SB pharmacophores taken individually have a predictive power limited by the data used to train the model, defining their applicability domain. The protocol combining pharmacophore models with other *in silico* methods are described in the reviewed articles and may help overcoming this limit.

4.1.4 QSAR models

QSAR (Quantitative Structure Activity Relationship) is a LB method relying on statistical models to predict the biological activity of compounds. Since their introduction, QSAR models

have known several evolutions from simplistic linear and classification models to the use of more elaborated algorithms like artificial neural networks (ANN) and deep learning (DL). In total, 48 of the reviewed articles used QSAR models with the majority (19) adopting classical 2D machine learning methods like Random Forest or SVM. The remaining studies present models obtained using ANN (4), DNN (3) or 3D QSAR methods such as CoMFA and CoMSIA (7). The classical QSAR models rely on a set of descriptors that encode chemical structures, describe physiochemical properties and that can be obtained from quantum chemical calculations (14, 88). Because a huge number of predefined descriptors can be generated, it is important to correctly select the ones that will accurately translate the link between the chemical structure and the associated activity (160). A classical approach is illustrated in the publication of Wang et al. (88) in which QSAR models dedicated to the identification of agonist and antagonist ligands of TR are presented. In this study, the software Dragon was used to initially generate over 1600 descriptors. This number was then reduced by discarding the descriptors highly correlated and selecting only those directly related to the structure, intuitive and easy to understand. Conversely, DL methods emancipate from descriptors computation and selection. Instead, DL methods ensures automatic feature extractions during the training phase. DL methods are thus particularly adapted for areas where existing predefined descriptors have not been crafted yet like macrocycles or the modeling of therapeutic peptides (161). However, DL methods also present some drawbacks, the first being the so-called "black box effect". DL methods compared to more simplistic algorithms are not easily understandable and deciphering the molecular mechanisms associated with the prediction may be a tough task. Moreover, DL requires higher computing resources which can be a limiting factor.

Similarly to the pharmacophore models, QSAR models' predictions are only accurate for compounds that are similar to the ones used to train them (21, 45, 162). An applicability domain (AD) can thus be defined with descriptors used to build the QSAR model (21, 25).. Implementing the AD has proven to enhance the confidence in the models by reducing the number of false negatives (25). However, prediction will only be achieved for compounds that are within the limit of the AD, i.e. within a defined descriptors similarity threshold with their nearest neighbors of the training set (45). In order to enhance the chemical space for which prediction can be made, QSAR models can be combined with other ones with complementary ADs or with a SB model such as docking (67, 68, 74, 81).

4.1.5 Data splitting

An important step prior the generation of LB models is to split the available ligands data into train and test set (160). The stake of a good data split is to obtain a test set that is somehow representative of the training set chemical space. In most of the reviewed article, data was split randomly as it is commonly done to avoid biased evaluation of the model. However, this randomness may lead to an uneven distribution of active and inactive compounds between the train and the test set especially if the active/inactive ratio is initially low. In this sense, Capuzzi et al. (101) build QSAR models relying on two sets of databases: (1) the native Tox21 set of data with a ratio of active: inactive equal to 1:10 and (2) a balanced set generated down-sampled from the inactive data of the Tox21 native dataset i.e. for each active compounds an equivalent inactive was selected either randomly or based on the highest Tanimoto score within the inactive pool of compounds. It is to note that using the balanced set leads to a decrease in the accuracy of the model in comparison to when they used the native (unbalanced) dataset. Randomly splitting data may also result in a test set that is not representative of the train especially when the initial data are structurally diverse. Wang, Xing et al. (88) used self-organizing map (an ANN algorithm) to project an ensemble of structurally diverse compounds on a low dimensionality grid. This helped visualizing the space and the selection of a test set representative of the overall dataset and lead to average accuracies of 83.1–97.2%. In other studies, the trade-off between randomness and unbiased subdivision was achieved with bootstrapping. Several splits were done, and models were built an evaluated with the associated train/test output. The split with the best results was kept for the following model optimization (29, 163).

4.1.6 Combinations of methods

It is to note that each computational method has limitations. For example, QSAR methods are not suited on their own to be used for HTS. They are more efficient at generating focused data and requires exhaustive model training and validation steps. Moreover, QSAR and pharmacophores modelling share the common particularity of being only efficient on molecules falling within a domain of applicability (162). A crucial point in the docking workflow is the choice of the scoring function as no universal scoring function exists yet (164). Each individual method can be at the origin of new structural information that can be combined to enable a better understanding of the mechanism of action. A solution to overcome each method limitations and take advantage of their complementarity is to use a combination of methods (28, 165, 166). Computational methods can be combined in integrative approaches using hierarchical or consensus screening (27–29, 48, 60, 100, 163) (for further references c.f. Table 1–4).

4.2 Databases

A chemical database (DB) is an organized collection of compounds with relevant information on their chemical structures together with activity data collected from *in vivo* and/or *in vitro* experiments and sometimes *in silico* predicted

activities. Several DB dedicated to NR exist and we decided in this review, to focus on the ones available in open access as presented in Table 5.

In silico studies focusing on NR are numerous and the choice of the databases to use to construct predictive models depends on the aim of the project. Moreover, and especially for drug design projects, some research teams chose to use in house databases i.e. a collection of compounds issued as a result of experimental work in the laboratory. These databases can be composed of one or several chemical series originating from a single hit or a known drug scaffold or, in some cases focused libraries. Finally, some projects rely on several existing databases, either to combine the molecules resulting in larger databases for the same purpose (therapeutic or toxicological) or use a database for the training step and another for the external validation. For example, Réau et al. (29) developed a docking and a pharmacophore modeling strategy for identifying AR agonist compounds relying on the NR-DBIND data and used as an external validation set the compounds from the tox21 challenge.

4.2.1 Assay consideration

In the scientific literature, biological data available for NRs ligands represent various assays endpoints sometimes measured from different laboratories and companies. Indeed, two main assays were used in the reviewed studies to assess NRs ligands potency: binding assay and gene reporter assay (34). These assays attributed to each compound a quantitative value that may vary according to the laboratory in charge of the experiment. In the *in silico* modelling field, the experimental data are used to compute predictive models for molecules with no related experimental information. For some methods such as QSAR, the numerical value associated with the experimental test can directly be integrated to construct the models. For other method, such as docking and pharmacophoric modeling, these data usually need to be converted into a binary variable with 2 possible values: "active compounds" and "inactive compounds". This is necessary to optimize and select the prediction models able to distinguish between both categories. Indeed, to evaluate the predictive power of a model, the number of correctly predicted active and inactive (TP and TN) as well as the number of wrongly predicted compounds (FP and FN) are

TABLE 5 Example of databases including or dedicated to nuclear receptors.

Database	Link	Composition	Specific to NR only
Binding DB	https://www.bindingdb.org/bind/index.jsp	As of November 8, 2021, BindingDB contains 2,369,418 binding data for 8,634 protein targets and 1,023,385 small molecules	No
ChEMBL	https://www.ebi.ac.uk/chembl/	manually curated database: 2.1 M compounds	No
Drugbank	https://go.drugbank.com/	14,585 drugs and several targets like enzyme, transporters and carriers	No
ZINC database	https://zinc.docking.org/	contains over 230 million purchasable compounds in ready-to-dock, 3D formats.	No
Tox21	https://tripod.nih.gov/tox21	The list of ToxCast and Tox21 chemicals suspected to be a hazard for human	No
ToxCast	https://www.epa.gov/chemical-research/toxcast-chemicals	and environmental health and associated information for 9,403 unique substances.	No
DUD-E	http://dude.docking.org/	22,886 active compounds and their affinities against 102 targets, an average of 224 ligands per target and 50 decoys for each active having similar physico-chemical properties but dissimilar 2-D topology.	No
DSSTox	https://comptox.epa.gov/dashboard/chemical-lists/tox21sl	launched in 2004, currently exceeds 875K substances spanning hundreds of lists of interest.	No
EDKB	https://www.fda.gov/science-research/endocrine-disruptor-knowledge-base/accessing-edkb-database	Data for more than 3200 chemicals	Yes (ER and AR)
EABD	https://www.fda.gov/science-research/bioinformatics-tools/estrogenic-activity-database-eadb	18,114 estrogenic-activity data points collected for 8,212 chemicals tested in 1,284 binding assays, reporter-gene assays, cell-proliferation assays, and in-vivo assays in 11 different species.	Yes (ER)
NR-DBIND	http://nr-dbind.drugdesign.fr/	15,116 positive and negative interactions data are provided for 28 NRs together with 593 PDB structures	Yes
NR-List BDB	http://nrlist.drugdesign.fr/	9,905 compounds and 339 structures of the NRLiSt BDB	Yes
ONRLDB	https://academic.oup.com/database/article/doi/10.1093/database/bav112/2433243	~11 000 ligands, of which ~6500 are unique.	Yes
NURA	https://www-sciencedirect-com.proxybib-pp.cnam.fr/science/article/pii/S0041008X20303707?via%3Dihub	bioactivity data for 15,247 molecules and 11 NRs	Yes (ERα andβ, PPARGαγ, and δ, AR, GR,PR, FXR, RXR and PXR)

used to compute metrics such as sensitivity, sensitivity and enrichment for example. The binarization of the data is a crucial but not straightforward step because a threshold must be defined to separate the data. Additionally, because various biological endpoints are considered, the concordance between the thresholds of activities defined for each assay should be assessed. Lunghini et al. (167) defined the "degree of agreement" parameter to pinpoint the differences between various assays used to label compounds that interact with ER and AR in the context of EDCs models. To do so, they performed multiple pairwise comparisons of various binding assays result among 4 data sources for ER binding compounds and 3 data sources for AR binding compounds. The degree of agreement was calculated as the average number of sources agreeing on a given label for each compound. Their analysis showed that there is a lack of concordance between experiments for 42% of the compounds. This study highlighted the danger of merging assays with different biological meanings (167) and considering the positive outputs as interacting compounds regardless of the mechanism of action. However, it is a commonly used protocol, often driven by the lack of appropriate data or the blurry definition of the mechanism of action. This is especially the case in the toxicological context of the study related to EDCs and in this review, we came across several studies performing the merging of data obtained from different assays and different sources. Some studies, aware of this issue, also developed various approaches to limit the associated bias. Sakkiah et al. (56) calculated the concordance between the train and the test set since originated respectively from binding and amplification assays. Manganelli et. al (28), compared the train and the test data to define concordance and exclude a part of the data that falls under a "dangerous" segment. Finally, Zhang et al. (128) normalized the activity of their test set to their training data when constructing their models.

4.2.2 Data balance

Along with active compounds, it is also important to construct, optimize and evaluate NRs prediction models to include negative data (168). Positive data are usually resulting from biological or cellular *in vivo assays* but, negative data has long been constituted by presumed inactive compounds (121). Since negative data have a crucial impact in influencing the performance of a model (168), recent efforts have been made to include carefully selected and experimentally validated inactive compounds [NR-DBIND (169)]. An additional issue is the proportionality between active and inactive compounds in DB. When collecting data from scientific literature, very few inactive compounds are retrieved. Conversely, in databases presenting high throughput screening results, inactive compounds usually outnumbered the active ones. In this case, the active substances represent a low proportion out of all tested chemicals (95). A performant model is dependent on clean, diverse and accurate data (95, 165), and unbalanced data (either in favor of the active or the inactive counterpart) can impact prediction models building (55). Clustering methods can be used to select smaller subsets of representative active or inactive compounds. For example, Another solution is the use of a structural similarity filter based on fingerprints and the application of a similarity metric cutoff (for example Tanimoto) in order to select sparse compounds to scour the entire chemical space (170). In the case of unbalanced data sets, the classical metrics are not the best suited to evaluate the predictive power of a model. In this case some metrics like the MCC and the F1 score can help in assessing the ability of a model to correctly predict each category of compound.

4.3 Level of reproducibility

Reproductible science is a quality standard that the scientific community values as it helps producing high quality, reliable and efficient research project (171). It has been proven that reproducibility together with methods, reporting, dissemination, evaluation and incentives are key elements for the scientific process (172). However, it has also been asserted that the field of preclinical research suffers from the inability to replicate findings published in high-profile journals (171). Although it is difficult to exactly replicate results in biology systems due to their inherent variability, some recommendations on good practice could be listed to alleviate the trustworthiness of the scientific work and the validity of the major conclusions (171). Reinforcing the policies on data and code sharing is one of these recommendations (171, 173).

Through this review, we decided to evaluate this point in the NRs *in silico* research field. For each reviewed model, the level of reproducibility was graded according to the availability of databases and the model procedure's parameters and/or code. The "high" grade was used to describe models providing both components within the article, the "medium" was associated with articles with partial available data. Finally, the reproducibility was described as "low" for models where no data were available. For most of the reviewed articles' reproducibility was rated "medium" (52 out of 89 articles) and only a low proportion was described to have a "high" reproducibility (16 out of 89). These results focused on the NRs are in concordance with other large-scale studies (173). For example, in a study of 2011, the analysis of 500 papers published in the top 50 journals across scientific fields showed that only 9% of these papers enclosed full primary data (171). The reinforcement of collaborations between research teams following the example of COMPARA (20) and CERAPP (36) consortiums or the Tox21 challenge (174) can improve the reproducibility since it creates accountability in exchange of a

benchmarked data (172). However, according to our classification, no particular tendency was observed over the 10 year covered by this review neither towards increase of the number of articles labeled with "high" reproducibility nor towards decrease of the number of articles associated with "low" reproducibility.

5 Conclusion

Nuclear receptors are a large family of transcription factors involved in several biological process and their impairment result in several pathologies. Moreover, this protein family can be targeted by toxic compounds leading to the disruption of the normal functioning of NRs and especially the disruption of the endocrine system by a family of compounds called endocrine disrupting chemicals or EDCs. Being in the center of both therapeutic and toxicological concerns, NRs are widely studied to find new cures but also to unravel the potential toxicity of environmental compounds such as pesticides, cosmetics or additives. In the last decades and with the emergence of bioinformatics and virtual screening techniques, computational models dedicated to NR were developed. The computational capabilities were put to use to interpretate and analyze the experimental data and build predictive models to unravel new drug hit and forecast potential toxicants. This article is a review of the studies dedicated to NR ligands prediction for both therapeutic and toxicological purposes, published in the last decade. 89 articles concerning 14 NR subfamilies were carefully read and analyzed in order to retrieve the most commonly used computational methods to develop predictive models, to retrieve the databases deployed in the model building process. Some issues facing the model building process were addressed like the assays endpoint discrepancies and data balance were also addressed and how authors managed to overcome them. This review emerged from the need to identify most of the *in silico* initiatives undertaken on NR and can be used as a starting point for future investigations on the subject not only to appreciate the importance of NR in both therapeutic and toxicological fields, but also to learn from previous experiences, encourage the elaboration of more accurate predictions and motivate collaborations.

Author contributions

Conceptualization: AS, MR, and NL. Methodology: AS, MR, and NL. Validation: NL, MR, and MM. Formal analysis: AS and NL. Investigation: AS. Resources: AS, MR, NL. Data collection and curation: AS. Writing—original draft preparation: AS. Writing—review and editing: AS, MR, NL, and MM. Supervision: NL. Project administration: MM. Funding acquisition: AS, NL, and MM. All authors contributed to the article and approved the submitted version.

Funding

AS is recipient of a MESRI (Ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation) fellowship.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fendo.2022.986016/full#supplementary-material

References

1. Weikum ER, Liu X, Ortlund EA. The nuclear receptor superfamily: A structural perspective. *Protein Sci* (2018) 27(11):1876–92. doi: 10.1002/pro.3496

2. Zhang Y, Luo X, Wu Dh, Xu Y. ROR nuclear receptors: structures, related diseases, and drug discovery. *Acta Pharmacol Sin* (2015) 36(1):71–87. doi: 10.1038/aps.2014.120

3. Sladek FM. Nuclear receptors as drug targets: New developments in coregulators, orphan receptors and major therapeutic areas. *Expert Opin Ther Targets* (2003) 7(5):679–84. doi: 10.1517/14728222.7.5.679

4. Kinch MS, Hoyer D, Patridge E, Plummer M. Target selection for FDA-approved medicines. *Drug Discovery Today* (2015) 20(7):784–9. doi: 10.1016/j.drudis.2014.11.001

5. Santos R, Ursu O, Gaulton A, Bento AP, Donadi RS, Bologa CG, et al. A comprehensive map of molecular drug targets. *Nat Rev Drug Discovery* (2017) 16(1):19–34. doi: 10.1038/nrd.2016.230

6. Bartolini D, De Franco F, Torquato P, Marinelli R, Cerra B, Ronchetti R, et al. Garcinoic acid is a natural and selective agonist of pregnane X receptor. *J Med Chem* (2020) 63(7):3701–12. doi: 10.1021/acs.jmedchem.0c00012

7. Schug TT, Johnson AF, Birnbaum LS, Colborn T, Guillette LJJr, Crews DP, et al. Minireview: Endocrine disruptors: Past lessons and future directions. *Mol Endocrinol* (2016) 30(8):833–47. doi: 10.1210/me.2016-1096

8. La Merrill MA, Vandenberg LN, Smith MT, Goodson W, Browne P, Patisaul HB, et al. Consensus on the key characteristics of endocrine-disrupting chemicals

as a basis for hazard identification. *Nat Rev Endocrinol* (2020) 16(1):45–57. doi: 10.1038/s41574-019-0273-8

9. Combarnous Y, Nguyen TMD. Comparative overview of the mechanisms of action of hormones and endocrine disruptor compounds. *Toxics* (2019) 7(1):5. doi: 10.3390/toxics7010005

10. *Search the TEDX list [Internet]. TEDX - the endocrine disruption exchange* (2021). Available at: https://endocrinedisruption.org/interactive-tools/tedx-list-of-potential-endocrine-disruptors/search-the-tedx-list.

11. Audouze K, Sarigiannis D, Alonso-Magdalena P, Brochot C, Casas M, Vrijheid M, et al. Integrative strategy of testing systems for identification of endocrine disruptors inducing metabolic disorders–an introduction to the oberon project. *Int J Mol Sci* (2020) 21(8):2988. doi: 10.3390/ijms21082988

12. Vedani A, Dobler M, Smieško M. VirtualToxLab — a platform for estimating the toxic potential of drugs, chemicals and natural products. *Toxicol Appl Pharmacol* (2012) 261(2):142–53. doi: 10.1016/j.taap.2012.03.018

13. Marroqui L, Tudurí E, Alonso-Magdalena P, Quesada I, Nadal Á, Dos Santos RS. Mitochondria as target of endocrine-disrupting chemicals: Implications for type 2 diabetes. *J Endocrinol* (2018) 239(2):R27–45. doi: 10.1530/JOE-18-0362

14. Chen Y, Cheng F, Sun L, Li W, Liu G, Tang Y. Computational models to predict endocrine-disrupting chemical binding with androgen or oestrogen receptors. *Ecotoxicol Environ Safety* (2014) 110:280–7. doi: 10.1016/j.ecoenv.2014.08.026

15. *Computational approaches to nuclear receptors [Internet]* (2012). Available at: https://pubs.rsc.org/en/content/ebook/978-1-84973-364-9.

16. Gellatly N, Sewell F. Regulatory acceptance of in silico approaches for the safety assessment of cosmetic-related substances. *Comput Toxicol* (2019) 11:82–9. doi: 10.1016/j.comtox.2019.03.003

17. Mazaira GI, Zgajnar NR, Lotufo CM, Daneri-Becerra C, Sivils JC, Soto OB, et al. The nuclear receptor field: A historical overview and future challenges. *Nucl Receptor Res* (2018) 5:101320. doi: 10.11131/2018/101320

18. Zhao L, Zhou S, Gustafsson JÅ. Nuclear receptors: Recent drug discovery for cancer therapies. *Endocrine Rev* (2019) 40(5):1207–49. doi: 10.1210/er.2018-00222

19. Sun L, Yang H, Cai Y, Li W, Liu G, Tang Y. In silico prediction of endocrine disrupting chemicals using single-label and multilabel models. *J Chem Inf Model* (2019) 59(3):973–82. doi: 10.1021/acs.jcim.8b00551

20. Mansouri K, Kleinstreuer N, Abdelaziz AM, Alberga D, Alves VM, Andersson PL, et al. CoMPARA: Collaborative modeling project for androgen receptor activity. *Environ Health Perspect* (2020) 128(2):027002. doi: 10.1289/EHP5580

21. Li J, Gramatica P. Classification and virtual screening of androgen receptor antagonists. *J Chem Inf Model* (2010) 50(5):861–74. doi: 10.1021/ci100078u

22. Wang X, Li X, Shi W, Wei S, Giesy JP, Yu H, et al. Docking and CoMSIA studies on steroids and non-steroidal chemicals as androgen receptor ligands. *Ecotoxicol Environ Safety* (2013) 89:143–9. doi: 10.1016/j.ecoenv.2012.11.020

23. Wu Y, Doering JA, Ma Z, Tang S, Liu H, Zhang X, et al. Identification of androgen receptor antagonists: *In vitro* investigation and classification methodology for flavonoid. *Chemosphere* (2016) 158:158:72–9. doi: 10.1016/j.chemosphere.2016.05.059

24. Yang X, Liu H, Yang Q, Liu J, Chen J, Shi L. Predicting anti-androgenic activity of bisphenols using molecular docking and quantitative structure-activity relationships. *Chemosphere* (2016) 163:373–81. doi: 10.1016/j.chemosphere.2016.08.062

25. Trisciuzzi D, Alberga D, Mansouri K, Judson R, Novellino E, Mangiatordi GF, et al. Predictive structure-based toxicology approaches to assess the androgenic potential of chemicals. *J Chem Inf Model* (2017) 57(11):2874–84. doi: 10.1021/acs.jcim.7b00420

26. Wahl J, Smieško M. Endocrine disruption at the androgen receptor: Employing molecular dynamics and docking for improved virtual screening and toxicity prediction. *Int J Mol Sci* (2018) 19(6):1784. doi: 10.3390/ijms19061784

27. Grisoni F, Consonni V, Ballabio D. Machine learning consensus to predict the binding to the androgen receptor within the compara project. *J Chem Inf Model* (2019) 59(5):1839–48. doi: 10.1021/acs.jcim.8b00794

28. Manganelli S, Roncaglioni A, Mansouri K, Judson RS, Benfenati E, Manganaro A, et al. Development, validation and integration of in silico models to identify androgen active chemicals. *Chemosphere* (2019) 220:204–15. doi: 10.1016/j.chemosphere.2018.12.131

29. Réau M, Lagarde N, Zagury JF, Montes M. Hits discovery on the androgen receptor: In silico approaches to identify agonist compounds. *Cells* (2019) 8 (11):1431. doi: 10.3390/cells8111431

30. Zorn KM, Foil DH, Lane TR, Hillwalker W, Feifarek DJ, Jones F, et al. Comparison of machine learning models for the androgen receptor. *Environ Sci Technol* (2020) 54(21):13690–700. doi: 10.1021/acs.est.0c03984

31. Li J, Gramatica P. QSAR classification of estrogen receptor binders and pre-screening of potential pleiotropic EDCs. *SAR QSAR Environ Res* (2010) 21(7–8):657–69. doi: 10.1080/1062936X.2010.528254

32. Rybacka A, Rudén C, Tetko IV, Andersson PL. Identifying potential endocrine disruptors among industrial chemicals and their metabolites–development and evaluation of in silico tools. *Chemosphere* (2015) 139:372–8. doi: 10.1016/j.chemosphere.2015.07.036

33. Bhhatarai B, Wilson DM, Price PS, Marty S, Parks AK, Carney E. Evaluation of OASIS QSAR models using ToxCastTM in vitro estrogen and androgen receptor binding data and application in an integrated endocrine screening approach. *Environ Health Perspect* (2016) 124(9):1453–61. doi: 10.1289/EHP184

34. Tan H, Wang X, Hong H, Benfenati E, Giesy JP, Gini GC, et al. Structures of endocrine-disrupting chemicals determine binding to and activation of the estrogen receptor α and androgen receptor. *Environ Sci Technol* (2020) 54 (18):11424–33. doi: 10.1021/acs.est.0c02639

35. Allen TEH, Nelms MD, Edwards SW, Goodman JM, Gutsell S, Russell PJ. In silico guidance for in vitro androgen and glucocorticoid receptor ToxCast assays. *Environ Sci Technol* (2020) 54(12):7461–70. doi: 10.1021/acs.est.0c01105

36. Mansouri K, Abdelaziz A, Rybacka A, Roncaglioni A, Tropsha A, Varnek A, et al. CERAPP: Collaborative estrogen receptor activity prediction project. *Environ Health Perspect* (2016) 124(7):1023–33. doi: 10.1289/ehp.1510267

37. Stojić N, Erić S, Kuzmanovski I. Prediction of toxicity and data exploratory analysis of estrogen-active endocrine disruptors using counter-propagation artificial neural networks. *J Mol Graphics Modelling* (2010) 29(3):450–60. doi: 10.1016/j.jmgm.2010.09.001

38. Zhang L, Sedykh A, Tripathi A, Zhu H, Afantitis A, Mouchlis VD, et al. Identification of putative estrogen receptor-mediated endocrine disrupting chemicals using QSAR- and structure-based virtual screening approaches. *Toxicol Appl Pharmacol* (2013) 272(1):67–76. doi: 10.1016/j.taap.2013.04.032

39. Zang Q, Rotroff DM, Judson RS. Binary classification of a large collection of environmental chemicals from estrogen receptor assays by quantitative structure-activity relationship and machine learning methods (2021) 53(12):3244–61. doi: 10.1021/ci400527b

40. Trisciuzzi D, Alberga D, Mansouri K, Judson R, Cellamare S, Catto M, et al. Docking-based classification models for exploratory toxicology studies on high-quality estrogenic experimental data. *Future Med Chem* (2015) 7(14):1921–36. doi: 10.4155/fmc.15.103

41. Zhang Q, Yan L, Wu Y, Ji L, Chen Y, Zhao M, et al. A ternary classification using machine learning methods of distinct estrogen receptor activities within a large collection of environmental chemicals. *Sci total Environ* (2017) 580:1268–75. doi: 10.1016/j.scitotenv.2016.12.088

42. Russo DP, Zorn KM, Clark AM, Zhu H, Ekins S. Comparing multiple machine learning algorithms and metrics for estrogen receptor binding prediction. *Mol Pharm* (2018) 15(10):4361–70. doi: 10.1021/acs.molpharmaceut.8b00546

43. Balabin IA, Judson RS. Exploring non-linear distance metrics in the structure-activity space: QSAR models for human estrogen receptor. *J Cheminform* (2018) 10(1):47. doi: 10.1186/s13321-018-0300-0

44. Zorn KM, Foil DH, Lane TR, Russo DP, Hillwalker W, Feifarek DJ, et al. Machine learning models for estrogen receptor bioactivity and endocrine disruption prediction. *Environ Sci Technol* (2020) 54(19):12202–13. doi: 10.1021/acs.est.0c03982

45. Ciallella HL, Russo DP, Aleksunes LM, Grimm FA, Zhu H. Predictive modeling of estrogen receptor agonism, antagonism, and binding activities using machine- and deep-learning approaches. *Lab Invest* (2021) 101(4):490–502. doi: 10.1038/s41374-020-00477-2

46. Chang YH, Chen JY, Hor CY, Chuang YC, Yang CB, Yang CN. Computational study of estrogen receptor-alpha antagonist with three-dimensional quantitative structure-activity relationship, support vector regression, and linear regression methods. *Int J Medicinal Chem* (2013) 2013: e743139. doi: 10.1155/2013/743139

47. Ng HW, Zhang W, Shu M, Luo H, Ge W, Perkins R, et al. Competitive molecular docking approach for predicting estrogen receptor subtype α agonists and antagonists. *BMC Bioinf* (2014) 15 Suppl 11:S4. doi: 10.1186/1471-2105-15-S11-S4

48. McRobb FM, Kufareva I, Abagyan R. In silico identification and pharmacological evaluation of novel endocrine disrupting chemicals that act *via* the ligand-binding domain of the estrogen receptor α. *Toxicological Sci* (2014) 141 (1):188–97. doi: 10.1093/toxsci/kfu114

49. TilakVijay J, Vivek Babu K, Uma A. Virtual screening of novel compounds as potential ER-alpha inhibitors. *Bioinformation* (2019) 15(5):321–32. doi: 10.6026/97320630015321

50. Ng HW, Doughty SW, Luo H, Ye H, Ge W, Tong W, et al. Development and validation of decision forest model for estrogen receptor binding prediction of chemicals using large data sets. *Chem Res Toxicol* (2015) 28(12):2343–51. doi: 10.1021/acs.chemrestox.5b00358

51. Agatonovic-Kustrin S, Alexander M, Morton DW, Turner JV. Pesticides as estrogen disruptors: QSAR for selective ERα and ERβ binding of pesticides. *Comb Chem High Throughput Screen* (2011) 14(2):85–92. doi: 10.2174/138620711794474097

52. Taha MO, Tarairah M, Zalloum H, Abu-Sheikha G. Pharmacophore and QSAR modeling of estrogen receptor beta ligands and subsequent validation and in silico search for new hits. *J Mol Graph Model* (2010) 28(5):383–400. doi: 10.1016/j.jmgm.2009.09.005

53. Chen L, Wu D, Bian Hp, Kuang Gl, Jiang J, Li Wh, et al. Selective ligands of estrogen receptor β discovered using pharmacophore mapping and structure-based virtual screening. *Acta Pharmacol Sin* (2014) 35(10):1333–41. doi: 10.1038/aps.2014.69

54. Tuccinardi T, Poli G, Dell'Agnello M, Granchi C, Minutolo F, Martinelli A, et al. Receptor-based virtual screening evaluation for the identification of estrogen receptor β ligands. *J Enzym Inhibition and Med Chem* (2015) 30(4):662–70. doi: 10.3109/14756366.2014.959946

55. Niu Aq, Xie Lj, Wang H, Zhu B, Wang S. Prediction of selective estrogen receptor beta agonist using open data and machine learning approach. *Drug Des Devel Ther* (2016) 10:2323–31. doi: 10.2147/DDDT.S110603

56. Sakkiah S, Selvaraj C, Gong P, Zhang C, Tong W, Hong H. Development of estrogen receptor beta binding prediction model using large sets of chemicals. *Oncotarget* (2017) 8(54):92989–3000. doi: 10.18632/oncotarget.21723

57. Zarezade V, Abolghasemi M, Rahim F, Veisi A, Behbahani M. In silico assessment of new progesterone receptor inhibitors using molecular dynamics: A new insight into breast cancer treatment. *J Mol Model* (2018) 24(12):337. doi: 10.1007/s00894-018-3858-6

58. Schuster D, Markt P, Grienke U, Mihaly-Bison J, Binder M, Noha SM, et al. Pharmacophore-based discovery of FXR agonists. part I: Model development and experimental validation. *Bioorganic Medicinal Chem* (2011) 19(23):7168–80. doi: 10.1016/j.bmc.2011.09.056

59. Grienke U, Mihály-Bison J, Schuster D, Afonyushkin T, Binder M, et al. Pharmacophore-based discovery of FXR-agonists. part II: Identification of bioactive triterpenes from ganoderma lucidum. *Bioorganic Medicinal Chem* (2011) 19(22):6779–91. doi: 10.1016/j.bmc.2011.09.039

60. Sindhu T, Srinivasan P. Identification of potential dual agonists of FXR and TGR5 using e-pharmacophore based virtual screening. *Mol Biosyst* (2015) 11(5):1305–18. doi: 10.1039/C5MB00137D

61. Chen Y, Yang H, Wu Z, Liu G, Tang Y, Li W. Prediction of farnesoid X receptor disruptors with machine learning methods. *Chem Res Toxicol* (2018) 31(11):1128–37. doi: 10.1021/acs.chemrestox.8b00162

62. Merk D, Grisoni F, Schaller K, Friedrich L, Schneider G. Discovery of novel molecular frameworks of farnesoid X receptor modulators by ensemble machine learning. *ChemistryOpen* (2019) 8(1):7–14. doi: 10.1002/open.201800156

63. von Grafenstein S, Mihaly-Bison J, Wolber G, Bochkov VN, Liedl KR, Schuster D. Identification of novel liver X receptor activators by structure-based modeling. *J Chem Inf Model* (2012) 52(5):1391–400. doi: 10.1021/ci300096c

64. Heitel P, Achenbach J, Moser D, Proschak E, Merk D. DrugBank screening revealed alitretinoin and bexarotene as liver X receptor modulators. *Bioorganic Medicinal Chem Letters* (2017) 27(5):1193–8. doi: 10.1016/j.bmcl.2017.01.066

65. Salum LB, Andricopulo AD, Honório KM. A fragment-based approach for ligand binding affinity and selectivity for the liver X receptor beta. *J Mol Graphics Modelling* (2012) 32:19–31. doi: 10.1016/j.jmgm.2011.09.007

66. Wang X, Lu K, Luo H, Liang D, Long X, Yuan Y, et al. In silico identification of small molecules as novel LXR agonists for the treatment of cardiovascular disease and cancer. *J Mol Model* (2018) 24(3):57. doi: 10.1007/s00894-018-3578-y

67. Chen M, Yang F, Kang J, Gan H, Yang X, Lai X, et al. Identfication of potent LXRβ-selective agonists without LXRα activation by in silico approaches. *Molecules* (2018) 23(6):1349. doi: 10.3390/molecules23061349

68. Carrieri A, Giudici M, Parente M, De Rosas M, Piemontese L, Fracchiolla G, et al. Molecular determinants for nuclear receptors selectivity: Chemometric analysis, dockings and site-directed mutagenesis of dual peroxisome proliferator-activated receptors α/γ agonists. *Eur J Medicinal Chem* (2013) 63:321–32. doi: 10.1016/j.ejmech.2013.02.015

69. Verma N, Chouhan U. Chemometric modelling of PPAR-α and PPAR-γ dual agonists for the treatment of type-2 diabetes. *Curr Science* (2016) 111(2):356–67. doi: 10.18520/cs/v111/i2/356-367

70. Liu X, Jing Z, Jia WQ, Wang SQ, Ma Y, Xu WR, et al. Identification of novel PPARα/γ dual agonists by virtual screening, ADMET prediction and molecular dynamics simulations. *J Biomol Struct Dyn* (2018) 36(11):2988–3002. doi: 10.1080/07391102.2017.1373706

71. Nath V, Agrawal R, Kumar V. Structure based docking and molecular dynamics studies: Peroxisome proliferator-activated receptors –α/γ dual agonists for treatment of metabolic disorders. *J Biomolecular Structure Dynamics* (2020) 38(2):511–23. doi: 10.1080/07391102.2019.1581089

72. Feng XY, Ding TT, Liu YY, Xu WR, Cheng XC. In-silico identification of peroxisome proliferator-activated receptor (PPAR)α/γ agonists from ligand expo components database. *J Biomolecular Structure Dynamics* (2021) 39(5):1853–64. doi: 10.1080/07391102.2020.1745279

73. Sonawane LV, Bari SB. Ligand-based in silico 3D-QSAR study of PPAR-γ agonists. *Med Chem Res* (2011) 20(7):1005–14. doi: 10.1007/s00044-010-9428-9

74. Chen KC, Chen CYC. In silico identification of potent PPAR- agonists from traditional Chinese medicine: A bioactivity prediction, virtual screening, and molecular dynamics study. *Evidence-Based Complementary Altern Med* (2014) 2014:e192452. doi: 10.1155/2014/192452

75. Maltarollo VG, Togashi M, Nascimento AS, Honorio KM. Structure-based virtual screening and discovery of new PPARδ/γ dual agonist and PPARδ and γ agonists. *PloS One* (2015) 10(3):e0118790. doi: 10.1371/journal.pone.0118790

76. Kaserer T, Obermoser V, Weninger A, Gust R, Schuster D. Evaluation of selected 3D virtual screening tools for the prospective identification of peroxisome proliferator-activated receptor (PPAR) γ partial agonists. *Eur J Medicinal Chem* (2016) 124:49–62. doi: 10.1016/j.ejmech.2016.07.072

77. Zhang GL, Liu FY, Zhang J, Wang LP, Jia EX, Lv SM. Integrated in silico–*in vitro* screening of ovarian cancer peroxisome proliferator-activated receptor-γ agonists against a biogenic compound library. *Med Chem Res* (2018) 27(1):341–9. doi: 10.1007/s00044-017-2060-1

78. Kotha S, Swapna B, Kulkarni VM, Setty. RS, Kumar BH, Harisha R. An in-silico approach: identification of PPAR-γ agonists from seaweeds for the management of alzheimer's disease. *J Biomolecular Structure Dynamics* (2021) 39(6):2210–29. doi: 10.1080/07391102.2020.1747543

79. Lee K, You H, Choi J, No KT. Development of pharmacophore-based classification model for activators of constitutive androstane receptor. *Drug Metab Pharmacokinetics* (2017) 32(3):172–8. doi: 10.1016/j.dmpk.2016.11.005

80. Verma G, Khan MF, Shaquiquzzaman M, Akhtar W, Akhter M, Hasan SM, et al. Molecular interactions of dioxins and DLCs with the xenosensors (PXR and CAR): An in silico risk assessment approach. *J Mol Recognit* (2017) 30(12):e2651. doi: 10.1002/jmr.2651

81. Kortagere S, Krasowski MD, Reschly EJ, Venkatesh M, Mani S, Ekins S. Evaluation of computational docking to identify pregnane X receptor agonists in the ToxCast database. *Environ Health Perspect* (2010) 118(10):1412–7. doi: 10.1289/ehp.1001930

82. Matter H, Anger LT, Giegerich C, Güssregen S, Hessler G, Baringhaus KH. Development of in silico filters to predict activation of the pregnane X receptor (PXR) by structurally diverse drug-like molecules. *Bioorganic Medicinal Chem* (2012) 20(18):5352–65. doi: 10.1016/j.bmc.2012.04.020

83. Dybdahl M, Nikolov NG, Wedebye EB, Jónsdóttir SÓ, Niemelä JR. QSAR model for human pregnane X receptor (PXR) binding: Screening of environmental chemicals and correlations with genotoxicity, endocrine disruption and teratogenicity. *Toxicol Appl Pharmacol* (2012) 262(3):301–9. doi: 10.1016/j.taap.2012.05.008

84. Ratajewski M, Grzelak I, Wiśniewska K, Ryba K, Gorzkiewicz M, Walczak-Drzewiecka A, et al. Screening of a chemical library reveals novel PXR-activating pharmacologic compounds. *Toxicol Letters* (2015) 232(1):193–202. doi: 10.1016/j.toxlet.2014.10.009

85. Cui Z, Kang H, Tang K, Liu Q, Cao Z, Zhu R. Screening ingredients from herbs against pregnane X receptor in the study of inductive herb-drug interactions: Combining pharmacophore and docking-based rank aggregation. *BioMed Res Int* (2015) 2015:e657159. doi: 10.1155/2015/657159

86. Torimoto-Katori N, Huang R, Kato H, Ohashi R, Xia M. In silico prediction of hPXR activators using structure-based pharmacophore modeling. *J Pharm Sci* (2017) 106(7):1752–9. doi: 10.1016/j.xphs.2017.03.004

87. Chen Q, Wang X, Shi W, Yu H, Zhang X, Giesy JP. Identification of thyroid hormone disruptors among HO-PBDEs: *In vitro* investigations and coregulator involved simulations. *Environ Sci Technol* (2016) 50(22):12429–38. doi: 10.1021/acs.est.6b02029

88. Wang F, Xing J. Classification of thyroid hormone receptor agonists and antagonists using statistical learning approaches. *Mol Divers* (2019) 23(1):85–92. doi: 10.1007/s11030-018-9857-9

89. Zhang J, Li Y, Gupta AA, Nam K, Andersson PL. Identification and molecular interaction studies of thyroid hormone receptor disruptors among household dust contaminants. *Chem Res Toxicol* (2016) 29(8):1345–54. doi: 10.1021/acs.chemrestox.6b00171

90. Wang FF, Yang W, Shi YH, Cheng XR, Le GW. Structure-based approach for the study of thyroid hormone receptor binding affinity and subtype selectivity. *J Biomol Struct Dyn* (2016) 34(10):2251–67. doi: 10.1080/07391102.2015.1113384

91. Li F, Xie Q, Li X, Li N, Chi P, Chen J, et al. Hormone activity of hydroxylated polybrominated diphenyl ethers on human thyroid receptor-beta: *in vitro* and in silico investigations. *Environ Health Perspect* (2010) 118(5):602–6. doi: 10.1289/ehp.0901457

92. Shen XL, Takimoto-Kamimura M, Wei J, Gao QZ. Computer-aided *de novo* ligand design and docking/molecular dynamics study of vitamin d receptor agonists. *J Mol Model* (2012) 18(1):203–12. doi: 10.1007/s00894-011-1066-8

93. Yadav DK, Kumar S, Teli MK, Kim MH. Ligand-based pharmacophore modeling and docking studies on vitamin d receptor inhibitors. *J Cell Biochem* (2020) 121(7):3570–83. doi: 10.1002/jcb.29640

94. Jayaraj JM, Reteti E, Kesavan C, Muthusamy K. Structural insights on vitamin d receptor and screening of new potent agonist molecules: Structure and ligand-based approach. *J Biomolecular Structure Dynamics* (2021) 39(11):4148–59. doi: 10.1080/07391102.2020.1775122

95. Klimenko K. In silico identification of endogenous and exogenous agonists of estrogen-related receptor α. *Comput Toxicol* (2019) 10:105–12. doi: 10.1016/j.comtox.2019.01.005

96. Chitrala KN, Yeguvapalli S. Prediction and analysis of ligands against estrogen related receptor alpha. *Asian Pac J Cancer Prev* (2013) 14(4):2371–5. doi: 10.7314/APJCP.2013.14.4.2371

97. Benod C, Carlsson J, Uthayaruban R, Hwang P, Irwin JJ, Doak AK, et al. Structure-based discovery of antagonists of nuclear receptor LRH-1*. *J Biol Chem* (2013) 288(27):19830–44. doi: 10.1074/jbc.M112.411686

98. Rauhamäki S, Postila PA, Lätti S, Niinivehmas S, Multamäki E, Liedl KR, et al. Discovery of retinoic acid-related orphan receptor γt inverse agonists *via* docking and negative image-based screening. *ACS Omega* (2018) 3(6):6259–66. doi: 10.1021/acsomega.8b00603

99. Meijer FA, Doveston RG, de Vries RMJM, Vos GM, Vos AAA, Leysen S, et al. Ligand-based design of allosteric retinoic acid receptor-related orphan receptor γt (RORγt) inverse agonists. *J Med Chem* (2020) 63(1):241–59. doi: 10.1021/acs.jmedchem.9b01372

100. Matsuzaka Y, Uesawa Y. Molecular image-based prediction models of nuclear receptor agonists and antagonists using the deepsnap-deep learning approach with the tox21 10k library. *Molecules* (2020) 25(12):2764. doi: 10.3390/molecules25122764

101. Capuzzi SJ, Politi R, Isayev O, Farag S, Tropsha A. QSAR modeling of Tox21 challenge stress response and nuclear receptor signaling toxicity assays. *Front Environ Sci* (2016) 4:3/full. doi: 10.3389/fenvs.2016.00003/full

102. Mayr A, Klambauer G, Unterthiner T, Hochreiter S. DeepTox: Toxicity prediction using deep learning. *Front Environ Sci* (2016) 3:80. doi: 10.3389/fenvs.2015.00080

103. Banerjee P, Eckert AO, Schrey AK, Preissner R. ProTox-II: A webserver for the prediction of toxicity of chemicals. *Nucleic Acids Res* (2018) 46(W1):W257–63. doi: 10.1093/nar/gky318

104. Kolšek K, Mavri J, Sollner Dolenc M, Gobec S, Turk S. Endocrine disruptome–an open source prediction tool for assessing endocrine disruption potential through nuclear receptor binding. *J Chem Inf Modeling* (2014) 54(4):1254–67. doi: 10.1021/ci400649p

105. Vedani A, Dobler M, Hu Z, Smieško M. OpenVirtualToxLab–a platform for generating and exchanging in silico toxicity data. *Toxicol Letters* (2015) 232(2):519–32. doi: 10.1016/j.toxlet.2014.09.004

106. Park SJ, Kufareva I, Abagyan R. Improved docking, screening and selectivity prediction for small molecule nuclear receptor modulators using conformational ensembles. *J Comput Aided Mol Des* (2010) 24(5):459–71. doi: 10.1007/s10822-010-9362-4

107. Toporova L, Balaguer P. Nuclear receptors are the major targets of endocrine disrupting chemicals. *Mol Cell Endocrinol* (2020) 502:110665. doi: 10.1016/j.mce.2019.110665

108. Porter BA, Ortiz MA, Bratslavsky G, Kotula L. Structure and function of the nuclear receptor superfamily and current targeted therapies of prostate cancer. *Cancers (Basel)* (2019) 11(12):E1852. doi: 10.3390/cancers11121852

109. Jia M, Dahlman-Wright K, Gustafsson JÅ. Estrogen receptor alpha and beta in health and disease. *Best Pract Res Clin Endocrinol Metab* (2015) 29(4):557–68. doi: 10.1016/j.beem.2015.04.008

110. Matthews J, Gustafsson JA. Estrogen signaling: a subtle balance between ER alpha and ER beta. *Mol Interv* (2003) 3(5):281–92. doi: 10.1124/mi.3.5.281

111. Tan ME, Li J, Xu HE, Melcher K, Yong El. Androgen receptor: structure, role in prostate cancer and drug discovery. *Acta Pharmacol Sin* (2015) 36(1):3–23. doi: 10.1038/aps.2014.18

112. Fujita K, Nonomura N. Role of androgen receptor in prostate cancer: A review. *World J Mens Health* (2018) 37(3):288–95. doi: 10.5534/wjmh.180040

113. Schumacher M, Zhu X, Guennoun R. 3.11 - progesterone: Synthesis, metabolism, mechanism of action, and effects in the nervous system. In: DW Pfaff and M Joëls, editors. *Hormones, brain and behavior*, *3rd ed*. Oxford: Academic Press (2017). p. 215–44. Available at: https://www.sciencedirect.com/science/article/pii/B9780128035924000547.

114. Grimm SL, Hartig SM, Edwards DP. Progesterone receptor signaling mechanisms. *J Mol Biol* (2016) 428(19):3831–49. doi: 10.1016/j.jmb.2016.06.020

115. Kadmiel M, Cidlowski JA. Glucocorticoid receptor signaling in health and disease. *Trends Pharmacol Sci* (2013) 34(9):518–30. doi: 10.1016/j.tips.2013.07.003

116. Oakley RH, Cidlowski JA. Cellular processing of the glucocorticoid receptor gene and protein: New mechanisms for generating tissue-specific actions of glucocorticoids. *J Biol Chem* (2011) 286(5):3177–84. doi: 10.1074/jbc.R110.179325

117. Bledsoe RK, Montana VG, Stanley TB, Delves CJ, Apolito CJ, McKee DD, et al. Crystal structure of the glucocorticoid receptor ligand binding domain reveals a novel mode of receptor dimerization and coactivator recognition. *Cell* (2002) 110(1):93–105. doi: 10.1016/S0092-8674(02)00817-6

118. Sharma S, Shen T, Chitranshi N, Gupta V, Basavarajappa D, Sarkar S, et al. Retinoid X receptor: Cellular and biochemical roles of nuclear receptor with a focus on neuropathological involvement. *Mol Neurobiol* (2022) 59(4):2027–50. doi: 10.1007/s12035-021-02709-y

119. Tanaka T, De Luca LM. Therapeutic potential of "rexinoids" in cancer prevention and treatment. *Cancer Res* (2009) 69(12):4945–7. doi: 10.1158/0008-5472.CAN-08-4407

120. Cramer PE, Cirrito JR, Wesson DW, Lee CYD, Karlo JC, Zinn AE, et al. ApoE-directed therapeutics rapidly clear β-amyloid and reverse deficits in AD mouse models. *Science* (2012) 335(6075):1503–6. doi: 10.1126/science.1217697

121. Mysinger MM, Carchia M, JohnJ I, Shoichet BK. Directory of useful decoys, enhanced (dud-e): Better ligands and decoys for better benchmarking. *J Medicinal Chem* (2012) 55(14):6582–94. doi: 10.1021/jm300687e

122. Kahremany S, Livne A, Gruzman A, Senderowitz H, Sasson S. Activation of PPARδ: From computer modelling to biological effects. *Br J Pharmacol* (2015) 172(3):754–70. doi: 10.1111/bph.12950

123. Pyper SR, Viswakarma N, Yu S, Reddy JK. PPARalpha: Energy combustion, hypolipidemia, inflammation and cancer. *Nucl Recept Signal* (2010) 8:e002. doi: 10.1621/nrs.08002

124. Astapova O, Leff T. Adiponectin and PPARγ: Cooperative and interdependent actions of two key regulators of metabolism. *Vitam Horm* (2012) 90:143–62. doi: 10.1016/B978-0-12-398313-8.00006-3

125. Grienke U, Mihály-Bison J, Schuster D, Afonyushkin T, Binder M, Guan S, et al. Identification of novel PPARα/γ dual agonists by virtual screening, ADMET prediction and molecular dynamics simulations. *J Biomolecular Structure Dynamics* (2021) 36(11):2988–3002. doi: 10.1080/07391102.2017.1373706?journalCode=tbsd20

126. Porskjær Christensen L, Bahij El-Houri R. Development of an *in vitro* screening platform for the identification of partial pparγ agonists as a source for antidiabetic lead compounds. *Molecules* (2018) 23(10):2431. doi: 10.3390/molecules23102431

127. Todorov MP. In silico identification of human pregnane x receptor activators. *Ecol Safety* (2015) 9(1):9–17.

128. Zhang Ym, Chang Mj, Yang Xs, Han X. In silico investigation of agonist activity of a structurally diverse set of drugs to hPXR using HM-BSM and HM-PNN. *J Huazhong Univ Sci Technol [Med Sci]* (2016) 36(3):463–8. doi: 10.1007/s11596-016-1609-4

129. Balaguer P, Delfosse V, Grimaldi M, Bourguet W. Structural and functional evidences for the interactions between nuclear hormone receptors and endocrine disruptors at low doses. *Comptes Rendus Biologies* (2017) 340(9–10):414–20. doi: 10.1016/j.crvi.2017.08.002

130. Ma Z, Deng C, Hu W, Zhou J, Fan C, Di S, et al. Liver X receptors and their agonists: Targeting for cholesterol homeostasis and cardiovascular diseases. *Curr Issues Mol Biol* (2017) 22:41–64. doi: 10.21775/cimb.022.041

131. Buñay J, Fouache A, Trousson A, de Joussineau C, Bouchareb E, Zhu Z, et al. Screening for liver X receptor modulators: Where are we and for what use? *Br J Pharmacol* (2021) 178(16):3277–93. doi: 10.1111/bph.15286

132. Piccinin E, Cariello M, Moschetta A. Lipid metabolism in colon cancer: Role of liver X receptor (LXR) and stearoyl-CoA desaturase 1 (SCD1). *Mol Aspects Med* (2021) 78:100933. doi: 10.1016/j.mam.2020.100933

133. Auerbach SS, Ramsden R, Stoner MA, Verlinde C, Hassett C, Omiecinski CJ. Alternatively spliced isoforms of the human constitutive androstane receptor. *Nucleic Acids Res* (2003) 31(12):3194–207. doi: 10.1093/nar/gkg419

134. Forman BM, Tzameli I, Choi HS, Chen J, Simha D, Seol W, et al. Androstane metabolites bind to and deactivate the nuclear receptor CAR-beta. *Nature* (1998) 395(6702):612–5. doi: 10.1038/26996

135. Tzameli I, Moore DD. Role reversal: New insights from new ligands for the xenobiotic receptor CAR. *Trends Endocrinol Metab* (2001) 12(1):7–10. doi: 10.1016/S1043-2760(00)00332-5

136. Liu YY, Brent GA. Thyroid hormone crosstalk with nuclear receptor signaling in metabolic regulation. *Trends Endocrinol Metab* (2010) 21(3):166–73. doi: 10.1016/j.tem.2009.11.004

137. Mullur R, Liu YY, Brent GA. Thyroid hormone regulation of metabolism. *Physiol Rev* (2014) 94(2):355–82. doi: 10.1152/physrev.00030.2013

138. Saponaro F, Sestito S, Runfola M, Rapposelli S, Chiellini G. Selective thyroid hormone receptor-beta (TRβ) agonists: New perspectives for the treatment of metabolic and neurodegenerative disorders. *Front Med (Lausanne)* (2020) 7:331. doi: 10.3389/fmed.2020.00331

139. Zoeller RT. Environmental chemicals as thyroid hormone analogues: new studies indicate that thyroid hormone receptors are targets of industrial chemicals? *Mol Cell Endocrinol* (2005) 242(1–2):10–5. doi: 10.1016/j.mce.2005.07.006

140. Kim MJ, Park YJ. Bisphenols and thyroid hormone. *Endocrinol Metab (Seoul)* (2019) 34(4):340–8. doi: 10.3803/EnM.2019.34.4.340

141. Pal S, Kumar V, Kundu B, Bhattacharya D, Preethy N, Reddy MP, et al. Ligand-based pharmacophore modeling, virtual screening and molecular docking studies for discovery of potential topoisomerase i inhibitors. *Comput Struct Biotechnol J* (2019) 17:291–310. doi: 10.1016/j.csbj.2019.02.006

142. Yaghmaei S, Roberts C, Ai R, Mizwicki MT, Chang CEA. Agonist and antagonist binding to the nuclear vitamin d receptor: Dynamics, mutation effects and functional implications. *In Silico Pharmacol* (2013) 1:2. doi: 10.1186/2193-9616-1-2

143. Giguère V. To ERR in the estrogen pathway. *Trends Endocrinol Metab* (2002) 13(5):220–5. doi: 10.1016/S1043-2760(02)00592-1

144. Gallet M, Vanacker JM. ERR receptors as potential targets in osteoporosis. *Trends Endocrinol Metab* (2010) 21(10):637–41. doi: 10.1016/j.tem.2010.06.008

145. Kumar N, Solt LA, Conkright J, Wang Y, Istrate MA, Busby SA, et al. Campaign to identify novel modulators of the retinoic acid receptor-related orphan receptors (ROR). In: *Probe reports from the NIH molecular libraries program [Internet]*. Bethesda (MD: National Center for Biotechnology Information (US (2010). Available at: http://www.ncbi.nlm.nih.gov/books/NBK56239/.

146. Nishiyama Y, Nakamura M, Misawa T, Nakagomi M, Makishima M, Ishikawa M, et al. Structure–activity relationship-guided development of retinoic acid receptor-related orphan receptor gamma (RORγ)-selective inverse agonists with a phenanthridin-6(5H)-one skeleton from a liver X receptor ligand. *Bioorganic Medicinal Chem* (2014) 22(9):2799–808. doi: 10.1016/j.bmc.2014.03.007

147. Balaguer P, Delfosse V, Bourguet W. Mechanisms of endocrine disruption through nuclear receptors and related pathways. *Curr Opin Endocrine Metab Res* (2019) 7:1–8. doi: 10.1016/j.coemr.2019.04.008

148. Vogel CFA, Van Winklea LS, Esser C, Haarmann-Stemmann T. The aryl hydrocarbon receptor as a target of environmental stressors – implications for pollution mediated stress and inflammatory responses. *RedOx Biol* (2020) 34:101530. doi: 10.1016/j.redox.2020.101530

149. Gu C, Goodarzi M, Yang X, Bian Y, Sun C, Jiang X. Predictive insight into the relationship between AhR binding property and toxicity of polybrominated diphenyl ethers by PLS-derived QSAR. *Toxicol Lett* (2012) 208(3):269–74. doi: 10.1016/j.toxlet.2011.11.010

150. Cao F, Li X, Ye L, Xie Y, Wang X, Shi W, et al. Molecular docking, molecular dynamics simulation, and structure-based 3D-QSAR studies on the aryl hydrocarbon receptor agonistic activity of hydroxylated polychlorinated biphenyls. *Environ Toxicol Pharmacol* (2013) 36(2):626–35. doi: 10.1016/j.etap.2013.06.004

151. Wu D, Potluri N, Kim Y, Rastinejad F. Structure and dimerization properties of the aryl hydrocarbon receptor pas-a domain. *Mol Cell Biol* (2013) 33(21):4346–56. doi: 10.1128/MCB.00698-13

152. Giani Tagliabue S, Faber SC, Motta S, Denison MS, Bonati L. Modeling the binding of diverse ligands within the ah receptor ligand binding domain. *Sci Rep* (2019) 9:10693. doi: 10.1038/s41598-019-47138-z

153. Motto I, Bordogna A, Soshilov AA, Denison MS, Bonati L. New aryl hydrocarbon receptor homology model targeted to improve docking reliability. *J Chem Inf Modeling* (2021) 51(11):2868–81. doi: 10.1021/ci2001617

154. Parks AJ, Pollastri MP, Hahn ME, Stanford EA, Novikov O, Franks DG, et al. In silico identification of an aryl hydrocarbon receptor antagonist with biological activity in vitro and *In vivo*. *Mol Pharmacol* (2014) 86(5):593–608. doi: 10.1124/mol.114.093369

155. Seidel T, Ibis G, Bendix F, Wolber G. Strategies for 3D pharmacophore-based virtual screening. *Drug Discovery Today: Technologies* (2010) 7(4):e221–8. doi: 10.1016/j.ddtec.2010.11.004

156. Morris GM, Lim-Wilby M. Molecular docking. In: A Kukol, editor. *Molecular modeling of proteins*. Totowa: NJ: Humana Press (2008). p. 365–82. doi: 10.1007/978-1-59745-177-2_19

157. Ben Nasr N, Guillemain H, Lagarde N, Zagury JF, Montes M. Multiple structures for virtual ligand screening: defining binding site properties-based criteria to optimize the selection of the query. *J Chem Inf Model* (2013) 53 (2):293–311. doi: 10.1021/ci3004557

158. International union of pure and applied chemistry. IUPAC. Available at: https://iupac.org/

159. Yang SY. Pharmacophore modeling and applications in drug discovery: Challenges and recent advances. *Drug Discovery Today* (2010) 15(11):444–50. doi: 10.1016/j.drudis.2010.03.013

160. Gedeck P, Rohde B, Bartels C. QSAR – how good is it in practice? comparison of descriptor sets on an unbiased cross section of corporate data sets. *J Chem Inf Model* (2006) 46(5):1924–36. doi: 10.1021/ci050413p

161. Jiménez-Luna J, Grisoni F, Weskamp N, Schneider G. Artificial intelligence in drug discovery: recent advances and future perspectives. *Expert Opin Drug Discov* (2021) 16(9):949–59. doi: 10.1080/17460441.2021.1909567

162. Weaver S, Gleeson MP. The importance of the domain of applicability in QSAR modeling. *J Mol Graphics Modelling* (2008) 26(8):1315–26. doi: 10.1016/j.jmgm.2008.01.002

163. Sellami A, Montes M, Lagarde N. Predicting potential endocrine disrupting chemicals binding to estrogen receptor α (ERα) using a pipeline combining structure-based and ligand-based in silico methods. *Int J Mol Sci* (2021) 22 (6):2846. doi: 10.3390/ijms22062846

164. Hein M, Zilian D, Sotriffer CA. Docking compared to 3D-pharmacophores: the scoring function challenge. *Drug Discovery Today: Technologies* (2010) 4(7):e229–36. doi: 10.1016/j.ddtec.2010.12.003

165. Vuorinen A, Odermatt A, Schuster D. In silico methods in the discovery of endocrine disrupting chemicals. *J Steroid Biochem Mol Biol* (2013) 137:137:18–26. doi: 10.1016/j.jsbmb.2013.04.009

166. Vázquez J, López M, Gibert E, Herrero E, Luque FJ. Merging ligand-based and structure-based methods in drug discovery: An overview of combined virtual screening approaches. *Molecules* (2020) 25(20):4723. doi: 10.3390/molecules25204723

167. Lunghini F, Marcou G, Azam P, Bonachera F, Enrici MH, Van Miert E, et al. Endocrine disruption: The noise in available data adversely impacts the models' performance. *SAR QSAR Environ Res* (2021) 32(2):111–31. doi: 10.1080/1062936X.2020.1864468

168. Réau M, Langenfeld F, Zagury JF, Lagarde N, Montes M. Decoys selection in benchmarking datasets: Overview and perspectives. *Front Pharmacol* (2018) 9:11. doi: 10.3389/fphar.2018.00011

169. Réau M, Lagarde N, Zagury JF, Montes M. Nuclear receptors database including negative data (nr-dbind): a database dedicated to nuclear receptors binding data including negative data and pharmacological profile. *J Med Chem* (2019) 62(6):2894–904. doi: 10.1021/acs.jmedchem.8b01105

170. Goya-Jorge E, Giner RM, Sylla-Iyarreta Veitía M, Gozalbes R, Barigye SJ. Predictive modeling of aryl hydrocarbon receptor (AhR) agonism. *Chemosphere* (2020) 256:127068. doi: 10.1016/j.chemosphere.2020.127068

171. Begley CG, Ioannidis JPA. Reproducibility in science. *Circ Res* (2015) 116 (1):116–26. doi: 10.1161/CIRCRESAHA.114.303819

172. Munafò MR, Nosek BA, Bishop DVM, Button KS, Chambers CD, du Sert NP, et al. A manifesto for reproducible science. *Nat Hum Behav* (2017) 1:0021. doi: 10.1038/s41562-016-0021

173. Stodden V, Guo P, Ma Z. Toward reproducible computational research: An empirical analysis of data and code policy adoption by journals. *PloS One* (2013) 8 (6):e67111. doi: 10.1371/journal.pone.0067111

174. Thomas RS, Paules RS, Simeonov A, Fitzpatrick SC, Crofton KM, Casey WM, et al. The US federal Tox21 program: A strategic and operational plan for continued leadership. *ALTEX* (2018) 35(2):163–8. doi: 10.14573/altex.1803011

1.1.3. Discussion

À la suite de la réalisation de cette revue, une première constatation peut être faite sur la disparité des études réalisées entre les différents NR. Alors que certains constituent un sujet d'intérêt de longue date comme les ER et les AR, pour d'autres le nombre d'études qui leur est consacrés est relativement réduit voire nul. Cela s'explique essentiellement par la variabilité de la quantité de données disponibles pour chaque NR, les connaissances acquises sur le mécanisme d'action et la physiopathologie étant disparate. Ceci est aussi vrai du point de vue structural puisque le nombre et la qualité des structures 3D résolues expérimentalement est aussi très variable. La connaissance moléculaire approfondie découlant de l'étude de ces structures et la capacité à réaliser des études SB est donc impactée. Le premier NR à avoir été cloné et dont le mécanisme a été étudié depuis 1962 est ER (qui sera plus précisément renommé ERα plus tard) [312]. 20 ans plus tard, la structure du récepteur complémentaire à l'ADN (cDNA) de GR a été élucidée [313], suivie par celle de ER [314] et MR[315]. L'ensemble de ces découvertes a permis de faire émerger le domaine de recherche sur les récepteurs aux stéroïdes. Par la suite d'autres structures « non stéroïdes » ont été identifiées et apparentées aux NR dont TR[316] et RAR [317]. Durant la décennie qui a suivi, le nombre croissant de recherche sur les NR a permis d'affiner les connaissances sur les récepteurs ER, AR, TR, GR et VDR mais a surtout permis de découvrir une nouvelle catégorie de NR : les récepteurs nucléaires orphelins [318] et leurs chefs de file ERRα et β[319]. Aujourd'hui, les NR figurent parmi les cibles thérapeutiques les plus utilisées et les plus fructueuses avec de nombreuses molécules destinées au traitement de maladies métaboliques (diabète, dyslipidémies etc..) mais aussi des pathologies plus graves notamment les cancers du sein et de la prostate. Cependant, depuis plusieurs années maintenant, les NR sont étudiés dans un contexte toxicologique, notamment pour leur capacité à lier des composés environnementaux ; les perturbateurs endocriniens. Ce nouveau contexte d'étude a aussi été retrouvé dans cette revue puisque la majorité des modèles créés depuis 2010 l'ont été pour une application toxicologique (**Figure 23**).

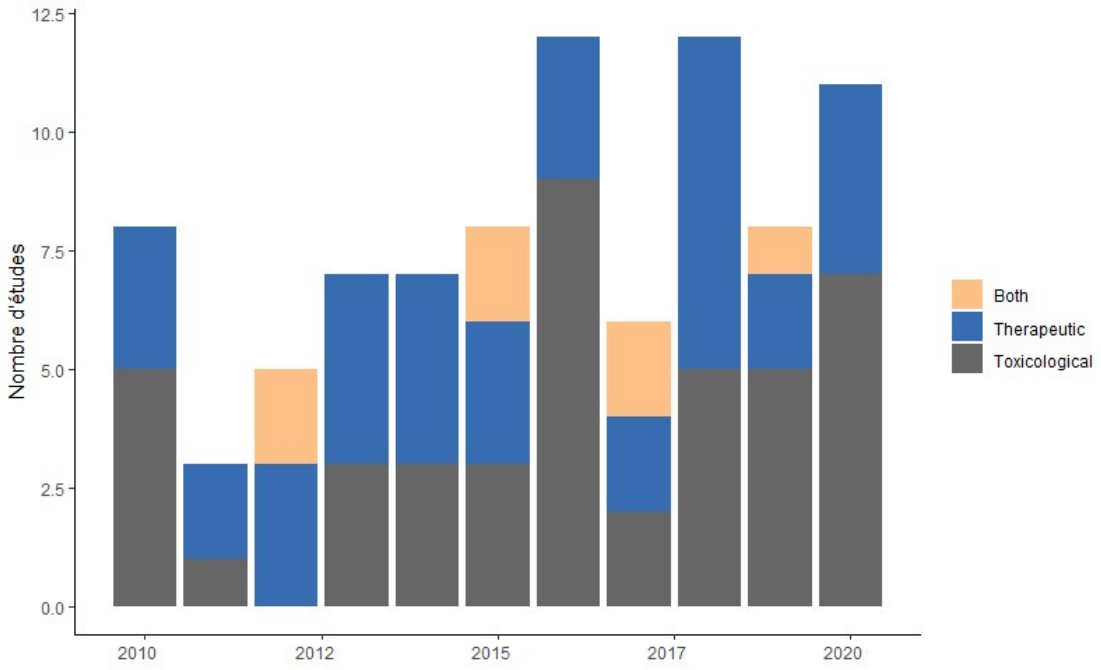


Figure 23 : Répartition des études incluses dans la revue selon le contexte toxicologique ou thérapeutique

Une autre constatation importante est la prépondérance des modèles créés à partir de méthodes QSAR (et de machine learning) pour l'application toxicologique. 48 études QSAR ont été recueillies dont 31 en toxicologie. Par ailleurs, nous avons aussi remarqué que les méthodes de docking étaient aussi bien présentes que ce soit pour les applications en toxicologie qu'en thérapeutique (17 études à visée toxicologique et 28 à visée thérapeutique). La disponibilité et le traitement des données étaient enfin au centre de la revue et une attention particulière a été portée sur l'évaluation de la reproductibilité des données dans le domaine des NR. Ce dernier a été évalué en considérant quelques critères comme la disponibilité des données chimiques, la diffusion des codes, l'utilisation de logiciels en libre accès et le détail des protocoles pour classer les études analysées en trois catégories selon le niveau de reproductibilité « fort », « moyen » et « faible ». Seule une faible proportion d'articles (~18 %) était décrite par un fort niveau de reproductibilité. Pour la majorité des études, la reproductibilité était moyenne généralement caractérisée par une absence de mise à disposition des modèles développé ou bien à cause d'une description très sommaire du protocole suivi. Or le détail de toutes les étapes suivies pour nettoyer et préparer les données ainsi que les étapes pour obtenir les résultats sont tout aussi importants que les données elles-mêmes. Ces détails peuvent inclure par exemple des informations sur la manière de gérer les *outlyers* ou les données manquantes qui sont généralement omis par la communauté scientifique [320].

La revue ici présentée s'est concentrée essentiellement sur les études impliquant le mécanisme d'actions de liaison directe et donc sur la liaison au LBD. Toutefois, il aurait pu être intéressant d'aborder les études s'intéressant aussi au DBD (DNA binding demain). En effet, cette région présente beaucoup d'intérêts pour de nombreux NR notamment les AR dans les études prospective à la recherche de composé contre le cancer de la prostate [321].

1.1.4. Conclusion et perspectives

Il n'existe pas encore à notre connaissance d'outil en libre accès qui permet de cribler un grand nombre de molécule en une seule fois pour la prédiction de leur potentiel PE. Les outils comme VirtualToxlab[322,323] et endocrine disruptome [324] ont la même vocation que la méthode que nous voulons mettre à disposition présentée dans la deuxième partie de cette partie résultats. Cependant les prédictions doivent se faire molécule par molécule. D'après les analyses de la revue, 3 grandes méthodes prédominent à savoir les modèles QSAR, le docking et les pharmacophores. Afin d'avoir une vue mécanistique, nous avons décidé de nous concentrer sur les deux dernières approches à savoir le docking et les modèles de pharmacophores. De plus, cette revue nous a permis de relever le grand nombre d'études destinées ainsi que l'abondance des données relatives au récepteur ERα que nous avons sélectionné pour réaliser la preuve de concept de notre projet.

1.2. Préparation des bases de données de docking

1.2.1. Introduction

Lors de la réalisation de la revue présentée précédemment, nous avons relevé les différentes bases de données utilisées lors des différentes études analysées dont certaines étaient spécifiquement dédiées aux NR existantes. Nous avons remarqué une grande disparité dans la manière dont l'information sur les molécules, que ce soient leurs identifiants, leurs noms ou encore leurs activités biologiques, était présentée. Afin de construire des jeux de données sur les NR, puis des modèles de prédiction de haute qualité, il est donc nécessaire d'extraire, nettoyer et préparer minutieusement ces molécules. À la suite d'une invitation à participer à un livre s'inscrivant dans la thématique du docking et de la conception *in silico* de molécules thérapeutiques, nous avons choisi de proposer un chapitre traitant cette problématique. En effet, en plus de la nécessité d'une bonne connaissance de la cible protéique étudiée [198], le choix des bases de données virtuelles est tout aussi crucial avant d'employer les méthodes de criblage virtuel et en particulier les méthodes de docking. L'article " *Virtual Libraries for Docking Methods: Guidelines for the Selection and the Preparation* ", est le 5ème chapitre d'un ouvrage collaboratif intitulé " *Molecular Docking for Computer-Aided Drug Design* " publié en 2021.
Dans ce chapitre, nous présentons les différentes catégories de bases de données de criblage pouvant être utilisées directement pour le docking dans le cadre d'une campagne de *drug design*. Nous distinguons entre bases de données de composés déjà synthétisés et les bases de données de composés virtuels. Pour chaque catégorie, nous avons listé quelques exemples avec le nombre de composés inclus ainsi que la nature commerciale ou libre. Nous évoquons aussi la difficulté de choisir une ou plusieurs bases de données. En effet, les bases de données sont généralement très denses et structurellement diverses, il est encore difficile, malgré l'évolution récente des moyens de calcul et des méthodes de docking à très large échelle [325], de les cribler en entier dans des délais raisonnables et surtout avec des moyens computationnels réduits. Ensuite, nous proposons un protocole débutant par la sélection des données à intégrer à la base de données et leur nettoyage à l'utilisation de différents filtres (ADME/Tox, PAINS). Enfin, nous listons dans un tableau différents outils de préparation de bases de données en détaillant pour chacun les étapes de préparation inclues ou non dans l'outil.

