Importance des données inactives dans les modèles : application aux méthodes de criblage virtuel en santé humaine et environnementale

Manon Réau

le cnam

École doctorale Sciences des Métiers de l'Ingénieur

Génomique Bioinformatique et Chimie Moléculaire

THÈSE présentée par :

Manon REAU

Soutenue le : **29 octobre 2019**

pour obtenir le grade de : **Docteur du Conservatoire National des Arts et Métiers**

Discipline : Biochimie et biologie moléculaire / Spécialité : Bioinformatique

Importance des données inactives dans les modèles

Application aux méthodes de criblage virtuel en santé humaine et environnementale

THÈSE dirigée par :

M. MONTES Matthieu	Professeur des Universités, Cnam
M. ZAGURY Jean-François	Professeur du Cnam, Titulaire de chaire, Cnam

RAPPORTEURS :

Mme. AUDOUZE Karine	Maître de conférences, université Paris Descartes
M. LANGER Thierry	Professeur des Universités, Université de Vienne

---

JURY :

Mme. SOPKOVA Jana	Professeur des Universités, Université de Caen
M. BAADEN Marc	PhD, Directeur de laboratoire, CNRS
M. PORT Marc	Professeur du Cnam, Titulaire de chaire, Cnam

Rien n’est à craindre, tout est à comprendre.

Marie Curie

Remerciements

J'adresse toute ma reconnaissance et mes remerciements au Professeur Jean-François Zagury pour m'avoir fait confiance et accueillie dans son laboratoire. Merci infiniment de m'avoir permis de réaliser mon rêve en m'ouvrant les portes de la recherche.

Un grand merci au Professeur Matthieu Montes qui a dirigé mes travaux de thèse avec bienveillance. Je le remercie pour l'autonomie qu'il m'a accordée et pour les multiples discussions scientifiques qui m'ont permis d'avancer et de grandir en tant que chercheuse. Je n'oublie pas les discussions extra-scientifiques qui sont certainement l'élément fondateur de l'esprit d'équipe qui règne au laboratoire.

Je remercie profondément le Professeur Thierry Langer et le Docteur Karine Adouze d'avoir accepté de juger mon travail de thèse en tant que rapporteurs. Je remercie également le Professeur Jana Sopovka, le Professeur Marc Port et le Docteur Marc Baaden de m'avoir fait l'honneur de participer au jury de cette thèse.

Je remercie toutes les personnes qui ont jalonnées mon chemin et m'ont donné le goût de la recherche scientifique. Tout d'abord, l'équipe pédagogique et administrative de SupBiotech Paris qui m'ont permi de suivre mon cursus d'ingénierie des biotechnologies tout en m'orientant progressivement vers la bio/chemoinformatique au travers de stage et d'un double diplôme à l'Université Paris Diderot. Je remercie le Docteur Jean-Yves Trosset, qui m'a fait confiance et a encadré mon premier stage de chémoinformatique au sein du laboratoire BIRL alors que je ne savais pas ouvrir un terminal. Merci au Professeur Amedeo Caflisch, au Docteur Marc Baaden et au Professeur Anne Claude Camproux de m'avoir à leur tour donné ma chance dans leur laboratoire respectif et de m'avoir inculqué les valeurs de la recherche académique.

Je souhaite remercier le Docteur Anne Badel, qui m'a permis d'effectuer du monitorat à l'Université Paris Diderot en parallèle à mes travaux de thèse et de mettre un premier pied dans l'enseignement. Je remercie particulièrement le Docteur Gautier Moroy avec qui j'ai réalisé la plus grande partie de mon monitorat et qui a été mon mentor en pédagogie. Merci de m'avoir fait confiance, de m'avoir toujours écoutée, aidée et conseillée.

Je tiens à remercier l'intégralité du laboratoire GBCM et assimilés. En particulier : mes JoMaCh, Chloé et Josselin, comme qui dirait « les rayons de mon soleil », mes compères d'arts en tout genre, la brigade du pinceau ! Merci pour votre bonne humeur contagieuse et votre créativité débordante ! ; Nathalie, pour ses heures consacrées à relire et corriger mon manuscrit, ses debriefs football, sa passion Escape Room, pour sa générosité et sa patience ; Florent, mon co-bureau, pour les debriefs de (vrai/faux) rugby et pour me rappeler autant que faire se peut qu'une thèse c'est avant tout « l'apprentissage de la patience »; Sigrid, pour vivre à 200% les Escape Games, tout comme les verres au Léonard, pour ta franchise et ton

ouverture d'esprit ; Taoufik, notre athlète en herbe, 3 participations à l'Ekiden à son inactif, premier sur radio planète, merci pour ta générosité et ton aide technique ; Cédric, le numberphile du GBCM, pour tous les débats éclairés, et pour m'avoir mis un chapron rouge alors même que j'étais dernière de la ligue ! ; Machat, pour les histoires rocambolesques des « ses amis » ; Benjamin et Jérémy, pour leur calme absolu, leur générosité, et leurs conseils toujours précieux; Asma, pour sa bonne humeur et son esprit d'équipe ; Hervé, pour son répertoire musical, de Schubert à Kerry James ; Maxime, pour apporter (furtivement !) un peu de Poitou dans ce laboratoire ; Vincent dit « Vincent des Maths », pour ses vannes/uppercuts habilement placées ; Sophia et Christiane, pour leur gentillesse, leur efficacité et leur bienveillance ; et enfin ceux que j'ai croisé quelques mois ou années : Marc, consommateur numéro 1 de tablettes Lindt, et Vincent dit « Vincent V », grâce à qui je n'ai loupé aucun tube de PNL. Je remercie également l'équipe de Chimie Moléculaire, et en particulier Aïda, le trait d'union de nos équipes. J'en arrive à nos collègues et amis de Cochin : Hadley, Lucille, Barbara et Gaby, merci pour tous ces bons moments partagés à Ermenonville, Chissay, lors d'Escape Games et bien sûr, au Léonard. Je terminerai par les stagiaires que j'ai eu l'occasion d'encadrer, Narjes et Léon, et les autres stagiaires qui ont croisé mon chemin, Vincent, Célia, Myriam et Nicolas, je leur souhaite une belle carrière.

Je remercie tous mes amis qui m'ont apporté du soutien tout au long de cette thèse. Mes fidèles, mes compagnons de voyage qui me supportent depuis des années et me soutiennent dans « ma quête d'inactivité » : Béné, Maud, Audrey, Orane et Tangui. Merci à tous les taggers du PTRC, grâce à qui j'ai pu me défouler chaque lundi soir sur le pré ; remerciements tous particuliers à Arnault, Flavien, JB, Julien (x2), Laure et Thomas pour avoir rendu l'aventure PTRC encore plus belle. Enfin, mes plus profonds remerciements à Pascal, pour les pains perdus qui ont égayé la rédaction de ce manuscrit, les injections de soleil Lotois qui fonctionnent mieux qu'une cure d'aubépine, pour le soutien quotidien et la confiance qu'il m'apporte.

Je termine par les plus importants, ma famille. Merci à mes parents qui m'ont sans cesse soutenue et encouragée à aller de l'avant. Merci de m'avoir fait confiance et de m'avoir laissée très tôt voler de mes propres ailes. Merci à mes frères et sœurs, Julie, Aurélien et Bastien, qui sont bien souvent trop loin de moi, mais qui m'apportent réconfort et fierté chaque jour.

Résumé

Le criblage virtuel est utilisé dans la recherche de médicaments et la construction de modèle de prédiction de toxicité. L'application d'un protocole de criblage est précédée par une étape d'évaluation sur une banque de données de référence. La composition des banques d'évaluation est un point critique ; celles-ci opposent généralement des molécules actives à des molécules supposées inactives, faute de publication des données d'inactivité. Les molécules inactives sont néanmoins porteuses d'information. Nous avons donc créé la banque NR-DBIND composée uniquement de molécules actives et inactives expérimentalement validées et dédiées aux récepteurs nucléaires. L'exploitation de la NR-DBIND nous a permis d'étudier l'importance des molécules inactives dans l'évaluation de modèles de docking et dans la construction de modèles de pharmacophores. L'application de protocoles de criblage a permis d'élucider des modes de liaison potentiels de petites molécules sur FXR et NRP-1.

Mots clés : Molécules inactives, modèles, criblage virtuel, docking, pharmacophores, récepteurs nucléaires, banque de données, benchmark, FXR, NRP-1**.**

Résumé en anglais

Virtual screening is widely used in early stages of drug discovery and to build toxicity prediction models. Commonly used protocols include an evaluation of the performances of different tools on benchmarking databases before applying them for prospective studies. The content of benchmarking tools is a critical point; most benchmarking databases oppose active data to putative inactive due to the scarcity of published inactive data in the literature. Nonetheless, experimentally validated inactive data also bring information. Therefore, we constructed the NR-DBIND, a database dedicated to nuclear receptors that contains solely experimentally validated active and inactive data. The importance of the integration of inactive data in docking and pharmacophore models construction was evaluated using the NR-DBIND data. Virtual screening protocols were used to resolve the potential binding mode of small molecules on FXR and NRP-1.

Keywords : Inactive molecules, models, virtual screening, docking, pharmacophores, nuclear receptors, database, benchmarking, FXR, NRP-1**.**

Table des matières

Liste des tableaux

Liste des figures

Liste des équations

Liste des annexes

Liste des abréviations

1D/2D/3D	1-Dimension/2-Dimensions/3-Dimensions
AAV2	*Adeno-associated virus serotype 2 variant*
ACD	*Advanced Chemical Directory*
ADME	Administration/Distribution/Métabolisme/Dégradation
ADN	Acide DésoxyriboNucléique
Ag	Agoniste(s)
AMM	Autorisation de Mise sur le Marché
ANN	*Artificial Neurone Network* - Réseau de neurones artificiel
ANSM	Agence nationale de sécurité du médicament et des produits de santé
Ant	Antagoniste(s)
AR	Récepteur aux androgènes
ARN	Acide RiboNucléique
AUC	*Area Under the Curve* - Aire sous la courbe ROC
BEDROC	*Boltzmann-Enhanced Discrimination of ROC*
BM	Bemis Murcko
CADD	*Computer Aided Drug Design* - Conception de médicament assistée par ordinateur
CAR	*Constitutive androgen receptor*
CATS	*Chemically Advanced Template Search*
ClogP	Coefficient de partition octanol/eau
CNN	*Common-Nearest Neighbor algorithm* - Algorithme du plus proche voisin commun
CoMFA	*Comparative Molecular Field Analysis*
CoRNR	CoRepresseur des Récepteurs Nucléaires
COUP-TFI/II	*Chicken ovalbumin upstream promoter transcription) factor I/II*
CRISPR	*Clustered Regularly Interspaced Short Palindromic Repeat* - Courtes répétirions palindromiques groupes et régulièrement espacées
CSD	Cambridge Structural Database
DBD	*DNA-Binding Domain* - Site de liaison de l'ADN
DFT	*Density Functional Theory* - Théorie de la fonctionelle de la densité
DG	*Distance Geometry*
DL	Dose léthale
DR	Séquences Directes Répétées
DT	*Decision Tree* - Arbres de décision
DXR	1-deoxy-D-xylulose-5-phosphate reductoisomerase
EBI	Institut Européen de Bioinformatique
EC50	Concentration Efficace médiane
ECFP4	*Extended Connectivity Fingerprint 4*
EF	Facteur d'Enrichissement
EM	Microscopie Électronique
EPA	*Environmental Protection Agency (U.S.)*

ER α/β/γ	Récepteurs aux estrogènes α/β/γ
ErG	Extended reduced Graph
ERR α/β/γ	*Estrogen related receptors α/β/γ*
FDA	*Food and Drug Administration (U.S.)*
FP	Polarisation de fluorescence
FRET	Förster resonance energy transfer
FXR	*Farnersoid X receptor*
GA	Algorithme Génétique
GB	Born Généralisé
GCNF	*Germ cell nuclear factor*
GPCR	Récepteurs couplés aux protéines G
GR	Récepteur aux glucocorticoïdes
GFP	*Green Fluorescent Protein* - Protéine fluorescente verte
HAMW	*Heavy Atom Moolecular Weigth* - Poids moléculaire des atomes lourds
HBA	*H-bond acceptor* - Accepteur de liaison hydrogène
HBD	*H-bond donor* - Donneur de liaison hydrogène
HDAC	Histone Désacétylase
hERG	*Human Ether-à-go-go Related Gene*
HRE	Elément de Eéponse aux Hormones
HSP	*Heat-shock protein* - Protéine de choc thermique
HTS	*High throughput screening* - Criblage à haut débit
IC50	Concentration Inhibitrice médiane
IDSS	*Inner Distance Shape Signature*
IP	Séquences de Palindromes Inversés
IR	Séquences Répétées Inversées
IUPAC	Union Internationale de Chimie Pure et Appliquée
Kd	Constante de dissociation
Ki	Constante d'inhibition
kNN	k plus proches voisins
LADS	*Latent Actives in the Decoys Set*
LBD	*Ligand Binding Domain* - Site de liaison du ligand
LIE	*Linear Interaction Energy*
LOO	*Leave One Out*
LR	Régression Linéaire
LRA	*Linear Response Approximation*
LXR α/β	Récepteurs des oxystérols α/β
MAE	Erreur absolue moyenne
MC	Monte Carlo
MCS	*Maximum Common Substructures* - Sous-structures communes maximales
MD	Dynamique moléculaire
MM-GBSA	*Molecular Mechanichs - Generalized Born (GB) model augmented with the hydrophobic solvent accessible surface area (SA)*
MM-PBSA	*Molecular Mechanichs - Poisson Boltzmann (GB) model augmented with the hydrophobic solvent accessible surface area (SA)*

MR	Récepteur aux minéralocorticoïdes
MW	*Molecular Weight* - poids moléculaire
NB	*Non-binder(s)* - Qui n'a pas ou peu d'affinité pour la cible étudée
NGFI-B	*Nerve growth factor IB*
NIH	*National Institut of Health (U.S.)*
NOAEL	Dose maximale sans effet néfaste observable
NOEL	Dose sans effet observable
NR	Récepteur Nucléaire
NRP-1	Neuropiline 1
OMS	Organisation Mondiale de la Santé
P	Séquences Palindromes
PAINS	*Pan-assay interference compounds*
PB	Poisson Botlzmann
PCBs	Polychlorobiphényles
PCR	Régression des Composantes Principales
PDB	*Protein Data Bank*
pH	Potentiel hydrogène
pKa	Constante d'acidité
PLS	Régression des moindres carrés partiels
PNR	*Photoreceptor-specific nuclear receptor*
PPAR α/β/γ	Récepteurs activés par les proliférateurs de peroxysomes α/β/γ
PR	Récepteur à la progestérone
PXR	*pregnane X receptor*
QSAR	Relations quantitative structure à activité
RAR α/β/γ	Récepteurs à l'acide rétinoïque α/β/γ
RCP	Résumé des Caractéristiques du Produit
REACH	*Registration, Evaluation, Authorization and Restriction of Chemicals*
RIE	*Robust Initial Enhancement*
RMN	Résonance Magnétique Nucléaire
RMSD	*Root Mean Square Deviation* - Ecart quadratique moyen des distances entre atomes
RMSE	*Root Mean Square Error* - Erreur quadratique moyenne
Ro3	*Rule of 3* - Règle de 3
Ro5	*Rule of 5* - Règle de 5 de Lipinski
ROC	*Receiver operating characteristic* - Courbe sensibilité/spécificité
ROCS	*Rapid Overlay of Chemical Structures*
ROR α/β/γ	*Retinoid related orphan receptors α/β/γ*
RP	Partitionnement Récursif
RSR	*Real space R-factor*
RXR α/β/γ	Récepteur X des rétinoïdes α/β/γ
SAR	Relations Structure à Activité
SAXS	*Small Angle X- rays Scattering* - Diffusion des rayons X aux petits angles
SF-1	*Steroidogenic factor 1*

SMILES	*Simplified Molecular Input Line Entry Specification*
SPA	Essai de scintillation par proximité
SPR	Résonance des plasmons de surface
SVM	Machines à Vecteur de Support
TBT	Tributyl étain
TGFβ	*Transforming growth factor β - Facteur de croissance transformant β*
TNFα	*Tumor Necrosis Factor α*
TPSA	Aire de la surface polaire
TPT	Dibutyl étain
TR α/β	Récepteurs des hormones thyroïdiennes α/β
TR2/4	Récepteurs testiculaires 2/4
USR	Utrafast Shape Recognition
VDR	Récepteur de la vitamine D
WDI	*World Drug Index*

Vue d’ensemble du travail de thèse

Le criblage virtuel est utilisé aussi bien dans un objectif de recherche de médicament que pour anticiper des effets néfastes de petites molécules chimiques (effets secondaires, toxicité etc.). Le choix du protocole de criblage dépend principalement des données disponibles. Malgré l'expansion du nombre de méthodes de criblage basées sur le ligand et sur la structure de la protéine, il n'existe aucune méthode qui présente des performances régulières et satisfaisantes sur l'intégralité des systèmes biologiques étudiés. Les études de criblages sont donc très souvent précédées d'une étape de calibrage et de comparaison des performances de différents protocoles sur un jeu de données de référence. Ces études rétrospectives peuvent être menées de deux manières ; 1) le protocole de criblage est testé sur un ensemble varié de cibles protéiques, l'objectif étant de s'assurer de ses performances globales, ou bien 2) le protocole de criblage est testé sur un système proche de celui étudié, l'objectif étant d'identifier l'outil le plus adapté à une étude prospective précise. Dans les deux cas, des banques de données de référence, communément appelées banque de données d'évaluation peuvent être utilisées. Certaines banques de données comme la Database of Useful (Docking) Decoys — Enhanced (DUD-E)[1], la Maximum Unbiased Validation database (MUV)[2], et la Demanding Evaluation Kits for Objective In Silico Screening (DEKOIS)[3] possèdent des données couvrant plusieurs familles protéiques. D'autres banques de données sont dédiées à une seule famille de protéines, comme la GPCR Ligand (GLL)/Decoys Database (GDD)[4], la Nuclear Receptors Ligands and Structures Benchmarking DataBase (NRLiSt BDB)[5], Maximal Unbiased Benchmarking Data sets for Histone Deacetylases (MUBD-HDAC)[6] respectivement consacrées aux récepteurs couplés aux protéines G, aux récepteurs nucléaires et aux histone désacétylases. La composition des banques de données d'évaluation est critique ; la sélection des molécules actives et des molécules inactives va très souvent de pair avec l'introduction de biais qui doit impérativement être prise en compte pour l'interprétation des résultats. Les molécules identifiées comme inactives dans des tests biologiques ne sont que rarement publiées, elles sont donc très souvent substituées partiellement ou complètement par des molécules supposées inactives appelées leurres ou *decoys*. La sélection de ces molécules peut conduire aussi bien à la surestimation qu'à la sous-estimation des performances d'un outil de criblage. La surestimation peut être due à une facilité de discrimination des données actives et des *decoys* liée à un faible recouvrement de l'espace chimique entre les deux jeux de données, ou bien à une complexité globalement plus élevée chez les molécules actives que chez les *decoys*. La sous-estimation peut être liée à la présence de faux négatifs dans le jeu de données. Le risque de surestimation est généralement limité en imposant une similarité physico-chimique entre les molécules actives et les *decoys* sélectionnés, alors que la sous-estimation est limitée en imposant des distances topologiques.