1.2.2. Publication

CHAPTER 5

Virtual Libraries for Docking Methods: Guidelines for the Selection and the Preparation

ASMA SELLAMI • MANON RÉAU • FLORENT LANGENFELD • NATHALIE LAGARDE[a] • MATTHIEU MONTES[a]

1 INTRODUCTION

Current drug discovery relies on massive screening of chemical libraries against various extracellular and intracellular molecular targets to identify novel chemotypes with the desired mode of action. Protein–ligand molecular docking methods are nowadays widely implemented in these drug discovery pipelines as an in silico analogy of experimental high-throughput screening of large databases of compounds (Kontoyianni, 2017).

The success of this approach in identifying new potent therapeutic compounds lies for many parts in the availability of accurate docking algorithms and scoring functions and of high-quality data on the structure of the macromolecular target. The choice of the compounds database to screen is as important as the choice of the parameters mentioned above. In fact, the quality of the input is key to ensure reliable docking results, and it is very unlikely that active compounds could be identified within poorly selected compound collections, despite the use of an ideal docking tool (Forli, 2015). Several studies have shown that the size (Lyu et al., 2019), the composition, and the preparation (Corbeil et al., 2012) of virtual databases impact the success of docking predictions.

In recent years, high-throughput technologies for combinatorial and multiparallel chemical synthesis, automation technologies for the isolation of natural products, and the availability of large compound collections from commercial sources have substantially increased the size and diversity of synthesizable compound collections available for virtual screening. However, navigating and cherry-picking compounds within this huge number of compounds (Williams et al., 2012) is a challenging task and no gold standard protocol has currently been defined (Gally et al., 2019).

This chapter aims at helping choose and construct virtual screening databases dedicated to docking methods. We will, in the first part, provide insights about the possible sources of compounds and how they can be combined to create a customized database adapted to a drug discovery campaign. In the second part, we will provide an extensive description of the steps required for compounds preparation.

2 DATA COLLECTION WITHIN THE VAST CHEMICAL SPACE

The chemical space is defined as a "comprehensive collection of all possible small molecules under some reasonable restrictions considering size and composition" (Vogt, 2020). Practically, all possibly existing compounds are mapped on a mathematical space and their positions are defined according to their properties (Arús-Pous et al., 2019; Awale et al., 2017). Even if thousands to billions of natural and synthetic compounds are listed in commercial or academic, bioactivity, and natural products databases, this represents only a very small coverage of the entire possible chemical space. Various teams are thus working toward the identification of virtual compounds belonging to areas of this chemical space not yet covered by commercially available compound collections (Kontijevskis, 2017). To do so, fragment libraries are used as a starting point for building new molecules. Known chemical reactions

[a]These authors contributed equally to this work.

Molecular Docking for Computer-Aided Drug Design. https://doi.org/10.1016/B978-0-12-822312-3.00017-5
Copyright © 2021 Elsevier Inc. All rights reserved.

are applied on known building blocks and are used to generate synthetically feasible compounds or all possible molecules according to these chemical rules.

In this section, we will present different sources for collecting compounds to build virtual screening libraries. It is to note that the purpose here is not to provide a complete enumeration of all available compound libraries. The databases will be divided into actual compound collections, i.e., physically available molecules representing the total output of the pharmaceutical field (both academic and commercial collections) and virtual compound collections, i.e., theoretically tangible and synthesizable molecules (Reymond, 2015). We will also provide guidelines to choose the appropriate database tailored for a drug design project.

2.1 Actual Compounds Collections

2.1.1 Commercial and academic databases

Commercial databases represent an important source of compounds for virtual screening. They often contain more than 1 million molecules (Table 5.1). The major advantage of these commercial collections of compounds lies in the possibility for the customer to cherry-pick and rapidly obtain selected compounds to be used in experimental assays. The average price is estimated between US$50 and US$200 for 5 mg of a given compound but it can vary according to the supplier, the amount of product ordered, and the complexity of the molecule (Rognan & Bonnet, 2014). Despite their commercial purpose, suppliers usually offer free access to the molecule's structures, provided in various file formats (2D or 3D). These databases undergo frequent updates, new products being added while other being either removed or out of stock. Virtual screening libraries constructed from these databases should ideally be prepared when the whole virtual screening protocol is settled and ready to be used.

Academic laboratories have always been a source for innovative compounds that chemical suppliers acquire and integrate in their commercial databases (Rognan & Bonnet, 2014). Initiatives to gather these academic databases within an institution, a country, or an ensemble of countries succeed in the creation of large

TABLE 5.1
Examples of Actual Compounds Collections.

	Database Name	Access	Number of Compounds	Compounds Purchasability	Website
Commercial	ChemSpider	Free	~67 millions	Link to vendors	http://www.chemspider.com/
	Ambinter	Free	~7 millions (screening compounds)	Yes	http://www.ambinter.com/
	eMolecules	Free	~7 millions	Link to vendors	https://www.emolecules.com/
	MolPort	Free	~7 millions	Yes	https://www.molport.com/
	Enamine	Free upon registration	~2.7 millions	Yes	https://enamine.net/
	ChemDiv	Free (e-mail address requested)	~1.6 millions	Yes	https://www.chemdiv.com/
	ChemBridge	Free	~1.3 millions	Yes	https://www.chembridge.com/
	IBS	Free upon registration	~550,000	Yes	https://www.ibscreen.com/
	Life Chemicals	Free (e-mail address requested)	~490,000	Yes	https://lifechemicals.com/
	Specs	Free upon registration	~350,000	Yes	https://www.specs.net/
	Asinex	Free	~260,000	Yes	https://www.asinex.com/
	Maybridge	Free upon registration	~53,000	Yes	https://www.maybridge.com/

TABLE 5.1
Examples of Actual Compounds Collections.—cont'd

	Database Name	Access	Number of Compounds	Compounds Purchasability	Website
Bioactivity	ZINC15	Free	~750 millions	Link to vendors	https://zinc15.docking.org/
	PubChem	Free	~103 millions	Link to vendors when available	https://pubchem.ncbi.nlm.nih.gov/
	ChEMBL	Free	~1,96 millions	Link to vendors when available	https://www.ebi.ac.uk/chembl/
	BindingDB	Free	~800,000	List of purchasable compounds	https://www.bindingdb.org/bind/index.jsp
	CARLSBAD	Free	~430,000	No	http://carlsbad.health.unm.edu/carlsbad/
	GLASS	Free	~340,000	No	https://zhanglab.ccmb.med.umich.edu/GLASS/index.html
	DrugBank	Free	~22,000	No	https://www.drugbank.ca/
	NR-DBIND	Free	7593	No	http://nr-dbind.drugdesign.fr/
	DrugCentral	Free	4052	No	http://drugcentral.org/
	SuperDrug2	Free	3992	No	http://cheminfo.charite.de/superdrug2/
Natural products	Super Natural II	Free	325,508	Yes	http://bioinf-applied.charite.de/supernatural_new/
	Dictionary of Natural Products	Commercial	~230,000	No	http://dnp.chemnetbase.com/
	Reaxys	Commercial	~220,000	No	https://www.reaxys.com/
	Antibase	Commercial	~40,000	No	https://www.wiley.com/en-us/AntiBase%3A+The+Natural+Compound+Identifier-p-9783527343591
	MarinLit	Commercial	~29,000	No	http://pubs.rsc.org/marinlit/
	The Natural Product Atlas	Free	~25,500	No	https://www.npatlas.org/joomla/
	AfroDB	Free	~900	No	http://african-compounds.org/nanpdb/

public databases such as the Molecular Libraries Probe Production Centers Network (MLPCN) database (National Center for Biotechnology Information (US), 2010) or the European Chemical Biology Library (ECBL) (Horvath et al., 2014). Using compounds belonging to academic databases in virtual screening protocols presents a dual interest (Rognan & Bonnet, 2014). The first advantage is that compounds gathered from different academic databases should guarantee a certain chemical diversity because each laboratory is usually focused on specific chemical scaffolds. The second one is the possibility of settling a collaboration with the academic laboratory that synthesized a given hit compound because its expertise on this chemical series could facilitate a future hit-to-lead optimization.

2.1.2 Bioactivity databases

Bioactivity databases are particularly important for drug discovery processes not only to get knowledge on biological targets and their modulation mechanisms but also to construct benchmarking datasets to design docking protocols prior to prospective virtual screenings and to construct predictive models of activity (Huang et al., 2006; Lagarde et al., 2015; Mysinger et al., 2012; Réau et al., 2018). This category includes databases such as the ZINC (Sterling & Irwin, 2015) or PubChem (Kim et al., 2019) (Table 5.1).

ZINC (ZINC Is Not Commercial) is a freely available compound collection that was initially developed to give access to the chemical structures of molecules included in commercial vendors catalogs encoded in file formats suitable for virtual screening experiments (SMILES, mol2, 3D SDF, and DOCK flexibase formats) (Irwin & Shoichet, 2005). The ZINC database was thus particularly designed to facilitate virtual screening databases preparation (Irwin, 2008). However, following the requests of ZINC investigators, the new version of the ZINC released in 2015, and thus named ZINC15, still provides purchasable data and ready-to-dock 3D formats but now also encompasses, whenever available, biological data on the proteins and biological processes modulated for more than 230 million compounds (Sterling & Irwin, 2015).

The PubChem (Kim et al., 2019) and ChEMBL databases (Gaulton et al., 2017) are two other examples of public bioactivity databases including both biological and structural information. PubChem is an open chemistry database, hosted by the US National Center for Biotechnology Information (NCBI), that gathers data from a wide variety of sources, among which are government agencies, pharmaceutical companies, chemical vendor catalogs, and scientific literature (Kim et al., 2016). PubChem is divided into three interconnected databases named "Substance," "Compound," and "BioAssay." The PubChem Compound database includes, as on July 2020, more than 100 million unique chemical structures. ChEMBL (Gaulton et al., 2017) is a manually curated open database that receives data from public and commercial organizations and scientific literature (patents and publications). ChEMBL includes over 16 million bioactivity data for over 1.96 million distinct compounds and more than 13,000 targets.

Focused bioactivity libraries are also available. They can be dedicated to a given protein family like the nuclear receptors (NRLiSt BDB (Lagarde et al., 2014), NR-DBIND (Réau et al., 2019)), the G-protein–coupled receptor (GLL) (Gatica & Cavasotto, 2012), GLASS (Chan et al., 2015), or protein kinases (PKIDB) (Carles et al., 2018).

Finally, the databases collecting information about molecules already approved as drugs or currently evaluated in clinical trials (the "drug" subset of the ChEMBL database (Gaulton et al., 2012), DrugBank (Wishart et al., 2018), DrugCentral (Ursu et al., 2019), SuperDrug2 (Siramshetty et al., 2018), and Drugs-lib (Lagarde et al., 2018)) can also be of interest. These compounds can be used for repositioning studies, i.e., find new biological target and therapeutic indications for approved or investigational drugs (Pushpakom et al., 2019). Drug repositioning is one of the emergent strategies to overcome drug attrition rates and to speed up the drug discovery process. It was shown of particular importance for cancer (Würth et al., 2016), rare diseases applications (Delavan et al., 2018) and also in situations where an urgent need of a therapeutic solution that can be immediately administered to patients is needed, such as the recent COVID-19 pandemic (Serafin et al., 2020).

The main drawback of these bioactivity databases is that synthesized samples of compounds of interest may not be easily or directly available, unlike in commercial and academic databases because bioactivity databases encompass data from a wide range of sources, including scientific publications and patents.

2.1.3 Natural products

Natural products were the first drugs ever used and have always been a source of drugs. In a recent retrospective study, it has been reported that between January 1, 1981 and September 30, 2019, 23.5% of all new approved drugs and 33.6% of new approved small molecules drugs were natural products or derivatives of natural products (Newman & Cragg, 2020). The chemical space covered by natural products is quite dissimilar to the one occupied by synthetic drug-like compounds (Morrison & Hergenrother, 2014) and natural products are believed to constitute promising starting point for drug discovery (Rodrigues et al., 2016). Natural products databases can thus be used as a source for virtual screening libraries. Numerous databases of natural products are available (for a review, see Sorokina and Steinbeck (2020), Chen et al. (2017)) that can be commercial or freely accessible (Table 5.1). Natural products databases can be comprehensive or focused on natural products used in traditional medicine (such as the AfroDb (Ntie-Kang et al., 2013) or the TCM database@Taiwan (Chen, 2011)). The most complete database of natural products is to date the Super Natural II with more than 320,000 compounds (Banerjee et al., 2015).

2.2 Virtual Compounds Collections

To access unexplored and intellectual property-free area of the chemical space, an emerging field is to use

virtual compounds, i.e., possible molecules but with undescribed synthetic access. Virtual compounds can be designed using fragment-based docking methods by assembling known building block using known chemical reactions or by enumerating all molecules respecting defined rules with meaningful chemical structures.

2.2.1 Fragment-based databases

Fragment-based drug discovery has shown efficiency in providing new drug candidates in the last 20 years (Erlanson et al., 2016). The success of fragment-based docking methods relies on the availability and quality of fragment libraries. The size and the composition of the fragment library directly influence the outcomes of fragment-based docking and it has been shown that a diverse library with size limited to 2000 fragments should be preferred (Böttcher et al., 2019; Messick et al., 2019; Shi & von Itzstein, 2019). Fragments included in the virtual screening libraries mainly comply to the "Rule of Three" (Congreve et al., 2003), derived from "Lipinski's rule of five" (Ro5), with a molecular weight inferior to 300 Da and a value inferior to 3 for the numbers of both hydrogen bond donors and hydrogen bond acceptors and for the partition coefficient between n-octanol and water (clogP) (a number of rotatable bond inferior to 3 and a polar surface area inferior to 60 $Å^2$ are also recommended). Fragments from large commercial libraries can be purchased for experimental testing and chemistry synthesis as shown in Table 5.2.

More than 10 million virtual fragments complying to the "Rule of Three" are provided in the FDB-17 (Visini et al., 2017). These fragments present the advantage to cover a large chemical space with a wide range of molecular size, polarity, and complexity. However, fragments of the FDB-17 identified as hits in virtual screening protocols are not directly purchasable for experimental validation.

TABLE 5.2
Examples of Purchasable Fragment-Based Databases.

Database Name	Number of Fragments	Web Interface
Enamine Fragment Collection	172,723	https://enamine.net/fragments/fragment-collection
Asinex's Fragments	20,117	http://www.asinex.com/fragments/
OTAVAchemicals General Fragments Library	13,685	https://otavachemicals.com/products/fragment-libraries/general-fragment-library
ChemBridge Fragment Library	13,500	https://www.chembridge.com/screening_libraries/fragment_library/,
Maybridge Ro3 Library	2500	https://www.maybridge.com/
Prestwick Drug-Fragment Library	1456	http://www.prestwickchemical.com
Prestwick F2L Library	530	

2.2.2 Tangible compounds

Tangible compounds (Hann & Oprea, 2004) are virtual compounds that present a high synthetic feasibility using well-known or in-house building blocks and reactions. The Screenable Chemical Universe Based on Intuitive Data Organization (SCUBIDOO) (Chevillard & Kolb, 2015) tries to enhance the synthetic feasibility criterion. The SCUBIDOO database gathers more than 21 million virtual compounds generated by exhaustively reacting 7805 building blocks with the 58 most commonly used reactions in the pharmaceutical field. Commercial databases are also providing lists of tangible compounds. For example, Life Chemicals (n.d.) provides a database of more than 500,000 tangible compounds estimated to be able to be synthesized at 90% through in-house developed and validated synthetic procedures. The REAL Space of Enamine comprises more than 3.8 billion virtual compounds that should be successfully synthesized and delivered to the customer within approximatively 3 weeks with a probability over 80% (Hoffmann & Gastreich, 2019).

2.2.3 Possible compounds

Possible compounds are virtual compounds which structures are obtained using the rules of chemistry to guide the assembling of atoms through covalent bonds. The generated databases (GDBs) initiative (Blum & Reymond, 2009; Fink & Reymond, 2007; Reymond, 2015; Ruddigkeit et al., 2012) represents the largest collection of possible compounds. GDBs aims to enumerate all possible molecules from a chemical space defined by a set of rules among which are geometrical strain, functional group stability criteria, and simple

synthetic feasibility rules. The latest version, the GDB-17 (Ruddigkeit et al., 2012), includes 166 billion compounds with a maximum of 17 heavy atoms (C, N, O, S, F, Cl, Br, and I).

2.2.4 Virtual compounds and virtual screenings

These databases of virtual compounds represent a gold mine to discover new hits, especially for difficult biological targets. However, despite the recent advances toward ultralarge virtual screening using ligand-based methods (Boehm et al., 2008; Hoffmann & Gastreich, 2019; Lessel et al., 2009) and structure-based methods (Gorgulla et al., 2020; Lyu et al., 2019), screening of these enormous databases of virtual compounds using docking methods remains marginal and poorly accessible to numerous teams. The best option to explore these virtual compounds databases remains to select smaller but representative subsets. In this way, the GDBs provide smaller subsets of compounds with less chemical complexity and thus enhanced feasibility probability (Meier et al., 2020) and the SCUBIDOO includes three representative subsets named S, M, and L that contained, respectively, 9994, 99,977, and 999,794 compounds (Chevillard & Kolb, 2015). Finally, despite the efforts put to guarantee the synthetical feasibility of the virtual compounds, the compounds availability for experimental testing could be challenging.

2.3 Library Selection and Customization

As aforementioned, there is a vast number of diverse databases, and the tricky question of the one(s) to choose to construct an appropriate virtual screening database may arise. To guide this selection, two approaches are possible. The first one is to screen all possible compounds and libraries, as it has been shown that raising the number of compounds that are screened improves the true positive rates (Lyu et al., 2019). However, as previously mentioned with virtual compounds databases, ultralarge databases screening achieved using docking methods is still an exception and is limited by computing and storage capacities. The second choice is thus to select a subset of compounds among all available. The choice of the compounds to be prioritized in the database can be made based on the information available for the targeted protein. Focused libraries can be designed by filtering the database to enrich it with compounds that present structural and physicochemical similarity with known modulators of the query target or complementarity to the binding site of the query target. Another option is to promote diversity in the retained database by removing molecules that are too similar to those already included in the database (Yosipof & Senderowitz, 2014). This selection procedure is based on the assumption that similar compounds present similar biological activities. Thus, testing a single representative compound from a group of similar ones should allow insights on the potential activity of other compound members of the same cluster (Bayada et al., 2010).

Various clustering methods, divided into supervised and unsupervised approaches, are commonly used for data selection (Downs & Barnard, 2002; Shemetulskis et al., 1995). Clustering methods are efficient to reduce the database sizes and thus the associated docking computational times; however, a recent study demonstrated that electing a single representative of one cluster, regardless of the selection procedure, was negatively impacting the docking scores (Lyu et al., 2019). Dissimilarity-based methods can also be used to select the compounds to include in the databases (Hassan et al., 1996; Lajiness & Watson, 2008). These approaches include the application of distance metrics, such as the widely used Tanimoto coefficient on a set of appropriate descriptors (Bayada et al., 2010) and diversity algorithms among which are the maximum dissimilarity algorithms (Kennard & Stone, 1969), the Kohonen neural network (Li et al., 1993), or the sphere exclusion algorithms (Hudson et al., 1996).

3 DATABASE PREPARATION

Most of the databases presented in the previous section, with few exceptions such as the ZINC15, are not dedicated to virtual screening approaches and the chemical structures provided by these databases cannot directly be used as input for docking methods. A rigorous and careful step of ligand preparation is thus necessary to ensure the success of the virtual screening protocol. Even if there is no widely accepted protocol for virtual screening databases preparation, common guidelines are found in the scientific literature (Bologa et al., 2019; Fourches et al., 2010; Gally et al., 2019; Gorgulla et al., 2020; Lagorce, Bouslama, et al., 2017; Rognan & Bonnet, 2014; Williams et al., 2012). In this section, we will detail each step of the preparation process as seen in Fig. 5.1 and provide a summary available in Table 5.3 that lists commonly used tools for one, several, or all steps of the database preparation.

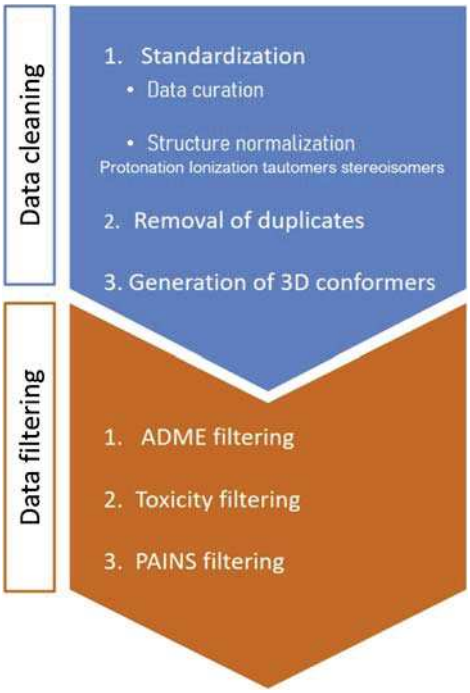


FIG. 5.1 Graphical representation of the main steps of database preparation. *ADME*, absorption, distribution, metabolism, elimination; *PAINS*, pan-assay interference compounds.

3.1 Database Cleaning

3.1.1 Standardization

3.1.1.1 Data curation. The first step of the cleaning procedure is the standardization of chemical structures, especially when the molecules are retrieved from different data sources. Compounds can be prepared with different protocols and encoded in different formats. Compact 1D and 2D formats, such as SMILES and SDF, are often privileged for large databases, but already prepared and cleaned structures in PDB or MOL2 formats can also be available. All compounds should thus be converted to the same format, and for SMILES inputs, canonical SMILES should be generated to avoid duplicates. Once the conversion is done, a major step is to identify and resolve issues such as correcting or filtering out compounds with incorrect or missing structures. Salted molecules are also problematic inputs for docking softwares and a neutralization step with elimination of the corresponding counterion is necessary. Most of the docking methods are only able to compute one molecule at a time, and mixture components should be separated. Finally, inorganics compounds, which usually represent a small fraction of drug discovery–oriented databases, should also be removed because they often cannot be handled by docking software (Fourches et al., 2010). All these steps can rapidly be achieved with different tools (Table 5.3).

3.1.1.2 Structure normalization (protonation, ionization, tautomerization, and stereochemistry). Compounds can present different protonation or ionization states and different tautomeric forms (tautomers) according to the physiological conditions. Depending on the state and form used, the possible interactions between the ligand and the binding site may vary and impact the docking outcomes. Experimentally determined protonation, ionization, and tautomeric states are rarely provided in the databases and are mainly determined through predictions (Ten Brink & Exner, 2009). Identification of the correct protonation state is crucial for docking calculations as the presence or the absence of a specific hydrogen may drastically influence both the sampling and the scoring part of the docking. Correct placement of these atoms is important to generate and identify the right docking poses (Polgár et al., 2007) and the compounds should be protonated at the physiological pH (Knox et al., 2005) with appropriate tools (Table 5.3). Tautomers are the result of a formal migration of a hydrogen atom or proton, accompanied by a switch of a single bond and an adjacent double bond. Consequently, properties such as hydrophobicity, pKa, 3D shapes, and ability to form hydrogen bonds can vary among the different tautomers of the same molecule. Careful selection of tautomers is thus crucial for accurate docking predictions (Martin, 2009) and several methods for the enumeration of tautomers have been published (Milletti et al., 2009; Oellien et al., 2006; Sitzmann et al., 2010; Trepalin et al., 2003). The level of enumeration that should be performed depends on the objective of the docking (Martin, 2009); a reasonable number of tautomers should be selected for a large virtual screening campaign, while a full enumeration of the tautomers is required for accurate prediction of one ligand/protein complex. In the same way, stereochemistry can impact the relative binding affinity. Proper enumeration of the relevant stereoisomers and pairs of enantiomers for chiral compounds is important for databases preparation (Brooks et al., 2008).

TABLE 5.3
Examples of Tools Available for Database Preparation.

Tool	Ref[1]	Type[2]	Cleaning										Filtering			
			Standardization	Inorganics	Neutralization	Mixtures	Checking structures	Duplicates	Tautomers	Stereoisomers	Protonation	Conformers	PhysChem	Toxicity	PAINS	BBB[3]
CDK	[1]	OS	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✓	✗	✗	✗
DataWarrior	[2]	F	✓	✓	✓	✓	✓	✓	✗	✓	✓[4]	✓	✓	✓	✗	✗
DockingServer	[3]	C	✓	✗	✗	✗	✗	✗	✗	✗	✓	✓	✗	✗	✗	✗
FAF-Drugs4	[4]	F	✓	✓	✓	✓	✓	✓	✗	✗	✓	✗	✓	✓	✓	✓
FILTER	[5]	C	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✓	✓	✗	✗
Indigo	[6]	OS	✓	✗	✗	✗	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗
Instant JChem	[7]	C(FA)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗

LigPrep	[8]	C	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗
LigQ	[9]	F	✓	✗	✗	✗	✗	✗	✓	✓	✓	✓	✗	✗	✗	✗
Open Babel	[10]	OS	✓	✗	✗	✗	✗	✓	✗	✓	✓	✓	✗	✗	✗	✗
RDKit	[11]	OS	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗
Screening Assistant	[12]	F	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✓	✓	✗
Standardizer	[13]	C(FA)	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✗	✗	✗	✗
SwissADME	[14]	F	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	✓	✓
UNICON	[15]	C(FA)	✓	✓	✗	✗	✓	✗	✓	✗	✓	✓	✗	✗	✗	✗
VFLP	[16]	F	✓	✗	✓	✗	✗	✗	✓		✓	✓	✗	✗	✗	✗
VSPred[5]	[17]	F	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✗

[1] References: [1] (Willighagen et al., 2017), [2] (Sander et al., 2015),[3] (Bikadi & Hazai, 2009), [4] (Lagorce, Oliveira, et al., 2017), [5] (FILTER, n.d.), [6] (Pavlov et al., 2011), [7] (Instant JChem, n.d.), [8] (LigPrep, n.d.), [9] (Radusky et al., 2017), [10] (OpenBabel, n.d.), [11] (RDKit, n.d.), [12] (Le Guilloux et al., 2012), [13] (Standardizer, n.d.), [14] (Daina et al., 2017), [15] (Sommer et al., 2016), [16] (Gorgulla et al., 2020), [17] (Gally et al., 2019).

[2] *C*, commercial; *C(FA)*, commercial but free for academics; *F*, freely accessible; *OS*, open source.

[3] BBB (blood–brain barrier) permeability filters.

[4] Only if the ChemAxon pKa Plugin is installed.

[5] Workflow using RDKit, ChemAxon, and Indigo features.

3.1.2 Removal of duplicates

Duplicated molecules should be removed to avoid unnecessary calculations. Various sources of duplicated molecules are possible, i.e., a same compound can be incorporated from different input databases (sometimes without the same identifier) or included in the database in different mixtures or salted forms. The standardization step is thus critical to highlight duplicates that were included in different forms and not recognized as duplicated.

3.1.3 Generation of 3D conformers

Most of the time, according to data storage consideration, compound structures are collected in a 2D format. A conversion step to generate 3D conformations is often required for docking algorithms. Because the bioactive conformation is generally unknown, a common strategy is to include a representative ensemble of conformers of each compound to avoid missing potential hits (Gimeno et al., 2019). This is of primary importance when running both rigid docking and flexible docking; in the former case, the ensemble of conformers should ideally cover the binding conformation(s) of the small molecules, and in the latter case, different conformations should be enumerated to cover the different conformational orientations of nonrotatable bonds and chemical groups (Hawkins, 2017). For example, the cyclohexane can adopt either a "chair" or a "boat" conformation, and both should be considered if no experimental data support the choice of one conformation over the other. The safe number of conformers to generate depends both on the tool used and on the number of rotatable bonds of the small molecules composing the database. Different commercial (e.g., OMEGA (Hawkins et al., 2010), ConfGen (Watts et al., 2010), or iCon (Poli et al., 2018)) or free (e.g., RDKIT (RDKit, n.d.), Frog2 (Miteva et al., 2010), or Tinker (Rackers et al., 2018)) ensemble generator tools are available. They are often included within software pipelines and are then not necessarily available as standalone tools. They fall into two categories: the stochastic approaches that sample random values of torsion angles with respect to predefined rules (e.g., molecular dynamics or Monte Carlo simulations or distance geometry) and the systematic approaches that sample all authorized torsion angles of a molecule (e.g., brute force and rule-based enumeration). Unlike the stochastic approaches, the systematic ones always return the same lowest-score conformation but are limited to molecules with few rotatable bonds because of the combinatorial explosion of solutions. Most tools make use of knowledge-based or force field–based scoring functions to estimate the internal energy of the generated conformers. In the latter case, the coulombic electrostatic term is generally omitted or tuned to avoid favoring conformers with intramolecular hydrogen bonds (Hawkins, 2017; Wang et al., 2020).

3.2 Database Filtering

Filtering procedures have been developed to enhance the chance that a hit identified by virtual screening will be successfully optimized as a potential drug candidate. A lot of ADME-Tox (absorption, distribution, metabolism, elimination, and toxicity) filters have been used in this purpose and will be presented here together with PAINS (pan-assay interference compounds) alerts and knowledge-based filters.

3.2.1 ADME (absorption, distribution, metabolism, elimination) filtering

In the mid-1990s, drug candidates' failure in clinical trials was mainly attributed to nonoptimal pharmacokinetics and bioavailability parameters (Kola & Landis, 2004). To overcome this issue, physicochemical filters were proposed to select compounds with pharmacokinetics values similar to actual approved drugs. The most used filters were established in 1997 by Lipinski and his colleagues from Pfizer (Lipinski et al., 1997). They selected in the World Drug Index 2245 orally available compounds that had reached at least phase II of clinical trials, hence displaying presumed adequate water solubility and intestinal permeability. The analysis of the common properties of these compounds led to the "Lipinski's Ro5," in which thresholds for molecular weight ($\leq$500 Da), number of hydrogen bonds donors ($\leq$5) and acceptors ($\leq$10), and water/octanol partition coefficient ($\leq$5) were defined. According to this rule, a compound that is not compliant to more than one of these criteria is associated with a higher risk of poor oral bioavailability.

The Ro5 has been largely used and integrated in numerous virtual screening databases preparation protocols. However, it is now of common knowledge that the Ro5 should not be considered as an absolute drug-likeness criterion because different toxic compounds are Ro5-compliant whereas some approved drugs are not (Bickerton et al., 2012) and that compounds compliant to this rule are not suitable to modulate half of the targets involved in human disease pathways (Surade & Blundell, 2012). Additionally, there is still some oral space beyond the Ro5 (Doak et al., 2014; Poongavanam et al., 2018; Tyagi et al., 2020). Virtual screening protocols aim to identify potential hits/leads for a biological target that will be modified

during the hit-to-lead and lead-to-candidate drug optimization processes. Applying drug-like filters on potential hits/leads seems untimely as the subsequent optimized drug candidates may not, at the end, be Ro5 compliant. Hann and Oprea suggested to adjust the Ro5 threshold to obtain "lead-like" filters deduced from the comparison of the properties of 176 leads compounds and 532 drugs (Alvarez & Shoichet, 2005). These "lead-like" filters presented not only modified threshold values compared to the Ro5 properties but also additional criteria such as the number of aromatic rings, Caco-2 intestinal permeability, and water solubility.

Drug permeability through the blood–brain barrier (BBB) is also a major property to monitor. Drugs targeting the central nervous system should be able to cross the BBB while peripheral acting drugs should not, in order to avoid any psychotropic side effect (Di & Kerns, 2015). The two gold-standard experimental measures of BBB permeability are logBB (the concentration of drug in the brain divided by concentration in the blood) and logPS (permeability surface area product). Both values can be obtained either through experimental measures or predicted via in silico methods (Carpenter et al., 2014). Another approach to estimate the BBB penetration is to use the BBB score (Gupta et al., 2019), a simple model that is based on five physicochemical descriptors.

Numerous methods to improve "drug-like," "lead-like," and other ADME filters have been developed (Korkmaz et al., 2015; Oprea et al., 2007; Petit et al., 2012; Ridder et al., 2011; Veber et al., 2002), but the interest for "drug-like" or "lead-like" filters has considerably decreased. The control of the physicochemical properties of the compounds during the optimization process remains important (Waring et al., 2015) and databases developed for docking purposes should at least be filtered using soft physicochemical thresholds to remove compounds not suitable for docking tools (Lagarde et al., 2018).

3.2.2 Toxicity filtering

Toxicity of drug candidates can be discovered at very late stages of drug development, during clinical trials or worse after its commercialization, ruining years of efforts and billions of dollars investments. Safety and toxicity have been identified as the major sources of failure in an analysis of attrition of drug candidates from four major pharmaceutical companies (Waring et al., 2015) and represent major concerns in drug discovery processes. Many computational toxicology methods have been developed to predict potential toxicity. These methods can be used as a prefiltering step to remove undesirable compounds from the virtual screening databases or after the virtual screening to flag predicted hits with toxicity alerts and to ensure that the flagged chemical moiety will be modified during the optimization process. Computational toxicology methods are mainly divided into three categories (Hevener, 2018): algorithms and models, chemical filters, and structural alerts tools.

In the first category, quantitative structure–activity relationship and quantitative structure–toxicity relationship models have long been the reference (Benigni & Bossa, 2008; Gini, 2016; Lapenna et al., 2010; Myshkin et al., 2012). These mathematical models can be established by correlating experimentally measured toxicity with the structure of the compounds encoded as physicochemical and geometrical descriptors. Other methods include deep learning methods (Gawehn et al., 2016; Goh et al., 2017), structure-based methods (Jing et al., 2015; Moroy et al., 2012), and toxicophore mapping (Kar & Roy, 2013; Pramanik & Roy, 2014; Singh et al., 2016).

The second category gathers a large variety of filters to identify promiscuous compounds, i.e., drug compounds that can act on multiple molecular targets, exhibiting similar or different pharmacological effects (Mei & Yang, 2018). Limited promiscuity might be a desirable property for polypharmacological drug discovery (Feldmann et al., 2019) but promiscuous compounds can also present undesirable side effect because of the modulation of unwanted targets (Mei & Yang, 2018). Filters are either rules built on physiochemical properties or structural alerts. Among the physiochemical properties evaluated (Bruns & Watson, 2012), specific attention was also given to molecular weight and lipophilicity (Bowes et al., 2012; Hughes et al., 2008; Morphy & Rankovic, 2007; Tarcsay & Keserű, 2013).

The last category includes structural alerts developed to highlight structural motifs and substructures often occurring among toxic compounds (Bruns & Watson, 2012; Pizzo et al., 2015).

A large number of freely available tools and web servers implement these computational toxicology method (ToxiPred (Mishra et al., 2014), DeepTox (Mayr et al., 2016), admetSAR (Cheng et al., 2012), ToxiM (Sharma et al., 2017), VirtualToxLab (Vedani et al., 2012), OpenTox (Tcheremenskaia et al., 2012), FAF-Drug4 (Lagorce, Bouslama, et al., 2017); ToxAlerts (Sushko et al., 2012), Screening Assistant (Le Guilloux et al., 2012)).

3.2.3 PAINS (pan-assay interference compounds) alerts

PAINS are compounds frequently identified as hits in any given assay due to the presence of common chemotypes able to interfere in biochemical assays (Baell & Nissink, 2018). PAINS behavior is linked to intrinsic redox activity, instability, reactivity, aggregation potency, covalent labeling of proteins, metal chelation, membrane disruption, fluorescence interference, and structural decomposition (Lagorce, Oliveira, et al., 2017). These compounds raised issues because numerous efforts to optimize PAINS identified as hits remained unsuccessful (Baell & Nissink, 2018). Several filters have been developed to flag PAINS (Baell & Holloway, 2010; Lagorce, Bouslama et al., 2017; Saubern et al., 2011; Sterling & Irwin, 2015) and were used to discard hits flagged with PAINS alerts prior to experimental validation. However, different studies point out that these filters are not perfect as 6%–7% of approved drugs display PAINS chemotypes (Capuzzi et al., 2017; Senger et al., 2016) and that some compounds flagged as PAINS were not confirmed by experimental assays, whereas compounds not flagged as PAINS revealed experimentally a PAINS behavior (Capuzzi et al., 2017). PAINS filters should then be used as flag alerts but not to remove compounds from virtual screening databases (Baell & Nissink, 2018; Lagorce, Oliveira, et al., 2017).

3.3 Automated Tools for Virtual Screening Databases Preparation

Numerous free or commercial methods are available for assisting the virtual screening databases preparation (Villoutreix et al., 2007). Table 5.3 lists the most popular tools with details about the corresponding monitored step of the database preparation. It is also worth mentioning that each step can be scripted by users with different programming languages. For example, the desalting procedure can be done with a simple text editing program by searching for "full stops" separating two fragments on SMILES and removing the one without any carbon atom. Finally, despite the precision of automated tools, a final manual inspection of each compound is recommended to ensure that no errors were left after the cleaning process (Fourches et al., 2010).

4 CONCLUSION

Docking methods are used in virtual screening protocols to identify hits that could be optimized into drug candidates. This can be successfully achieved by carefully selecting and preparing compounds that are included in the virtual screening compound collections. Numerous sources of compounds and databases are available, and we described the main categories in this chapter. Due to limited resources in computer power and data storage, ultralarge databases screening using docking methods has remained poorly accessible for a long time. However, recent advances could enhance its democratization within a few years. The compounds included in a virtual screening database must be correctly prepared for docking methods. This is particularly important to avoid missing potential hits and providing unnecessary efforts to optimize inactive, toxic, or promiscuous compounds that could have been identified beforehand. No current gold standard for database preparation exists, and different docking software require different molecular file formats. We detail in this chapter the main steps of compounds cleaning filtering. Filters may be applied to enrich the database in compounds that are less likely destined to fail in the drug discovery process and to reduce the virtual screening computational times. The strategy to select compounds and to filter the database before docking depends on the virtual screening purpose and on the importance of the virtual and experimental screening facilities.

LIST OF ABBREVIATIONS

1D	One-dimensional
2D	Two-dimensional
3D	Three-dimensional
ADME/Tox	Absorption, distribution, metabolism, elimination, toxicity
Å	Ångström
BBB	Blood–brain barrier
Br	Bromine
C	Carbon
ChEMBL	Chemical database of the European Molecular Biology Laboratory
Cl	Chlorine
clogP	Partition coefficient between n-octanol and water
Da	Dalton
F	Fluorine
FDB-17	Fragment database 17
GDBs	Generated databases
GLASS	GPCR-ligand associations
GLL	G-protein–coupled receptor (GPCR) ligand library
I	Iodine
logBB	Logarithm of the ratio of the concentration of drug in the brain divided by concentration of the drug in the blood
logPS	Permeability surface area product

mg	Milligram
N	Nitrogen
NCBI	National Center for Biotechnology Information
NR-DBIND	Nuclear Receptors Database Including Negative Data
NRLiSt BDB	Nuclear Receptors Ligands and Structures Benchmarking Database
O	Oxygen
PAINS	Pan-assay interference compounds
PDB	Protein Data Bank
pH	Potential of hydrogen
pKa	Acid dissociation constant
PKIDB	Protein Kinase Inhibitor Database
S	Sulfur
SCUBIDOO	Screenable Chemical Universe Based on Intuitive Data Organization
SDF	Structure data file
SMILES	Simplified Molecular Input Line Entry Specification
TCM DB	Traditional Chinese Medicine Database
US	United States of America

REFERENCES

Alvarez, J., & Shoichet, B. (2005). *Virtual screening in drug discovery*. CRC Press.

Arús-Pous, J., Awale, M., Probst, D., & Reymond, J.-L. (2019). Exploring chemical space with machine learning. *Chimia*, *73*(12), 1018–1023. https://doi.org/10.2533/chimia.2019.1018.

Awale, M., Visini, R., Probst, D., Arús-Pous, J., & Reymond, J.-L. (2017). Chemical space: Big data challenge for molecular diversity. *Chimia*, *71*(10), 661–666. https://doi.org/10.2533/chimia.2017.661.

Baell, J. B., & Holloway, G. A. (2010). New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *Journal of Medicinal Chemistry*, *53*(7), 2719–2740. https://doi.org/10.1021/jm901137j.

Baell, J. B., & Nissink, J. W. M. (2018). Seven year itch: pan-assay interference compounds (PAINS) in 2017—utility and limitations. *ACS Chemical Biology*, *13*(1), 36–44. https://doi.org/10.1021/acschembio.7b00903.

Banerjee, P., Erehman, J., Gohlke, B.-O., Wilhelm, T., Preissner, R., & Dunkel, M. (2015). Super natural II—a database of natural products. *Nucleic Acids Research*, *43*(D1), D935–D939. https://doi.org/10.1093/nar/gku886.

Bayada, D. M., Hamersma, H., & van Geerestein, V. J. (2010). ChemInform abstract: Molecular diversity and representativity in chemical databases. *ChemInform*, *30*(17). https://doi.org/10.1002/chin.199917285.

Benigni, R., & Bossa, C. (2008). Predictivity and reliability of QSAR models: The case of mutagens and carcinogens. *Toxicology Mechanisms and Methods*, *18*(2–3), 137–147. https://doi.org/10.1080/15376510701857056.

Bickerton, G. R., Paolini, G. V., Besnard, J., Muresan, S., & Hopkins, A. L. (2012). Quantifying the chemical beauty of drugs. *Nature Chemistry*, *4*(2), 90–98. https://doi.org/10.1038/nchem.1243.

Bikadi, Z., & Hazai, E. (2009). Application of the PM6 semi-empirical method to modeling proteins enhances docking accuracy of AutoDock. *Journal of Cheminformatics*, *1*(1), 15. https://doi.org/10.1186/1758-2946-1-15.

Blum, L. C., & Reymond, J.-L. (2009). 970 million druglike small molecules for virtual screening in the chemical universe database GDB-13. *Journal of the American Chemical Society*, *131*(25), 8732–8733. https://doi.org/10.1021/ja902302h.

Boehm, M., Wu, T.-Y., Claussen, H., & Lemmen, C. (2008). Similarity searching and Scaffold Hopping in synthetically accessible combinatorial chemistry spaces. *Journal of Medicinal Chemistry*, *51*(8), 2468–2480. https://doi.org/10.1021/jm0707727.

Bologa, C. G., Ursu, O., & Oprea, T. I. (2019). How to prepare a compound collection prior to virtual screening. In R. S. Larson, & T. I. Oprea (Eds.), *Bioinformatics and drug discovery* (Vol. 1939, pp. 119–138). Springer New York. https://doi.org/10.1007/978-1-4939-9089-4_7.

Böttcher, J., Dilworth, D., Reiser, U., Neumüller, R. A., Schleicher, M., Petronczki, M., Zeeb, M., Mischerikow, N., Allali-Hassani, A., Szewczyk, M. M., Li, F., Kennedy, S., Vedadi, M., Barsyte-Lovejoy, D., Brown, P. J., Huber, K. V. M., Rogers, C. M., Wells, C. I., Fedorov, O., … McConnell, D. B. (2019). Fragment-based discovery of a chemical probe for the PWWP1 domain of NSD3. *Nature Chemical Biology*, *15*(8), 822–829. https://doi.org/10.1038/s41589-019-0310-x.

Bowes, J., Brown, A. J., Hamon, J., Jarolimek, W., Sridhar, A., Waldron, G., & Whitebread, S. (2012). Reducing safety-related drug attrition: The use of in vitro pharmacological profiling. *Nature Reviews Drug Discovery*, *11*(12), 909–922. https://doi.org/10.1038/nrd3845.

Ten Brink, T., & Exner, T. E. (2009). Influence of protonation, tautomeric, and stereoisomeric states on Protein–Ligand docking results. *Journal of Chemical Information and Modeling*, *49*(6), 1535–1546. https://doi.org/10.1021/ci800420z.

Brooks, W. H., Daniel, K. G., Sung, S.-S., & Guida, W. C. (2008). Computational validation of the importance of absolute stereochemistry in virtual screening. *Journal of Chemical Information and Modeling*, *48*(3), 639–645. https://doi.org/10.1021/ci700358r.

Bruns, R. F., & Watson, I. A. (2012). Rules for identifying potentially reactive or promiscuous compounds. *Journal of Medicinal Chemistry*, *55*(22), 9763–9772. https://doi.org/10.1021/jm301008n.

Capuzzi, S. J., Muratov, E. N., & Tropsha, A. (2017). Phantom PAINS: Problems with the utility of alerts for Pan-Assay INterference compoundS. *Journal of Chemical Information and Modeling*, *57*(3), 417–427. https://doi.org/10.1021/acs.jcim.6b00465.

Carles, F., Bourg, S., Meyer, C., & Bonnet, P. (2018). PKIDB: A curated, annotated and updated database of protein kinase inhibitors in clinical trials. *Molecules (Basel, Switzerland)*, *23*(4). https://doi.org/10.3390/molecules23040908.

Carpenter, T. S., Kirshner, D. A., Lau, E. Y., Wong, S. E., Nilmeier, J. P., & Lightstone, F. C. (2014). A method to predict blood-brain barrier permeability of drug-like compounds using molecular dynamics simulations. *Biophysical Journal*, *107*(3), 630–641. https://doi.org/10.1016/j.bpj.2014.06.024.

Chan, W. K. B., Zhang, H., Yang, J., Brender, J. R., Hur, J., Özgür, A., & Zhang, Y. (2015). GLASS: A comprehensive database for experimentally validated GPCR-ligand associations. *Bioinformatics (Oxford, England)*, *31*(18), 3035–3042. https://doi.org/10.1093/bioinformatics/btv302.

Chen, C. Y.-C. (2011). TCM Database@Taiwan: The world's largest traditional Chinese medicine database for drug screening in silico. *PLoS One*, *6*(1), e15939. https://doi.org/10.1371/journal.pone.0015939.

Chen, Y., de Bruyn Kops, C., & Kirchmair, J. (2017). Data resources for the computer-guided discovery of bioactive natural products. *Journal of Chemical Information and Modeling*, *57*(9), 2099–2111. https://doi.org/10.1021/acs.jcim.7b00341.

Cheng, F., Li, W., Zhou, Y., Shen, J., Wu, Z., Liu, G., Lee, P. W., & Tang, Y. (2012). admetSAR: A comprehensive source and free tool for assessment of chemical ADMET properties. *Journal of Chemical Information and Modeling*, *52*(11), 3099–3105. https://doi.org/10.1021/ci300367a.

Chevillard, F., & Kolb, P. (2015). SCUBIDOO: A large yet screenable and easily searchable database of computationally created chemical compounds optimized toward high likelihood of synthetic tractability. *Journal of Chemical Information and Modeling*, *55*(9), 1824–1835. https://doi.org/10.1021/acs.jcim.5b00203.

Congreve, M., Carr, R., Murray, C., & Jhoti, H. (2003). A 'Rule of Three' for fragment-based lead discovery? *Drug Discovery Today*, *8*(19), 876–877. https://doi.org/10.1016/S1359-6446(03)02831-9.

Corbeil, C. R., Williams, C. I., & Labute, P. (2012). Variability in docking success rates due to dataset preparation. *Journal of Computer-Aided Molecular Design*, *26*(6), 775–786. https://doi.org/10.1007/s10822-012-9570-1.

Daina, A., Michielin, O., & Zoete, V. (2017). SwissADME: A free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. *Scientific Reports*, *7*(1), 42717. https://doi.org/10.1038/srep42717.

Delavan, B., Roberts, R., Huang, R., Bao, W., Tong, W., & Liu, Z. (2018). Computational drug repositioning for rare diseases in the era of precision medicine. *Drug Discovery Today*, *23*(2), 382–394. https://doi.org/10.1016/j.drudis.2017.10.009.

Di, L., & Kerns, E. H. (2015). *Blood-brain barrier in drug discovery: Optimizing brain exposure of CNS drugs and minimizing brain side effects for peripheral drugs*. Wiley.

Doak, B. C., Over, B., Giordanetto, F., & Kihlberg, J. (2014). Oral druggable space beyond the rule of 5: Insights from drugs and clinical candidates. *Chemistry and Biology*, *21*(9), 1115–1142. https://doi.org/10.1016/j.chembiol.2014.08.013.

Downs, G. M., & Barnard, J. M. (2002). Clustering methods and their uses in computational chemistry. In K. B. Lipkowitz, & D. B. Boyd (Eds.), *Reviews in computational chemistry* (Vol. 18, pp. 1–40). John Wiley & Sons, Inc. https://doi.org/10.1002/0471433519.ch1.

Erlanson, D. A., Fesik, S. W., Hubbard, R. E., Jahnke, W., & Jhoti, H. (2016). Twenty years on: The impact of fragments on drug discovery. *Nature Reviews Drug Discovery*, *15*(9), 605–619. https://doi.org/10.1038/nrd.2016.109.

Feldmann, C., Miljković, F., Yonchev, D., & Bajorath, J. (2019). Identifying promiscuous compounds with activity against different target classes. *Molecules*, *24*(22), 4185. https://doi.org/10.3390/molecules24224185.

FILTER. (n.d.). OpenEye Scientific Software. Retrieved July 27, 2020, from http://www.eyesopen.com.

Fink, T., & Reymond, J.-L. (2007). Virtual exploration of the chemical Universe up to 11 atoms of C, N, O, F: Assembly of 26.4 million structures (110.9 million stereoisomers) and analysis for new ring systems, stereochemistry, physicochemical properties, compound classes, and drug discovery. *Journal of Chemical Information and Modeling*, *47*(2), 342–353. https://doi.org/10.1021/ci600423u.

Forli, S. (2015). Charting a path to success in virtual screening. *Molecules (Basel, Switzerland)*, *20*(10), 18732–18758. https://doi.org/10.3390/molecules201018732.

Fourches, D., Muratov, E., & Tropsha, A. (2010). Trust, but verify: On the importance of chemical structure curation in cheminformatics and QSAR modeling research. *Journal of Chemical Information and Modeling*, *50*(7), 1189–1204. https://doi.org/10.1021/ci100176x.

Gally, J.-M., Bourg, S., Fogha, J., Do, Q.-T., Aci-Sèche, S., & Bonnet, P. (2019). VSPrep: A KNIME workflow for the preparation of molecular databases for virtual screening. *Current Medicinal Chemistry*. https://doi.org/10.2174/0929867326666190614160451.

Gatica, E. A., & Cavasotto, C. N. (2012). Ligand and decoy sets for docking to G protein-coupled receptors. *Journal of Chemical Information and Modeling*, *52*(1), 1–6. https://doi.org/10.1021/ci200412p.

Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B., & Overington, J. P. (2012). ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic Acids Research*, *40*(Database issue), D1100–D1107. https://doi.org/10.1093/nar/gkr777.

Gaulton, A., Hersey, A., Nowotka, M., Bento, A. P., Chambers, J., Mendez, D., Mutowo, P., Atkinson, F., Bellis, L. J., Cibrián-Uhalte, E., Davies, M., Dedman, N., Karlsson, A., Magariños, M. P., Overington, J. P., Papadatos, G., Smit, I., & Leach, A. R. (2017). The ChEMBL database in 2017. *Nucleic Acids Research*, *45*(D1), D945–D954. https://doi.org/10.1093/nar/gkw1074.

Gawehn, E., Hiss, J. A., & Schneider, G. (2016). Deep learning in drug discovery. *Molecular Informatics, 35*(1), 3–14. https://doi.org/10.1002/minf.201501008.

Gimeno, A., Ojeda-Montes, M., Tomás-Hernández, S., Cereto-Massagué, A., Beltrán-Debón, R., Mulero, M., Pujadas, G., & Garcia-Vallvé, S. (2019). The light and dark sides of virtual screening: What is there to know? *International Journal of Molecular Sciences, 20*(6), 1375. https://doi.org/10.3390/ijms20061375.

Gini, G. (2016). QSAR methods. In E. Benfenati (Ed.), *In silico methods for predicting drug toxicity* (Vol. 1425, pp. 1–20). Springer New York. https://doi.org/10.1007/978-1-4939-3609-0_1.

Goh, G. B., Hodas, N. O., & Vishnu, A. (2017). Deep learning for computational chemistry. *Journal of Computational Chemistry, 38*(16), 1291–1307. https://doi.org/10.1002/jcc.24764.

Gorgulla, C., Boeszoermenyi, A., Wang, Z.-F., Fischer, P. D., Coote, P. W., Padmanabha Das, K. M., Malets, Y. S., Radchenko, D. S., Moroz, Y. S., Scott, D. A., Fackeldey, K., Hoffmann, M., Iavniuk, I., Wagner, G., & Arthanari, H. (2020). An open-source drug discovery platform enables ultra-large virtual screens. *Nature, 580*(7805), 663–668. https://doi.org/10.1038/s41586-020-2117-z.

Gupta, M., Lee, H. J., Barden, C. J., & Weaver, D. F. (2019). The blood–brain barrier (BBB) score. *Journal of Medicinal Chemistry, 62*(21), 9824–9836. https://doi.org/10.1021/acs.jmedchem.9b01220.

Hann, M. M., & Oprea, T. I. (2004). Pursuing the leadlikeness concept in pharmaceutical research. *Current Opinion in Chemical Biology, 8*(3), 255–263. https://doi.org/10.1016/j.cbpa.2004.04.003.

Hassan, M., Bielawski, J. P., Hempel, J. C., & Waldman, M. (1996). Optimization and visualization of molecular diversity of combinatorial libraries. *Molecular Diversity, 2*(1–2), 64–74. https://doi.org/10.1007/BF01718702.

Hawkins, P. C. D. (2017). Conformation generation: The state of the art. *Journal of Chemical Information and Modeling, 57*(8), 1747–1756. https://doi.org/10.1021/acs.jcim.7b00221.

Hawkins, P. C. D., Skillman, A. G., Warren, G. L., Ellingson, B. A., & Stahl, M. T. (2010). Conformer generation with OMEGA: Algorithm and validation using high quality structures from the protein databank and Cambridge structural database. *Journal of Chemical Information and Modeling, 50*(4), 572–584. https://doi.org/10.1021/ci100031x.

Hevener, K. E. (2018). Computational toxicology methods in chemical library design and high-throughput screening hit validation. In O. Nicolotti (Ed.), *Computational toxicology* (Vol. 1800, pp. 275–285). Springer New York. https://doi.org/10.1007/978-1-4939-7899-1_13.

Hoffmann, T., & Gastreich, M. (2019). The next level in chemical space navigation: Going far beyond enumerable compound libraries. *Drug Discovery Today, 24*(5), 1148–1156. https://doi.org/10.1016/j.drudis.2019.02.013.

Horvath, D., Lisurek, M., Rupp, B., Kühne, R., Specker, E., von Kries, J., Rognan, D., Andersson, C. D., Almqvist, F., Elofsson, M., Enqvist, P.-A., Gustavsson, A.-L., Remez, N., Mestres, J., Marcou, G., Varnek, A., Hibert, M., Quintana, J., & Frank, R. (2014). Design of a general-purpose European compound screening library for EU-OPENSCREEN. *ChemMedChem, 9*(10), 2309–2326. https://doi.org/10.1002/cmdc.201402126.

Huang, N., Shoichet, B. K., & Irwin, J. J. (2006). Benchmarking sets for molecular docking. *Journal of Medicinal Chemistry, 49*(23), 6789–6801. https://doi.org/10.1021/jm0608356.

Hudson, B. D., Hyde, R. M., Rahr, E., Wood, J., & Osman, J. (1996). Parameter based methods for compound selection from chemical databases. *Quantitative Structure-Activity Relationships, 15*(4), 285–289. https://doi.org/10.1002/qsar.19960150402.

Hughes, J. D., Blagg, J., Price, D. A., Bailey, S., DeCrescenzo, G. A., Devraj, R. V., Ellsworth, E., Fobian, Y. M., Gibbs, M. E., Gilles, R. W., Greene, N., Huang, E., Krieger-Burke, T., Loesel, J., Wager, T., Whiteley, L., & Zhang, Y. (2008). Physiochemical drug properties associated with in vivo toxicological outcomes. *Bioorganic and Medicinal Chemistry Letters, 18*(17), 4872–4875. https://doi.org/10.1016/j.bmcl.2008.07.071.

Instant JChem. ChemAxon. Retrieved July 31, 2020, from https://chemaxon.com/products/instant-jchem.

Irwin, J. J. (2008). Using ZINC to acquire a virtual screening library. *Current Protocols in Bioinformatics*. https://doi.org/10.1002/0471250953.bi1406s22 (Chapter 14), Unit 14.6.

Irwin, J. J., & Shoichet, B. K. (2005). ZINC—a free database of commercially available compounds for virtual screening. *Journal of Chemical Information and Modeling, 45*(1), 177–182. https://doi.org/10.1021/ci049714+.

Jing, Y., Easter, A., Peters, D., Kim, N., & Enyedy, I. J. (2015). In silico prediction of hERG inhibition. *Future Medicinal Chemistry, 7*(5), 571–586. https://doi.org/10.4155/fmc.15.18.

Kar, S., & Roy, K. (2013). First report on predictive chemometric modeling, 3D-toxicophore mapping and in silico screening of in vitro basal cytotoxicity of diverse organic chemicals. *Toxicology in Vitro, 27*(2), 597–608. https://doi.org/10.1016/j.tiv.2012.10.015.

Kennard, R. W., & Stone, L. A. (1969). Computer aided design of experiments. *Technometrics, 11*(1), 137–148. https://doi.org/10.1080/00401706.1969.10490666.

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B. A., Thiessen, P. A., Yu, B., Zaslavsky, L., Zhang, J., & Bolton, E. E. (2019). PubChem 2019 update: Improved access to chemical data. *Nucleic Acids Research, 47*(D1), D1102–D1109. https://doi.org/10.1093/nar/gky1033.

Kim, S., Thiessen, P. A., Bolton, E. E., Chen, J., Fu, G., Gindulyte, A., Han, L., He, J., He, S., Shoemaker, B. A., Wang, J., Yu, B., Zhang, J., & Bryant, S. H. (2016). PubChem substance and compound databases. *Nucleic Acids Research, 44*(D1), D1202–D1213. https://doi.org/10.1093/nar/gkv951.

Knox, A. J. S., Meegan, M. J., Carta, G., & Lloyd, D. G. (2005). Considerations in compound database preparation"Hidden" impact on virtual screening results. *Journal of Chemical Information and Modeling, 45*(6), 1908–1919. https://doi.org/10.1021/ci050185z.

Kola, I., & Landis, J. (2004). Can the pharmaceutical industry reduce attrition rates? *Nature Reviews Drug Discovery, 3*(8), 711–716. https://doi.org/10.1038/nrd1470.

Kontijevskis, A. (2017). Mapping of drug-like chemical Universe with reduced complexity molecular frameworks. *Journal of Chemical Information and Modeling, 57*(4), 680–699. https://doi.org/10.1021/acs.jcim.7b00006.

Kontoyianni, M. (2017). Docking and virtual screening in drug discovery. *Methods in Molecular Biology (Clifton, N.J.), 1647*, 255–266. https://doi.org/10.1007/978-1-4939-7201-2_18.

Korkmaz, S., Zararsiz, G., & Goksuluk, D. (2015). MLViS: A web tool for machine learning-based virtual screening in early-phase of drug discovery and development. *PLoS One, 10*(4), e0124600. https://doi.org/10.1371/journal.pone.0124600.

Lagarde, N., Ben Nasr, N., Jérémie, A., Guillemain, H., Laville, V., Labib, T., Zagury, J.-F., & Montes, M. (2014). NRLiSt BDB, the manually curated nuclear receptors ligands and structures benchmarking database. *Journal of Medicinal Chemistry, 57*(7), 3117–3125. https://doi.org/10.1021/jm500132p.

Lagarde, N., Rey, J., Gyulkhandanyan, A., Tufféry, P., Miteva, M. A., & Villoutreix, B. O. (2018). Online structure-based screening of purchasable approved drugs and natural compounds: Retrospective examples of drug repositioning on cancer targets. *Oncotarget, 9*(64), 32346–32361. https://doi.org/10.18632/oncotarget.25966.

Lagarde, N., Zagury, J.-F., & Montes, M. (2015). Benchmarking data sets for the evaluation of virtual ligand screening methods: Review and perspectives. *Journal of Chemical Information and Modeling, 55*(7), 1297–1307. https://doi.org/10.1021/acs.jcim.5b00090.

Lagorce, D., Bouslama, L., Becot, J., Miteva, M. A., & Villoutreix, B. O. (2017). FAF-Drugs4: Free ADME-tox filtering computations for chemical biology and early stages drug discovery. *Bioinformatics, 33*(22), 3658–3660. https://doi.org/10.1093/bioinformatics/btx491.

Lagorce, D., Oliveira, N., Miteva, M. A., & Villoutreix, B. O. (2017). Pan-assay interference compounds (PAINS) that may not be too painful for chemical biology projects. *Drug Discovery Today, 22*(8), 1131–1133. https://doi.org/10.1016/j.drudis.2017.05.017.

Lajiness, M., & Watson, I. (2008). Dissimilarity-based approaches to compound acquisition. *Current Opinion in Chemical Biology, 12*(3), 366–371. https://doi.org/10.1016/j.cbpa.2008.03.010.

Lapenna, S., Fuart Gatnik, M., & Worth, A. (2010). *Review of QSAR models and software tools for predicting acute and chronic systemic toxicity*. Publications Office of the European Union. https://op.europa.eu/en/publication-detail/-/publication/940acf32-3e4d-47cf-b4a1-eebe67c79ae1/language-en.

Le Guilloux, V., Arrault, A., Colliandre, L., Bourg, S., Vayer, P., & Morin-Allory, L. (2012). Mining collections of compounds with screening assistant 2. *Journal of Cheminformatics, 4*(1), 20. https://doi.org/10.1186/1758-2946-4-20.

Lessel, U., Wellenzohn, B., Lilienthal, M., & Claussen, H. (2009). Searching fragment spaces with feature trees. *Journal of Chemical Information and Modeling, 49*(2), 270–279. https://doi.org/10.1021/ci800272a.

Li, X., Gasteiger, J., & Zupan, J. (1993). On the topology distortion in self-organizing feature maps. *Biological Cybernetics, 70*(2), 189–198. https://doi.org/10.1007/BF00200832.

Life Chemicals (n.d.). Leading supplier of HTS compounds, building blocks. Retrieved July 25, 2020, from https://lifechemicals.com/.

LigPrep (Version Schrödinger Release 2020-3). (n.d.). Schrödinger, LLC, New York, NY, 2020.

Lipinski, C. A., Lombardo, F., Dominy, B. W., & Feeney, P. J. (1997). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews, 23*(1–3), 3–25. https://doi.org/10.1016/S0169-409X(96)00423-1.

Lyu, J., Wang, S., Balius, T. E., Singh, I., Levit, A., Moroz, Y. S., O'Meara, M. J., Che, T., Algaa, E., Tolmachova, K., Tolmachev, A. A., Shoichet, B. K., Roth, B. L., & Irwin, J. J. (2019). Ultra-large library docking for discovering new chemotypes. *Nature, 566*(7743), 224–229. https://doi.org/10.1038/s41586-019-0917-9.

Martin, Y. C. (2009). Let's not forget tautomers. *Journal of Computer-Aided Molecular Design, 23*(10), 693–704. https://doi.org/10.1007/s10822-009-9303-2.

Mayr, A., Klambauer, G., Unterthiner, T., & Hochreiter, S. (2016). DeepTox: Toxicity prediction using deep learning. *Frontiers in Environmental Science, 3*. https://doi.org/10.3389/fenvs.2015.00080.

Meier, K., Bühlmann, S., Arús-Pous, J., & Reymond, J.-L. (2020). The generated databases (GDBs) as a source of 3D-shaped building blocks for use in medicinal chemistry and drug discovery. *CHIMIA International Journal for Chemistry, 74*(4), 241–246. https://doi.org/10.2533/chimia.2020.241.

Mei, Y., & Yang, B. (2018). Rational application of drug promiscuity in medicinal chemistry. *Future Medicinal Chemistry, 10*(15), 1835–1851. https://doi.org/10.4155/fmc-2018-0018.

Messick, T. E., Smith, G. R., Soldan, S. S., McDonnell, M. E., Deakyne, J. S., Malecka, K. A., Tolvinski, L., van den Heuvel, A. P. J., Gu, B.-W., Cassel, J. A., Tran, D. H., Wassermann, B. R., Zhang, Y., Velvadapu, V., Zartler, E. R., Busson, P., Reitz, A. B., & Lieberman, P. M. (2019). Structure-based design of small-molecule inhibitors of EBNA1 DNA binding blocks Epstein-Barr virus latent infection and tumor growth. *Science Translational Medicine, 11*(482), eaau5612. https://doi.org/10.1126/scitranslmed.aau5612.

Milletti, F., Storchi, L., Sforna, G., Cross, S., & Cruciani, G. (2009). Tautomer enumeration and stability prediction for virtual screening on large chemical databases. *Journal of Chemical Information and Modeling, 49*(1), 68–75. https://doi.org/10.1021/ci800340j.

Mishra, N. K., Singla, D., Agarwal, S., & Raghava, G. P. S. (2014). ToxiPred: A server for prediction of aqueous toxicity of small chemical molecules in *T. Pyriformis*. *Journal of Translational Toxicology, 1*(1), 21–27. https://doi.org/10.1166/jtt.2014.1005.

Miteva, M. A., Guyon, F., & Tufféry, P. (2010). Frog2: Efficient 3D conformation ensemble generator for small compounds. *Nucleic Acids Research, 38*(Web Server issue), W622–W627. https://doi.org/10.1093/nar/gkq325.

Moroy, G., Martiny, V. Y., Vayer, P., Villoutreix, B. O., & Miteva, M. A. (2012). Toward in silico structure-based ADMET prediction in drug discovery. *Drug Discovery Today, 17*(1–2), 44–55. https://doi.org/10.1016/j.drudis.2011.10.023.

Morphy, R., & Rankovic, Z. (2007). Fragments, network biology and designing multiple ligands. *Drug Discovery Today, 12*(3–4), 156–160. https://doi.org/10.1016/j.drudis.2006.12.006.

Morrison, K. C., & Hergenrother, P. J. (2014). Natural products as starting points for the synthesis of complex and diverse compounds. *Natural Product Reports, 31*(1), 6–14. https://doi.org/10.1039/c3np70063a.

Myshkin, E., Brennan, R., Khasanova, T., Sitnik, T., Serebriyskaya, T., Litvinova, E., Guryanov, A., Nikolsky, Y., Nikolskaya, T., & Bureeva, S. (2012). Prediction of organ toxicity endpoints by QSAR modeling based on precise chemical-histopathology annotations: Prediction of organ toxicity endpoints by QSAR modeling. *Chemical Biology and Drug Design, 80*(3), 406–416. https://doi.org/10.1111/j.1747-0285.2012.01411.x.

Mysinger, M. M., Carchia, M., Irwin, J. J., & Shoichet, B. K. (2012). Directory of useful decoys, enhanced (DUD-E): Better ligands and decoys for better benchmarking. *Journal of Medicinal Chemistry, 55*(14), 6582–6594. https://doi.org/10.1021/jm300687e.

National Center for Biotechnology Information (US). (2010). *Probe reports from the NIH molecular libraries program*.

Newman, D. J., & Cragg, G. M. (2020). Natural products as sources of new drugs over the nearly four decades from 01/1981 to 09/2019. *Journal of Natural Products, 83*(3), 770–803. https://doi.org/10.1021/acs.jnatprod.9b01285.

Ntie-Kang, F., Zofou, D., Babiaka, S. B., Meudom, R., Scharfe, M., Lifongo, L. L., Mbah, J. A., Mbaze, L. M., Sippl, W., & Efange, S. M. N. (2013). AfroDb: A select highly potent and diverse natural product library from African medicinal plants. *PLoS One, 8*(10), e78085. https://doi.org/10.1371/journal.pone.0078085.

Oellien, F., Cramer, J., Beyer, C., Ihlenfeldt, W.-D., & Selzer, P. M. (2006). The impact of tautomer forms on pharmacophore-based virtual screening. *Journal of Chemical Information and Modeling, 46*(6), 2342–2354. https://doi.org/10.1021/ci060109b.

OpenBabel. (n.d.). Retrieved July 27, 2020, from http://openbabel.org/wiki/Main_Page.

Oprea, T. I., Allu, T. K., Fara, D. C., Rad, R. F., Ostopovici, L., & Bologa, C. G. (2007). Lead-like, drug-like or "Pub-like": How different are they? *Journal of Computer-Aided Molecular Design, 21*(1–3), 113–119. https://doi.org/10.1007/s10822-007-9105-3.

Pavlov, D., Rybalkin, M., Karulin, B., Kozhevnikov, M., Savelyev, A., & Churinov, A. (2011). Indigo: Universal cheminformatics API. *Journal of Cheminformatics, 3*(S1), P4. https://doi.org/10.1186/1758-2946-3-S1-P4.

Petit, J., Meurice, N., Kaiser, C., & Maggiora, G. (2012). Softening the rule of five—where to draw the line? *Bioorganic and Medicinal Chemistry, 20*(18), 5343–5351. https://doi.org/10.1016/j.bmc.2011.11.064.

Pizzo, F., Gadaleta, D., Lombardo, A., Nicolotti, O., & Benfenati, E. (2015). Identification of structural alerts for liver and kidney toxicity using repeated dose toxicity data. *Chemistry Central Journal, 9*(1), 62. https://doi.org/10.1186/s13065-015-0139-7.

Polgár, T., Magyar, C., Simon, I., & Keserü, G. M. (2007). Impact of ligand protonation on virtual screening against β-secretase (BACE1). *Journal of Chemical Information and Modeling, 47*(6), 2366–2373. https://doi.org/10.1021/ci700223p.

Poli, G., Seidel, T., & Langer, T. (2018). Conformational sampling of small molecules with iCon: Performance assessment in comparison with OMEGA. *Frontiers in Chemistry, 6*, 229. https://doi.org/10.3389/fchem.2018.00229.

Poongavanam, V., Doak, B. C., & Kihlberg, J. (2018). Opportunities and guidelines for discovery of orally absorbed drugs in beyond rule of 5 space. *Current Opinion in Chemical Biology, 44*, 23–29. https://doi.org/10.1016/j.cbpa.2018.05.010.