Cette méthode de sélection n'expose pas les outils aux phénomènes de gain ou de perte d'affinité dus à de faibles changements structuraux qui sont observés lors d'études de pharmaco-modulation de petites molécules. Seule l'intégration de molécules expérimentalement validées comme inactives peut pallier ce manque et apporter une information supplémentaire à l'utilisation de *decoys* dans l'évaluation et la calibration de méthode de criblage. Afin d'étudier l'importance de l'utilisation de molécules inactives dans l'évaluation et la construction de modèles de criblage, nous avons choisi de nous focaliser sur la famille des récepteurs nucléaires pour lesquelles de nombreuses données sont disponibles : nous avons (1) étudié l'évolution de la sélection de *decoys* dans les banques de données (Cf Résultats 1), (2) construit et publié une banque de données dédiées aux récepteurs nucléaires contenant des informations d'inactivité expérimentalement validées et extraites de la littérature (Cf Résultats 2), (3) comparé les performances d'outils de criblages virtuels en se basant uniquement sur des données expérimentalement validées et comparé ces performances en remplaçant les molécules inactives par des *decoys* (Cf Résultats 3), et enfin (4) construit des modèles de pharmacophore en tirant profit de l'information apportée par les molécules inactives (Cf Résultats 4). Les résultats de ces études seront détaillés dans ce manuscrit et précédés d'une introduction sur le criblage virtuel organisée en 7 parties. La première partie est une introduction sur l'utilisation du criblage virtuel en recherche médicale et en santé publique. Elle est suivie d'une partie concernant les chimiothèques existantes dans laquelle les étapes de préparation des petites molécules sont présentées, de deux parties faisant l'état de l'art des approches de criblage virtuel basé sur le ligand et basé sur la structure, puis d'une partie concernant les méthodes d'évaluation des outils ou protocoles de criblage virtuel. L'avant dernière partie concerne la famille des récepteurs nucléaires qui a été étudiée dans le cadre de ce projet de thèse, et les enjeux de son étude. Enfin, les objectifs de cette thèse concluent cette introduction. La partie résultat regroupre l'ensemble des études conduites dans le cadre de ce projet de thèse et deux études annexes d'application de protocoles de criblage virtuel sur les protéines FXR et NRP-1 sont présentées.

Introduction

1 Introduction au criblage virtuel en recherche médicale et en santé publique

1.1 Découverte de médicament

1.1.1 Processus général du développement de médicament

De tout temps les Hommes ont usé de procédés variés pour soigner les blessures, les douleurs, les piqûres d'insectes, les infections etc. Les traces de médication les plus anciennes datent de 5000 ans av J.-C. ; 12 recettes de remèdes faisant appel à 250 plantes différentes ont été découvertes sur une dalle d'argile Sumérienne à Nagpur, au cœur de l'Inde[7]. Si aujourd'hui la plupart des médicaments utilisés sont issus de l'industrie chimique, la nature a très longtemps été la première ressource de molécules thérapeutiques. Les plantes[8,9], les algues[10,11], les micro-organismes[12,13] ainsi que les animaux possèdent une multitude de molécules complexes jouant un rôle dans leur protection et leur survie. Leur utilisation chez l'Homme s'est avérée bénéfique dans de nombreux cas[14]. Dioscorides a énuméré dès 77 av J.-C. un ensemble de

Figure 1 Extrait d'une copie de Materia Medica – schéma de l'Aigremoine vulgaire

944 remèdes préparés à partir de 657 plantes dans un ouvrage de référence appelé « De Materia Medica » utilisé jusqu'à la Renaissance.

Pendant très longtemps, les découvertes des propriétés médicinales ont découlé d'observations directes et de tests sur l'Homme, non sans risques. Aujourd'hui, le développement de médicament, allant de la conception d'une molécule active jusqu'à sa mise sur le marché, est un processus complexe soumis à de nombreuses réglementations. Il se décompose en 6 grandes étapes[15] : l'identification d'une cible thérapeutique, la découverte de molécules actives, les études précliniques, les études cliniques, la validation du médicament par les agences agréés, ainsi que la phase post-approbation de la mise sur le marché.

1.1.1.1 Identification d'une cible thérapeutique

Un produit est qualifié de « médicament » s'il est « présenté comme possédant des propriétés curatives ou préventives à l'égard des maladies humaines ou animales, ou s'il peut être utilisé ou administré en vue d'établir un diagnostic médical » d'après l'Agence nationale de sécurité du médicament et des produits de santé (ANSM). La recherche de médicament doit donc débuter par l'identification d'une maladie pour laquelle le défaut de traitement efficace constitue un réel besoin. Afin d'orienter les choix de traitement envisageables, il est nécessaire d'étudier les mécanismes impliqués dans le développement de la maladie via des études biologiques et biochimiques dans le but de définir une ou plusieurs cibles présentant un intérêt thérapeutique. Une cible est adaptée à l'étude si 1) son activité peut être quantifiée de manière expérimentale ; on peut ainsi mesurer un changement de comportement stimulé par sa mise en contact avec une molécule testée, et si 2) sa pertinence dans la maladie étudiée est prouvée. La modulation de différentes cibles biologiques comme l'ADN, l'ARN ou encore les protéines peut s'avérer intéressante dans un cadre thérapeutique. Dans ce manuscrit, nous nous limiterons aux cibles protéiques. Ces cibles protéiques peuvent être identifiées par des méthodes biochimiques (ex : spectrométrie de masse[16]), biologiques (ex : utilisation d'anticorps, mesure de l'expression des ARN messagers), génétiques (ex : invalidation génique par « *gene knockout* », ARN interférent, CRISPR-Cas9), génomiques (ex : recherche d'associations génétiques), ou encore par déduction informatique (ex : biologie des systèmes). Les méthodes biochimiques, biologiques, génétiques sus-citées permettent également de valider l'intérêt de la cible en s'assurant de l'efficacité de sa modulation sur la non-progression de la maladie. Il faut également s'assurer que la modulation de la cible ne soit pas être délétère pour l'organisme. Un ultime critère non négligeable à la sélection d'une cible est sa « druggabilité » c'est-à-dire sa capacité à interagir avec une molécule thérapeutique. Cette notion reste complexe à évaluer expérimentalement et des méthodes de prédictions virtuelles sont de plus en plus utilisées[17,18].

Dans ce manuscrit, nous limiterons la notion de molécules thérapeutiques aux petites molécules chimiques, excluant les peptides, les protéines recombinantes, les anticorps, les protéines solubles, les oligonucléotides etc.

1.1.1.2 Recherche de « hits »

La seconde étape de la recherche de médicament consiste à identifier une touche, appelée « hit », c'est-à-dire une molécule capable de se lier et de moduler l'activité de la cible étudiée. Des molécules sont généralement qualifiées « hits » lorsqu'elles atteignent une concentration inhibitrice médiane (IC50) de l'ordre de 10 μM[19]. Elles sont généralement identifiées via des méthodes de criblage *in vitro*, comme le criblage à haut débit, des méthodes *in* silico, comme le criblage virtuel (Cf 3 et 4), ou encore par repositionnement, c'est-à-dire en utilisant un médicament existant ayant déjà validé les tests de toxicité lors d'essais précliniques et cliniques sur une cible nouvelle.

1.1.1.3 Génération et optimisation des leads

Les « hits » identifiés font ensuite l'objet de modifications structurales afin d'évoluer vers des molécules présentant une meilleure affinité et sélectivité pour la cible thérapeutique ainsi que de meilleures propriétés ADME/Tox (administration, distribution, métabolisme, élimination et toxicité). Ces molécules sont appelées « têtes de série » ou « leads ». Une molécule est généralement qualifiée de *lead* lorsque son affinité pour la cible est de l'ordre du nM[19]. Pour obtenir ces *leads*, des études de relation structure activité (SAR) sont réalisées au cours desquelles de nouvelles molécules sont dérivées du *hit* initial via la modification de groupements fonctionnels, tout en conservant le squelette du *hit*. L'influence des modifications sur l'affinité pour la cible, la sélectivité et les propriétés ADME/Tox des nouveaux composés est mesurée et permet de guider de manière rationnelle la génération de *lead*. Des méthodes informatiques permettent d'assister l'analyse des études de SAR. Les *leads* obtenus sont ensuite optimisés de manière à améliorer des paramètres nécessaires à l'administration d'un médicament, à savoir la solubilité du composé, sa faisabilité technique, ainsi que sa sélectivité pour la cible d'étude afin de minimiser les potentiels effets indésirables liés à l'interaction avec des cibles secondaires.

1.1.1.4 Études précliniques

L'étude préclinique constitue une étape fondamentale d'évaluation de l'efficacité et de la toxicité de la molécule candidate sur des modèles animaux (*in vitro* et *in vivo*), avant de l'administrer à l'Homme. C'est au cours de cette étape que la faisabilité synthétique à grande échelle des composés, les problèmes liés à leur formulation, leur pharmacologie et leur toxicologie sont évalués. En particulier, le mode d'action du composé est caractérisé grâce à des tests biologiques et biochimiques. Les tests toxicologiques permettent quant à eux d'évaluer les effets secondaires possibles (mutagénèse, cancérogénèse, etc.), la dose létale (DL 50, dose à laquelle 50 % des animaux sont tués), la dose à laquelle la molécule n'a aucun effet (NOEL, no effect level) et la dose à laquelle la molécule n'a aucun effet secondaire (NOAEL, no adverse effect level). Cette étape constitue l'un des prérequis pour déposer un dossier de demande d'autorisation d'essai clinique et pour obtenir une Autorisation de Mise sur le Marché (AMM).

1.1.1.5 Études cliniques

Les études cliniques représentent l'étape la plus sensible du développement d'un médicament ; il s'agit de tester sur l'Homme une molécule jusqu'alors uniquement testée *in vitro* ou *in vivo* sur des modèles animaux. Les études cliniques se divisent en 4 phases consécutives et dépendantes de la réussite de la phase précédente. La phase 1 correspond à l'évaluation de la toxicité et du seuil de tolérance du médicament chez un nombre limité de volontaires sains, excepté pour les anti-cancéreux qui sont testés directement sur des individus malades. La phase 2 correspond à l'étude de l'efficacité et de la relation dose-effet chez un nombre limité de patients malades. La phase 3 s'étend à une échelle plus large, excepté dans le cas de maladies rares ; elle est effectuée sur un panel de patients malades représentatif de la diversité des patients à qui le médicament sera destiné. C'est au cours de ces deux dernières phases que les médicaments sont comparés à des placebos lors de tests en double aveugle lors desquels ni le patient, ni le médecin ne sait quelle molécule est administrée. Cette étape permet d'établir un rapport bénéfice/risque et doit permettre d'estimer l'intérêt de l'indication de la molécule testée en comparaison des traitements de référence. Une demande d'AMM est effectuée auprès de ANSM pour les composés ayant franchi avec succès les 3 étapes des études cliniques. En cas d'autorisation, l'ANSM valide un Résumé des Caractéristiques du Produit (RCP) définissant les conditions d'utilisation du médicament, la notice d'utilisation aux patients et les conditions d'étiquetage. Enfin, la phase 4, dite phase de pharmaco-vigilance a lieu en aval de l'AMM. Elle correspond à un suivi du médicament commercialisé notamment pour surveiller l'efficacité du médicament et la déclaration d'éventuels effets secondaires.
Au total, le développement d'un médicament s'étend sur 6 à 15 ans[20,21] et atteint des coûts astronomiques de l'ordre du milliard d'euros[21]. L'utilisation de méthodes *in silico*, moins coûteuses que les méthodes expérimentales, vise à réduire ces coûts et le temps de développement en intervenant lors de la phase de découverte de molécules actives. Outre leur utilisation pour l'identification de hits, elles servent également à modéliser la cible thérapeutique en cas d'absence de structure expérimentale, à simuler son comportement *in vivo* pour comprendre des mécanismes d'action, ou encore à anticiper la toxicité des molécules ainsi que leurs effets secondaires. Dans ce manuscrit, nous nous intéresserons particulièrement aux méthodes de criblage virtuel.

1.2 Les effets indésirables des petites molécules

Généralement, un médicament a fait l'objet d'études orientées vers une cible thérapeutique impliquée dans une maladie. En réalité, une petite molécule peut interagir avec plusieurs cibles et on estime qu'un médicament possède en moyenne 6 cibles protéiques[22]. Cette pluralité des cibles est responsable d'effets secondaires. Dans certains cas, les effets secondaires peuvent être bénéfiques, et amplifier l'action d'une molécule thérapeutique. Par exemple, la communauté scientifique a récemment commencé à mettre à profit cette non sélectivité des médicaments pour développer des molécules capables d'agir simultanément sur plusieurs cibles thérapeutiques d'intérêt ; cette discipline s'appelle la poly-pharmacologie. Dans d'autre cas, elle peut avoir un impact délétère sur des voies de signalisation non impliquées dans la maladie traitée et essentielles au bon maintien des fonctions d'une cellule, d'un organe ou d'un tissu. On appelle ceci des effets indésirables. Ces effets indésirables peuvent également être liés à des molécules exogènes d'origine naturelle ou artificielle (produits issus de l'industrie chimiques, les pesticides, les additifs alimentaires, et autres polluants) issues de l'environnement d'un individu. C'est le cas des perturbateurs endocriniens définis par l'Organisation mondiale de la santé (OMS) comme « des substances ou mélanges de substances, qui altèrent les fonctions du système endocrinien et de ce fait induisent des effets néfastes dans un organisme intact, chez sa progéniture ou au sein de (sous)- populations ».

1.2.1 Effets indésirables des médicaments

En 2009, les effets secondaires des médicaments constituaient la 4ème cause de mortalité aux États-Unis, avec environ 2 millions de patients affectés et 100 000 décès chaque année[23]. Avec le manque d'efficacité du candidat médicament, ils sont responsables d'environ 30% du taux d'abandon dans les essais cliniques. Les essais mis en place lors des phases précliniques de développement de médicament ne permettent pas d'anticiper l'ensemble des effets secondaires pour plusieurs raisons : 1) de nombreuses protéines liées au effets secondaires observés n'ont pas été caractérisées[24], 2) certains effets secondaires ne sont décelables que lors d'expositions prolongées à un traitement, 3) les résultats de tests de toxicité sur modèles animaux ne sont pas systématiquement transposables à l'homme[25], 4) les mêmes effets secondaires peuvent différer d'un organisme à l'autre au sein d'une même espèce. La stratégie la plus commune pour éviter la survenue de ces effets secondaires consiste à rechercher des médicaments capables d'agir sur leur cible à très faible dose, en espérant que cette forte affinité pour la cible lui confère de la sélectivité. Une autre stratégie consiste à mesurer l'affinité des leads sur un petit ensemble de récepteurs à risque avant de faire évoluer les molécules en phases précliniques[26]. Néanmoins, pour appréhender efficacement les effets secondaires d'un médicament, il faudrait mesurer son activité sur une grande quantité de cibles potentielles, ce qui reste très peu accessible à la grande majorité des laboratoires si l'on considère seulement la faisabilité logistique et économique. De plus en plus d'outils virtuels de prédiction des effets secondaires sont donc mis en place pour amoindrir les coûts. Parmi les

méthodes appliquées, nous retrouvons des méthodes de biologie des systèmes et de criblage virtuel qui ont déjà porté leurs fruits[24,27].

1.2.2 Les perturbateurs endocriniens

Les perturbateurs endocriniens, quant à eux, font l'objet d'une attention particulière depuis le début des années 2000[28,29]. Leur interférence avec différents étapes du cycle des hormones naturelles (production, sécrétion, transport, métabolisme, liaison, action ou élimination) et la découverte de leur impact sur le développement, la reproduction et sur les fonctions neurologiques, cardiovasculaires, métaboliques et immunitaires[29] chez l'ensemble des êtres vivants a alerté la communauté scientifique et l'attention collective. La compréhension des mécanismes de perturbation est aujourd'hui au cœur des recherches de santé publique. L'historique de la recherche sur les perturbateurs endocriniens rappelle qu'il n'existe pas de relation linéaire entre l'exposition aux perturbateurs endocriniens et la déclaration d'un phénotype. Contrairement aux effets toxiques habituels, les perturbateurs endocriniens semblent avoir des effets plus néfastes à faible et forte dose qu'à des doses moyennes[30]. La faible connaissance de leur mode d'action explique la mise en place de travaux de recherche à échelles macroscopiques[31] (études cliniques et *in vivo*) et microscopiques (*in vitro* et *in silico*) ; leur objectif est de résoudre leur mécanisme d'action et d'estimer les risques associés à des expositions plus ou moins prolongées et à plus ou moins forte dose, afin de mettre en œuvre des solutions visant à minimiser ces risques. Parmi les initiatives émergentes, le programme « Toxicology in the 21st Century » (Tox21)[32], qui résulte d'une collaboration entre l'agence américaine de protection de l'environnement (EPA) et l'institut national de santé (NIH), et l'agence américaine des produits alimentaires et médicamenteux (FDA), a pour objectif de mettre en place des méthodes d'évaluation de la toxicité de 10 000 composés jugés potentiellement dangereux (pesticides, additifs alimentaires, produits chimiques commerciaux, médicaments etc.). L'Organisation for Economic Co-operation and Development (OECD) propose un ensemble de protocoles validés pour évaluer le risque d'écotoxicité de composés chimiques (http://www.oecd.org/env/ehs/testing/seriesontestingandassessmentecotoxicitytesting.htm). Plusieurs outils comme l'Endocrine Disruptome[33] et VirtualToxLab[34] proposent des méthodes de criblage virtuel et de QSAR pour évaluer les interactions possibles entre les petites molécules testées et des récepteurs dérégulés par les perturbateurs endocriniens, parmi lesquels se trouvent les récepteurs nucléaires.

1.3 Généralités sur le criblage virtuel dans la recherche de médicament et la prédiction d'effets indésirables

Le criblage virtuel peut être visualisé comme un entonnoir ou un tamis dans lequel nous versons une grande quantité de molécules, et duquel ressort idéalement une quantité réduite de molécules ayant une affinité probable pour la cible étudiée. En d'autres termes, il a pour objectif d'identifier, à partir de larges collections de composés chimiques, des molécules appelées hits capables d'interagir avec la cible étudiée et de moduler son activité. Le criblage virtuel est généralement utilisé en phase précoce de la

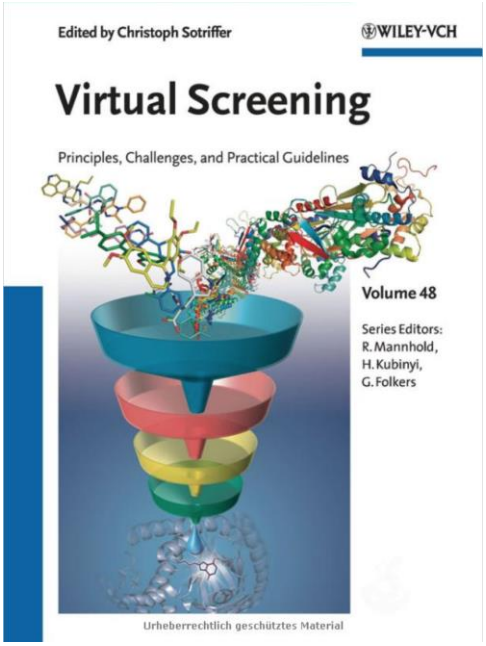


Figure 2 Couverture du livre intitulé « Virtual Screening : Principles, Challenges, and Practical Guidelines » publié en 2011 chez Wiley VCH

recherche de médicament. En 1981 le magazine Fortune titrait « Next Industrial Revolution: Designing Drugs by Computer at Merck », soulignant le potentiel de ces nouvelles approches[35]. Plus récemment, la capacité des méthodes de criblage virtuel à modéliser des interactions ligand/protéine a été utilisée dans le cadre de la prédiction d'effets indésirables en étudiant la capacité des molécules (médicaments ou perturbateurs endocriniens) à se lier à des cibles sensibles comme les récepteurs nucléaires, les cytochromes P450, le canal hERG, ou simplement des cibles autres que la cible principale.