Pramanik, S., & Roy, K. (2014). Exploring QSTR modeling and toxicophore mapping for identification of important molecular features contributing to the chemical toxicity in *Escherichia coli*. *Toxicology in Vitro, 28*(2), 265–272. https://doi.org/10.1016/j.tiv.2013.11.002.

Pushpakom, S., Iorio, F., Eyers, P. A., Escott, K. J., Hopper, S., Wells, A., Doig, A., Guilliams, T., Latimer, J., McNamee, C., Norris, A., Sanseau, P., Cavalla, D., & Pirmohamed, M. (2019). Drug repurposing: Progress, challenges and recommendations. *Nature Reviews Drug Discovery, 18*(1), 41–58. https://doi.org/10.1038/nrd.2018.168.

Rackers, J. A., Wang, Z., Lu, C., Laury, M. L., Lagardère, L., Schnieders, M. J., Piquemal, J.-P., Ren, P., & Ponder, J. W. (2018). Tinker 8: Software tools for molecular design. *Journal of Chemical Theory and Computation, 14*(10), 5273–5289. https://doi.org/10.1021/acs.jctc.8b00529.

Radusky, L., Ruiz-Carmona, S., Modenutti, C., Barril, X., Turjanski, A. G., & Martí, M. A. (2017). LigQ: A webserver to select and prepare ligands for virtual screening. *Journal of Chemical Information and Modeling, 57*(8), 1741–1746. https://doi.org/10.1021/acs.jcim.7b00241.

RDKit. (n.d.). Retrieved July 27, 2020, from https://www.rdkit.org/.

Réau, M., Lagarde, N., Zagury, J.-F., & Montes, M. (2019). Nuclear receptors database including negative data (NR-DBIND): A database dedicated to nuclear receptors binding data including negative data and pharmacological profile. *Journal of Medicinal Chemistry, 62*(6), 2894−2904. https://doi.org/10.1021/acs.jmedchem.8b01105.

Réau, M., Langenfeld, F., Zagury, J.-F., Lagarde, N., & Montes, M. (2018). Decoys selection in benchmarking datasets: Overview and perspectives. *Frontiers in Pharmacology, 9*, 11. https://doi.org/10.3389/fphar.2018.00011.

Reymond, J.-L. (2015). The chemical space project. *Accounts of Chemical Research, 48*(3), 722−730. https://doi.org/10.1021/ar500432k.

Ridder, L., Wang, H., de Vlieg, J., & Wagener, M. (2011). Revisiting the rule of five on the basis of pharmacokinetic data from rat. *ChemMedChem, 6*(11), 1967−1970. https://doi.org/10.1002/cmdc.201100306.

Rodrigues, T., Reker, D., Schneider, P., & Schneider, G. (2016). Counting on natural products for drug design. *Nature Chemistry, 8*(6), 531−541. https://doi.org/10.1038/nchem.2479.

Rognan, D., & Bonnet, P. (2014). Chemical databases and virtual screening. *Medecine Sciences: M/S, 30*(12), 1152−1160. https://doi.org/10.1051/medsci/20143012019.

Ruddigkeit, L., van Deursen, R., Blum, L. C., & Reymond, J.-L. (2012). Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *Journal of Chemical Information and Modeling, 52*(11), 2864−2875. https://doi.org/10.1021/ci300415d.

Sander, T., Freyss, J., von Korff, M., & Rufener, C. (2015). DataWarrior: An open-source program for chemistry aware data visualization and analysis. *Journal of Chemical Information and Modeling, 55*(2), 460−473. https://doi.org/10.1021/ci500588j.

Saubern, S., Guha, R., & Baell, J. B. (2011). KNIME workflow to assess PAINS filters in SMARTS format. Comparison of RDKit and Indigo cheminformatics libraries. *Molecular Informatics, 30*(10), 847−850. https://doi.org/10.1002/minf.201100076.

Senger, M. R., Fraga, C. A. M., Dantas, R. F., & Silva, F. P. (2016). Filtering promiscuous compounds in early drug discovery: Is it a good idea? *Drug Discovery Today, 21*(6), 868−872. https://doi.org/10.1016/j.drudis.2016.02.004.

Serafin, M. B., Bottega, A., Foletto, V. S., da Rosa, T. F., Hörner, A., & Hörner, R. (2020). Drug repositioning is an alternative for the treatment of coronavirus COVID-19. *International Journal of Antimicrobial Agents, 55*(6), 105969. https://doi.org/10.1016/j.ijantimicag.2020.105969.

Sharma, A. K., Srivastava, G. N., Roy, A., & Sharma, V. K. (2017). ToxiM: A toxicity prediction tool for small molecules developed using machine learning and chemoinformatics approaches. *Frontiers in Pharmacology, 8*, 880. https://doi.org/10.3389/fphar.2017.00880.

Shemetulskis, N. E., Dunbar, J. B., Dunbar, B. W., Moreland, D. W., & Humblet, C. (1995). Enhancing the diversity of a corporate database using chemical database clustering and analysis. *Journal of Computer-Aided Molecular Design, 9*(5), 407−416. https://doi.org/10.1007/BF00123998.

Shi, Y., & von Itzstein, M. (2019). How size matters: Diversity for fragment library design. *Molecules, 24*(15), 2838. https://doi.org/10.3390/molecules24152838.

Singh, P. K., Negi, A., Gupta, P. K., Chauhan, M., & Kumar, R. (2016). Toxicophore exploration as a screening technology for drug design and discovery: Techniques, scope and limitations. *Archives of Toxicology, 90*(8), 1785−1802. https://doi.org/10.1007/s00204-015-1587-5.

Siramshetty, V. B., Eckert, O. A., Gohlke, B.-O., Goede, A., Chen, Q., Devarakonda, P., Preissner, S., & Preissner, R. (2018). SuperDRUG2: A one stop resource for approved/marketed drugs. *Nucleic Acids Research, 46*(D1), D1137−D1143. https://doi.org/10.1093/nar/gkx1088.

Sitzmann, M., Ihlenfeldt, W.-D., & Nicklaus, M. C. (2010). Tautomerism in large databases. *Journal of Computer-Aided Molecular Design, 24*(6−7), 521−551. https://doi.org/10.1007/s10822-010-9346-4.

Sommer, K., Friedrich, N.-O., Bietz, S., Hilbig, M., Inhester, T., & Rarey, M. (2016). UNICON: A powerful and easy-to-use compound library converter. *Journal of Chemical Information and Modeling, 56*(6), 1105−1111. https://doi.org/10.1021/acs.jcim.6b00069.

Sorokina, M., & Steinbeck, C. (2020). Review on natural products databases: Where to find data in 2020. *Journal of Cheminformatics, 12*(1), 20. https://doi.org/10.1186/s13321-020-00424-9.

Standardizer. (n.d.). ChemAxon. https://chemaxon.com/products/chemical-structure-representation-toolkit.

Sterling, T., & Irwin, J. J. (2015). Zinc 15—ligand discovery for everyone. *Journal of Chemical Information and Modeling, 55*(11), 2324−2337. https://doi.org/10.1021/acs.jcim.5b00559.

Surade, S., & Blundell, T. L. (2012). Structural biology and drug discovery of difficult targets: The limits of ligandability. *Chemistry and Biology, 19*(1), 42−50. https://doi.org/10.1016/j.chembiol.2011.12.013.

Sushko, I., Salmina, E., Potemkin, V. A., Poda, G., & Tetko, I. V. (2012). ToxAlerts: A web server of structural alerts for toxic chemicals and compounds with potential adverse reactions. *Journal of Chemical Information and Modeling, 52*(8), 2310−2316. https://doi.org/10.1021/ci300245q.

Tarcsay, Á., & Keserű, G. M. (2013). Contributions of molecular properties to drug promiscuity: Miniperspective. *Journal of Medicinal Chemistry, 56*(5), 1789−1795. https://doi.org/10.1021/jm301514n.

Tcheremenskaia, O., Benigni, R., Nikolova, I., Jeliazkova, N., Escher, S. E., Batke, M., Baier, T., Poroikov, V., Lagunin, A., Rautenberg, M., & Hardy, B. (2012). OpenTox predictive toxicology framework: Toxicological ontology and semantic media wiki-based OpenToxipedia. *Journal of Biomedical Semantics, 3*(Suppl. 1), S7. https://doi.org/10.1186/2041-1480-3-S1-S7.

Trepalin, S. V., Skorenko, A. V., Balakin, K. V., Nasonov, A. F., Lang, S. A., Ivashchenko, A. A., & Savchuk, N. P. (2003). Advanced exact structure searching in large databases of chemical compounds. *Journal of Chemical Information and Computer Sciences*, *43*(3), 852–860. https://doi.org/10.1021/ci025582d.

Tyagi, M., Begnini, F., Poongavanam, V., Doak, B. C., & Kihlberg, J. (2020). Drug syntheses beyond the rule of 5. *Chemistry - A European Journal*, *26*(1), 49–88. https://doi.org/10.1002/chem.201902716.

Ursu, O., Holmes, J., Bologa, C. G., Yang, J. J., Mathias, S. L., Stathias, V., Nguyen, D.-T., Schürer, S., & Oprea, T. (2019). DrugCentral 2018: An update. *Nucleic Acids Research*, *47*(D1), D963–D970. https://doi.org/10.1093/nar/gky963.

Veber, D. F., Johnson, S. R., Cheng, H.-Y., Smith, B. R., Ward, K. W., & Kopple, K. D. (2002). Molecular properties that influence the oral bioavailability of drug candidates. *Journal of Medicinal Chemistry*, *45*(12), 2615–2623. https://doi.org/10.1021/jm020017n.

Vedani, A., Dobler, M., & Smieško, M. (2012). VirtualToxLab—a platform for estimating the toxic potential of drugs, chemicals and natural products. *Toxicology and Applied Pharmacology*, *261*(2), 142–153. https://doi.org/10.1016/j.taap.2012.03.018.

Villoutreix, B. O., Renault, N., Lagorce, D., Montes, M., & Miteva, M. A. (2007). Free resources to assist structure-based virtual ligand screening experiments. *Current Protein and Peptide Science*, *8*(4), 381–411. https://doi.org/10.2174/138920307781369391.

Visini, R., Awale, M., & Reymond, J.-L. (2017). Fragment database FDB-17. *Journal of Chemical Information and Modeling*, *57*(4), 700–709. https://doi.org/10.1021/acs.jcim.7b00020.

Vogt, M. (2020). How do we optimize chemical space navigation? *Expert Opinion on Drug Discovery*, *15*(5), 523–525. https://doi.org/10.1080/17460441.2020.1730324.

Wang, S., Witek, J., Landrum, G. A., & Riniker, S. (2020). Improving conformer generation for small rings and macrocycles based on distance geometry and experimental torsional-angle preferences. *Journal of Chemical Information and Modeling*, *60*(4), 2044–2058. https://doi.org/10.1021/acs.jcim.0c00025.

Waring, M. J., Arrowsmith, J., Leach, A. R., Leeson, P. D., Mandrell, S., Owen, R. M., Pairaudeau, G., Pennie, W. D., Pickett, S. D., Wang, J., Wallace, O., & Weir, A. (2015). An analysis of the attrition of drug candidates from four major pharmaceutical companies. *Nature Reviews Drug Discovery*, *14*(7), 475–486. https://doi.org/10.1038/nrd4609.

Watts, K. S., Dalal, P., Murphy, R. B., Sherman, W., Friesner, R. A., & Shelley, J. C. (2010). ConfGen: A conformational search method for efficient generation of bioactive conformers. *Journal of Chemical Information and Modeling*, *50*(4), 534–546. https://doi.org/10.1021/ci100015j.

Williams, A. J., Ekins, S., & Tkachenko, V. (2012). Towards a gold standard: Regarding quality in public domain chemistry databases and approaches to improving the situation. *Drug Discovery Today*, *17*(13–14), 685–701. https://doi.org/10.1016/j.drudis.2012.02.013.

Willighagen, E. L., Mayfield, J. W., Alvarsson, J., Berg, A., Carlsson, L., Jeliazkova, N., Kuhn, S., Pluskal, T., Rojas-Chertó, M., Spjuth, O., Torrance, G., Evelo, C. T., Guha, R., & Steinbeck, C. (2017). The Chemistry Development Kit (CDK) v2.0: Atom typing, depiction, molecular formulas, and substructure searching. *Journal of Cheminformatics*, *9*(1), 33. https://doi.org/10.1186/s13321-017-0220-4.

Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., Assempour, N., Iynkkaran, I., Liu, Y., Maciejewski, A., Gale, N., Wilson, A., Chin, L., Cummings, R., Le, D., … Wilson, M. (2018). DrugBank 5.0: A major update to the DrugBank database for 2018. *Nucleic Acids Research*, *46*(D1), D1074–D1082. https://doi.org/10.1093/nar/gkx1037.

Würth, R., Thellung, S., Bajetto, A., Mazzanti, M., Florio, T., & Barbieri, F. (2016). Drug-repositioning opportunities for cancer therapy: Novel molecular targets for known compounds. *Drug Discovery Today*, *21*(1), 190–199. https://doi.org/10.1016/j.drudis.2015.09.017.

Yosipof, A., & Senderowitz, H. (2014). Optimization of molecular representativeness. *Journal of Chemical Information and Modeling*, *54*(6), 1567–1577. https://doi.org/10.1021/ci400715n.

1.2.3. Discussion

L'objectif de cet article est de fournir un guide pour les utilisateurs sur la manière d'appréhender les bases de données disponibles pour une utilisation dans le cadre d'un projet de *drug design*. Ainsi nous décomposons la démarche en deux étapes : (1) l'état de l'art des chimiothèques existantes et (2) détailler un protocole pour la sélection et la préparation des composés inclus dans la base de données.

1.2.3.1. Etat de l'art des chimiothèques

Par l'énumération des différentes chimiothèques et leurs classifications, nous cherchions à citer quelques exemples populaires et pertinents à utiliser et non à être exhaustif sur le sujet. Nous avons choisi dans cet article de grouper les chimiothèques présentées en chimiothèques «existantes » et chimiothèques virtuelles. Dans la première catégorie nous avons fait la distinction entre les chimiothèques commerciales, les chimiothèques à bioactivités et celles de composés naturels. Pour les chimiothèques virtuelles, nous avons distingué entre les collections de fragments, les composés tangibles et les composés possibles ou « synthétisables ». Il existe toutefois différentes autres manières de catégoriser les bases de données, par exemple en fonction des cibles thérapeutiques. En effet, les bases de données peuvent être généralistes ou dédiées à une famille de cibles particulière. Les composés inclus dans de telles bases de données sont des composés testés expérimentalement (*in vitro* ou *in vivo*) pour une cible protéique. Nous pouvons par exemple citer la NR-DBIND[216] et la NRLiSt BDB [326] toutes les deux développées au laboratoire GBCM et dédiées au récepteurs nucléaires, la base de données GLASS[327] spécifique de récepteurs couplés à la protéine G (RCPG) ou encore la base de données KLIF[328] dédiée aux protéines kinases. Le travail réalisé lors de cette thèse portant sur une thématique plutôt orientée toxicologie, il aurait pu être intéressant d'aborder aussi les librairies de criblage disponibles pour cette application. Nous pensons ici aux librairies de métabolites ou encore aux librairies biologiques répertoriant des toxiques ou des composés supposés toxiques pour l'Homme et l'environnement [4,16,329] Cependant, afin de rester dans le périmètre des thématiques définies pour ce livre, nous ne les avons pas mentionnées.

1.2.3.2. Protocole de préparation des bases de données de criblage

Le protocole présenté dans ce chapitre ne prend en considération que la préparation d'un point de vue structural. Ainsi, la sélection de composés est focalisée sur le principe de similarité et/ou

diversité, et la standardisation est faite dans un but d'élimination des doublons et de préparation des formats utilisables par les logiciels de criblage. Toutefois, nous n'avons pas abordé ces mêmes points de sélection et de standardisation en fonction de l'activité biologique des composés. Ceci est très important lorsque la collection de composés est préparée à partir de bases de données incluant des données de bioactivités telles que la ChEMBL ou d'autres répertoires et sources incluant les résultats expérimentaux (HTS) des composés chimiques. En effet, la qualité des données biologiques est primordiale pour assurer la pertinence des modèles *in silico* [330] et les erreurs qui peuvent exister sont nombreuses comme par exemple des différences d'activité dues au variabilité expérimentales inter et intra-laboratoires [331]. Cette thématique, bien que non abordé dans ce chapitre de livre, a toutefois était traitée par Fourches et al [330] qui proposent un guide pratique pour préparer les données issues de répertoires chemogénomiques.

1.2.4. Conclusion et perspectives

Ce chapitre constitue un guide pratique destiné à une communauté plus ou moins novice désirant entamer un projet de criblage virtuel à l'aide de méthodes de docking dans le cadre d'un projet de *drug design*. L'article énumère quelques librairies virtuelles existantes et présente un protocole simple à adopter pour préparer la base de données de criblage. Bien que ce protocole ait été mentionné dans le cadre d'un criblage SB, il est important de mentionner que ces étapes sont similaires à celles que nous conseillerons de suive pour la construction de modèles LB. Nous avons d'ailleurs appliqué un protocole similaire à celui présenté dans ce chapitre afin de construire les bases de données d'entraînement et de test pour construire nos modèles dédiés aux NR en excluant la partie « filtrage » des données.

2. Prédiction de la capacité des composés chimiques à se lier aux NR : Application aux Perturbateurs endocriniens

Il est important de pouvoir identifier les perturbateurs endocriniens surtout pour des secteurs comme l'industrie cosmétique, pharmaceutique et agro-alimentaire. Cela permet d'implémenter des mesures de gestion du risque et donc de protéger l'homme et l'environnement d'éventuelles expositions [332]. Même si leur dangerosité n'est plus à prouver, il n'existe malheureusement aucune procédure expérimentale standard qui permet de prédire avec confiance si un composé est un PE [245] en raison de leur diversité et de leurs mécanismes d'actions variables et pas encore complétement élucidés (c.f. Paragraphe I.4. Mécanismes d'action de PE). Ainsi parmi les nombreux mécanismes, il est admis que certains PE agissent par liaison directe au site d'action des hormones, perturbant ainsi la liaison du ligand endogène et modulant la biologie du système endocrinien. Ce mécanisme d'action est très facile à modéliser *in vitro* via les tests biologiques de *binding* (c.f. paragraphe I.4.5 Caractérisation *in vitro* des PE). Ainsi de nombreuses instances ont mis au point un programme de criblage *in vitro* afin de tester massivement les composés suspectés d'être des PE et de partager les retombées [329,333]. Notons aussi de nombreux autres tests biologiques permettent d'étudier d'autres mécanismes d'actions mais ces derniers ne seront pas étudiés pour cette thèse.

En raison des limitations de temps et de coût, les tests biologiques ne peuvent être utilisés pour tester rapidement tous les produits chimiques retrouvés dans les écosystèmes. De plus, de nouveaux composés sont synthétisés quotidiennement, se rajoutant à la liste déjà existante des composés qu'il faudra tester pour garantir leur innocuité. Les approches de modélisation *in silico* nous permettent de surmonter ce problème [334]. Ces dernières sont en effet rapides, peu coûteuses et permettent de tester aussi bien des composés existants qu'hypothétiques [245]. Les modèles *in silico* sont construits à partir des données biologiques disponibles et de nombreuses méthodes peuvent être employées pour le faire (c.f. paragraphe 2. Evaluation des risques en toxicologie). Parmi ces méthodes nous emploierons pour ce travail les méthodes de criblage virtuel dont les méthodes LB comme les modèles de pharmacophores et les méthodes SB comme le docking.

Cette partie de la section Résultats présente les résultats obtenus pour l'étape de modélisation réalisée sur les récepteurs nucléaires. Elle est décomposée en deux. Tout d'abord, une preuve de concept a été faite afin de mettre au point le protocole. Ensuite, la deuxième partie des travaux de thèse a consisté à l'application de ce dernier pour d'autres récepteurs nucléaires impliqués dans le mécanisme de perturbation endocrinienne.

2.1. Preuve de concept : Prédiction de potentiel perturbateurs endocriniens agissant su ERα

2.1.1. Introduction

La preuve de concept de la capacité des méthodes *in silico* à prédire l'interaction de potentiels perturbateurs endocriniens avec les NR a été réalisée en utilisant des données relatives au récepteur aux œstrogènes alpha (ERα). Le choix s'est porté sur ce récepteur en raison de l'abondance de donnés existantes, en partie explicable par le fait que cette protéine a été le premier NR à avoir été cloné en 1962 [312]. En effet, la vaste majorité des études menées sur les PE étaient concentrées surtout sur ce récepteur [335] comme le montre la **Figure 24** issues de la revue présentée précédemment. Aujourd'hui on compte plus de 600 structures de ERα répertoriées dans la PDB[90] dont 329 humaines. Lors de précédents travaux au laboratoire [216,326], 33 structures humaines holo sans mutation ont été référencées. Ces structures montrent le récepteur en complexe avec des ligands agonistes, des ligands antagonistes et ou d'autres modulateurs (SERM). Au-delà des nombreuses études menées pour découvrir de nouveaux candidats thérapeutiques pour ERα [336,337], des études relatives à l'influence des PE et leur mécanisme d'action sur ERα ont aussi été menées [335]. Il a ainsi été démontré que les PE peuvent agir directement au niveau de la poche de liaison (Ligand Binding Pocket : LBP) de l'estrogène (E2), le ligand endogène des récepteurs ER. Parmi ces PE, nous pouvons citer le Bisphénol-A, les phytoœstrogènes, et autres pesticides organochlorés [335].

Pour réaliser cette preuve de concept, nous avons utilisé des données issues de l'EPA regroupant plusieurs initiatives de criblage à haut débit (HTS) de composés toxiques ou suspectés toxiques pour l'Homme et l'environnement. Le choix de cette base de données a été motivé par les multitudes études répertoriées dans la littérature scientifique qui l'utilisaient (c.f. partie 2.1 revue) mais aussi par le fait qu'elle inclue des données expérimentales aussi bien positives/actives que négatives/inactives. Ainsi, en exploitant les résultats de HTS obtenus par des tests de binding (affinité de liaison) contre ERα, nous avons pu construire un protocole combinant des méthodes de criblage virtuel SB et LB.

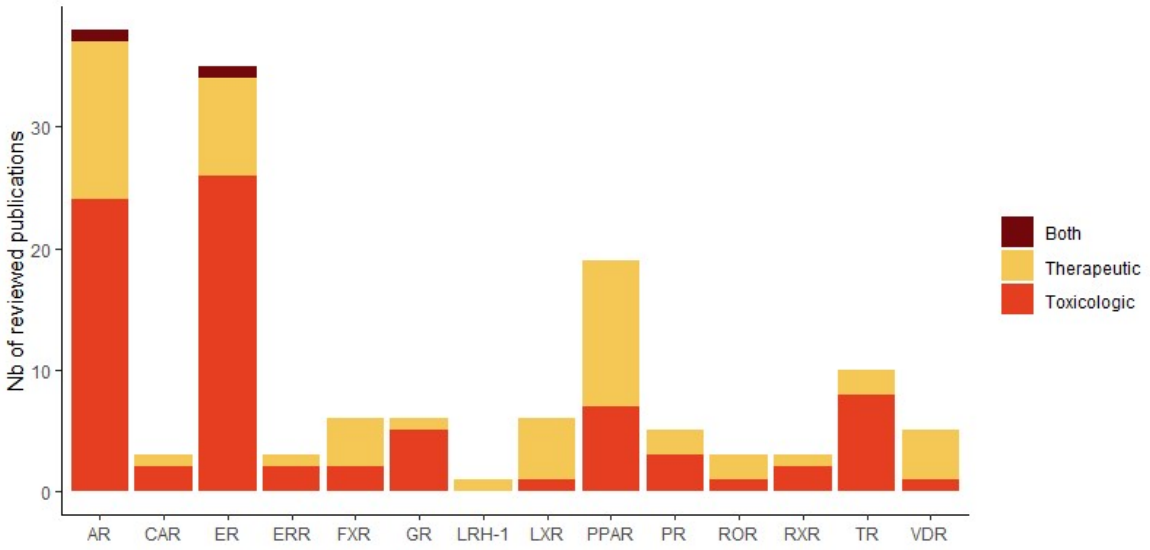


Figure 24 : Distribution du nombre de publication analysées entre les différents NR et le contexte toxicologique ou thérapeutique

2.1.2. Publication

International Journal of
Molecular Sciences

MDPI

Article

Predicting Potential Endocrine Disrupting Chemicals Binding to Estrogen Receptor α (ERα) Using a Pipeline Combining Structure-Based and Ligand-Based in Silico Methods

Asma Sellami, Matthieu Montes *,† and Nathalie Lagarde *,†

Laboratoire GBCM, EA 7528, Conservatoire National des Arts et Métiers, Hésam Université, 2 rue Conté, F-75003 Paris, France; asma.sellami@lecnam.net

* Correspondence: matthieu.montes@cnam.fr (M.M.); nathalie.lagarde@lecnam.net (N.L.)

† These authors contributed equally to this work.

Abstract: The estrogen receptors α (ERα) are transcription factors involved in several physiological processes belonging to the nuclear receptors (NRs) protein family. Besides the endogenous ligands, several other chemicals are able to bind to those receptors. Among them are endocrine disrupting chemicals (EDCs) that can trigger toxicological pathways. Many studies have focused on predicting EDCs based on their ability to bind NRs; mainly, estrogen receptors (ER), thyroid hormones receptors (TR), androgen receptors (AR), glucocorticoid receptors (GR), and peroxisome proliferator-activated receptors gamma (PPARγ). In this work, we suggest a pipeline designed for the prediction of ERα binding activity. The flagged compounds can be further explored using experimental techniques to assess their potential to be EDCs. The pipeline is a combination of structure based (docking and pharmacophore models) and ligand based (pharmacophore models) methods. The models have been constructed using the Environmental Protection Agency (EPA) data encompassing a large number of structurally diverse compounds. A validation step was then achieved using two external databases: the NR-DBIND (Nuclear Receptors DataBase Including Negative Data) and the EADB (Estrogenic Activity DataBase). Different combination protocols were explored. Results showed that the combination of models performed better than each model taken individually. The consensus protocol that reached values of 0.81 and 0.54 for sensitivity and specificity, respectively, was the best suited for our toxicological study. Insights and recommendations were drawn to alleviate the screening quality of other projects focusing on ERα binding predictions.

Keywords: nuclear receptors; ERα; endocrine disrupting chemicals; docking; pharmacophores; virtual screening

check for updates

Citation: Sellami, A.; Montes, M.; Lagarde, N. Predicting Potential Endocrine Disrupting Chemicals Binding to Estrogen Receptor α (ERα) Using a Pipeline Combining Structure-Based and Ligand-Based in Silico Methods. *Int. J. Mol. Sci.* **2021**, *22*, 2846. https://doi.org/10.3390/ijms22062846

Academic Editor: Akira Sugawara

Received: 27 January 2021
Accepted: 8 March 2021
Published: 11 March 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Estrogens are hormones involved in many physiological processes such as growth, development, the female reproductive system, and homeostasis [1]. They can exert their activity through binding to particular transcription factors: the estrogen receptors (ER). As members of the nuclear receptor protein family (NRs), ER are composed of three functional domains, the NH2-terminal domain (NTD), the DNA-binding domain (DBD), and the COOH-terminal ligand-binding domain (LBD) [2]. Two isoforms of the receptor exist, ERα and ERβ. Both isoforms share a high degree of sequence identity within their LBDs and exhibit similar affinities for the main endogenous ligand, 17β-estradiol [3], but different affinities for other compounds, given that each subtype displays a unique role in estrogenic activity in vivo. Since its discovery [4], several therapeutic applications have emerged for ERα ligands, in particular in breast cancer therapies [5,6]. Consequently, a large number of small molecules were developed with the purpose of ERα activity modulation. However, some compounds belonging to a particular category of exogenous molecules called endocrine disrupting chemicals (EDCs) are also able to bind to ERα [7].

EDCs have the ability to penetrate the body through ingestion, inhalation, or skin and to mimic the endogenous hormones, leading to the disruption of the endocrine system in both human and animal species. The first reported EDCs harmful effects were related to estrogens [8] such as breast cancer, endometriosis, fertility problems, and learning disability. EDCs are now considered a public health threat [9–11], as human exposure to these compounds can increase the risk of impairment of several biological functions such as the reproductive [12], cognitive [13], and metabolic [14] functions (for a review of associations between EDC exposures and risk to diseases, see Table 1 in [15]). However, the knowledge about possible adverse effects of EDCs is still incomplete and numerous studies have focused on better understanding their mechanism of action.

EDCs have been shown to act through direct or indirect mechanisms. In the direct mechanism, EDCs directly bind to a receptor of the NRs family (estrogen receptors ER, thyroid hormones receptors TR, androgen receptors AR, glucocorticoid receptors GR, and peroxisome proliferator-activated receptors gamma PPARγ) or the aryl hydrocarbon receptor, leading to activation or inhibition of its signaling pathway. In the indirect mechanism, EDCs affect other transcription factors or hormone metabolism through interaction with components of the hormone signaling pathway, stimulation or inhibition of endogenous hormones biosynthesis, binding to circulating hormone-binding protein, stimulation or inhibition of hormone-binding protein synthesis or degradation, stimulation or inhibition of hormone receptor expression [16,17]. Other potential targets of EDCs include the membrane-associated NRs and the G protein-coupled receptor GPR30/G protein-coupled estrogen receptor [18]. Experimental campaigns are conducted to identify potential EDCs and better understand their mechanism of action.

With the large and increasing number of compounds suspected to be EDCs, an intermediate step is needed to prioritize or reduce the number of compounds to be assessed. Several in silico methods are providing prediction and estimation of the potential endocrine disrupting activity of chemicals [19–22]. The majority of the in silico studies dedicated to EDCs focused on the direct mechanism. These studies are dedicated to NR binding prediction and most studies available are related to ERα [21–25]. These studies considered that a compound predicted to be able to bind to ERα can be a potential EDC that should be further investigated experimentally. In silico predictions of EDCs are mostly done through QSAR models and machine learning methods that provide a quantitative estimation of the binding affinity or a classification of the potential hazard. Docking methods are also used but to a lesser extent despite the advantage of providing insights on the molecular mechanism of binding [19].

In the present work, we designed a pipeline for the prediction of compounds binding to ERα. These flagged compounds can be further explored using experimental techniques to assess their potential to be EDCs. This pipeline combines structure-based (SB) and ligand-based (LB) methods, i.e., docking, SB, and LB pharmacophore models. To select the optimal docking protocol for ERα binding (B) compounds prediction, the performance of different docking software was evaluated and docking scores thresholds were defined. A combination of 26 pharmacophore models was designed to guarantee a maximum coverage of the chemical space of ERα B compounds. Individual performances of LB and SB models to discriminate between B and non-binding (NB) compounds were evaluated. Finally, different combination approaches were also explored to define the best protocol for the prediction of ERα binding potential. We conclude our work with recommendations for future ERα (and other NRs) binding prediction studies.

2. Results

2.1. Compounds and Database Preparation

2.1.1. Database Preparation

After filtering and cleaning, the Environmental Protection Agency (EPA) database is a collection of 2442 chemical compounds experimentally tested for ERα binding comprising 2219 non-binding (NB) compounds and 223 binding (B) compounds (see Material and

Methods section). The distribution of the physiochemical and constitutional descriptors of B and NB compounds is represented in Figure 1.

Two external validation sets were used, i.e., the NR-DBIND (Nuclear Receptors DataBase Including Negative Data) ERα set that comprises 732 compounds, divided into 554 B compounds and 178 NB compounds, and the EADB (Estrogenic Activity DataBase) set comprising 131 B compounds and 101 NB compounds for a total of 232 molecules. Distributions of the 15 constitutional, physiochemical, and molecular descriptors for each dataset are presented in Supplementary Figures S1 and S2.

2.1.2. Databases Comparison

Pairwise similarities were calculated using the Tanimoto coefficient (Tc) between each pair of topological fingerprints for: 1.) the EPA database and the NR-DBIND and 2.) the EPA and the EADB (see Figure S3A). The analysis of similarity values shows that the Tc are globally very low with a mean of 0.181 for the pairing with NR-DBIND and 0.174 for the pairing with EADB. Only 2% and 0.6% of the total calculated Tc for EADB and NR-DBIND, respectively (as shown in Figure S3B), are higher than 0.5. Finally, the chemical space of the three databases was mapped using a SALI (Structure Activity Landscape Index) map for the whole databases (Figure 2). The map illustrates that all three databases share the same chemical space.

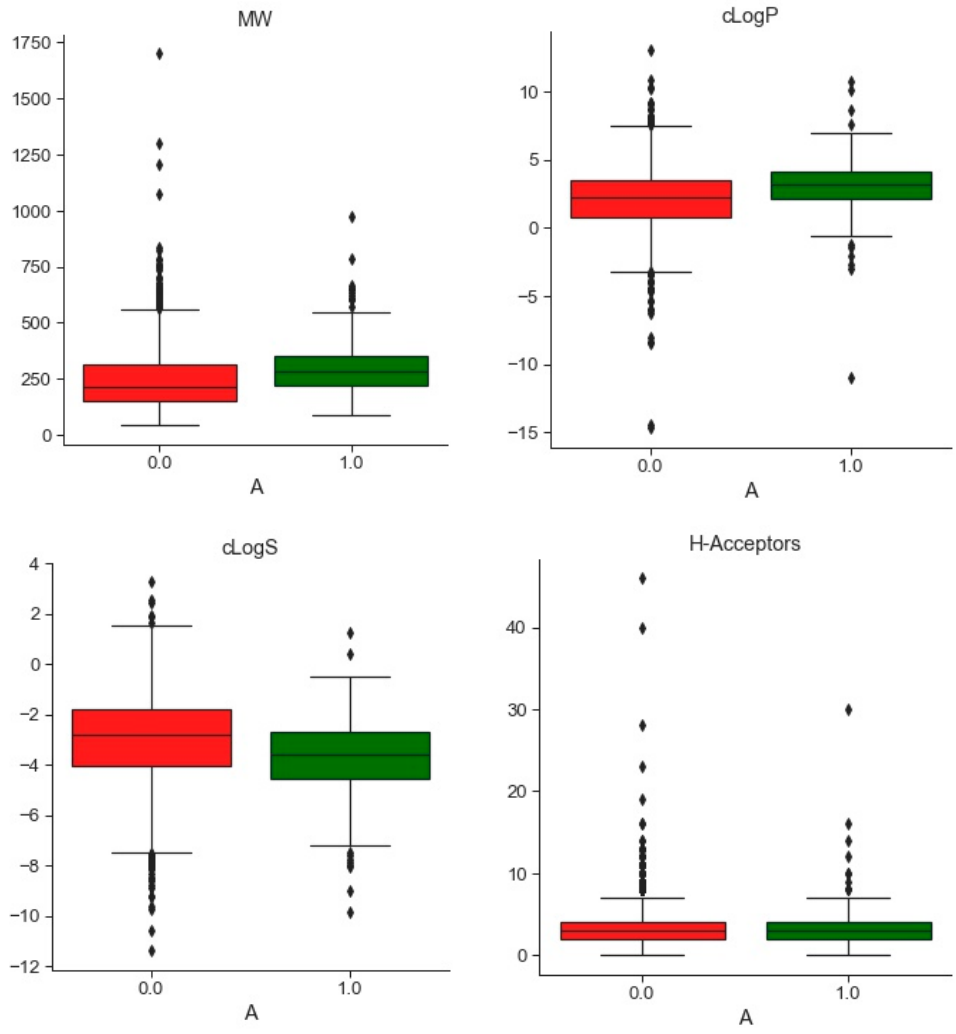


Figure 1. *Cont.*

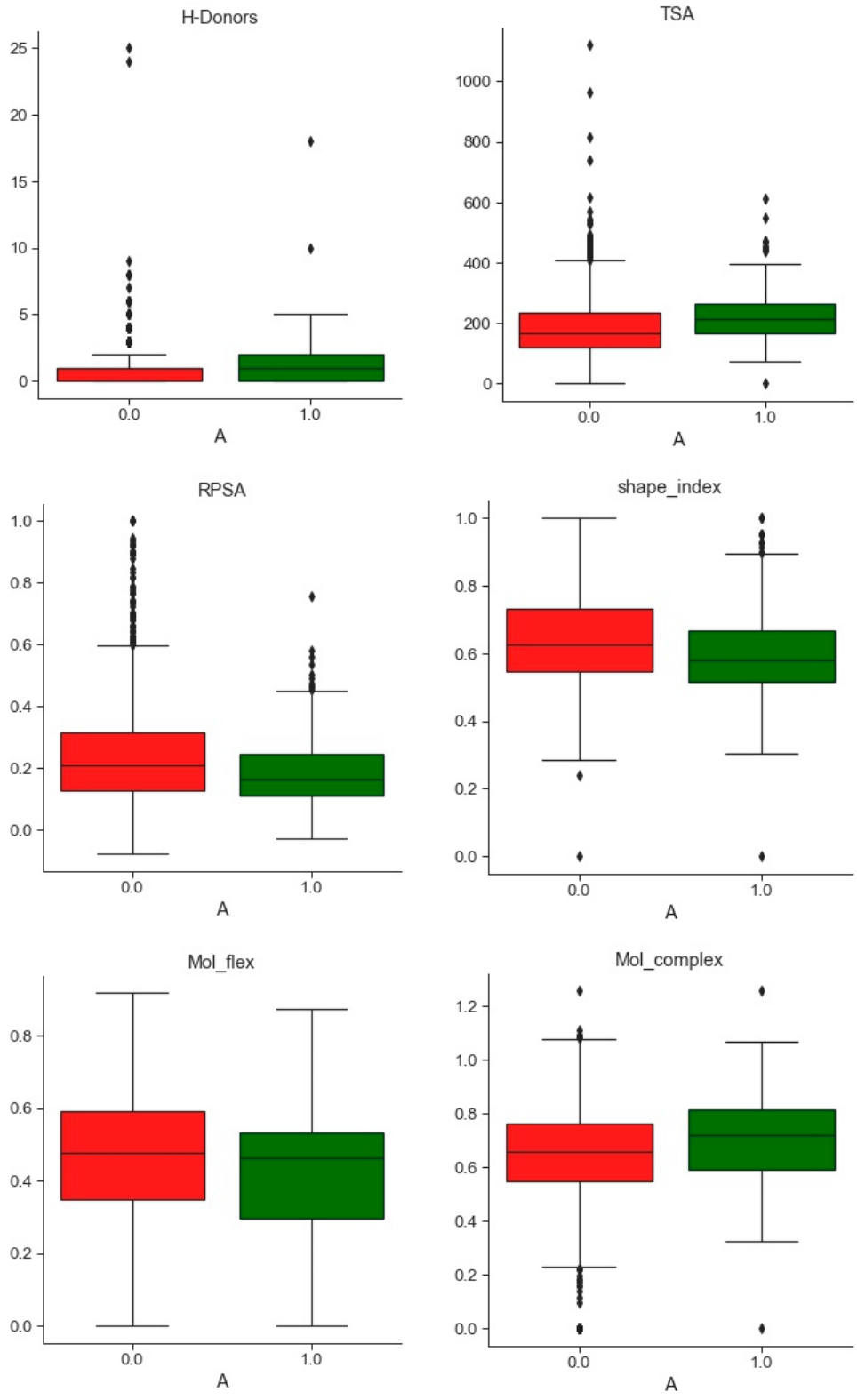


Figure 1. *Cont.*

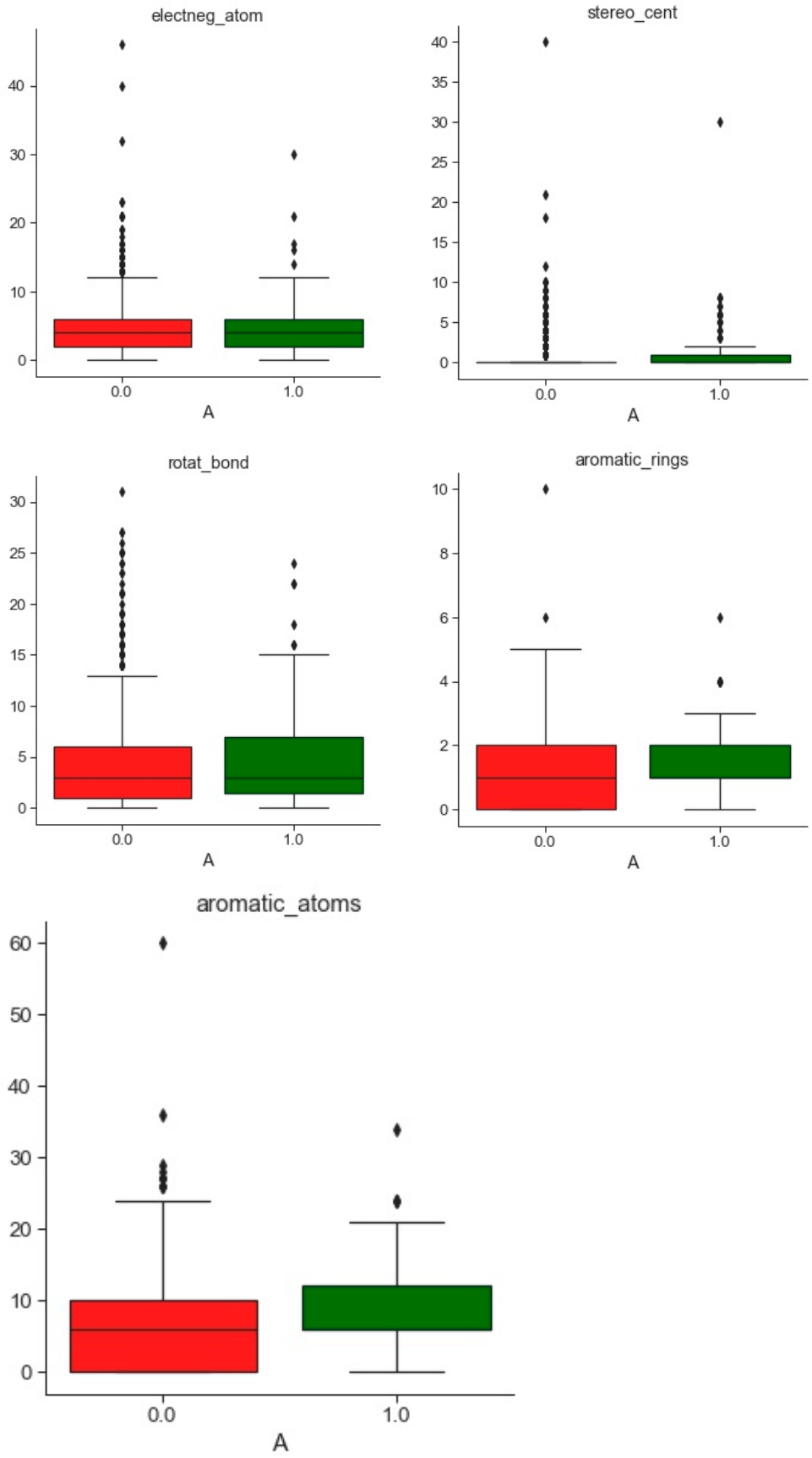


Figure 1. Boxplots representing the distribution of physiochemical descriptors computed with Datawarrior [26] for binding compounds (B) in green and non-binding (NB) compounds in red.

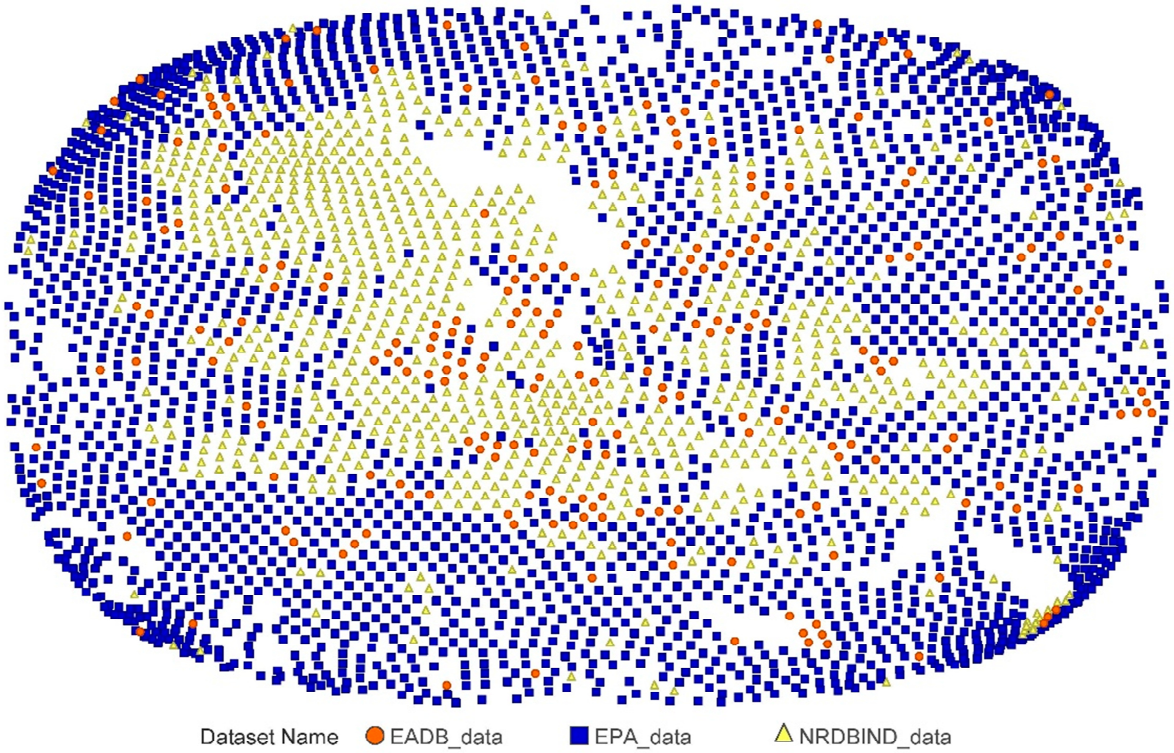


Figure 2. Structure Activity Landscape Index (SALI) maps for all three databases (B and NB compounds): Environmental Protection Agency (EPA) (blue), Nuclear Receptors DataBase Including Negative Data (NR-DBIND) (yellow) and Estrogenic Activity DataBase (EADB) (orange).

2.2. Docking

2.2.1. Docking Outcome

In order to determine the optimal protocol for discriminating ERα B from NB compounds, 7 molecular docking tools (smina-vina, smina-vinardo, smina-dkoes_scoring, smina-adt4, Protein–Ligand ANT System (PLANTS), and Surflex-dock) were explored using 2 approaches: single structure docking and ensemble docking. Docking performance in predicting B compounds was evaluated using the area under the ROC (Receiver Operating Characteristic) curve (AUC) values (Table 1). For the single structure docking approach, mean AUC are comprised between 0.576 for Surflex-dock (with the largest standard deviation between AUCs) and 0.704 for both smina-dkoes (with the smallest standard-deviation between AUCs) and smina-vinardo. The best performance is obtained using smina with the scoring function dkoes for the 1qku structure with an AUC of 0.708. For all the scoring functions, the structure associated with the best performance displays an agonist-bound conformation.

For the ensemble docking approach, all ensemble sizes, from 2 to 7 structures, were tested but no amelioration in the AUC values was observed with ensembles of more than 3 structures. Table 1 summarizes the results obtained for both single structure and ensemble docking approaches for ensembles of 2 and 3 structures (results for the ensembles of size superior to 3 are presented in Supplementary Table S1). The best mean AUC (0.703) and max AUC (0.710) values are associated with the smina_dkoes scoring function for ensemble of 2 and 3 structures, respectively. The lowest mean AUC (0.594) and max AUC (0.616) were obtained for an ensemble of 2 structures using Surflex-dock.

No significant improvement was observed between single structure and ensemble docking approaches. This is particularly true for both smina-dkoes and PLANTS, for which the best AUC obtained using the ensemble docking approach is almost equal to those obtained with single structure docking. It is to note that for all six scoring functions,

the structure associated with the best AUC performance for single structure docking is always present in the best ensemble of 2 and 3 structures.

Table 1. Docking performances (Max area under the ROC curve (AUC), min, mean, and standard deviation (SD)) calculated for the different scoring functions and for the different docking approaches.

Software	Docking Approach	Best Performances		Min AUC	Mean AUC	SD
		AUC	PDB			
smina-dkoes	Single	0.708	[1qku]	0.700	0.704	0.003
	Ensemble of 2	0.709	[2yja-1qku]	0.702	0.703	0.003
	Ensemble of 3	0.710	[2yja-1qku-1g50]	0.704	0.702	0.003
smina-vina	Single structure	0.699	[1a52]	0.643	0.676	0.02
	Ensemble of 2	0.696	[1xp9-1a52]	0.642	0.67	0.017
	Ensemble of 3	0.695	[1xp9-1xp1-1a52]	0.642	0.667	0.014
smina-vinardo	Single structure	0.68	[1a52]	0.686	0.704	0.018
	Ensemble of 2	0.676	[1xp9-1a52]	0.619	0.650	0.019
	Ensemble of 3	0.673	[1xp9-1xp1-1a52]	0.618	0.644	0.018
smina-ad4	Single structure	0.656	[1a52]	0.613	0.639	0.0154
	Ensemble of 2	0.654	[1x7e-1a52]	0.618	0.641	0.009
	Ensemble of 3	0.650	[1x7e-1qku-1a52]	0.623	0.640	0.007
PLANTS	Single structure	0.659	[1x7e]	0.598	0.634	0.019
	Ensemble of 2	0.660	[1x7e-1a52]	0.647	0.62	0
	Ensemble of 3	0.659	[1x7e-1qku-1a52]	0.620	0.642	0.009
Surflex-dock	Single structure	0.604	[1a52]	0.547	0.576	0.027
	Ensemble of 2	0.616	[1xp1-1x7e]	0.556	0.594	0.020
	Ensemble of 3	0.623	[1xp1-1x7e-1a52]	0.562	0.605	0.015

2.2.2. Predictiveness Curve

Predictiveness curve (PC) was used to define docking score thresholds (TH) associated with a high P(active), i.e., the probability of having active compounds in the screened fraction. For each scoring function and for both docking approaches, i.e., single structure and ensemble docking, TH associated with the highest P(active) were defined. For these TH, sensitivity and specificity values were also deduced. The highest P(active) value is the one of smina-dkoes (~0.3) followed closely by PLANTS (see Table S2). However, these values of P(active)max are associated with a low hit rate. As presented in Table S3, the highest P(active)max is associated with a TH of −10 using smina-dkoes and yields a low hit rate (14 hits out of 2442 compound at start). The same tendency is observed for PLANTS for which the screened subset with the highest probability of activity encompasses few molecules: 5 hits in total without any B among them. Thus, we chose to explore TH associated with various sensitivity levels. Table 2 displays the performances for various sensitivity values for both scoring functions smina_dkoes and PLANTS and for both single structure and ensemble docking approaches. The P(active) and enrichment factor (EF) deducted for these TH yielded better results for smina-dkoes than PLANTS. Regardless, trends are the same for both: the higher the sensitivity, the lower are the specificity, the P(active), and the EF. The behavior is the same for single structure and ensemble docking.

In the light of the docking results, we decided to select for the rest of the study the smina-dkoes scoring function and the single structure docking approach (using the 1QKU PDB structure) and to select two potential scoring TH (−6 and −7). Table 3 presents the performance of the selected protocols on the EPA database and the external validation sets (EADB and NR-DBIND) in terms of specificity, sensitivity, and binders retrieval rate.

Table 2. P(active), scoring threshold (TH), Specificity (Sp), Enrichment factor (EF), and the positive predictive value (PPV) calculated for different values of sensitivity (Se) (0.25/0.5 and 0.75) for all the docking approaches and for the scoring function smina-dkoes and Protein–Ligand ANT System (PLANTS).

	Docking Approach	Performances	Se = 0.25	Se = 0.5	Se = 0.75
smina_dkoes	Single (1qku)	P(active)	0.137	0.094	0.094
		TH	−7	−6	−6
		Sp	0.918	0.766	0.601
		EF	1.9	1.65	1.65
		PPV	56/237	111/631	167/1052
	Ensemble de 2 (2yja-1qku)	P(active)	0.134	0.094	0.094
		TH	−7	−6	−6
		Sp	0.916	0.759	0.597
		EF	1.89	1.63	1.63
		PPV	56/242	111/645	167/1061
	Ensemble de 3 (2yja-1qku-1g50)	P(active)	0.137	0.13	0.091
		TH	−8	−7	−6
		Sp	0.915	0.777	0.599
		EF	2.37	1.9	1.59
		PPV	56/244	111/605	167/1057
PLANTS	Single (1x7e)	P(active)	0.127	0.103	0.081
		TH	−79	−72	−64
		Sp	0.876	0.723	0.501
		EF	1.9	1.69	1.42
		PPV	55/328	110/719	165/1261
	Ensemble of 2 (1x7e-1a52)	P(active)	0.123	0.097	0.08
		TH	−82	−73	−66
		Sp	0.86	0.707	0.49
		EF	1.69	1.58	1.42
		PPV	55/362	110/753	165/1287
	Ensemble of 3 (1x7e-1a52-1qku)	P(active)	0.122	0.096	0.079
		TH	−82	−73	−66
		Sp	0.857	0.701	0.493
		EF	1.65	1.6	1.41
		PPV	55/369	110/767	165/1279

Table 3. Sensitivities (Se), specificities (Sp), and positive predictive value (PPV) calculated for the single docking approach with smina_dkoes scoring function screening for both TH = −6 and TH = −7 scoring thresholds.

Scoring Threshold (TH)	Performances	EPA	Estrogenic Activity DataBase (EADB)	Nuclear Receptors DataBase Including Negative Data (NR-DBIND)
TH = −7	Se	0.79	0.48	0.93
	Sp	0.55	0.58	0.03
	PPV	176/2442	63/232	513/732
TH = −6	Se	0.46	0.77	0.99
	Sp	0.78	0.198	0.001
	PPV	103/2442	101/232	553/732

2.3. Pharmacophore Modeling

2.3.1. LB Pharmacophore Models

Since the compounds of the active training set belong to different chemical series, their alignment to derive a single LB pharmacophore is not feasible. To overcome this issue, all the compounds were clustered to obtain subsets of similar compounds for which pharmacophores can be generated. Distance between each cluster was fixed to 0.4 to ensure balanced groups and to minimize the number of singletons. In total, 14 clusters were obtained containing a minimum of 3 and a maximum of 69 compounds per cluster.

6 molecules could not be fitted in any cluster and were not used to generate the pharmacophores models. The maximum number of pharmacophores generated per cluster was set to 10. Each pharmacophore was used to screen the training subset of the EPA database. Based on individual hit retrieval performances, the best pharmacophore of each cluster was optimized according to the procedure described in the methods section. In the case where the optimization protocol failed, i.e., the optimized pharmacophore was not associated with a high rate of B/NB, the other pharmacophores generated for this cluster were considered in the descent order of their individual performances until one pharmacophore could be successfully optimized. If none out of the 10 generated pharmacophores or the corresponding optimized were associated with a high rate of B/NB, no pharmacophore was conserved for this cluster. In total, 11 unique (non-redundant) LB pharmacophores were obtained. Their performances in terms of selectivity and sensitivity are described in Table 4. These 11 LB pharmacophores were combined and used to screen the training subset of the EPA database. High specificity and relatively low sensitivity values were obtained with 30% of the total of binders retrieved against only 2.7% of the total of NB for the training set (Figure 3). To ensure that the performance is not biased towards the ligands of the training set, the 11 LB pharmacophore models were used to screen the test subset of the EPA database. Specificity and sensitivity values obtained were similar to those obtained with the training set and 27% of all B compounds were retrieved against 3% of all NB compounds (Figure 3).

Table 4. Sensitivity (Se) and specificity (Sp) of ligand-based (LB), structure-based (SB), and combination LB and SB pharmacophores, for the training set and the test set of the EPA database, the EADB and the NRDBIND.

		EPA Database		EADB	NR-DBIND
	Performances	**Train Set**	**Test Set**	**Validation Set**	**Validation Set**
LB pharmacophores	Se (B/total_B)	0.305 (51/167)	0.232 (13/56)		
	Sp (NB/total_NB)	0.973 (45/1664)	0.960 (22/555)		
SB pharmacophores	Se (B/total_B)	0.251 (42/167)	0.232 (13/56)		
	Sp (NB/total_NB)	0.990 (16/1664)	0.987 (7/555)		
SBLB pharmacophores	Se (B/total_B)	0.371 (62/1664)	0.321 (18/56)	0.557 (73/131)	0.819 (458/554)
	Sp (NB/total_NB)	0.968 (53/167)	0.595 (25/555)	0.871 (13/101)	0.629 (66/178)

2.3.2. SB Pharmacophore Models

In addition to LB pharmacophores, 31 SB pharmacophores were generated from the holo structures of ERα available in the NR-DBIND. All these pharmacophores were used to screen the training set and were optimized according to the protocol described in the methods section. Redundant pharmacophores were removed, and 15 SB pharmacophores were retained. Screening of the EPA training and test subsets using the 15 SB pharmacophores led to low sensitivity values and high specificity values (Table 4). The percentage of B compounds retrieved with SB pharmacophores is similar to those obtained with the LB pharmacophores, but the percentage of NB compounds retrieved with the SB pharmacophore is lower.

2.3.3. SBLB Pharmacophore Models

Results for both SB and LB selective pharmacophores were combined into a set of SBLB pharmacophores for ERα binding compounds. Redundant pharmacophores were removed to obtain a total of 26 unique SBLB pharmacophores. Performance in terms of

sensitivity and specificity of this ensemble of pharmacophores is shown in Table 4. The set of SBLB pharmacophores is able to retrieve almost 40% of B against only 3% of NB.

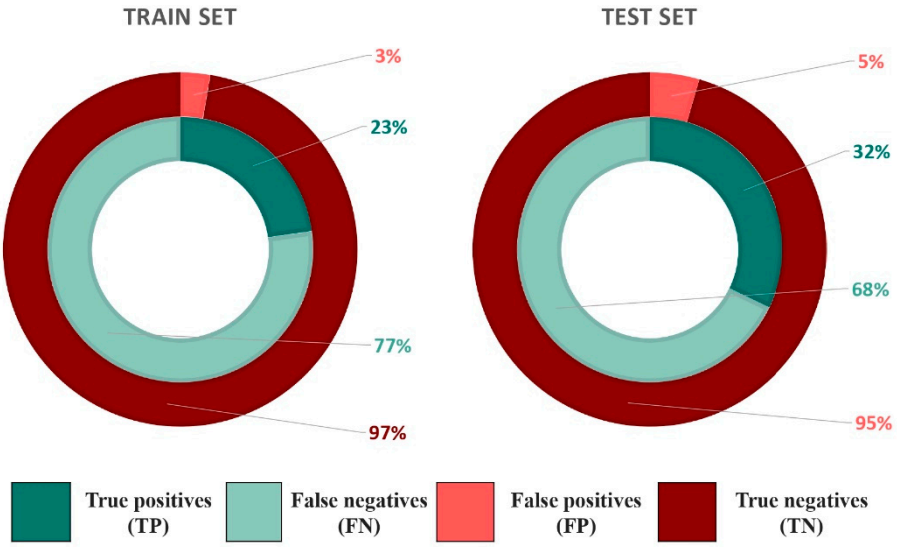


Figure 3. Pie charts displaying the performance of the combination of structure-based and ligand-based SBLB pharmacophores for the train and the test sets.

The 26 SBLB pharmacophores were also used to screen the two external validation sets, i.e., the EADB and the NR-DBIND ERα sets, and the results are shown in Table 4. For EADB, similarly to the results associated with the EPA database, high specificity and low sensitivity values were obtained. The opposite is observed with the NR-DBIND ERα set, for which the sensitivity value is higher than the specificity.

2.4. Combination of Docking and Pharmacophore Models

Individual performances for docking (AUC, Se, and Sp) and pharmacophore models (Se, Sp, and hits retrieval rate) remain moderate, since sensitivities are hardly higher than 50% and the specificities equal or superior to 50% are associated with a low hit rate. For this reason, we evaluated the performance of the combination of docking and pharmacophore models in accurately predicting the binding profile of the compounds to ERα.

Two different protocols for performing this combination were explored, i.e., the consensus and the hierarchical protocols, detailed in the method section.

2.4.1. Consensus Protocol

Using the consensus protocol, each molecule predicted as active using the docking or the pharmacophores models will be identified as an active compound in the consensus protocol results. The remaining compounds will be predicted as inactive. Performances obtained using this protocol for the EPA database and the validation datasets, i.e., the EADB and the NR-DBIND ERα set are depicted in Table 5.

Two docking TH defined using the PC were studied. For TH = −7, a sensitivity of 0.56 and a specificity of 0.76 are obtained for the EPA database. Conversely, for each validation set, the consensus protocol yields higher sensitivity (0.832 and 0.495, respectively) against lower specificities (0.495 and 0.029). When TH = −6 is chosen, the corresponding sensitivities are high: 0.81, 0.937, and 1 corresponding to the EPA database, the EADB, and the NR-DBIND, respectively. Recorded specificities are very low: 0.51, 0.158, and 0.005 for the EPA, the EADB, and the NR-DBIND. The higher positive predictive value (PPV) for the EPA database is reached by applying the TH = −7, with a PPV value around 19%. The same trend is observed with the EADB external validation set, whereas quite similar PPV are obtained for both threshold using the NR-DBIND set. The PPV obtained with the

external validation sets using both TH = −6 and TH = −7 were largely superior to those obtained with the EPA database.

Table 5. Sensitivities (Se), specificities (Sp), and B/Total ratio calculated for the consensus and hierarchical screening method for two different thresholds (TH) of docking scores.

	TH	**−7**			**−6**		
		Se	**Sp**	**PPV**	**Se**	**Sp**	**PPV**
Consensus protocol	EPA database	0.56	0.76	124/652	0.81	0.54	180/1205
	EADB	0.832	0.495	109/160	0.931	0.158	122/207
	NR-DBIND	0.986	0.029	546/719	1.0	0.005	554/731
Hierarchical protocol	EPA database	0.25	0.99	55/84	0.32	0.98	72/117
	EADB	0.206	0.960	27/31	0.370	0.911	52/61
	NR-DBIND	0.756	0.635	419/484	0.814	0.635	451/516

For equal specificity values between both TH, the TH = −6 yields better sensitivities for the EPA database as well as for the validation datasets. This is why our choice of docking TH is set at −6 for the consensus protocol.

2.4.2. Hierarchical Protocol

We first evaluated the impact of using hierarchical screening with the pharmacophore models prior to or after the molecular docking models on the performance in enrichment.

Since both protocols displayed similar performances in terms of sensitivity and specificity, we relied on computational times to select the protocol. We thus decide to first screen using the pharmacophore models and then using the optimal docking protocol previously defined. On a desktop computer with 8x Intel(R) Xeon(R) CPU L5520 @ 2.27 GHz it takes ~75 min to dock the 2442 molecules against one ERα structure versus ~5 min to screen the same number of compounds on the 26 SBLB pharmacophore models.

Results depicted in Table 5 are those obtained using this hierarchical screening, i.e., the entire database is screened using the pharmacophore models and the compounds thereby identified as hits are used as the screening database for the docking method. The docking outcomes are then analyzed using the 2 docking scores TH previously identified and corresponding to different sensitivity values. For both TH values, the same trend is observed, i.e., high specificities (0.99 and 0.98) and low sensitivities (0.25 and 0.32).

Table 4 also presents sensitivity, specificity, and PPV obtained using the hierarchical protocol on the validation sets. The performance associated with the EADB is very similar to those obtained with the EPA database whereas the hierarchical protocol applied on the NR-DBIND ERα set lead to high values of sensitivity and specificity for both thresholds. Based on the hierarchical protocol outcomes, in particular the sensitivity values, on both the EPA database and the external validation sets, we selected the TH = −6 as the threshold to be used for docking scores using the hierarchical protocol.

3. Discussion

Through this work, we aim at finding the best in silico protocol(s) to discriminate B from NB compounds for ERα. Both SB and LB methods were evaluated, together with two different protocol to combine them.

3.1. Compounds and Database Preparation

The comparison of the distribution of the 15 constitutional descriptors for the three databases, i.e., EPA, EADB, and NR-DBIND, was performed in order to ensure that the difference in activity was not solely explained by the difference in physiochemical properties.

In order to assess the prediction performance of our models, we used external validation sets. Pairwise comparison of topological fingerprints between the EPA database and each external validation sets verifies the structural dissimilarity between those sets and

thus the possible use of the EADB and NR-DBIND ERα sets as external validation sets. Moreover, the SALI map confirms that the three databases belong to the same chemical space, which was recommended for pharmacophore models validation [27].

3.2. Docking

For the docking approach, both single structure and ensemble docking were explored. Three software with free academic licenses, accounting for 6 scoring functions, were used, i.e., smina (smina-ad4, smina-dkoes, smina-vina, smina-vinardo), Surflex-dock, and PLANTS. Although different magnitudes of AUC were obtained, most of them agreed on the elected structure yielding the best single structure docking results: 4 out of the 6 docking methods associated the best outcomes with the 1a52 structure. However, the highest AUC were obtained with different structures, 1qku and 1x7e for smina-dkoes and PLANTS, respectively. Interestingly, the 1a52 structure presents an artifactual position of the helix that is extending away from the body of the ligand binding domain. The resulting conformation is more similar to an antagonist-bound ERα structure than an agonist-bound one [4]. This observation leads to discard 1a52 despite its selection by most of the software and reinforces the choice of the 1qku structure and smina-dkoes as the optimal single structure docking protocol. It is to note that 1qku is co-crystallized with the native ligand 17β-estradiol. Furthermore, it was shown that smina_dkoes was very proficient at sampling low RMSD poses compared to Vina [28].

No major performance improvement as evaluated by the AUC values was brought by ensemble docking over the single structure strategy. This was true using either only agonist-bound structures ensembles or combinations of agonist and antagonist-bound structures. When considering only agonist compounds as positives and the remaining compounds (antagonists and experimental non binders) as negatives, both agonist-bound and antagonist-bound structures were associated with similar AUC values (results not shown). Similarly, no significant differences in docking performance were noted among agonist- and antagonist-bound structures when only antagonists were set as positives and all the remaining compounds as negatives. This could be explained by the fact that ERα conformations used in this study are very similar, as shown by the RMSD values obtained among all structures (Table S4). The structures used display a limited flexibility, explaining the similar performances obtained in terms of AUC values, regardless of the pharmacological profile of the co-crystallized ligands or of the binding compounds. This limited flexibility sampling can also explain the lack of significant performances improvement observed using the ensemble docking strategy. Furthermore, previous studies also showed that ensemble docking did not always outperform the single structure docking approach especially when the single structure is rationally selected [29]. Finally, and although displaying several advantages, such as accounting for the flexibility of the target, ensemble docking presents also noteworthy drawbacks. Docking a database against more than one protein structure requires more computational resources and/or time. Ensemble docking can also lead to inaccurate predictions due to a favored inaccurate interaction with a particular protein conformation included in the ensemble [30].

3.3. Predictiveness Curve

Molecular docking is a valuable method often used to elucidate a mechanism of action or to predict the nature of interactions established between a ligand and a target protein. It can also be used as a screening tool to filter a database according to docking scores. In a virtual screening protocol using molecular docking, the ranked list of compounds according to the docking scores is generated. Then, a fraction of the top scoring compounds (1%, 5%, 10% . . .) is tested experimentally depending on the budget and experimental facilities. For this type of protocol, defining a docking score threshold is not necessarily a priority. In our study, we preferred to rationally select an optimal docking threshold rather than selecting an arbitrary fraction of the top scoring compounds. Endocrine disruptome [21], for example, is an online tool based on docking calculations that also established docking

scores thresholds to differentiate between binding and non-binding compounds for a set of NRs. In an ideal case where all B compounds would have better docking scores than the NB compounds (Figure 4, left panel), the threshold would simply be defined as the value separating the docking score values of the last ranked B compound and the first ranked NB compound. However, in reality, some B and NB compounds present very similar docking score values and the distribution of the profiles of scores between B and NB compounds are often overlapping. In our study, both distribution curves for B (green) and NB compounds (red) overlap (Figure 4, right panel), preventing a straightforward manual definition of a perfect score threshold. To help the definition of a score threshold, we used Screening Explorer [31], an interactive tool for the analysis of screening results, based on the predictiveness curve (PC) metric [32].

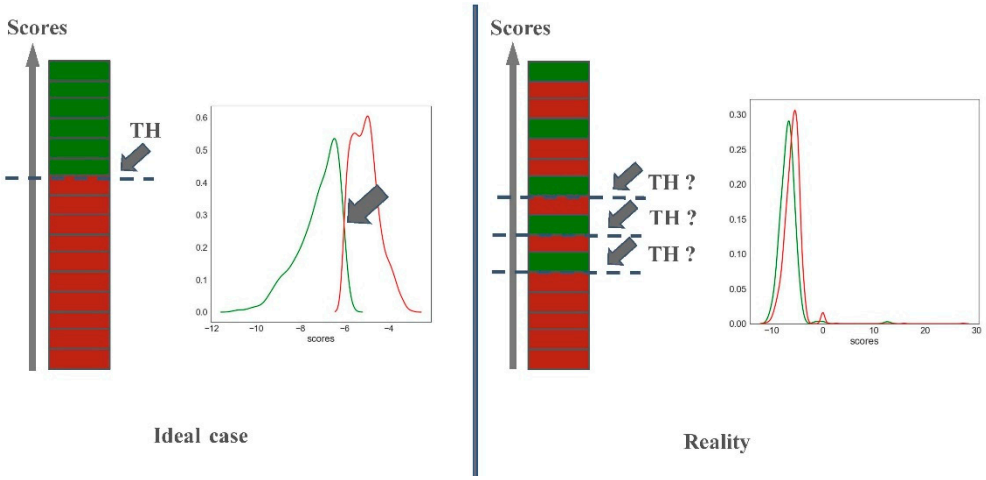


Figure 4. Docking scores distribution between B (green) and NB compounds (red) of the EPA database.

Although newly introduced in the virtual screening field, the PC has already been applied in different studies [27,33–37]. This metric is usually used altogether with ROC curves and enrichment factors to assess the ability of a given method to discriminate active compounds from inactive ones [38]. PC have been used in the literature to define a score threshold to discriminate agonist from antagonist compounds for androgen receptors [27]. As in [27], we assessed the predictiveness of the single structure and ensemble docking approaches as well as each docking/scoring scheme. Using the Screening Explorer tool, 2 potential docking score thresholds were identified to differentiate ERα B from NB. We thus chose to evaluate these 2 docking score thresholds for the combination of the docking procedure and the pharmacophores modeling.