1.3.1 Criblage virtuel versus criblage à haut débit *in vitro*

Malgré l'essor de nouvelles méthodes informatiques, l'augmentation des capacités techniques de calcul[59], et le développement de formations universitaires dédiées[37], l'apparition du criblage à haut débit,

autrement nommé le *High Throughput Screening* (HTS) a dépassé les méthodes de prédiction informatique pour la recherche de molécules thérapeutiques[38]. Le HTS est une automatisation des essais de criblage *in vitro* pour un grand nombre de molécules (de l'ordre du millier voire du million[39]) qui vise à identifier les composés provoquant une réponse biologique recherchée. En 2004, il était estimé qu'un laboratoire équipé pouvait cribler plus de 100 000 composés en une journée, et réaliser une campagne de HTS incluant un criblage primaire ainsi qu'un criblage de confirmation en moins de 2 mois[40]. La possibilité d'obtenir des mesures expérimentales comparables sur une large diversité de composés a permis d'identifier des hits dont certains ont été optimisés et commercialisés (ex : le maraviroc (Celsentri®, Pfizer Product Inc) [41], l'eltrombopag (Promacta®, GlaxoSmithKline) [42], et le BMS-790052 (Daclatasvir®, Bristol-Myers Squibb) [43]). Le taux de succès, par rapport au nombre de molécules testées, reste néanmoins très faible (de l'ordre de 0 à 4%[44]) et la sélection sans *a priori* des molécules criblées, couvrant parfois un espace chimique très restreint, ou *a contrario* très large et éparse, fait l'objet de nombreuses critiques[45]. Ces observations ont mené à un regain d'intérêt pour les méthodes *in silico* qui ont à plusieurs reprises démontré un taux de succès supérieur aux méthodes HTS (de l'ordre de 10% à 70%[46,47]). Le potentiel des méthodes informatiques a été illustré de manière frappante par la recherche d'inhibiteur du facteur de croissance transformant TGFβ. En 2003, le laboratoire Eli Lilly a mis en place un protocole classique de HTS suivi d'une étude *in vitro* de relation structure/activité[48] alors que Biogen Idec a utilisé un protocole de criblage informatique[49] basé sur la recherche de molécules partageant une similarité de forme et des points pharmacophoriques imposés avec un faible inhibiteur de référence, SB203580. Cette dernière méthode a permis d'identifier, parmi les 87 hits obtenus, un composé identique au meilleur inhibiteur potentiel retenu par Eli Lilly, tout en impliquant des coûts de recherche moindres.

Le criblage virtuel apparaît comme une approche alternative et complémentaire aux méthodes d'HTS expérimentales. Il permet notamment de cribler des composés originaux non disponibles commercialement ou purement issus de la conception informatique[50], en supplément des banque de composés commerciales[51,52]. Selon les données à disposition, différentes méthodes de criblage virtuel sont applicables. Lorsque la structure tridimensionnelle de la protéine étudiée est résolue expérimentalement (cristallographie, RMN, Cryo-EM), des méthodes basées sur la structure peuvent être utilisées. Lorsque la structure de la cible n'est pas ou que partiellement élucidée, deux solutions se présentent : 1) des méthodes basées sur les ligands connus de la cible peuvent être appliquées, et 2) la structure de la cible peut être modélisée via des approches *in silico* afin d'appliquer des méthodes basées sur la structure. Il faut noter que les études de criblage virtuel donnent des résultats théoriques qui doivent systématiquement être confirmée par des tests experimentaux. Le manque de données expérimentales concernant aussi bien la structure d'une cible que ses ligands limite fortement les stratégies de recherche applicables *in silico*. En 2013, on comptait 41% de familles protéiques dépourvues de structures tridimensionnelles[53].

1.3.2 Criblage virtuel basé sur la structure

Le criblage basé sur la structure s'appuie essentiellement sur l'estimation d'une affinité entre une molécule donnée et le site de liaison d'une protéine via le calcul d'un score. Le site de liaison peut être déduit de données expérimentales, telles que la co-cristallisation d'une protéine avec un ligand connu[54,55,56] et la mutagénèse dirigée, ou d'estimations informatiques[17,57,58]. La méthode de criblage basé sur la structure la plus communément utilisée est l'amarrage moléculaire, plus souvent appelé *docking*[59]. Le docking s'effectue en deux temps : une première étape consiste à échantillonner un ensemble de poses possibles d'une petite molécule dans le site de liaison de la protéine, la deuxième étape consiste à associer un score aux poses prédites. Dans un cas idéal, la pose associée au meilleur score correspond au mode de liaison *in vivo*, et le score reflète l'affinité réelle de la petite molécule pour la cible étudiée. Cependant, les modèles et les fonctions de scores étant systématiquement simplifiés, il n'existe pas de score parfait ; l'objectif est d'avoir une fonction de score qui corrèle au mieux avec les données expérimentales. La combinaison du docking et de tests *in vitro* et *in vivo* s'est avérée fructueuse pour la découverte de hits dans de nombreux cas[60] ; elle a notamment permis l'identification d'un inhibiteur du TNFα possédant un mode d'action en deux temps avec Kd de K_{d1} = 4.79 ± 1.12 μM and K_{d2} = 2.31 ± 1.03 nM, représentant la seule petite molécule anti- TNFα non toxique publiée à ce jour[61]. D'autres études de *docking* ont permis l'identification d'inhibiteurs de la 1-deoxy-D-xylulose-5-phosphate reductoisomerase (DXR) de *Mycobacterium tuberculosis* et du *Plasmodium falciparum* – protéine nécessaire à la survie des parasites –

dont le mode de liaison a été validé par cristallographie[62,63] et dont l'optimisation a conduit à la synthèse de molécules possédant un Ki de l'ordre du nM[64]. Le docking montre également un intérêt grandissant dans la prédiction du risque perturbateur endocrinien[65,66] (Endocrine Disruptome[33], VirtualToxLab[34]) et dans la compréhension de leur mode d'action[67]. En 2008, Celik et al. ont ainsi décrit des modes de liaison probables de perturbateurs endocriniens (polychorobiphényles (PCBs), plastifiants, et pesticides) dans le récepteur ERα, révélant des affinités théoriques fortes et la susceptibilité d'ERα à être ciblé par nombre d'entre eux[67].

D'autres approches telles que la conception de pharmacophores basés sur la structure[68,69,70], la conception de molécule de novo[71] ou encore la dynamique moléculaire[72,73] permettent d'identifier ou de synthétiser virtuellement des hits et d'évaluer et comprendre leur affinité pour la cible étudiée.

1.3.3 Criblage virtuel basé sur les ligands

Les méthodes basées sur les ligands présentent l'avantage de s'affranchir de données structurales de la protéine. Elles nécessitent néanmoins des données d'affinité et/ou d'activité expérimentales entre au minimum un ligand et la cible étudiée. Lorsque peu de données sont disponibles, des *hits* peuvent être identifiés par recherche de molécules analogues grâce à des méthodes de recherche par similarité de forme[74], de structure[75] ou de propriétés physicochimiques 1D, 2D ou 3D (ex : pharmacophore)[7677].

Lorsque les données sont plus abondantes, des modèles de relation structure/activité peuvent être construits. Ces méthodes ont contribué à la découverte de nombreux inhibiteurs de protéines[60]; par exemple, un modèle de pharmacophore dérivé d'antagonistes connus du récepteur muscarinique M3 a permis d'identifier 172 antagonistes potentiels issus de la chimiothèque d'Astra Charnwood parmi lesquels plusieurs composés ont montré des affinités expérimentales de l'ordre du µM pour le récepteurs M3[78].

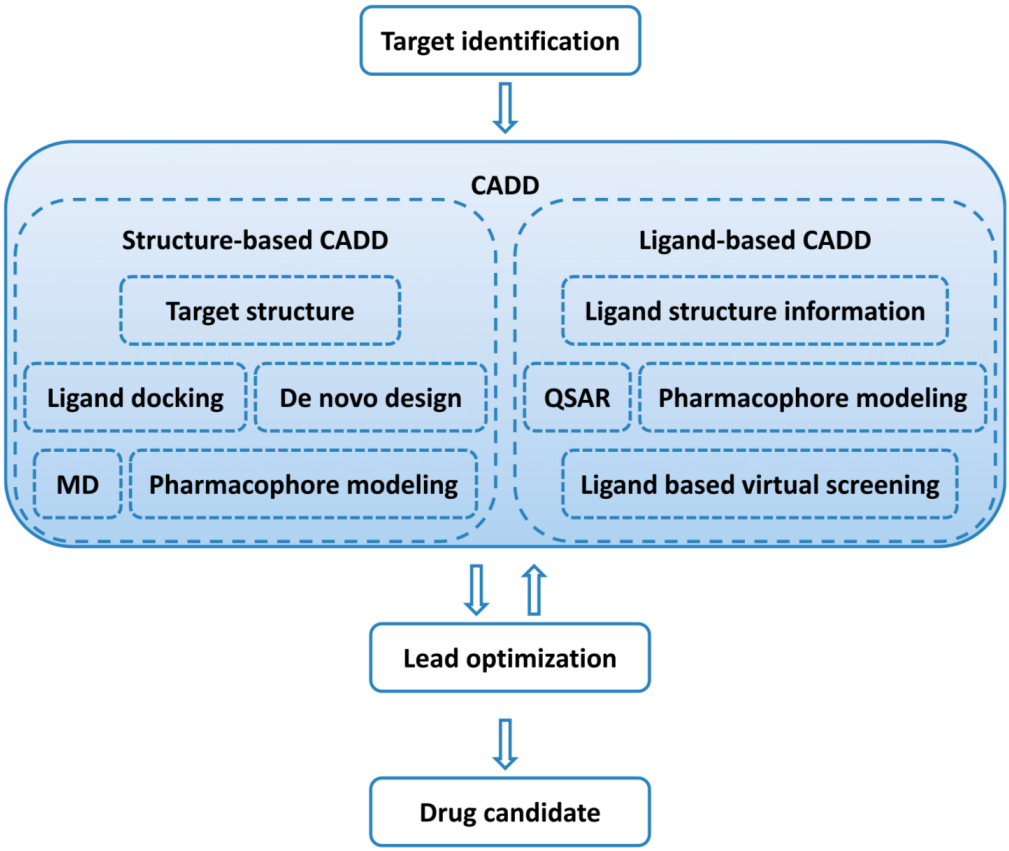


Figure 3 Les méthodes de recherche de médicament assistées par ordinateurs, communément appelées méthodes de *Computer-Aided Drug Design* (CADD) sont applicables dès lors qu'une cible thérapeutique est identifiée. En fonction des données structurales et des données d'interaction petite molécule/protéine cible, des approches basées sur la structure ou basées sur le ligand peuvent être privilégiées. Un criblage réussi aboutit à l'identification d'une « têtes de série » (« *hit* ») qui subira des cycles d'optimisation afin de proposer des composés « phares » (« *lead* »). Ces composés « phares » sont ensuite testés in vivo pour identifier des candidats médicament. D'après Sliwoski, 2014

2 Les chimiothèques

2.1 Les différentes chimiothèques

Les collections de petites molécules, appelées chimiothèques, sont essentielles au criblage virtuel. Ce sont elles qui vont être criblées de sorte à identifier les molécules affines pour la cible étudiée. Dans le cas du criblage basé sur le ligand, elles sont également utilisées pour la construction de modèles.

Les chimiothèques disponibles publiquement ou commercialement se divisent en 4 grandes classes : les chimiothèques de données bioactives qui recensent des données d'activité et/ou d'affinité de couples ligands/protéines issues de tests expérimentaux, les chimiothèques de molécules commerciales, les chimiothèques de fragment et les chimiothèques de composés virtuels (Cf Annexe 4). Les chimiothèques de données bioactives (ChEMBL[79], PubChem[80], NRLiSt BDB[5], NR-DBIND[81]) sont utilisées pour la construction, l'évaluation et le calibrage de modèles de criblage aussi bien basés sur le ligand que sur la structure. Lors de l'étape de criblage, l'information sur la bioactivité n'est pas nécessaire et davantage de chimiothèques peuvent être utilisées (ZINC[52], eMolecules, ChemSpider, ChemBridge, Asinex, Enamine, Maybridge). Le nombre de molécules « réelles », c'est à dire ayant déjà été synthétisées ou extraites de bioressources, recensées dans ces banques de données se rapproche du milliard[5]. Bien que cette quantité paraisse importante, la découverte de nouvelles molécules thérapeutiques stagne[82]. Une explication très discutée dans la communauté scientifique se trouve dans le manque de repréentativité de l'espace chimique connu (10^9 molécules) par rapport à l'espace chimique total qui serait composé de 10^{33} molécules en se limitant aux molécules répondant à la règle de Lipinski[83] (Cf 2.2.3.1). Pour pallier ce manque, la construction de molécules par des approches *de novo* permet d'apporter davantage de diversité et de couvrir un plus grand espace chimique. Il existe ainsi des chimiothèques de fragments qui offrent la possibilité de construire de nouvelles molécules en partant de points d'accroche, et des chimiothèques de composés virtuels.

2.1.1 Chimiothèques de molécules bioactives

Les chimiothèques de molécules bioactives recensent des données d'activité ou d'affinité entre un ligand et la cible protéique contre laquelle il a été testé. C'est vers ces banques de données que s'oriente le chémoinformaticien pour enrichir son jeu de molécules actives et en tirer des informations. On y retrouve des valeurs d'activité et/ou d'affinité sous forme de :

- Constantes (Ki : constante d'inhibition, Kd : constante de dissociation),
- Concentration nécessaire pour inhiber 50% de l'activité basale d'une protéine (IC50 : concentration inhibitrice médiane) ou pour stimuler une activité à mi-chemin entre l'activité basale et l'activité maximal d'une protéine (EC50 : concentration efficace médiane),
- Pourcentage d'activité (%activité) par rapport à une molécule de référence ou par rapport à l'activité basale d'une protéine.

La plupart des valeurs d'affinité proviennent de test de transfert d'énergie entre molécules fluorescentes (FRET), de polarisation de fluorescence (FP) ou d'analyse de scintillation par proximité (SPA) ou de Résonance des plasmons de surface (SPR). Les valeurs d'activité proviennent de tests cellulaires parmi lesquels le plus fréquemment utilisé est l'utilisation de gène rapporteur.

2.1.1.1 Tests d'affinité ligand/protéine

Les tests permettant de mesurer l'affinité entre un ligand et sa cible se divisent en deux catégories : les tests qui nécessitent un ligand marqué et ceux qui n'en nécessite pas (Tableau 1). Les tests dépendant de ligands marqués sont très fréquemment utilisés et faciles à mettre en place, certains d'entre eux requièrent cependant des ligands radio-marqués et font l'objet de discussions quant à leur utilisation peu écologique.

Tableau 1 Liste de tests d'affinité fréquemment utilisés.

Tests d'affinité ligand/protéine			
Méthode	Abréviation	Label	Propriétés exploités
Tests dépendants de ligands labellisés			
Transfert d'énergie	FRET	Fluorophores	Propriétés optiques des fluorophores - transfert

entre molécules fluorescentes			d'énergie d'un fluorophore A excité à une fluorophore B
Polarisation de fluorescence	FP	Fluorophores	Polarisation des fluorophores - phénomène de dépolarisation lors de la rotation rapide d'un fluorophore/ polarisation lors de sa stabilisation
Analyse de la scintillation par proximité	SPA	Radiomarqueurs (ex : ^{3}H, ^{14}C, ^{32}P, ^{35}S ou ^{125}I)	Adsorption de protéines sur un métal scintillant excité par proximité aux ligands radiomarqués
Tests indépendants de ligands labellisés			
Résonance des plasmons de surface	SPR	-	Protéines adsorbées sur une surface métallique - changement d'amplitude et de phase d'une onde de surface lors de la l'interaction de molécules avec les protéines adsorbées

2.1.1.2 Tests d'activité ligand/protéine

Il existe une très grande variété de test d'activité applicable selon le système d'étude. Parmi ces tests, le plus connu est l'utilisation de gène rapporteur. Un gène rapporteur dont l'expression est facilement quantifiable (ex : gène codant pour la protéine fluorescente verte (GFP), la luciférase, la β-glucuronidase ou la β-galactosidase) est placé à proximité du promoteur d'un gène induit soit 1) directement par la protéine étudiée s'il s'agit d'un facteur de transcription soit 2) indirectement par une voie de signalisation impliquant la protéine d'intérêt. La quantification de l'expression du gène rapporteur permet de mesurer l'activité de la protéine étudiée. Par exemple, la luciférase est une protéine bioluminescente ; son expression peut être mesurée par un photomètre et permet d'évaluer l'activité basale d'une protéine, son activité maximale, et la variation d'activité liée à l'interaction avec une petite molécule. On peut ainsi déduire l'IC_{50} ou l'EC_{50} d'un ligand, et son pourcentage d'activité basale ou d'activité maximale.

2.1.1.3 Banques de données de référence

Parmi ces chimiothèques de molécules bioactives existantes, la base de données ChEMBL, gérée par l'Institut européen de bioinformatique (EBI), fait office de référence ; elle répertorie plus de 15 millions de données d'activités pour plus de 1,6 millions de composés différents[79,84]. Elle fournit des données d'activité pour ~11 000 cibles, parmi lesquelles 9052 sont des protéines dont 4255 sont humaines. La ChEMBL a été initialement construite pour regrouper les données de bioactivité

issues des publications de chimie médicinale et s'est élargie au fil des années avec des données issues de laboratoires académiques, d'institutions gouvernementales ou de l'industrie pharmaceutique. Par exemple, suite à un appel à la publication des données concernant les maladies négligées comme la malaria en 2010, de nombreux groupes académiques et privés ont rendu les données issues de leurs projets de recherche gratuitement disponibles sur la ChEMBL[84] (Cf Annexe 3).

La ChEMBL a été utilisée comme base pour la création de banques de données focalisées sur des familles de protéines. C'est le cas de la Nuclear Receptors Ligands and Structures Benchmarking DataBase (NRLiSt DBD) et de la Nuclear Receptors DataBase Including Negative Data (NR-DBIND) développées dans notre laboratoire et pour lesquelles les données issues de la ChEMBL ont été validées ou corrigées par inspection manuelle de leur publication d'origine.

Les données de ces banques ne sont pas homogènes : la diversité des tests expérimentaux et des mesures (Ki, Kd, IC_{50}, EC_{50}, %activité) rend difficile la comparaison des données. Il est par exemple dangereux de comparer des valeurs d'affinité exprimées en Ki avec des valeurs exprimées en IC_{50}. Le Ki représente une constante qui renseigne sur la capacité d'un ligand à se fixer sur sa cible et à y rester alors que l'IC_{50} dépend de la concentration de la protéine d'étude et du ligand de référence. Généralement l'IC_{50} est mesurée par des tests de compétition et peut être convertie en Ki grâce à l'équation de Cheng-Prusoff :

$$Ki = \frac{IC_{50}}{1 + \frac{[ref]}{EC_{50}}}$$

Équation 2 Équation de Cheng-Prusoff permettant de convertir l'IC_{50} en Ki dans le cas d'un test de compétition

où [ref] est la concentration en composé de référence utilisé pour la compétition et EC_{50} est la concentration du composé de référence qui déclenche 50% de l'activité maximale du récepteur. Bien que des études montrent que des IC_{50} comparables sont obtenues en utilisant des techniques expérimentales différentes comme la FRET, la SPA et la FP[85], il est toujours hasardeux de comparer les tests réalisés dans des conditions expérimentales différentes et par des équipes de recherche différentes. Leur utilisation en criblage virtuel nécessite donc une étape préalable de nettoyage afin d'obtenir des données les plus homogènes et comparables possibles. C'est au

chercheur de trouver la balance idéale entre la quantité et la qualité des données et de pouvoir justifier ce choix. En ce sens, les données issues de HTS présentent l'avantage d'être 1) nombreuses et 2) soumises à des protocoles identiques. Certaines de ces données sont désormais accessible sur la PubChem Bioassay (et par conséquent sur la ChEMBL) : plus de 7000 jeux de données issus d'études HTS y sont répertoriés. On y retrouve notamment les données de HTS du projet Tox21 initié par l'agence américaine de protection de l'environnement (EPA), l'institut national de santé (NIH), et l'agence américaine des produits alimentaires et médicamenteux (FDA) qui a permis d'évaluer l'affinité et l'activité d'environ 10 000 composés présentant des risques potentiels de perturbation endocrinienne (médicament, pesticides, additifs alimentaires et autres polluants) sur 30 cibles incluant de nombreux récepteurs nucléaires.

2.1.2 Chimiothèques de molécules commerciales

La plupart des chimiothèques de molécules commerciales (Cf Annexe 4) sont mises à disposition gratuitement par les fournisseurs (ZINC[52], ChemSpider, Enamine). L'avantage de cet accès libre est qu'il rend la recherche de médicament tangible pour tous[52] et présente un argument commercial fort pour les vendeurs. Les molécules en question sont disponibles à l'achat en vrac ou en microplaques. La ZINC compte parmi les plus fournies (>750 millions de molécules) et les plus utilisées pour le criblage virtuel.

Certaines banques commerciales possèdent des sous-jeux de données focalisés sur des familles de protéines qui contiennent des molécules sélectionnées par similarité avec leurs ligands (Enamine, ChemBridge, Asinex). Par exemple la banque ChemBridge contient des chimiothèques spécifiques de la famille des protéines kinases, des récepteurs couplés aux protéines G (GPCR), des canaux ioniques et des récepteurs nucléaires. Les molécules sont sélectionnées par criblage de pharmacophores définis grâce au site de liaison des protéines, à leurs ligands connus ou les deux.

2.1.3 Chimiothèques de fragments

Les chimiothèques de fragments permettent d'augmenter l'espace chimique exploré[86] ; par exemple, l'assemblage de 3 fragments permet d'obtenir un million de composés[87]. Leur criblage permet d'identifier des fragments ayant une forte affinité pour des points d'accroche. Ils pourront

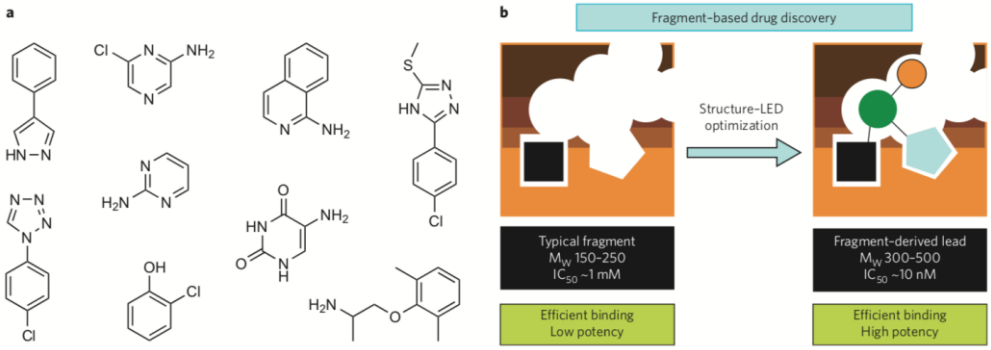


Figure 4 Illustration du principe du criblage basé sur des fragments. Des fragments sont dockés dans le site de liaison de la cible étudiés. Ceux présentant un bon score d'affinité servent de base pour la recherche ou la conception de molécule plus grandes. D'après [81]

être assemblés en petites molécules par des approches *de novo* ou utilisés pour identifier des molécules « parentes » grâce à des recherches de sous-structures (Figure 4). L'analyse des *hits* obtenus grâce à des méthodes basées sur les fragments[88] a permis de proposer une « règle de 3 » dérivée de la « règle de 5 » de Lipinski (Cf 2.2.3.1) pour guider la construction de banques de fragments. Cette règle de 3 exige :

- Un poids moléculaire < 300 Da,
- Un nombre de donneurs et d'accepteurs de liaison hydrogène ≤ 3,
- Un coefficient de partage clogP ≤ 3,
- Un nombre de liaisons rotatives ≤ 3,
- Une aire de surface polaire (SASA) < 60 Å2.

Chaque fournisseur possède sa banque de fragments (Prestwick Fragment Library, Asinex's BioFragments, ChemBridge Fragment Library, OTAVA Fragment Library, Enamine Fragment Library, Maybridge Ro3 Library), disponible à l'achat pour réaliser des synthèses chimiques. Au

contraire, la banque FDB-17[89] propose plus de 10 millions de fragments virtuels extraits de la GDB-17[50], dans l'objectif de sonder au maximum l'espace chimique explorable.

2.1.4 Chimiothèques de composés virtuels

Les chimiothèques de composés virtuels répondent à un besoin de représentativité de l'espace chimique explorable. L'ensemble de ces chimiothèques est listé dans l'Annexe 4. Elles se divisent en deux catégories :

- Celles qui contiennent des molécules synthétisées à partir de *building blocks* et d'un ensemble de réaction chimiques connues (SCUBIDOO, SAVI, CH*I*PMUNK)
- Celles qui énumèrent l'ensemble des molécules d'un espace chimique sans prendre en compte leur faisabilité expérimentale (GDB-11, GDB-13, GDB-17)

Par exemple la banque SCUBIDOO est construite à partir de 58 réactions chimiques répertoriées au format SMARTS et de 18 561 *building blocks* provenant du filtrage de la ChemBridge (poids moléculaire ≤ 250 Da, nombre de liaisons rotatives ≤ 2, nombre de centre chiraux ≤ 1) afin de réduire le nombre de produits générés. Elle fournit ainsi plus de 21 millions de molécules ainsi que leur voie de rétro-synthèse théorique.

Les molécules de la GDB-17 sont quant à elle générées via un algorithme issu de la théorie des graphes[90,50] qui tient compte de nombreuses contraintes géométriques et de groupement fonctionnels. Des exemples de molécules virtuelles de la GDB-17 sont présentés dans la Figure 5. On y retrouve des isomères de médicaments jusqu'alors jamais recensés dans une banque de molécule.

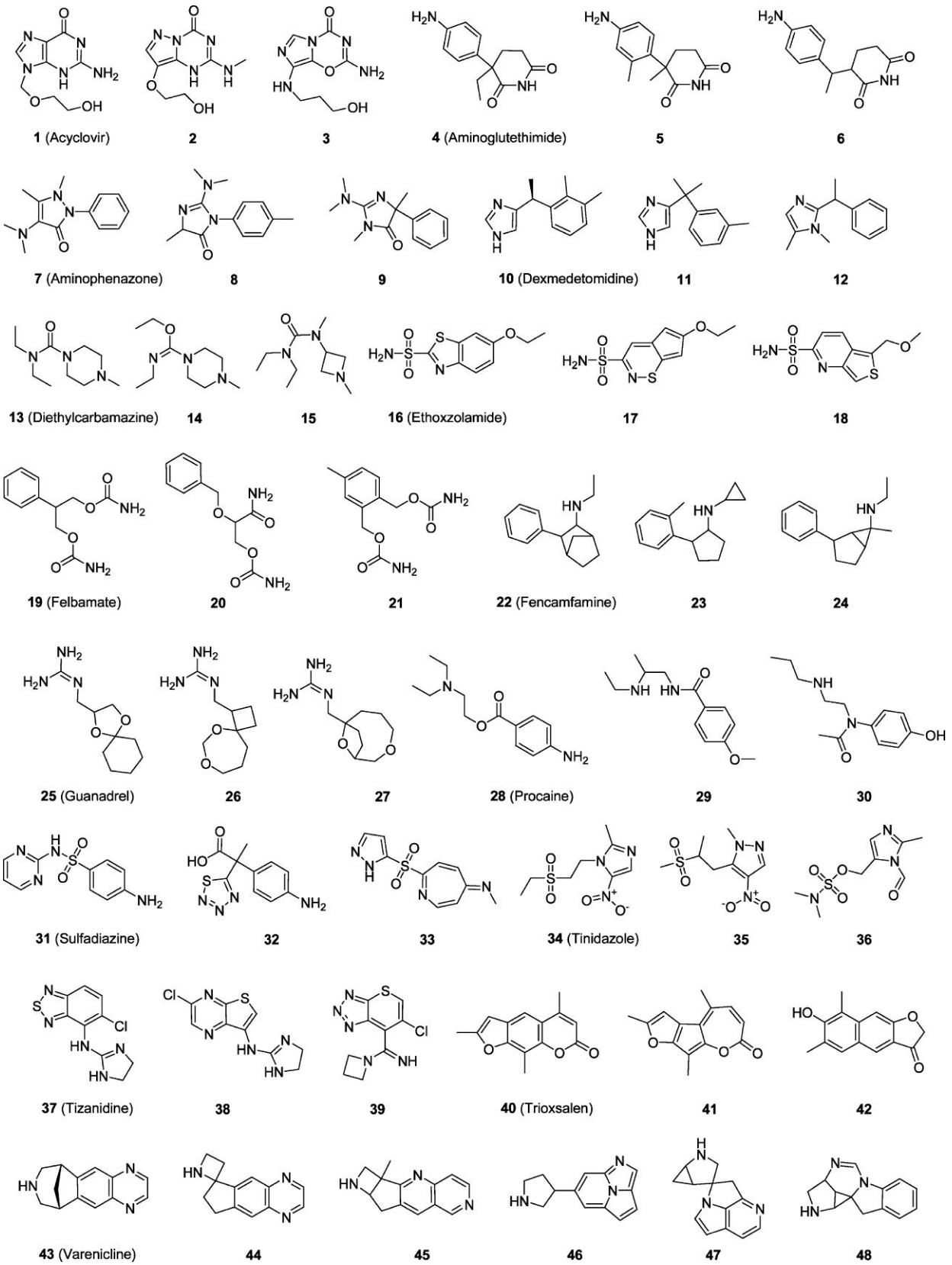


Figure 5 Médicaments retrouvés parmi les composés virtuels de la GDB-17 associés à des isomères jamais répertoriés dans des bases de données. D'après [44]

Il est à noter que l'exhaustivité de ces banques de molécules les rend inappropriées pour du criblage virtuel à haut débit rapide. Certaines banques comme la SCUBIDOO proposent donc des sous-ensembles de données réduits qui couvrent le même espace chimique que la banque entière[91].

2.2 Préparation des chimiothèques

Pour garantir la réussite d'un criblage virtuel, il est important de s'assurer des paramètres adéquats de chaque outil et de la qualité des données utilisées. En ce qui concerne les chimiothèques de criblage, elles doivent être minutieusement préparées en s'assurant de générer les différents états d'ionisation, les formes mésomères et tautomères et un nombre suffisant de conformères de chaque molécule. Les chimiothèques sont aussi filtrées afin d'exclure des molécules potentiellement toxiques (groupement réactifs, molécules toxiques répertoriées), et anticiper les interférences avec les tests d'affinité et ayant tendance à interagir avec de nombreuses cibles (Pan-assay interference compounds PAINS) mais aussi éventuellement pour restreindre la chimiothèque à des molécules répondant aux critères *drug-like* ou *lead-like*. Nous verrons dans les paragraphes suivant que le filtrage au sens large des données fait l'objet de nombreux débats et doit être appliqué à bon escient.

2.2.1 États d'ionisation, tautomérie, mésomérie

Selon les conditions physiologiques, une molécule chimique peut subir des phénomènes d'ionisation, de tautomérie ou de mésomérie. L'ionisation se caractérise par le gain ou la perte

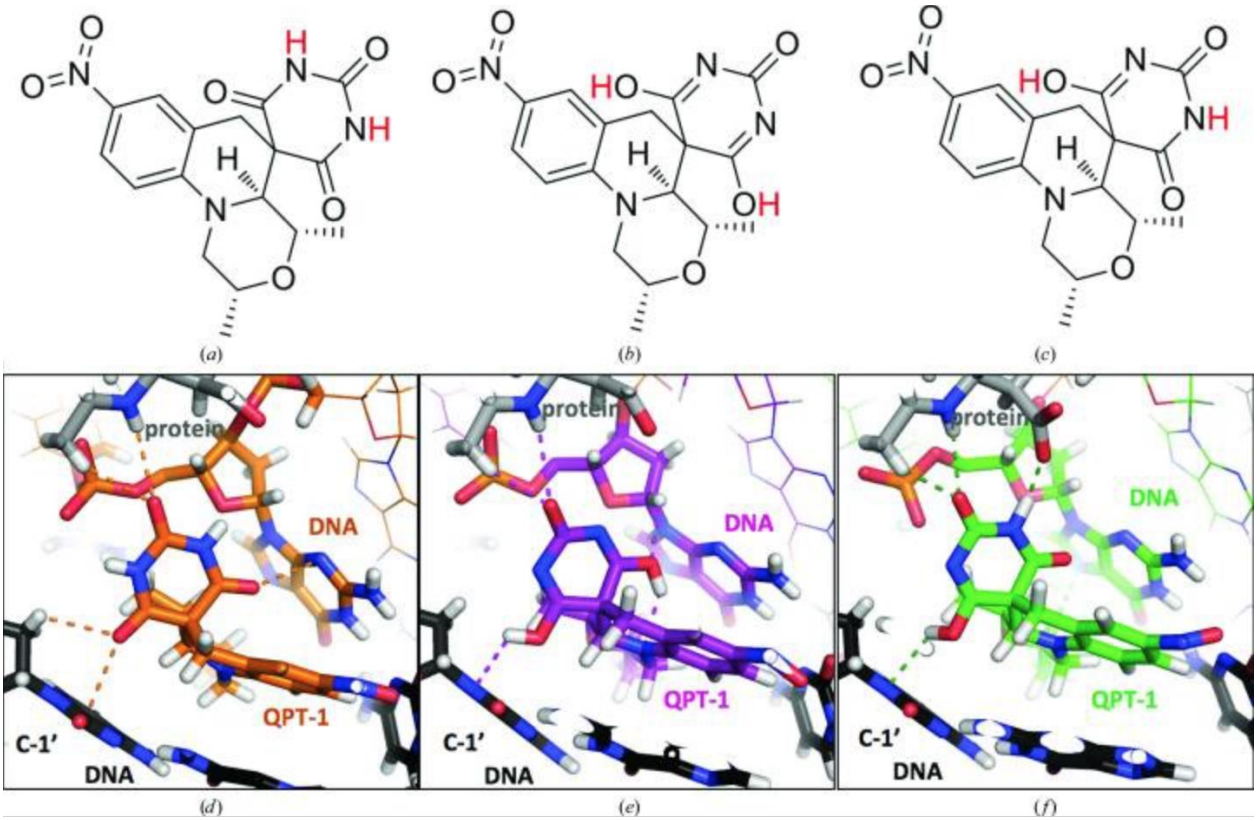


Figure 6 Exemple de tautomères de QPT-1 dockés dans un site de liaison entre d'ADN gyrase et l'ADN. Les différents tautomères sont dockés de la même manière mais impliquent des interactions différentes. D'après [87]

d'une charge sur un atome d'une molécule, la mésomérie par une délocalisation d'un électron dans le cas d'un groupement conjugué, et la tautomérie par un déplacement concomitant d'un atome d'hydrogène et d'une liaison π au sein d'une molécule (Figure 6). La considération de ces phénomènes lors d'études de criblage virtuel basé sur le ligand ou sur la structure est primordiale. Le gain ou la perte d'une charge, tout comme une délocalisation d'un atome d'hydrogène ou d'un électron modifie la polarité locale et affecte positivement ou négativement l'établissement de liaisons entre la molécule et la cible étudiée (Figure 6)[92]. Une molécule peut adopter différentes propriétés en fonction du pH, de la température et du solvant. Dans certains cas, une forme est largement prioritaire à pH neutre ; c'est cette forme stable qui sera considérée pour les études de

criblage. Par exemple, l'acide carboxylique présente un pKa compris entre 4 et 5, à pH neutre (pH=7.4) il est donc indiscutablement déprotoné. Dans d'autres cas, lorsque le pKa d'un groupement fonctionnel se rapproche du pH neutre, plusieurs formes cohabitent ; toutes les formes cohabitantes doivent alors être considérées pour les études de criblage.

De nombreux outils *in silico* sont capables de prédire les différents tautomères possibles ainsi que les états d'ionisation des molécules à un pH donné (QuacPac[93], ProtoPlex[94], Pipeline Pilot[95], TAUTOMER[96], AGENT[97], and LigPrep[98], Protoss[99], Epik[100]), mais seul Marvin (ChemAxon) prend en compte la stabilité des tautomères en estimant la constante d'équilibre entre deux formes tautomériques via des méthodes empiriques. Protoss énumère l'ensemble des tautomères possibles en respectant des combinaisons de valence chimiquement correctes[99].

2.2.2 Génération des conformères

Lorsque des modèles 1D ou 2D basés sur les ligands sont construits, la structure tridimensionnelle de la molécule est superflue. En revanche les autres méthodes basées sur les ligands et basées sur la structure de la protéine requièrent une conformation 3D de départ. Ces structures peuvent être générées par des approches stochastiques ou systématiques[101] (Tableau 3).

Parmi les méthodes stochastiques, la construction de molécules par *Distance Géométrie* (DG) est fréquemment utilisée. La méthode DG consiste à générer des points aléatoires pour chaque atome d'une molécule et à optimiser leurs positions relatives grâce à un ensemble de contraintes géométriques imposées. Des méthodes de dynamiques moléculaires, de modes normaux, de Monte Carlo ou encore des algorithmes génétiques (GA) sont parfois appliquées sur une conformation initiale déjà générée afin d'explorer des conformations de plus basse énergie.

Parmi les méthodes systématiques, les méthodes de force brute et celles basées sur la connaissance permettent de générer des conformations 3D. Les méthodes de force brute énumèrent l'ensemble des conformations possibles d'une molécule alors que les méthodes basées sur la connaissance s'appuient sur des données d'angles de torsion et de chemins interatomiques autorisés issus de données expérimentales (Protein Data Bank (PDB)[102], Cambridge Structural Database (CSD)[103,104]), voire sur des bases de données contenant une ou plusieurs conformations 3D des fragments.

Selon sa taille et sa flexibilité, une multitude de conformations peut être générée pour une molécule unique. En fonction de l'utilisation finale, il peut être judicieux de réduire le nombre de conformations retenues en ne conservant que les conformations de plus basse énergie ou les conformations les plus probables. Ces conformations peuvent être identifiées via le calcul d'un score basé sur 1) le calcul des structures électroniques (*Density Functional Theory* (DFT)), 2) des champs de forces moléculaires (MMFF94[105]; OPLS[106]; CHARMm[107], TRIPOS[108]) ou 3) la

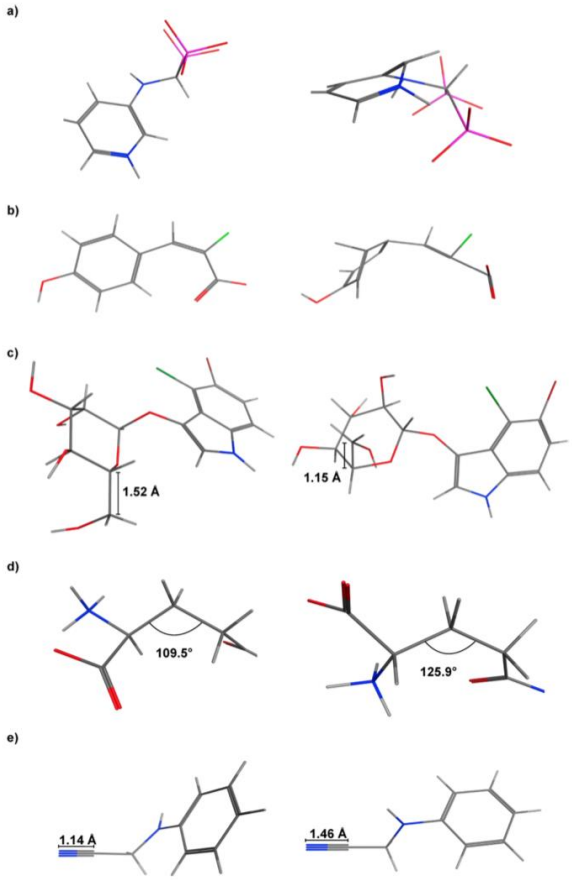


Figure 7 Exemples d'erreurs de géométrie introduites par les générateurs de conformations 3D de molécules. Les conformations expérimentales sont représentées à gauche, celles générées informatiquement à droite. On observe a) une mauvaise planéité du cycle aromatique généré (JD5, PDB : 4NFK) et b) avec Balloon DG (FHC, PDB : 2OPA), c) une mauvaise planéité du cycle aromatique et une mauvais longueur de liaison interatomique (XMM, PDB : 2JE7) avec RDKIT, d) un mauvais angle au niveau du carbone sp3 (ZBF, PDB : 4OPN) avec RDKIT ETKDG et e) une mauvaise longueur de liaison interatomique (264, PDB : 2RBN) avec Frog2. D'après [110]

connaissance. Généralement, l'énergie électrostatique n'est pas considérée pour le calcul de l'énergie potentielle d'une conformation 3D puisqu'elle a tendance à favoriser un repliement de la molécule sur elle-même. La plupart des méthodes libres d'accès et sous licence (Cf Tableau 2) présentent des performances similaires en terme de temps de calcul, de nombre de conformations générées et de capacité à retrouver une pose proche des données de cristallographie (RMSD ≤ 2 (Cf 5.3.1))[109,110,111]. Cependant, des erreurs de géométrie (longueur des liaisons, angles et planéité des cycles aromatiques) dans les conformations générées par plusieurs logiciels ont été détectées (Balloon DG[112], Balloon GA, RDKIT[113] (DG et ETKDG[114])) (Figure 7) [13]. Lors de cette même étude, seuls les outils iCon[115] et OMEGA[116], qui par ailleurs présentent des performances très proches[115], ont permis de générer 100% de conformations géométriquement correctes[111].