3.4. Pharmacophores

Several studies already focused on generating pharmacophores for NRs ligands [27,39–42]. In this work, numerous SB and LB pharmacophores targeting ERα were generated and optimized. A large number of B were retrieved by both SB and LB pharmacophores, but some were specifically identified by only one or the other class of pharmacophores. Consequently, all non-redundant pharmacophores were merged in the SBLB ensemble that contains approximately as much LB (11) as SB (15) pharmacophores. The SBLB ensemble of pharmacophores achieve better sensitivity over a slight drop in the specificity compared to SB pharmacophores or LB pharmacophores. Hits retrieved by the SB and LB pharmacophores are represented in Figure 5 together with the yield of the SBLB pharmacophores.

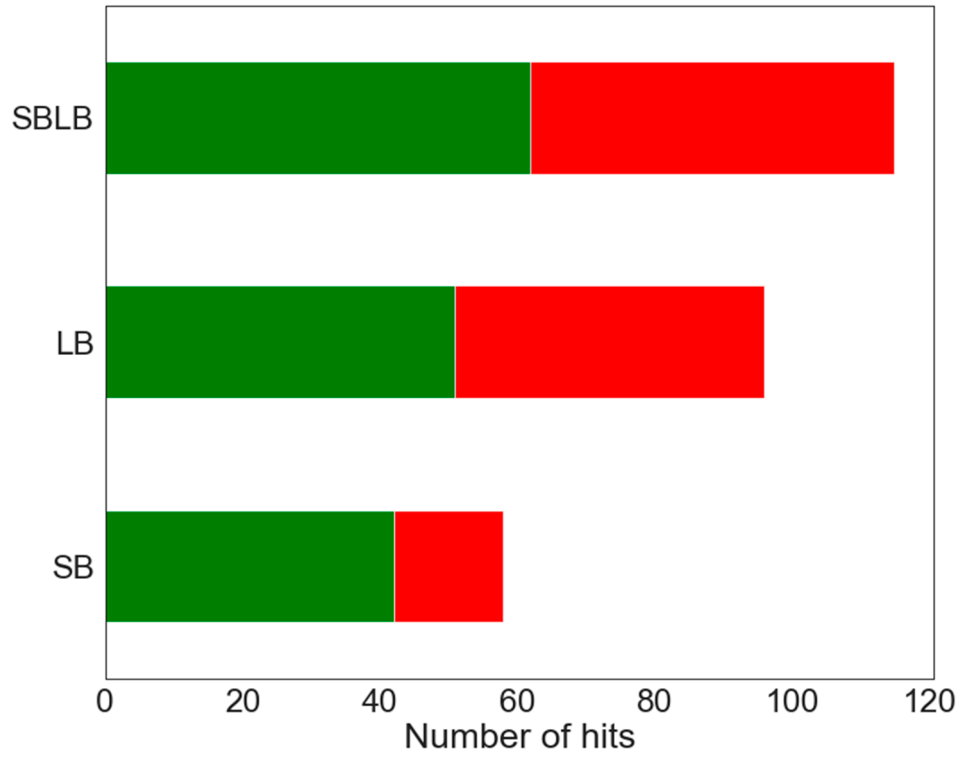


Figure 5. Barplot displaying the proportion of B (green) and NB (red) retrieved as hits using SB and LB pharmacophore models individually and with the SBLB combination.

Interestingly, our SBLB pharmacophores applied to the external validation data yielded very good sensitivities and lower specificities. This is similar to the results obtained by Réau et al. [27] with pharmacophores models generated using the NR-DBIND AR set. This study also suggests that pharmacophores are only suited for data filtering as long as the compounds belong to the same chemical space as the molecules used to build the model. The SALI map of all the databases (the EPA training database and the EADB and NR-DBIND ERα external validation sets) in Figure 2 shows that our data fit this requirement and supports the use of pharmacophores for this study. The lower sensitivities obtained with the EPA database compared to those obtained with the external validation sets may be explained by the imbalance in the number of B and NB that exists in the EPA database (223 B and 2219 NB) compared to the validation sets which present lower proportions of inactive data. The SBLB pharmacophores present better performance in discarding true negatives than in identifying true positives. To overcome this issue, we decided to evaluate the ERα B prediction performances obtained when combining SBLB pharmacophores and docking approaches.

3.5. Combination of Methods

Combining several bioinformatic methods is often used for various purposes such as extending the knowledge about a drug–target interaction or refining screening results [43–47]. Docking methods are usually successful in poses prediction but fail at distinguishing active from inactive compounds yielding low sensitivities. Pharmacophore methods on the other hand, used in the appropriate applicability domain [27], succeed at discarding molecules which structures misfit the requirements to interact with the binding site. In accordance with these results, our study shows that the 2 types of combinations we evaluated enhance performances towards better specificities for the hierarchical protocol and better sensitivities for the consensus protocol (Figure 6).

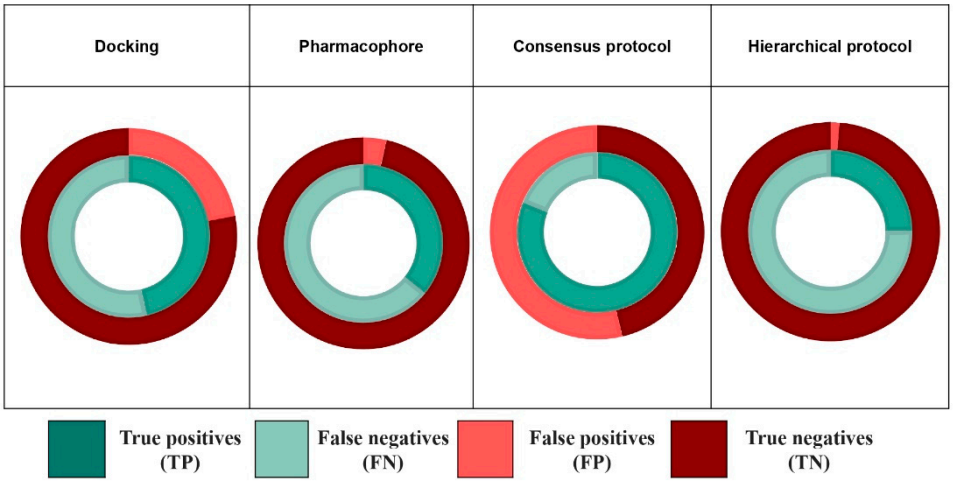


Figure 6. Pie charts illustrating the performance of each individual model (docking and pharmacophores) and the combination of both using the consensus and the hierarchical protocol.

Furthermore, a review of studies dedicated to NR, and more specifically to the prediction of EDCs able to bind ERα, enabled us to better assess the performances obtained with our models. We obtained high sensitivities values, 0.81 for EPA and 0.93 and 1 for EADB and NR-DBIND, respectively, associated with low specificities. The different studies herein undermentioned can be divided into studies relying on docking models and others that are mostly based on machine learning and QSAR (Quantitative structure activity relation) models [20,23,24,48–53]. Docking methods of the studies of the former class [21,22,54–57] present AUC values similar to those obtained with our selected scoring function and receptor structure. It should be noted that these docking studies used various ERα structures, and especially the 1a52 we chose to discard because of its artifactual position of the helix 12 [4]. Studies of the latter category are the most abundant, and present high AUC values around 0.8 with good overall sensitivities and specificities. These good prediction performances are not surprising since classification and QSAR models are known for their ability to well predict structural analogs. However, these methods can suffer from overfitting bias which can lead to lower performances if applied on a different dataset as they will be unable to predict completely new/different molecules [58]. Moreover, outliers are frequently discarded in this kind of study, but these compounds may introduce a category of yet unrepresented compounds. Nevertheless, these LB methods perform better than our LB pharmacophores and should be investigated for future integration in the protocol.

Some sources of bias that may have affected the performances should be taken into consideration. Annotation errors of biological assays are possible, and compounds identified with binding assays may bind on different ERα binding sites. Furthermore, the compounds of the EPA database are mostly compounds suspected to be toxic and not therapeutic compounds. Even if our models were validated with external sets dedicated to therapeutic compounds, it is important to enrich databases with more compounds relative to both therapeutic and toxicological explorations according to the purpose of the study [59,60].

Previous studies [27,49,55,61] suggested that the pharmacological profile of the ligands should be considered to better discriminate agonist from antagonist compounds. Endocrine disrupting chemicals act in several ways including agonism and antagonism [17] and it is important to be able to retrieve ERα B regardless of their pharmacological profiles. The structure identified to be optimal for the docking study is in an agonist-bound conformation. We thus verified that our protocol was not biased towards agonist ligands and that we were also able to identify ERα B with different pharmacological profiles.

We compared the distribution of pharmacological profiles within the starting database, i.e., the EPA database, as well as within the hits obtained for each screening protocol (see Figure 7). In the EPA database, the pharmacological profile annotation was achieved using agonist and antagonist experimental assays. Among the 223 compounds, 58 are agonist

(26%), 50 antagonist (22.4%), and 66 agonist–antagonist (29.6%) compounds. No pharmacological profile annotation was available for 49 molecules (22%). Interestingly, the relative proportion of each pharmacological profile observed in the initial EPA database was maintained among the hits of both consensus and hierarchical protocols. This highlights the fact that the screening protocol presented herein is able to identify ERα B, regardless of their pharmacological profile and is thus not overfitted towards any pharmacological profile.

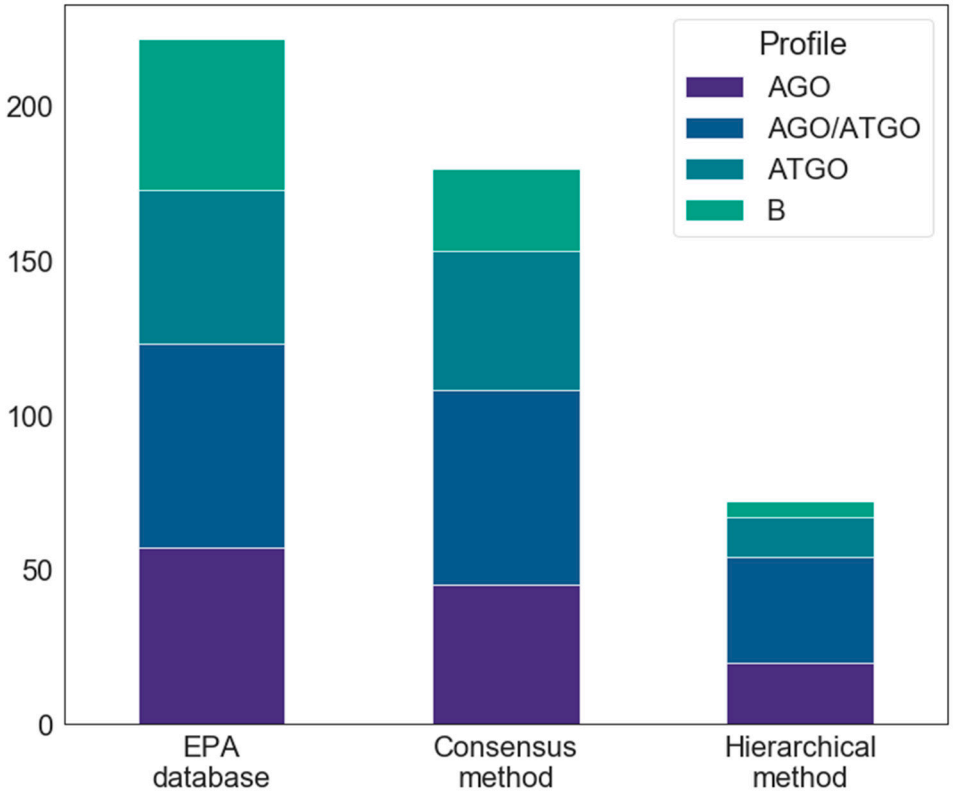


Figure 7. Barplots illustrating the relative proportion of compounds of each pharmacological profile (AGO: compounds with ERα agonist activity, ATGO: compounds with ERα antagonist activity, AGO/ATGO: compounds with both ERα agonist and antagonist activities, B: ERα binders without pharmacological profile annotation) in the EPA database and among the hits identified using the hierarchical and consensus protocols.

Finally, it is important to mention that both sensitivity and specificity are valuable for assessing the screening quality. However, and depending on the purpose of the study, one value tends to be more meaningful than the other. Therapeutic studies favor good specificities as they are an indicator of the ability to discard true negatives, which is more important to reduce the number of molecules to be tested in vivo. For toxicological studies, high sensitivities are preferred, as the goal is to identify the maximum of potentially undesired compounds. These observations are supported by the results obtained for validation sets. In this way, we suggest that the consensus protocol is better tailored for our study and the hierarchical protocol could better suit drug design projects. Both protocols provide a list of compounds that are predicted to bind ERα. These predictions must be confirmed and the estrogenic activity modulation and potential endocrine disruption effects should be further experimentally assessed.

4. Materials and Methods

4.1. Compounds, Databases Preparation, and Annotation

Two types of dataset were used, i.e., a set formed by EPA compounds used to build the different individual methods and two external data sets (the NR-DBIND, the EADB database) used for validation.

4.1.1. EPA Dataset

Compounds and biological data used to build the training dataset were extracted from the United States Environmental Protection Agency (EPA). Chemical compounds and their associated biological data were downloaded from the DSSTox dashboard in February 2019. The platform has been removed since then and compounds can now be found under the Comptox dashboard [62]. This dashboard gathers high throughput screening data of a large and structurally diverse chemical library of compounds sus-pected to be of risk for humankind and for the environment against a wide spectrum of biological targets involved in toxicity pathways [63]. Compounds included in training dataset were obtained by filtering the DSSTox/Comptox database to only keep compounds that have undergone binding assays on ERα receptor. All compounds were available in csv files where each molecule was identified by its SMILES and CAS number. Binding compounds were selected to form the active subset (activity annotated 1) and the non-binding molecules constituted the inactive subset (activity annotated 0). This data-base will be referred to as the "EPA database". The EPA database is available in the Sup-plementary Materials in SMILES format.

4.1.2. Validation Sets

- NR-DBIND

The NR-DBIND (Nuclear Receptors DataBase Including Negative Data) is a non-commercial manually curated benchmarking database that provides affinity data for small molecules that were experimentally tested against 28 nuclear receptors [64]. For this study, a filter was applied to extract compounds tested against ERα. All compounds were directly downloaded from the website (http://nr-dbind.drugdesign.fr/, accessed on 20 November 2019) in SMILES format and annotated by their CAS names. Binding compounds were selected to form the active subset and the non-binding molecules constituted the inactive subset.

- EADB

The Estrogenic Activity Database (EADB) developed by the NCTR (National center for toxicological research) assembles a large number of estrogenic activities data from various sources [56,65,66]. It contains 18.114 estrogenic-activity data points collected for 8212 chemicals tested in 1284 binding assays, reporter-gene assays, cell-proliferation assays, and in vivo assays in 11 different species. The database has been directly downloaded from the website (https://www.fda.gov/science-research/bioinformatics-tools/estrogenic-activity-database-eadb, accessed on 25 November 2019) and filtered to only keep data relative to human ERα.

4.1.3. Molecule Curation and Preparation

The same molecule curation and preparation protocol was applied for the EPA database, the NR-DBIND, and the EADB validation sets. SMILES were standardized using Standardizer from the ChemAxon suite [67] and salts and fragments were removed together with duplicates and small molecules containing less than 5 atoms. Conformations were generated using i-Con [16], the conformer generation tool of LigandScout [68], with BEST settings except for the maximum number of conformations per molecule that was set to 25. Compounds containing certain metal atoms (e.g., Pb or Hg) were removed from the docking collection mainly because the software used were unable to process these molecules. Finally, molecules were converted into the appropriate format for the different

software at use, i.e., pdbqt for docking with smina, mol2 for PLANTS, and Surflex_dock and ldb for pharmacophore model generations.

In order to assess the accuracy of the data, 15 constitutional, physiochemical, and molecular descriptors were computed for each molecule of the three databases, namely, molecular weight (MW), ClogP, ClogS, number of HBond-Acceptors (H-Acc), number of HBond-Donors (H-Don), Total Surface Area (TSA), Relative Polar Surface Area (RPSA), Shape Index, Molecular flexibility (Mol_Flex), Molecular Complexity (Mol_Comp), number of Electronegative atoms (Elect_atom), number of Stereo Centers (Stereo_cent), number of rotatable bonds (rotat_bond), number of aromatic rings (aromatic_rings), and number of aromatic atoms (aromatic_atom). Descriptors were computed using the DataWarrior software [26]. Moreover, topological fingerprints were computed using the rdkit library [69] for python and pairwise Tanimoto coefficient (Tc) were calculated between compounds of the EPA database and the EADB on one side and EPA database and the NR-DBIND on the other side.

4.2. Structures Preparation

ERα structures were selected according to 3 criteria: (1) human structures; (2) without mutations nor residue's deletion in the ligand binding domain; (3) referenced by a scientific article. Accordingly, 31 holo structures were used for SB pharmacophore building. Among these 31 structures, only 7 holo (Protein-ligand) crystal structures of human ERα were used for docking (Table S5). The 24 remaining structures were discarded since they presented residues deletion in the binding site that can affect docking results more than pharmacophore building. Among these 7 structures, 2 are classified as antagonist bound as they are co-crystallized with an antagonist molecule. The remaining 5 structures are agonist-bound and 4 of them share the same co-crystallized ligand, the 17β-estradiol. For the docking procedure, the structures were directly downloaded from the NR-DBIND database [64] since they are already enumerated, annotated, and cleaned. Format conversion from PDB to the appropriate docking format was done accordingly to the requirements of the software, i.e., PDB were converted into mol2 format using the software chimera [70], into pdbqt with the prepare_receptor4.py python script available with the MGLTool [18]. In order to generate the structure-based pharmacophores, structures were directly downloaded from the RCSB website [71] via the LigandScout graphical interface.

4.3. Docking

4.3.1. Protocol

Docking is a structure-based virtual screening method that aims at predicting the pose of a ligand inside a protein [17]. All docking calculations were performed with 3 different software with free academic licenses, i.e., smina [72], PLANTS [73], and Surflex-dock [74]. The same binding site was used with the 3 software that was delimited using the co-crystallized ligands. For each software, 5 docking runs were performed.

Smina is a fork of AutoDock Vina [28] that is designed for scoring function development and minimization workflows [72]. It relies on the same sampling algorithm as vina, the latter being the succession of stochastic mutations steps, but integrates several scoring functions. For this study, we relied on 4 scoring functions already implemented within smina, i.e., vina [28], the Vina RaDii Optimized (vinardo) [75], dkoes [72], and ad4 scoring functions. All dockings were performed using the default options of smina and num_modes = 20 and exhaustiveness of 8. The bounding box coordinates were determined based on the crystal structure of 1a52 used as reference to align the remaining structures. The box parameters were chosen based on the co-crystallized ligand position with a spacing of 1 Angstrom. A cubic box was delimited with size_x, size_y, and size_z set to 20 and the following coordinates center_x = 107.175, center_y = 14.983, and center_z = 96.009. PLANTS relies on the docking algorithm carrying the same name. This Protein–Ligand ANT System (PLANTS) algorithm is based on ant colony optimization, a class of stochastic optimization. An artificial ant colony must find the minimum energy conformation of the ligand within

the receptor through a trail of pheromone whenever an ideal low energy conformation is found. This marking is iteratively changed until the lowest energy conformation is found [73,76,77]. The binding site coordinates were the same that were used for smina. Regarding other parameters, the binding site_radis was set to 18, the cluster_structures to 10, the cluster_RMSD to 2, and the search speed to "speed2".

Surflex-dock is a docking methodology that combines Hammerhead's empirical scoring function with a molecular similarity method to generate putative poses of ligand fragments [74]. The search approach is based on an incremental construction and a fragment assembly method similar to the genetic algorithm. Surflex-Dock uses a pseudo-molecule, a protomol, as a target to align fragments of the ligands. Protomols were generated starting from the holo structures.

4.3.2. Docking Performances Analyses

- Single structure docking and ensemble docking

In the single structure docking approach, docking performance for each PDB structure was evaluated individually by calculating the area under the ROC curve (AUC). AUC values were computed with python using the *scikitlearn* library [78] and the package *sklearn metrics*. In the ensemble docking approach, docking performances of all the possible ensembles of 2, 3, 4, 5, 6, and 7 structures were computed. In this approach, each ligand was sequentially docked into several protein structures. The results were post processed to keep only, for each ligand, the best docking score among all structures. All ligands are then ranked according to these new scores and the corresponding AUC are computed. Python version 3.8.1 was used to prepare data and analyze the results.

- Predictiveness curves

Although docking scores are continuous values, they can be transformed into a binary classifier to discriminate between ERα B and NB using the predictiveness curve (PC) [27,32]. The predictiveness curve is a metric usually used in clinical epidemiology to evaluate the ability of a biological marker to assess the fit of risk models and to estimate the clinical utility of a model when applied to a population [32]. Transferred to the field of Chemoinformatics, this metric can be used to assess the predictive power of a screening methods as well as defining a score threshold retrieving best candidates to be tested experimentally. In this way, PC was used to define a docking scoring threshold for which we can compute the probability that a compound with this given score will be a B compound and define associated sensitivity (Se) and specificity (Sp) (*cf* Equations (1) and (2)). Enrichment factor $EF_{x\%}$ and positive predictive value (PPV) were also calculated following the Equations (3) and (4) where $Hits_{x\%}$ is the number of active compounds in the top x% of the ranked dataset, $Hits_t$ is the total of active compounds, $N_{x\%}$ is the number of compounds contained in the x% of the dataset, and N_t is the total number of compounds in the dataset.

$$Sensitivty = \frac{Nb\ of\ True\ Positives}{Nb\ of\ True\ Positives + Nb\ of\ False\ Negatives} \quad (1)$$

$$Specificity = \frac{Nb\ of\ True\ Negatives}{Nb\ of\ True\ Negatives + Nb\ of\ False\ Positives} \quad (2)$$

$$EF_{x\%} = \frac{\frac{Hits_{x\%}}{N_{x\%}}}{\frac{Hits_t}{N_t}} \quad (3)$$

$$PPV = \left[\frac{Nb\ of\ True\ Positives}{Nb\ of\ True\ Positives + Nb\ of\ False\ Positives}\right] \times 100 \quad (4)$$

The aim of the study is to select as much positive data as possible (toxic compounds). It is then interesting to identify a TH associated with a high probability of activity P(active) but also a high value of sensitivity (Se).

Various TH values and their P(active), Sp, PPV, and EF were calculated for different sensitivity values (0.25/0.5/0.75). The highest P(actives)max was calculated beforehand for each scoring function and for the different ensemble sizes.

4.4. Pharmacophore Modeling Protocol

Structure based (SB) and ligand based (LB) pharmacophores were generated using LigandScout software version 4.4 [68].

4.4.1. Ligand Based Approach (LB) Models Protocol

In order to generate LB-pharmacophores, active compounds from one side and inactive compounds on the other were both divided into training and test sets. 75% of the active compounds and 75% of the inactive compounds were gathered to form the training set. The remaining 25% of active compounds and inactive compounds were used to form the test set. The active compounds of the training set were clustered using the i-cluster [79] tool provided with LigandScout software and pharmacophores were generated for each of the resulting clusters. Default parameters of the I-cluster tool were used, i.e., cluster_dis = 0.4 with average method, except for the maximum number of conformations set to 3. In order to derive a LB-pharmacophore dedicated to a particular cluster of compounds, LigandScout operates in several steps: (1) conformations of the ERα ligands included in the cluster are generated using the ICON algorithm; (2) molecules are ranked according to their flexibility and the best alignments; (3) for each compound, the generated conformations are used to create intermediate pharmacophores that are ranked using several scoring functions; (4) common features are aligned to all the conformations of the next molecule and so on until all the molecules are processed [80]. Each final pharmacophore obtained with this protocol was used to screen the train set on which global and individual performances were assessed. In order to make sure that data separation into training and test does not affect the performance, the whole procedure (from training and test set separation to pharmacophores generation and evaluation) was repeated 25 times. The iteration yielding the best global performances was kept and used during the pharmacophore optimization set and the composition of each set.

4.4.2. Structure Based Approach (SB) Models Protocol

3D SB pharmacophores were automatically generated from the PDB structures of ERα included in the NR-DBIND [64]. In this approach, the LigandScout algorithm tags the key features of the ligands that are interacting with the residues of the receptor. To complete the pharmacophore, an ensemble of exclusion volume spheres is generated to represent the shape of the active site [42].

- Pharmacophore model optimization

In order to optimize the pharmacophore, we followed literature recommendations, especially a screening protocol that succeeded in generating selective pharmacophores for NR agonist ligands and selective pharmacophores for NR antagonist ligands [42]. This protocol was applied on both SB and LB pharmacophores. The generated 3D pharmacophores were used to screen the training set and the test set. All the ligands were converted into ldb format using the idbgen tool provided with LigandScout. For each pharmacophore, a first screening was made with LigandScout default settings and particularly the Max. number of omitted features set to 0. Two case scenarios were possible. If after the first screening, the ratio PPV was high, i.e., few non binders are retrieved but a large number of binders are matching the pharmacophore, a second screening was performed with the same pharmacophore but setting the Max. number of omitted features to 1. This way, non-essential features could be identified to be removed or set as optional possibly leading to the retrieval of more active compounds and less inactive molecules. After that, a third screening was performed with Max. number of omitted features set to 0 again. If the ratio of PPV decreased, this pharmacophore was not validated, and another round of feature identification was performed. If the ratio increased, the pharmacophore was

validated, and other potential non-essential features were investigated. This protocol was applied to each pharmacophore until 3 pharmacophoric features were retained or until no non-essential features could be identified.

4.4.3. Combination of SB and LB Pharmacophores Models

Once a collection of optimal SB and LB pharmacophores was obtained, redundant pharmacophores were removed. Redundant pharmacophores are pharmacophores that can be removed without decreasing the recall, i.e., pharmacophores that only retrieved ligands that are also retrieved with other pharmacophores of the set. To remove these redundant pharmacophores, all generated pharmacophores were ranked according to the number of hits they retrieved. Then, each pharmacophore was removed sequentially, starting from the pharmacophore associated with the smallest number of hits. For each removal, the impact on the recall was evaluated. If the recall was not affected, the pharmacophore was dismissed and, in the opposite, if the recall decreased, the pharmacophore was conserved.

The SBLB pharmacophores used in this study are available in the Supplementary Materials in pml format.

4.5. Pipelines Construction

Two different ways of combining pharmacophore models and docking were explored, the consensus and the hierarchical protocols. The first protocol consists in the analysis of the union of the results belonging to each model. Each molecule predicted as active by docking or pharmacophore will be predicted as active compound by the consensus protocol. The remaining compounds will be predicted as inactive. The second approach used a hierarchical protocol in which the database undergoes a sequence of screening methods. Two possible sequences exist: [pharmacophore-docking] or [docking-pharmacophore].

5. Conclusions

In the present work, we present a pipeline designed for the prediction of potential EDCs acting through the binding to ERα. Optimized protocols for docking studies and SB and LB pharmacophore models' generation were evaluated together with the best approach to combine them. Both combination approaches that were investigated here, i.e., consensus protocol and the hierarchical protocol, yielded good results. However, we recommend favoring the consensus protocol for toxicological studies and the hierarchical protocol for the identification of therapeutic compounds. Results were validated using two external datasets. Using our pipeline, we show that combining several in silico methods can enhance the prediction performances for compounds binding to ERα. Additional methods should be evaluated and implemented in this pipeline such as classification models.

Supplementary Materials: Supplementary Materials can be found at https://www.mdpi.com/1422-0067/22/6/2846/s1, Figures S1 and S2: Boxplot of the distribution of the 15 physiochemical properties computed with Datawarrior for EADB (S1) and NR-DBIND (S2), Figure S3: Boxplot representing the distribution of pairwise calculated Tanimoto coefficient between EPA database and EADB (in blue) and NR-DBIND (green) topological fingerprints, Table S1: Docking performances for both single and ensemble docking approach illustrated with the AUC of the best and the worst ensembles; Table S2: Maximum values of predictiveness (P(active)) associated to each scoring function and each docking approach; Table S3: Sensitivity (Se), specificity (Sp), scoring threshold (TH), Enrichment factor (EF) and PPV calculated for the best scoring function (Smina-dkoes and PLANTS) and corresponding to P(active)max for different docking; Table S4: Pairwise RMSD computed between all the protein structures; Table S5: PDB Structures used for structure based model building. All 31 structures were used to generate SB pharmacophores and only those colored in blue were used for docking; EPA database in SMILES format; SBLB pharmacophores in pml format.

Author Contributions: Conceptualization, A.S., M.M. and N.L.; methodology, A.S. and N.L.; validation, A.S., N.L. and M.M.; formal analysis, A.S. and N.L.; investigation, A.S.; resources, A.S., N.L. and M.M.; software, A.S.; data curation, A.S.; writing—original draft preparation, A.S.; writing—review and editing, A.S., N.L. and M.M.; visualization, A.S.; supervision, N.L.; project administration, M.M.;

funding acquisition, A.S., N.L. and M.M. All authors have read and agreed to the published version of the manuscript.

Funding: A.S. is recipient of a MESRI (Ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation) fellowship.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The reported data are provided in the Supplementary Materials.

Acknowledgments: We would like to thank T. Langer and Inte:Ligand for the LigandScout 4.4 software license.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

AR	Androgen receptors
AUC	Area under the ROC curve
B	Binding compounds
CAS	Chemical Abstracts Service
DBD	DNA-binding domain
DNA	deoxyribonucleic acid
DSSTox	Distributed Structure-Searchable Toxicity
EADB	Estrogenic activity database
EDCs	Endocrine disrupting chemicals
EF	Enrichment factor
EPA	United states Environmental protection agency
ER	Estrogen receptors
FIX	Factor IX
GR	Glucocorticoid receptors
LB	Ligand based
LBD	Ligand binding domain
NB	Non-Binding compounds
NCTR	National center for toxicological research USA
NR	Nuclear receptor
NR-DBIND	Nuclear Receptors Database Including Negative Data
NTD	NH2-terminal domain
PC	Predictiveness curve
PDB	Protein data bank
PPAR	Peroxisome proliferator-activated receptors
PPV	Positive Predictive value
PLANTS	Protein-ligand ANTSystem
QSAR	Quantitative structure activity relationship
RMSD	Root-mean-square deviation
ROC	Receiver operating curve
SB	Structure based
SD	Standard deviation
Se	Sensitivity
SMILES	Simplified molecular-input line-entry system
Sp	Specificity
TH	scoring Threshold
TR	Thyroid hormones receptors

References

1. Brzozowski, A.M.; Pike, A.C.W.; Dauter, Z.; Hubbard, R.E.; Bonn, T.; Engström, O.; Öhman, L.; Greene, G.L.; Gustafsson, J.-Å.; Carlquist, M. Molecular basis of agonism and antagonism in the oestrogen receptor. *Nat. Cell Biol.* **1997**, *389*, 753–758. [CrossRef] [PubMed]
2. Jia, M.; Dahlman-Wright, K.; Gustafsson, J.Å. Estrogen receptor alpha and beta in health and disease. *Best Pract. Res. Clin. Endocrinol. Metab.* **2015**, *29*, 557–568. [CrossRef]
3. Matthews, J.; Gustafsson, J.-A. Estrogen Signaling: A Subtle Balance between ER Alpha and ER Beta. *Mol. Interv.* **2003**, *3*, 281–292. [CrossRef] [PubMed]
4. Tanenbaum, D.M.; Wang, Y.; Williams, S.P.; Sigler, P.B. Crystallographic comparison of the estrogen and progesterone receptor's ligand binding domains. *Proc. Natl. Acad. Sci. USA* **1998**, *95*, 5998–6003. [CrossRef]
5. Shao, W.; Brown, M. Advances in estrogen receptor biology: Prospects for improvements in targeted breast cancer therapy. *Breast Cancer Res.* **2003**, *6*, 39–52. [CrossRef] [PubMed]
6. Minutolo, F.; Macchia, M.; Katzenellenbogen, B.S.; Katzenellenbogen, J.A. Estrogen receptor β ligands: Recent advances and biomedical applications. *Med. Res. Rev.* **2009**, *31*, 364–442. [CrossRef]
7. Comparing Multiple Machine Learning Algorithms and Metrics for Estrogen Receptor Binding Prediction/Molecular Pharmaceutics. Available online: https://pubs.acs.org/doi/10.1021/acs.molpharmaceut.8b00546 (accessed on 2 November 2020).
8. Golden, R.J.; Noller, K.L.; Titus-Ernstoff, L.; Kaufman, R.H.; Mittendorf, R.; Stillman, R.; Reese, E.A. Environmental Endocrine Modulators and Human Health: An Assessment of the Biological Evidence. *Crit. Rev. Toxicol.* **1998**, *28*, 109–227. [CrossRef]
9. Schug, T.T.; Johnson, A.F.; Birnbaum, L.S.; Colborn, T.; Guillette, L.J., Jr.; Crews, D.P.; Collins, T.; Soto, A.M.; vom Saal, F.S.; McLachlan, J.A.; et al. Minireview: Endocrine Disruptors: Past Lessons and Future Directions. *Mol. Endocrinol.* **2016**, *30*, 833–847. [CrossRef]
10. Fillol, C.; Oleko, A.; Saoudi, A.; Zeghnoun, A.; Balicco, A.; Gane, J.; Rambaud, L.; Leblanc, A.; Gaudreau, É.; Marchand, P.; et al. Exposure of the French population to bisphenols, phthalates, parabens, glycol ethers, brominated flame retardants, and perfluorinated compounds in 2014–2016: Results from the Esteban study. *Environ. Int.* **2021**, *147*, 106340. [CrossRef] [PubMed]
11. Audouze, K.; Sarigiannis, D.; Alonso-Magdalena, P.; Brochot, C.; Casas, M.; Vrijheid, M.; Babin, P.J.; Karakitsios, S.; Coumoul, X.; Barouki, R. Integrative Strategy of Testing Systems for Identification of Endocrine Disruptors Inducing Metabolic Disorders—An Introduction to the OBERON Project. *Int. J. Mol. Sci.* **2020**, *21*, 2988. [CrossRef]
12. Johansson, H.K.L.; Svingen, T.; Fowler, P.A.; Vinggaard, A.M.; Boberg, J. Environmental influences on ovarian dysgenesis—Developmental windows sensitive to chemical exposures. *Nat. Rev. Endocrinol.* **2017**, *13*, 400–414. [CrossRef]
13. Ghassabian, A.; Trasande, L. Disruption in Thyroid Signaling Pathway: A Mechanism for the Effect of Endocrine-Disrupting Chemicals on Child Neurodevelopment. *Front. Endocrinol.* **2018**, *9*, 204. [CrossRef]
14. Cano-Sancho, G.; Salmon, A.G.; La Merrill, M.A. Association between Exposure to p,p′-DDT and Its Metabolite p,p′-DDE with Obesity: Integrated Systematic Review and Meta-Analysis. *Environ. Health Perspect.* **2017**, *125*, 096002. [CrossRef]
15. Kumar, M.; Sarma, D.K.; Shubham, S.; Kumawat, M.; Verma, V.; Prakash, A.; Tiwari, R. Environmental Endocrine-Disrupting Chemical Exposure: Role in Non-Communicable Diseases. *Front. Public Health* **2020**, *8*, 553850. [CrossRef] [PubMed]
16. Shanle, E.K.; Xu, W. Endocrine Disrupting Chemicals Targeting Estrogen Receptor Signaling: Identification and Mechanisms of Action. *Chem. Res. Toxicol.* **2010**, *24*, 6–19. [CrossRef]
17. Combarnous, Y.; Nguyen, T.M.D. Comparative Overview of the Mechanisms of Action of Hormones and Endocrine Disruptor Compounds. *Toxics* **2019**, *7*, 5. [CrossRef]
18. Balaguer, P.; Delfosse, V.; Bourguet, W. Mechanisms of endocrine disruption through nuclear receptors and related pathways. *Curr. Opin. Endocr. Metab. Res.* **2019**, *7*, 1–8. [CrossRef]
19. Schneider, M.; Pons, J.-L.; Labesse, G.; Bourguet, W. In Silico Predictions of Endocrine Disruptors Properties. *Endocrinology* **2019**, *160*, 2709–2716. [CrossRef]
20. Sun, L.; Yang, H.; Cai, Y.; Li, W.; Liu, G.; Tang, Y. In Silico Prediction of Endocrine Disrupting Chemicals Using Single-Label and Multilabel Models. *J. Chem. Inf. Model.* **2018**, *59*, 973–982. [CrossRef]
21. Kolšek, K.; Mavri, J.; Dolenc, M.S.; Gobec, S.; Turk, S. Endocrine Disruptome—An Open Source Prediction Tool for Assessing Endocrine Disruption Potential through Nuclear Receptor Binding. *J. Chem. Inf. Model.* **2014**, *54*, 1254–1267. [CrossRef]
22. Vedani, A.; Dobler, M.; Smieško, M. VirtualToxLab—A platform for estimating the toxic potential of drugs, chemicals and natural products. *Toxicol. Appl. Pharmacol.* **2012**, *261*, 142–153. [CrossRef]
23. Mayr, A.; Klambauer, G.; Unterthiner, T.; Hochreiter, S. DeepTox: Toxicity Prediction using Deep Learning. *Front. Environ. Sci.* **2016**, *3*. [CrossRef]
24. Banerjee, P.; Eckert, A.O.; Schrey, A.K.; Preissner, R. ProTox-II: A webserver for the prediction of toxicity of chemicals. *Nucleic Acids Res.* **2018**, *46*, W257–W263. [CrossRef]
25. Mansouri, K.; Abdelaziz, A.; Rybacka, A.; Roncaglioni, A.; Tropsha, A.; Varnek, A.; Zakharov, A.; Worth, A.; Richard, A.M.; Grulke, C.M.; et al. CERAPP: Collaborative Estrogen Receptor Activity Prediction Project. *Environ. Health Perspect.* **2016**, *124*, 1023–1033. [CrossRef]
26. Sander, T.; Freyss, J.; Von Korff, M.; Rufener, C. DataWarrior: An Open-Source Program for Chemistry Aware Data Visualization and Analysis. *J. Chem. Inf. Model.* **2015**, *55*, 460–473. [CrossRef]

27. Réau, M.; Lagarde, N.; Zagury, J.-F.; Montes, M. Hits Discovery on the Androgen Receptor: In Silico Approaches to Identify Agonist Compounds. *Cells* **2019**, *8*, 1431. [CrossRef]
28. Trott, O.; Olson, A.J. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **2010**, *31*, 455–461. [CrossRef] [PubMed]
29. Ben Nasr, N.; Guillemain, H.; Lagarde, N.; Zagury, J.-F.; Montes, M. Multiple Structures for Virtual Ligand Screening: Defining Binding Site Properties-Based Criteria to Optimize the Selection of the Query. *J. Chem. Inf. Model.* **2013**, *53*, 293–311. [CrossRef] [PubMed]
30. Craig, I.R.; Essex, J.W.; Spiegel, K. Ensemble Docking into Multiple Crystallographically Derived Protein Structures: An Evaluation Based on the Statistical Analysis of Enrichments. *J. Chem. Inf. Model.* **2010**, *50*, 511–524. [CrossRef]
31. Empereur-Mot, C.; Zagury, J.-F.; Montes, M. Screening Explorer–An Interactive Tool for the Analysis of Screening Results. *J. Chem. Inf. Model.* **2016**, *56*, 2281–2286. [CrossRef] [PubMed]
32. Empereur-Mot, C.; Guillemain, H.; Latouche, A.; Zagury, J.-F.; Viallon, V.; Montes, M. Predictiveness curves in virtual screening. *J. Chemin.* **2015**, *7*, 1–17. [CrossRef]
33. Gheyouche, E.; Launay, R.; Lethiec, J.; Labeeuw, A.; Roze, C.; Amossé, A.; Téletchéa, S. DockNmine, a Web Portal to Assemble and Analyse Virtual and Experimental Interaction Data. *Int. J. Mol. Sci.* **2019**, *20*, 5062. [CrossRef]
34. Danishuddin; Madhukar, G.; Malik, M.; Subbarao, N. Development and rigorous validation of antimalarial predictive models using machine learning approaches. *SAR QSAR Environ. Res.* **2019**, *30*, 543–560. [CrossRef]
35. Klingspohn, W.; Mathea, M.; Ter Laak, A.; Heinrich, N.; Baumann, K. Efficiency of different measures for defining the applicability domain of classification models. *J. Cheminf.* **2017**, *9*, 1–17. [CrossRef] [PubMed]
36. Myrianthopoulos, V.; Lozach, O.; Zareifi, D.; Alexopoulos, L.; Meijer, L.; Gorgoulis, V.G.; Mikros, E. Combined Virtual and Experimental Screening for CK1 Inhibitors Identifies a Modulator of p53 and Reveals Important Aspects of in Silico Screening Performance. *Int. J. Mol. Sci.* **2017**, *18*, 2102. [CrossRef]
37. Furlan, V.; Konc, J.; Bren, U. Inverse Molecular Docking as a Novel Approach to Study Anticarcinogenic and Anti-Neuroinflammatory Effects of Curcumin. *Molecules* **2018**, *23*, 3351. [CrossRef] [PubMed]
38. Réau, M.; Langenfeld, F.; Zagury, J.-F.; Lagarde, N.; Montes, M. Decoys Selection in Benchmarking Datasets: Overview and Perspectives. *Front. Pharmacol.* **2018**, *9*, 11. [CrossRef]
39. Onnis, V.; Kinsella, G.K.; Carta, G.; Fayne, D.; Lloyd, D.G. Rational ligand-based virtual screening and structure–activity relationship studies in the ligand-binding domain of the glucocorticoid receptor-α. *Futur. Med. Chem.* **2009**, *1*, 483–499. [CrossRef] [PubMed]
40. Taha, M.O.; Tarairah, M.; Zalloum, H.; Abu-Sheikha, G. Pharmacophore and QSAR modeling of estrogen receptor β ligands and subsequent validation and in silico search for new hits. *J. Mol. Graph. Model.* **2010**, *28*, 383–400. [CrossRef]
41. Verma, N.; Chouhan, U. Chemometric Modelling of PPAR-α and PPAR-γ Dual Agonists for the Treatment of Type-2 Diabetes. *Curr. Sci.* **2016**, *111*, 356. [CrossRef]
42. Lagarde, N.; Delahaye, S.; Zagury, J.-F.; Montes, M. Discriminating agonist and antagonist ligands of the nuclear receptors using 3D-pharmacophores. *J. Cheminf.* **2016**, *8*, 43. [CrossRef] [PubMed]
43. Pal, S.; Kumar, V.; Kundu, B.; Bhattacharya, D.; Preethy, N.; Reddy, M.P.; Talukdar, A. Ligand-based Pharmacophore Modeling, Virtual Screening and Molecular Docking Studies for Discovery of Potential Topoisomerase I Inhibitors. *Comput. Struct. Biotechnol. J.* **2019**, *17*, 291–310. [CrossRef] [PubMed]
44. Vittorio, S.; Seidel, T.; Germanò, M.P.; Gitto, R.; Ielo, L.; Garon, A.; Rapisarda, A.; Pace, V.; Langer, T.; De Luca, L. A Combination of Pharmacophore and Docking-based Virtual Screening to Discover new Tyrosinase Inhibitors. *Mol. Inform.* **2020**, *39*, e1900054. [CrossRef]
45. Li, P.; Peng, J.; Zhou, Y.; Li, Y.; Liu, X.; Wang, L.; Zuo, Z. Discovery of FIXa inhibitors by combination of pharmacophore modeling, molecular docking, and 3D-QSAR modeling. *J. Recept. Signal Transduct.* **2018**, *38*, 213–224. [CrossRef] [PubMed]
46. Wang, F.; Shi, Y.; Le, G. Statistical methods and molecular docking for the prediction of thyroid hormone receptor subtype binding affinity and selectivity. *Struct. Chem.* **2017**, *28*, 833–847. [CrossRef]
47. Lu, S.-H.; Wu, J.W.; Liu, H.-L.; Zhao, J.-H.; Liu, K.-T.; Chuang, C.-K.; Lin, H.-Y.; Tsai, W.-B.; Ho, Y. The discovery of potential acetylcholinesterase inhibitors: A combination of pharmacophore modeling, virtual screening, and molecular docking studies. *J. Biomed. Sci.* **2011**, *18*, 8. [CrossRef] [PubMed]
48. Capuzzi, S.J.; Epoliti, R.; Eisayev, O.; Efarag, S.; Etropsha, A. QSAR Modeling of Tox21 Challenge Stress Response and Nuclear Receptor Signaling Toxicity Assays. *Front. Environ. Sci.* **2016**, *4*, 4. [CrossRef]
49. Russo, D.P.; Zorn, K.M.; Clark, A.M.; Zhu, H.; Ekins, S. Comparing Multiple Machine Learning Algorithms and Metrics for Estrogen Receptor Binding Prediction. *Mol. Pharm.* **2018**, *15*, 4361–4370. [CrossRef] [PubMed]
50. Chang, Y.-H.; Chen, J.-Y.; Hor, C.-Y.; Chuang, Y.-C.; Yang, C.-B.; Yang, C.-N. Computational Study of Estrogen Receptor-Alpha Antagonist with Three-Dimensional Quantitative Structure-Activity Relationship, Support Vector Regression, and Linear Regression Methods. Available online: https://www.hindawi.com/journals/ijmc/2013/743139/ (accessed on 17 November 2020).
51. Bhhatarai, B.; Wilson, D.M.; Price, P.S.; Marty, S.; Parks, A.K.; Carney, E. Evaluation of OASIS QSAR Models Using ToxCast™ in Vitro Estrogen and Androgen Receptor Binding Data and Application in an Integrated Endocrine Screening Approach. *Environ. Health Perspect.* **2016**, *124*, 1453–1461. [CrossRef]

52. Rybacka, A.; Rudén, C.; Tetko, I.V.; Andersson, P.L. Identifying potential endocrine disruptors among industrial chemicals and their metabolites—Development and evaluation of in silico tools. *Chemosphere* **2015**, *139*, 372–378. [CrossRef]
53. Zorn, K.M.; Foil, D.H.; Lane, T.R.; Russo, D.P.; Hillwalker, W.; Feifarek, D.J.; Jones, F.; Klaren, W.D.; Brinkman, A.M.; Ekins, S. Machine Learning Models for Estrogen Receptor Bioactivity and Endocrine Disruption Prediction. *Environ. Sci. Technol.* **2020**, *54*, 12202–12213. [CrossRef]
54. Trisciuzzi, D.; Alberga, D.; Mansouri, K.; Judson, R.S.; Cellamare, S.; Catto, M.; Carotti, A.; Benfenati, E.; Novellino, E.; Mangiatordi, G.F.; et al. Docking-based classification models for exploratory toxicology studies on high-quality estrogenic experimental data. *Futur. Med. Chem.* **2015**, *7*, 1921–1936. [CrossRef]
55. Zhang, L.; Sedykh, A.; Tripathi, A.; Zhu, H.; Afantitis, A.; Mouchlis, V.D.; Melagraki, G.; Rusyn, I.; Tropsha, A. Identification of putative estrogen receptor-mediated endocrine disrupting chemicals using QSAR- and structure-based virtual screening approaches. *Toxicol. Appl. Pharmacol.* **2013**, *272*, 67–76. [CrossRef] [PubMed]
56. Ng, H.W.; Zhang, W.; Shu, M.; Luo, H.; Ge, W.; Perkins, R.; Tong, W.; Hong, H. Competitive molecular docking approach for predicting estrogen receptor subtype α agonists and antagonists. *BMC Bioinform.* **2014**, *15*, S4. [CrossRef] [PubMed]
57. Tan, H.; Wang, X.; Hong, H.; Benfenati, E.; Giesy, J.P.; Gini, G.C.; Kusko, R.; Zhang, X.; Yu, H.; Shi, W. Structures of Endocrine-Disrupting Chemicals Determine Binding to and Activation of the Estrogen Receptor α and Androgen Receptor. *Environ. Sci. Technol.* **2020**, *54*, 11424–11433. [CrossRef] [PubMed]
58. Balaguer, P.; Delfosse, V.; Grimaldi, M.; Bourguet, W. Structural and functional evidences for the interactions between nuclear hormone receptors and endocrine disruptors at low doses. *Comptes Rendus Biol.* **2017**, *340*, 414–420. [CrossRef]
59. Wassermann, A.M.; Bajorath, J.; Binding, D.B. ChEMBL: Online compound databases for drug discovery. *Expert Opin. Drug Discov.* **2011**, *6*, 683–687. [CrossRef]
60. Valsecchi, C.; Grisoni, F.; Motta, S.; Bonati, L.; Ballabio, D. NURA: A curated dataset of nuclear receptor modulators. *Toxicol. Appl. Pharmacol.* **2020**, *407*, 115244. [CrossRef]
61. Lagarde, N.; Delahaye, S.; Jérémie, A.; Ben Nasr, N.; Guillemain, H.; Empereur-Mot, C.; Laville, V.; Labib, T.; Réau, M.; Langenfeld, F.; et al. Discriminating Agonist from Antagonist Ligands of the Nuclear Receptors Using Different Chemoinformatics Approaches. *Mol. Inform.* **2017**, *36*, 1700020. [CrossRef] [PubMed]
62. Williams, A.J.; Grulke, C.M.; Edwards, J.; McEachran, A.D.; Mansouri, K.; Baker, N.C.; Patlewicz, G.; Shah, I.; Wambaugh, J.F.; Judson, R.S.; et al. The CompTox Chemistry Dashboard: A community data resource for environmental chemistry. *J. Chemin.* **2017**, *9*, 1–27. [CrossRef] [PubMed]
63. Richard, A.M.; Judson, R.S.; Houck, K.A.; Grulke, C.M.; Volarath, P.; Thillainadarajah, I.; Yang, C.; Rathman, J.F.; Martin, M.T.; Wambaugh, J.F.; et al. ToxCast Chemical Landscape: Paving the Road to 21st Century Toxicology. *Chem. Res. Toxicol.* **2016**, *29*, 1225–1251. [CrossRef]
64. Réau, M.; Lagarde, N.; Zagury, J.-F.; Montes, M. Nuclear Receptors Database Including Negative Data (NR-DBIND): A Database Dedicated to Nuclear Receptors Binding Data Including Negative Data and Pharmacological Profile. *J. Med. Chem.* **2019**, *62*, 2894–2904. [CrossRef]
65. Shen, J.; Xu, L.; Fang, H.; Richard, A.M.; Bray, J.D.; Judson, R.S.; Zhou, G.; Colatsky, T.J.; Aungst, J.L.; Teng, C.; et al. EADB: An Estrogenic Activity Database for Assessing Potential Endocrine Activity. *Toxicol. Sci.* **2013**, *135*, 277–291. [CrossRef]
66. Ng, H.W.; Perkins, R.; Tong, W.; Hong, H. Versatility or Promiscuity: The Estrogen Receptors, Control of Ligand Selectivity and an Update on Subtype Selective Ligands. *Int. J. Environ. Res. Public Health* **2014**, *11*, 8709–8742. [CrossRef]
67. ChemAxon—Software Solutions and Services for Chemistry & Biology. Available online: https://chemaxon.com/ (accessed on 27 July 2020).
68. Wolber, G.; Langer, T. LigandScout: 3-D Pharmacophores Derived from Protein-Bound Ligands and Their Use as Virtual Screening Filters. *J. Chem. Inf. Model.* **2004**, *45*, 160–169. [CrossRef]
69. RDKit. Available online: https://www.rdkit.org/ (accessed on 27 July 2020).
70. Pettersen, E.F.; Goddard, T.D.; Huang, C.C.; Couch, G.S.; Greenblatt, D.M.; Meng, E.C.; Ferrin, T.E. UCSF Chimera—A visualization system for exploratory research and analysis. *J. Comput. Chem.* **2004**, *25*, 1605–1612. [CrossRef] [PubMed]
71. Bank, R.P.D. RCSB PDB: Homepage. Available online: https://www.rcsb.org/ (accessed on 6 May 2020).
72. Koes, D.R.; Baumgartner, M.P.; Camacho, C.J. Lessons Learned in Empirical Scoring with smina from the CSAR 2011 Benchmarking Exercise. *J. Chem. Inf. Model.* **2013**, *53*, 1893–1904. [CrossRef] [PubMed]
73. Korb, O.; Stützle, T.; Exner, T.E. PLANTS: Application of Ant Colony Optimization to Structure-Based Drug Design. In *Ant Colony Optimization and Swarm Intelligence*; Dorigo, M., Gambardella, L.M., Birattari, M., Martinoli, A., Poli, R., Stützle, T., Eds.; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2006; Volume 4150, pp. 247–258. ISBN 978-3-540-38482-3.
74. Jain, A.N. Surflex: Fully Automatic Flexible Molecular Docking Using a Molecular Similarity-Based Search Engine. *J. Med. Chem.* **2003**, *46*, 499–511. [CrossRef]
75. Quiroga, R.; Villarreal, M.A. Vinardo: A Scoring Function Based on Autodock Vina Improves Scoring, Docking, and Virtual Screening. *PLoS ONE* **2016**, *11*, e0155183. [CrossRef]
76. Korb, O.; Stützle, T.; Exner, T.E. An ant colony optimization approach to flexible protein–ligand docking. *Swarm Intell.* **2007**, *1*, 115–134. [CrossRef]
77. Korb, O.; Stützle, T.; Exner, T.E. Empirical Scoring Functions for Advanced Protein−Ligand Docking with PLANTS. *J. Chem. Inf. Model.* **2009**, *49*, 84–96. [CrossRef] [PubMed]

78. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Pret-tenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
79. Kainrad, T.; Hunold, S.; Seidel, T.; Langer, T. LigandScout Remote: A New User-Friendly Interface for HPC and Cloud Resources. *J. Chem. Inf. Model.* **2018**, *59*, 31–37. [CrossRef] [PubMed]
80. Vuorinen, A.; Schuster, D. Methods for generating and applying pharmacophore models as virtual screening filters and for bioactivity profiling. *Methods* **2015**, *71*, 113–134. [CrossRef] [PubMed]

2.1.3. Discussion

2.1.3.1. Choix des méthodes

L'objectif de cette étude était de réaliser une preuve de concept sur la capacité des méthodes de docking et des modèles de pharmacophores LB et SB à distinguer entre composés « Binders » (B) et « Non-Binders » (NB) des NR dans la perspective de prédire leur potentiels PE. Pour cette étude, le postulat de départ était qu'un composé identifié comme étant capable de se lier aux NR était un potentiel perturbateur endocrinien et devait donc être plus amplement étudié pour sa toxicité potentielle. Cette preuve de concept a été réalisée sur le récepteur ERα, dont l'implication dans le mécanisme de perturbation endocrinienne a été démontrée[335]. De nombreuses études *in silico* ont déjà été menées dans ce sens et avec ce même récepteur (c.f. paragraphe 1.1). La revue recensant celles menées entre 2010 et 2020 liste 27 études sur ERα dont la large majorité (21) a utilisé des méthodes de QSAR. Ceci est cohérent avec la popularité de ces méthodes en toxicologie prédictive. Toutefois, elles ont certaines limites dont la principale est leur dépendance aux données de départ, leur pertinence limitée à leur domaine d'applicabilité [311] mais aussi parfois une difficulté à traduire les modèles d'un point de vue mécanistique. Nous avons donc choisi d'utiliser des méthodes LB (modèles de pharmacophores) et SB (docking et pharmacophores) complémentaires pour créer notre preuve de concept.

2.1.3.2. Différences entre les composés PE et les composés thérapeutiques

Les PE, contrairement aux composés thérapeutiques, ne sont pas conçus et optimisés pour interagir directement avec les NR et ils présentent donc une grande variabilité structurale [311]. En conséquence, pour la réalisation des méthodes de pharmacophores, les données de départ ont été séparées en groupes distincts (ou clusters) selon leur similarité de structure pour pouvoir générer les différents modèles pharmacophoriques représentatifs de chaque groupe.

2.1.3.3. Performances obtenues

Le succès des méthodes de docking dans la prédiction de la toxicité de composés a déjà été démontré et sa place est confirmée comme un système d'experts de référence [12]. Pour cette preuve de docking, nous avons évalué différents protocoles afin de choisir le plus adapté à la prédiction des ligands des NR. Pour cela, nous avons comparé différents logiciels de docking, différentes fonctions de score et différentes approches (une seule structure de ERα vs un ensemble de structure). Nous avons aussi cherché comment choisir rationnellement un score de

docking seuil pour distinguer les molécules prédites comme actives des molécules prédites comme inactives. Finalement, le protocole retenu est celui du docking à une structure (*single structure docking*) par opposition au docking d'ensemble qui n'a pas entrainé de grandes améliorations au niveau des performances.

Nous pouvons comparer ce protocole à celui utilisé par Endocrine disruptome [324], un outils disponible gratuitement en ligne, qui est basé sur le même mécanisme d'action des PE, le mécanisme direct. Il utilise aussi une méthode de docking, plus précisément le logiciel AutoDock Vina, pour choisir une structure de référence pour chacun des NR étudié en se basant sur les AUC des courbes ROC pour chaque structure de chaque NR étudié. Tout comme dans notre protocole, Endocrine Disruptome fonctionne à l'aide d'un score seuil de docking choisi pour différentes valeurs de sensibilité 0,25 – 0,5 – 0,75 correspondant respectivement à des scores de docking de -9,3, -8,8 et -8,2. Ainsi, avec Endocrine disruptome, les scores de docking obtenus pour chaque composé permettent de les classer en 4 catégories : rouge, orange, jaune et verte. La classe rouge correspond à la catégorie la plus probable d'inclure des PE et vert la catégorie dites « sans danger » incluant les composés à faible risque d'être des PE. Toutefois, Endocrine disruptome présente des faiblesses qui nous ont poussé à proposer notre propre modèle plutôt que d'utiliser directement cet outil disponible en libre accès. En effet, le choix du protocole de docking présenté dans Endocrine disruptome est basé sur des données de la DUD-E et de la ChEMBL et contient donc des composés inactifs supposés (*decoys*) et des composés bioactifs à caractères *drug-like* à la différence des données de l'EPA utilisées au cours de cette étude incluant des composés toxiques et suspectés dans les mécanismes de toxicité. De plus, endocrine disruptome ne permet pas d'automatiser les sélections sur une liste de composés, mais nécessite la recherche pour chaque composé de façon distincte. En ce qui concerne les performances de prédiction, même s'il est difficile de comparer directement nos performances aux leurs car les jeux de données ne sont pas les mêmes, nous avons remarqué que les AUCs pour ERα étaient comparables à celles que nous avons obtenu avec le logiciel Smina.

Par ailleurs, des modèles de pharmacophores SBLB ont été générés et affinés grâce à un protocole d'optimisation qui a permis d'obtenir de fortes Sp (0,95 et 0,97 sur les jeux d'entrainement et de test) et des Se moyennes de 0,23 et 0,32. Malgré la popularité des modèles de pharmacophores, il n'existe pas à notre connaissance d'outils qui l'utilise dans un but de criblage des PE et qu'on pourrait utiliser pour comparer nos performances.

Compte tenu de la bonne Se des modèles de docking et de la haute Sp des modèles de pharmacophores, nous avons pensé à les combiner afin de tirer profit de leur combinaison afin d'enrichir nos résultats en palliant les déficits de chacun. Grâce au protocole de combinaison sélectionné, nous avons pu atteindre une Se de 0,8.

2.1.3.4. Composition du jeu de données

Parmi les profils pharmacologiques des composés issus de l'EPA et inclus dans notre jeu de données, nous retrouvons 29% de composés agissant comme agoniste et antagoniste à la fois. Notons que les tests considérés sont des tests à gène rapporteur analysés dans un seul sens pour chaque (*single endpoint*). Un test est donc effectué pour évaluer le profil agoniste des ligands en recherchant un éventuel gain de signal pour lequel le ligand est classifié comme agoniste ou non. Ce test est répété pour cette fois évaluer le profil antagoniste des ligands en évaluant la perte de signal qui est interprétée comme le reflet de l'antagonisme du composé. Cela suggèrerait que ces PE i.e. positifs pour les deux tests, peuvent être aussi bien agonistes qu'antagonistes. Ceci est en accord avec les données de la littérature qui affirment que certains PE impliqués dans le mécanisme d'action direct pour le ER agissait en tant que modulateur (SERM : selective estrogen receptor modulator) agissant en tant qu'agonistes pour certains tissus et antagonistes pour d'autres. L'exemple le plus connu des SERM est le tamoxifène, un médicament utilisé contre le cancer du sein médié par les ERα. Le tamoxifène agit comme antagoniste du ERα au niveau des cellules mammaires inhibant ainsi la progression des cellules cancéreuses, et agit comme agoniste de ce même récepteur au niveau de l'endomètre contribuant à augmenter le risque de cancers de l'endomètre après une exposition prolongée [338].

3.1.3. Analyse critique et perspectives

3.1.3.4.Protocole

Bien que le protocole de docking soit automatisé, ce n'était pas le cas complètement pour le protocole des modèles pharmacophoriques. Plus spécifiquement, l'étape d'optimisation était faite manuellement en se basant sur les recommandations de la littérature et en suivant un protocole précédemment développé au laboratoire [339]. Cela pose deux problèmes, le premier est le biais apporté par l'intervention humaine lié aux erreurs de manipulation ou aux oublis en plus de l'impossibilité d'assurer avec certitude la reproductibilité de ce protocole. Le deuxième problème est l'incapacité à généraliser ce protocole pour un plus grand nombre de récepteurs

comme nous entendons de le faire par la suite. Ainsi, un protocole automatisé a été mis au point après cette preuve de concept.

3.1.3.5.Métriques d'évaluation

Afin d'évaluer la performance de nos différents modèles, nous nous sommes basés sur deux métriques couramment utilisées qui sont la sensibilité (Se) et la spécificité (Sp). Toutefois, ces métriques ne prennent pas en compte un déséquilibre potentiel des bases de données entre composés actifs et inactifs et d'autres métriques devront être considérées pour apprécier à la fois la capacité de distinguer entre B et NB mais aussi à correctement retrouver les actifs parmi la totalité des B et de même pour les NB. Par exemple, afin de choisir le meilleur protocole, l'équipe à l'origine de Endocrine disruptome [324], en plus de s'intéresser aux Se et Sp, a privilégié la NPV (*negative predictive value*) avec le justificatif que ce paramètre était plus important en toxicologie car il reflète la capacité du protocole à correctement éliminer les composés non toxiques. Dans la suite de ce travail et la généralisation du protocole à d'autres NR, nous avons donc élargi la gamme des métriques utilisées pour évaluer les performances.

3.1.3.6.Prise en compte de la flexibilité

Habituellement, les études de docking réalisées se basent sur des outils où le ligand est flexible mais la protéine est considérée comme rigide. Hors l'importance de la prise en compte de la flexibilité de la protéine a été démontrée, surtout pour les études relatives aux NR [311].

Il est possible de prendre en compte la flexibilité de façon implicite dans le cadre des méthodes de docking (c.f. paragraphe 3.2.3.2.2.3 Prise en compte de la flexibilité). Parmi ces méthodes, nous avons exploré l'approche de docking d'ensemble lors de nos travaux. Pour ce faire, nous avons uniquement considéré les 7 structures d'ERα choisies pour réaliser les docking sur des critères connus comme pouvant affecter les performances des méthodes de docking (mutation, structure apo vs structure holo etc…). Or, il a été suggéré que considérer la totalité des structures cristallisées connues permettaient d'échantillonner la totalité de l'espace conformationnel du LBD du récepteur ERα [337]. Nous aurions alors pu inclure aussi par exemple les structures apo pour l'approche de docking d'ensemble.

Parmi les autres approches, le docking à résidus flexible permet à l'utilisateur de définir les résidus du site actif pour lesquels il souhaite explorer l'influence de la mobilité sur la performance au cours de la simulation de docking. Ainsi, le logiciel Smina utilisé au laboratoire propose la possibilité de faire cela en suggérant les résidus à explorer à une distance spécifiée autour du ligand, ou en permettant à l'utilisateur de faire sa propre sélection. Cependant, les

temps de calcul associés sont proportionnels au nombre de résidus flexibles sélectionnés et peuvent donc rapidement devenir incompatible avec le criblage d'une grande base de données. De ce fait, dans le cas où des résidus dont la flexibilité est importante à prendre en compte sont connus et documentés, il est possible d'utiliser directement cette approche. Dans le cas contraire, un protocole devra être monté afin de (1) identifier les résidus dont la flexibilité influence les performances de docking et (2) déterminer quelle(s) combinaison(s) sont les plus pertinentes pour cela. Cette approche est actuellement étudiée au laboratoire dans le cadre d'un travail de thèse.

3.1.4. Conclusion

Ce travail nous a permis de valider la conception d'un protocole associant les méthodes LB et SB afin de discriminer les composés capables de se lier aux NR de ceux qui ne le sont pas. Les performances des modèles individuels et combinés nous ont permis d'identifier la complémentarité des deux méthodes pour notre étude : les méthodes de docking étaient moins spécifiques mais atteignait des valeurs satisfaisantes de sensibilité ; les modèles de pharmacophores étaient caractérisés par une forte spécificité. Ainsi, la combinaison des deux modèles a permis d'améliorer les performances et le modèle choisi était un modèle de prédictions consensus qui permettait ainsi de prendre en considération le contexte toxicologique dans lequel s'inscrit l'étude. De plus, les modèles de pharmacophores permettent de relever les propriétés physicochimiques importantes pour la liaison aux ERα et les méthodes de docking permettent ainsi d'apprécier l'affinité de la molécule au niveau du site d'action. Si un composé est prédit B par les deux méthodes, la probabilité pour que ce dernier le soit, augmente par rapport des modèles individuels [67].

L'analyse critique de notre étude a permis de relever certains points à améliorer. Cela a conduit pour la généralisation à d'autres NR à la mise en place d'un protocole d'optimisation automatisé pour les modèles de pharmacophores ainsi qu'une prise en compte de différentes autres métriques pertinentes pour évaluer les performances.

2.1. Généralisation du protocole à d'autres récepteurs nucléaires

2.1.1. Introduction

Bien que les premières hypothèses concernant les PE incriminent les récepteurs nucléaires aux hormones sexuelles [335], il s'est avéré que les PE pouvaient aussi agir par l'intermédiaire d'autres NR. Au cours de la suite de ces travaux, nous avons choisi d'étendre l'application du protocole précédemment établi lors de la preuve de concept à deux autres récepteurs aux hormones sexuelles à savoir ERβ et AR en plus de deux autres récepteurs hormonaux, TR et GR et un autre isoforme de PPAR incriminés aussi dans le mécanisme de la perturbation endocrinienne à savoir PPARγ[340].

Ainsi, des jeux de données pour chacun des récepteurs considéré a été préparé à partir du tableau de bord Comptox contenant les composés testés pour différentes activités biologiques y compris la capacité à se lier aux NR. Différents protocoles, logiciels et approches de dockings ont été explorés pour sélectionner un modèle par NR. De plus des modèles SBLB pharmacophoriques ont été générés et affinés grâce à un nouveau protocole d'optimisation automatisé. Les performances individuelles et en combinaison des modèles ont été évalués pour sélectionner finalement le meilleur modèle pour chaque NR. Dans cette partie, nous présentons les différents protocoles suivis et les résultats engendrées. La partie discussion aborde les différents aspects de la modélisation et enfin l'analyse critique relève les points faibles du travail et apporte des perspectives et des pistes à explorer pour l'améliorer

2.1.2. Matériel et méthodes

2.1.2.1. Bases de données

Les composés ainsi que leurs données biologiques associées ont été téléchargés depuis le tableau de bord de la Comptox [333]. Ce dernier permet l'accession aux DSSTox qui présentent les résultats de tests biologiques et cellulaires contre plusieurs cibles biologiques dont les NR[89] pour des composés dangereux ou suspectés d'être à risque pour l'Homme et l'environnement. En utilisant les données disponibles dans le Comptox Dashboard pour AR, ERα, ERβ, GR, PPARγ et TRα, 6 jeux de données ont été générés en sélectionnant uniquement les composés pour lesquels les résultats des tests de liaison (*binding*) étaient disponibles pour les NR recherchés. Les composés actifs ont été directement annotés comme B pour les composés capables de se lier à la cible étudiée et NB pour les composés qui ne le sont pas. En outre, les

résultats des tests d'agonisme et d'antagonisme ont été utilisés pour annoter le profil pharmacologique des B lorsque ces données étaient disponibles. Chacun des 6 jeux de données a été soumis au même protocole de préparation. Tout d'abord, les SMILES manquants ont été générés à partir des noms IUPAC en utilisant la suite Chemaxon [341] et les composés pour lesquels il était impossible de générer les SMILES ont été écartés. Ensuite, toutes les structures ont été normalisées à l'aide du Standardizer de la suite Chemaxon. Les sels ont été supprimés, les fragments éliminés, la stéréochimie absolue et la tautomérie ont été prises en compte dans la structure. Ensuite, tous les doublons ont été supprimés. Tous les composés ont été protonés à pH 7,4 à l'aide de la commande cxcalc des *Calculator Plugins* de la suite Chemaxon. Les conformations 3D de chaque molécule ont été générées à l'aide du générateur de conformation iCon intégré à LigandScout. Tous les formats moléculaires appropriés ont été générés pour chaque base de données en fonction des besoins de la méthode utilisée : pdbqt et mol2 pour le docking, sdf et ldb pour les modèles de pharmacophores. Ces 6 jeux de données seront appelés bases de données de l'EPA par la suite. Différentes propriétés et descripteurs physico-chimiques ont été calculées avec le logiciel dragon [80] et utilisées pour visualiser la distribution des B et NB dans l'espace. La liste est la suivante : Poids moléculaire, ClogP, ClogS, nombre de donneurs de liaison H, nombre d'accepteurs de liaisons H, *Total surface area* (TSA), *Polar surface area* (PSA), nombre d'atomes sp3, nombre d'atomes symétriques, nombre d'amides, nombre d'amines, nombre d'alkyl amines, nombre d'amines aromatiques, nombre d'atomes d'azote aromatiques et nombre d'atomes d'oxygène acides.

En plus de ces jeu de l'EPA, 6 ensembles de validation pour les mêmes NR, i.e. AR, ERα, ERβ, GR, PPARγ et TRα, ont été préparés à partir de la base de données NR-DBIND[216], une base de données non-commerciale, qui fournit des données d'affinité et d'activité pour des petites molécules testées expérimentalement contre 28 NR, y compris des données négatives. Les ensembles de données correspondants ont été directement téléchargés du site web (http://nr-dbind.drugdesign.fr/) au format SMILES. Les composés pour lesquels aucune donnée de liaison n'était disponible ont été supprimés et les ensembles B et NB ont été définis en fonction du profil de *binding*. Les jeux de données ont été préparés selon le même protocole que celui décrit précédemment pour les jeux de données de l'EPA. Ces jeux de données seront désignés par le nom de jeux NR-DBIND.