Tableau 3 Liste de logiciels de génération de conformation 3D ainsi que les algorithmes qu'ils emploient pour échantillonner (*sampling*) différentes conformations d'une molécule et calculer leur potentiel d'énergie. Les méthodes surlignées en bleu sont des méthodes libres d'accès, les autres requièrent une licence.

	Échantillonnage	Potentiel	Composantes	Optimisation
Ballon GA	GA	MMFF94	Liaisons, van der Waals	Oui
RDKIT (ETKDG)	DG + basé sur la connaissance	Potentiel de torsion dérivé de CSD	Angles de torsion	Oui
Frog2	Monte Carlo	MMFF94	Van der Waals	Oui
RDKIT (DG)	DG			Oui
MCDOCK	Force brute ; ancrage et agrandissement	Amber	Van der Waals et électrostatique	Non
OMEGA	Basé sur la connaissance, énumération complète	MMFF94	Pas de terme de van der Waals attractif et pas de potentiel électrostatique	Oui

ConfGen	Basé sur la connaissance	OPLS	Angles de torsion et van der Waals	Oui
iCon	Basé sur la connaissance	MMFF94s		Oui

2.2.3 Filtrage des molécules

2.2.3.1 Molécules drug-like et lead-like

Lorsque le criblage virtuel est utilisé à des fins de recherche de médicament, les chimiothèques sont souvent filtrées pour répondre à la fameuse règle de 5 de Lipinski. Cette règle a initialement été mise en place pour essayer de garantir des composés absorbables par voie orale et ainsi pallier le fort taux d'abandon des essais cliniques. Il paraissait raisonnable, pour obtenir des candidats-médicaments fiables, de tirer des connaissances des molécules ayant déjà franchi la phase I d'essais cliniques. En se basant sur l'analyse de plus de 2000 composés du WDI (World Drug Index) répondant à ces critères, Lipinski a identifié 4 propriétés communes, les Ro5, qui sont aujourd'hui largement utilisées pour qualifier une molécule de *drug-like* :

- Poids moléculaire ≤ 500Da
- Nombre de donneur de liaison hydrogène ≤ 5
- Nombre d'accepteurs de liaison hydrogène ≤ 10
- Coefficient de partition eau/octanol (ClogP) ≤ 5

Les composés répondant à la règle de Lipinski sont supposés posséder une bonne biodisponibilité orale. A l'inverse, la violation de deux de ces critères suffit à rejeter une molécule comme non *drug-like*. Bickerton et al.[117] jugent cette règle trompeuse puisque certaines molécules toxiques valident tous les critères alors que certains médicaments commercialisés ne seraient pas « drug-like » selon ces critères. Ils ont donc proposé d'autres règles dérivées des Ro5, plus flexibles et plus robustes[118], et prenant en compte des paramètres supplémentaires comme le nombre de liaisons rotatives, de cycle aromatiques, et la surface polaire accessible. A chaque descripteur a été associé une fonction de « désirabilité » et un poids correspondant à sa contribution à la *drug-likeliness*. Ainsi, la violation de deux critères de Lipinski n'est pas rédhibitoire pour disqualifier

une molécule de *drug-like*. D'autres équipes contestent radicalement la Ro5 la qualifiant de « biais historique »[119]. Le premier argument de cette contestation est son aspect trop contraignant qui tend à restreindre le nombre de molécules testées[119]. Le second argument concerne la différence d'espace chimique couvert entre les hits, les leads et les candidats médicaments : les filtres sont dérivés de candidats médicaments et tendent à rejeter les molécules *hit-like* et *lead-like* qui sont moins complexes mais qui offrent néanmoins des points de départ pour le développement de médicament[119]. De nouveaux paramètres ont été proposés pour inclure les molécules *lead-like* (poids moléculaire ≤ 350 Da et ClogP ≥ 3), leur utilisation devant rester consultative et non strictement décisionnelle. Le concept a été amélioré à plusieurs reprises[120,121,122,123,119], avec des prises en compte du nombre de donneur et d'accepteur de liaison d'hydrogène, de liaisons rotatives, de cycles aromatiques ainsi que la lipophilie (ClogP), les changements majeurs étant les poids moléculaire (≤ 460 Da) et un élargissement de la fenêtre d'inclusion du ClogP (-4 ≤ ClogP ≤ 4.2)[124,125]. Ces règles présentent de moins en moins de succès auprès de la communauté scientifique malgré une constance remarquée dans le ClogP des médicaments approuvés avant 1983 et ceux approuvés entre 1983 et 2002.

2.2.3.2 Molécules indésirables

2.2.3.2.1 Toxicité

Selon le but escompté de la campagne de criblage, le filtrage des molécules toxiques peut s'avérer d'intérêt majeur. Cette étape est exclue lors d'études de criblage ayant pour but de prédire des effets secondaires ou encore le risque de perturbation endocrinienne. En revanche, il s'agit d'une étape cruciale dans la recherche de médicaments. La toxicité des molécules est l'une des causes majeures et indiscutables d'abandons lors d'essais cliniques, un abandon si tardif du développement d'un médicament représentant un échec financièrement dramatique. Il existe une multitude de toxicités possibles selon la cible protéique affectée (toxicité sur les récepteurs nucléaires, sur le canal potassique human Ether-à-go-go-Related Gene (hERG) etc.), l'organe (hépatotoxicité, cardiotoxicité, néphrotoxicité, pneumotoxicité etc.) ou l'effet produit (carcinotoxicité, génotoxicité etc.). Des sources d'information, comme l'Adverse Outcome Pathways Knowledge Base (AOP-KB) qui regroupe des informations de toxicité connues issues de laboratoires internationaux et modérés par l'OECD, permettent de centraliser la connaissance des risques et de les anticiper. L'analyse *in silico* des AOPs, qui associent un effet indésirable à un ensemble d'évènements clés

causés par une perturbation d'un état biologique stable via l'interaction d'une molécule, devrait permettre d'améliorer la prédiction des risques[126]. Par ailleurs, des outils de prédiction *in silico* ont été mis en place pour anticiper ces risques, c'est ce que l'on appelle la toxicologie *in silico*. Cette discipline répond notamment à la réglementation européenne REACH (Registration, Evaluation, Authorization and Restriction of Chemicals) qui vise à réduire l'utilisation de molécules chimiques dangereuses pour l'homme et l'environnement ainsi que les tests de toxicité exercés sur les animaux[127]. La plupart des outils de prédiction existant utilisent des méthodes de relation quantitative structure activité, plus connues sous le nom (*Quantitative*) *Structure Activity Relationship* ((Q)SAR) (Cf 3.2). Il s'agit de modèles mathématiques associant des descripteurs physicochimiques et géométriques de molécules avérées toxiques avec leur toxicité expérimentale mesurée. La plupart des outils de prédiction de toxicité ont été construits à partir de données de toxicité expérimentales (Cf Annexe 5[128,129]), et nombres d'entre eux sont disponibles gratuitement (ex : ToxiPred[130], DeepTox, admetSAR[131], ToxiM[132], VirtualToxLab[34], OpenTox[127], FAF-Drug4[133]). Parmi ces outils, admetSAR présente 25 modèles QSAR pour prédire la toxicité d'une molécule (Tableau 4). Des études plus nombreuses encore ont permis de publier des modèles adaptés à des jeux de données réduits, peu appropriés à la prédiction sur des espaces chimiques différents, mais néanmoins très performants[134].

Tableau 4 Exemples de modèles de prédiction de la toxicité disponibles sur AdmetSAR

Class	Model
Drug Induced Toxicity	Drug-Induced Liver Injury
	Human Ether-a-go-go-Related Gene (hERG) Inhibition
	Rat Acute Toxicity
	Skin Sensitivity
Genomic Toxicity	AMES Toxicity
	Carcinogens
Aquatic & Terrestrial Toxicity	Phytoplankton toxicity
	Fish (eg. Fathead Minnow) toxicity
	Tetrahymena Pyriformis toxicity
	Honey Bee Toxicity

	Quail Toxicity
	Rodent (human, rat, mouse, hamster etc.) animals toxicity
Reproductive Toxicity	Estrogen Receptor α
	Estrogen Receptor β
	Androgen Receptor
	Peroxisome Proliferator-Activated Receptor γ
	Peroxisome Proliferator-Activated Receptor δ
	Farnesoid X Receptor
	Glucocorticoid Receptor
	Retinoid X Receptor α
	Thyroid Hormone Receptor β
	Vitamin D Receptor
	Human Pregnane X Receptor
Environmental Factor	Biodegradability
	Bioconcentration Factors

2.2.3.2.2 Frequent hitters et PAINS

Les *frequent hitters* sont des molécules fréquemment associées à des affinités significatives peu importe la cible étudiée. Ce terme opaque est très souvent utilisé par abus de langage pour évoquer différents types de composés que Baell et Nissink ont récemment proposé de reclasser et renommer pour s'affranchir de toute ambiguïté de nomenclature[135]. On y trouve :

- Des vrais *hits*, capables d'interagir sur la ou les cibles testées de manière non covalente
- Des mauvais *hits*, ou *bad actors*, qui correspondent aux molécules capables de moduler la cible mais qui interfèrent avec celle-ci de par leur réactivité
- Des faux positifs qui comprennent :
    - Les molécules interférentes du point de vue de la cible, qui interagissent de manière non spécifique avec la cible, les détergents ou les molécules réactives
    - Les molécules qui interfèrent avec l'essai mis en place

Les mauvais *hits* et faux positifs sont regroupés sous le nom de PAINS (Pan Assay Interference Compounds). Ces molécules sont redoutées en chimie médicinale puisqu'elles montrent des résultats expérimentaux encourageant bien qu'elles ne soient aucunement capables de moduler l'activité de la cible. Ainsi, malgré tous les efforts déployés, l'optimisation de PAINS est vouée à l'échec[135]. Si certains mécanismes d'interférences sont connus, d'autres restent non élucidés. Par exemple, nous savons que des composés azo, des quinones, et des rhodanines absorbent la lumière à des longueurs d'ondes comprises entre 570 et 620 nm et interfèrent avec les signaux d'essais photométriques. En revanche, une étude montre que de nombreux PAINS n'interfèrent pas avec 100% des essais de HTS faisant appel aux mêmes technologies, ce qui ne simplifie pas la compréhension du mécanisme[135]. Des essais de HTS ont permis de mettre en avant des sous-structures impliquées dans ces interférences[136,137,138,139,140]. Des outils *in silico* ont été mis en place pour filtrer les PAINS en s'appuyant sur les données de la littérature[141,142]. Par exemple, FAF-Drug[143] possède un filtre qui inclut 326 *frequent hitters* (15 sous-structures différentes), des agrégateurs, des ligands dits « *promiscuous* » c'est à dire non sélectifs d'une ou quelques cibles, et enfin 511 sous-structures associées aux PAINS tels que décrit dans l'étude de Baell et Holloway[136]. Cependant, l'étude de Baell et Nissink précédemment mentionnée[135] rappelle que l'identification expérimentale des PAINS utilisés pour la prédiction présente de nombreuses faiblesses :

- Les PAINS sont uniquement issus d'observations et d'analyses statistiques simples de données de HTS et ne sont pas dérivés de toxicophores connus, de groupements chimiques défavorables ou de données de pharmacocinétiques ; il est donc incorrect d'associer un PAIN à de mauvaises propriétés pharmacocinétiques tout comme il est incorrect d'associer un toxicophore à un PAIN

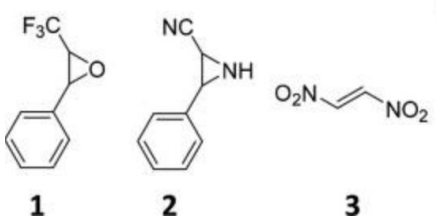


Figure 8 Structures des époxydes (1), des aziridines (2) et des nitroalkènes (3) retirés des données HTS utilisé pour la caractérisation de PAINS. D'après [129]

- Les données de HTS utilisées pour la mise en place de prédicteurs de PAINS ont elles même été filtrées notamment pour retirer les molécules possédant des groupements réactifs tels que les époxydes, les aziridines et les nitros (Figure 8), qui sont par conséquent automatiquement omis des filtres en découlant
- Un changement de plateforme de HTS ou de protocole avec une même plateforme de HTS peut induire de nouveaux modes d'interférences
- Les molécules ont été testées à forte concentration (50µM) ce qui a tendance à amplifier le comportement négatif observé qui peut disparaître à faible dose
- La présence de sous-structure qualifiées de PAINS dans une molécule n'implique pas systématiquement que la molécule est elle-même un PAIN[144]

La bonne conduite à suivre selon les auteurs est d'être précautionneux quant à l'interprétation d'un *hit* : un *hit* ne devrait pas être considéré comme utile s'il n'est pas appuyé par des études de SAR. En d'autres termes un *hit* apporte peu d'information s'il ne peut être amélioré. Les filtres de PAINS peuvent apporter de l'information sur la propension à donner des résultats de criblage anormaux mais l'interprétation doit tenir compte de la structure du PAINS identifié et de la cause de l'interférence. Dans ce sens, le *Journal of Medicinal Chemistry* exige depuis 2016 davantage de données pour accepter de publier des composés reconnus comme PAINS par des filtres de PAINS basés sur des SMARTS[135].

3 Criblage virtuel basé sur le ligand

Dès lors que nous disposons d'au moins une molécule active sur la cible d'étude et d'une banque de données de molécules à tester, nous pouvons appliquer des méthodes de criblage basé sur le ligand. Il existe de nombreuses méthodes que l'on peut classer en fonction de la quantité de données nécessaire à leur application. Lorsqu'un seul ligand de la cible est connu, nous pouvons appliquer des méthodes de recherche de similarité qui s'appuient sur le principe que les molécules structurellement similaires ont des propriétés similaires[74]. Ces approches sont nombreuses et variées, la similarité pouvant être évaluées en comparant les molécules en termes de surface, de propriétés physicochimiques 2D (empreinte moléculaire, connectivité entre groupements fonctionnels), ou de propriétés dans un espace 3D (similarité de forme, modèles de pharmacophores)[145]. Des études de relation structure activité (SAR) peuvent être appliquées lorsque davantage de ligands de la cible sont connus. L'utilisation d'une seule molécule (ou d'un faible nombre de molécules) comme point de départ peut se révéler très limitant puisque cela ne permet pas de mettre en avant les propriétés pertinentes pour assurer une activité avec la cible ; les propriétés communes entre plusieurs molécules actives sont plus informatives. Il faut cependant se méfier des redondances entre plusieurs molécules actives lorsqu'elles appartiennent à une même série chimique, ou des séries chimiques très proches. Dans ce cas précis, les redondances observées sont généralement liées au squelette de la série et non pas aux propriétés impliquées dans l'activité. L'idéal est d'avoir des molécules structurellement différentes desquelles nous pouvons extraire des informations communes.

3.1 Recherche de similarité

La similarité est un concept très subjectif ; établir un lien de similarité entre deux molécules dépend entièrement de la manière dont une molécule est décrite et dont la similarité est quantifiée. Les différentes stratégies utilisées seront décrites dans cette partie et nous verrons qu'elles requièrent souvent un certain niveau d'abstraction pour décrire une molécule ou un ensemble de molécules.

Mesure	Expression mathématique	Intervalle
Coefficient de Tanimoto/Jaccard	$\frac{c}{a+b-c}$	0 à 1
Distance Euclidienne	$\sqrt{a+b-2c}$	0 à N
Distance de Manhattan	$a+b-2c$	0 à N
Indice de Sørensen-Dice	$\frac{2c}{a+b}$	0 à 1
Similarité cosinus	$\frac{c}{m}$	0 à 1
Coefficient de Russel-RAO	$\frac{cm}{ab}$	0 à 1
Coefficient de Forbes	$\frac{cm}{ab}$	0 à 1
Distance de Soergel	$\frac{a+b-2c}{a+b-c}$	0 à 1

Figure 9 Illustration des différentes distances utilisées pour calculer la similarité entre deux empreintes moléculaires (a indique le nombre de propriétés propres à la molécule A, b indique le nombre de propriétés propres à la molécule B, c le nombre de propriétés communes aux deux molécules et m =$\sqrt{ab}$)

3.1.1 Similarité d'empreinte moléculaire

Les empreintes moléculaires, ou *fingerprints*, sont les modèles de molécules les plus populaires et les plus développés. Il s'agit de chaînes de *bits* qui contiennent des informations (généralement 2D) sur la structure d'une molécule. Il existe des empreintes moléculaires basées sur des sous-structures clés (Figure 10), sur les différents chemins intramoléculaires (Figure 11), ou sur l'environnement de chaque atome (empreinte moléculaire circulaire)[145]. Les empreintes moléculaires basées sur des sous-structures (MACCS[146], PubChem fingerprints[147], BCI fingerprints[148], TGD et TGT fingerprints[149]) dépendent d'un ensemble de « clés structurales », ou sous-structures, répertoriées dans une liste propre à chaque outil. La longueur de la chaîne de *bits* dépend du nombre de clés de référence, et chaque *bit* renseigne sur la présence (1) ou l'absence (0) de la sous-structure (Figure 10). L'empreinte MACCS, qui fait partie des plus plébiscitées compte deux variantes, une avec 960 bits, l'autre avec 166 bits. C'est cette dernière qui est la plus souvent utilisée puisque les sous-structures qu'elle contient sont jugées suffisante pour décrire l'espace

chimique des molécules destinées au criblage virtuel[145]. Les modèles basés sur des chemins sont très similaires aux empreintes moléculaires basées sur des sous-structures, à la différence que les données contenues dans la chaîne de *bits* représentent des chemins intramoléculaires d'une

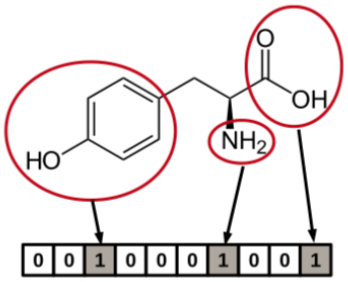


Figure 10 Exemple simplifié d'empreinte moléculaire basé sur des sous-structures à 10 bits. Les 3 bit-index correspondant aux groupements entourés en rouge prennent la valeur de 1. D'après [144]

longueur inférieure à une valeur limite imposée (Figure 11)[150]. Il arrive qu'un bit corresponde à plusieurs chemins, c'est ce que l'on appelle les collisions. Par exemple, les empreintes Daylight[151] possèdent 2048 bits qui énumèrent l'ensemble des chemins connectant au maximum 31 atomes (7 par défaut) d'une molécule. Les empreintes circulaires, elles, stockent des informations sur l'environnement de chaque atome jusqu'à un rayon prédéfini (Molprint2D[152,153], ECFP[154,155]).