2.1.2.2. Structures

Pour chacun des 6 NR étudiés, (1) toutes les structures humaines disponibles, (2) sans mutation ni délétion de résidus au niveau du LBD et (3) référencées par un article scientifique ont été utilisées. Pour le protocole de docking, les structures ont été directement téléchargées de la PDB et seules les structures sans délétions de résidus au niveau du site d'action ont été retenues. Les identifiants des structures utilisées pour les deux méthodes sont listés en annexe .

2.1.2.3. Protocole et modèles de docking

Tous les calculs de docking ont été effectués avec 3 logiciels smina [196], PLANTS[342] et Surflexdock[155]. Smina est une branche d'AutoDock Vina[343], conçu pour la personnalisation de fonctions de scoring. L'algorithme d'échantillonnage, hérité d'AutoDock Vina, utilise une succession d'étapes de mutations stochastiques pour générer des poses de liaison. Quatre fonctions de scores intégrées sont disponibles dans smina et sont utilisées ici : vina[343], Vina RaDii Optimized (vinardo) [344], dkoes [196] et ad4 [345]. Tous les dockings ont été réalisés en utilisant les paramètres par défaut de smina. Le site de liaison et les coordonnées de la boîte ont été déterminés pour chaque NR à partir des structures déjà alignées. Les paramètres de la boîte ont été définis en fonction de la position du ligand co-cristallisé avec un espacement de 1 Angstrom. Une boîte cubique a été délimitée avec une taille_x, une taille_y et une taille_z fixées à 20 et un centre_x, un centre_y et un centre_z appropriés pour chaque NR. Pour le logiciel PLANTS, l'algorithme d'échantillonnage portant le même nom PLANTS (*Protein-Ligand ANT System*) est basé sur l'optimisation par colonies de fourmis. Pour cet algorithme, les molécules sont assimilées à une colonie de fourmis artificielles qui doivent trouver la conformation à plus basse énergie du ligand au niveau du récepteur. En effectuant la recherche conformationnelle, une « traînée de phéromone » est libérée à chaque fois qu'une conformation idéale à faible énergie est trouvée. Ce marquage est modifié de manière itérative jusqu'à ce que la conformation de plus basse énergie soit trouvée. Les coordonnées du site de liaison sont les mêmes que celles utilisées pour smina. En ce qui concerne les autres paramètres, le rayon du site de liaison a été fixé à 18, le *cluster_structures* à 10, le *cluster_RMSD* à 2, et la vitesse de recherche à "*speed2*". Surflex-dock est un outil de docking qui combine la fonction de score empirique de Hammerhead et une méthode d'échantillonnage basée sur une construction incrémentale et un assemblage de fragments similaire à l'algorithme génétique. Surflex-dock utilise une pseudo-molécule, un *protomol*, comme cible pour aligner les fragments de ligands. Les *protomols* ont été générés à partir des structures *holo*.

Les approches docking à structure unique ou « *single structure docking* » et docking d'ensemble ou « *ensemble structure docking* » ont été explorées pour chaque cible en utilisant les structures PDB déjà sélectionnées. Au cours de l'approche *single structure docking*, chaque structure PDB a été évaluée individuellement grâce à l'aire sous la courbe ROC (AUC). Les valeurs AUC ont été calculées avec python 3.8 en utilisant la bibliothèque scikitlearn[346]. Dans l'approche de docking d'ensemble, les résultats du *single structure docking* pour chaque structure sont post-traités pour chaque ligand afin de ne conserver que la pose associée au meilleur score parmi toutes les structures PDB de la requête. Tous les ligands sont ensuite classés en fonction des scores sélectionnés et les AUC correspondants sont calculés. Seuls les ensembles de 2 et 3 structures ont été considérés suivant les recommandations précédentes[199,347].

La courbe de prédictivité (PC) a été utilisée pour définir un seuil de score (TH) capable de discriminer les composés B des composés NB. Utilisée à l'origine pour l'épidémiologie clinique, Empereur et. al [348] ont transposé l'utilisation de la PC au domaine de la chémoinformatique pour évaluer le pouvoir prédictif d'une méthode de criblage. À chaque score correspond une probabilité, appelée prédictivité (P), pour qu'un composé docké avec ce score donné soit un composé B. Pour chaque P, la sensibilité (Se) et la spécificité (Sp) associées peuvent être calculées et utilisées pour sélectionner le meilleur TH. De plus, l'évolution de plusieurs métriques (Se, Sp, PPV, NPV, Accuracy, score F1 et MCC) a été évaluée en fonction de tous les scores moyens pour chaque base de données afin de s'assurer que le TH sélectionné combine l'équilibre parfait entre toutes ces métriques.

Le modèle final de docking sélectionné pour chaque NR est constitué d'une ou plusieurs structure(s) protéique(s), avec des paramètres de docking optimisés (coordonnées de la boîte englobante, fonction de score et TH de score de docking). Ainsi, lorsqu'une molécule devra être prédite pour sa capacité à se lier à un NR, elle sera dockée avec les paramètres sélectionnés contre la cible définie et évaluée avec la fonction de score sélectionnée. Si elle obtient un meilleur score que le TH, elle sera prédite comme B. Sinon, elle sera prédite comme NB.

2.1.2.4. Modèles de pharmacophore

Les bases de données EPA d'une part et les structures PDB sélectionnées d'autre part ont été utilisées pour générer respectivement des modèles de pharmacophore LB et SB, à l'aide du logiciel LigandScout version 4.5.[56] Pour le modèle pharmacophore LB, chaque jeu de composés

a été divisé en un jeu d'entrainement et un jeu de test suivant un ratio de 75%/25% . Les composés de l'ensemble d'entraînement ont été regroupés sur la base de leurs similarités structurelles en utilisant l'outil "*icluster*" de LigandScout en gardant les paramètres par défaut. Des modèles pharmacophoriques ont été générés pour chaque groupe et utilisés pour cribler l'ensemble d'entraînement correspondant. Pour éviter tout biais causé par la séparation des données, ces 3 étapes ont été répétées 25 fois, après quoi les performances individuelles et générales des pharmacophores ont été évaluées pour ne conserver que l'itération donnant les meilleures performances. Les pharmacophores 3D SB ont été générés automatiquement à partir des structures PDB *holo* sélectionnées pour chaque NR avec le même logiciel.

Chacun des pharmacophores LB et SB générés pour chaque NR ont ensuite été optimisés individuellement selon plusieurs recommandations de la littérature [60,60,199,339,349] (Figure 25 : Protocole d'optimisation des modèles de pharmacophores. Pour chaque pharmacophore, avec un nombre initial de points pharmacophoriques T, un premier criblage a été effectué en fixant un nombre maximum de points omis (OMF pour *Maximum Omitted features*) à 0. Cela signifie qu'un composé doit correspondre à tous les points du pharmacophore pour être considéré comme un hit et les pharmacophores générant une liste de hits vide ont été éliminés. Ensuite, pour chacun des pharmacophores restants, un deuxième tour de criblage a été effectué avec le nombre de OMF fixé à T- 3. Cela permet de générer toutes les combinaisons possibles de points pharmacophoriques ainsi que leurs résultats associés. Chaque pharmacophore correspondant à l'une de ces combinaisons de points a été généré et utilisé pour cribler à nouveau l'ensemble d'entraînement avec le nombre OMF à nouveau fixé à 0. Si dans la liste de résultats le ratio B/NB est égal ou supérieur à 1, le modèle de pharmacophore est conservé, sinon il était éliminé. A la fin du processus d'optimisation, les pharmacophores redondants, c'est-à-dire ceux qui retrouvent des composés déjà retrouvés par d'autres pharmacophores du même groupe, ont été supprimés. Le même protocole d'optimisation a été appliqué pour les pharmacophores SB et LB, à la différence que pour les LB, une étape préliminaire supplémentaire a été effectuée pour définir comme "obligatoires" les points pharmacophoriques originairement "optionnels" pour éliminer le bruit [350]. Pour chaque NR, les modèles de pharmacophores SB et LB uniques optimisés ont ensuite été combinés pour former un groupe unique de pharmacophores SBLB après élimination des modèles redondants entre les deux groupes i.e modèles LB et modèles SB.

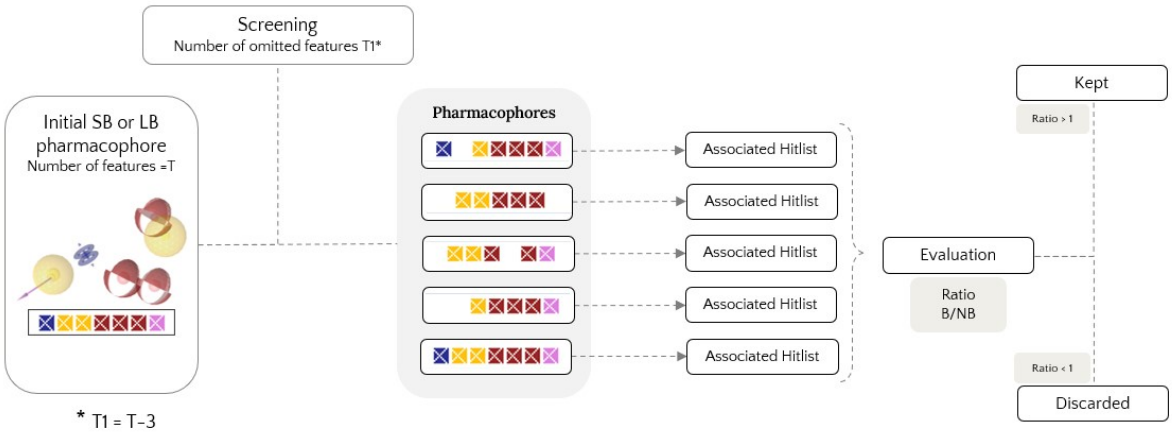


Figure 25 : Protocole d'optimisation des modèles de pharmacophores

2.1.2.5. Protocole de combinaison et sélection des modèles

Une fois les modèles de docking et de pharmacophores générés, la combinaison des deux méthodes a été évaluée pour chaque NR. De manière similaire à ce qui a été fait dans nos travaux précédents [347], les protocoles consensus et hiérarchiques ont été évalués (**Figure 26**). Dans le protocole de consensus, un composé est prédit comme B pour un NR s'il est prédit ainsi par au moins un des modèles construit pour ce NR (docking ou pharmacophores). Inversement, pour le protocole hiérarchique, seuls les composés prédits comme B par les modèles de pharmacophores et de docking à la fois seront prédits comme B.

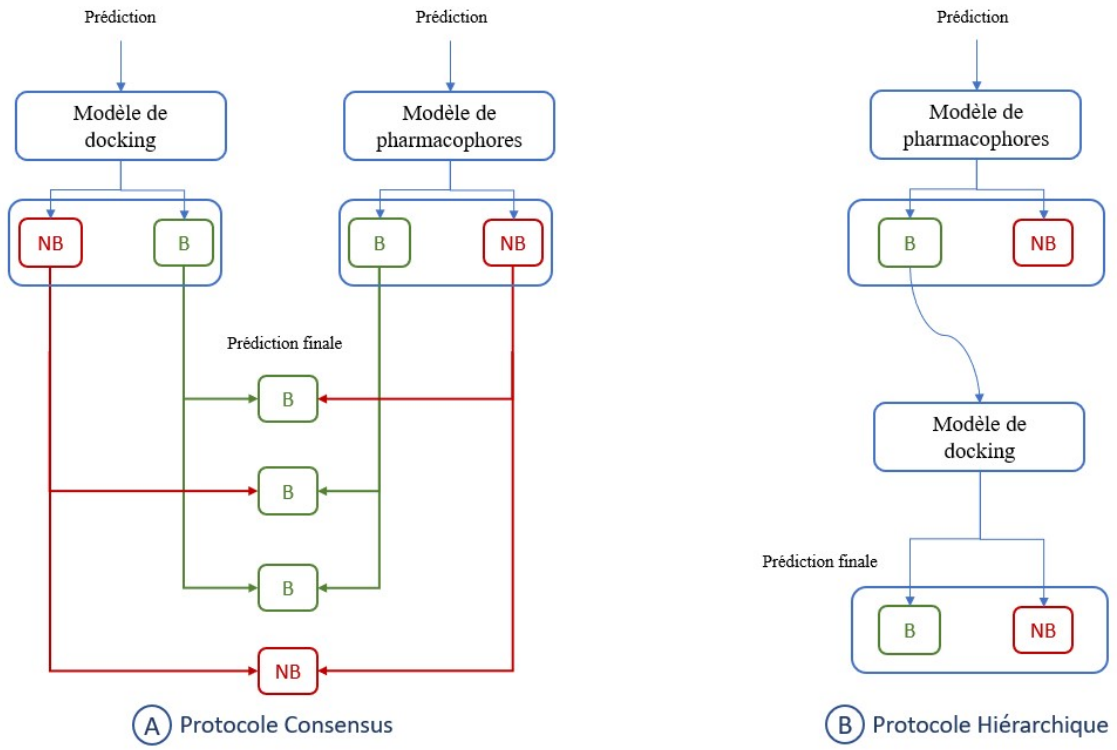


Figure 26 : Schématisation des prédictions réalisées selon le protocole Consensus (A) et hiérarchique en (B)

2.1.2.6. Métriques d'évaluations

Tous les modèles ont été conçus dans le but de pouvoir distinguer les B des NB pour les différents NR étudiés. Afin d'évaluer les performances des modèles à atteindre ce but, les prédictions obtenues pour chaque composé sont décrites comme l'une de ces 4 catégories de la matrice de contingence.

Vrai positif (VP), c'est-à-dire B confirmé expérimentalement et correctement prédit comme B.

Vrai négatif (VN) c.-à-d. NB confirmé expérimentalement correctement prédit comme NB

Faux positif (FP) : NB confirmé expérimentalement et prédit à tort comme B.

Faux négatif (FN), c'est-à-dire B confirmé expérimentalement prédit à tort comme NB.

Les métriques suivantes ont été calculées pour évaluer la qualité des prédictions :

- Sensibilité (*Se*), également appelée rappel, dont les valeurs se situent dans l'intervalle [0,1]. La *Se* représente la probabilité pour le modèle de prédire des

composés comme B pour des composés qui sont confirmés expérimentalement B (VP).

$$Se = \frac{VP}{VP + FN}$$

- La spécificité (*Sp*) , avec des valeurs comprises dans l'intervalle [0,1]. La *Sp* représente la probabilité que le modèle prédise les composés comme NB pour les composés qui sont confirmés expérimentalement NB (VN).

$$Sp = \frac{VN}{VN + FP}$$

- La valeur prédictive positive (PPV), également appelée précision, dont les valeurs sont comprises dans l'intervalle [0,1].

$$PPV = \frac{VP}{VP + FP}$$

- La valeur prédictive négative (NPV) , comprise aussi entre [0,1]

$$NPV = \frac{VN}{VN + FN}$$

- L'exactitude (*Acc*), dont les valeurs se situent dans l'intervalle [0,1]. L'*Acc* permet d'apprécier le rapport entre les composés correctement prédits et tous les composés prédits.

$$Acc = \frac{VP + VN}{VP + VN + FP + FN}$$

- Le score F1 (F1) dont les valeurs se situent dans l'intervalle [0,1]. F1 est défini comme la moyenne harmonique de la précision et du rappel.

$$F1 = \frac{2 * VP}{2 * VP + FP + FN} = 2 * \frac{PPV * Se}{PPV + Se}$$

- Le coefficient de corrélation de Matthews (MCC) dont les valeurs se situent dans l'intervalle [-1,1]. Les valeurs MCC de -1 et +1 sont atteintes lorsque les deux catégories (B et NB) sont respectivement parfaitement mal classées et

parfaitement classées. La MCC atteint 0 lorsque la classification est similaire à l'aléatoire.

$$MCC = \frac{VP * VN - FP * FN}{\sqrt{(VP + FP) * (VP + FN) * (VN + FP) * (VN + FN)}}$$

2.1.3. Résultats

2.1.3.1. Données

Le tableau de bord Comptox a été filtré pour chaque NR étudié pour ne garder que les composés testés pour le *binding*. Au total, 6 jeux de données ont été constitués pour AR, ERα, ERβ, GR, PPARγ et TRα comprenant entre 54 B pour TRα et 371 B pour AR (**Tableau 4**).

Tableau 4 : Composition des différents jeux de données de l'EPA

NR	Binders (B)	Non binders (NB)
AR	371	1401
ERα	223	2219
ERβ	242	1495
GR	266	126
PPARγ	108	1653
TRα	54	168

Différentes propriétés physico-chimiques et descripteurs (listés dans le paragraphe II.2.1.2. matériel et méthodes) ont été calculées et utilisées pour visualiser la distribution des B et NB dans l'espace (**Figure 27 :** Distribution des composés des différents jeux de données en fonction des propriétés physico-chimiques. Les composés en vert représentent les B (actifs) et en rouge les NB (inactifs)).

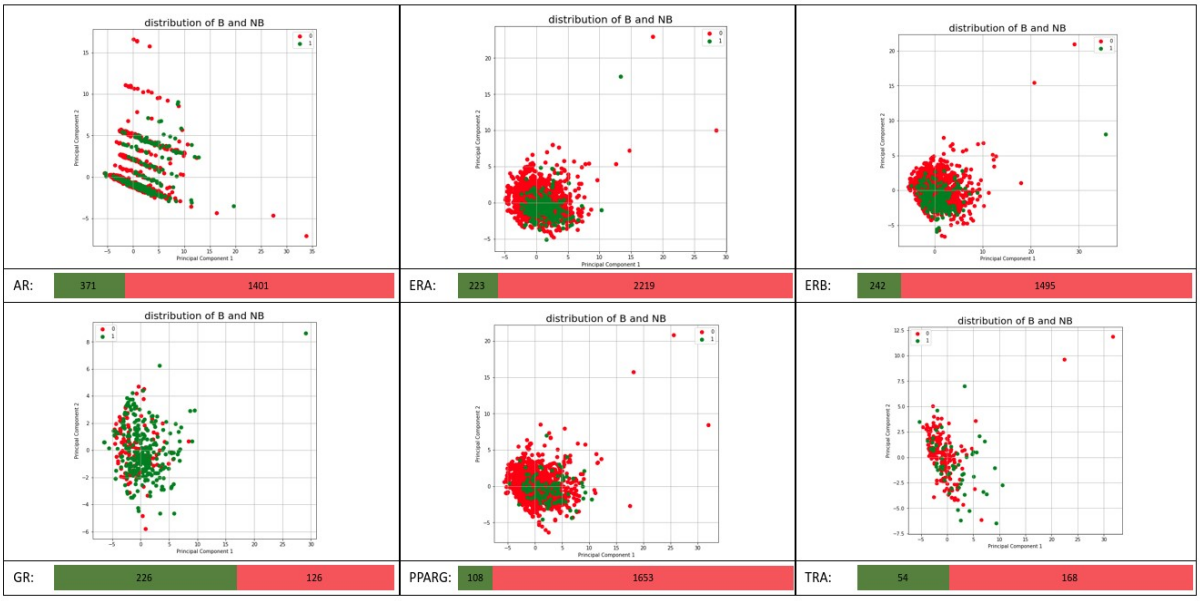


Figure 27 : Distribution des composés des différents jeux de données en fonction des propriétés physico-chimiques. Les composés en vert représentent les B (actifs) et en rouge les NB (inactifs)

La composition des différents jeux en B a été étudiée et les résultats sont présentés au niveau de l'*Upsetplot* suivant

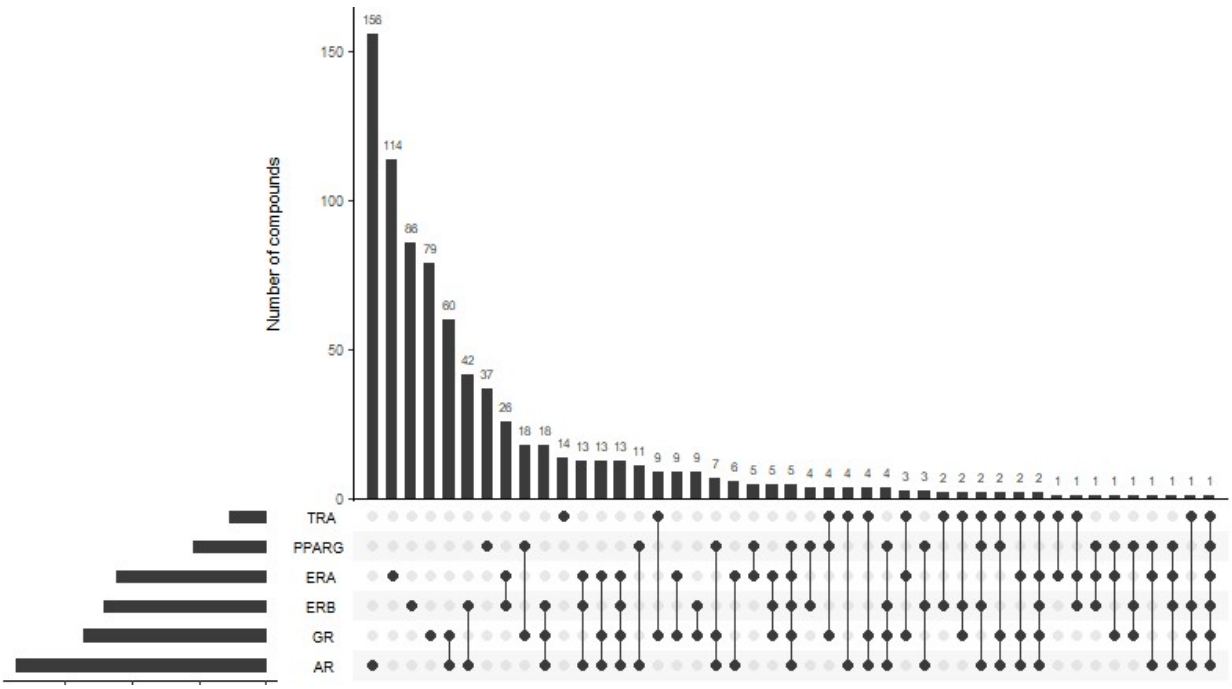


Figure 28 : *Upset plot* des composés B communs entre les différents jeux de données de l'EPA

Nous remarquons que les composés individuels pour chaque NR sont plus nombreux que les composés en communs. En effet, le jeu de AR à lui seul contient 156 B uniques suivi de ERα avec 154 B uniques. Nous remarquons aussi les couples de jeux de données AR-GR, AR-ERα présentent beaucoup de B en commun avec respectivement 60 et 42 B en communs. Les deux isoformes ERα et ERβ partagent entre eux 26 composés uniques et 68 en tenant compte des composés communs avec d'autres récepteurs. Seul 1 composés B est commun unique pour les 6 jeux de données à la fois (**Figure 28**).

Par ailleurs, des jeux de validation issues de la NR-DBIND ont été préparés et leur composition est présentée dans **Tableau 5**. Ici encore, le plus petit nombre de B est associé au jeu de données de TRα (89 B) mais cette fois le plus grand nombre de B (1200) est obtenu pour le jeu de données de GR. Les NB, contrairement aux jeux de données de l'EPA, sont beaucoup moins nombreux, entre 14 NB (TRα) et 188 NB (ERα).

Tableau 5 : Composition des différents jeux de données de la NR-DBIND

NR	B	NB
AR	583	150
ERα	685	188
ERβ	676	81
GR	1200	63
PPARγ	423	168
TRα	89	14

2.1.3.2. Docking

Pour chaque NR, 3 logiciels de docking (smina, PLANTS et Surflex-dock) ont été utilisés. Pour le logiciel smina, les 4 fonctions de score implémentées dans le logiciel ont été testées (vinardo, vina, dkoes, ad4). Au total, pour chaque NR, le jeu de données de l'EPA a ainsi été criblé à l'aide de 7 protocoles de docking sur chacune des structures du NR précédemment sélectionnées. Les résultats ont été analysés avec l'approche *single structure docking* et *ensemble structure docking et* les performances ont été évaluées à l'aide de l'AUC de la courbe

ROC (Receiver Operating Characteristic). Pour chaque récepteurs, les tableaux ci-dessous résument les meilleures performances obtenues (en terme d'AUC) pour les 6 fonctions de scores et les structures PDB associées.

Tableau 6: Meilleures AUC de docking et leurs structures associées pour AR

	Fct score	Single		Ensemble de 2			Ensemble de 3		
		Structure	**AUC**	**Structures**	**AUC**	**Gain**	**Structures**	**AUC**	**Gain**
Smina	**dkoes**	3b66	0,751	3b66,1xj7	**0,752**	**0,001**	3b66,1xj7,2pkl	**0,752**	**0,001**
	vinardo	2pir	0,624	2pir,3b67	0,636	**0,019**	2pir,3b67,2pip	0,64	**0,025**
	vina	2pir	0,656	2pir,3b67	0,665	**0,014**	2pir,3b67,2pix	0,667	**0,016**
	ad4_scoring	5v8q	0,5	5v8q,5t8j	0,5	**0,000**	5v8q,5t8j,5t8e	0,5	**0,000**
PLANTS	PLANTS	2pir	0,691	2pir,2pix	0,694	**0,004**	2pir,2pix,2pit	0,66	**-0,047**
Surflex-dock	Surflex-dock	5t8e	0,623	5t8e,3rlj	0,635	**0,019**	5t8e,3rlj,3b65	0,64	**0,027**

Tableau 7 : Meilleures AUC de docking et leurs structures associées pour ERα

	Fct score	Single		Ensemble de 2			Ensemble de 3		
		Structure	**AUC**	**Structures**	**AUC**	**Gain**	**Structures**	**AUC**	**Gain**
Smina	**dkoes**	1qku	0,708	1qku,2yja	0,709	**0,001**	1qku,2yja,1g50	0,71	**0,003**
	vinardo	1a52	0,68	1a52,1xp9	0,676	**-0,006**	1a52,1xp9,1xp1	0,673	**-0,010**
	vina	1a52	0,699	1a52,1xp9	0,696	**-0,004**	1a52,1xp9,1xp1	0,695	**-0,006**
	ad4_scoring	1a52	0,656	1a52,1x7e	0,654	**-0,003**	1a52,1qku,1x7e	0,65	**-0,009**
PLANTS	PLANTS	1x7e	0,659	1x7e,1a52	0,66	**0,002**	1x7e,1qku,1a52	0,659	**0,000**
Surflex-dock	Surflex-dock	1a52	0,604	1xp1,1x7e	0,616	**0,019**	1xp1,1x7e,1a52	0,623	**0,030**

Tableau 8 : Meilleures AUC de docking et leurs structures associées pour ERβ

	Fct score	Single		Ensemble de 2			Ensemble de 3		
		Structure	**AUC**	**Structures**	**AUC**	**Gain**	**Structures**	**AUC**	**Gain**
Smina	dkoes	3omo	0,647	3omo,3omq	0,643	**-0,006**	3omo,3omq, 3omp	0,64	**-0,011**
	vinardo	3ols	0,686	3omq,3ols	0,686	**0,000**	3omq,3omp, 3ols	0,686	**0,000**
	vina	3ols	**0,705**	3omp,3ols	0,704	**-0,001**	3omq,3ols, 3omo	0,704	**-0,001**
	ad4_scoring	3ols	0,622	3omq,3ols	0,622	**0,000**	5toa,3omq, 3ols	0,621	**-0,002**
PLANTS	PLANTS	3ols	0,638	3ols,1u3q	0,643	**0,008**	3omp,3ols, 1u3q	0,643	**0,008**
Surflex-dock	Surflex-dock	3omp	0,547	3omp,3omq	0,561	**0,025**	3omp,3omo, 1u3q	0,562	**0,027**

Tableau 9 : Meilleures AUC de docking et leurs structures associées pour le récepteur nucléaire GR

	Fonction de score	Single docking		Ensemble de 2			Ensemble de 3		
		Structure	**AUC**	**Structures**	**AUC**	**Gain**	**Structures**	**AUC**	**Gain**
Smina	**dkoes**	4mdd	0,667	4udc,6dxk	0,677	**0,015**	4udc,6dxk, 4udd	**0,678**	**0,016**
	vinardo	6dxk	0,633	4mdd,3cld	0,65	**0,026**	4mdd,3cld, 4udc	0,652	**0,029**
	vina	6dxk	0,643	6dxk,3cld	0,643	**0,000**	4mdd,6dxk, 4udc	0,642	**-0,002**
	ad4_scoring	6dxk	0,633	4mdd,4udd	0,64	**0,011**	4mdd,4udd, 6dxk	0,64	**0,011**
PLANTS	PLANTS	4mdd	0,604	4mdd,5nfp	0,62	**0,026**	4mdd,6dxk, 4p6k	0,626	**0,035**
Surflex-dock	Surflex-dock	5nft	0,532	5nft,4udc	0,583	**0,087**	5nft,4udc, 5g3j	0,596	**0,107**

Tableau 10: Meilleures AUC de docking et leurs structures associées pour le récepteur nucléaire PPARγ

	Fonction de score	Single docking		Ensemble de 2			Ensemble de 3		
		Structure	**AUC**	**Structures**	**AUC**	**Gain**	**Structures**	**AUC**	**Gain**
Smina	dkoes	2zk5	0,718	4prg,3hod	0,72	**0,003**	4prg,3hod, 2zk2	0,72	**0,003**
	vinardo	2prg	0,737	2prg,1fm6	0,738	**0,001**	2prg,4e4q, 2i4z	0,74	**0,004**
	vina	2prg	0,733	2prg,3adw	0,738	**0,007**	2prg,3adw, 1fm6	0,74	**0,009**
	ad4_scoring	2prg	0,727	2zk2,2zk6	0,729	**0,003**	4jaz,2ath, 3cdp	0,731	**0,005**
PLANTS	**PLANTS**	2prg	0,742	2prg,3zk3	0,745	**0,004**	2i4j,2prg, 2zk3	**0,746**	**0,005**
Surflex-dock	Surflex-dock	2ath	0,456	2ath;3adw	0,502	**0,092**	2ath;3adw; 3hod	0,518	**0,120**

Tableau 11 : Meilleures AUC de docking et leurs structures associées pour le récepteur nucléaire TRα

	Fonction de score	Single Docking		Ensemble de 2			Ensemble de 3		
		Structure	**AUC**	**Structures**	**AUC**	**Gain**	**Structures**	**AUC**	**Gain**
Smina	dkoes	3ilz	0,5	3jzb,3ilz	0,503	**0,006**	3jzb,3ilz,3hzf	0,504	**0,008**
	vinardo	1nav	0,358	1nav, 3hzf	0,351	**-0,020**	1nav, 3hzf,3ilz	0,346	**-0,035**
	vina	1nav	0,347	3ilz,3hzf	0,346	**-0,003**	3ilz,3hzf, 3jzb	0,344	**-0,009**
	ad4_scoring	1nav	0,321	1nav,3jzb	0,323	**0,006**	1nav,3jzb, 3hzf	0,323	**0,006**
PLANTS	PLANTS	3jzb	0,37	1nav, 3jzb	0,37	**0,000**	1nav,3hzf, 3ilz	0,346	**-0,069**
Surflex-dock	**Surflex-dock**	3hzf	**0,711**	1nav, 3hzf	0,701	**-0,014**	1nav,3hzf, 3jzb	0,689	**-0,032**

Approche *single structure docking*

Pour le logiciel smina, toutes les structures de tous les NR ont été associées à une AUC supérieure à 0,6 à l'exception du TRα pour lequel la meilleure AUC ne dépasse pas 0,5 pour toutes les fonctions de score évaluées. Les meilleures AUC ont été obtenues avec les structures de PPARγ, toutes associées à des valeurs supérieures à 0,7. Les meilleures valeurs d'AUC ont été obtenues avec la fonction de score dkoes pour 3 des 6 NR, à savoir AR, ERα et GR. Pour le logiciel PLANTS, les mêmes tendances que celles observées pour smina se dégagent : les AUC sont supérieures à 0,600, les meilleures performances sont obtenues pour PPARγ et les pires, inférieures à 0,500, pour TRα. A l'inverse, les AUC des résultats générés avec Surflex-dock sont très faibles, sauf pour TRα pour lequel des AUC supérieures à 0,700 sont obtenues. Il est aussi à noter que la structure associée à la meilleure AUC est la même pour 4 NR parmi les 6 lorsque l'on compare les résultats de smina (toutes fonctions de score confondues) et PLANTS, mais jamais pour Surflex-dock.

Approche *single structure docking* vs *ensemble structure docking*

Afin de mieux estimer l'apport du docking d'ensemble, nous avons calculé le gain des deux approches à 2 et à 3 structures tel que $Gain = \frac{(AUC_{ens} - AUC_{sing})}{AUC_{ens}}$.

La meilleure AUC est obtenue avec l'approche *ensemble structure docking* (ensemble de 3 structures) pour 4 NR sur les 6. Pour les 2 autres NR, la meilleure AUC est associée à l'approche *single structure docking*. Cependant, en utilisant l'approche de docking d'ensemble, aucune amélioration majeure n'a été constatée pour aucun des NR pour les fonctions de scores de smina et de PLANTS. Pour la fonction de score de Surflex-dock, nous remarquons une amélioration des AUC (gain > 0,1) pour GR et PPARγ mais les AUC restent tout de même faibles. Par ailleurs, dans la très large majorité des cas, la structure associée à la meilleure AUC pour l'approche *single structure docking* est aussi présente dans l'ensemble de 2 et de 3 structures associé aux meilleures AUC.

Ainsi, le **Tableau 12** résume les différents modèles de docking sélectionnés pour chaque NR. Les justifications sont détaillées dans la partie discussion.

Tableau 12 : Protocoles de docking sélectionnés pour chaque NR

NR	PDB	Profile	Coordonnées de la boite	Algorithme	F. Score	TH
AR	3b66	SARM	x_center =1,725 , y_center = 32,527 ,z_center = 4,659	smina	Dkoes	-7
ERα	1qku	Ago	x_center = 107.175 , y_center = 14.938 ,z_center = 96.009	smina	Dkoes	-6
ERβ	3ols	Ago	x_center = 33.383, y_center =80.731 ,z_center = -9.056	smina	Vina	-6.5
GR	4mdd	Atgo	x_center = 32.811, y_center = 7.088, z_center = 3.008	smina	Dkoes	-6
PPARγ	2prg	Ago	x_center = 17.836 , y_center = -21.534, z_center = 10.007	smina	Vinardo	-7

2.1.3.3. Modèles de pharmacophores

En appliquant notre protocole de génération des modèles de pharmacophores entre 9 et 37 modèles SB et de 34 à 253 modèles LB ont été créés. Ces modèles ont ensuite été optimisés suivant le protocole décrit plus haut.

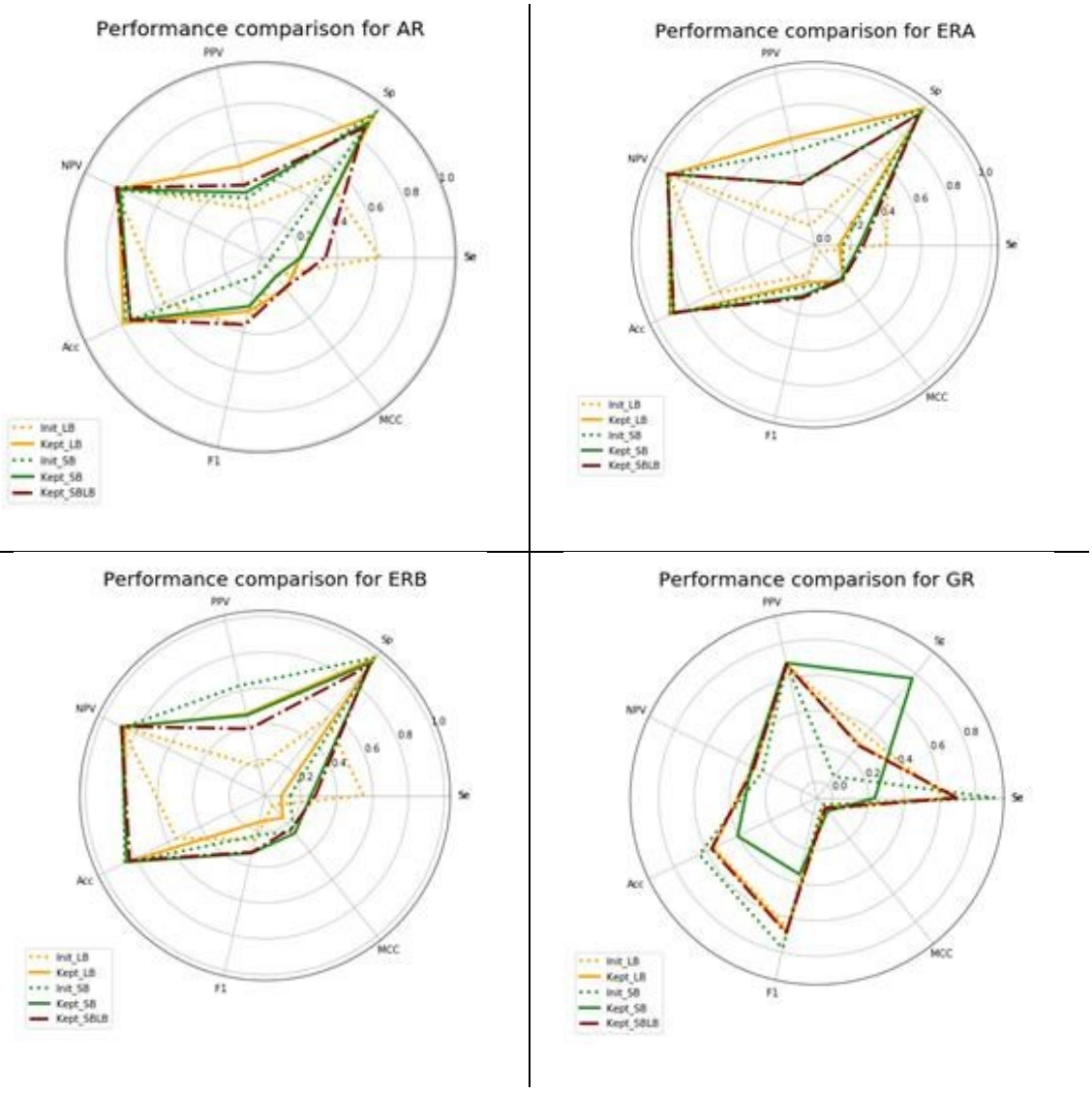
Tableau 13 : Evolution du nombre de modèles pharmacophoriques avant et après le protocole d'optimisation

	Nombre de modèles initiaux	Nombre de modèles après optimisation	Nombre de modèles finaux (sans redondances)
AR LB	253	15	13
AR SB	29	12	10
ERA LB	140	8	4
ERA SB	15	18	17
ERB LB	120	4	4
ERB SB	24	10	10
GR LB	232	8	8
GR SB	9	2	1
PPARG LB	100	3	3
PPARG SB	37	1	1
TRA LB	34	3	3
TRA SB	5	1	0

Le protocole d'optimisation permet de réduire le nombre initial de modèles en le divisant par un facteur 13 en moyenne (**Tableau 13**). La liste des pharmacophores optimisés dans leur format *pml* sont présentés en annexe. Afin d'évaluer l'impact de l'optimisation, nous avons comparé l'évolution de plusieurs métriques avant et après, comme le montre la **Figure 29**.

Aucune tendance générale d'amélioration des performances évaluées d'après les valeurs des métriques n'est observée. Pour les modèles SB, une augmentation de la Se et du facteur F1 est notée pour les modèles générés pour AR, ERα et ERβ alors que la Sp et la NPV restent inchangés. A l'inverse, pour les modèles construits pour GR et PPARγ, la Se est diminuée mais la Sp est augmentée ainsi que, même si de manière moins marquée, le MCC. Pour TRα, aucune évolution des métriques n'est observée. Pour les modèles LB, le protocole d'optimisation a

permis d'améliorer la spécificité, le MCC et le PPV des modèles pharmacophoriques développés pour 4 NR (AR, ERα, ERβ et PPARγ). Pour TRα, l'optimisation des pharmacophores n'a eu aucun impact sur la Sp et la Se, mais le MCC et le PPV étaient dans ce cas aussi augmentés. L'optimisation des pharmacophores de GR n'a modifié que la Se et la Sp, mais dans le sens inverse de ce qui a été observé pour les autres pharmacophores puisque la Sp est diminuée et la Se plus élevée.



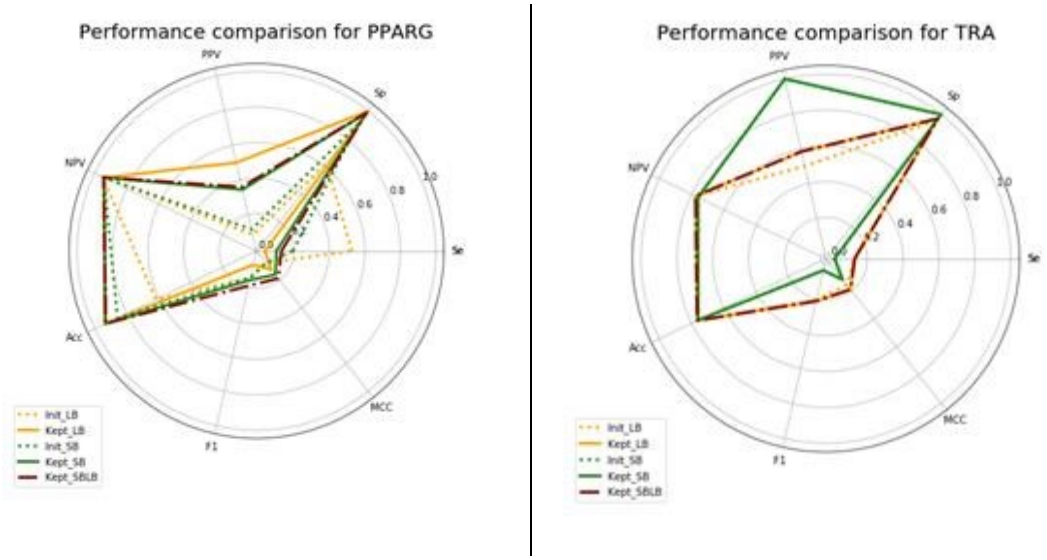


Figure 29 : Evolution des différentes métriques avant et après optimisation des modèles de pharmacophores

La combinaison des modèles de pharmacophores SB et LB a conduit à des performances assez similaires à celles obtenues avec les pharmacophores LB optimisés (GR, TRα) ou les pharmacophores SB (ERα, ERβ, PPARγ). La combinaison des pharmacophores SB et LB est globalement associée à une amélioration de la Se et du F1. En effet, la Se associée aux pharmacophores SBLB est plus élevée que celles des pharmacophores LB optimisés pour 4 NR et des pharmacophores SB optimisés pour les 6 NR. Le F1 est quant à lui amélioré pour 4 NR par rapport aux pharmacophores LB et aux pharmacophores SB optimisés. A l'inverse, la Sp associée aux pharmacophores SBLB est soit égale soit diminuée par rapport aux pharmacophores LB et aux pharmacophores SB optimisés et aucune tendance claire n'est visible pour les autres métriques.

2.1.3.4. Protocole de combinaison

Une fois un modèle sélectionné pour chacune des méthodes i.e. docking et modèles de pharmacophores, les combinaisons de ces 2 méthodes ont été évaluées pour les 5 NR étudiés (Le protocole de docking étant éliminé pour TRα). Deux manières de combiner les modèles ont été explorées respectivement appelées protocole consensus et hiérarchique. La **Figure 30** montre les performances obtenues pour chacun des modèles de docking et de pharmacophore sélectionnés ainsi que pour les deux protocoles de combinaisons.

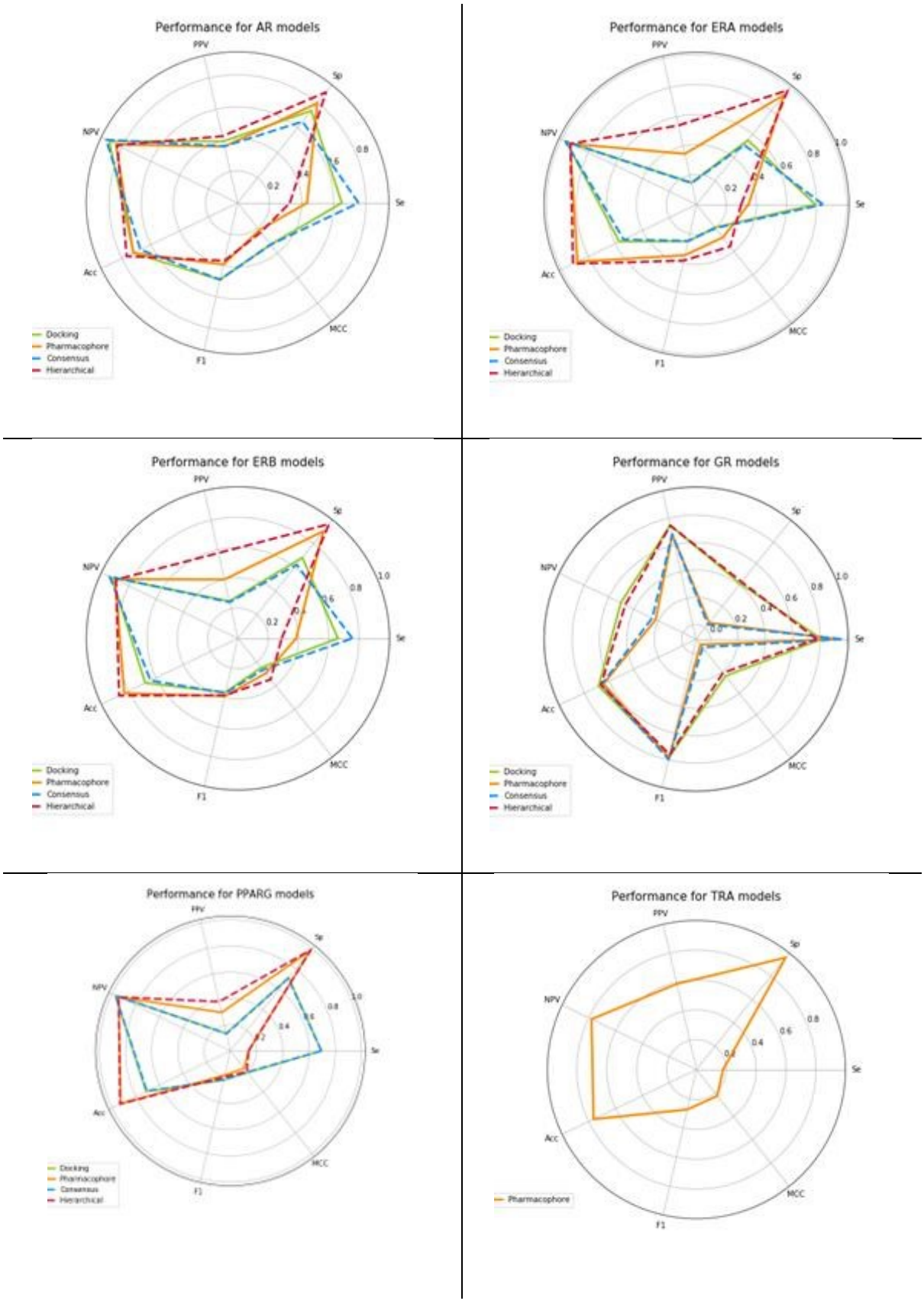


Figure 30 : Performances des différents modèles de docking, de pharmacophores et combinés (consensus et hiérarchique) pour chaque NR

Indépendamment du protocole suivi, la combinaison permet d'obtenir de meilleures performances ou des performances similaires. Le protocole consensus est associé aux plus hautes valeurs de Se supérieures à 0,69. La NPV est supérieurs à 0,9 sauf pour le modèle consensus de GR avec une NPV de 0,25. Pour les valeurs de Sp, elles sont variables entre les différents modèles mais restent tout de même faibles. Le protocole consensus de GR est associé à la plus faible valeur de Sp (0,032). Les meilleures spécificités sont celles des modèles consensus de AR (0,650) et PPARγ (0,7). A l'inverse le protocole hiérarchique permet d'atteindre de très hautes valeurs de Sp toutes supérieurs à 0,9 sauf pour GR où elle est égale à 0,429. Par ailleurs le protocole hiérarchique permet d'atteindre les meilleures valeurs de MCC avec un maximum de 0,359 pour le jeu de données de ERα.

Ainsi, pour chaque NR, un modèle a été choisi entre les 4 options docking, pharmacophore, consensus ou hiérarchique. Les différents modèles correspondant à chaque NR sont présentés dans le **Tableau 14** avec les performances associées.

Tableau 14 : Performance des modèles sélectionnés sur les jeux de données de l'EPA

Model	Modèle	Se	Sp	PPV	NPV	Acc	F1	MCC
AR	Consensus	0,755	0,650	0,364	0,909	0,672	0,491	0,332
ERA	Consensus	0,842	0,509	0,146	0,970	0,539	0,249	0,202
ERB	Consensus	0,756	0,623	0,245	0,940	0,641	0,370	0,265
GR	Docking	0,812	0,413	0,748	0,505	0,685	0,779	0,238
PPARG	Consensus	0,694	0,707	0,134	0,694	0,706	0,225	0,175
TRA	Pharmaco.	0,179	0,959	0,588	0,779	0,764	0,274	0,224

Enfin, les performances de nos modèles ont été évalués sur les différents jeux de données de la NR-DBIND et les performances sont exposées dans le **Tableau 15**. Nous remarquons que pour tous les jeux de données sauf TRα, les valeurs de Se, d'Acc et de F1 sont hautes. En effet pour TRα, seules la Sp et la PPV sont élevées.

Tableau 15 : Performance des modèles sélectionnés sur les jeux de données de la NR-DBIND

Model	Se	Sp	PPV	NPV	Acc	F1	MCC
AR	0.784	0.1333	0.780	0.137	0.651	0.781	-0.084
ERA	0.983	0.024	0.766	0.308	0.758	0.861	0.024
ERB	0.895	0.432	0.929	0.330	0.845	0.912	0.291
GR	0.998	0.0	0.95	0.0	0.948	0.973	-0.011
PPARG	0,998	0,042	0,724	0,875	0,726	0,839	0,153
TRA	0.090	0.786	0.727	0.120	0.184	0.16	-0.138

2.1.4. Discussion

L'hypothèse sur laquelle repose cette étude découle du mécanisme d'action direct des PE (c.f paragraphe I.4.3 Mécanismes d'action). Ce mécanisme stipule que cette catégorie de composés est capable de se lier au récepteur nucléaire bloquant ainsi l'action du ligand endogène ou entrainant l'activation non désirée du récepteur. Biologiquement, cela peut se traduire soit (1) par la compétition du PE avec le ligand endogène au niveau du site orthostérique ou (2) par la liaison du PE au niveau d'un autre site allostérique du récepteur nucléaire empêchant le ligand de correctement induire son activité. Pour ce projet, le parti pris de se concentrer uniquement sur le site actif était motivé par deux éléments principaux. Le premier est la disponibilité de structures de PE co-cristallisés dans le site actif de NR dans la littérature et pour lesquels les mécanismes moléculaires ont été étudiés. Le second repose sur le principe de similarité qui stipule que deux composés dont la structure est similaire auront la même activité biologique [37]. Ce concept est largement utilisé non seulement en thérapeutique et en Drug design mais aussi en toxicologie dans les méthodes de « Read-across » [27]. En effet, il est reconnu que de nombreux PE agissent au niveau du site d'action des hormones de par la similarité de structure chimique[224] . C'est le cas notamment des phytoœstrogènes, la première catégorie de PE à avoir mis la lumière sur le mécanisme de perturbation endocrinienne [351].

Les modèles construits dans cette étude permettent donc d'étudier la capacité des PE à se lier au LBD de 6 NR et ne couvrent donc pas tous les PE et tous les mécanismes de perturbation

endocrinienne. Les résultats engendrés devront être analysés dans un cadre plus complet de modélisation d'autres mécanismes d'action en parallèle mais surtout de validation biologique[352]. L'objectif de cette étude est de proposer des modèles qui pourront être utilisés pour établir une liste de composés à investiguer plus amplement en priorité mais ces modèles ne suffisent pas pour statuer définitivement sur le caractère PE d'un composé.

Un autre parti pris réalisé pour cette étude est le choix des NR étudiés à savoir AR, ERα, ERβ, GR, PPARγ et TRα. Historiquement, les PE ont été liés aux manifestations hormonales sexuelles, et très vite, ERα, Erβ, AR ont été mis en cause. Plus tard, d'autres NR hormonaux ont été aussi incriminés incluant GR et TRα [353]. Finalement, le nombre croissant d'études reliant les PE à PPARγ [340] nous a décidé à inclure aussi ce récepteur.

2.1.4.1. Données

La **Figure 27** montrant la distribution des propriétés physicochimiques tel que le poids moléculaire, ou le ClogP (pour la liste complète c.f paragraphe II.2.1.2. matériel et méthodes) pour les B et NB pour les différents jeux de NR nous permet de conclure que les différences d'activités (B/NB) ne sont pas dues uniquement à des différences de ces propriétés mais bien à des caractéristiques structurales.

Nous avons utilisé un *Upset plot* (**Figure 28**) pour illustrer les composés B communs entre les jeux de données des différents NR étudiés. Ce graphique montre qu'un peu moins de la moitié des composés sont capables de se lier à au moins 2 NR (307 sur 796 composés B tous NR confondus) et que 1 seul composé est B pour les 6 jeux de données utilisés cette étude. Outre le fait que les PE soient connus pour leur caractère ubiquitaire, une hypothèse que nous souhaitons explorer est la possibilité que les poches de certains des NR étudiés présentent une similarité en termes de propriétés physicochimiques.

Un grand nombre de B est commun entre les jeux de données de ERα et β : 68 B en commun au total en comptabilisant les 26 B communs uniques à ERα et ERβ et ceux en communs avec d'autres récepteurs. Ceci est logique puisque les 2 isoformes présentent une grande similarité et beaucoup de composés sont capable de se lier aux deux isoformes avec différents niveaux d'affinités[354]. C'est aussi le cas de certains PE comme par exemple les phytoœstrogènes, comme la genistéine, contenues dans des aliments comme le soja qui sont capables d'interagir avec les 2 isoformes mais avec une meilleure affinité pour ERβ [335,354].

Le jeu de données de AR est celui contenant le plus grand nombre de B et celui de TRα est celui présentant le moins de composés B et ceux-ci sont essentiellement des antagonistes (**Figure 31**). Le jeu de données de PPARγ présente relativement peu de B alors que dans les

bases de données dédiées aux NR comme la NR-DBIND [216], c'est l'un des NR pour lequel le plus de données sont disponibles. Ceci pourrait s'expliquer par l'intérêt plus récent de ces récepteurs en tant que cible de PE[355] .

La distribution des profils pharmacologiques des B pour chacun des NR étudiés est présentée au niveau de la **Figure 31**.

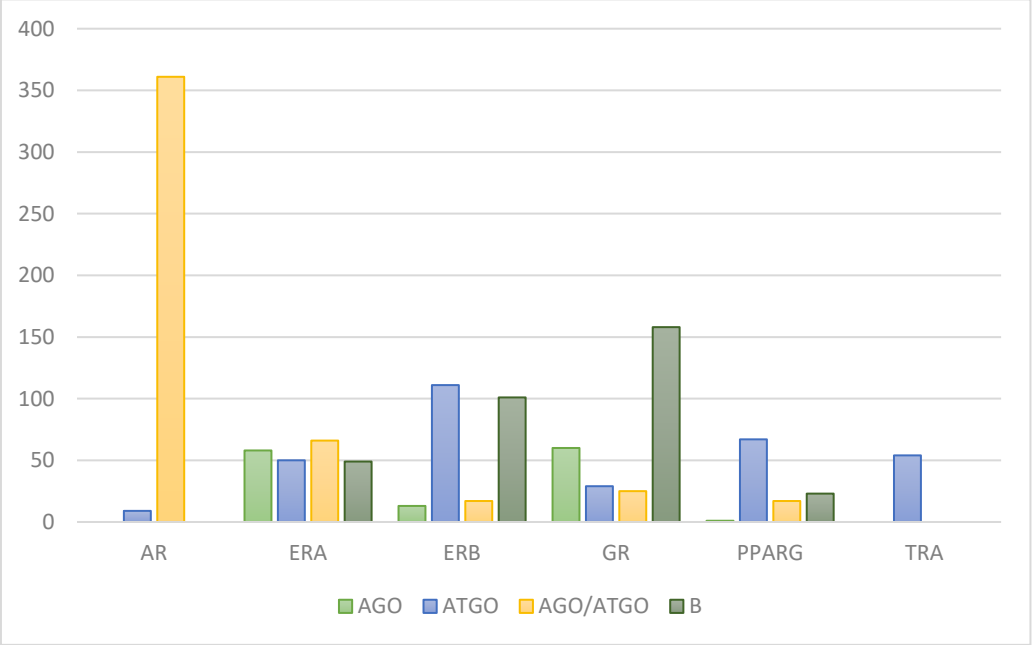


Figure 31: Distribution des profils agonistes (AGO), antagonistes (ATGO), agonistes-antagonistes (AGO/ATGO) et non testés pour l'agonisme des différents B des jeux de données étudiés

A l'exception de AR et TRα, beaucoup de composés testés pour leur affinité de liaison aux NR n'ont pas encore été évalués avec des tests à gènes rapporteurs pour déterminer leurs profils pharmacologiques. Les autres jeux de données présentent aussi bien des agonistes, des antagonistes que des agonistes/antagonistes. Il est à noter qu'après la constitution des jeux de données utilisés pour cette étude (entre 2019 et 2020), la base de données a migré et a été remplacée par la Comptox[333]. Les résultats de tests expérimentaux ont été rajoutés au fur et à mesure et il est possible que de nouvelles données (nouveaux composés et résultats biologiques) soient maintenant disponibles.

Enfin, les proportions des différents jeux de données révèlent un déséquilibre entre B et NB. Pour tous les NR, le nombre de NB est largement supérieur au B, ce qui est classique et attendu, sauf pour GR pour lequel c'est le contraire. Ce déséquilibre, dans un sens ou l'autre, est à prendre en considération au moment d'analyser les résultats. Ainsi, une attention particulière a

été consacrée à évaluer les performances en incluant différentes métriques qui incorporent dans leurs formules les 4 descripteurs du tableau de contingence à savoir les vrais positifs (VP), les vrais négatifs (VN), les faux positifs (FP) et les faux négatifs (FN).

2.1.4.2. Docking

Les études de docking ont été réalisées avec 3 logiciels de docking à savoir PLANTS, Surflex-Dock et smina avec les 4 fonctions de scores implémentées par défaut i.e dkoes, vina, vinardo et ad4. Le but était d'explorer les performances d'outils de docking présentant différents algorithmes d'échantillonnage et différentes fonctions de score pour évaluer leur performance, rechercher un éventuel compromis parmi les résultats sur la ou les structures à utiliser et l'approche donnant les meilleurs résultats et ainsi renforcer le processus de sélection de ces 2 points critiques.

2.1.4.2.1. Sélection des structures, fonctions de scores et approches de docking

En comparant les AUC obtenues avec les deux approches, single et *ensemble structure docking*, nous ne remarquons aucune amélioration significative qui nous incite à sélectionner la dernière approche (du **Tableau 6** au **Tableau 11**). En effet, l'approche *ensemble structure docking* nécessitant des temps de calcul plus long, nous avons décidé de continuer avec l'approche *single structure docking* pour tous les NR. Pour chaque NR, une structure, un algorithme de docking et une fonction de score ont été choisis. Le rationnel de ce choix est détaillé ci-dessous. Pour ERα qui a fait l'objet d'un article dédié, les résultats ne sont pas reprécisés ici.

Pour AR, nous avons sélectionné la structure 3b66, l'algorithme d'échantillonnage de smina et la fonction de score dkoes pour lesquels une AUC de 0,751 a été obtenue. Il est à noter que la meilleure AUC toute approche confondue était de 0,752 pour un ensemble de 3 structures (incluant 3b66), avec le même algorithme de recherche et la même fonction de score. La structure associée à la meilleure AUC pour chacun des 6 protocoles de docking étudiés n'était pas la même. Cette différence s'explique probablement par le profil des ligands co-cristallisés. En effet, il a été décrit dans la littérature que l'adéquation entre le profil pharmacologique des ligands ciblés et du ligand co-cristallisé de la structure est important pour assurer de bonnes performances de prédiction des ligands des NR avec des méthodes de docking [198]. Dans le paragraphe précédent, nous avons mentionné que la majorité des ligands du jeu de données AR présente un profil agoniste/antagoniste. Or, la structure 3b66 associée à la meilleure AUC pour AR est en complexe avec un ligand à profil SARM (c'est aussi le cas de la structure 5v8q

associée à la meilleure AUC pour smina-ad4_scoring). A l'inverse, 2pir, la structure associée aux meilleurs résultats pour 3 protocoles sur les 6 évalués (smina-vinardo, smina-vina et PLANTS) mais avec des AUC plus faibles, est co-cristallisée avec la testostérone (DHT), ligand endogène et agoniste de AR.

Pour ERβ, 4 fonctions de scores sur 6 s'accordent sur le choix de la meilleure structure : 3ols. La meilleure AUC tout protocole de docking confondu est de 0,705 et elle obtenue avec cette structure 3ols en utilisant l'algorithme de recherche de smina et la fonction de score vina. Il est à noter que pour les 2 protocoles de docking restants, smina-dkoes et Surflex-dock, les structures associées aux meilleures AUC sont respectivement 3omo et 3omp [356]. Ces 3 structures de ERβ sont toutes en conformation agoniste. Alors que 3omp et 3omo sont en complexe avec des quinolones synthétiques, 3ols est associée à l'estradiol et permet d'étudier l'interaction du récepteur avec le ligand endogène [357].

Toutes les structures de GR disponibles dans la PDB présentent des mutations nous contraignant à faire une entorse à nos critères de sélections. Historiquement, ces mutations sont dues aux problèmes de solubilité rendant le LBD difficile à faire exprimer[358]. Ainsi, nous avons examiné les différentes mutations des différentes structures humaines et nous avons uniquement gardé celles où il n'y avait pas de mutations autour de 4 A du ligand co-cristallisé. Le protocole sélectionné pour GR est smina-dkoes avec la structure 4mdd qui est associée à une AUC de 0,667. Tout comme pour AR, la meilleure AUC toute structure confondue, d'une valeur de 0,678, a été obtenue avec le même protocole mais pour un ensemble de 3 structures. La structure 4mdd (ayant aussi donné la meilleure AUC pour le logiciel PLANTS) et la structure 6dxk, associée aux meilleures AUC pour les 3 autres fonctions des score de smina, vinardo, vina et ad_4, sont en complexe avec un ligand antagoniste alors que la structure 5nft, sélectionnée par Surflex-Dock, est en complexe avec un modulateur (SGRMs)[359]. Les performances obtenues, un peu plus faibles que pour AR et ERβ, pourraient s'expliquer par le nombre réduit d'antagonistes au niveau du jeu de données GR[360].

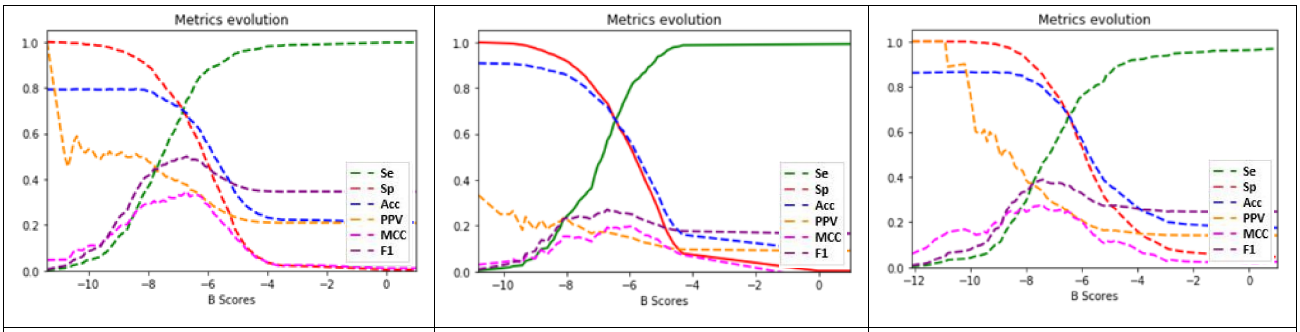
Pour PPARγ, la majorité des protocoles s'accordent sur le choix de la structure 2prg. Le protocole choisi est smina-vinardo avec cette même structure, pour lequel une AUC de 0,737 a été obtenue. Il est à noter que des AUC légèrement supérieurs, 0,746 et 0,742, ont été atteintes avec le logiciel PLANTS pour un ensemble de 3 structures et l'approche *single structure*

docking respectivement. Cependant, l'objectif final étant de proposer un outil de criblage en ligne, nous avons choisi pour faciliter l'implémentation de l'outil, le protocole avec smina-vinardo. La structure 2prg est associée à une thiazolidinedione (TDZ), un anti-diabétique oral (profil agoniste) et a l'avantage par rapport aux autres structures choisies de présenter une poche assez large [361]. Cela expliquerait l'amplitude des AUC un petit peu plus élevé que le reste des autres NR (**Tableau 10**).

Seules 4 structures répondant à nos critères de sélection étaient disponibles pour TRα dans la PDB. Il est important de mentionner que ces 4 structures sont en complexe avec un ligand spécifique de TRβ. Seul le logiciel Surflex-dock a permis d'obtenir des résultats satisfaisants avec une AUC de 0,711 pour la structure 3hzf. Pour smina (et ses 4 fonctions de score) et PLANTS, tous les composés, B et NB, ont obtenus des scores élevés avec des moyennes de scores égales pour les 2 catégories. La distinction entre B et NB n'était donc pas possible ce qui se reflète sur les performances de docking (les AUC < à 0,5). Le logiciel Surflex-dock au contraire associe de très mauvais scores aux poses (pour les B les moyennes vont de -4 à 0,14 alors qu'un bon score est associé à des valeurs positives) et entachent la confiance dans les résultats obtenus. Cela pourrait s'expliquer par le fait que les 4 structures utilisées sont en complexe avec des ligands spécifiques de TRβ entrainant peut être une conformation non favorable de la poche. Pour la suite, nous avons décidé d'éliminer le modèle de docking pour le récepteur TRα afin de ne pas impacter les performances finales pour ce NR.

2.1.4.2.2. Sélection du score seuil

Une fois la sélection du protocole optimal réalisée pour chaque NR, un score de docking seuil (TH) a été défini en utilisant la courbe de prédictivité. A la différence de ce qui a été fait dans la preuve de concept, nous n'avons pas sélectionné les différents TH correspondant aux Se de 0,25 – 0,5- et 0,75 mais nous avons plutôt cherché à chaque fois à maximiser la Prédictivité (P) et la Se simultanément. En effet, l'outil screening explorer permet une utilisation interactive de la courbe de prédictivité afin d'obtenir au niveau de chaque point de la courbe les différentes Se, Sp et Prédictivé (P) associées. Une fois quelques hypothèses de TH sélectionnées, nous avons affiné cette sélection en suivant l'évolution des différentes autres métriques (Acc, MCC, F1, PPV, NPV) en fonction des moyennes de score pour les B. Les différents modèles de docking choisis sont présentés dans le **Figure 32**.



AR| dkoes | TH = -7	ERA| dkoes | TH = -6	ERB| vina | TH = -6,5
GR | dkoes | TH = -6	PPARG | vinardo | TH = -7	

Figure 32 : Evolution des différentes métriques en fonction des scores de docking des composé B

2.1.4.3. Modèle de Pharmacophore

Modèles SB

Des modèles SB de pharmacophores ont été générés à partir des structures expérimentales des complexes ligand-récepteur sélectionnés pour chacun des 6 NR. Les modèles obtenus permettent ainsi de représenter les propriétés stériques et électroniques impliquées dans l'interaction ligand-récepteur au niveau du site de liaison. Pour ERα, ERβ et AR, nous remarquons qu'un grand nombre de modèles SB sont retenus après optimisation et élimination des pharmacophores SB redondants. A l'inverse, pour les 3 autres NR, 2 ou 1 seul pharmacophore SB optimisé a été retenu après évaluation de leurs performances (GR, 2 modèles SB retenu sur 9 générés ; PPARγ, 1 modèle SB retenu sur 37 générés ; TRα, 1 modèle SB retenu sur 5 générés). Pour PPARγ, au premier tour de criblage 25 sur 37 des modèles générés ne sélectionnaient pas de touches (*hits*), sur les 12 restants soumis au protocole d'optimisation, les pharmacophores donnant un ratio B/NB supérieur à 1 sont au nombre de 8 dont 7 tous redondants avec le seul modèle SB retenu. Pour GR, les 9 modèles de départ possèdent beaucoup de points pharmacophoriques (entre 5 et 10 points) hydrophobes pour la majorité. Ainsi cette similarité dans la nature des points pharmacophoriques fait que beaucoup sont redondants inclus dans le 2 seuls SB conservés. Pour TRα, dès le premier tour 4 sur les 5 modèles ne sélectionnaient aucune touche. Ce dernier contenant 3 points pharmacophoriques n'a pas pu être plus optimisé. Il est intéressant de constater que les modèles SB retenus pour AR et ERβ sont tous issus de structures en conformation agoniste alors que des structures en complexe avec un modulateur pour AR et 2 structures co-cristallisées avec des antagonistes pour ERβ, faisaient partie de la liste initiale de structures. Le contraire est observé pour ERα pour lequel les pharmacophores SB retenus sont issues aussi bien de structures en complexe avec un agoniste, un antagoniste mais surtout dans la majorité des cas un modulateur (SERM). Après optimisation des pharmacophores SB, certains pharmacophores retenus dérivent d'un même pharmacophore initial. C'est notamment le cas pour ERα avec 3 structures (2bj4, 2iog, et 3ert) qui ont chacune permis de générés plusieurs pharmacophores optimisés complémentaires non redondants.

Enfin, aucune des structures retenues pour le docking pour chaque récepteur ne figure dans la liste des structures ayant donné lieu à un pharmacophore SB optimisé et retenu. Ceci est en

faveur de la complémentarité des méthodes étudiés et est intéressant à noter dans l'optique de combiner ces méthodes.

Modèles LB

Les structures des molécules pour chaque jeu de données étant différentes, une étape de « *clustering* » préliminaire était nécessaire. Pour chaque groupe (*cluster*), 10 modèles de pharmacophores au maximum ont été générés, ce qui explique le nombre élevé de modèles LB initiaux. Conceptuellement, les modèles LB permettent d'identifier, pour chaque groupe de molécules structurellement similaires et alignées, le maximum de propriétés communes supposées nécessaires pour interagir avec le récepteur.

Protocole d'optimisation

Le protocole d'optimisation a permis de diminuer le nombre de pharmacophores inclus dans chaque groupes SB et LB pour chaque NR étudié ce qui représente un atout pour la durée du criblage tout en améliorant les performances associées. Ceci est particulièrement visible pour les modèles pharmacophoriques LB pour lesquels le protocole d'optimisation a permis d'améliorer la Sp, le MCC et le PPV des modèles pour 4 NR. Pour les modèles SB, le protocole d'optimisation a permis d'améliorer la Se pour 3 NR sur 6 (AR, ERα et β) sans modifier la Sp. Pour ces récepteurs , les pharmacophores LB optimisés présentent donc de meilleures valeurs de spécificité alors que les pharmacophores SB ont tendance à être associés à de meilleures valeurs de sensibilité. Pour le modèle GR, il s'agit du cas contraire où les modèles SB sont associés à une meilleure Sp et les modèles LB génèrent de meilleures Se. Dans les deux cas, cela démontre l'intérêt de combiner ses 2 types de pharmacophores afin d'obtenir des performances optimales pour toutes les métriques étudiées.

Modèles combinés SBLB

Les modèles de pharmacophores SB optimisés et de pharmacophores LB optimisés pour chaque NR ont été combinés ensemble dans un groupe appelé pharmacophores SBLB. Les modèles de pharmacophores redondants, capables de retrouver des B déjà retrouvés par un autre modèle plus inclusif, ont été éliminés au sein de chaque ensemble SBLB. Le cas de TRα est singulier, puisqu'après combinaison et élimination des redondants, aucun pharmacophore SB n'est conservé. L'ensemble SBLB pour TRα correspond donc en fait aux pharmacophores LB. Pour les 5 NR restants (AR, ERα, ERβ, GR et PPARγ) les pharmacophores SBLB sont associés à de meilleures valeurs de Se, MCC et F1 par rapport aux pharmacophores SB et LB optimisés avec

des valeurs de Sp, Acc et NPV légèrement diminuées mais qui restent très similaires de celles des modèles LB. La PPV chute pour tous les modèles SBLB sauf pour GR qui maintient une PPV élevée. Cette même constatation a été faite lors de précédents travaux qui ont permis d'obtenir des modèles SBLB dont les performances dépassaient largement celles des modèles LB et SB pris individuellement [339]. Au total, 74 pharmacophores SBLB sont donc retenus avec des ensembles de 3 (TRα) à 23 (AR) pharmacophores SBLB constituant les modèles de pharmacophore dont la combinaison avec les méthodes de docking a été évaluée.

2.1.4.4. Protocole de combinaisons et sélection des modèles

Cette étude vise à créer des modèles capables de prédire la capacité de composés à interagir avec 6 NR. Le but est donc d'obtenir les modèles avec le plus fort pouvoir prédictif possible, estimés grâce aux valeurs de Se et Sp qui devraient être les plus hautes possibles [352]. Cependant, nous avons observé au cours de cette étude qu'un gain de Sp est généralement associé à une perte de Se et inversement. Dans un contexte toxicologique tel que celui de cette étude, il est plus important de garantir que le modèle retrouve bien tous les actifs i.e. les molécules toxiques [352,362] que d'éliminer les vrais inactifs. La Se doit donc être optimisée en priorité. Un autre descripteur de la matrice de contingence important à évaluer est le nombre de FN qui doit être le plus faible possible dans un contexte toxicologique [324]. Les indicateurs que nous considérons en priorité pour sélectionner nos modèles sont la Se associée à la NPV.

Il existe différentes manières de rationaliser le choix du modèle et son évaluation [352] et il est difficile de trouver un protocole standard à suivre systématiquement pour obtenir un bon modèle pour chaque NR. Intuitivement, l'idée de combiner les prédictions de docking et des modèles pharmacophoriques était attrayante dans l'objectif de compenser les "défaillances" de chacune et d'améliorer les Se et les Sp. Deux manières de procéder à cette combinaison ont été suivies, appelées protocoles consensus et hiérarchique. Ces protocoles ont été appliqués sur 5 des 6 NR étudiés. En effet, pour TRα, les prédictions obtenues avec les méthodes de docking ont été écartées car les performances étaient trop faibles. Il est à noter aussi que, comme mentionné précédemment, aucun pharmacophore SB n'a été conservé après élimination des pharmacophores redondants lors de la création de l'ensemble de pharmacophores SBLB. Pour TRα, le modèle sélectionné est donc uniquement le modèle de pharmacophore LB. Ceci peut s'expliquer par le faible nombre et la qualité des 4 structures de TRα retenues. Une étude plus approfondie de la structure et l'étude d'un protocole de docking protéine flexible avec smina par exemple semble particulièrement pertinent pour ce NR.

Pour AR, ERα, ERβ et PPARγ nous remarquons deux tendances parmi les modèles, d'un côté les pharmacophores SBLB et le modèle hiérarchique caractérisés par de fortes Sp et de l'autre les résultats de docking et le modèle consensus associés à de fortes Se. En effet, les modèles hiérarchiques sont plus stricts puisqu'il s'agit de l'intersection des prédictions des modèles de pharmacophores et de docking entrainant ainsi une augmentation du nombre de vrais négatifs (VN) associée malheureusement à une augmentation des FN. A l'opposé, les modèles consensus sont plus permissifs entrainant en conséquence une diminution des FN. Nous avons calculé différentes métriques pour évaluer et comparer les modèles générés. Cependant, les pharmacophores et le protocole hiérarchique sont associées à des valeurs de Se extrêmement basses (autour de 0,3). Dans le contexte toxicologique des PE, cela signifierait d'éliminer de la liste des composés à tester en priorité de nombreux composés qui sont en fait des B des NR, ce qui n'est pas acceptable. Nous avons donc choisi de prioriser la Se et la NPV au détriment des autres métriques en acceptant le risque d'inclure dans la liste des composés à tester en priorité des composés NB pour les NR.

Pour AR, ERα et ERβ, nous avons donc choisi comme modèle de prédiction final le protocole consensus. Pour ces 3 NR, le protocole consensus a permis de générer les meilleures valeurs de Se, MCC et F1 associées à de plus faibles valeurs de Sp. Pour PPARγ, le protocole de combinaison consensus et les résultats de docking permettent d'obtenir des performances identiques incluant la meilleure Se. Il est à noter que pour ce récepteur, les valeurs de MCC et F1 restent les mêmes quel que soit le protocole utilisé. Même si les performances associées sont identiques, nous avons cependant décidé de garder le protocole consensus comme modèle final puisque le criblage par les pharmacophores est relativement rapide (4 pharmacophores sont inclus dans l'ensemble SBLB). Pour les 4 NR mentionnés ci-dessus, la valeur de la NPV est similaire pour les 4 protocoles comparés et est supérieure à 0,8.

Pour GR, la tendance est inversée et les performances obtenues avec le protocole consensus et hiérarchique sont similaires à celles associées respectivement aux pharmacophores SBLB et aux résultats de docking. L'ensemble des protocoles permettent d'obtenir de hautes valeurs de Se, ce qui pourrait être lié à la composition du jeu de données incluant en majorité des B. Les résultats de docking et le protocole hiérarchique permettent d'obtenir des valeurs de Se légèrement plus faibles que les pharmacophores SBLB et le protocole consensus mais de meilleures valeurs de Sp, NPV et PPV et de MCC. En examinant les performances des modèles pharmacophoriques de GR, nous remarquons qu'ils sont très permissifs puisque le nombre de *hits* trouvé est de 371 sur 398 composés de départ. C'est pour cela que les performances du

modèle de docking sont superposables celles du modèle hiérarchique. A performances similaires, nous sélectionnant le modèle de docking par manque de confiance dans le modèle pharmacophorique. Il est à noter que les 4 protocoles pour GR permettent d'obtenir les mêmes performances en termes de F1 et Acc.

2.1.1.1. Contextualisation

Dans le but de réaliser une validation des modèles finaux choisis pour chacun des 6 NR, nous avons décidé de créer des jeux de données pour chaque récepteur étudié à partir de la base de données NR-DBIND[216] et de les cribler avec le modèle correspondant. Les performances obtenues sur les jeux de données de NR-DBIND présentent des valeurs de Se élevées pour tous les NR sauf pour TRα associé à une Sp faible. Le F1 est élevé pour tous sauf pour TRα montrant une bonne harmonie entre la spécificité et la précision (PPV). Toutefois, contrairement aux performances obtenues avec le jeu EPA, le NPV est assez faible pour tous les NR (et même égale à 0 pour GR), indiquant un nombre élevé de FN i.e un grand nombre de composés B prédits comme NB ce qui est en accord avec la Sp faible et les valeurs de MCC négatives. Ceci est un résultat décevant puisque nous avons mis l'accent dans la sélection des modèles finaux sur leur capacité à ne pas exclure de B. Pour TRα, le modèle final est constitué des seuls pharmacophores LB ce qui peut expliquer la forte spécificité obtenue.

Pour justifier ces faibles performances, deux hypothèses sont envisagées. La première est liée au nombre très faible des NB inclus dans les jeux de données extraits de la NR-DBIND avec un ratio B/NB inversé par rapport aux données de la NR-DBIND sauf pour GR. En effet, les NB inclus dans les jeux de données de la NR-DBIND ont été collectés après une revue intensive de la littérature, mais le manque de publications de résultats négatifs limite le nombre de composés disponibles à cet effet [215,216]. Une solution afin de pallier ce problème serait d'augmenter le nombre de NB afin de réaliser un nouveau criblage avec les modèles finaux et analyser les performances obtenues. Il serait par exemple possible d'inclure des NB supposés ou decoys en utilisant la base de données et le générateur de leurres (*decoys)* de la DUD-E[214]. Malheureusement une limitation technique du générateur de *decoys* de la DUD-E nous a contraint à ne pas pouvoir générer les *decoys* à temps. Nous remettons ce travail d'évaluation à une date ultérieure.

Une deuxième hypothèse serait que les jeux de données de la NR-DBIND et de l'EPA sont trop différents au niveau des structures inclues. En effet, les composés inclus dans cette base de

données i.e les B et NB, ont une vocation *druglike* contrairement aux PE qui eux incluent une grande variabilité structurale. Pour ce faire, nous avons décidé de visualiser l'espace chimique dans lequel sont contenus les deux jeux de données ainsi que la distribution des B et des NB de chaque jeu dans cet espace. Nous nous sommes pour cela basés sur les visualisations t-SNE [363] (*t-distributed stochastic neighbor embedding*) pour chaque NR réalisées avec le logiciel DataWarrior [364] à partir des structures SMILES des différentes molécules. Cela a confirmé en partie notre hypothèse et a permis de révéler que les B et NB de la NR-DBIND étaient très similaires et ne recouvraient qu'une partie de l'espace des composés B issues des jeux de l'EPA. Cela remet en question la pertinence de la SALI map utilisée pour représenter les espaces chimiques dans la preuve du concept.

Ainsi, cette analyse a permis de mettre la lumière la nécessité de contraindre le contexte d'utilisation de nos modèles d'une part dans son contexte d'interprétabilité (thérapeutique ou toxicologique) mais surtout concernant le domaine d'applicabilité (AD) qui affecte en particulier les modèles LB.

De plus, les résultats et les interprétations doivent être faits en considérant que ce que nous cherchons : la capacité des composés à se lier aux NR et non leur innocuité en absolue. Si les composés de départ sont connus pour être toxiques ou des PE reconnus les modèles de prédiction de liaison aux NR présentés dans cette étude permettront de prédire s'ils agissent via un mécanisme direct ou non. Dans le cas d'une campagne de criblage à la recherche de potentiels PE, les modèles présentés permettront d'établir une liste de composés à tester expérimentalement. Enfin, dans un contexte de *drug design*, les modèles ici décrits pourraient renseigner sur la possibilité qu'un composé se lie aux NR qui pourrait être soit la cible thérapeutique soit une anti-cible, mais les performances obtenues avec la NR-DBIND n'encourage pas cette application.

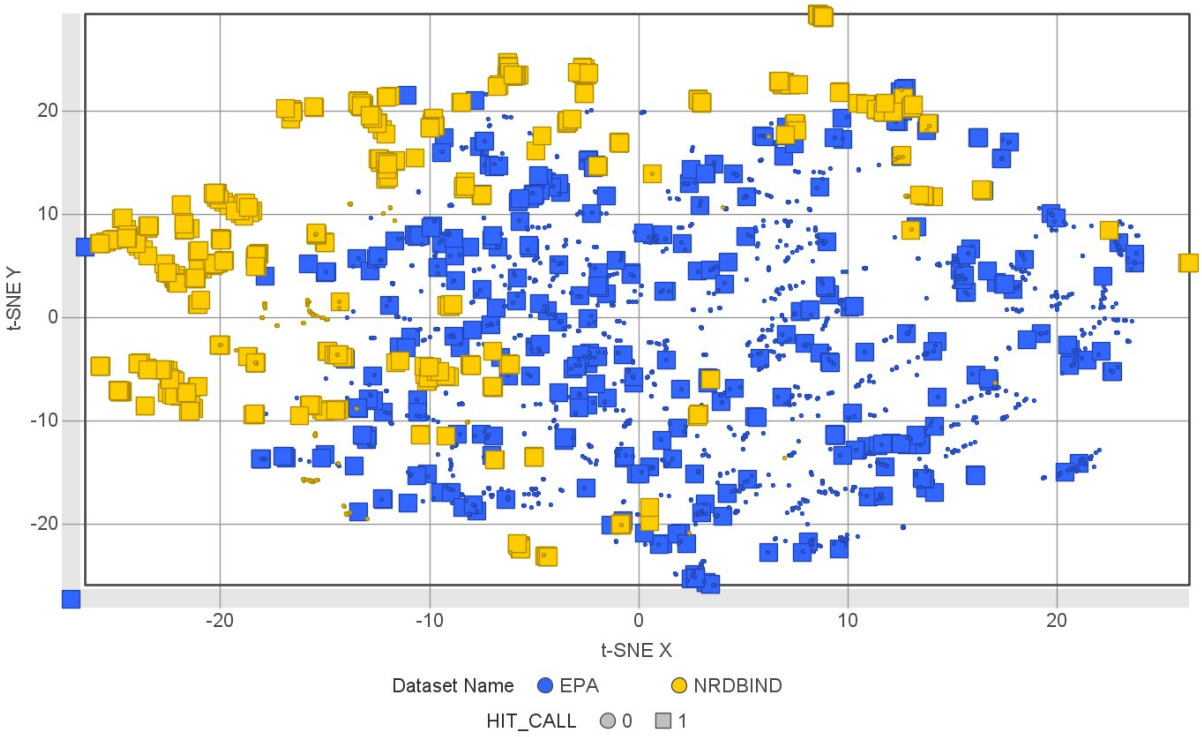


Figure 33 : Distribution des composés B et NB du jeu de données de l'EPA et celui de la NR-DBIND pour le récepteur AR. Pour chaque jeu de données, Les carrés représentent les B (1) et les ronds représentent les NB (0)

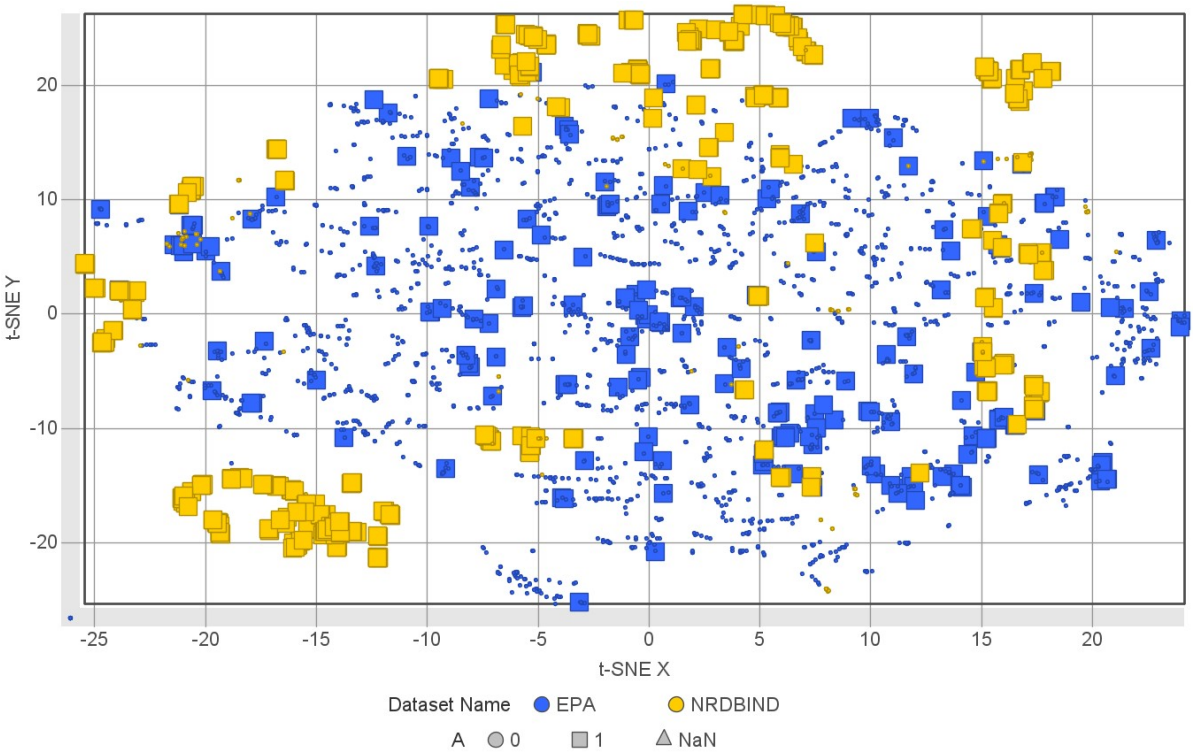


Figure 34 : Distribution des composés B et NB du jeu de données de l'EPA et celui de la NR-DBIND pour le récepteur ERα. Pour chaque jeu de données, Les carrés représentent les B (1) et les ronds représentent les NB (0)

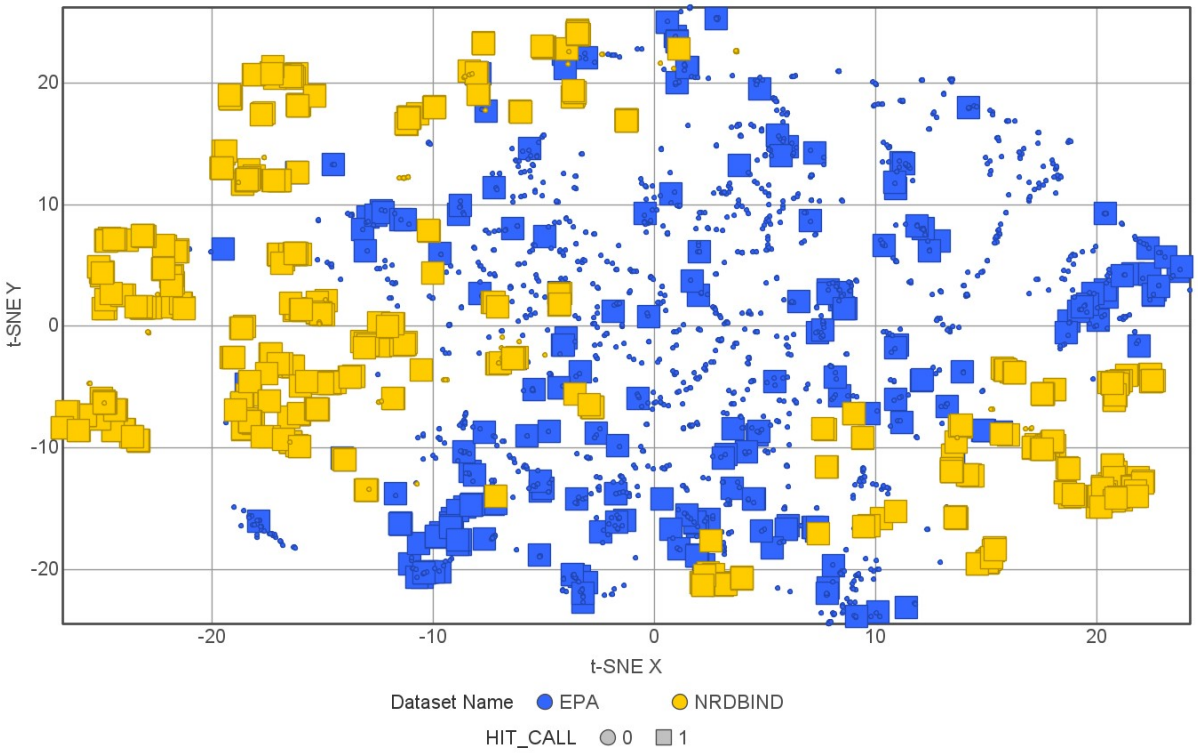


Figure 35 : Distribution des composés B et NB du jeu de données de l'EPA et celui de la NR-DBIND pour le récepteur ERβ. Pour chaque jeu de données, Les carrés représentent les B (1) et les ronds représentent les NB (0)

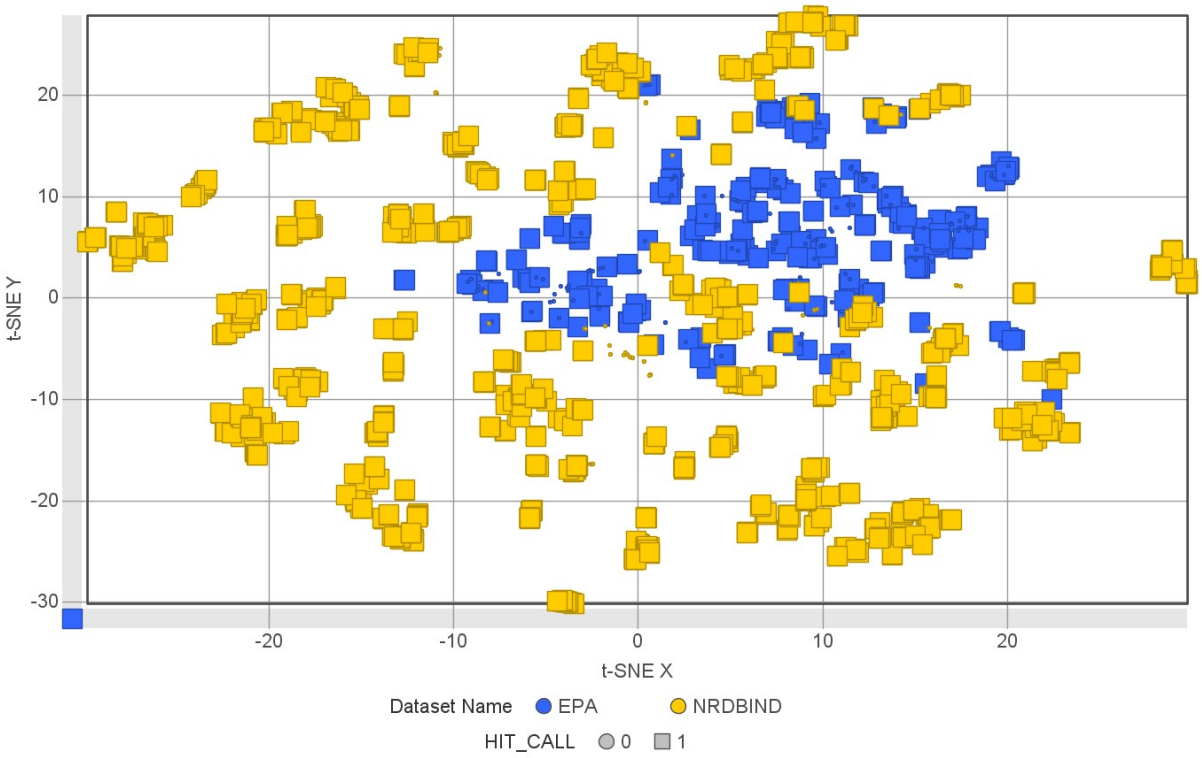


Figure 36 : Distribution des composés B et NB du jeu de données de l'EPA et celui de la NR-DBIND pour le récepteur GR. Pour chaque jeu de données, Les carrés représentent les B (1) et les ronds représentent les NB (0)

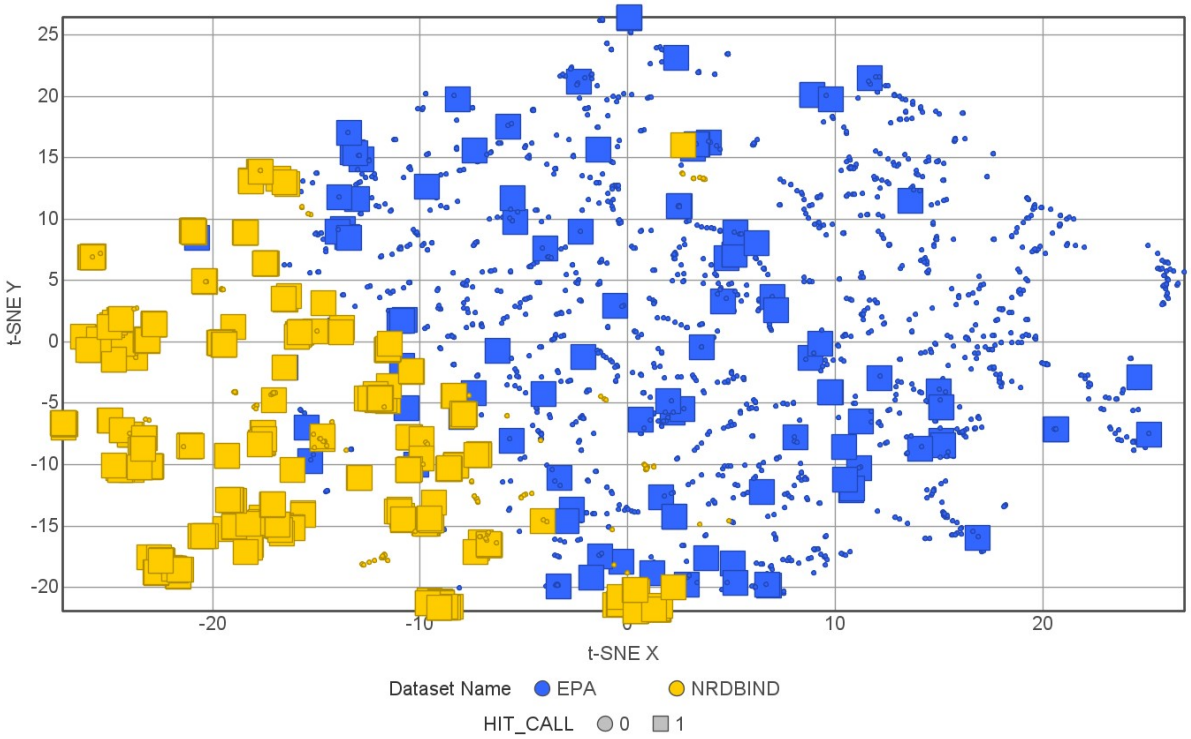


Figure 37 : Distribution des composés B et NB du jeu de données de l'EPA et celui de la NR-DBIND pour le récepteur PPARγ. Pour chaque jeu de données, Les carrés représentent les B (1) et les ronds représentent les NB (0)

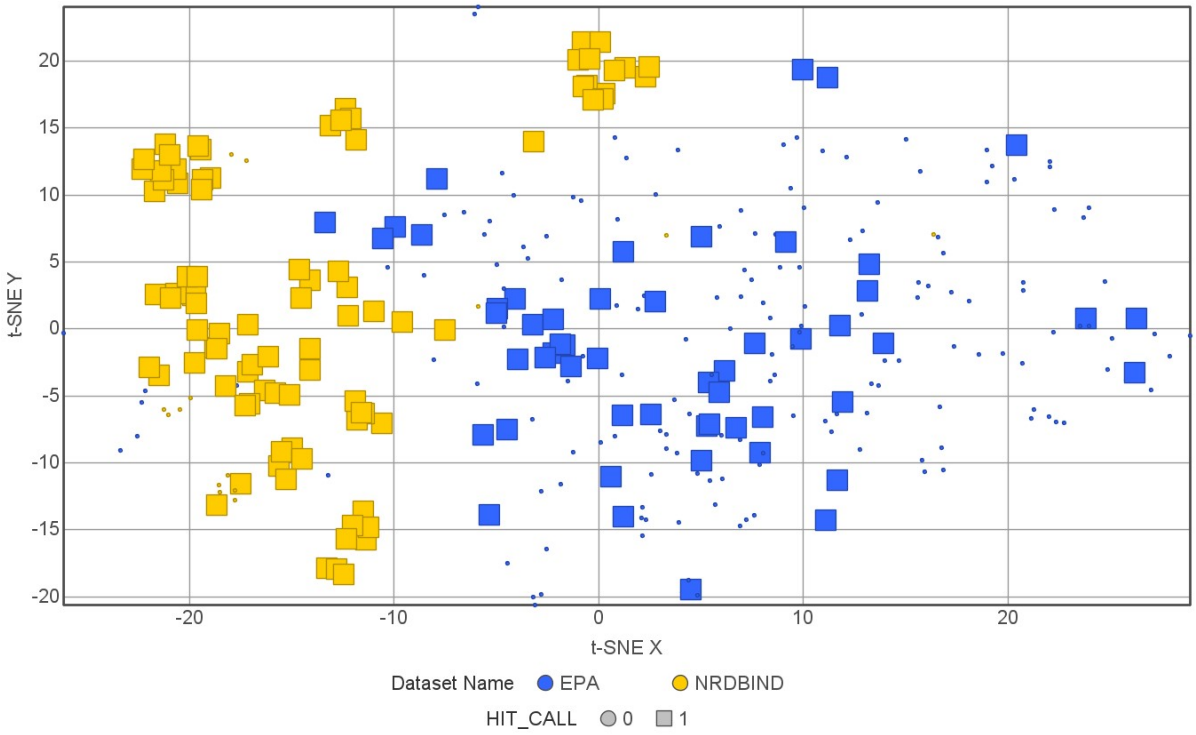


Figure 38 : Distribution des composés B et NB du jeu de données de l'EPA et celui de la NR-DBIND pour le récepteur TRα. Pour chaque jeu de données, Les carrés représentent les B (1) et les ronds représentent les NB (0)

2.1.1.2. Mise en application

L'objectif de ce projet était de proposer un outil regroupant des modèles SB et LB prédisant la capacité des composés chimiques à se lier aux NR, dans l'optique d'identifier rapidement de potentiels PE devant être plus amplement investigués. Les NR étant nombreux (on en compte 48 chez l'homme) et par contrainte de temps, nous avons restreint ce travail à 6 NR principalement liés à la problématique des PE. L'outil de prédiction, une fois terminé, permettra de fournir une prédiction qualitative relative à chaque NR se matérialisant sous forme de liste de priorités à trois niveaux « Haut », « intermédiaire » et « faible ». Cette liste pourra être utilisée comme « thermomètre » pour savoir quels composés sont à tester en urgence permettant ainsi de mieux orienter les ressources et d'optimiser le temps nécessaire à l'identification de nouveaux PE. Afin de montrer un exemple d'application de l'outil de prédiction, nous présentons ici les résultats préliminaires obtenus avec la EDList I^{365}. Cette liste contient des substances qui ont été évaluées pour leurs propriétés perturbatrices du système endocrinien et légalement identifiées comme PE selon différentes réglementations de l'UE i.e PPPR (*Plant*

Protection Product Regulation), BPR (*Biocidal Product Regulation*) ou REACH (*Registration, Evaluation, Authorisation and Restriction of Chemicals*).

Il est à noter qu'il est très difficile de trouver des données sur les PE non incluses dans les données de l'EPA. La plupart des composés de l'EDList I ont d'ailleurs déjà été testés pour le binding et sont inclus dans nos données d'entrainement. Nous avons donc écarté ces composés et nous avons criblé le reste avec nos modèles. Au total le criblage s'est fait sur 733 molécules et les différentes prédictions sont illustrées dans le **Tableau 16**.

Tableau 16 : Bilan des catégories prédites pour les différents PE de la EDList

DB	Probabilité forte « *High* »	Probabilité moyenne « *Moderate* »	Probabilité faible « *Low* »
AR	108	228	397
ERA	149	374	210
ERB	160	328	245
GR	0	430	303
PPARG	7	321	405
TRA	0	43	690

Bien que ces résultats soient le reflet de la capacité de prédiction de nos modèles, nous remarquons toutefois que beaucoup de composés ont été prédits comme ayant une probabilité faible de se lier aux différents NR. L'*Upset plot* de la **Figure 39**, montre la distribution de ces composés prédits « *low* » par les différents modèles de NR. Ainsi, 122 composés sur les 733 testés sont aussi prédits comme NB par la totalité des NR. De plus, le modèle de TRα est associé au plus grand nombre de prédictions « *Low* » des différents modèles avec 144 composés. Ceci peut s'expliquer par un plus faible nombre de PE capable d'interagir avec TRα mais aussi parce que le modèle de TRα repose sur les pharmacophores LB uniquement avec des performances assez réduites.

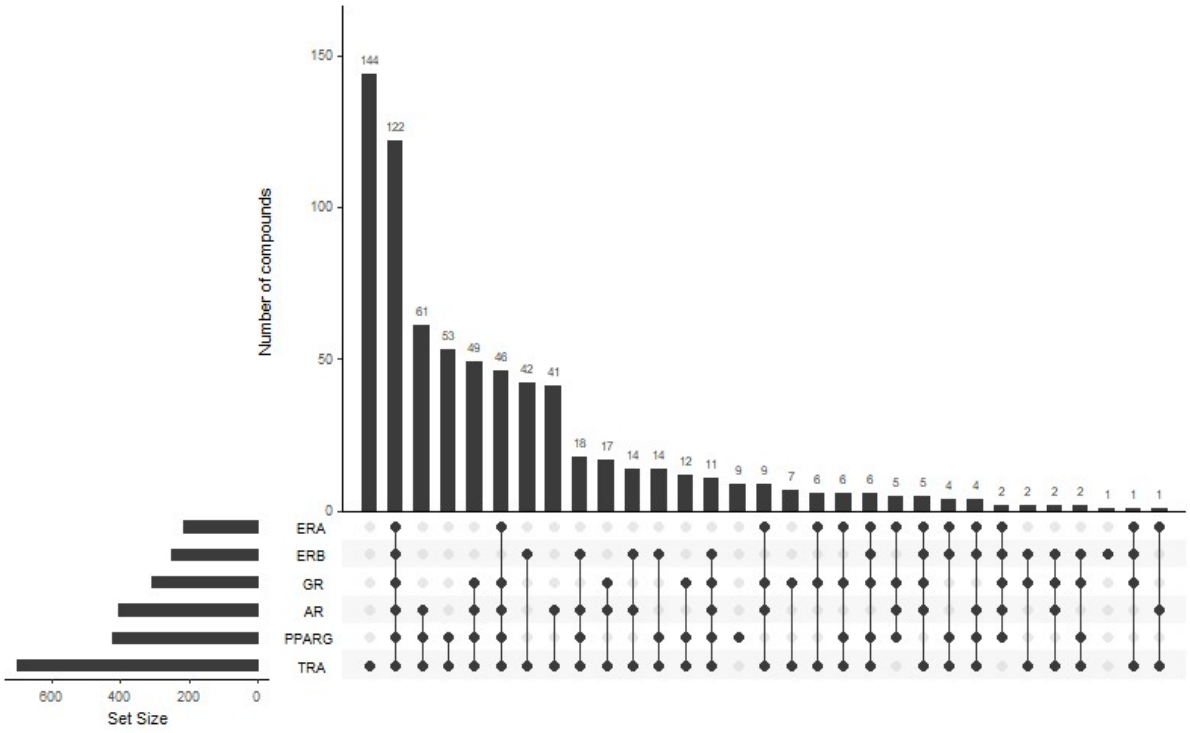


Figure 39: *Upset* plot des distributions des composés prédits comme ayant un faible probabilité (*LOW*) de se lier au NR correspondants à nos modèles

Dans ce contexte d'analyse, outre la mise en cause des performances de nos modèles, cela pourrait montrer que ces 122 composés, agissent via le mécanisme direct sur d'autres NR non modélisés ici ou avec d'autres mécanismes différents de la liaison directe. Ainsi il serait intéressant de réaliser comme suite de ce travail des modèles de prédiction correspondants aux autres NR pertinents dans le contexte des PE et aux différents mécanismes connus et modélisables. Le détail des prédictions obtenues pour les 786 molécules de l'EDList I est présenté en annexes.

2.1.5. Perspectives

2.1.5.1. Constitution et composition des jeux de données

La collecte des données constitue une difficulté majeure lors de la réalisation du projet. En effet, le but était de baser nos travaux sur un ensemble de composés à caractères toxicologique et d'utiliser des vrais inactifs. Pour ce faire, nous avons utilisé au départ le tableau de bord Toxcast à partir duquel nous avons téléchargé notre base de données pour le modèle ERα. Cette base a

cependant migré vers la Comptox toujours en cours de développement. Cela a conduit à un manque de données pour certains NR et pose un souci pour la reproductibilité du protocole de préparation des bases de données puisque ces données ont évolué entre le moment de la collecte et aujourd'hui.

2.1.5.1.1. Composition d'un jeu de données

Beaucoup de composés figurant dans la base de données sont suspectés ou avérés dangereux. Toutefois, le test de binding sur le composé peut parfois être biaisé car il ne tient pas compte du facteur de métabolisme. En effet, certains PE avérés ne sont pas capables de se lier aux NR directement, ce sont leurs métabolites qui sont responsables de la PE en se liant aux NR. C'est notamment le cas du Méthoxychlor, métabolisé *in vivo* en (2,2-bis-(p-hydroxyphényl)- 1,1,1-trichloroethane (HPTE) and 2,2-bis-(p-hydroxyphényl)- 1,1,1-dichloroethane (HPDE)). Ces métabolites et notamment le HPTE interagissent avec les ER et AR [366].

Par ailleurs il s'est avéré qu'à la suite du binding des PE aux NR, deux voies pouvaient être activées : la voie génomique et la non-génomique respectivement liées à des fortes et faibles affinités de liaison du ligand. La voie génomique est celle qui affecte la liaison de ER avec l'ADN ou à d'autres NR se liant par la suite à l'ADN nécessitant ainsi une forte affinité du composé pour le récepteur. Lors de la voie non génomique, l'exposition aux estrogènes entraine une activation rapide d'une cascade de kinases qui peut se manifester même avec une faible affinité [335]. En effet, une étude menée sur une large gamme de composés environnementaux a révélé que les PE étaient caractérisés par une diversité structurale en plus du fait d'avoir plusieurs intervalles d'affinités allant du sub-nanomolaire au micromolaire [367]. Ainsi, il serait intéressant d'établir deux modèles sur les bases des affinités afin de mieux cerner les caractéristiques des PE et mieux comprendre leur mécanisme d'action[335]. Toutefois, une vigilance importante doit être accordée lors de l'interprétation de ces résultats de modélisation par affinité car les concentrations des composées pour les tests ne sont pas mentionnées.

Enfin, il faut mentionner que les profils agonistes/ antagonistes des différents composés testés pour le *binding* sont issus de concaténation des résultats des différents tests biologiques disponibles au niveau du tableau de bord Comptox. Les profils ont été relevés dans un but prospectif pour potentiellement observer une tendance dominante aux niveaux des composés B. Toutefois, à l'avenir pour modéliser ces activités génomiques, il faudra prêter une attention particulière aux différents tests et résultats selon les recommandations de la littérature [330,331,368].

2.1.5.1.2. Déséquilibre des jeux de données et validation

Pour nos différents modèles sauf GR, nous comptons un important déséquilibre entre les données actives et inactives. Bien que ce déséquilibre soit représentatif de la réalité, il amène un biais au niveau des performances qu'il faut interpréter avec soin. Ce problème a été analysé au niveau de la revue présentée dans la section XX, dans laquelle nous avons listé les différentes méthodes abordées pour le surmonter. Ici, nous avons choisi de nous baser sur un ensemble de métriques parmi lesquelles certaines (MCC et F1) permettent de prendre en compte le déséquilibre dans le jeu de données. Une autre piste d'amélioration aurait pu être de réduire le nombre d'inactifs à proportion avec une approche d'*« Under sampling »* [369]. Le principe de cette méthode est de diviser aléatoirement les composés inactifs en *n* sous-groupes dont le nombre est proportionnel au nombre d'actifs. Ainsi, *n* jeux d'entrainement sont obtenues et utilisables pour générer *n* modèles. Le modèle le plus récurrent ou le plus pertinent est ensuite choisi selon le contexte.

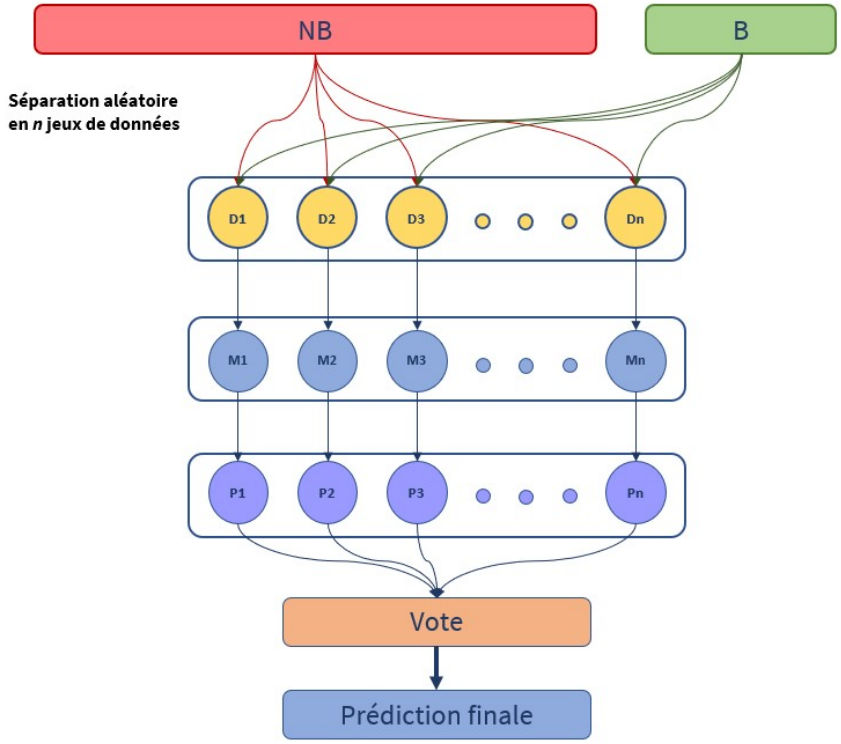


Figure 40 : Protocole d' « *undersampling »* permettant de remédier au déséquilibre entre actifs et inactifs d'après [369].

2.1.5.2. Modèle de docking

Les méthodes de docking se sont imposées depuis leurs création comme la référence en termes de méthodes SB [201]. Elles reposent comme leur nom l'indique sur une structure de la protéine. Le choix de la ou des structures à utiliser comme point de départ, critique en docking protéine rigide reste difficile.

Il existe dans la PDB des PE confirmés co-cristallisés avec les NR sur lesquels ils agissent. Certaines des structures considérées pour notre étude font même partie de cette catégorie. Par exemple la structure 1x7j représente le cristal du complexe ERα avec la Genistein, un phytoœstrogène connu pour être un PE. La structure 3erd représente aussi le même récepteur en complexe avec le diéthylstilbestrol (DES). Cependant, les résultats de docking ont montré que ces structures n'étaient pas celles qui donnaient les meilleurs résultats.

Par ailleurs, les PE ayant une affinité de liaison plus faible que les ligands endogènes[335], il aurait été intéressant d'explorer aussi des structures en complexe avec des agonistes ou antagonistes partiels et en l'occurrence inclure des structures humaines même avec des mutations comme par exemple le mutant Y537S de ERα [370]. La mutation pratiquée sur cette structure permet de stabiliser la conformation active du récepteur et facilite donc la cristallisation d'agonistes à faibles affinités sans pour autant changer l' « architecture » globale du LBD et sans affecter me mode de liaison [367].

Lors de l'étape de docking, nous avons sélectionné les structures de façon « simpliste ». L'idée des travaux était de trouver un protocole standard à suivre pour les 6 NR étudiés, transposable à d'autres NR. Malheureusement, parmi les paramètres qui affectent les performances de docking, se trouve le choix de la structure et du profil pharmacologique du ligand associé. Cela a été démontré par des études précédentes au laboratoire [198,360] mais peut être observé au niveau de l'étude de docking pour AR. En effet, pour ce NR, la structure ayant la meilleure performance est une structure associée à SARM ce qui fait sens puisque la majorité des ligands étudiés possèdent un profil agoniste et antagoniste en même temps. Certains PE se liant à ER agissent aussi en tant que modulateurs. Malheureusement, les deux structures sélectionnées pour les deux isoformes ne correspondent pas à des structures associées aux SERMs, mais plutôt des agonistes. Le choix de la structure pour le modèle de docking est donc un critère important qu'il est difficile de rationaliser et réaliser de façon automatique en se basant uniquement sur le critère des performances.

Par ailleurs il serait intéressant de comparer les performances obtenues dans cette étude avec les performances obtenues en créant des modèles spécifiques de chaque profil pharmacologique. Ainsi, des jeux de données spécifiques pourraient être créés :

(1) un jeu où les agonistes seraient actifs et le reste des molécules (antagonistes, autres profils pharmacologiques et inactifs expérimentaux) seraient considérés inactifs pour les structures en conformation agoniste

(2) un jeu où les antagonistes seraient actifs et le reste (agonistes, autres profils pharmacologiques et inactifs expérimentaux) seraient considérés inactifs pour les structures en conformation antagoniste

(3) un jeu où les molécules agonistes et antagonistes seraient actives et le reste (autres profils pharmacologiques et inactifs expérimentaux) seraient considérés inactifs pour les structures co-cristallisées avec un modulateur.

Notons que cette exploration a été faite lors de la preuve de concept pour ERα, les meilleures performances ont été obtenues lorsque la base de données était considérée comme le cas (2), cependant la meilleure structure est la même pour les deux jeux de données qu'avec la base de données ou tous les B sont considérés comme actifs. Cela peut essentiellement s'expliquer par les proportions d'agonistes prédominantes sur cette base de données biaisant potentiellement les résultats. Une piste à explorer serait pour les modèles de réaliser pour chaque NR les études de docking sur les différents profils de structures disponibles et de sélectionner la meilleure de chacune afin de combiner leurs résultats ultérieurement.

Chaque logiciel de docking présente des inconvénients qu'il est essentiel d'appréhender lors de la modélisation. C'est pour cela que nous avons réalisés nos docking avec différents logiciels arborant différents algorithmes d'échantillonnage et différentes fonctions de scores.
Nous aurions pu aller au-delà de la comparaison des performances des différents logiciels de docking et essayer d'explorer deux autres approches à savoir le « rescoring » et le docking consensus. Le rescoring fait référence au traitement d'une liste de poses de docking avec une fonction de score qui n'est pas utilisée lors du protocole de docking, ce qui conduit généralement à de meilleures performances globales. Les fonctions de scores *knowledge based* sont normalement plus performantes pour la prédiction de la pose [371], tandis que les fonctions empiriques sont mieux adaptées à la prédiction de l'affinité de liaison si des poses natives sont

disponibles [371,372]. Ainsi, la meilleure pose élue par une fonction de score sera réévaluée avec une autre. En ce qui concerne le docking consensus, il s'agit d'un consensus aussi bien pour les étapes d'échantillonnage que les étapes de scoring. Cette méthode consensus a déjà fait ses preuves en compensant les faiblesses des fonctions de scores prises individuellement [201,373].
Cependant malgré cette diversité, toutes les fonctions de scores abordées dans ce projet appartiennent à la catégorie des fonctions de score empirique et donc dépendent du jeu d'entrainement avec lequel elles ont été développées. Ces dernières sont construites sur la base des affinités de liaison entre un ensemble de complexes ligand-récepteur connus et tentent de reproduire les données expérimentales [155,196,374–376]. La question se pose sur leur adéquation pour représenter/prédire les interactions entre un PE et les NR surtout que les PE montrent expérimentalement des affinités plus faibles que les ligands endogènes ou les médicaments par exemple. Ce point est à explorer plus en détails et pourrait expliquer les performances limitées de nos modèles. La perspective majeure de ce projet est de développer une fonction de score adaptée pour les NR, qui sera facilitée par le fait que nous disposons déjà d'une base de données sur les NR [216].

2.1.5.3. Modèles de pharmacophores

Les structures employées pour la modélisation de pharmacophore SB étaient soumises à des critères de sélection assez strictes i.e. structure humaine, sans mutation et référencé par un article au moins. Beaucoup de ces mutations existent pour un but de stabilisation du cristal et ont permis la mise en évidence de protéines en complexes avec des ligands intéressants pour notre étude à savoir des PE. Il aurait pu être intéressant d'explorer toutes les structures PDB humaines référencées afin d'explorer tous les modèles pharmacophoriques possibles.
Par ailleurs, une autre critique concerne le critère d'astringence appliqués à nos modèles de pharmacophore. En effet, lors de l'étape d'optimisation, un modèle de pharmacophore est retenu lorsqu'il engendre une *hitlist* où le ratio B/NB est supérieur ou égal à 1 ce qui est très permissif et se reflète sur les performances. Dans le processus de découverte de médicament, le criblage est un succès même si on n'identifier pas la totalité des composés actifs. En effet, le but est de réduire le nombre total de molécules à tester en une petite fraction dans laquelle on espère identifier des *hits* qu'on pourra plus tard optimiser. En contrepartie, le criblage virtuel en toxicologie vise à retrouver le maximum de composés actifs, puisque « rater » un de ses composés entraînerait un risque de sécurité très important. Beaucoup de travaux dans la littérature ont abordé l'utilisation des modèles pharmacophoriques en toxicologie et deux grandes stratégies ressortent. La première est de construire des modèles assez permissifs

(généralement à 3 ou 4 points pharmacophoriques) ou incorporant des points pharmacophoriques optionnels[50]. Cela entraine l'augmentation du nombre de FP comme pour notre cas mais garantie la diversité structurale des composés retrouvés. La deuxième stratégie serait de réaliser des criblages parallèles de plusieurs modèles restrictifs i.e à forte Sp [377]. Pour aller dans ce sens, l'augmentation du seuil du ratio B/NB n'a pas abouti à de bons résultats et a conduit à des modèles qui n'arrivaient pas à détecter les B. Une autre approche possible aurait aussi pu être de grouper les composés selon leur profil agoniste ou antagoniste et de créer des modèles spécifiques de chaque profil. Une précédente étude au laboratoire a permis de faire cela à partir de la base de donnée NRListBDB conduisant à l'obtention de modèles à hautes performances capables de discriminer les agonistes des antagonistes [339]. Les modèles générés pourraient ainsi être groupés pour élargir leurs capacités de prédictions. De plus ils pourraient être comparés aux modèles précédemment fait pour en déduire les différences inhérentes aux composés toxiques. Cependant, la diversité des profils pharmacologiques autres qu'agoniste et antagoniste et parfois le manque de données biologiques à ce sujet rendent cette étude compliquée.

2.1.5.4. Combinaison des modèles

Notre travail a conduit à l'obtention de 6 modèles chacun spécifique d'un NR étudié à partir de données toxicologiques issues de tests de liaison. Pour 4 de ces récepteurs i.e AR, ERα, ERβ et PPARγ ces modèles sont un consensus entre des modèles de docking et de pharmacophores. Ces derniers étaient associés à des valeurs de Se et de NPV satisfaisantes mais à de plus faibles valeurs de Sp. Pour les deux NR restants i.e GR et TRα, les modèles sélectionnés sont respectivement un modèle de docking et un modèle de pharmacophore entrainant des performances moins satisfaisantes que les autres. Beaucoup de paramètres expliquent ces performances notamment le nombre de composés inclus dans chaque jeu, le nombre de B supérieur au nombre des NB pour le jeu de GR et aussi la qualité et le nombre de structures de la protéine disponibles pour ces 2 NR. Toutes ces constations nous poussent à envisager de ne pas inclure ces modèles ou essayer de pallier les problèmes décrits afin d'améliorer le modèle final de prédiction.

2.1.5.5. Autres méthodes *in silico*

Ce travail s'est concentré essentiellement sur deux grandes méthodes de criblage virtuel i.e. le docking et les modèles de pharmacophores. Il serait intéressant dans une optique d'ouverture

d'explorer l'apport d'autres méthodes *in silico* comme la dynamique moléculaire qui donne accès à un plus large panel conformationnel et permet par exemple d'étudier le temps de résidence d'un ligand ou encore d'effectuer des calculs d'énergie libre qui peuvent affiner l'estimation de l'évaluation de l'affinité du ligand pour les différentes cibles. Les méthodes QSAR, largement employées en toxicologie prédictive pourront aussi intégrées aux protocoles pour améliorer les performances de prédiction des B des NR.

2.1.6. Conclusion

Pour donner suite à la réalisation de la preuve de concept sur ERα, l'étude a été élargie à 5 autres NR incriminés dans le mécanisme de perturbation endocrinienne. Des modèles de docking et de pharmacophores SBLB ont été construits et leurs performances ont été évaluées individuellement et en combinaison. Cela a donné lieu à 6 modèles pouvant être utilisés séparément ou collectivement pour cribler une liste de composés suspectés toxiques pour prédire leurs capacités à se lier aux NR. Toutefois, les performances de ces modèles sont à interpréter en fonction du contexte et surtout de la nature et la structure des composés à cribler. La principale difficulté de ce travail est la diversité des jeux de données aussi bien dans la nature et l'origine des composés, la structure et le profil pharmacologique constituant un frein à la réalisation d'un modèle standard et efficace. Dans ce sens, une perspective intéressante à explorer serait de créer des modèles plus spécifiques et d'associer ainsi leurs performances [67,352]. Nous pensons à des modèles construits à partir de jeux de composés agonistes et de jeux de données de composés antagonistes, des modèles construits à partir de sous-ensembles définis selon des intervalles de Ki ou des modèles à partir de catégories spécifiques de composé comme les pesticides ou les additifs.

Par ailleurs, il est important de noter que ce projet à caractère exploratoire constitue la première tentative du laboratoire en matière de toxicologie prédictive et c'est pour cette raison que nous n'intégrons pas tous les NR. Une fois un protocole satisfaisant obtenu, nous pourrons l'étendre à d'autres et spécialement, PR pour compléter le groupe des récepteurs aux hormones sexuelles, MR pour compléter GR, les autres isotypes de PPAR.

Malgré l'efficacité et la popularité croissante des méthodes de criblage virtuel, il faut être conscient de leurs limites. Par exemple, l'approche présentée ici ne permettent pas de modéliser le pouvoir de perturbation endocrinienne globale. Elle ne permet que de couvrir que le volet du mécanisme d'action direct. Pour développer un modèle de prédiction au sens large, il faudra

modéliser les autres mécanismes d'action comme l'effet sur les protéines de transport ou encore la génotoxicité et les intégrer dans le cadre de prédictions hiérarchisées. Une limite de la modélisation *in silico* du potentiel PE est liée aux mécanismes particuliers des PE comme leur capacité à réaliser de liaisons covalentes avec les NR ou encore l'effet cocktail encore difficile à modéliser [378].

Troisième Partie
Conclusion

Les méthodes de criblage virtuel constituent un atout majeur d'accélération de la recherche. Dans le domaine pharmaceutique, elles sont quasiment incontournables en phases pré-cliniques. Elles permettent notamment de cribler des bases de données afin de prédire des hits pour une cible thérapeutique d'intérêt mais aussi pour déceler des composés à potentiel effet délétère qu'il est nécessaire d'écarter de la liste des candidats thérapeutiques avant de s'engager dans des phases d'optimisations chimiques et de tests *in vitro* et *in vivo*. En toxicologie, elles sont aussi très utilisées, souvent intégrées à des systèmes experts et couplées avec d'autres méthodes *in silico* et d'autres connaissances empiriques afin d'identifier les risques des composés chimiques des différentes industries. Elles peuvent notamment être utilisées dans un cadre de toxicologie environnementale où le but est d'assurer la sécurité de l'Homme et de l'environnement en décelant des substances dangereuses qu'on retrouverait dans les aliments, les objets du quotidien ou même l'air que l'on respire. Les méthodes de criblage virtuel permettent alors d'associer une probabilité de danger aux différents composés à évaluer et donc de créer une liste de priorité orientant ainsi les travaux expérimentaux. La toxicologie environnementale s'intéresse à de nombreux phénomènes et catégories de composés dont les perturbateurs endocriniens (PE).

Les PE constituent un problème majeur de santé humaine et environnementale[369]. Ils ont été associés à des manifestations pathologiques relatives aux fonctions hormonales notamment les hormones mâles et femelles et très vite leurs mécanismes d'action ont été reliés aux récepteurs nucléaires ER et AR [224]. L'intérêt grandissant pour ces substances a permis de révéler que le périmètre d'action des PE s'entendait à plusieurs autres membres de la famille des récepteurs nucléaires[378] entrainant des répercussions sur le métabolisme, sur l'obésité, l'immunité et l'apparition de certains cancers. Le mode d'action des PE est assez complexe et n'est jusqu'à aujourd'hui pas complétement résolu. Toutefois, l'étude des PE se focalise le plus souvent sur l'étude de leur mécanisme d'action directe à savoir leur capacité à se fixer aux NR et à bloquer ou suractiver leur fonctionnement normal [311]. Dans ce contexte, les méthodes de criblage virtuel permettent d'aider dans l'élucidation du mécanisme d'interaction et dans la découverte de nouveaux PE.

Ce travail s'inscrit dans une initiative de réalisation d'un outil de criblage *in silico* combinant des méthodes SB et LB dans le but de prédire la capacité de composés chimiques à interférer avec les NR et donc leur potentiel effet PE.

Tout d'abord une évaluation de l'état de l'art dédié à l'emploi des méthodes *in silico* dans le champ d'application des NR a été réalisée. Cette étude a révélé le déséquilibre des travaux sur

les différents récepteurs humains et leurs focalisations sur quelques membres. En effet, notre investigation a mené à l'identification d'études pour 14 NR uniquement (sur 48 NR humains). Par ailleurs, l'écart se creuse encore plus entre ces derniers, puisque la majorité de travaux se concentre sur les récepteurs hormonaux AR et ER spécialement. Au cours de cette étude, nous avons relevé la manière avec laquelle les méthodes de criblage virtuel étaient employées aussi bien pour des applications thérapeutiques que toxicologiques, nous avons réalisé une analyse critique de différents aspects de la modélisation allant de la nature des données à la reproductibilité des méthodes.

Ensuite, nous avons entrepris la réalisation d'une preuve de concept pour notre protocole de criblage. Pour cela, nous avons utilisé les données de toxicité de ERα au regard de son implication dans les mécanismes liés aux PE et de la quantité des données disponibles. Une fois cette étude publiée et présentée au cours de deux communications orales, nous avons décidé d'élargir ce protocole à d'autres NR incriminés dans le mécanisme de perturbation endocrinienne. Notre choix s'est porté sur 5 NR à savoir ERβ, AR, GR, PPARγ et TRα. Ainsi, au total, nos travaux ont permis de construire 6 modèles *in silico* combinant des méthodes SB (docking et modèles de pharmacophores SB) et une méthode LB (modèle de pharmacophores LB). Ces modèles appliqués dans l'identification de composés capables de se lier à ces NR et donc à la prédiction de potentiels PE ont permis d'obtenir des bonnes performances de prédiction pour 4 des 6 modèles avec des valeurs de sensibilité élevées. *In fine*, ces modèles pourront être utilisés pour cribler des collections de données et établir une liste de composés à tester expérimentalement en priorité. Le domaine d'application de ces modèles doit être considéré avant leur utilisation pour établir de nouvelles prédictions.

Ce travail constitue une première expérience du laboratoire dans l'application des méthodes de criblage virtuel en toxicologie humaine et en santé environnementale et a donc nécessité un changement de paradigme dans l'interprétation des résultats. Plusieurs pistes d'améliorations ont été suggérées comprenant la gestion du déséquilibre des données et l'optimisation des différents protocoles. Une perspective de modélisation du mécanisme de perturbation endocrinienne plus générale serait aussi à envisager en incluant ces modèles dans une hiérarchie d'autres modèles concentrés sur d'autres mécanismes d'action «indirects ».

Bibliographie

(1) Gilbert, S.; Mohapatra, A.; Bobst, S.; Hayes, A.; Humes, S. T. *Information Resources in Toxicology, Volume 1: Background, Resources, and Tools*; Academic Press, 2020.

(2) *Toxicology: a primer - ScienceDirect*. https://www-sciencedirect-com.proxybib-pp.cnam.fr/science/article/pii/B9780128137246000013 (accessed 2022-06-26).

(3) Kavlock, R. J.; Ankley, G.; Blancato, J.; Breen, M.; Conolly, R.; Dix, D.; Houck, K.; Hubal, E.; Judson, R.; Rabinowitz, J.; Richard, A.; Setzer, R. W.; Shah, I.; Villeneuve, D.; Weber, E. Computational Toxicology—A State of the Science Mini Review. *Toxicol. Sci.* **2008**, *103* (1), 14–27. https://doi.org/10.1093/toxsci/kfm297.

(4) Seidle, T.; Stephens, M. L. Bringing Toxicology into the 21st Century: A Global Call to Action. *Toxicol. In Vitro* **2009**, *23* (8), 1576–1579. https://doi.org/10.1016/j.tiv.2009.06.012.

(5) Zurlo, J.; Rudacille, D.; Goldberg, A. M. The Three Rs: The Way Forward. *Environ. Health Perspect.* **1996**, *104* (8), 878–880. https://doi.org/10.1289/ehp.96104878.

(6) Watson, K. D.; Wexler, P.; Everitt, J. M. CHAPTER 1 - History. In *Information Resources in Toxicology (Third Edition)*; Wexler, P., Ed.; Academic Press: San Diego, 2000; pp 1–25. https://doi.org/10.1016/B978-012744770-4/50042-1.

(7) Raies, A. B.; Bajic, V. B. In Silico Toxicology: Computational Methods for the Prediction of Chemical Toxicity. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2016**, *6* (2), 147–172. https://doi.org/10.1002/wcms.1240.

(8) Waring, M. J.; Arrowsmith, J.; Leach, A. R.; Leeson, P. D.; Mandrell, S.; Owen, R. M.; Pairaudeau, G.; Pennie, W. D.; Pickett, S. D.; Wang, J.; Wallace, O.; Weir, A. An Analysis of the Attrition of Drug Candidates from Four Major Pharmaceutical Companies. *Nat. Rev. Drug Discov.* **2015**, *14* (7), 475–486. https://doi.org/10.1038/nrd4609.

(9) Coumoul, X.; Massicot, F.; Pairon, J.-C. *Toxicologie*; Dunod, 2020.

(10) Alasuvanto, T.; Gissi, A.; Sobanski, T.; Karamertzanis, P.; Rasenberg, M. (Q)SARs as Adaptations to REACH Information Requirements. *Methods Mol. Biol. Clifton NJ* **2018**, *1800*, 107–115. https://doi.org/10.1007/978-1-4939-7899-1_4.

(11) Zbinden, G.; Flury-Roversi, M. Significance of the LD50-Test for the Toxicological Evaluation of Chemical Substances. *Arch. Toxicol.* **1981**, *47* (2), 77–99. https://doi.org/10.1007/BF00332351.

(12) Hemmerich, J.; Ecker, G. F. In Silico Toxicology: From Structure–Activity Relationships towards Deep Learning and Adverse Outcome Pathways. *WIREs Comput. Mol. Sci.* **2020**, *10* (4), e1475. https://doi.org/10.1002/wcms.1475.

(13) Kim, J. H.; Scialli, A. R. Thalidomide: The Tragedy of Birth Defects and the Effective Treatment of Disease. *Toxicol. Sci. Off. J. Soc. Toxicol.* **2011**, *122* (1), 1–6. https://doi.org/10.1093/toxsci/kfr088.

(14) The Principles of Humane Experimental Technique. *Med. J. Aust.* **1960**, *1* (13), 500–500. https://doi.org/10.5694/j.1326-5377.1960.tb73127.x.

(15) Szymański, P.; Markowicz, M.; Mikiciuk-Olasik, E. Adaptation of High-Throughput Screening in Drug Discovery—Toxicological Screening Tests. *Int. J. Mol. Sci.* **2012**, *13* (1), 427. https://doi.org/10.3390/ijms13010427.

(16) Dix, D. J.; Houck, K. A.; Martin, M. T.; Richard, A. M.; Setzer, R. W.; Kavlock, R. J. The ToxCast Program for Prioritizing Toxicity Testing of Environmental Chemicals. *Toxicol. Sci.* **2007**, *95* (1), 5–12. https://doi.org/10.1093/toxsci/kfl103.
(17) Zhu, H.; Zhang, J.; Kim, M. T.; Boison, A.; Sedykh, A.; Moran, K. Big Data in Chemical Toxicity Research: The Use of High-Throughput Screening Assays To Identify Potential Toxicants. *Chem. Res. Toxicol.* **2014**, *27* (10), 1643–1651. https://doi.org/10.1021/tx500145h.
(18) Dix, D. ToxCast and Tox21: High Throughput Screening for Hazard & Risk of Environmental Chemicals. 34.
(19) Amini, A.; Muggleton, S. H.; Lodhi, H.; Sternberg, M. J. E. A Novel Logic-Based Approach for Quantitative Toxicology Prediction. *J. Chem. Inf. Model.* **2007**, *47* (3), 998–1006. https://doi.org/10.1021/ci600223d.
(20) Reisfeld, B.; Mayeno, A. N. What Is Computational Toxicology? *Methods Mol. Biol. Clifton NJ* **2012**, *929*, 3–7. https://doi.org/10.1007/978-1-62703-050-2_1.
(21) Rusyn, I.; Daston, G. P. Computational Toxicology: Realizing the Promise of the Toxicity Testing in the 21st Century. *Environ. Health Perspect.* **2010**, *118* (8), 1047–1050. https://doi.org/10.1289/ehp.1001925.
(22) Bal-Price, A.; Meek, M. E. (Bette). Adverse Outcome Pathways: Application to Enhance Mechanistic Understanding of Neurotoxicity. *Pharmacol. Ther.* **2017**, *179*, 84–95. https://doi.org/10.1016/j.pharmthera.2017.05.006.
(23) Escher, S. E.; Kamp, H.; Bennekou, S. H.; Bitsch, A.; Fisher, C.; Graepel, R.; Hengstler, J. G.; Herzler, M.; Knight, D.; Leist, M.; Norinder, U.; Ouédraogo, G.; Pastor, M.; Stuard, S.; White, A.; Zdrazil, B.; van de Water, B.; Kroese, D. Towards Grouping Concepts Based on New Approach Methodologies in Chemical Hazard Assessment: The Read-across Approach of the EU-ToxRisk Project. *Arch. Toxicol.* **2019**, *93* (12), 3643–3667. https://doi.org/10.1007/s00204-019-02591-7.
(24) Pyzer-Knapp, E. O.; Suh, C.; Gómez-Bombarelli, R.; Aguilera-Iparraguirre, J.; Aspuru-Guzik, A. What Is High-Throughput Virtual Screening? A Perspective from Organic Materials Discovery. *Annu. Rev. Mater. Res.* **2015**, *45* (1), 195–216. https://doi.org/10.1146/annurev-matsci-070214-020823.
(25) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Deliv. Rev.* **1997**, *23* (1), 3–25. https://doi.org/10.1016/S0169-409X(96)00423-1.
(26) *A generalizable definition of chemical similarity for read-across | Journal of Cheminformatics | Full Text*. https://jcheminf.biomedcentral.com/articles/10.1186/s13321-014-0039-1 (accessed 2022-06-30).
(27) Floris, M.; Olla, S. Molecular Similarity in Computational Toxicology. In *Computational Toxicology: Methods and Protocols*; Nicolotti, O., Ed.; Methods in Molecular Biology; Springer: New York, NY, 2018; pp 171–179. https://doi.org/10.1007/978-1-4939-7899-1_7.
(28) Moriwaki, H.; Tian, Y.-S.; Kawashita, N.; Takagi, T. Mordred: A Molecular Descriptor Calculator. *J. Cheminformatics* **2018**, *10* (1), 4. https://doi.org/10.1186/s13321-018-0258-y.
(29) Rester, U. From Virtuality to Reality - Virtual Screening in Lead Discovery and Lead Optimization: A Medicinal Chemistry Perspective. *Curr. Opin. Drug Discov. Devel.* **2008**, *11* (4), 559–568.
(30) Seifert, M. H. J.; Lang, M. Essential Factors for Successful Virtual Screening. *Mini Rev. Med. Chem.* **2008**, *8* (1), 63–72. https://doi.org/10.2174/138955708783331540.

(31) Kirchmair, J.; Distinto, S.; Schuster, D.; Spitzer, G.; Langer, T.; Wolber, G. Enhancing Drug Discovery through in Silico Screening: Strategies to Increase True Positives Retrieval Rates. *Curr. Med. Chem.* **2008**, *15* (20), 2040–2053. https://doi.org/10.2174/092986708785132843.
(32) Köppen, H. Virtual Screening - What Does It Give Us? *Curr. Opin. Drug Discov. Devel.* **2009**, *12* (3), 397–407.
(33) *Concepts and Applications of Molecular Similarity | Wiley*. Wiley.com. https://www-wiley-com.proxybib-pp.cnam.fr/en-us/Concepts+and+Applications+of+Molecular+Similarity-p-9780471621751 (accessed 2022-06-14).
(34) *Bioactive Diversity and Screening Library Selection via Affinity Fingerprinting | Journal of Chemical Information and Modeling*. https://pubs-acs-org.proxybib-pp.cnam.fr/doi/full/10.1021/ci980105%2B (accessed 2022-06-14).
(35) *Flexsim-X: A Method for the Detection of Molecules with Similar Biological Activity | Journal of Chemical Information and Modeling*. https://pubs-acs-org.proxybib-pp.cnam.fr/doi/10.1021/ci990439e (accessed 2022-06-14).
(36) Mauser, H.; Guba, W. Recent Developments in de Novo Design and Scaffold Hopping. *Curr. Opin. Drug Discov. Devel.* **2008**, *11* (3), 365–374.
(37) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38* (6), 983–996. https://doi.org/10.1021/ci9800211.
(38) Brown, R. D. Descriptors for Diversity Analysis. *Perspect. Drug Discov. Des.* **1996**, *7* (1), 31–49. https://doi.org/10.1007/BF03380180.
(39) Lovrics, A.; Pape, V. F. S.; Szisz, D.; Kalászi, A.; Heffeter, P.; Magyar, C.; Szakács, G. Identifying New Topoisomerase II Poison Scaffolds by Combining Publicly Available Toxicity Data and 2D/3D-Based Virtual Screening. *J. Cheminformatics* **2019**, *11* (1), 67. https://doi.org/10.1186/s13321-019-0390-3.
(40) *Daylight Theory: Fingerprints*. https://www.daylight.com/dayhtml/doc/theory/theory.finger.html (accessed 2022-06-14).
(41) Todeschini, R.; Consonni, V.; Xiang, H.; Holliday, J.; Buscema, M.; Willett, P. Similarity Coefficients for Binary Chemoinformatics Data: Overview and Extended Comparison Using Simulated and Real Data Sets. *J. Chem. Inf. Model.* **2012**, *52* (11), 2884–2901. https://doi.org/10.1021/ci300261r.
(42) *Virtual Screening Software for Drug Discovery | ROCS*. https://www.eyesopen.com/rocs (accessed 2022-06-30).
(43) Yan, X.; Li, J.; Liu, Z.; Zheng, M.; Ge, H.; Xu, J. Enhancing Molecular Shape Comparison by Weighted Gaussian Functions. *J. Chem. Inf. Model.* **2013**, *53* (8), 1967–1978. https://doi.org/10.1021/ci300601q.
(44) Li, S.; Song, Y.; Liu, X.; Li, H. A Rapid Python-Based Methodology for Target-Focused Combinatorial Library Design. *Comb. Chem. High Throughput Screen. 19* (1), 25–35.
(45) Puertas-Martín, S.; Redondo, J. L.; Ortigosa, P. M.; Pérez-Sánchez, H. OptiPharm: An Evolutionary Algorithm to Compare Shape Similarity. *Sci. Rep.* **2019**, *9* (1), 1398. https://doi.org/10.1038/s41598-018-37908-6.
(46) Grant, J. A.; Gallardo, M. A.; Pickup, B. T. A Fast Method of Molecular Shape Comparison: A Simple Application of a Gaussian Description of Molecular Shape. *J. Comput. Chem.* **1996**, *17* (14), 1653–1666. https://doi.org/10.1002/(SICI)1096-987X(19961115)17:14<1653::AID-JCC7>3.0.CO;2-K.
(47) *Molecular Shape and Medicinal Chemistry: A Perspective | Journal of Medicinal Chemistry*. https://pubs.acs.org/doi/10.1021/jm900818s (accessed 2022-06-30).