La description de molécules sous forme d'empreinte 2D a l'avantage d'être très rapide et de ne nécessiter aucun calcul complexe pour la comparaison de molécules, notamment en

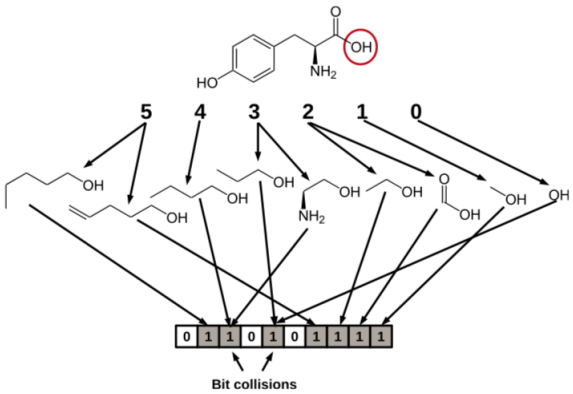


Figure 11 Exemple simplifié d'empreinte moléculaire basé sur des chemins à 10 bits. Les bit-index correspondant aux chemins intramoléculaires partant du groupement hydroxy- entouré de rouge prennent la valeur de 1. D'après [144]

s'affranchissant de l'étape de génération de conformation 3D et des erreurs de description qui peuvent apparaître avec la génération et la sélection de mauvaises conformations[156]. Une fois les empreintes moléculaires générées, elles sont comparées deux à deux via un calcul de distance. Parmi les différentes distances pouvant être mesurées (Figure 9), le coefficient de Tanimoto fait office de référence. Il représente le nombre de *bits* égaux à 1 communs à la molécule A et la molécule B divisé par le nombre de *bits* égaux à 1 des molécules A et B. Le coefficient de Tanimoto est donc compris entre 0 (aucune similarité entre les empreintes) et 1 (identité complète d'empreinte moléculaire).

3.1.2 Représentations abstraites

Les méthodes de recherche de similarité précédemment introduites partagent un même point commun : elles tendent à décrire de manière fidèle les configurations de molécules, c'est à dire leur enchaînement d'atomes et de liaisons. Ces méthodes permettent d'identifier rapidement des molécules partageant des structures similaires, en revanche 1) elles ne permettent pas d'identifier des molécules structurellement différentes mais présentant des propriétés physicochimiques communes pouvant leur conférer des activités similaires et 2) elles ne sont pas adaptées pour des études de *scaffold hopping*[157], c'est à dire de recherche de modification du squelette d'une molécule qui permettent de conserver ou d'améliorer son activité biologique. Il existe des méthodes plus abstraites pour se dégager de cette représentation formelle, par exemple, les méthodes de similarité de forme et les méthodes décrivant une molécule comme un ensemble de groupements fonctionnels représentés dans un espace à 2 ou 3 dimensions.

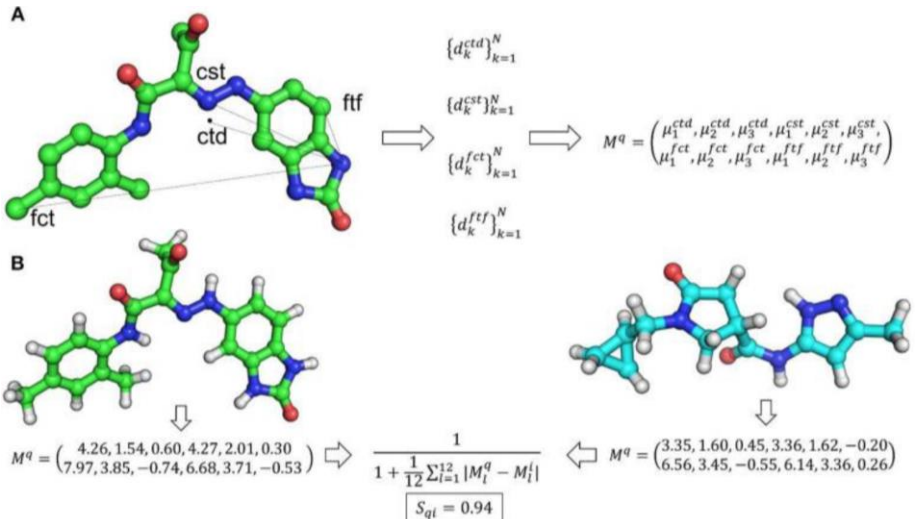


Figure 12 Schéma de la méthode Ultrafast shape recognition (USR). La distribution des distances entre chaque atome et 4 points de référence (le centroïde de la molécule (ctd), 2) l'atome le plus proche du centroïde (cst), l'atome le plus éloigné du centroïde (fct) et celui le plus éloigné du fct (ftf)) est calculée. La moyenne, la variance et l'asymétrie des 4 distributions sont mesurées et stockées dans un vecteur de 12 descripteurs. La similarité est ensuite estimée via la distance de Manhattan. D'après [152]

3.1.2.1 Similarité de forme

Les méthodes de similarité de forme se divisent en trois catégories : les méthodes 1) basées sur la distances interatomiques, 2) basées sur le volume et 3) basées sur la surface moléculaire.

Les méthodes basées sur les distances interatomiques décrivent une molécule comme un ensemble de distances entre 2 ou 3 atomes[158,159]. Ces méthodes s'affranchissent d'une étape d'alignement des molécules, ce qui la rend plus facilement applicable à de grands jeux de données. L'*Utrafast shape recognition* (USR)[160,161] est la méthode de distance atomique la plus populaire (Figure 12). Les distribution des distances entre chaque atome et 4 points de référence sont calculées : 1) le centroïde de la molécule (ctd), 2) l'atome le plus proche du centroïde (cst), l'atome le plus éloigné du centroïde (fct) et celui le plus éloigné du fct (ftf) (Figure 12)[158]. La moyenne, la variance et l'asymétrie des 4 distributions sont mesurées et stockées dans un vecteur de 12 descripteurs. La similarité est ensuite estimée via la distance de Manhattan (Cf 3.1).

Les méthodes basées sur le volume décrivent le volume d'une molécule comme l'union de sphères dites « dures » [162], correspondant au rayon de van der Waals de chaque atome, ou de sphères (ou densités) gaussiennes. Le chevauchement maximal entre les volumes de deux molécules est calculé

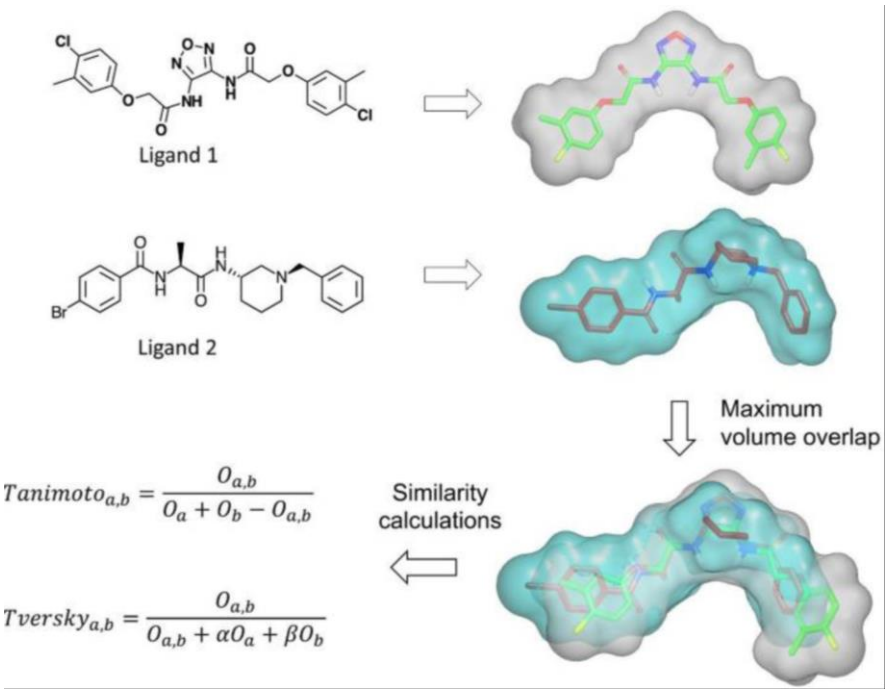


Figure 13 Illustration du calcul de la similarité de forme selon le logiciel ROCS. Deux molécules sont superposées de sorte à maximiser le recouvrement des volumes. Une distance de Tanimoto ou de Tversky permet de quantifier la similarité (O_a et O_b correspondent aux volumes non chevauchant de la molécule a et b respectivement, Oa,b représente le volume chevauchant). D'après [99]

pour estimer la distance volumique entre ces molécules. Le très utilisé algorithme de *Rapid Overlay*

of Chemical Structures (ROCS), par exemple, utilise des sphères gaussiennes de même rayon et ignore les atomes d'hydrogènes pour estimer le volume d'une molécule (Figure 13). Cette méthode nécessite un alignement des molécules afin d'atteindre un chevauchement maximal de leurs volumes. La distance entre molécules est ensuite calculée grâce au coefficient de Tanimoto – qui dans ce cas correspond au volume de chevauchement total divisé par la somme des volumes de chaque molécule moins le volume de chevauchement – et la distance de Tversky, très proche du coefficient de Tanimoto, qui correspond au volume de chevauchement total divisé par le volume de chevauchement auquel s'additionnent les volumes non chevauchés de chaque molécule. L'algorithme ROCS intègre également des données physicochimiques (donneur de liaison

Figure 1

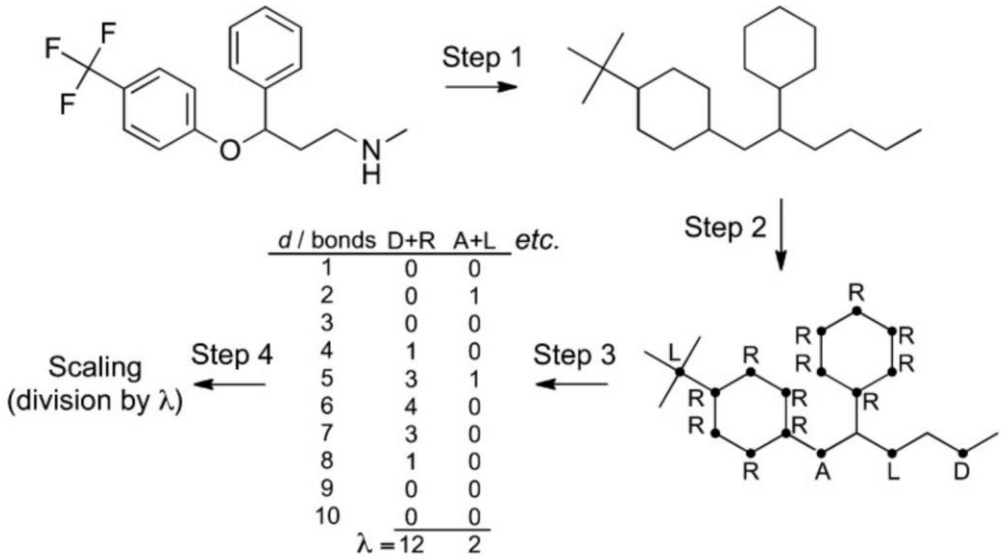


Figure 14 Illustration de la génération de descripteur CATS. 1) La structure moléculaire est réduite à un graphe moléculaire ; 2) à chaque atome est assigné des types de fonction (R= aromatique, L = lipophile/hydrophobe, A = accepteur de liaison hydrogène, D = donneur de liaison hydrogène) ; 3) les paires d'atomes espacées de *d* liaisons sont comptées et 4) chaque valeur obtenue est divisée par le nombre de paires observées toutes distances confondues. La matrice obtenue correspond au descripteur CATS. D'après [158]

hydrogène, accepteur de liaison hydrogène etc.) pour améliorer l'alignement et la recherche de similarité.

Les méthodes basées sur la surface moléculaires reposent sur l'hypothèse que des molécules de surfaces similaires peuvent avoir des propriétés physiques et biologiques similaires. Les surfaces

moléculaires sont comparées à l'aide de descripteurs de surface parmi lesquels, l'IDSS (Inner Distance Shape Signature)[163] qui présente l'avantage de ne pas être sensible à la flexibilité des molécules.

3.1.2.2 Représentations 2D des groupements fonctionnels

Les méthodes de représentation 2D des groupements fonctionnels s'affranchissent de la connectivité intramoléculaire. Ce mode de représentation considère comme équivalent l'ensemble des groupements chimiques présentant des propriétés physicochimiques similaires. Par exemple, une chaîne aliphatique et un groupement trifluorométhyl partagent tous deux des propriétés hydrophobes. Ce concept a été introduit en 1999 par Schneider et al. avec le descripteur Chemically Advanced Template Search (CATS)[164]. Les descripteurs CATS proposent une représentation topologique d'une molécule qui consiste en une représentation simplifiée de la molécule en « graphe moléculaire » auquel des propriétés fonctionnelles sont assignées (Figure 14). Certains

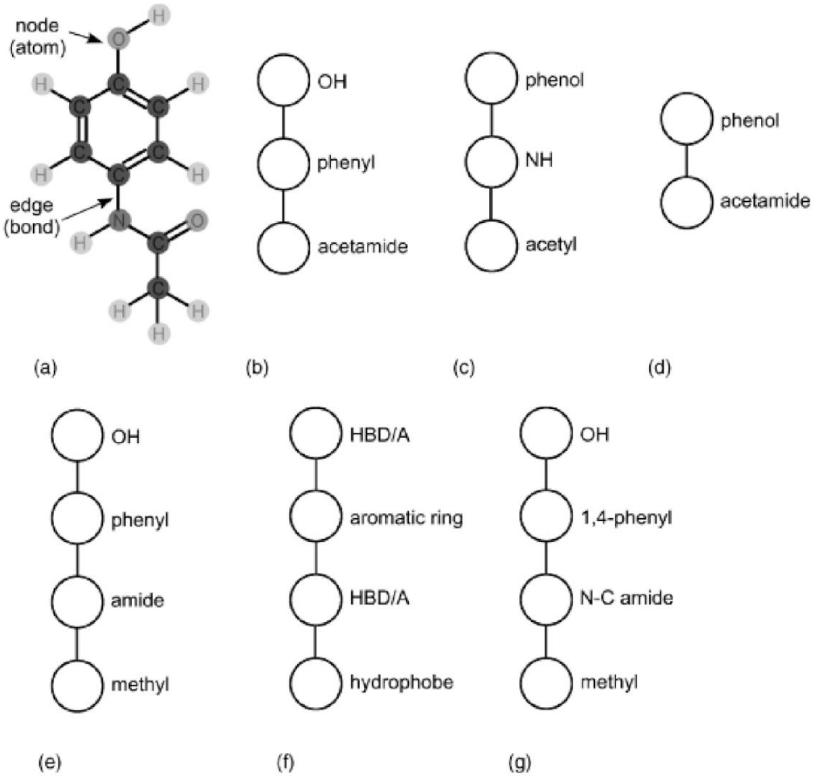


Figure 15 Représentations du paracétamol sous forme de graphe réduit. Des nombres différents de nœuds peuvent être obtenus en fonction de la nature et du niveau de détail encodé dans chaque nœud. D'après [151]

atomes n'apparaissent pas dans le descripteur final lorsqu'ils ne sont associés à aucune propriété fonctionnelle ou lorsqu'ils sont compris dans la propriété associée à l'atome voisin. Une approche similaire consiste à utiliser des graphes réduits pour décrire une molécule. La réduction de graphes consiste à condenser des sous-structures en des nœuds individuels tout en conservant la connectivité entre sous-structures[157]. Différentes représentations en graphes réduits du paracétamol sont illustrées Figure 15.

L'Extended reduced Graph (ErG) développé par Stiefl et al[165] est un exemple de graphe réduit. La procédure de construction du graphe est la suivante : les accepteurs et donneurs de liaison hydrogène ainsi que les groupements chargés sont encodés, les centroïdes des cycles sont positionnés, les atomes de fusion inter-cycles sont conservés, et chaque groupement retenu est connecté aux centroïdes via le chemin le plus court. Les atomes des cycles non impliqués dans les chemins les plus courts et dans les zones de fusion cycle-cycle sont éliminés. Le graphe obtenu est une représentation générique de la molécule initiale et permet de reconnaître une plus grande diversité de molécules que les méthodes faisant appel aux empreintes 2D. La méthode plus populaire est la méthode appelée Feature Tree[166]. Elle permet d'obtenir un graphe réduit avec deux informations supplémentaires ajoutées à chaque nœud : le nombre d'atome qu'il contient, ainsi que son volume moyen approximé (Figure 16).

Les représentations simplifiées ainsi obtenues (graphes moléculaires et graphes réduits), peuvent être utilisées comme intermédiaires pour déduire une empreinte (*fingerprint*) qui est utilisée pour

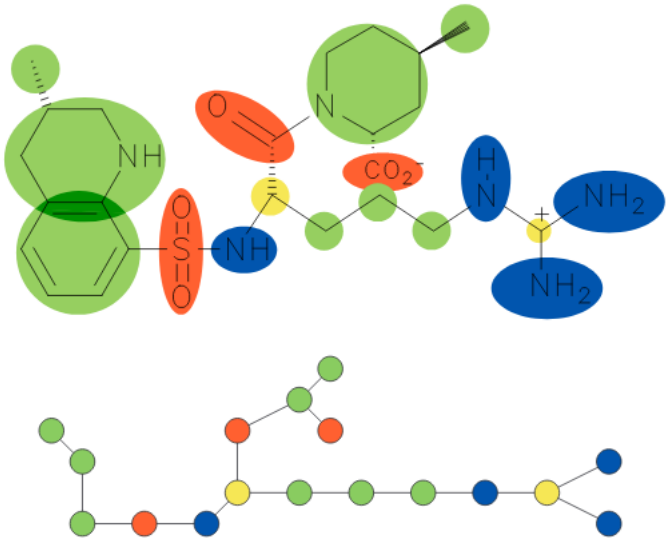


Figure 16 Exemple de Feature Tree. Les cercles de couleurs montrent les atomes qui sont condensés en un nœud. Le Feature Tree en découlant est représenté en dessous avec les propriétés correspondant aux nœuds : hydrophobe (vert), donneur de liaison hydrogène (rouge), accepteur de liaison hydrogène (bleu), pas d'interaction directe (jaune). D'après [161]

le calcul de distance (à l'aide du coefficient de Tanimoto ou d'autres métriques – comme dans CATS et ErG) ou être utilisées directement pour calculer la similarité entre deux molécules (Feature Tree[166,167] ou DeCAF[168]). Les premières évaluations de l'algorithme de Feature Tree ont montré qu'il était aussi performant sinon plus que l'empreinte moléculaire DAYLIGHT pour l'identification de *hits*. Les deux approches ont permis de retrouver environ 50% de *hits* communs sur 5 cas d'études, Feature Tree permettant de retrouver des molécules structurellement plus diverses. Ces résultats montrent la complémentarité des approches dans la recherche de *hits*.

3.1.3 Modélisation de pharmacophores 3D

Un pharmacophore est défini selon l'IUPAC comme un « ensemble d'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 »[169]. En d'autres termes, le pharmacophore d'une molécule est un ensemble de propriétés physicochimiques localisées dans l'espace qui s'affranchit de toute notion de connectivité, et qui confère à une molécule sa capacité

à interagir avec une cible et à moduler son activité. Identifier ces points clés représente le Graal non seulement dans la recherche de médicaments mais aussi pour prédire tous les phénomènes impliquant une interaction ligand/protéine (ex : effets secondaires, action de perturbateurs endocriniens). Du point de vue de la chémoinformatique, il est possible de modéliser un pharmacophore en s'appuyant sur un ensemble de ligands connus d'une protéine qui présente le même profil pharmacologique[170] (ex : agoniste ou antagoniste) ou bien à partir du site de liaison connu de la cible (Cf 3.1.3 et 4.4.1). La connaissance de plusieurs ligands d'une cible peut permettre d'identifier un arrangement spatial de groupements fonctionnels communs responsables de l'interaction. La quantité et diversité des données de départ sont néanmoins à étudier avec attention ; elles peuvent apporter ou faire perdre de l'information. La perte d'information liée à la diversité des molécules peut être liée au fait que différentes molécules peuvent avoir un mode de liaison différent tout en déclenchant une activité similaire. C'est pour cela que la plupart des études de pharmacophore débutent avec un alignement et un *clustering* des molécules similaires. S'ensuit l'attribution de points pharmacophoriques à chaque molécule puis l'identification d'un modèle de pharmacophore commun qui constitue une étape clé et peu triviale : il s'agit de déterminer quel(s) modèle(s) de pharmacophore concorde(nt) au mieux avec les molécules actives connues. En général, ceci revient à déterminer un modèle de pharmacophore qui contient un maximum de points pharmacophoriques communs aux molécules actives lorsqu'elles sont dans des conformations de basse énergie[171]. L'étape finale consiste à associer un score à chaque modèle de pharmacophore généré. Le ou les modèles de pharmacophore retenus servent de base au criblage de chimiotèques. Il existe de nombreux logiciel qui automatisent l'élucidation d'un modèle pharmacophorique (DiscoveryStudio[172], LigandScout[173,174,175], MOE[176], PHASE[177,178], Pharmer[179], PharmaGist[180], QUASI)

3.1.3.1 Préparation des molécules de départ

La plupart des logiciels de génération de modèle de pharmacophore utilisent comme molécules de référence des ligands présentant une activité validée expérimentalement pour la cible. Il est impératif de s'assurer que les ligands pris en compte agissent au niveau d'un même site de liaison. Généralement, les molécules fournies possèdent des activités comparables pour la cible biologique étudiée ; dans ce cas le criblage du modèle de pharmacophore finalement généré aboutira à un classement binaire (active/inactive) des molécules criblées. Certains outils comme PHASE[178] offrent la possibilité de fournir des données d'activité, auquel cas un modèle de relation structure

activité 3D sera généré. Certains modèles acceptent également des molécules inactives dans le jeu d'apprentissage. Par exemple l'algorithme HipHop de Discovery Studio génère des modèles basés sur les molécules actives, aligne les molécules inactives sur le modèle généré, puis ajoute des volumes d'exclusion de sorte à exclure les molécules inactives. L'algorithme HypoGen de Discovery Studio quant à lui génère des modèles de pharmacophores basés sur les molécules actives et exclut tous ceux sur lesquels peuvent s'aligner des molécules inactives du jeu d'apprentissage.

Idéalement, un modèle de pharmacophore doit être construit à partir des conformations bioactives de molécules. Dans la plupart des cas, cette conformation n'est pas disponible et il faut générer des conformations biologiquement plausibles. De nombreux outils permettent de générer ces conformations (Cf 2.2.2) et plusieurs études montrent qu'ils proposent des poses proches des modes de liaison observés en cristallographie pour la plupart des ligands testés[109,111]. Si certains logiciels fonctionnent à partir de conformations 3D préalablement générées (HipHop, PHASE, LigandScout), d'autres les génèrent pendant la phase de génération du modèle de pharmacophore en utilisant des algorithmes génétiques (GASP, GALAHAD[181]).

3.1.3.2 Représentation des points pharmacophoriques

Chaque logiciel possède sa propre définition d'un point pharmacophorique qui peut être cruciale pour assurer la performance finale du modèle généré. Ces points pharmacophoriques possèdent un centre et une sphère de tolérance (Figure 17)[171] ; ils contiennent 3 types d'information : le type de

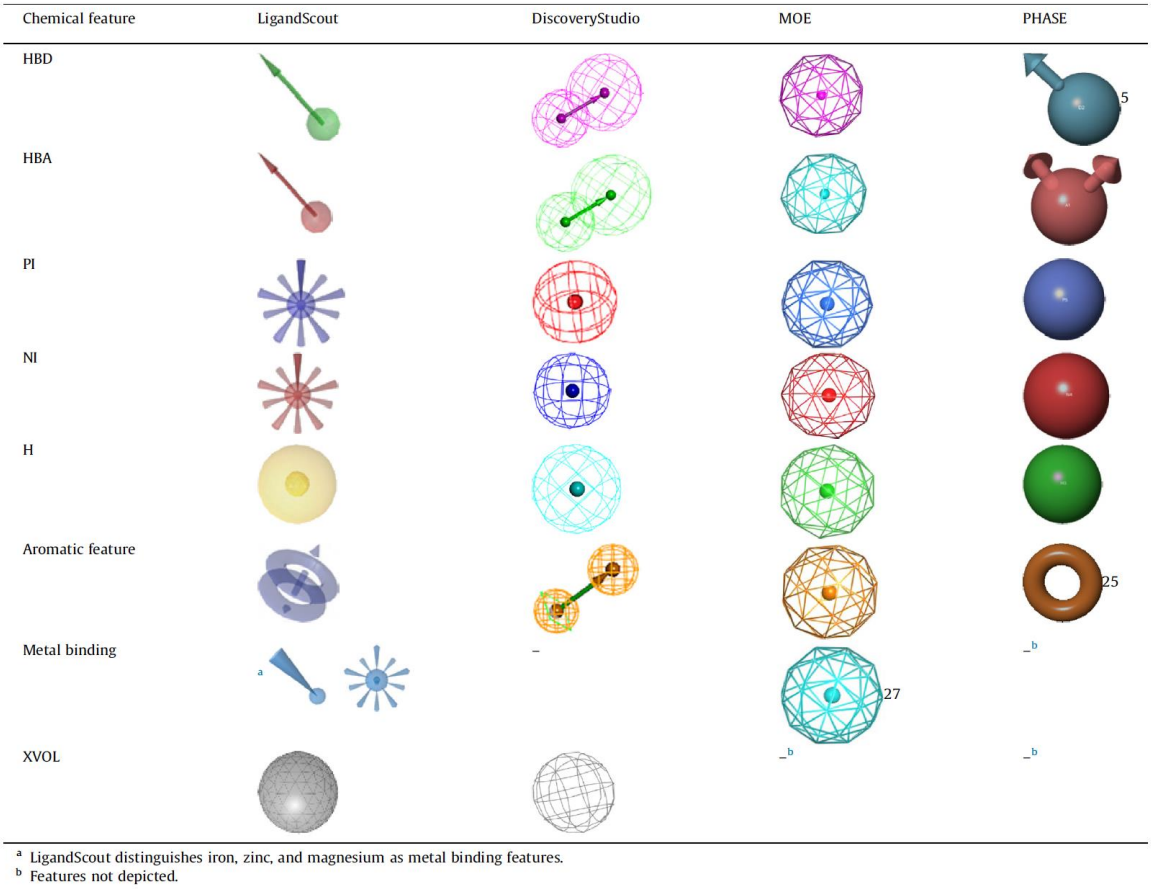


Figure 17 Illustration des points pharmacophoriques implémentésdans LigandScout, DiscoveryStudio, MOE et PHASE. D'après [165]

point pharmacophorique, sa position (coordonnées xyz) et un poids qui informe sur l'importance du point pharmacophorique[171]. Dans le cas des donneurs et accepteurs de liaison hydrogène un vecteur et une zone de tolérance indiquent généralement la direction du groupement complémentaire (LigandScout, DiscoveryStudio, PHASE). Il en est de même pour cycles aromatiques qui sont parfois représentés sous forme de tore (ou donut) avec un vecteur de direction. Le poids associé à chaque point pharmacophorique peut être calculé en fonction de la proportion de molécules actives qui le décrit et est utilisé au moment du criblage pour l'attribution d'un score.

3.1.3.3 Génération de modèles de pharmacophore

Une fois les points pharmacophorique assignés à chaque molécule de référence, il reste à identifier les points qui sont communs à plusieurs voire toutes les molécules actives. Pour ceci on applique un procédé appelé *mapping* de pharmacophore dont l'objectif est d'identifier des sous-structures communes maximales (MCS). Les techniques suivantes sont communément utilisées pour le *mapping* de pharmacophore[157] :

- La recherche systématique avec contraintes explore de manière systématique l'espace conformationel des molécules : dans un premier temps, la molécule la plus rigide est sélectionnée et son espace conformationel est exploré ; c'est pendant cette phase que les distances entre chaque paire de points pharmacophoriques sont enregistrées. L'espace de la seconde molécule la plus rigide est ensuite exploré en prenant en compte les contraintes de l'étape précédente, ce qui limite l'exploration des angles de torsion. Ainsi, lorsque l'algorithme arrive aux molécules les plus flexibles, un nombre limité de torsion est autorisé, ce qui facilite l'exploration conformationelle.
- La détection de clique, qui correspond à un sous-graphe connecté, fonctionne en 3 grandes étapes : 1) toutes les paires de points pharmacophoriques possibles entre deux molécules sont générées, 2) une arête est créée entre toutes les paires pour lesquelles la distance entre les points pharmacophoriques est la même dans les deux molécules, 3) chaque ensemble de paires pleinement connectées (chaque paire doit être connectée à au moins deux autres paires) défini une clique qui correspond à un pharmacophore commun. DISCO utilise cette approche et part des molécules les moins flexibles pour débuter sa recherche de clique.
- La méthode de ressemblance maximum (*maximum likelyhood*) consiste à prendre tour à tour chaque molécule active comme molécule de référence ainsi que ses conformations préalablement générées et à identifier l'arrangement spatial de ses points pharmacophoriques le plus fréquemment retrouvé. De nombreux modèles de pharmacophores sont ainsi générés et un score leur est attribué en fonction de la quantité de molécules *mappées* et de la rareté du modèle pharmacophorique. Selon les méthodes utilisées une sphère de tolérance plus ou moins large permet d'être plus ou moins stricte dans le *mapping* de pharmacophore.

- Les algorithmes génétiques consistent à générer des populations de solutions potentielles qui évoluent vers de meilleures solutions. Une conformation est aléatoirement générée pour chaque molécule et chaque angle de torsion est stocké dans un byte (8 bits - correspondant à des valeurs d'entiers comprises entre à et 255), l'ensemble des bytes formant un « *chromosome* ». L'ensemble des molécules est superposé à la molécule présentant le moins de points pharmacophoriques et un score est attribué. Ce score dépend généralement du nombre de points pharmacophoriques similaires alignés, du recouvrement en terme de volume, et de l'énergie de van der Waals des conformations (GASP, GAMMA)[182]. A chaque itération, de nouvelles conformations sont proposées via des cross-over (échange de bytes entre deux chromosomes) et des mutations (modification d'un bit). Un nombre défini de cross-over a lieu à chaque itération, chaque conformation mère étant choisie aléatoirement avec une chance de tirage proportionnelle au score auquelle celle-ci est associée. Ainsi les meilleures conformations ont plus de chance de donner de nouvelles conformations. L'algorithme s'arrête après un nombre défini de cycles ou bien lorsqu'il converge vers une conformation qu'il n'arrive pas à optimiser davantage.

Les différents outils qui intègrent ces méthodes ne proposent généralement pas une mais plusieurs solutions. A chaque solution est attribué un score qui permet de les classer entre elles. Ce score dépend du nombre et de la qualité de superposition des points pharmacophoriques des ligands de référence, des volumes de recouvrement, et de la rareté du modèle de pharmacophore de sorte à ce qu'un modèle de pharmacophore très générique et peu spécifique des ligands de référence ne pas soit privilégié. Les algorithmes HipHop (DiscoveryStudio) et PHASE (Schrödinger) prennent en compte cette rareté. Il faut noter que lorsque beaucoup de molécules sont utilisées comme point de départ, il est possible que le modèle de pharmacophore résultant possède peu de points pharmacophoriques du fait de la difficulté à satisfaire l'ensemble des ligands de référence. Pour éviter ce genre de situation, il existe des logiciels qui permettent de générer des modèles de pharmacophore qui prennent en compte la fréquence d'observation d'un point pharmacophorique chez les molécules de référence. Par exemple LigandScout oppose aux pharmacophores partagés (« *shared* » - qui correspondent aux points pharmacophoriques communs aux molécules de référence), les pharmacophores fusionnés (« *merged* ») qui regroupent l'ensemble des points

pharmacophoriques observés au delà d'une certaine fréquence, les points les plus fréquents étant associés aux poids les plus forts[173]. La plupart des outils permettent également une customisation manuelle, ce qui permet d'intégrer l'expertise du chercheur à la conception d'un modèle de pharmacophore.

3.1.3.3.1 Criblage de chimiothèques

Une fois un modèle de pharmacophore généré, des collections de composés sont criblées afin d'identifier des *hits,* autrement dit des molécules présentant les mêmes propriétés pharmacophoriques et par conséquent, ayant des chances d'avoir une affinité pour la cible étudiée[182].

Il est extrêmement important de prendre en compte la flexibilité des molécules lors de l'étape de criblage : la prise en compte d'une seule ou de peu de conformations d'une molécule peut être insuffisante pour observer un alignement représentatif de la conformation bioactive de la molécule. Deux stratégies sont utilisées en ce sens, la première consistant à générer préalablement les conformations 3D des molécules (LigandScout[173,175], CATALYST[183]), l'autre consistant à prendre en compte la flexibilité pendant le criblage en utilisant le modèle de pharmacophore pour guider la génération d'une conformation (UNITY, CFS 167, 3DFS 168). Dans le cas de la génération préalable de conformation, il est important de choisir les bons paramètres pour s'assurer que la conformation optimale a été générée. Par exemple, Poli et al. ont montré qu'il est préférable de générer un minimum de 50 conformations pour maximiser les chances d'obtenir une conformation proche de la conformation bioactive (<2 Å) avec iCon[115].

La phase de criblage s'effectue généralement en deux temps. Dans un premier temps, seules les molécules possédant le même nombre et type de points pharmacophoriques que le modèle de pharmacophore sont conservées, sans nécessité d'alignement. Il s'agit d'une étape rapide de filtrage qui permet d'éliminer rapidement des molécules assurément inadéquates avant l'étape suivante du criblage qui est plus complexe et chronophage. Il est généralement possible d'accepter qu'un ou plusieurs points pharmacophoriques soient omis dans les molécules criblées, auquel cas le nombre total de points pharmacophorique moins le nombre d'omissions possibles est utilisé pour ce filtrage primaire. La seconde étape consiste à identifier les molécules ayant passé le premier filtre qui satisfont les critères du pharmacophore[184]. Pour ceci, des méthodes d'alignements sont utilisées, comme la recherche de clique ou la méthode de ressemblance maximale explicitées précédemment. Dans certains cas, la recherche s'arrête dès lors qu'un conformère répond à l'ensembles des

contraintes du modèle de pharmacophore, dans d'autres cas, l'ensemble des conformères est évalué afin de trouver celui qui présente la meilleure superposition. Des scores sont attribués aux conformères acceptés pour les classer entre eux, puis parmi les autres molécules criblées. Il existe deux types de scores : ceux qui se basent sur un calcule de distance entre les groupements chimiques de la molécule et le centre du point pharmacophorique correspondant (« RMSD »), et ceux qui prennent en compte le rayons de van der Waals des atomes en question et du points pharmacophorique (« recouvrement »)[171] (Tableau 5). Une molécule est rejetée lorsqu'aucun de ses conformères ne répond aux contraintes du modèle de pharmacophore.

La recherche de pharmacophores permet donc de discerner les propriétés communes aux ligands de référence et par conséquent, d'identifier les propriétés d'intérêt qui expliquent l'interaction entre une molécule et une cible donnée. Les modèles de pharmacophores peuvent donc être utilisés pour prédire la capacité d'autres molécules à interagir avec la même cible biologique, que ce soit pour identifier de nouvelles molécules (*hits*) thérapeutiques éventuellement structurellement différents des ligands connus (ce qui peut ouvrir la voie à des optimisations de nouvelles séries chimiques dans le cadre de la recherche de médicament) ou pour identifier les composés interagissant avec la cible de manière non désirée (ex : effets secondaires, perturbateurs endocriniens, etc.). Généralement, les modèle de pharmacophores ne permettent pas de détailler la contribution d'une propriété physicochimique à l'activité d'une molécule, contrairement aux modèles QSAR.

Tableau 5 Liste des outils de génération de modèles de pharmacophores. D'après [171]

	Méthode d'attribution de score	Fournisseur	Type d'outils
DiscoveryStudio (HipHop/ HypoGen/ HypoGenRefine)	Recouvrement	Biovia (anciennement Accelrys)	Logiciel
LigandScout (espresso)	Recouvrement	Inte:Ligand	Logiciel
MOE	RMSD	Chemical Computing Group	Logiciel
PHASE	RMSD	Schrödinger	Logiciel
GASP	Recouvrement	Tripos (Certara)	Logiciel
DISCOTech	RMSD	Tripos (Certara)	Logiciel
Pharmer	RMSD	Camacho Lab	Logiciel (gratuit)

ZINCPharmer	RMSD	Camacho Lab	Outils en ligne
PharmaGist	Recouvrement	Tel Aviv University	Outils en ligne
QUASI	Recouvrement	DeNovo Pharmaceuticals	Logiciel

3.2 Méthodes QSAR

3.2.1 Généralités

De manière générale les méthodes de relation structure-activité (QSAR) ont pour objectif d'associer un ensemble pondéré de propriétés géométriques et physicochimiques de molécules (encodées sous forme de descripteurs) à leurs activités biologiques (ou propriétés chimiques), et de l'utiliser pour prédire les activités biologiques (ou propriétés chimiques) d'un ensemble de molécules pour lesquelles nous ne disposons pas de données expérimentales. Il existe une multitude de méthodes QSAR allant du 0D au 4D (Tableau 6). Ces méthodes reposent toutes sur le même principe : elles sont des applications de méthodes mathématiques ou statistiques qui visent à identifier la relation $Pi = \hat{k}(D_1, D_2, \ldots, D_n)$ où Pi est l'activité biologique, $D_1, D_2, \ldots, D_n$ sont des propriétés calculées ou mesurées expérimentalement (descripteurs) et $\hat{k}$ est une transformation mathématique empirique qui doit être appliquée à l'ensemble des descripteurs pour permettre de calculer l'activité biologique des molécules étudiées[185] (Figure 18).

Tableau 6 Liste des différentes méthodes QSAR. D'après [185]

Dimensions	Méthodes QSAR	Refs.
0D-QSAR	Modèles basés sur les descripteurs issus de la formule moléculaires des molécules (poids moléculaire etc.)	[186]
1D-QSAR	Modèles basés sur les propriétés physicochimiques des structures moléculaires (solubilité, hydrophobicité etc.)	[187]
2D-QSAR	Activité corrélée avec des motifs structuraux (topologie des molécules) sans considération de la structure 3D des molécules	[188]
3D-QSAR	Activitée corrélée avec des propriétés structurales tridimensionnelles des ligands.	[189,190,191]

4D-QSAR	Plusieurs configurations d'un même ligand sont considérées	[192,193,194,195]
5D-QSAR	= 4D-QSAR + représentation explicite de modèles induced-fit	[196]
6D-QSAR	= 5D-QSAR + considération de plusieurs modèles de solvatation	[197]

Parmi les différences entre les nombreuses méthodologies QSAR existantes, la première réside dans le type d'activité/propriété cible (ou variable dépendante) qui représente un critère majeur dans le choix de la démarche à adopter et des descripteurs à sélectionner[185]. Cette activité/propriété cible peut être de type 1) quantitative – c'est le cas des données d'activité ou d'affinité de type IC_{50} ou Ki –, 2) qualitative continue – les molécules peuvent être classées active ou inactives ou encore stable/modérément stable/instable dans le cas de la stabilité métabolique –, ou 3) qualitative discontinue – c'est le cas des classes pharmacologiques et du caractère *druglike* par exemple. Les variables quantitatives orientent vers l'utilisation de modèles de régression linéaire, alors que les variables qualitatives orientent vers l'utilisation de méthodes de classification. Les descripteurs utilisés (variables indépendantes) varient aussi selon la méthode QSAR utilisée. Ils peuvent eux aussi être quantitatifs continus (ex : poids moléculaire, cLogP, volume etc.) ou quantitatifs discrets (nombre de groupement fonctionnels, nombre de liaisons rotatives etc.) et ils peuvent être issus de représentations 2D ou 3D des molécules, donnant lieu à des QSAR respectivement 2D et 3D. Dans le cas de recherche de médicament et de la prédiction de la toxicité, il est important de comprendre les descripteurs utilisés afin d'en tirer des informations judicieuses pour optimiser des séries chimiques.