(48) Molecular Descriptors. In *An Introduction To Chemoinformatics*; Leach, A. R., Gillet, V. J., Eds.; Springer Netherlands: Dordrecht, 2007; pp 53–74. https://doi.org/10.1007/978-1-4020-6291-9_3.
(49) Wermuth, C. G.; Ganellin, C. R.; Lindberg, P.; Mitscher, L. A. Glossary of Terms Used in Medicinal Chemistry (IUPAC Recommendations 1998). *Pure Appl. Chem.* **1998**, *70* (5), 1129–1143. https://doi.org/10.1351/pac199870051129.
(50) Tyagi, R.; Singh, A.; Chaudhary, K. K.; Yadav, M. K. Chapter 17 - Pharmacophore Modeling and Its Applications. In *Bioinformatics*; Singh, D. B., Pathak, R. K., Eds.; Academic Press, 2022; pp 269–289. https://doi.org/10.1016/B978-0-323-89775-4.00009-2.
(51) Richmond, N. J.; Abrams, C. A.; Wolohan, P. R. N.; Abrahamian, E.; Willett, P.; Clark, R. D. GALAHAD: 1. Pharmacophore Identification by Hypermolecular Alignment of Ligands in 3D. *J. Comput. Aided Mol. Des.* **2006**, *20* (9), 567–587. https://doi.org/10.1007/s10822-006-9082-y.
(52) Sadowski, Jens.; Gasteiger, Johann. From Atoms and Bonds to Three-Dimensional Atomic Coordinates: Automatic Model Builders. *Chem. Rev.* **1993**, *93* (7), 2567–2581. https://doi.org/10.1021/cr00023a012.
(53) *Molecular Modeling Software | OpenEye Scientific*. https://www.eyesopen.com (accessed 2022-06-30).
(54) Hawkins, P. C. D.; Skillman, A. G.; Warren, G. L.; Ellingson, B. A.; Stahl, M. T. Conformer Generation with OMEGA: Algorithm and Validation Using High Quality Structures from the Protein Databank and Cambridge Structural Database. *J. Chem. Inf. Model.* **2010**, *50* (4), 572–584. https://doi.org/10.1021/ci100031x.
(55) *OMEGA Theory — Applications, vDev build*. https://docs.eyesopen.com/applications/omega/theory/omega_classic_theory.html (accessed 2022-07-02).
(56) Wolber, G.; Langer, T. LigandScout: 3-D Pharmacophores Derived from Protein-Bound Ligands and Their Use as Virtual Screening Filters. *J. Chem. Inf. Model.* **2005**, *45* (1), 160–169. https://doi.org/10.1021/ci049885e.
(57) *Pharmacophore and Ligand-Based Design - BIOVIA - Dassault Systèmes®*. https://www.3ds.com/products-services/biovia/products/molecular-modeling-simulation/biovia-discovery-studio/pharmacophore/ (accessed 2022-07-02).
(58) *Chemical Computing Group (CCG) | Computer-Aided Molecular Design*. https://www.chemcomp.com/ (accessed 2022-07-02).
(59) Dixon, S. L.; Smondyrev, A. M.; Rao, S. N. PHASE: A Novel Approach to Pharmacophore Modeling and 3D Database Searching. *Chem. Biol. Htmlent Glyphamp Asciiamp Drug Des.* **2006**, *67* (5), 370–372. https://doi.org/10.1111/j.1747-0285.2006.00384.x.
(60) Vuorinen, A.; Schuster, D. Methods for Generating and Applying Pharmacophore Models as Virtual Screening Filters and for Bioactivity Profiling. *Methods* **2015**, *71*, 113–134. https://doi.org/10.1016/j.ymeth.2014.10.013.
(61) Wolber, G.; Dornhofer, A. A.; Langer, T. Efficient Overlay of Small Organic Molecules Using 3D Pharmacophores. *J. Comput. Aided Mol. Des.* **2006**, *20* (12), 773–788. https://doi.org/10.1007/s10822-006-9078-7.
(62) Poptodorov, K.; Luu, T.; Hoffmann, R. D. Pharmacophore Model Generation Software Tools. In *Pharmacophores and Pharmacophore Searches*; John Wiley & Sons, Ltd, 2006; pp 15–47. https://doi.org/10.1002/3527609164.ch2.
(63) *Algorithm 457: finding all cliques of an undirected graph | Communications of the ACM*. https://dl.acm.org/doi/10.1145/362342.362367 (accessed 2022-07-02).

(64) Martin, Y. C.; Bures, M. G.; Danaher, E. A.; DeLazzer, J.; Lico, I.; Pavlik, P. A. A Fast New Approach to Pharmacophore Mapping and Its Application to Dopaminergic and Benzodiazepine Agonists. *J. Comput. Aided Mol. Des.* **1993**, *7* (1), 83–102. https://doi.org/10.1007/BF00141577.
(65) Jones, G.; Willett, P.; Glen, R. C. A Genetic Algorithm for Flexible Molecular Overlay and Pharmacophore Elucidation. *J. Comput. Aided Mol. Des.* **1995**, *9* (6), 532–549. https://doi.org/10.1007/BF00124324.
(66) Barnum, D.; Greene, J.; Smellie, A.; Sprague, P. Identification of Common Functional Configurations Among Molecules. *J. Chem. Inf. Comput. Sci.* **1996**, *36* (3), 563–571. https://doi.org/10.1021/ci950273r.
(67) Schuster, D. Pharmacophore Models for Toxicology Prediction. In *Computational Toxicology*; John Wiley & Sons, Ltd, 2018; pp 121–144. https://doi.org/10.1002/9781119282594.ch5.
(68) Hewitt, M.; Przybylak, K. In Silico Models for Hepatotoxicity. *Methods Mol. Biol. Clifton NJ* **2016**, *1425*, 201–236. https://doi.org/10.1007/978-1-4939-3609-0_11.
(69) Alves, V. M.; Braga, R. C.; Andrade, C. H. Computational Approaches for Predicting HERG Activity. In *Computational Toxicology*; John Wiley & Sons, Ltd, 2018; pp 69–91. https://doi.org/10.1002/9781119282594.ch3.
(70) Banerjee, P.; Eckert, A. O.; Schrey, A. K.; Preissner, R. ProTox-II: A Webserver for the Prediction of Toxicity of Chemicals. *Nucleic Acids Res.* **2018**, *46* (W1), W257–W263. https://doi.org/10.1093/nar/gky318.
(71) Zhu, R.; Hu, L.; Li, H.; Su, J.; Cao, Z.; Zhang, W. Novel Natural Inhibitors of CYP1A2 Identified by in Silico and in Vitro Screening. *Int. J. Mol. Sci.* **2011**, *12* (5), 3250–3262. https://doi.org/10.3390/ijms12053250.
(72) Hochleitner, J.; Akram, M.; Ueberall, M.; Davis, R. A.; Waltenberger, B.; Stuppner, H.; Sturm, S.; Ueberall, F.; Gostner, J. M.; Schuster, D. A Combinatorial Approach for the Discovery of Cytochrome P450 2D6 Inhibitors from Nature. *Sci. Rep.* **2017**, *7* (1), 8071. https://doi.org/10.1038/s41598-017-08404-0.
(73) Parthasarathi, R.; Dhawan, A. Chapter 5 - In Silico Approaches for Predictive Toxicology. In *In Vitro Toxicology*; Dhawan, A., Kwon, S., Eds.; Academic Press, 2018; pp 91–109. https://doi.org/10.1016/B978-0-12-804667-8.00005-5.
(74) Gini, G. QSAR: What Else? *Methods Mol. Biol. Clifton NJ* **2018**, *1800*, 79–105. https://doi.org/10.1007/978-1-4939-7899-1_3.
(75) Ajjarapu, S. M.; Tiwari, A.; Ramteke, P. W.; Singh, D. B.; Kumar, S. Chapter 15 - Ligand-Based Drug Designing. In *Bioinformatics*; Singh, D. B., Pathak, R. K., Eds.; Academic Press, 2022; pp 233–252. https://doi.org/10.1016/B978-0-323-89775-4.00018-3.
(76) Grisoni, F.; Ballabio, D.; Todeschini, R.; Consonni, V. Molecular Descriptors for Structure–Activity Applications: A Hands-On Approach. In *Computational Toxicology: Methods and Protocols*; Nicolotti, O., Ed.; Methods in Molecular Biology; Springer: New York, NY, 2018; pp 3–53. https://doi.org/10.1007/978-1-4939-7899-1_1.
(77) Guha, R.; Willighagen, E. A Survey of Quantitative Descriptions of Molecular Structure. *Curr. Top. Med. Chem.* **2012**, *12* (18), 1946–1956.
(78) Bajorath, J. Integration of Virtual and High-Throughput Screening. *Nat. Rev. Drug Discov.* **2002**, *1* (11), 882–894. https://doi.org/10.1038/nrd941.
(79) Willighagen, E. L.; Mayfield, J. W.; Alvarsson, J.; Berg, A.; Carlsson, L.; Jeliazkova, N.; Kuhn, S.; Pluskal, T.; Rojas-Chertó, M.; Spjuth, O.; Torrance, G.; Evelo, C. T.; Guha, R.; Steinbeck, C. The Chemistry Development Kit (CDK) v2.0: Atom Typing, Depiction, Molecular Formulas, and Substructure Searching. *J. Cheminformatics* **2017**, *9* (1), 33. https://doi.org/10.1186/s13321-017-0220-4.

(80) *Molecular descriptors calculation - Dragon - Talete srl*. http://www.talete.mi.it/products/dragon_description.htm (accessed 2022-07-03).
(81) Koeppen, H.; Kriegl, J.; Lessel, U.; Tautermann, C. S.; Wellenzohn, B. Ligand-Based Virtual Screening. In *Virtual Screening*; John Wiley & Sons, Ltd, 2011; pp 61–85. https://doi.org/10.1002/9783527633326.ch3.
(82) Roy, K.; Ghosh, G. QSTR with Extended Topochemical Atom (ETA) Indices. VI. Acute Toxicity of Benzene Derivatives to Tadpoles (Rana Japonica). *J. Mol. Model.* **2006**, *12* (3), 306–316. https://doi.org/10.1007/s00894-005-0033-7.
(83) Bonche, D.; Trinajstič, N. Overall Molecular Descriptors. 3. Overall Zagreb Indices. *SAR QSAR Environ. Res.* **2001**, *12* (1–2), 213–236. https://doi.org/10.1080/10629360108035379.
(84) Fourches, D.; Muratov, E.; Tropsha, A. Trust, but Verify: On the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research. *J. Chem. Inf. Model.* **2010**, *50* (7), 1189–1204. https://doi.org/10.1021/ci100176x.
(85) Kar, S.; Roy, K.; Leszczynski, J. Applicability Domain: A Step Toward Confident Predictions and Decidability for QSAR Modeling. *Methods Mol. Biol. Clifton NJ* **2018**, *1800*, 141–169. https://doi.org/10.1007/978-1-4939-7899-1_6.
(86) Sahigara, F.; Ballabio, D.; Todeschini, R.; Consonni, V. Defining a Novel K-Nearest Neighbours Approach to Assess the Applicability Domain of a QSAR Model for Reliable Predictions. *J. Cheminformatics* **2013**, *5* (1), 27. https://doi.org/10.1186/1758-2946-5-27.
(87) Aniceto, N.; Freitas, A. A.; Bender, A.; Ghafourian, T. A Novel Applicability Domain Technique for Mapping Predictive Reliability across the Chemical Space of a QSAR: Reliability-Density Neighbourhood. *J. Cheminformatics* **2016**, *8* (1), 69. https://doi.org/10.1186/s13321-016-0182-y.
(88) Tetko, I. V.; Sushko, I.; Pandey, A. K.; Zhu, H.; Tropsha, A.; Papa, E.; Oberg, T.; Todeschini, R.; Fourches, D.; Varnek, A. Critical Assessment of QSAR Models of Environmental Toxicity against Tetrahymena Pyriformis: Focusing on Applicability Domain and Overfitting by Variable Selection. *J. Chem. Inf. Model.* **2008**, *48* (9), 1733–1746. https://doi.org/10.1021/ci800151m.
(89) Grulke, C. M.; Williams, A. J.; Thillanadarajah, I.; Richard, A. M. EPA's DSSTox Database: History of Development of a Curated Chemistry Resource Supporting Computational Toxicology Research. *Comput. Toxicol.* **2019**, *12*, 100096. https://doi.org/10.1016/j.comtox.2019.100096.
(90) Bank, R. P. D. *RCSB PDB: Homepage*. https://www.rcsb.org/ (accessed 2020-05-06).
(91) Ilari, A.; Savino, C. Protein Structure Determination by X-Ray Crystallography. In *Bioinformatics: Data, Sequence Analysis and Evolution*; Keith, J. M., Ed.; Methods in Molecular Biology™; Humana Press: Totowa, NJ, 2008; pp 63–87. https://doi.org/10.1007/978-1-60327-159-2_3.
(92) *Resonance Magnétique Nucléaire (RMN): Principe de Base*.
(93) Prospects for High-Throughput Structure Determination of Proteins by NMR Spectroscopy. In *Protein Structure*; Chasman, D., Ed.; CRC Press, 2003.
(94) *Dynamic folding modulation generates FGF21 variant against diabetes | EMBO reports*. https://www.embopress.org/doi/full/10.15252/embr.202051352 (accessed 2022-06-16).
(95) Callaway, E. The Revolution Will Not Be Crystallized: A New Method Sweeps through Structural Biology. *Nature* **2015**, *525* (7568), 172–174. https://doi.org/10.1038/525172a.
(96) *La cryo-microscopie électronique en biologie structurale*. CultureSciences-Chimie. https://culturesciences.chimie.ens.fr/thematiques/chimie-du-vivant/la-cryo-microscopie-electronique-en-biologie-structurale (accessed 2022-06-16).

(97) Yip, K. M.; Fischer, N.; Paknia, E.; Chari, A.; Stark, H. Atomic-Resolution Protein Structure Determination by Cryo-EM. *Nature* **2020**, *587* (7832), 157–161. https://doi.org/10.1038/s41586-020-2833-4.
(98) *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. - PMC*. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC146917/ (accessed 2022-06-16).
(99) Fidler, D. R.; Murphy, S. E.; Courtis, K.; Antonoudiou, P.; El-Tohamy, R.; Ient, J.; Levine, T. P. Using HHsearch to Tackle Proteins of Unknown Function: A Pilot Study with PH Domains. *Traffic Cph. Den.* **2016**, *17* (11), 1214–1226. https://doi.org/10.1111/tra.12432.
(100) Ghouzam, Y.; Postic, G.; Guerin, P.-E.; de Brevern, A. G.; Gelly, J.-C. ORION: A Web Server for Protein Fold Recognition and Structure Prediction Using Evolutionary Hybrid Profiles. *Sci. Rep.* **2016**, *6*, 28268. https://doi.org/10.1038/srep28268.
(101) *About MODELLER*. https://salilab.org/modeller/ (accessed 2022-06-16).
(102) *Robetta: full-chain protein structure prediction server*. http://robetta.bakerlab.org/ (accessed 2022-06-16).
(103) Zwanzig, R.; Szabo, A.; Bagchi, B. Levinthal's Paradox. *Proc. Natl. Acad. Sci. U. S. A.* **1992**, *89* (1), 20–22.
(104) Anfinsen, C. B. Principles That Govern the Folding of Protein Chains. *Science* **1973**, *181* (4096), 223–230. https://doi.org/10.1126/science.181.4096.223.
(105) Dill, K. A.; MacCallum, J. L. The Protein-Folding Problem, 50 Years On. *Science* **2012**, *338* (6110), 1042–1046. https://doi.org/10.1126/science.1219021.
(106) *Atomic-Level Characterization of the Structural Dynamics of Proteins*. https://www.science.org/doi/10.1126/science.1187409 (accessed 2022-06-16).
(107) *Constraint methods that accelerate free-energy simulations of biomolecules: The Journal of Chemical Physics: Vol 143, No 24*. https://aip.scitation.org/doi/10.1063/1.4936911 (accessed 2022-06-16).
(108) Senior, A. W.; Evans, R.; Jumper, J.; Kirkpatrick, J.; Sifre, L.; Green, T.; Qin, C.; Žídek, A.; Nelson, A. W. R.; Bridgland, A.; Penedones, H.; Petersen, S.; Simonyan, K.; Crossan, S.; Kohli, P.; Jones, D. T.; Silver, D.; Kavukcuoglu, K.; Hassabis, D. Improved Protein Structure Prediction Using Potentials from Deep Learning. *Nature* **2020**, *577* (7792), 706–710. https://doi.org/10.1038/s41586-019-1923-7.
(109) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2.
(110) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc., 2017; Vol. 30.
(111) *AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models | Nucleic Acids Research | Oxford Academic*. https://academic.oup.com/nar/article/50/D1/D439/6430488 (accessed 2022-06-16).
(112) Morgan, G. J. Emile Zuckerkandl, Linus Pauling, and the Molecular Evolutionary Clock, 1959-1965. *J. Hist. Biol.* **1998**, *31* (2), 155–178. https://doi.org/10.1023/a:1004394418084.

(113) Appel, R. D.; Bairoch, A.; Hochstrasser, D. F. A New Generation of Information Retrieval Tools for Biologists: The Example of the ExPASy WWW Server. *Trends Biochem. Sci.* **1994**, *19* (6), 258–260. https://doi.org/10.1016/0968-0004(94)90153-8.
(114) Celniker, G.; Nimrod, G.; Ashkenazy, H.; Glaser, F.; Martz, E.; Mayrose, I.; Pupko, T.; Ben-Tal, N. ConSurf: Using Evolutionary Data to Raise Testable Hypotheses about Protein Function. *Isr. J. Chem.* **2013**, *53* (3–4), 199–206. https://doi.org/10.1002/ijch.201200096.
(115) Gao, M.; Skolnick, J. A Comprehensive Survey of Small-Molecule Binding Pockets in Proteins. *PLOS Comput. Biol.* **2013**, *9* (10), e1003302. https://doi.org/10.1371/journal.pcbi.1003302.
(116) Laskowski, R. A.; Watson, J. D.; Thornton, J. M. ProFunc: A Server for Predicting Protein Function from 3D Structure. *Nucleic Acids Res.* **2005**, *33* (suppl_2), W89–W93. https://doi.org/10.1093/nar/gki414.
(117) Shulman-Peleg, A.; Nussinov, R.; Wolfson, H. J. Recognition of Functional Sites in Protein Structures. *J. Mol. Biol.* **2004**, *339* (3), 607–633. https://doi.org/10.1016/j.jmb.2004.04.012.
(118) Hendlich, M. Databases for Protein-Ligand Complexes. *Acta Crystallogr. D Biol. Crystallogr.* **1998**, *54* (Pt 6 Pt 1), 1178–1182. https://doi.org/10.1107/s0907444998007124.
(119) Stark, A.; Sunyaev, S.; Russell, R. B. A Model for Statistical Significance of Local Similarities in Structure. *J. Mol. Biol.* **2003**, *326* (5), 1307–1316. https://doi.org/10.1016/s0022-2836(03)00045-7.
(120) Kinoshita, K.; Furui, J.; Nakamura, H. Identification of Protein Functions from a Molecular Surface Database, EF-Site. *J. Struct. Funct. Genomics* **2002**, *2* (1), 9–22. https://doi.org/10.1023/a:1011318527094.
(121) Zhang, Y.; Skolnick, J. TM-Align: A Protein Structure Alignment Algorithm Based on the TM-Score. *Nucleic Acids Res.* **2005**, *33* (7), 2302–2309. https://doi.org/10.1093/nar/gki524.
(122) *COFACTOR: improved protein function prediction by combining structure, sequence and protein–protein interaction information | Nucleic Acids Research | Oxford Academic*. https://academic.oup.com/nar/article/45/W1/W291/3787871 (accessed 2022-06-16).
(123) Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. A Geometric Approach to Macromolecule-Ligand Interactions. *J. Mol. Biol.* **1982**, *161* (2), 269–288. https://doi.org/10.1016/0022-2836(82)90153-x.
(124) DesJarlais, R. L.; Sheridan, R. P.; Seibel, G. L.; Dixon, J. S.; Kuntz, I. D.; Venkataraghavan, R. Using Shape Complementarity as an Initial Screen in Designing Ligands for a Receptor Binding Site of Known Three-Dimensional Structure. *J. Med. Chem.* **1988**, *31* (4), 722–729. https://doi.org/10.1021/jm00399a006.
(125) Levitt, D. G.; Banaszak, L. J. POCKET: A Computer Graphics Method for Identifying and Displaying Protein Cavities and Their Surrounding Amino Acids. *J. Mol. Graph.* **1992**, *10* (4), 229–234. https://doi.org/10.1016/0263-7855(92)80074-n.
(126) Laskowski, R. A. SURFNET: A Program for Visualizing Molecular Surfaces, Cavities, and Intermolecular Interactions. *J. Mol. Graph.* **1995**, *13* (5), 323–330, 307–308. https://doi.org/10.1016/0263-7855(95)00073-9.
(127) Laurie, A. T. R.; Jackson, R. M. Methods for the Prediction of Protein-Ligand Binding Sites for Structure-Based Drug Design and Virtual Ligand Screening. *Curr. Protein Pept. Sci.* **2006**, *7* (5), 395–406. https://doi.org/10.2174/138920306778559386.

(128) Goodford, P. J. A Computational Procedure for Determining Energetically Favorable Binding Sites on Biologically Important Macromolecules. *J. Med. Chem.* **1985**, *28* (7), 849–857. https://doi.org/10.1021/jm00145a002.
(129) Ruppert, J.; Welch, W.; Jain, A. N. Automatic Identification and Representation of Protein Binding Sites for Molecular Docking. *Protein Sci. Publ. Protein Soc.* **1997**, *6* (3), 524–533. https://doi.org/10.1002/pro.5560060302.
(130) Sanders, M. P. A.; McGuire, R.; Roumen, L.; Esch, I. J. P. de; Vlieg, J. de; Klomp, J. P. G.; Graaf, C. de. From the Protein's Perspective: The Benefits and Challenges of Protein Structure-Based Pharmacophore Modeling. *MedChemComm* **2012**, *3* (1), 28–38. https://doi.org/10.1039/C1MD00210D.
(131) Hu, B.; Lill, M. A. PharmDock: A Pharmacophore-Based Docking Program. *J. Cheminformatics* **2014**, *6*, 14. https://doi.org/10.1186/1758-2946-6-14.
(132) Mortier, J.; Dhakal, P.; Volkamer, A. Truly Target-Focused Pharmacophore Modeling: A Novel Tool for Mapping Intermolecular Surfaces. *Molecules* **2018**, *23* (8), 1959. https://doi.org/10.3390/molecules23081959.
(133) Kratochwil, N. A.; Malherbe, P.; Lindemann, L.; Ebeling, M.; Hoener, M. C.; Mühlemann, A.; Porter, R. H. P.; Stahl, M.; Gerber, P. R. An Automated System for the Analysis of G Protein-Coupled Receptor Transmembrane Binding Pockets: Alignment, Receptor-Based Pharmacophores, and Their Application. *J. Chem. Inf. Model.* **2005**, *45* (5), 1324–1336. https://doi.org/10.1021/ci050221u.
(134) Miranker, A.; Karplus, M. Functionality Maps of Binding Sites: A Multiple Copy Simultaneous Search Method. *Proteins* **1991**, *11* (1), 29–34. https://doi.org/10.1002/prot.340110104.
(135) Alexov, E. G.; Gunner, M. R. Incorporating Protein Conformational Flexibility into the Calculation of PH-Dependent Protein Properties. *Biophys. J.* **1997**, *72* (5), 2075–2093. https://doi.org/10.1016/S0006-3495(97)78851-9.
(136) Chen, J.; Lai, L. Pocket v.2: Further Developments on Receptor-Based Pharmacophore Modeling. *J. Chem. Inf. Model.* **2006**, *46* (6), 2684–2691. https://doi.org/10.1021/ci600246s.
(137) Goto, J.; Kataoka, R.; Hirayama, N. Ph4Dock: Pharmacophore-Based Protein−Ligand Docking. *J. Med. Chem.* **2004**, *47* (27), 6804–6811. https://doi.org/10.1021/jm0493818.
(138) Sydow, D. Dynophores: Novel Dynamic Pharmacophores. **2015**. https://doi.org/10.18452/14267.
(139) Wieder, M.; Garon, A.; Perricone, U.; Boresch, S.; Seidel, T.; Almerico, A. M.; Langer, T. Common Hits Approach: Combining Pharmacophore Modeling and Molecular Dynamics Simulations. *J. Chem. Inf. Model.* **2017**, *57* (2), 365–385. https://doi.org/10.1021/acs.jcim.6b00674.
(140) Perricone, U.; Wieder, M.; Seidel, T.; Langer, T.; Padova, A.; Almerico, A. M.; Tutone, M. A Molecular Dynamics–Shared Pharmacophore Approach to Boost Early-Enrichment Virtual Screening: A Case Study on Peroxisome Proliferator-Activated Receptor α. *ChemMedChem* **2017**, *12* (16), 1399–1407. https://doi.org/10.1002/cmdc.201600526.
(141) Hu, B.; Lill, M. A. Protein Pharmacophore Selection Using Hydration-Site Analysis. *J. Chem. Inf. Model.* **2012**, *52* (4), 1046–1060. https://doi.org/10.1021/ci200620h.
(142) Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and Scoring in Virtual Screening for Drug Discovery: Methods and Applications. *Nat. Rev. Drug Discov.* **2004**, *3* (11), 935–949. https://doi.org/10.1038/nrd1549.
(143) Fischer, E. Einfluss Der Configuration Auf Die Wirkung Der Enzyme. *Berichte Dtsch. Chem. Ges.* **1894**, *27* (3), 2985–2993. https://doi.org/10.1002/cber.18940270364.

(144) Koshland Jr., D. E. The Key–Lock Theory and the Induced Fit Theory. *Angew. Chem. Int. Ed. Engl.* **1995**, *33* (23–24), 2375–2378. https://doi.org/10.1002/anie.199423751.
(145) Antunes, D. A.; Devaurs, D.; Kavraki, L. E. Understanding the Challenges of Protein Flexibility in Drug Design. *Expert Opin. Drug Discov.* **2015**, *10* (12), 1301–1313. https://doi.org/10.1517/17460441.2015.1094458.
(146) Morris, G. M.; Lim-Wilby, M. Molecular Docking. In *Molecular Modeling of Proteins*; Kukol, A., Ed.; Methods Molecular Biology™; Humana Press: Totowa, NJ, 2008; pp 365–382. https://doi.org/10.1007/978-1-59745-177-2_19.
(147) *Thermodynamique. Plan du cours 3. Théorie cinétique des gaz : Calcul de la pression. Température et Energie. Dégrés de liberté d'une molécule - PDF Téléchargement Gratuit*. https://docplayer.fr/18599771-Thermodynamique-plan-du-cours-3-theorie-cinetique-des-gaz-calcul-de-la-pression-temperature-et-energie-degres-de-liberte-d-une-molecule.html (accessed 2022-06-17).
(148) Leach, A. R.; Kuntz, I. D. Conformational analysis of flexible ligands in macromolecular receptor sites. *J. Comput. Chem.* **1992**, *13* (6), 730–748. https://doi.org/10.1002/jcc.540130608.
(149) Miller, M. D.; Kearsley, S. K.; Underwood, D. J.; Sheridan, R. P. FLOG: A System to Select 'Quasi-Flexible' Ligands Complementary to a Receptor of Known Three-Dimensional Structure. *J. Comput. Aided Mol. Des.* **1994**, *8* (2), 153–174. https://doi.org/10.1007/BF00119865.
(150) McGann, M. FRED Pose Prediction and Virtual Screening Accuracy. *J. Chem. Inf. Model.* **2011**, *51* (3), 578–596. https://doi.org/10.1021/ci100436p.
(151) *Computational Approaches to Nuclear Receptors*; 2012. https://doi.org/10.1039/9781849735353.
(152) Mangoni, M.; Roccatano, D.; Di Nola, A. Docking of flexible ligands to flexible receptors in solution by molecular dynamics simulation. *Proteins Struct. Funct. Bioinforma.* **1999**, *35* (2), 153–162. https://doi.org/10.1002/(SICI)1097-0134(19990501)35:2<153::AID-PROT2>3.0.CO;2-E.
(153) Ferreira, L. G.; Dos Santos, R. N.; Oliva, G.; Andricopulo, A. D. Molecular Docking and Structure-Based Drug Design Strategies. *Molecules* **2015**, *20* (7), 13384–13421. https://doi.org/10.3390/molecules200713384.
(154) Zsoldos, Z.; Reid, D.; Simon, A.; Sadjad, S. B.; Johnson, A. P. EHiTS: A New Fast, Exhaustive Flexible Ligand Docking System. *J. Mol. Graph. Model.* **2007**, *26* (1), 198–212. https://doi.org/10.1016/j.jmgm.2006.06.002.
(155) Jain, A. N. Surflex: Fully Automatic Flexible Molecular Docking Using a Molecular Similarity-Based Search Engine. *J. Med. Chem.* **2003**, *46* (4), 499–511. https://doi.org/10.1021/jm020406h.
(156) McGann, M. FRED and HYBRID Docking Performance on Standardized Datasets. *J. Comput. Aided Mol. Des.* **2012**, *26* (8), 897–906. https://doi.org/10.1007/s10822-012-9584-8.
(157) Ewing, T. J. A.; Makino, S.; Skillman, A. G.; Kuntz, I. D. DOCK 4.0: Search Strategies for Automated Molecular Docking of Flexible Molecule Databases. *J. Comput. Aided Mol. Des.* **2001**, *15* (5), 411–428. https://doi.org/10.1023/A:1011115820450.
(158) Dias, R.; de Azevedo, W. F. Molecular Docking Algorithms. *Curr. Drug Targets* **2008**, *9* (12), 1040–1047. https://doi.org/10.2174/138945008786949432.
(159) Gorelik, B.; Goldblum, A. High Quality Binding Modes in Docking Ligands to Proteins. *Proteins* **2008**, *71* (3), 1373–1386. https://doi.org/10.1002/prot.21847.
(160) Nicholas Metropolis, S. U. The Monte Carlo Method. *J. Am. Stat. Assoc.* **1949**.

(161) Morris, G. M.; Goodsell, D. S.; Huey, R.; Olson, A. J. Distributed Automated Docking of Flexible Ligands to Proteins: Parallel Applications of AutoDock 2.4. *J. Comput. Aided Mol. Des.* **1996**, *10* (4), 293–304. https://doi.org/10.1007/BF00124499.
(162) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and Validation of a Genetic Algorithm for Flexible Docking. *J. Mol. Biol.* **1997**, *267* (3), 727–748. https://doi.org/10.1006/jmbi.1996.0897.
(163) Jain, A. N. Scoring Functions for Protein-Ligand Docking. *Curr. Protein Pept. Sci.* **2006**, *7* (5), 407–420. https://doi.org/10.2174/138920306778559395.
(164) Charifson, P. S.; Corkery, J. J.; Murcko, M. A.; Walters, W. P. Consensus Scoring: A Method for Obtaining Improved Hit Rates from Docking Databases of Three-Dimensional Structures into Proteins. *J. Med. Chem.* **1999**, *42* (25), 5100–5109. https://doi.org/10.1021/jm990352k.
(165) Feher, M. Consensus Scoring for Protein-Ligand Interactions. *Drug Discov. Today* **2006**, *11* (9–10), 421–428. https://doi.org/10.1016/j.drudis.2006.03.009.
(166) Okamoto, M.; Masuda, Y.; Muroya, A.; Yasuno, K.; Takahashi, O.; Furuya, T. Evaluation of Docking Calculations on X-Ray Structures Using CONSENSUS-DOCK. *Chem. Pharm. Bull. (Tokyo)* **2010**, *58* (12), 1655–1657. https://doi.org/10.1248/cpb.58.1655.
(167) Bar-Haim, S.; Aharon, A.; Ben-Moshe, T.; Marantz, Y.; Senderowitz, H. SeleX-CS: A New Consensus Scoring Algorithm for Hit Discovery and Lead Optimization. *J. Chem. Inf. Model.* **2009**, *49* (3), 623–633. https://doi.org/10.1021/ci800335j.
(168) Teramoto, R.; Fukunishi, H. Supervised Consensus Scoring for Docking and Virtual Screening. *J. Chem. Inf. Model.* **2007**, *47* (2), 526–534. https://doi.org/10.1021/ci6004993.
(169) Betzi, S.; Suhre, K.; Chétrit, B.; Guerlesquin, F.; Morelli, X. GFscore: A General Nonlinear Consensus Scoring Function for High-Throughput Docking. *J. Chem. Inf. Model.* **2006**, *46* (4), 1704–1712. https://doi.org/10.1021/ci0600758.
(170) Wang, R.; Lai, L.; Wang, S. Further Development and Validation of Empirical Scoring Functions for Structure-Based Binding Affinity Prediction. *J. Comput. Aided Mol. Des.* **2002**, *16* (1), 11–26. https://doi.org/10.1023/a:1016357811882.
(171) Terp, G. E.; Johansen, B. N.; Christensen, I. T.; Jørgensen, F. S. A New Concept for Multidimensional Selection of Ligand Conformations (MultiSelect) and Multidimensional Scoring (MultiScore) of Protein-Ligand Binding Affinities. *J. Med. Chem.* **2001**, *44* (14), 2333–2343. https://doi.org/10.1021/jm001090l.
(172) Englebienne, P.; Moitessier, N. Docking Ligands into Flexible and Solvated Macromolecules. 5. Force-Field-Based Prediction of Binding Affinities of Ligands to Proteins. *J. Chem. Inf. Model.* **2009**, *49* (11), 2564–2571. https://doi.org/10.1021/ci900251k.
(173) Kramer, B.; Rarey, M.; Lengauer, T. Evaluation of the FLEXX Incremental Construction Algorithm for Protein-Ligand Docking. *Proteins* **1999**, *37* (2), 228–241. https://doi.org/10.1002/(sici)1097-0134(19991101)37:2<228::aid-prot8>3.0.co;2-8.
(174) Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. Automated Docking Using a Lamarckian Genetic Algorithm and an Empirical Binding Free Energy Function. *J. Comput. Chem.* **1998**, *19* (14), 1639–1662. https://doi.org/10.1002/(SICI)1096-987X(19981115)19:14<1639::AID-JCC10>3.0.CO;2-B.
(175) *Pairwise solute descreening of solute charges from a dielectric medium - ScienceDirect*. https://www.sciencedirect.com/science/article/pii/000926149501082K (accessed 2022-06-17).

(176) Murray, C. W.; Auton, T. R.; Eldridge, M. D. Empirical Scoring Functions. II. The Testing of an Empirical Scoring Function for the Prediction of Ligand-Receptor Binding Affinities and the Use of Bayesian Regression to Improve the Quality of the Model. *J. Comput. Aided Mol. Des.* **1998**, *12* (5), 503–519. https://doi.org/10.1023/a:1008040323669.
(177) Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D. Improved Protein–Ligand Docking Using GOLD. *Proteins Struct. Funct. Bioinforma.* **2003**, *52* (4), 609–623. https://doi.org/10.1002/prot.10465.
(178) Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. Empirical Scoring Functions: I. The Development of a Fast Empirical Scoring Function to Estimate the Binding Affinity of Ligands in Receptor Complexes. *J. Comput. Aided Mol. Des.* **1997**, *11* (5), 425–445. https://doi.org/10.1023/a:1007996124545.
(179) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A Fast Flexible Docking Method Using an Incremental Construction Algorithm. *J. Mol. Biol.* **1996**, *261* (3), 470–489. https://doi.org/10.1006/jmbi.1996.0477.
(180) Huang, S.-Y.; Zou, X. An Iterative Knowledge-Based Scoring Function to Predict Protein-Ligand Interactions: I. Derivation of Interaction Potentials. *J. Comput. Chem.* **2006**, *27* (15), 1866–1875. https://doi.org/10.1002/jcc.20504.
(181) Gohlke, H.; Hendlich, M.; Klebe, G. Knowledge-Based Scoring Function to Predict Protein-Ligand Interactions. *J. Mol. Biol.* **2000**, *295* (2), 337–356. https://doi.org/10.1006/jmbi.1999.3371.
(182) Lill, M. A. Efficient Incorporation of Protein Flexibility and Dynamics into Molecular Docking Simulations. *Biochemistry* **2011**, *50* (28), 6157–6169. https://doi.org/10.1021/bi2004558.
(183) Najmanovich, R.; Kuttner, J.; Sobolev, V.; Edelman, M. Side-Chain Flexibility in Proteins upon Ligand Binding. *Proteins* **2000**, *39* (3), 261–268. https://doi.org/10.1002/(sici)1097-0134(20000515)39:3<261::aid-prot90>3.0.co;2-4.
(184) Wei, B. Q.; Weaver, L. H.; Ferrari, A. M.; Matthews, B. W.; Shoichet, B. K. Testing a Flexible-Receptor Docking Algorithm in a Model Binding Site. *J. Mol. Biol.* **2004**, *337* (5), 1161–1182. https://doi.org/10.1016/j.jmb.2004.02.015.
(185) Kokh, D. B.; Wade, R. C.; Wenzel, W. Receptor Flexibility in Small-Molecule Docking Calculations. *WIREs Comput. Mol. Sci.* **2011**, *1* (2), 298–314. https://doi.org/10.1002/wcms.29.
(186) Lexa, K. W.; Carlson, H. A. Protein Flexibility in Docking and Surface Mapping. *Q. Rev. Biophys.* **2012**, *45* (3), 301–343. https://doi.org/10.1017/S0033583512000066.
(187) Jiang, F.; Kim, S. H. "Soft Docking": Matching of Molecular Surface Cubes. *J. Mol. Biol.* **1991**, *219* (1), 79–102. https://doi.org/10.1016/0022-2836(91)90859-5.
(188) Ferrari, A. M.; Wei, B. Q.; Costantino, L.; Shoichet, B. K. Soft Docking and Multiple Receptor Conformations in Virtual Screening. *J. Med. Chem.* **2004**, *47* (21), 5076–5084. https://doi.org/10.1021/jm049756p.
(189) Mizutani, M. Y.; Takamatsu, Y.; Ichinose, T.; Nakamura, K.; Itai, A. Effective Handling of Induced-Fit Motion in Flexible Docking. *Proteins* **2006**, *63* (4), 878–891. https://doi.org/10.1002/prot.20931.
(190) Desmet, J.; Wilson, I. A.; Joniau, M.; De Maeyer, M.; Lasters, I. Computation of the Binding of Fully Flexible Peptides to Proteins with Flexible Side Chains. *FASEB J. Off. Publ. Fed. Am. Soc. Exp. Biol.* **1997**, *11* (2), 164–172. https://doi.org/10.1096/fasebj.11.2.9039959.
(191) Schaffer, L.; Verkhivker, G. M. Predicting Structural Effects in HIV-1 Protease Mutant Complexes with Flexible Ligand Docking and Protein Side-Chain Optimization.

Proteins **1998**, *33* (2), 295–310. https://doi.org/10.1002/(sici)1097-0134(19981101)33:2<295::aid-prot12>3.0.co;2-f.
(192) Leach, A. R. Ligand Docking to Proteins with Discrete Side-Chain Flexibility. *J. Mol. Biol.* **1994**, *235* (1), 345–356. https://doi.org/10.1016/s0022-2836(05)80038-5.
(193) Abagyan, R.; Totrov, M.; Kuznetsov, D. ICM—A New Method for Protein Modeling and Design: Applications to Docking and Structure Prediction from the Distorted Native Conformation. *J. Comput. Chem.* **1994**, *15* (5), 488–506. https://doi.org/10.1002/jcc.540150503.
(194) Schnecke, V.; Swanson, C. A.; Getzoff, E. D.; Tainer, J. A.; Kuhn, L. A. Screening a Peptidyl Database for Potential Ligands to Proteins with Side-Chain Flexibility. *Proteins* **1998**, *33* (1), 74–87.
(195) Ravindranath, P. A.; Forli, S.; Goodsell, D. S.; Olson, A. J.; Sanner, M. F. AutoDockFR: Advances in Protein-Ligand Docking with Explicitly Specified Binding Site Flexibility. *PLoS Comput. Biol.* **2015**, *11* (12), e1004586. https://doi.org/10.1371/journal.pcbi.1004586.
(196) Koes, D. R.; Baumgartner, M. P.; Camacho, C. J. Lessons Learned in Empirical Scoring with Smina from the CSAR 2011 Benchmarking Exercise. *J. Chem. Inf. Model.* **2013**, *53* (8), 1893–1904. https://doi.org/10.1021/ci300604z.
(197) Claussen, H.; Buning, C.; Rarey, M.; Lengauer, T. FlexE: Efficient Molecular Docking Considering Protein Structure Variations. *J. Mol. Biol.* **2001**, *308* (2), 377–395. https://doi.org/10.1006/jmbi.2001.4551.
(198) Ben Nasr, N.; Guillemain, H.; Lagarde, N.; Zagury, J.-F.; Montes, M. Multiple Structures for Virtual Ligand Screening: Defining Binding Site Properties-Based Criteria to Optimize the Selection of the Query. *J. Chem. Inf. Model.* **2013**, *53* (2), 293–311. https://doi.org/10.1021/ci3004557.
(199) Réau, M.; Lagarde, N.; Zagury, J.-F.; Montes, M. Hits Discovery on the Androgen Receptor: In Silico Approaches to Identify Agonist Compounds. *Cells* **2019**, *8* (11), 1431. https://doi.org/10.3390/cells8111431.
(200) Bottegoni, G.; Kufareva, I.; Totrov, M.; Abagyan, R. Four-Dimensional Docking: A Fast and Accurate Account of Discrete Receptor Flexibility in Ligand Docking. *J. Med. Chem.* **2009**, *52* (2), 397–406. https://doi.org/10.1021/jm8009958.
(201) Berry, M.; Fielding, B.; Gamieldien, J. Chapter 27 - Practical Considerations in Virtual Screening and Molecular Docking. In *Emerging Trends in Computational Biology, Bioinformatics, and Systems Biology*; Tran, Q. N., Arabnia, H., Eds.; Emerging Trends in Computer Science and Applied Computing; Morgan Kaufmann: Boston, 2015; pp 487–502. https://doi.org/10.1016/B978-0-12-802508-6.00027-2.
(202) *Diverse, High-Quality Test Set for the Validation of Protein−Ligand Docking Performance | Journal of Medicinal Chemistry*. https://pubs.acs.org/doi/10.1021/jm061277y (accessed 2022-06-19).
(203) Lie, M. A.; Thomsen, R.; Pedersen, C. N. S.; Schiøtt, B.; Christensen, M. H. Molecular Docking with Ligand Attached Water Molecules. *J. Chem. Inf. Model.* **2011**, *51* (4), 909–917. https://doi.org/10.1021/ci100510m.
(204) Schneider, N.; Lange, G.; Hindle, S.; Klein, R.; Rarey, M. A Consistent Description of HYdrogen Bond and DEhydration Energies in Protein–Ligand Complexes: Methods behind the HYDE Scoring Function. *J. Comput. Aided Mol. Des.* **2013**, *27* (1), 15–29. https://doi.org/10.1007/s10822-012-9626-2.
(205) *Nuclear Hormone Receptor Targeted Virtual Screening | Journal of Medicinal Chemistry*. https://pubs-acs-org.proxybib-pp.cnam.fr/doi/10.1021/jm0300173 (accessed 2022-06-18).

(206) Genheden, S.; Ryde, U. The MM/PBSA and MM/GBSA Methods to Estimate Ligand-Binding Affinities. *Expert Opin. Drug Discov.* **2015**, *10* (5), 449–461. https://doi.org/10.1517/17460441.2015.1032936.
(207) Ashtawy, H. M.; Mahapatra, N. R. Machine-Learning Scoring Functions for Identifying Native Poses of Ligands Docked to Known and Novel Proteins. *BMC Bioinformatics* **2015**, *16* (6), S3. https://doi.org/10.1186/1471-2105-16-S6-S3.
(208) *OnionNet: a Multiple-Layer Intermolecular-Contact-Based Convolutional Neural Network for Protein–Ligand Binding Affinity Prediction | ACS Omega*. https://pubs.acs.org/doi/10.1021/acsomega.9b01997 (accessed 2022-07-04).
(209) Singh, S.; Bani Baker, Q.; Singh, D. B. Chapter 18 - Molecular Docking and Molecular Dynamics Simulation. In *Bioinformatics*; Singh, D. B., Pathak, R. K., Eds.; Academic Press, 2022; pp 291–304. https://doi.org/10.1016/B978-0-323-89775-4.00014-6.
(210) Hawkins, P. C. D.; Warren, G. L.; Skillman, A. G.; Nicholls, A. How to Do an Evaluation: Pitfalls and Traps. *J. Comput. Aided Mol. Des.* **2008**, *22* (3–4), 179–190. https://doi.org/10.1007/s10822-007-9166-3.
(211) Kirchmair, J.; Markt, P.; Distinto, S.; Wolber, G.; Langer, T. Evaluation of the Performance of 3D Virtual Screening Protocols: RMSD Comparisons, Enrichment Assessments, and Decoy Selection--What Can We Learn from Earlier Mistakes? *J. Comput. Aided Mol. Des.* **2008**, *22* (3–4), 213–228. https://doi.org/10.1007/s10822-007-9163-6.
(212) *Contact area difference (CAD): a robust measure to evaluate accuracy of protein models - ScienceDirect*. https://www-sciencedirect-com.proxybib-pp.cnam.fr/science/article/pii/S0022283697909943 (accessed 2022-07-18).
(213) *Maximum Unbiased Validation (MUV) Data Sets for Virtual Screening Based on PubChem Bioactivity Data | Journal of Chemical Information and Modeling*. https://pubs-acs-org.proxybib-pp.cnam.fr/doi/10.1021/ci8002649 (accessed 2022-06-19).
(214) Mysinger, M. M.; Carchia, M.; Irwin, John. J.; Shoichet, B. K. Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *J. Med. Chem.* **2012**, *55* (14), 6582–6594. https://doi.org/10.1021/jm300687e.
(215) Réau, M.; Langenfeld, F.; Zagury, J.-F.; Lagarde, N.; Montes, M. Decoys Selection in Benchmarking Datasets: Overview and Perspectives. *Front. Pharmacol.* **2018**, *9*, 11. https://doi.org/10.3389/fphar.2018.00011.
(216) Réau, M.; Lagarde, N.; Zagury, J.-F.; Montes, M. Nuclear Receptors Database Including Negative Data (NR-DBIND): A Database Dedicated to Nuclear Receptors Binding Data Including Negative Data and Pharmacological Profile. *J. Med. Chem.* **2019**, *62* (6), 2894–2904. https://doi.org/10.1021/acs.jmedchem.8b01105.
(217) Vihinen, M. How to Evaluate Performance of Prediction Methods? Measures and Their Interpretation in Variation Effect Analysis. *BMC Genomics* **2012**, *13* (4), S2. https://doi.org/10.1186/1471-2164-13-S4-S2.
(218) Fisher, J. S. Are All EDC Effects Mediated via Steroid Hormone Receptors? *Toxicology* **2004**, *205* (1), 33–41. https://doi.org/10.1016/j.tox.2004.06.035.
(219) Zoeller, R. T.; Brown, T. R.; Doan, L. L.; Gore, A. C.; Skakkebaek, N. E.; Soto, A. M.; Woodruff, T. J.; Vom Saal, F. S. Endocrine-Disrupting Chemicals and Public Health Protection: A Statement of Principles from The Endocrine Society. *Endocrinology* **2012**, *153* (9), 4097–4110. https://doi.org/10.1210/en.2012-1422.
(220) *Review on crosstalk and common mechanisms of endocrine disruptors: Scaffolding to improve PBPK/PD model of EDC mixture - ScienceDirect*. https://www-sciencedirect-com.proxybib-pp.cnam.fr/science/article/pii/S0160412016304433?casa_token=5t9gbZJ0u1cAAAAA:

8XTAvJrT2Ih_3vWsY-aN_JEhGgKrWKciDN9gH5c71U2-EerFm3UQe1JVekqVg7fU_4EhN0xl (accessed 2022-06-20).
(221) Montes-Grajales, D.; Olivero-Verbel, J. EDCs DataBank: 3D-Structure Database of Endocrine Disrupting Chemicals. *Toxicology* **2015**, *327*, 87–94. https://doi.org/10.1016/j.tox.2014.11.006.
(222) *Endocrine Disruptors*. National Institute of Environmental Health Sciences. https://www.niehs.nih.gov/health/topics/agents/endocrine/index.cfm (accessed 2022-06-20).
(223) *Sex ratio skew and breeding patterns of gulls: demographic and toxicological considerations – https://cascadiaresearch.org*. https://cascadiaresearch.org/publications/sex-ratio-skew-and-breeding-patterns-gulls-demographic-and-toxicological-considerations/ (accessed 2022-06-20).
(224) Combarnous, Y.; Nguyen, T. M. D. Comparative Overview of the Mechanisms of Action of Hormones and Endocrine Disruptor Compounds. *Toxics* **2019**, *7* (1). https://doi.org/10.3390/toxics7010005.
(225) Vandenberg, L. N.; Colborn, T.; Hayes, T. B.; Heindel, J. J.; Jacobs, D. R.; Lee, D.-H.; Myers, J. P.; Shioda, T.; Soto, A. M.; vom Saal, F. S.; Welshons, W. V.; Zoeller, R. T. Regulatory Decisions on Endocrine Disrupting Chemicals Should Be Based on the Principles of Endocrinology. *Reprod. Toxicol.* **2013**, *38*, 1–15. https://doi.org/10.1016/j.reprotox.2013.02.002.
(226) Yang, O.; Kim, H. L.; Weon, J.-I.; Seo, Y. R. Endocrine-Disrupting Chemicals: Review of Toxicological Mechanisms Using Molecular Pathway Analysis. *J. Cancer Prev.* **2015**, *20* (1), 12–24. https://doi.org/10.15430/JCP.2015.20.1.12.
(227) Hiller-Sturmhöfel, S.; Bartke, A. The Endocrine System. *Alcohol Health Res. World* **1998**, *22* (3), 153–164.
(228) Kleine, B.; Rossmanith, W. G. Hormones: Some Definitions. In *Hormones and the Endocrine System: Textbook of Endocrinology*; Kleine, B., Rossmanith, W. G., Eds.; Springer International Publishing: Cham, 2016; pp 11–17. https://doi.org/10.1007/978-3-319-15060-4_3.
(229) Novac, N.; Heinzel, T. Nuclear Receptors: Overview and Classification. *Curr. Drug Targets Inflamm. Allergy* **2004**, *3* (4), 335–346. https://doi.org/10.2174/1568010042634541.
(230) Mohan, R.; Heyman, R. A. Orphan Nuclear Receptor Modulators. *Curr. Top. Med. Chem.* **2003**, *3* (14), 1637–1647. https://doi.org/10.2174/1568026033451709.
(231) *Effects of three 19-nor steroids on human ovulation and menstruation - ScienceDirect*. https://www.sciencedirect.com/science/article/abs/pii/0002937858905556 (accessed 2022-06-23).
(232) Shao, W.; Brown, M. Advances in Estrogen Receptor Biology: Prospects for Improvements in Targeted Breast Cancer Therapy. *Breast Cancer Res.* **2003**, *6* (1), 39. https://doi.org/10.1186/bcr742.
(233) Di Croce, L.; Okret, S.; Kersten, S.; Gustafsson, J. A.; Parker, M.; Wahli, W.; Beato, M. Steroid and Nuclear Receptors. Villefranche-Sur-Mer, France, May 25-27, 1999. *EMBO J.* **1999**, *18* (22), 6201–6210. https://doi.org/10.1093/emboj/18.22.6201.
(234) Sever, R.; Glass, C. K. Signaling by Nuclear Receptors. *Cold Spring Harb. Perspect. Biol.* **2013**, *5* (3), a016709–a016709. https://doi.org/10.1101/cshperspect.a016709.
(235) Patel, S. R.; Skafar, D. F. Modulation of Nuclear Receptor Activity by the F Domain. *Mol. Cell. Endocrinol.* **2015**, *418*, 298–305. https://doi.org/10.1016/j.mce.2015.07.009.
(236) Weikum, E. R.; Liu, X.; Ortlund, E. A. The Nuclear Receptor Superfamily: A Structural Perspective. *Protein Sci. Publ. Protein Soc.* **2018**, *27* (11), 1876–1892. https://doi.org/10.1002/pro.3496.

(237) Moras, D.; Gronemeyer, H. The Nuclear Receptor Ligand-Binding Domain: Structure and Function. *Curr. Opin. Cell Biol.* **1998**, *10* (3), 384–391. https://doi.org/10.1016/s0955-0674(98)80015-x.
(238) Wurtz, J. M.; Bourguet, W.; Renaud, J. P.; Vivat, V.; Chambon, P.; Moras, D.; Gronemeyer, H. A Canonical Structure for the Ligand-Binding Domain of Nuclear Receptors. *Nat. Struct. Biol.* **1996**, *3* (1), 87–94. https://doi.org/10.1038/nsb0196-87.
(239) Nagy, L.; Schwabe, J. W. R. Mechanism of the Nuclear Receptor Molecular Switch. *Trends Biochem. Sci.* **2004**, *29* (6), 317–324. https://doi.org/10.1016/j.tibs.2004.04.006.
(240) Forman, B. M.; Tzameli, I.; Choi, H. S.; Chen, J.; Simha, D.; Seol, W.; Evans, R. M.; Moore, D. D. Androstane Metabolites Bind to and Deactivate the Nuclear Receptor CAR-Beta. *Nature* **1998**, *395* (6702), 612–615. https://doi.org/10.1038/26996.
(241) Gronemeyer, H.; Gustafsson, J.-A.; Laudet, V. Principles for Modulation of the Nuclear Receptor Superfamily. *Nat. Rev. Drug Discov.* **2004**, *3* (11), 950–964. https://doi.org/10.1038/nrd1551.
(242) Shi, Y. Orphan Nuclear Receptors in Drug Discovery. *Drug Discov. Today* **2007**, *12* (11–12), 440–445. https://doi.org/10.1016/j.drudis.2007.04.006.
(243) Togashi, M.; Borngraeber, S.; Sandler, B.; Fletterick, R. J.; Webb, P.; Baxter, J. D. Conformational Adaptation of Nuclear Receptor Ligand Binding Domains to Agonists: Potential for Novel Approaches to Ligand Design. *J. Steroid Biochem. Mol. Biol.* **2005**, *93* (2), 127–137. https://doi.org/10.1016/j.jsbmb.2005.01.004.
(244) Rastinejad, F.; Huang, P.; Chandra, V.; Khorasanizadeh, S. Understanding Nuclear Receptor Form and Function Using Structural Biology. *J. Mol. Endocrinol.* **2013**, *51* (3), T1–T21. https://doi.org/10.1530/JME-13-0173.
(245) Lee, H.-R.; Jeung, E.-B.; Cho, M.-H.; Kim, T.-H.; Leung, P. C. K.; Choi, K.-C. Molecular Mechanism(s) of Endocrine-Disrupting Chemicals and Their Potent Oestrogenicity in Diverse Cells and Tissues That Express Oestrogen Receptors. *J. Cell. Mol. Med.* **2013**, *17* (1), 1–11. https://doi.org/10.1111/j.1582-4934.2012.01649.x.
(246) Swedenborg, E.; Rüegg, J.; Mäkelä, S.; Pongratz, I. Endocrine Disruptive Chemicals: Mechanisms of Action and Involvement in Metabolic Disorders. *J. Mol. Endocrinol.* **2009**, *43* (1), 1–10. https://doi.org/10.1677/JME-08-0132.
(247) Schug, T. T.; Janesick, A.; Blumberg, B.; Heindel, J. J. Endocrine Disrupting Chemicals and Disease Susceptibility. *J. Steroid Biochem. Mol. Biol.* **2011**, *127* (3), 204–215. https://doi.org/10.1016/j.jsbmb.2011.08.007.
(248) Caron, -Beaudoin Élyse; Viau, R.; Sanderson, J. T. Effects of Neonicotinoid Pesticides on Promoter-Specific Aromatase (CYP19) Expression in Hs578t Breast Cancer Cells and the Role of the VEGF Pathway. *Environ. Health Perspect. 126* (4), 047014. https://doi.org/10.1289/EHP2698.
(249) Derouiche, S.; Mariot, P.; Warnier, M.; Vancauwenberghe, E.; Bidaux, G.; Gosset, P.; Mauroy, B.; Bonnal, J.-L.; Slomianny, C.; Delcourt, P.; Dewailly, E.; Prevarskaya, N.; Roudbaraki, M. Activation of TRPA1 Channel by Antibacterial Agent Triclosan Induces VEGF Secretion in Human Prostate Cancer Stromal Cells. *Cancer Prev. Res. Phila. Pa* **2017**, *10* (3), 177–187. https://doi.org/10.1158/1940-6207.CAPR-16-0257.
(250) Engeli, R. T.; Rohrer, S. R.; Vuorinen, A.; Herdlinger, S.; Kaserer, T.; Leugger, S.; Schuster, D.; Odermatt, A. Interference of Paraben Compounds with Estrogen Metabolism by Inhibition of 17β-Hydroxysteroid Dehydrogenases. *Int. J. Mol. Sci.* **2017**, *18* (9), E2007. https://doi.org/10.3390/ijms18092007.
(251) Sheikh, I. A.; Turki, R. F.; Abuzenadah, A. M.; Damanhouri, G. A.; Beg, M. A. Endocrine Disruption: Computational Perspectives on Human Sex Hormone-Binding Globulin and Phthalate Plasticizers. *PLOS ONE* **2016**, *11* (3), e0151444. https://doi.org/10.1371/journal.pone.0151444.

(252) Boas, M.; Feldt-Rasmussen, U.; Main, K. M. Thyroid Effects of Endocrine Disrupting Chemicals. *Mol. Cell. Endocrinol.* **2012**, *355* (2), 240–248. https://doi.org/10.1016/j.mce.2011.09.005.
(253) Fittipaldi, S.; Bimonte, V. M.; Soricelli, A.; Aversa, A.; Lenzi, A.; Greco, E. A.; Migliaccio, S. Cadmium Exposure Alters Steroid Receptors and Proinflammatory Cytokine Levels in Endothelial Cells in Vitro: A Potential Mechanism of Endocrine Disruptor Atherogenic Effect. *J. Endocrinol. Invest.* **2019**, *42* (6), 727–739. https://doi.org/10.1007/s40618-018-0982-1.
(254) *Investigation of the Effects of Subchronic Low Dose Oral Exposure to Bisphenol A (BPA) and Ethinyl Estradiol (EE) on Estrogen Receptor Expression in the Juvenile and Adult Female Rat Hypothalamus | Toxicological Sciences | Oxford Academic*. https://academic.oup.com/toxsci/article/140/1/190/1676105 (accessed 2022-06-24).
(255) Qiu, L.-L.; Wang, X.; Zhang, X.; Zhang, Z.; Gu, J.; Liu, L.; Wang, Y.; Wang, X.; Wang, S.-L. Decreased Androgen Receptor Expression May Contribute to Spermatogenesis Failure in Rats Exposed to Low Concentration of Bisphenol A. *Toxicol. Lett.* **2013**, *219* (2), 116–124. https://doi.org/10.1016/j.toxlet.2013.03.011.
(256) La Merrill, M. A.; Vandenberg, L. N.; Smith, M. T.; Goodson, W.; Browne, P.; Patisaul, H. B.; Guyton, K. Z.; Kortenkamp, A.; Cogliano, V. J.; Woodruff, T. J.; Rieswijk, L.; Sone, H.; Korach, K. S.; Gore, A. C.; Zeise, L.; Zoeller, R. T. Consensus on the Key Characteristics of Endocrine-Disrupting Chemicals as a Basis for Hazard Identification. *Nat. Rev. Endocrinol.* **2020**, *16* (1), 45–57. https://doi.org/10.1038/s41574-019-0273-8.
(257) Diamanti-Kandarakis, E.; Bourguignon, J.-P.; Giudice, L. C.; Hauser, R.; Prins, G. S.; Soto, A. M.; Zoeller, R. T.; Gore, A. C. Endocrine-Disrupting Chemicals: An Endocrine Society Scientific Statement. *Endocr. Rev.* **2009**, *30* (4), 293–342. https://doi.org/10.1210/er.2009-0002.
(258) Hamid, N.; Junaid, M.; Pei, D.-S. Combined Toxicity of Endocrine-Disrupting Chemicals: A Review. *Ecotoxicol. Environ. Saf.* **2021**, *215*, 112136. https://doi.org/10.1016/j.ecoenv.2021.112136.
(259) vom Saal, F. S.; Akingbemi, B. T.; Belcher, S. M.; Birnbaum, L. S.; Crain, D. A.; Eriksen, M.; Farabollini, F.; Guillette, L. J.; Hauser, R.; Heindel, J. J.; Ho, S.-M.; Hunt, P. A.; Iguchi, T.; Jobling, S.; Kanno, J.; Keri, R. A.; Knudsen, K. E.; Laufer, H.; LeBlanc, G. A.; Marcus, M.; McLachlan, J. A.; Myers, J. P.; Nadal, A.; Newbold, R. R.; Olea, N.; Prins, G. S.; Richter, C. A.; Rubin, B. S.; Sonnenschein, C.; Soto, A. M.; Talsness, C. E.; Vandenbergh, J. G.; Vandenberg, L. N.; Walser-Kuntz, D. R.; Watson, C. S.; Welshons, W. V.; Wetherill, Y.; Zoeller, R. T. Chapel Hill Bisphenol A Expert Panel Consensus Statement: Integration of Mechanisms, Effects in Animals and Potential to Impact Human Health at Current Levels of Exposure. *Reprod. Toxicol. Elmsford N* **2007**, *24* (2), 131–138. https://doi.org/10.1016/j.reprotox.2007.07.005.
(260) Wang, X.; Tian, J. Health Risks Related to Residential Exposure to Cadmium in Zhenhe County, China. *Arch. Environ. Health* **2004**, *59* (6), 324–330. https://doi.org/10.3200/AEOH.59.6.324-330.
(261) Farr, S. L.; Cooper, G. S.; Cai, J.; Savitz, D. A.; Sandler, D. P. Pesticide Use and Menstrual Cycle Characteristics among Premenopausal Women in the Agricultural Health Study. *Am. J. Epidemiol.* **2004**, *160* (12), 1194–1204. https://doi.org/10.1093/aje/kwi006.
(262) Woodruff, T. J.; Carlson, A.; Schwartz, J. M.; Giudice, L. C. Proceedings of the Summit on Environmental Challenges to Reproductive Health and Fertility: Executive Summary. *Fertil. Steril.* **2008**, *89* (2), 281–300. https://doi.org/10.1016/j.fertnstert.2007.10.002.

(263) Mendola, P.; Messer, L. C.; Rappazzo, K. Science Linking Environmental Contaminant Exposures with Fertility and Reproductive Health Impacts in the Adult Female. *Fertil. Steril.* **2008**, *89* (2 Suppl), e81-94. https://doi.org/10.1016/j.fertnstert.2007.12.036.
(264) Gibson, D. A.; Saunders, P. T. K. Endocrine Disruption of Oestrogen Action and Female Reproductive Tract Cancers. *Endocr. Relat. Cancer* **2014**, *21* (2), T13-31. https://doi.org/10.1530/ERC-13-0342.
(265) Costa, E. M. F.; Spritzer, P. M.; Hohl, A.; Bachega, T. A. S. S. Effects of Endocrine Disruptors in the Development of the Female Reproductive Tract. *Arq. Bras. Endocrinol. Metabol.* **2014**, *58* (2), 153–161. https://doi.org/10.1590/0004-2730000003031.
(266) Buck Louis, G. M.; Lynch, C. D.; Cooney, M. A. Environmental Influences on Female Fecundity and Fertility. *Semin. Reprod. Med.* **2006**, *24* (3), 147–155. https://doi.org/10.1055/s-2006-944421.
(267) Chang, S.-H.; Cheng, B.-H.; Lee, S.-L.; Chuang, H.-Y.; Yang, C.-Y.; Sung, F.-C.; Wu, T.-N. Low Blood Lead Concentration in Association with Infertility in Women. *Environ. Res.* **2006**, *101* (3), 380–386. https://doi.org/10.1016/j.envres.2005.10.004.
(268) Paasch, U.; Salzbrunn, A.; Glander, H. J.; Plambeck, K.; Salzbrunn, H.; Grunewald, S.; Stucke, J.; Vierula, M.; Skakkebaek, N. E.; Jørgensen, N. Semen Quality in Sub-Fertile Range for a Significant Proportion of Young Men from the General German Population: A Co-Ordinated, Controlled Study of 791 Men from Hamburg and Leipzig. *Int. J. Androl.* **2008**, *31* (2), 93–102. https://doi.org/10.1111/j.1365-2605.2007.00860.x.
(269) Boisen, K. A.; Chellakooty, M.; Schmidt, I. M.; Kai, C. M.; Damgaard, I. N.; Suomi, A.-M.; Toppari, J.; Skakkebaek, N. E.; Main, K. M. Hypospadias in a Cohort of 1072 Danish Newborn Boys: Prevalence and Relationship to Placental Weight, Anthropometrical Measurements at Birth, and Reproductive Hormone Levels at Three Months of Age. *J. Clin. Endocrinol. Metab.* **2005**, *90* (7), 4041–4046. https://doi.org/10.1210/jc.2005-0302.
(270) Boisen, K. A.; Kaleva, M.; Main, K. M.; Virtanen, H. E.; Haavisto, A.-M.; Schmidt, I. M.; Chellakooty, M.; Damgaard, I. N.; Mau, C.; Reunanen, M.; Skakkebaek, N. E.; Toppari, J. Difference in Prevalence of Congenital Cryptorchidism in Infants between Two Nordic Countries. *Lancet Lond. Engl.* **2004**, *363* (9417), 1264–1269. https://doi.org/10.1016/s0140-6736(04)15998-9.
(271) Main, K. M.; Skakkebaek, N. E.; Virtanen, H. E.; Toppari, J. Genital Anomalies in Boys and the Environment. *Best Pract. Res. Clin. Endocrinol. Metab.* **2010**, *24* (2), 279–289. https://doi.org/10.1016/j.beem.2009.10.003.
(272) *Association of endocrine disruptors and obesity: perspectives from epidemiological studies - Hatch - 2010 - International Journal of Andrology - Wiley Online Library*. https://onlinelibrary.wiley.com/doi/full/10.1111/j.1365-2605.2009.01035.x?casa_token=qU2uNOfGwiYAAAAA%3AzZMVq0H0Um0etXwgLqLezkR6FK3SCpRsqdnOisI3_mWpilotGdg2bxLKO-3tGHOt3ZV_G4XAwZrh1ko (accessed 2022-06-24).
(273) McAllister, E. J.; Dhurandhar, N. V.; Keith, S. W.; Aronne, L. J.; Barger, J.; Baskin, M.; Benca, R. M.; Biggio, J.; Boggiano, M. M.; Eisenmann, J. C.; Elobeid, M.; Fontaine, K. R.; Gluckman, P.; Hanlon, E. C.; Katzmarzyk, P.; Pietrobelli, A.; Redden, D. T.; Ruden, D. M.; Wang, C.; Waterland, R. A.; Wright, S. M.; Allison, D. B. Ten Putative Contributors to the Obesity Epidemic. *Crit. Rev. Food Sci. Nutr.* **2009**, *49* (10), 868–913. https://doi.org/10.1080/10408390903372599.
(274) Borrell, B. Toxicology: The Big Test for Bisphenol A. *Nature* **2010**, *464* (7292), 1122–1124. https://doi.org/10.1038/4641122a.
(275) Erickson, M. D. PCB PROPERTIES, USES, OCCURRENCE, AND REGULATORY HISTORY. 70.

(276) Giera, S.; Bansal, R.; Ortiz-Toro, T. M.; Taub, D. G.; Zoeller, R. T. Individual Polychlorinated Biphenyl (PCB) Congeners Produce Tissue- and Gene-Specific Effects on Thyroid Hormone Signaling during Development. *Endocrinology* **2011**, *152* (7), 2909–2919. https://doi.org/10.1210/en.2010-1490.
(277) Amano, I.; Miyazaki, W.; Iwasaki, T.; Shimokawa, N.; Koibuchi, N. The Effect of Hydroxylated Polychlorinated Biphenyl (OH-PCB) on Thyroid Hormone Receptor (TR)-Mediated Transcription through Native-Thyroid Hormone Response Element (TRE). *Ind. Health* **2010**, *48* (1), 115–118. https://doi.org/10.2486/indhealth.48.115.
(278) Gilbert, M. E.; Rovet, J.; Chen, Z.; Koibuchi, N. Developmental Thyroid Hormone Disruption: Prevalence, Environmental Contaminants and Neurodevelopmental Consequences. *Neurotoxicology* **2012**, *33* (4), 842–852. https://doi.org/10.1016/j.neuro.2011.11.005.
(279) *Current and Potential Rodent Screens and Tests for Thyroid Toxicants: Critical Reviews in Toxicology: Vol 37, No 1-2*. https://www.tandfonline.com/doi/abs/10.1080/10408440601123461 (accessed 2022-06-24).
(280) Kamrin, M. A. Phthalate Risks, Phthalate Regulation, and Public Health: A Review. *J. Toxicol. Environ. Health B Crit. Rev.* **2009**, *12* (2), 157–174. https://doi.org/10.1080/10937400902729226.
(281) Foster, P. M. D. Disruption of Reproductive Development in Male Rat Offspring Following in Utero Exposure to Phthalate Esters. *Int. J. Androl.* **2006**, *29* (1), 140–147; discussion 181-185. https://doi.org/10.1111/j.1365-2605.2005.00563.x.
(282) Gray, L. E.; Barlow, N. J.; Howdeshell, K. L.; Ostby, J. S.; Furr, J. R.; Gray, C. L. Transgenerational Effects of Di (2-Ethylhexyl) Phthalate in the Male CRL:CD(SD) Rat: Added Value of Assessing Multiple Offspring per Litter. *Toxicol. Sci. Off. J. Soc. Toxicol.* **2009**, *110* (2), 411–425. https://doi.org/10.1093/toxsci/kfp109.
(283) Noriega, N. C.; Howdeshell, K. L.; Furr, J.; Lambright, C. R.; Wilson, V. S.; Gray, L. E. Pubertal Administration of DEHP Delays Puberty, Suppresses Testosterone Production, and Inhibits Reproductive Tract Development in Male Sprague-Dawley and Long-Evans Rats. *Toxicol. Sci. Off. J. Soc. Toxicol.* **2009**, *111* (1), 163–178. https://doi.org/10.1093/toxsci/kfp129.
(284) Weber Lozada, K.; Keri, R. A. Bisphenol A Increases Mammary Cancer Risk in Two Distinct Mouse Models of Breast Cancer. *Biol. Reprod.* **2011**, *85* (3), 490–497. https://doi.org/10.1095/biolreprod.110.090431.
(285) *Exposure to the Endocrine Disruptor Bisphenol A Alters Susceptibility for Mammary Cancer - PubMed*. https://pubmed.ncbi.nlm.nih.gov/21687816/ (accessed 2022-06-24).
(286) Cheng, J.; Lee, E. J.; Madison, L. D.; Lazennec, G. Expression of Estrogen Receptor Beta in Prostate Carcinoma Cells Inhibits Invasion and Proliferation and Triggers Apoptosis. *FEBS Lett.* **2004**, *566* (1–3), 169–172. https://doi.org/10.1016/j.febslet.2004.04.025.
(287) Prins, G. S.; Birch, L.; Tang, W.-Y.; Ho, S.-M. Developmental Estrogen Exposures Predispose to Prostate Carcinogenesis with Aging. *Reprod. Toxicol. Elmsford N* **2007**, *23* (3), 374–382. https://doi.org/10.1016/j.reprotox.2006.10.001.
(288) Ho, S.-M.; Tang, W.-Y.; Belmonte de Frausto, J.; Prins, G. S. Developmental Exposure to Estradiol and Bisphenol A Increases Susceptibility to Prostate Carcinogenesis and Epigenetically Regulates Phosphodiesterase Type 4 Variant 4. *Cancer Res.* **2006**, *66* (11), 5624–5632. https://doi.org/10.1158/0008-5472.CAN-06-0516.
(289) Gore, A. C.; Patisaul, H. B. Neuroendocrine Disruption: Historical Roots, Current Progress, Questions for the Future. *Front. Neuroendocrinol.* **2010**, *31* (4), 395–399. https://doi.org/10.1016/j.yfrne.2010.07.003.

(290) Ahmed, O. M.; El-Gareib, A. W.; El-Bakry, A. M.; Abd El-Tawab, S. M.; Ahmed, R. G. Thyroid Hormones States and Brain Development Interactions. *Int. J. Dev. Neurosci. Off. J. Int. Soc. Dev. Neurosci.* **2008**, *26* (2), 147–209. https://doi.org/10.1016/j.ijdevneu.2007.09.011.
(291) Watson, C. S.; Alyea, R. A.; Cunningham, K. A.; Jeng, Y.-J. Estrogens of Multiple Classes and Their Role in Mental Health Disease Mechanisms. *Int. J. Womens Health* **2010**, *2*, 153–166.
(292) Palanza, P.; Gioiosa, L.; vom Saal, F. S.; Parmigiani, S. Effects of Developmental Exposure to Bisphenol A on Brain and Behavior in Mice. *Environ. Res.* **2008**, *108* (2), 150–157. https://doi.org/10.1016/j.envres.2008.07.023.
(293) Rasier, G.; Parent, A.-S.; Gérard, A.; Denooz, R.; Lebrethon, M.-C.; Charlier, C.; Bourguignon, J.-P. Mechanisms of Interaction of Endocrine-Disrupting Chemicals with Glutamate-Evoked Secretion of Gonadotropin-Releasing Hormone. *Toxicol. Sci. Off. J. Soc. Toxicol.* **2008**, *102* (1), 33–41. https://doi.org/10.1093/toxsci/kfm285.
(294) Toni, R. The Neuroendocrine System: Organization and Homeostatic Role. *J. Endocrinol. Invest.* **2004**, *27* (6 Suppl), 35–47.
(295) Gore, A. C. Neuroendocrine Systems as Targets for Environmental Endocrine-Disrupting Chemicals. *Fertil. Steril.* **2008**, *89* (2 Suppl), e101-102. https://doi.org/10.1016/j.fertnstert.2007.12.039.
(296) Herbstman, J.; Apelberg, B. J.; Witter, F. R.; Panny, S.; Goldman, L. R. Maternal, Infant, and Delivery Factors Associated with Neonatal Thyroid Hormone Status. *Thyroid Off. J. Am. Thyroid Assoc.* **2008**, *18* (1), 67–76. https://doi.org/10.1089/thy.2007.0180.
(297) Herbstman, J. B.; Sjödin, A.; Apelberg, B. J.; Witter, F. R.; Halden, R. U.; Patterson, D. G.; Panny, S. R.; Needham, L. L.; Goldman, L. R. Birth Delivery Mode Modifies the Associations between Prenatal Polychlorinated Biphenyl (PCB) and Polybrominated Diphenyl Ether (PBDE) and Neonatal Thyroid Hormone Levels. *Environ. Health Perspect.* **2008**, *116* (10), 1376–1382. https://doi.org/10.1289/ehp.11379.
(298) Baker, V. A. Endocrine Disrupters--Testing Strategies to Assess Human Hazard. *Toxicol. Vitro Int. J. Publ. Assoc. BIBRA* **2001**, *15* (4–5), 413–419. https://doi.org/10.1016/s0887-2333(01)00045-5.
(299) *Endocrine screening methods workshop report: Detection of estrogenic and androgenic hormonal and antihormonal activity for chemicals that act via receptor or steroidogenic enzyme mechanisms - ScienceDirect*. https://www.sciencedirect.com/science/article/pii/S0890623897000257 (accessed 2022-06-24).
(300) Molina-Molina, J.-M.; Escande, A.; Pillon, A.; Gomez, E.; Pakdel, F.; Cavaillès, V.; Olea, N.; Aït-Aïssa, S.; Balaguer, P. Profiling of Benzophenone Derivatives Using Fish and Human Estrogen Receptor-Specific in Vitro Bioassays. *Toxicol. Appl. Pharmacol.* **2008**, *232* (3), 384–395. https://doi.org/10.1016/j.taap.2008.07.017.
(301) Marin-Kuan, M.; Fussell, K. C.; Riederer, N.; Latado, H.; Serrant, P.; Mollergues, J.; Coulet, M.; Schilter, B. Differentiating True Androgen Receptor Inhibition from Cytotoxicity-Mediated Reduction of Reporter-Gene Transactivation in-Vitro. *Toxicol. In Vitro* **2017**, *45*, 359–365. https://doi.org/10.1016/j.tiv.2017.03.014.
(302) Dekant, W.; Colnot, T. Endocrine Effects of Chemicals: Aspects of Hazard Identification and Human Health Risk Assessment. *Toxicol. Lett.* **2013**, *223* (3), 280–286. https://doi.org/10.1016/j.toxlet.2013.03.022.
(303) Li, L.; Andersen, M. E.; Heber, S.; Zhang, Q. Non-Monotonic Dose-Response Relationship in Steroid Hormone Receptor-Mediated Gene Expression. *J. Mol. Endocrinol.* **2007**, *38* (5), 569–585. https://doi.org/10.1677/JME-07-0003.

(304) Villar-Pazos, S.; Martinez-Pinna, J.; Castellano-Muñoz, M.; Alonso-Magdalena, P.; Marroqui, L.; Quesada, I.; Gustafsson, J.-A.; Nadal, A. Molecular Mechanisms Involved in the Non-Monotonic Effect of Bisphenol-a on Ca2+ Entry in Mouse Pancreatic β-Cells. *Sci. Rep.* **2017**, *7* (1), 11770. https://doi.org/10.1038/s41598-017-11995-3.
(305) Vandenberg, L. N.; Colborn, T.; Hayes, T. B.; Heindel, J. J.; Jacobs, D. R.; Lee, D.-H.; Shioda, T.; Soto, A. M.; vom Saal, F. S.; Welshons, W. V.; Zoeller, R. T.; Myers, J. P. Hormones and Endocrine-Disrupting Chemicals: Low-Dose Effects and Nonmonotonic Dose Responses. *Endocr. Rev.* **2012**, *33* (3), 378–455. https://doi.org/10.1210/er.2011-1050.
(306) Borgert, C. J.; Baker, S. P.; Matthews, J. C. Potency Matters: Thresholds Govern Endocrine Activity. *Regul. Toxicol. Pharmacol. RTP* **2013**, *67* (1), 83–88. https://doi.org/10.1016/j.yrtph.2013.06.007.
(307) Solecki, R.; Kortenkamp, A.; Bergman, Å.; Chahoud, I.; Degen, G. H.; Dietrich, D.; Greim, H.; Håkansson, H.; Hass, U.; Husoy, T.; Jacobs, M.; Jobling, S.; Mantovani, A.; Marx-Stoelting, P.; Piersma, A.; Ritz, V.; Slama, R.; Stahlmann, R.; van den Berg, M.; Zoeller, R. T.; Boobis, A. R. Scientific Principles for the Identification of Endocrine-Disrupting Chemicals: A Consensus Statement. *Arch. Toxicol.* **2017**, *91* (2), 1001–1006. https://doi.org/10.1007/s00204-016-1866-9.
(308) Rajapakse, N.; Silva, E.; Kortenkamp, A. Combining Xenoestrogens at Levels below Individual No-Observed-Effect Concentrations Dramatically Enhances Steroid Hormone Action. *Environ. Health Perspect.* **2002**, *110* (9), 917–921. https://doi.org/10.1289/ehp.02110917.
(309) *Something from "Nothing" − Eight Weak Estrogenic Chemicals Combined at Concentrations below NOECs Produce Significant Mixture Effects | Environmental Science & Technology*. https://pubs-acs-org.proxybib-pp.cnam.fr/doi/full/10.1021/es0101227?casa_token=7UOqyV02VbIAAAAA%3AWDtgaT0bvODHDR_3IfRLAhWczk65HHh4OMot7yXT4jvIRwWnIqrkr1lniARlZchlLz-C2LvGkNSU (accessed 2022-06-25).
(310) US EPA, O. *Endocrine Disruptor Screening and Testing Advisory Committee (EDSTAC) Final Report*. https://www.epa.gov/endocrine-disruption/endocrine-disruptor-screening-and-testing-advisory-committee-edstac-final (accessed 2022-06-25).
(311) Schneider, M.; Pons, J.-L.; Labesse, G.; Bourguet, W. In Silico Predictions of Endocrine Disruptors Properties. *Endocrinology* **2019**, *160* (11), 2709–2716. https://doi.org/10.1210/en.2019-00382.
(312) *On the mechanism of estrogen action - PubMed*. https://pubmed.ncbi.nlm.nih.gov/13957617/ (accessed 2022-07-07).
(313) Hollenberg, S. M.; Weinberger, C.; Ong, E. S.; Cerelli, G.; Oro, A.; Lebo, R.; Thompson, E. B.; Rosenfeld, M. G.; Evans, R. M. Primary Structure and Expression of a Functional Human Glucocorticoid Receptor CDNA. *Nature* **1985**, *318* (6047), 635–641. https://doi.org/10.1038/318635a0.
(314) Green, S.; Walter, P.; Kumar, V.; Krust, A.; Bornert, J. M.; Argos, P.; Chambon, P. Human Oestrogen Receptor CDNA: Sequence, Expression and Homology to v-Erb-A. *Nature* **1986**, *320* (6058), 134–139. https://doi.org/10.1038/320134a0.
(315) Arriza, J. L.; Weinberger, C.; Cerelli, G.; Glaser, T. M.; Handelin, B. L.; Housman, D. E.; Evans, R. M. Cloning of Human Mineralocorticoid Receptor Complementary DNA: Structural and Functional Kinship with the Glucocorticoid Receptor. *Science* **1987**, *237* (4812), 268–275. https://doi.org/10.1126/science.3037703.
(316) Evans, R. M. The Steroid and Thyroid Hormone Receptor Superfamily. *Science* **1988**, *240* (4854), 889–895. https://doi.org/10.1126/science.3283939.

(317) Petkovich, M.; Brand, N. J.; Krust, A.; Chambon, P. A Human Retinoic Acid Receptor Which Belongs to the Family of Nuclear Receptors. *Nature* **1987**, *330* (6147), 444–450. https://doi.org/10.1038/330444a0.
(318) Mullican, S. E.; Dispirito, J. R.; Lazar, M. A. The Orphan Nuclear Receptors at Their 25-Year Reunion. *J. Mol. Endocrinol.* **2013**, *51* (3), T115-140. https://doi.org/10.1530/JME-13-0212.
(319) Giguère, V.; Yang, N.; Segui, P.; Evans, R. M. Identification of a New Class of Steroid Hormone Receptors. *Nature* **1988**, *331* (6151), 91–94. https://doi.org/10.1038/331091a0.
(320) Stodden, V.; McNutt, M.; Bailey, D. H.; Deelman, E.; Gil, Y.; Hanson, B.; Heroux, M. A.; Ioannidis, J. P. A.; Taufer, M. Enhancing Reproducibility for Computational Methods. *Science* **2016**, *354* (6317), 1240–1241. https://doi.org/10.1126/science.aah6168.
(321) Helsen, C.; Dubois, V.; Verfaillie, A.; Young, J.; Trekels, M.; Vancraenenbroeck, R.; De Maeyer, M.; Claessens, F. Evidence for DNA-Binding Domain–Ligand-Binding Domain Communications in the Androgen Receptor. *Mol. Cell. Biol.* **2012**, *32* (15), 3033–3043. https://doi.org/10.1128/MCB.00151-12.
(322) Vedani, A.; Dobler, M.; Smieško, M. VirtualToxLab — A Platform for Estimating the Toxic Potential of Drugs, Chemicals and Natural Products. *Toxicol. Appl. Pharmacol.* **2012**, *261* (2), 142–153. https://doi.org/10.1016/j.taap.2012.03.018.
(323) Vedani, A.; Dobler, M.; Hu, Z.; Smieško, M. OpenVirtualToxLab—A Platform for Generating and Exchanging in Silico Toxicity Data. *Toxicol. Lett.* **2015**, *232* (2), 519–532. https://doi.org/10.1016/j.toxlet.2014.09.004.
(324) Kolšek, K.; Mavri, J.; Sollner Dolenc, M.; Gobec, S.; Turk, S. Endocrine Disruptome—An Open Source Prediction Tool for Assessing Endocrine Disruption Potential through Nuclear Receptor Binding. *J. Chem. Inf. Model.* **2014**, *54* (4), 1254–1267. https://doi.org/10.1021/ci400649p.
(325) Warr, W. A.; Nicklaus, M. C.; Nicolaou, C. A.; Rarey, M. Exploration of Ultralarge Compound Collections for Drug Discovery. *J. Chem. Inf. Model.* **2022**, *62* (9), 2021–2034. https://doi.org/10.1021/acs.jcim.2c00224.
(326) *NRLiSt BDB, the Manually Curated Nuclear Receptors Ligands and Structures Benchmarking Database | Journal of Medicinal Chemistry*. https://pubs.acs.org/doi/abs/10.1021/jm500132p (accessed 2021-12-09).
(327) Chan, W. K. B.; Zhang, H.; Yang, J.; Brender, J. R.; Hur, J.; Özgür, A.; Zhang, Y. GLASS: A Comprehensive Database for Experimentally Validated GPCR-Ligand Associations. *Bioinforma. Oxf. Engl.* **2015**, *31* (18), 3035–3042. https://doi.org/10.1093/bioinformatics/btv302.
(328) *KLIFS: A Knowledge-Based Structural Database To Navigate Kinase–Ligand Interaction Space | Journal of Medicinal Chemistry*. https://pubs-acs-org.proxybib-pp.cnam.fr/doi/10.1021/jm400378w (accessed 2022-07-10).
(329) *Endocrine disruptor assessment list - ECHA*. https://echa.europa.eu/da/ed-assessment (accessed 2022-07-04).
(330) Fourches, D.; Muratov, E.; Tropsha, A. Trust, but Verify II: A Practical Guide to Chemogenomics Data Curation. *J. Chem. Inf. Model.* **2016**, *56* (7), 1243–1252. https://doi.org/10.1021/acs.jcim.6b00129.
(331) Kalliokoski, T.; Kramer, C.; Vulpetti, A.; Gedeck, P. Comparability of Mixed IC_{50} Data - a Statistical Analysis. *PloS One* **2013**, *8* (4), e61007. https://doi.org/10.1371/journal.pone.0061007.
(332) Website of the European Commission, Environment Section.
(333) Williams, A. J.; Grulke, C. M.; Edwards, J.; McEachran, A. D.; Mansouri, K.; Baker, N. C.; Patlewicz, G.; Shah, I.; Wambaugh, J. F.; Judson, R. S.; Richard, A. M. The

CompTox Chemistry Dashboard: A Community Data Resource for Environmental Chemistry. *J. Cheminformatics* **2017**, *9* (1), 61. https://doi.org/10.1186/s13321-017-0247-6.
(334) Devillers, J.; Marchand-Geneste, N.; Carpy, A.; Porcher, J. M. SAR and QSAR Modeling of Endocrine Disruptors. *SAR QSAR Environ. Res.* **2006**, *17* (4), 393–412. https://doi.org/10.1080/10629360600884397.
(335) Shanle, E. K.; Xu, W. Endocrine Disrupting Chemicals Targeting Estrogen Receptor Signaling: Identification and Mechanisms of Action. *Chem. Res. Toxicol.* **2011**, *24* (1), 6–19. https://doi.org/10.1021/tx100231n.
(336) Riggs, B. L.; Hartmann, L. C. Selective Estrogen-Receptor Modulators — Mechanisms of Action and Application to Clinical Practice. *N. Engl. J. Med.* **2003**, *348* (7), 618–629. https://doi.org/10.1056/NEJMra022219.
(337) Schneider, M.; Pons, J.-L.; Labesse, G. Exploring the Conformational Space of a Receptor for Drug Design: An ERα Case Study. *J. Mol. Graph. Model.* **2021**, *108*, 107974. https://doi.org/10.1016/j.jmgm.2021.107974.
(338) Swaby, R. F.; Sharma, C. G. N.; Jordan, V. C. SERMs for the Treatment and Prevention of Breast Cancer. *Rev. Endocr. Metab. Disord.* **2007**, *8* (3), 229–239. https://doi.org/10.1007/s11154-007-9034-4.
(339) Lagarde, N.; Delahaye, S.; Zagury, J.-F.; Montes, M. Discriminating Agonist and Antagonist Ligands of the Nuclear Receptors Using 3D-Pharmacophores. *J. Cheminformatics* **2016**, *8* (1), 43. https://doi.org/10.1186/s13321-016-0154-2.
(340) Casals-Casas, C.; Feige, J. N.; Desvergne, B. Interference of Pollutants with PPARs: Endocrine Disruption Meets Metabolism. *Int. J. Obes. 2005* **2008**, *32 Suppl 6*, S53-61. https://doi.org/10.1038/ijo.2008.207.
(341) *ChemAxon - Software Solutions and Services for Chemistry & Biology*. https://chemaxon.com/ (accessed 2020-07-27).
(342) Korb, O.; Stützle, T.; Exner, T. E. PLANTS: Application of Ant Colony Optimization to Structure-Based Drug Design. In *Ant Colony Optimization and Swarm Intelligence*; Dorigo, M., Gambardella, L. M., Birattari, M., Martinoli, A., Poli, R., Stützle, T., Eds.; Hutchison, D., Kanade, T., Kittler, J., Kleinberg, J. M., Mattern, F., Mitchell, J. C., Naor, M., Nierstrasz, O., Pandu Rangan, C., Steffen, B., Sudan, M., Terzopoulos, D., Tygar, D., Vardi, M. Y., Weikum, G., Series Eds.; Lecture Notes in Computer Science; Springer Berlin Heidelberg: Berlin, Heidelberg, 2006; Vol. 4150, pp 247–258. https://doi.org/10.1007/11839088_22.
(343) Trott, O.; Olson, A. J. AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading. *J. Comput. Chem.* **2009**, NA-NA. https://doi.org/10.1002/jcc.21334.
(344) Quiroga, R.; Villarreal, M. A. Vinardo: A Scoring Function Based on Autodock Vina Improves Scoring, Docking, and Virtual Screening. *PloS One* **2016**, *11* (5), e0155183. https://doi.org/10.1371/journal.pone.0155183.
(345) *AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility - Morris - 2009 - Journal of Computational Chemistry - Wiley Online Library*. https://onlinelibrary-wiley-com.proxybib-pp.cnam.fr/doi/full/10.1002/jcc.21256 (accessed 2022-07-26).
(346) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, É. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12* (85), 2825–2830.
(347) Sellami, A.; Montes, M.; Lagarde, N. Predicting Potential Endocrine Disrupting Chemicals Binding to Estrogen Receptor α (ERα) Using a Pipeline Combining Structure-

Based and Ligand-Based in Silico Methods. *Int. J. Mol. Sci.* **2021**, *22* (6), 2846. https://doi.org/10.3390/ijms22062846.
(348) Empereur-mot, C.; Guillemain, H.; Latouche, A.; Zagury, J.-F.; Viallon, V.; Montes, M. Predictiveness Curves in Virtual Screening. *J. Cheminformatics* **2015**, *7* (1), 52. https://doi.org/10.1186/s13321-015-0100-8.
(349) Seidel, T.; Ibis, G.; Bendix, F.; Wolber, G. Strategies for 3D Pharmacophore-Based Virtual Screening. *Drug Discov. Today Technol.* **2010**, *7* (4), e221–e228. https://doi.org/10.1016/j.ddtec.2010.11.004.
(350) *QPHAR: quantitative pharmacophore activity relationship: method and validation | Journal of Cheminformatics | Full Text*. https://jcheminf.biomedcentral.com/articles/10.1186/s13321-021-00537-9 (accessed 2022-07-18).
(351) Darbre, P. D. The History of Endocrine-Disrupting Chemicals. *Curr. Opin. Endocr. Metab. Res.* **2019**, *7*, 26–33. https://doi.org/10.1016/j.coemr.2019.06.007.
(352) Vuorinen, A.; Odermatt, A.; Schuster, D. In Silico Methods in the Discovery of Endocrine Disrupting Chemicals. *J. Steroid Biochem. Mol. Biol.* **2013**, *137*, 18–26. https://doi.org/10.1016/j.jsbmb.2013.04.009.
(353) Beg, M. A.; Sheikh, I. A. Endocrine Disruption: Molecular Interactions of Environmental Bisphenol Contaminants with Thyroid Hormone Receptor and Thyroxine-Binding Globulin. *Toxicol. Ind. Health* **2020**, *36* (5), 322–335. https://doi.org/10.1177/0748233720928165.
(354) Mal, R.; Magner, A.; David, J.; Datta, J.; Vallabhaneni, M.; Kassem, M.; Manouchehri, J.; Willingham, N.; Stover, D.; Vandeusen, J.; Sardesai, S.; Williams, N.; Wesolowski, R.; Lustberg, M.; Ganju, R. K.; Ramaswamy, B.; Cherian, M. A. Estrogen Receptor Beta (ERβ): A Ligand Activated Tumor Suppressor. *Front. Oncol.* **2020**, *10*.
(355) Rotman, N.; Haftek-Terreau, Z.; Lücke, S.; Feige, J.; Gelman, L.; Desvergne, B.; Wahli, W. PPAR Disruption: Cellular Mechanisms and Physiological Consequences. *CHIMIA* **2008**, *62* (5), 340–340. https://doi.org/10.2533/chimia.2008.340.
(356) Möcklinghoff, S.; van Otterlo, W. A. L.; Rose, R.; Fuchs, S.; Zimmermann, T. J.; Dominguez Seoane, M.; Waldmann, H.; Ottmann, C.; Brunsveld, L. Design and Evaluation of Fragment-Like Estrogen Receptor Tetrahydroisoquinoline Ligands from a Scaffold-Detection Approach. *J. Med. Chem.* **2011**, *54* (7), 2005–2011. https://doi.org/10.1021/jm1011116.
(357) Möcklinghoff, S.; Rose, R.; Carraz, M.; Visser, A.; Ottmann, C.; Brunsveld, L. Synthesis and Crystal Structure of a Phosphorylated Estrogen Receptor Ligand Binding Domain. *ChemBioChem* **2010**, *11* (16), 2251–2254. https://doi.org/10.1002/cbic.201000532.
(358) Bledsoe, R. K.; Montana, V. G.; Stanley, T. B.; Delves, C. J.; Apolito, C. J.; McKee, D. D.; Consler, T. G.; Parks, D. J.; Stewart, E. L.; Willson, T. M.; Lambert, M. H.; Moore, J. T.; Pearce, K. H.; Xu, H. E. Crystal Structure of the Glucocorticoid Receptor Ligand Binding Domain Reveals a Novel Mode of Receptor Dimerization and Coactivator Recognition. *Cell* **2002**, *110* (1), 93–105. https://doi.org/10.1016/S0092-8674(02)00817-6.
(359) Hemmerling, M.; Nilsson, S.; Edman, K.; Eirefelt, S.; Russell, W.; Hendrickx, R.; Johnsson, E.; Kärrman Mårdh, C.; Berger, M.; Rehwinkel, H.; Abrahamsson, A.; Dahmén, J.; Eriksson, A. R.; Gabos, B.; Henriksson, K.; Hossain, N.; Ivanova, S.; Jansson, A.-H.; Jensen, T. J.; Jerre, A.; Johansson, H.; Klingstedt, T.; Lepistö, M.; Lindsjö, M.; Mile, I.; Nikitidis, G.; Steele, J.; Tehler, U.; Wissler, L.; Hansson, T. Selective Nonsteroidal Glucocorticoid Receptor Modulators for the Inhaled Treatment of Pulmonary Diseases. *J. Med. Chem.* **2017**, *60* (20), 8591–8605. https://doi.org/10.1021/acs.jmedchem.7b01215.

(360) Lagarde, N.; Delahaye, S.; Jérémie, A.; Ben Nasr, N.; Guillemain, H.; Empereur-Mot, C.; Laville, V.; Labib, T.; Réau, M.; Langenfeld, F.; Zagury, J.-F.; Montes, M. Discriminating Agonist from Antagonist Ligands of the Nuclear Receptors Using Different Chemoinformatics Approaches. *Mol. Inform.* **2017**, *36* (10). https://doi.org/10.1002/minf.201700020.
(361) Nolte, R. T.; Wisely, G. B.; Westin, S.; Cobb, J. E.; Lambert, M. H.; Kurokawa, R.; Rosenfeld, M. G.; Willson, T. M.; Glass, C. K.; Milburn, M. V. Ligand Binding and Co-Activator Assembly of the Peroxisome Proliferator-Activated Receptor-γ. *Nature* **1998**, *395* (6698), 137–143. https://doi.org/10.1038/25931.
(362) *Development and Validation of an In Silico P450 Profiler Based on...: Ingenta Connect*. https://www.ingentaconnect.com/content/ben/cddt/2006/00000003/00000001/art00001 (accessed 2022-07-10).
(363) Maaten, L. van der; Hinton, G. Visualizing Data Using T-SNE. *J. Mach. Learn. Res.* **2008**, *9* (86), 2579–2605.
(364) Sander, T.; Freyss, J.; von Korff, M.; Rufener, C. DataWarrior: An Open-Source Program for Chemistry Aware Data Visualization and Analysis. *J. Chem. Inf. Model.* **2015**, *55* (2), 460–473. https://doi.org/10.1021/ci500588j.
(365) *Endocrine Disruptor List*. https://edlists.org/ (accessed 2022-07-13).
(366) Chen, G. Methoxychlor. In *Encyclopedia of Toxicology (Third Edition)*; Wexler, P., Ed.; Academic Press: Oxford, 2014; pp 254–255. https://doi.org/10.1016/B978-0-12-386454-3.00162-7.
(367) Delfosse, V.; Grimaldi, M.; Cavaill, ès V.; Balaguer, P.; Bourguet, W. Structural and Functional Profiling of Environmental Ligands for Estrogen Receptors. *Environ. Health Perspect.* **2014**, *122* (12), 1306–1313. https://doi.org/10.1289/ehp.1408453.
(368) Lunghini, F.; Marcou, G.; Azam, P.; Bonachera, F.; Enrici, M. H.; Van Miert, E.; Varnek, A. Endocrine Disruption: The Noise in Available Data Adversely Impacts the Models' Performance. *SAR QSAR Environ. Res.* **2021**, *32* (2), 111–131. https://doi.org/10.1080/1062936X.2020.1864468.
(369) Sun, L.; Yang, H.; Cai, Y.; Li, W.; Liu, G.; Tang, Y. In Silico Prediction of Endocrine Disrupting Chemicals Using Single-Label and Multilabel Models. *J. Chem. Inf. Model.* **2019**, *59* (3), 973–982. https://doi.org/10.1021/acs.jcim.8b00551.
(370) Nettles, K. W.; Bruning, J. B.; Gil, G.; Nowak, J.; Sharma, S. K.; Hahm, J. B.; Kulp, K.; Hochberg, R. B.; Zhou, H.; Katzenellenbogen, J. A.; Katzenellenbogen, B. S.; Kim, Y.; Joachmiak, A.; Greene, G. L. NFkappaB Selectivity of Estrogen Receptor Ligands Revealed by Comparative Crystallographic Analyses. *Nat. Chem. Biol.* **2008**, *4* (4), 241–247. https://doi.org/10.1038/nchembio.76.
(371) Cheng, T.; Li, X.; Li, Y.; Liu, Z.; Wang, R. Comparative Assessment of Scoring Functions on a Diverse Test Set. *J. Chem. Inf. Model.* **2009**, *49* (4), 1079–1093. https://doi.org/10.1021/ci9000053.
(372) *SFCscore: Scoring functions for affinity prediction of protein–ligand complexes - Sotriffer - 2008 - Proteins: Structure, Function, and Bioinformatics - Wiley Online Library*. https://onlinelibrary-wiley-com.proxybib-pp.cnam.fr/doi/10.1002/prot.22058 (accessed 2022-07-14).
(373) Chang, M. W.; Ayeni, C.; Breuer, S.; Torbett, B. E. Virtual Screening for HIV Protease Inhibitors: A Comparison of AutoDock 4 and Vina. *PLOS ONE* **2010**, *5* (8), e11955. https://doi.org/10.1371/journal.pone.0011955.
(374) Böhm, H. J. The Development of a Simple Empirical Scoring Function to Estimate the Binding Constant for a Protein-Ligand Complex of Known Three-Dimensional Structure. *J. Comput. Aided Mol. Des.* **1994**, *8* (3), 243–256. https://doi.org/10.1007/BF00126743.

(375) Hu, L.; Benson, M. L.; Smith, R. D.; Lerner, M. G.; Carlson, H. A. Binding MOAD (Mother Of All Databases). *Proteins Struct. Funct. Bioinforma.* **2005**, *60* (3), 333–340. https://doi.org/10.1002/prot.20512.
(376) Korb, O.; Stützle, T.; Exner, T. E. Empirical Scoring Functions for Advanced Protein−Ligand Docking with PLANTS. *J. Chem. Inf. Model.* **2009**, *49* (1), 84–96. https://doi.org/10.1021/ci800298z.
(377) Kratz, J. M.; Schuster, D.; Edtbauer, M.; Saxena, P.; Mair, C. E.; Kirchebner, J.; Matuszczak, B.; Baburin, I.; Hering, S.; Rollinger, J. M. Experimentally Validated HERG Pharmacophore Models as Cardiotoxicity Prediction Tools. *J. Chem. Inf. Model.* **2014**, *54* (10), 2887–2901. https://doi.org/10.1021/ci5001955.
(378) Balaguer, P.; Delfosse, V.; Bourguet, W. Mechanisms of Endocrine Disruption through Nuclear Receptors and Related Pathways. *Curr. Opin. Endocr. Metab. Res.* **2019**, *7*, 1–8. https://doi.org/10.1016/j.coemr.2019.04.008.

Annexes

Annexe 1 : Liste des structures PDB utilisées pour l'étude des modèles de docking et de pharmacophores pour les 6 NR étudiés

Docking	
AR	1e3g ;1t5z ;1t63 ;1xj7 ;2am9 ;2hvc ;2pio ;2pip ;2piq ;2pir ;2pit ;2piu ;2piv ;2piw ;2pix ; 2pkl ;2q7i ;2z4j ;3b5r ;3b65 ;3b66 ;3b67 ;3b68 ;3l3x ;3rlj ;5jjm ;5t8e ;5t8j ;5v8q
ERA	1a52 ;1g50 ;1qku ;1x7e ;1xp1 ;1xp9 ;2yja
ERB	1u3q ;1u3r ;1u3s ;3ols ;3omo ;3omp ;3omq ;5toa
GR	4p6x ;6dxk ;5g3j ;3cld ;4udc ;4udd ;5nfp ;4mdd ;5nft
PPARG	2hfp ;2zk6 ;3cdp ;2i4p ;1zgy ;2zk3 ;3adw ;3gbk ;2g0h ;3noa ;2prg ;3ho0 ;2zk5 ;1k74 ;5two ; 2zk2 ;3hod ;2zk4 ;3ty0 ;1fm6;3cds ;2zk1 ;2zk0 ;2ath ; 1fm9 ;2f4b ;3adx ;2i4j ;4jaz ;3d6d ;2i4z ;1prg ;4e4k ;2g0g ;2om9 ;4prg ;3r8i ;3b3k ;4e4q
TRA	1nav ;3hzf ;3ilz ;3jzb
Modèles de pharmacophores	
AR	1t5z; 1t63; 1t65; 1xj7; 2am9; 2hvc; 2pio; 2pip; 2piq; 2pir; 2pit; 2piu; 2piv; 2piw; 2pix; 2pkl; 2q7i; 2z4j; 3b5r; 3b65; 3b66; 3b67; 3b68; 3l3x; 3rlj; 5jjm; 5t8e; 5t8j; 5v8q
ERA	1gwq ;1x7e ;1xp6 ;1yim ;1yin ;2bj4 ;2iog ;2jf9 ;2jfa ;2ouz ;3dt3 ;3erd ;3ert ;5t92 ;5t97
ERB	1l2j; 1qkm; 1u3q; 1u3r; 1u3s; 1u9e; 1x76; 1x78; 1x7b; 1x7j; 1yy4; 1yye; 1zaf; 2giu; 2nv7; 3oll ; 3ols ;3omo; 3omp; 3omq; 4j24; 4j26; 5toa
GR	3cld, 4mdd, 4p6x, 4udc, 4udd, 5g3j, 5nfp, 5nft, 6dxk
PPARG	1fm6 ;1fm9 ;1k74 ;1zgy ;2ath ;2f4b ;2g0g ;2g0h ;2hfp ;2i4j ;2i4p ;2i4z ;2prg ;2zk1 ;2zk2 ;2zk3 ; 2zk4 ;2zk5 ;2zk6 ;3adw ;3adx ;3b3k ;3cdp ; 3cds ;3d6d ;3gbk ;3ho0 ;3hod ;3noa ;3r8i ;3ty0 ; 4e4k ;4e4q ;4jaz ;4prg ;5two
TRA	1nav ;3hzf ;3ilz ;3jzb

Annexe 2: Détails des performances de docking pour AR

Scoring function	Approach	Best PDB	AUC	Min	Mean	Std
smina-dkoes	Single structure	[3b66-]	0,751	0,734	0,745	0,004
	Ensemble of 2 structures	[3b66-1xj7-]	0,752	0,737	0,746	0,002
	Ensemble of 3 structures	[3b66-2pkl-1xj7-]	0,752	0,738	0,747	0,000
smina-vina	Single structure	[2pir-]	0,656	0,474	0,591	0,042
	Ensemble of 2 structures	[3b67-2pir-]	0,665	0,510	0,618	0,025
	Ensemble of 3 structures	[3b67-2pix-2pir-]	0,667	0,533	0,629	0,000
smina-vinardo	Single structure	[2pir-]	0,624	0,449	0,551	0,040
	Ensemble of 2 structures	[3b67-2pir-]	0,636	0,449	0,575	0,029
	Ensemble of 3 structures	[3b67-2pir-2pip-]	0,638	0,451	0,587	0,000
smina-ad4	Single structure	[5v8q-]	0,500	0,500	0,500	0,000
	Ensemble of 2 structures	[5v8q-5t8j-]	0,500	0,500	0,500	0,000
	Ensemble of 3 structures	[5v8q-5t8j-5t8e-]	0,500	0,500	0,500	0,000
PLANTS	Single structure	[2pir-]	0,691	0,630	0,665	0,015
	Ensemble of 2 structures	[2pix-2pir-]	0,694	0,643	0,674	0,008
	Ensemble of 3 structures	[2pix-2pit-2pir-]	0,695	0,654	0,678	0,000
Surflex-dock	Single structure	[5t8e-]	0,623	0,468	0,526	0,040
	Ensemble of 2 structures	[3rlj-5t8e]	0,635	0,482	0,556	0,035
	Ensemble of 3 structures	[3rlj-5t8e-3b65-]	0,640	0,490	0,571	0,000

Annexe 3 : Détails des performances de docking pour ERα

Scoring function	Approach	Best PDB	AUC	Min	Mean	Std
smina-dkoes	Single structure	[1qku]	0,708	0,7	0,704	0,003
	Ensemble of 2 structures	[2yja-1qku]	0,709	0,702	0,703	0,003
	Ensemble of 3 structures	[2yja-1qku-1g50]	0,71	0,704	0,702	0,003
smina-vina	Single structure	[1a52]	0,699	0,643	0,676	0,02
	Ensemble of 2 structures	[1xp9-1a52]	0,696	0,642	0,67	0,017
	Ensemble of 3 structures	[1xp9-1xp1-1a52]	0,695	0,642	0,667	0,014
smina-vinardo	Single structure	[1a52]	0,68	0,686	0,704	0,018
	Ensemble of 2 structures	[1xp9-1a52]	0,676	0,619	0,65	0,019
	Ensemble of 3 structures	[1xp9-1xp1-1a52]	0,673	0,618	0,644	0,018
smina-ad4	Single structure	[1a52]	0,656	0,613	0,639	0,0154
	Ensemble of 2 structures	[1x7e-1a52]	0,654	0,618	0,641	0,009
	Ensemble of 3 structures	[1x7e-1qku-1a52]	0,65	0,623	0,64	0,007
PLANTS	Single structure	[1x7e]	0,659	0,598	0,634	0,019
	Ensemble of 2 structures	[1x7e-1a52]	0,66	0,647	0,62	0
	Ensemble of 3 structures	[1x7e-1qku-1a52]	0,659	0,62	0,642	0,009
Surflex-dock	Single structure	[1a52]	0,604	0,547	0,576	0,027
	Ensemble of 2 structures	[1xp1-1x7e]	0,616	0,556	0,594	0,02
	Ensemble of 3 structures	[1xp1-1x7e-1a52]	0,623	0,562	0,605	0,015

Annexe 4 : Détails des performances de docking pour ERβ

Scoring function	Approach	Best PDB	AUC	Min	Mean	Std
smina_dkoes	Single structure	[3omo-]	0,647	0,634	0,640	0,004
	Ensemble of 2 structures	[3omq-3omo-]	0,643	0,631	0,637	0,003
	Ensemble of 3 structures	[3omq-3omp-3omo-]	0,640	0,630	0,635	0,000
smina_vina	Single structure	[3ols-]	0,705	0,642	0,680	0,018
	Ensemble of 2 structures	[3omp-3ols-]	0,704	0,666	0,688	0,009
	Ensemble of 3 structures	[3omp-3omo-3ols-]	0,704	0,680	0,690	0,000
smina_vinardo	Single structure	[3ols-]	0,686	0,628	0,660	0,017
	Ensemble of 2 structures	[3omq-3ols-]	0,686	0,649	0,668	0,008
	Ensemble of 3 structures	[3omq-3omp-3ols-]	0,685	0,659	0,671	0,000
smina_ad4	Single structure	[3ols-]	0,622	0,595	0,610	0,009
	Ensemble of 2 structures	[3omq-3ols-]	0,622	0,602	0,612	0,006
	Ensemble of 3 structures	[5toa-3omq-3ols-]	0,621	0,602	0,611	0,000
PLANTS	Single structure	[3ols-]	0,638	0,604	0,625	0,011
	Ensemble of 2 structures	[3ols-1u3q-]	0,643	0,607	0,625	0,010
	Ensemble of 3 structures	[3omp-3ols-1u3q-]	0,643	0,608	0,624	0,000
Surflex-dock	Single structure	[3omp]	0,547	0,331	0,468	0,088
	Ensemble of 2 structures	[3omp-3omq]	0,561	0,332	0,495	0,067
	Ensemble of 3 structures	[3omp-3omo-1u3q]	0,562	0,333	0,514	0,000

Annexe 5: Détails des performances de docking pour GR

Scoring function	Approach	Best PDB	AUC	Min	Mean	Std
smina_dkoes	Single structure	[4mdd-]	0,67	0,65	0,66	0,01
	Ensemble of 2 structures	[4udc-6dxk-]	0,68	0,65	0,66	0,01
	Ensemble of 3 structures	[4udd-4udc-6dxk-]	0,68	0,65	0,67	0,00
smina_vina	Single structure	[6dxk-]	0,64	0,50	0,56	0,05
	Ensemble of 2 structures	[3cld-6dxk-]	0,64	0,52	0,59	0,04
	Ensemble of 3 structures	[4mdd-4udc-6dxk-]	0,64	0,54	0,60	0,00
smina_vinardo	Single structure	[6dxk-]	0,63	0,48	0,56	0,05
	Ensemble of 2 structures	[4mdd-3cld-]	0,65	0,52	0,59	0,04
	Ensemble of 3 structures	[4mdd-4udc-3cld-]	0,65	0,55	0,61	0,00
smina_ad4	Single structure	[6dxk-]	0,63	0,55	0,59	0,03
	Ensemble of 2 structures	[4mdd-4udd-]	0,64	0,57	0,60	0,02
	Ensemble of 3 structures	[4mdd-4udd-6dxk-]	0,64	0,57	0,61	0,00
PLANTS	Single structure	[4mdd-]	0,60	0,51	0,56	0,03
	Ensemble of 2 structures	[4mdd-5nfp-]	0,62	0,54	0,58	0,03
	Ensemble of 3 structures	[4mdd-6dxk-4p6x-]	0,63	0,55	0,59	0,00
Surflex-dock	Single structure	[5nft]	0,532	0,426	0,487	0,035
	Ensemble of 2 structures	[5nft-4udc]	0,583	0,466	0,532	0,032
	Ensemble of 3 structures	[5g3j-5nft-4udc]	0,596	0,482	0,555	0,000

Annexe 6 : Détails des performances de docking pour PPARγ

Scoring function	Approach	Best PDB	AUC	Min	Mean	Std
smina_dkoes	Single structure	[2zk5-]	0,718	0,420	0,703	0,047
	Ensemble of 2 structures	[4prg-3hod-]	0,719	0,703	0,712	0,003
	Ensemble of 3 structures	[4prg-3hod-2zk2-]	0,720	0,703	0,712	0,000
smina_vina	Single structure	[2prg-]	0,733	0,487	0,696	0,037
	Ensemble of 2 structures	[2prg-3adw-]	0,738	0,659	0,707	0,009
	Ensemble of 3 structures	[1fm6-2prg-3adw-]	0,740	0,679	0,709	0,000
smina_vinardo	Single structure	[2prg-]	0,738	0,497	0,691	0,040
	Ensemble of 2 structures	[1fm6-2prg-]	0,739	0,641	0,707	0,016
	Ensemble of 3 structures	[4e4q-2i4z-2prg-]	0,740	0,652	0,712	0,000
smina_ad4	Single structure	[2prg-]	0,727	0,434	0,708	0,046
	Ensemble of 2 structures	[2zk2-2zk6-]	0,729	0,685	0,719	0,005
	Ensemble of 3 structures	[4jaz-2ath-3cdp-]	0,731	0,704	0,721	0,000
PLANTS	Single structure	[2prg-]	0,742	0,688	0,717	0,013
	Ensemble of 2 structures	[2prg-2zk3-]	0,745	0,688	0,722	0,010
	Ensemble of 3 structures	[2i4j-2prg-2zk3-]	0,746	0,324	0,392	0,039
Surflex-dock	Single structure	[2ath]	0,456	0,323	0,411	0,038

	Ensemble of 2 structures	[2ath-3adw-]	0,502	0,320	0,426	0,000
	Ensemble of 3 structures	[2ath-3hod-3adw]	0,519	0,490	0,426	0,036

Annexe 7 : Détails des performances de docking pour TRα

Scoring function	Approach	Best PDB	AUC	Min	Mean	Std
smina_dkoes	Single structure	[3ilz-]	0,500	0,484	0,491	0,007
	Ensemble of 2 structures	[3jzb-3ilz-]	0,504	0,495	0,498	0,003
	Ensemble of 3 structures	[3jzb-3ilz-3hzf-]	0,504	0,499	0,501	0,000
smina_vina	Single structure	[1nav-]	0,348	0,315	0,333	0,013
	Ensemble of 2 structures	[3ilz-3hzf-]	0,346	0,331	0,340	0,006
	Ensemble of 3 structures	[3jzb-3ilz-3hzf-]	0,345	0,336	0,342	0,000
smina_vinardo	Single structure	[1nav-]	0,358	0,325	0,341	0,013
	Ensemble of 2 structures	[3hzf-1nav-]	0,351	0,338	0,344	0,004
	Ensemble of 3 structures	[3ilz-3hzf-1nav-]	0,346	0,342	0,344	0,000
smina_ad4	Single structure	[1nav-]	0,321	0,271	0,291	0,021
	Ensemble of 2 structures	[3jzb-1nav-]	0,323	0,289	0,307	0,016
	Ensemble of 3 structures	[3jzb-3hzf-1nav-]	0,323	0,299	0,316	0,000
PLANTS	Single structure	[3jzb-]	0,370	0,320	0,339	0,023
	Ensemble of 2 structures	[3jzb-1nav-]	0,348	0,322	0,338	0,010
	Ensemble of 3 structures	[3ilz-3hzf-1nav-]	0,346	0,332	0,341	0,000
Surflex-dock	Single structure	[3hzf]	0,711	0,646	0,676	0,033

	Ensemble of 2 structures	[3hzf-1nav]	0,702	0,682	0,692	0,010
	Ensemble of 3 structures	[3jzb-3hzf-1nav]	0,689	0,689	0,689	0,000

Annexe 8 : Résultats préliminaires de prédiction des 6 modèles sur la base de données EDlistI

DTXSID	AR	ERA	ERB	GR	PPARG	TRA
DTXSID3020596	🔴	🔴	🔴	🔴	🔴	🔴
DTXSID2021284	🔴	🔴	🔴	🔴	🔴	🔴
DTXSID0020153	🔴	🔴	🔴	🔴	🔴	🔴
DTXSID8039241	🔴	🔴	🔴	🔴	🔴	🔴
DTXSID80143040	🟡	🔴	🟡	🔴	🟡	🔴
DTXSID90143258	🔴	🔴	🔴	🔴	🔴	🔴
DTXSID5023958	🔴	🔴	🔴	🔴	🔴	🔴
DTXSID30143474	🟡	🟢	🟢	🟡	🔴	🔴
DTXSID2035069	🔴	🔴	🔴	🔴	🔴	🔴
DTXSID9047962	🟡	🟢	🟢	🟡	🟡	🔴
DTXSID6022422	🔴	🟡	🟢	🔴	🔴	🔴
DTXSID3036654	🔴	🔴	🔴	🔴	🔴	🔴
DTXSID8024101	🟡	🟡	🔴	🟡	🟡	🔴
DTXSID8059269	🟡	🟢	🟢	🟡	🟡	🔴
DTXSID2020717	🟡	🟢	🟢	🔴	🟡	🔴
DTXSID90145656	🔴	🟢	🟢	🔴	🔴	🔴
DTXSID70145831	🔴	🔴	🟡	🔴	🔴	🔴
DTXSID90145833	🔴	🔴	🟡	🔴	🔴	🔴
DTXSID50145834	🔴	🔴	🟡	🔴	🔴	🔴
DTXSID6036885	🔴	🔴	🟡	🔴	🔴	🔴
DTXSID1036886	🔴	🔴	🟡	🔴	🔴	🔴
DTXSID4058601	🟡	🟡	🟡	🟡	🟡	🔴
DTXSID5029055	🟡	🟡	🟡	🟡	🟡	🔴
DTXSID1022508	🟡	🟡	🟡	🟡	🟡	🔴
DTXSID1024835	🔴	🔴	🔴	🔴	🔴	🔴
DTXSID00146270	🟡	🟡	🔴	🟡	🟡	🔴
DTXSID8025591	🔴	🔴	🔴	🔴	🔴	🟡
DTXSID70884490	🟢	🟢	🟢	🟡	🟡	🔴
DTXSID20872100	🔴	🔴	🟡	🔴	🔴	🔴
DTXSID3051543	🟡	🟡	🟡	🔴	🔴	🔴
DTXSID2021868	🔴	🔴	🔴	🔴	🔴	🔴
DTXSID1020566	🔴	🔴	🔴	🔴	🔴	🔴
DTXSID3020203	🔴	🔴	🔴	🔴	🔴	🔴
DTXSID5021883	🔴	🔴	🔴	🔴	🔴	🔴
DTXSID1024122	🔴	🔴	🔴	🔴	🔴	🔴
DTXSID4061464	🔴	🔴	🔴	🔴	🔴	🔴
DTXSID00148017	🟡	🟡	🟡	🟡	🔴	🔴
DTXSID60869478	🔴	🟡	🔴	🟡	🔴	🔴
DTXSID4024800	🔴	🟡	🟡	🔴	🔴	🔴
DTXSID9048354	🔴	🔴	🔴	🔴	🔴	🔴
DTXSID6020802	🔴	🔴	🔴	🔴	🔴	🔴
DTXSID7021360	🔴	🔴	🔴	🔴	🔴	🔴
DTXSID5041851	🔴	🟡	🟡	🟡	🔴	🔴
DTXSID1051563	🔴	🔴	🔴	🔴	🔴	🔴
DTXSID6021909	🔴	🔴	🔴	🔴	🔴	🔴
DTXSID1021328	🔴	🔴	🔴	🔴	🔴	🔴
DTXSID2074116	🔴	🟡	🔴	🔴	🟡	🔴
DTXSID8074445	🔴	🟡	🔴	🔴	🔴	🔴
DTXSID00148895	🔴	🟡	🔴	🔴	🔴	🔴

DTXSID3074446						
DTXSID6073784						
DTXSID60148896						
DTXSID8030138						
DTXSID0021917						
DTXSID60872987						
DTXSID3073052						
DTXSID0020101						
DTXSID5020100						
DTXSID5073793						
DTXSID0026882						
DTXSID3093945						
DTXSID60149742						
DTXSID70675860						
DTXSID50675863						
DTXSID90150003						
DTXSID50862559						
DTXSID5040673						
DTXSID8061557						
DTXSID5036684						
DTXSID90150528						
DTXSID6037728						
DTXSID1021950						
DTXSID50151112						
DTXSID60893753						
DTXSID7021316						
DTXSID9020376						
DTXSID60151527						
DTXSID90151530						
DTXSID50151531						
DTXSID8025094						
DTXSID7040734						
DTXSID1021958						
DTXSID2020682						
DTXSID2020266						
DTXSID7024372						
DTXSID50921594						
DTXSID2047644						
DTXSID40922893						
DTXSID0021963						
DTXSID2020767						
DTXSID5021465						
DTXSID4059548						
DTXSID3024499						
DTXSID2021658						
DTXSID6020351						
DTXSID00154512						
DTXSID60154831						
DTXSID00925204						
DTXSID70333115						

DTXSID5020316						
DTXSID4023884						
DTXSID2021319						
DTXSID1020722						
DTXSID7031531						
DTXSID6073861						
DTXSID4047880						
DTXSID4044533						
DTXSID5022483						
DTXSID5022352						
DTXSID5041805						
DTXSID90156681						
DTXSID3074492						
DTXSID10156784						
DTXSID5031133						
DTXSID3022403						
DTXSID6038875						
DTXSID4044795						
DTXSID50156925						
DTXSID7021156						
DTXSID60157529						
DTXSID2021446						
DTXSID7038319						
DTXSID40158115						
DTXSID00158116						
DTXSID1027394						
DTXSID5032076						
DTXSID00873771						
DTXSID0065482						
DTXSID7024291						
DTXSID7065548						
DTXSID0021462						
DTXSID1038878						
DTXSID2021999						
DTXSID8020701						
DTXSID90161811						
DTXSID6024127						
DTXSID40349642						
DTXSID301028876						
DTXSID00858720						
DTXSID60163004						
DTXSID3022877						
DTXSID9047881						
DTXSID8044836						
DTXSID0051732						
DTXSID00879854						
DTXSID20164178						
DTXSID9044164						
DTXSID90164347						
DTXSID7074539						

DTXSID30599058						
DTXSID00164434						
DTXSID20164436						
DTXSID00164499						
DTXSID20164517						
DTXSID80164518						
DTXSID00349320						
DTXSID90864552						
DTXSID70164663						
DTXSID60164874						
DTXSID5074551						
DTXSID70165175						
DTXSID9051743						
DTXSID8052853						
DTXSID60477017						
DTXSID5051444						
DTXSID20935648						
DTXSID3044594						
DTXSID10166368						
DTXSID70166369						
DTXSID80166370						
DTXSID40166371						
DTXSID00166372						
DTXSID5041019						
DTXSID5026207						
DTXSID0051782						
DTXSID3022160						
DTXSID8022161						
DTXSID70937798						
DTXSID30431262						
DTXSID60872583						
DTXSID1061942						
DTXSID20170012						
DTXSID70170784						
DTXSID30170967						
DTXSID6051809						
DTXSID1034397						
DTXSID4052685						
DTXSID6061999						
DTXSID40864820						
DTXSID0041775						
DTXSID701029290						
DTXSID30696536						
DTXSID0042080						
DTXSID9059751						
DTXSID4059752						
DTXSID40616285						
DTXSID50879965						
DTXSID70879861						
DTXSID30873481						

DTXSID60879895						
DTXSID4052689						
DTXSID50873928						
DTXSID60583561						
DTXSID80872267						
DTXSID30172360						
DTXSID80940470						
DTXSID40940471						
DTXSID6066442						
DTXSID1066443						
DTXSID9059753						
DTXSID70172622						
DTXSID3052690						
DTXSID3023764						
DTXSID60172792						
DTXSID20172793						
DTXSID80172875						
DTXSID7020184						
DTXSID1073282						
DTXSID3047893						
DTXSID4075459						
DTXSID40173262						
DTXSID4047753						
DTXSID7040287						
DTXSID201034434						
DTXSID8062105						
DTXSID40942430						
DTXSID3032626						
DTXSID8073130						
DTXSID3024811						
DTXSID3022453						
DTXSID9058600						
DTXSID1062122						
DTXSID00174477						
DTXSID8052691						
DTXSID0062139						
DTXSID9062140						
DTXSID4047541						
DTXSID6040298						
DTXSID3040300						
DTXSID5044869						
DTXSID50864952						
DTXSID8052693						
DTXSID1041683						
DTXSID8029864						
DTXSID8073895						
DTXSID60304394						
DTXSID9049245						
DTXSID3040279						
DTXSID0058693						

DTXSID80176196						
DTXSID3032129						
DTXSID6025800						
DTXSID8074811						
DTXSID9059759						
DTXSID3059761						
DTXSID4075372						
DTXSID50177251						
DTXSID3074288						
DTXSID50870492						
DTXSID3075044						
DTXSID1058680						
DTXSID0025654						
DTXSID3062366						
DTXSID80946694						
DTXSID8075205						
DTXSID8058031						
DTXSID7020186						
DTXSID1022477						
DTXSID5075442						
DTXSID0022513						
DTXSID90873927						
DTXSID2047569						
DTXSID8041745						
DTXSID30179617						
DTXSID60179685						
DTXSID7020215						
DTXSID401017149						
DTXSID3021857						
DTXSID7075214						
DTXSID5025992						
DTXSID2062535						
DTXSID1062542						
DTXSID90180587						
DTXSID8059766						
DTXSID7074826						
DTXSID20180919						
DTXSID0041806						
DTXSID5041308						
DTXSID8027903						
DTXSID60274048						
DTXSID8022458						
DTXSID9047960						
DTXSID4025082						
DTXSID2034259						
DTXSID6062599						
DTXSID00872587						
DTXSID80872585						
DTXSID7074822						
DTXSID70472472						

DTXSID70181976						
DTXSID6034554						
DTXSID80885397						
DTXSID80182217						
DTXSID5022354						
DTXSID0022482						
DTXSID7073141						
DTXSID4022521						
DTXSID0040549						
DTXSID0067428						
DTXSID2021074						
DTXSID80904259						
DTXSID90183917						
DTXSID50183938						
DTXSID2034885						
DTXSID5062766						
DTXSID50184187						
DTXSID8031861						
DTXSID7074030						
DTXSID90184803						
DTXSID801029300						
DTXSID501029301						
DTXSID5047956						
DTXSID4062850						
DTXSID4032116						
DTXSID2020684						
DTXSID80863135						
DTXSID2024246						
DTXSID3038309						
DTXSID5074133						
DTXSID5022514						
DTXSID8038306						
DTXSID3024786						
DTXSID2038314						
DTXSID6058055						
DTXSID7038313						
DTXSID8024947						
DTXSID5074135						
DTXSID0074136						
DTXSID50335975						
DTXSID8047399						
DTXSID10879865						
DTXSID7074878						
DTXSID3040302						
DTXSID50955657						
DTXSID1042077						
DTXSID90858719						
DTXSID4022313						
DTXSID50188145						
DTXSID8042349						

DTXSID7073480						
DTXSID5074137						
DTXSID0022511						
DTXSID2032180						
DTXSID8038300						
DTXSID6038299						
DTXSID2073481						
DTXSID7047566						
DTXSID7040150						
DTXSID7073482						
DTXSID3038305						
DTXSID4030045						
DTXSID20189306						
DTXSID10873929						
DTXSID70367798						
DTXSID5058068						
DTXSID8022402						
DTXSID80873557						
DTXSID6063143						
DTXSID4059916						
DTXSID30190948						
DTXSID5030030						
DTXSID90190949						
DTXSID8059920						
DTXSID6038324						
DTXSID6073491						
DTXSID8038304						
DTXSID401029322						
DTXSID20958973						
DTXSID3038301						
DTXSID3038307						
DTXSID1073498						
DTXSID9074141						
DTXSID0052706						
DTXSID6073499						
DTXSID9074143						
DTXSID4074778						
DTXSID40865913						
DTXSID4052059						
DTXSID9047885						
DTXSID70872019						
DTXSID8052067						
DTXSID50192489						
DTXSID6074041						
DTXSID3028005						
DTXSID50865964						
DTXSID4074146						
DTXSID70192805						
DTXSID5052709						
DTXSID2020896						

DTXSID7052078						
DTXSID8032627						
DTXSID40193960						
DTXSID9042049						
DTXSID3029364						
DTXSID4052710						
DTXSID4047886						
DTXSID70872296						
DTXSID8073508						
DTXSID6041688						
DTXSID90961893						
DTXSID7037478						
DTXSID3052858						
DTXSID4047961						
DTXSID3022532						
DTXSID1022429						
DTXSID1040320						
DTXSID80345027						
DTXSID5044944						
DTXSID60963162						
DTXSID40879898						
DTXSID5020861						
DTXSID3034799						
DTXSID30883388						
DTXSID6022399						
DTXSID8022450						
DTXSID10197379						
DTXSID1022392						
DTXSID6024832						
DTXSID6041684						
DTXSID7058701						
DTXSID9022318						
DTXSID0022220						
DTXSID4074988						
DTXSID7020637						
DTXSID2020139						
DTXSID2040236						
DTXSID2060123						
DTXSID8058110						
DTXSID10198604						
DTXSID6073524						
DTXSID30872030						
DTXSID8041583						
DTXSID40398771						
DTXSID0060147						
DTXSID0060149						
DTXSID6038326						
DTXSID0032314						
DTXSID3052147						
DTXSID8058112						

DTXSID101017940						
DTXSID5025231						
DTXSID30199855						
DTXSID50904160						
DTXSID1033325						
DTXSID4022319						
DTXSID00617379						
DTXSID20872584						
DTXSID5073535						
DTXSID0073536						
DTXSID2074160						
DTXSID7074161						
DTXSID5052832						
DTXSID7074163						
DTXSID0060191						
DTXSID60870589						
DTXSID0022309						
DTXSID3040930						
DTXSID7041469						
DTXSID70201623						
DTXSID60201773						
DTXSID1042360						
DTXSID60968394						
DTXSID60333083						
DTXSID9063821						
DTXSID0022351						
DTXSID30202176						
DTXSID90202177						
DTXSID10202179						
DTXSID1022502						
DTXSID30871129						
DTXSID3030056						
DTXSID1025302						
DTXSID5044495						
DTXSID0021331						
DTXSID5028039						
DTXSID3073557						
DTXSID0074180						
DTXSID6037512						
DTXSID8020250						
DTXSID8022537						
DTXSID0020862						
DTXSID8047973						
DTXSID0074184						
DTXSID70205388						
DTXSID80205415						
DTXSID40205416						
DTXSID1025512						
DTXSID7043823						
DTXSID6021290						

DTXSID0034776						
DTXSID10205726						
DTXSID7030066						
DTXSID90205749						
DTXSID60205751						
DTXSID40205754						
DTXSID7052234						
DTXSID2069155						
DTXSID4022816						
DTXSID7041912						
DTXSID0029765						
DTXSID3032179						
DTXSID7025001						
DTXSID60874027						
DTXSID20206401						
DTXSID90206493						
DTXSID40206726						
DTXSID2022462						
DTXSID2060387						
DTXSID7060417						
DTXSID4074067						
DTXSID60207533						
DTXSID20207534						
DTXSID60207538						
DTXSID10871440						
DTXSID70858838						
DTXSID5021174						
DTXSID6022395						
DTXSID00208589						
DTXSID6022393						
DTXSID10208772						
DTXSID3041039						
DTXSID2022387						
DTXSID70209144						
DTXSID2040650						
DTXSID4049327						
DTXSID8030423						
DTXSID7022461						
DTXSID0025571						
DTXSID9022079						
DTXSID7024247						
DTXSID3052276						
DTXSID20209997						
DTXSID1022425						
DTXSID3047974						
DTXSID1052290						
DTXSID2037508						
DTXSID9025299						
DTXSID9022445						
DTXSID1031591						

DTXSID40212266						
DTXSID6038320						
DTXSID60978973						
DTXSID1023815						
DTXSID2042220						
DTXSID2022460						
DTXSID1033325						
DTXSID3044253						
DTXSID3064302						
DTXSID0058148						
DTXSID70981711						
DTXSID9020584						
DTXSID90214336						
DTXSID50214337						
DTXSID5058149						
DTXSID1041926						
DTXSID30496952						
DTXSID7027629						
DTXSID801031721						
DTXSID5022487						
DTXSID7038317						
DTXSID9058238						
DTXSID4064397						
DTXSID50867160						
DTXSID7022465						
DTXSID80216167						
DTXSID2021604						
DTXSID3042261						
DTXSID8042260						
DTXSID5022512						
DTXSID40216800						
DTXSID60216802						
DTXSID20216803						
DTXSID8022327						
DTXSID8021482						
DTXSID1020306						
DTXSID6026294						
DTXSID7075262						
DTXSID00893171						
DTXSID90897481						
DTXSID00217893						
DTXSID5022350						
DTXSID10217965						
DTXSID70217966						
DTXSID6073605						
DTXSID90218051						
DTXSID30218116						
DTXSID20871456						
DTXSID8029282						
DTXSID6023997						

DTXSID9052393						
DTXSID80218409						
DTXSID60887314						
DTXSID4030047						
DTXSID7024110						
DTXSID00219716						
DTXSID10219984						
DTXSID8023426						
DTXSID2038310						
DTXSID0022301						
DTXSID9044829						
DTXSID6074207						
DTXSID6074209						
DTXSID30220936						
DTXSID6029915						
DTXSID20221215						
DTXSID3039242						
DTXSID3058167						
DTXSID1046055						
DTXSID70992569						
DTXSID80222127						
DTXSID301029331						
DTXSID50222275						
DTXSID70222277						
DTXSID5045960						
DTXSID90873487						
DTXSID3020253						
DTXSID90868151						
DTXSID8042478						
DTXSID0064624						
DTXSID7058173						
DTXSID5074216						
DTXSID9020538						
DTXSID30474634						
DTXSID9074226						
DTXSID8074233						
DTXSID6029044						
DTXSID0036633						
DTXSID40225952						
DTXSID5039224						
DTXSID0020600						
DTXSID5021207						
DTXSID2058178						
DTXSID0022484						
DTXSID60997588						
DTXSID10227448						
DTXSID8040727						
DTXSID6024961						
DTXSID80507683						
DTXSID4074770						

DTXSID3047003						
DTXSID1029704						
DTXSID30999059						
DTXSID4024515						
DTXSID2020761						
DTXSID3021516						
DTXSID3058826						
DTXSID60229856						
DTXSID80229939						
DTXSID2020844						
DTXSID7021368						
DTXSID8034873						
DTXSID6025690						
DTXSID1021405						
DTXSID00230845						
DTXSID20231516						
DTXSID201002392						
DTXSID00232026						
DTXSID2022127						
DTXSID8022404						
DTXSID50232536						
DTXSID10232537						
DTXSID30871467						
DTXSID80232544						
DTXSID40232545						
DTXSID10232552						
DTXSID70232558						
DTXSID20232563						
DTXSID00232566						
DTXSID9039789						
DTXSID3058903						
DTXSID6036546						
DTXSID3021778						
DTXSID9047966						
DTXSID10232936						
DTXSID4047963						
DTXSID8029602						
DTXSID7023938						
DTXSID8041408						
DTXSID101029361						
DTXSID301029367						
DTXSID80235338						
DTXSID50235340						
DTXSID001029370						
DTXSID0052592						
DTXSID2025004						
DTXSID6025983						
DTXSID8058958						
DTXSID5037527						
DTXSID3041582						

DTXSID2042359						
DTXSID60237451						
DTXSID20237452						
DTXSID80237453						
DTXSID40237454						
DTXSID60237456						
DTXSID20237457						
DTXSID2041547						
DTXSID2024791						
DTXSID1049643						
DTXSID8023971						
DTXSID2059069						
DTXSID1029120						
DTXSID8025701						
DTXSID0021387						
DTXSID5026209						
DTXSID2026101						
DTXSID70920504						
DTXSID7022417						
DTXSID8022531						
DTXSID5059117						
DTXSID8024498						
DTXSID7025005						
DTXSID5037571						
DTXSID0024183						
DTXSID4021341						
DTXSID0058227						
DTXSID8040274						
DTXSID5024687						
DTXSID0040703						
DTXSID1021827						
DTXSID4073889						
DTXSID3025203						
DTXSID6020303						

le cnam

Asma SELLAMI

Utilisation des méthodes de criblage virtuel dans un contexte de santé humaine et environnementale : application aux récepteurs nucléaires et aux perturbateurs endocriniens

Résumé

Les perturbateurs endocriniens (PE) constituent un problème de santé publique. En effet, l'exposition humaine à ces composés est associée à un risque accru de développement de plusieurs pathologies. Les PE sont capables de pénétrer dans l'organisme et d'interférer avec les fonctions du système endocrinien par divers mécanismes, dont la liaison directe aux récepteurs nucléaires (NR). La détection précoce de potentiels PE est donc nécessaire pour garantir la sécurité de l'Homme et de l'environnement. Ceci est possible en utilisant des tests expérimentaux sur les composés suspects, mais cela reste une tâche difficile notamment en raison du nombre considérable de composés à évaluer. Ainsi, les méthodes *in silico* peuvent être utilisées en amont de ces derniers pour prioriser les composés à tester expérimentalement. Avec cet objectif, nous avons utilisé des méthodes basées sur la structure de la cible (SB) et des méthodes basées sur les ligands (LB) pour prédire la liaison de composés aux NR. Nous avons établi une preuve de concept sur le récepteur ER alpha, le plus étudié des NR. Nous avons ensuite proposé un protocole combinant des modèles de docking et de pharmacophores pour prédire les potentiels PE en se basant sur leur capacité à se lier à six récepteurs nucléaires : AR, ERα, ERβ, GR, PPARγ et TRα. Les modèles développés permettent ainsi de catégoriser la probabilité de composés requêtes de se lier aux NR et donc, en fonction du mécanisme direct, d'être des PE.

Mots clés : In silico, criblage virtuel, docking, pharmacophore, perturbateurs endocriniens, récepteurs nucléaires

Résumé en anglais

Endocrine disrupting chemicals (EDCs) are considered as a public health threat as human exposure to these compounds have been associated with increased risk of several diseases. EDCs are able to penetrate the body and to interfere with the functions of the endocrine system through various mechanisms including a direct binding to nuclear receptors (NR). Early detection of potential EDCs becomes an imperative to prevent human and environmental safety issues. This can be achieved using experimental tests, but it remains a challenging task in particular because of the considerable number of compounds to be evaluated. *In silico* methods can then be used in complement, to prioritize compounds for experimental testing and to help elucidating toxicity mechanisms. In this work, we used structure-based (SB) and ligand-based (LB) *in silico* methods to predict compounds binding to NR. We established a first proof of concept on ER alpha, the most studied NR. We then proposed a pipeline combining docking and pharmacophore models to predict potential EDCs based on their ability to bind six nuclear receptors: AR, ERα, ERβ, GR, PPARγ and TRα. The pipeline output enables to categorize query compounds according to their probability of being NR binders and thus, accordingly to the direct mechanism, potential EDCs

Keywords : In silico, virtual screening, docking, Pharmacophores, endocrine disrupting chemicals