Le déroulement d'une étude QSAR débute généralement par une étape de « nettoyage » des descripteurs utilisés : sont ainsi rejetés 1) les descripteurs qui n'apportent aucune information du fait de leur valeur constante pour toutes les molécules de référence (écart type nul) et 2) les descripteurs moléculaires fortement corrélés à d'autres descripteurs plus informatifs (ex : descripteur plus compréhensible, ou descripteur représentatif d'un ensemble de descripteurs corrélés).

On applique ensuite des méthodes permettant d'établir une corrélation entre l'ensemble des

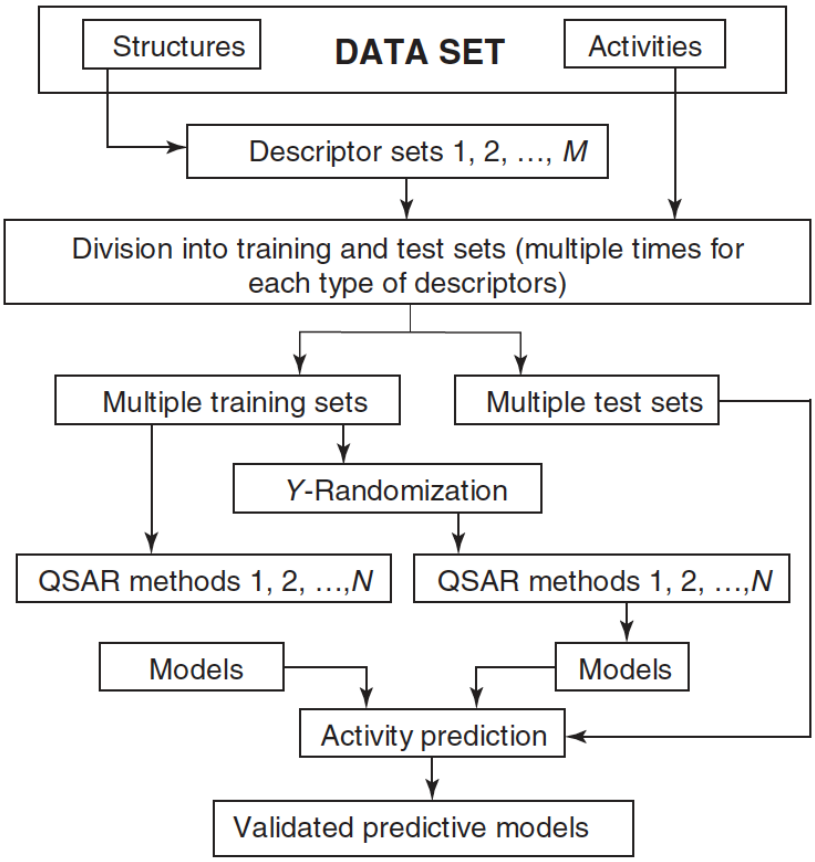


Figure 18 Schéma général d'un protocole QSAR. La première étape consiste à diviser les données initiales en un jeu de données d'apprentissage et un jeu de données tests. Des modèles sont ensuite générés et appliqués au jeu de données tests pour être évalués et éventuellement validés. Parfois, les descripteurs associés aux molécules sont aléatoirement redistribués afin de comparer la performance d'un modèle de aléatoirement générés aux autres modèles générés. Daprès [179]

descripteurs sélectionnés et les activités biologiques. Ces méthodes se divisent en deux catégories : les méthodes linéaires (régression linéaire (LR), régression des composantes principales (PCR), ou régression des moindres carrés partiels (PLS)) et les méthodes non linéaires (les k plus proches voisins (kNN), le partitionnement récursif (RP), les réseaux de neurones artificiels (ANNs), les arbres de décisions (DT), les forêts d'arbres de décision ou les machines à vecteur de support (SVMs))[185]. Généralement, la nature des données biologiques (variable dépendante), le choix des

descripteurs moléculaires (variables indépendantes) et le choix de la technique appliquée sont interdépendants.

3.2.2 Descripteurs moléculaires

D'après A. Tropsha, les 3 éléments clés pour construire un bon modèle QSAR sont : les descripteurs, les descripteurs et les descripteurs[185]. Il existe une vaste variété de descripteurs pouvant être utilisés[198] :

- Les descripteurs géométriques comme la surface moléculaire, le volume, la surface accessible au solvant, le moment d'inertie etc.
- Les descripteurs constitutionnels tels que le nombre et le type d'atomes, de liaisons, de groupes fonctionnels etc.
- Les indices topologiques qui sont basés sur les graphes moléculaires. On retrouve notamment l'index de Wiener qui renseigne sur la somme des liaisons entre tous les nœuds du graphe moléculaire, ou encore l'indice de Randic qui renseigne sur la connectivité des nœuds.
- Les descripteurs physicochimiques tels que le poids moléculaire, le nombre de donneurs/accepteurs de liaison hydrogène, le coefficient de partition octanol/eau etc.
- Les descripteurs électrostatiques comme la charge atomique.
- Les descripteurs issus de la chimie quantique qui se basent sur les orbitales atomiques des atomes de chaque molécule et leurs propriétés, les plus fréquemment utilisés étant les charges atomiques, les énergies d'orbitales moléculaires, les densités d'orbitales frontières, la polarisabilité, les indices de moments dipolaires et de polarité.
- Les empreintes moléculaires qui encodent l'absence ou la présence de structures d'intérêt comme détaillé précédemment (Cf 3.1.1)

Des logiciels comme CDK[199], AFGen[200], Vlife[201]et Dragon[202] permettent de générer plusieurs milliers de descripteurs parmi ces catégories. Un bon choix de descripteurs est impératif pour construire le modèle idéal. Lorsque trop peu de descripteurs sont sélectionnés, il y a un fort risque d'omettre soit le ou les descripteurs essentiels, soit celui qui permet de parfaire le modèle. A contrario, si trop de descripteurs sont pris en compte, il y a un risque d'identifier des corrélations

parasites qui ne sont pas porteuses d'information intelligible. Il convient également de sélectionner des descripteurs qui sont adaptés à la qualité et la quantité des molécules de référence. Des descripteurs 3D n'apportent aucun intérêt si la conformation 3D bioactive d'une molécule n'est pas correctement caractérisée. En revanche, contrairement aux descripteurs 2D, les descripteurs 3D rendent possible la prise en compte de la stéréochimie des molécules. La méthode CoMFA est la plus populaire des méthodes de QSAR 3D : elle repose sur le postulat que les différences d'activité biologique entre les molécules s'expliquent souvent par des différences dans la forme et la force des champs d'interactions non covalentes entourant les molécules[203]. Ainsi les molécules étudiées sont virtuellement alignées et placées dans une grille. Des sondes correspondant à des carbone sp3 et chargées +1 sont positionnées à chaque point de la grille[189] et permettent de calculer les énergies de van der Waals et électrostatiques entre chaque point de la grille et les atomes de chaque molécule. Les énergies calculées sont ensuite insérées dans une grille en tant que descripteurs, et servent de base à la construction d'un modèle. Ces descripteurs nécessitent un alignement des molécules et sont peu adaptables à des molécules structurellement différentes.

3.2.3 Validation des modèles QSAR

Les modèles QSAR sont finalement évalués sur les jeux de données d'apprentissage et de tests pour estimer leur pertinence et les utiliser ou non en tant qu'outils de prédiction de l'activité d'autres molécules. Dans un premier temps, le modèle est évalué sur le jeu d'apprentissage lui-même. Plusieurs métriques sont utilisées : la sensibilité, la spécificité, l'exactitude et la précision pour les méthodes de classification, l'erreur absolue moyenne (MAE), l'erreur quadratique moyenne (RMSE), le coefficient de détermination (R^2) pour les méthodes de régression. De nombreuses revues détaillent ces métriques[204,205]. Des méthodes de cross-validation sont utilisées pour vérifier la robustesse des résultats obtenus. Ceci consiste à retirer une ou plusieurs molécules du jeu de données d'apprentissage et de contrôler les performances sur ces sous-ensembles. Le Leave-One-Out (LOO), qui consiste à retirer une molécule du jeu d'apprentissage, le Leave-many-Out, qui consiste à en retirer plusieurs, et le Bootstrap, qui consiste à ré-échantillonner le jeu d'apprentissage par le biais de tirages avec remise sont les plus utilisés. Les modèles sont ensuite appliqués aux jeux de tests voire à des jeux de données externes pour validation. Un modèle idéal doit montrer des performances comparables sur le jeu d'apprentissage, en cross-validation et sur le jeu de test. Il faut néanmoins s'assurer que les molécules utilisées dans le jeu d'apprentissage et le

jeu de test couvrent un espace chimique relativement comparable aux molécules pour lesquels nous souhaitons prédire l'affinité, afin de rester dans le domaine d'applicabilité du modèle.

4 Criblage virtuel basé sur la structure

Les méthodes de criblage virtuel basées sur la structure s'appuient sur les informations apportées par la structure tridimensionnelle de la cible étudiée pour prédire des interactions possibles avec une petite molécule. Les structures peuvent être obtenues expérimentalement grâce à des méthodes comme la cristallographie aux rayons X, la Résonance Magnétique Nucléaire (RMN) et la cryo-Microscopie Électronique (cryo-EM). Lorsque la structure n'a pas été résolue expérimentalement, elle peut être modélisée via différentes techniques *in silico* qui seront expliquées dans ce chapitre. Certaines protéines possèdent des régions qui n'adoptent pas de conformation stable ; on appelle ceci des régions désordonnées. Ces régions sont plus ou moins longues et jouent un rôle majeur dans les interactions avec d'autres partenaires protéiques. L'existence de régions désordonnées à proximité du site de liaison de petites molécules rend peu propice le criblage virtuel. En effet, il est difficile d'identifier une conformation bioactive ponctuelle à partir de laquelle le criblage pourrait être mené et ce malgré les progrès considérables liés au couplage de méthodes expérimentales (SAXS et RMN) et de méthodes de dynamique moléculaire[206]. Dans ce chapitre, nous nous concentrerons sur les protéines possédant une structure secondaire stable au niveau de leur(s) site(s) de liaison. Une étape clé du criblage basé sur la structure est l'identification du ou des sites les plus propices à accueillir un ligand. L'incapacité d'une protéine ciblée à interagir avec la molécule testée est un des facteurs responsables des ~60% d'échecs observés[207] en phase d'identification ou d'optimisation de lead, un autre facteur étant la faible implication de la cible protéique dans le développement de la maladie[17]. Les sites d'interactions entre une petite molécule et une protéine se caractérisent le plus souvent par des cavités hydrophobes à la surface de la protéine qui peuvent se retrouver enfouies. Plusieurs méthodes d'identification de ces sites et d'estimation de leur *druggabilité* ont émergé et seront présentées dans ce chapitre. Il arrive que ces sites de liaison adoptent des conformations différentes en fonction de leur environnement[208,209], de la présence de co-activateurs/ de co-répresseurs[210], ou des propriétés du ligand[211]. Les changements conformationnels peuvent avoir lieu à une échelle locale (mouvement d'un ou plusieurs résidus qui ne modifie pas la structure globale du site de liaison) ou bien à une échelle globale, auquel cas on observe des changements de conformation du squelette de la protéine ou du domaine protéique. Il

est important de considérer cette flexibilité locale ou globale pour conduire des études rationnelles de criblage virtuel. De même qu'il est important de prendre en compte le solvant dans lequel se trouvent la protéine et la petite molécule pour prédire leur affinité. En effet, l'énergie nécessaire à la désolvatation du ligand et de la protéine joue un rôle majeur dans l'établissement d'une interaction[212,213]. Nous verrons que le solvant n'est pas simple à prendre en compte et que des approximations sont souvent utilisées.

Dans ce chapitre, nous présenterons 2 grandes méthodes de criblage : le criblage de pharmacophore basé sur la structure et l'amarrage moléculaire de fragments et de petites molécules. Très brièvement, le premier consiste à identifier des points d'ancrage à la surface du site de liaison impliqués dans l'établissement d'une interaction avec un ligand[175], les autres consistent à prédire le mode de liaison d'une petite molécule dans le site de liaison de la cible et d'y associer un score censé informer au mieux sur l'affinité biologique théorique entre ces deux entités[214,215,216,217,218].

4.1 Obtention des structures 3D

L'obtention d'une structure 3D est primordiale pour conduire des études de criblage virtuel basé sur la structure. Elles peuvent être obtenues grâce à des méthodes expérimentales ou prédites par des méthodes *in silico*.

4.1.1 Elucidation expérimentale des structures 3D

La Protein Data Bank[219] (PDB) recense plus de 150000 structures résolues expérimentalement (le 08/04/2019) parmi lesquelles plus de 140000 sont des protéines, le reste étant des acides nucléiques (ARN et ADN). Chaque année, cette base de données s'enrichie de nouvelles structures, la barre de plus de 10000 structures résolues par année ayant été franchie en 2016 (Figure 19). Environ 90% des strctures disponibles dans la PDB sont résolues par cristallographie aux rayons X, et 10% par RMN ou microscopie électronique (EM). Il est à noter que si le nombre de structures résolues par RMN dépasse encore largement le nombre de structures résolues par EM, cette tendance tend à s'inverser depuis 2016 : l'EM permet désormais de résoudre plus de structures par an que les méthodes RMN. En effet, la cryo-EM (Cf 4.1.1.3) permet désormais de déterminer la structure de protéines avec une haute résolution (< 3 Å), ce qui a notamment valu le prix Nobel de chimie au Pr Dubochet, au Pr Frank et au Dr. Henderson en 2017[220]. La cryo-EM qui permet de résoudre la

structure de macromolécules allant jusqu'à plusieurs milliers de résidus apparaît comme une nouvelle alternative à la RMN qui souffre de contraintes de taille puisqu'elle n'est applicable qu'aux protéines ayant généralement moins de 300 résidus. A titre d'exemple, la structure de l'*Adeno-associated virus serotype 2 variant* (AAV2) contenant plus de 44000 résidus a été résolue

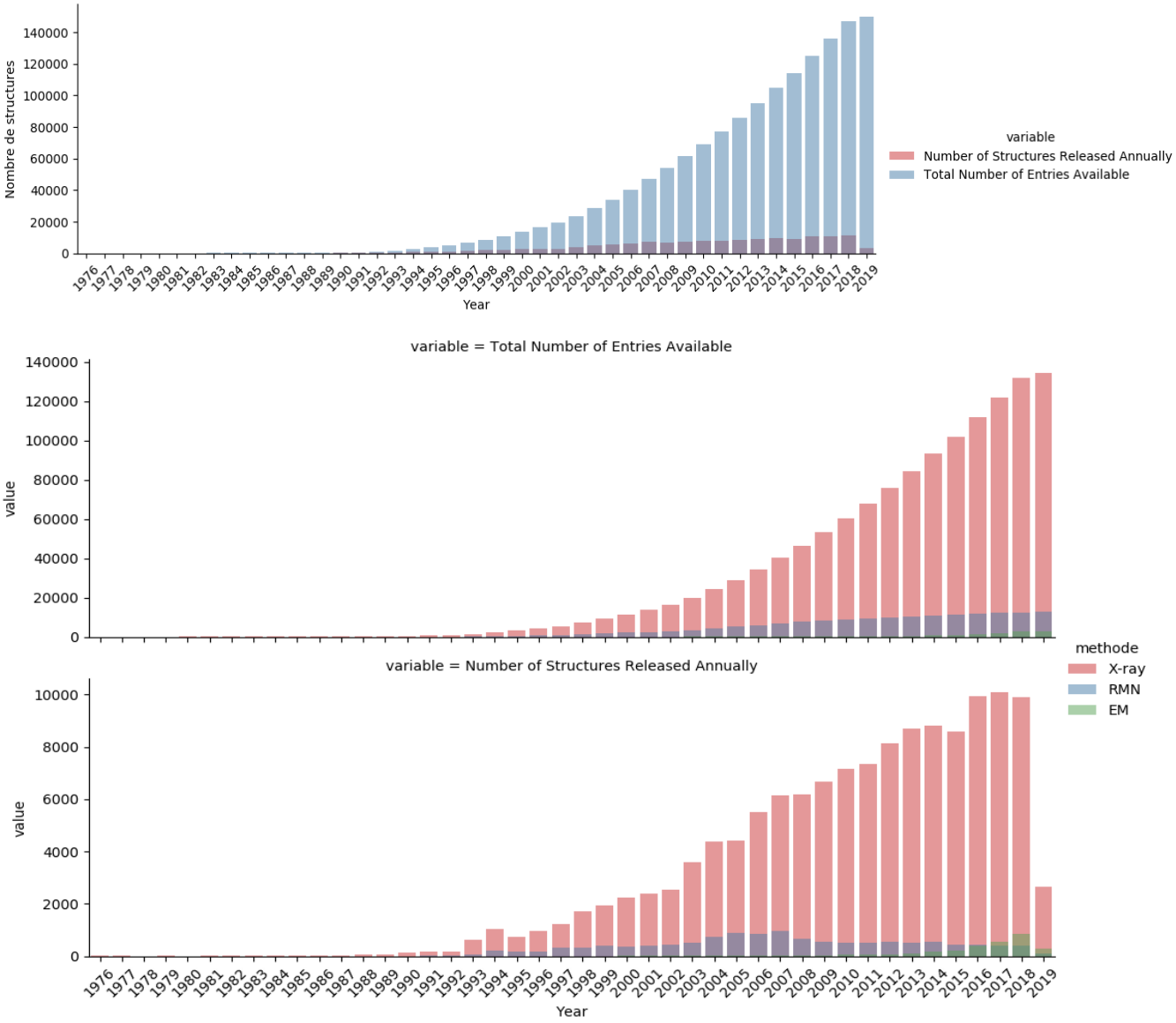


Figure 19 Histogramme du nombre de structures recensées dans la PDB de 1976 au 08/04/2019.

par cryo-EM en 2018 avec une résolution 1,86 Å.

4.1.1.1 Cristallographie aux rayons X

La cristallographie aux rayons X de protéines est divisée en 4 étapes.

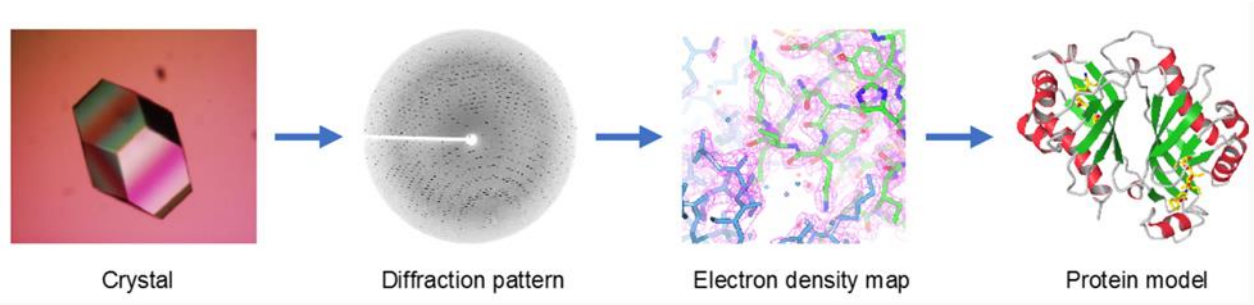


Figure 20 Illustration des grandes étapes de la cristallographie aux rayons X. D'après [216]

Les deux premières étapes sont clés et limitantes : la nucléation, ou initiation d'un cristal, et la croissance du cristal. La nucléation a lieu lorsque des protéines sont sursaturées dans leur solution et s'agrègent pour former un noyau solide stable. Pour réussir la nucléation d'une solution protéique, il faut identifier les paramètres expérimentaux optimaux (réactifs, pH du tampon, la bonne température, et excipients ou additifs) pour diminuer sa solubilité. La phase de croissance est la phase durant laquelle le noyau va grossir, c'est à dire que d'avantage de protéines vont venir former le noyau stable et un cristal va apparaître.

Différentes méthodes sont employées pour faire cristalliser une solution protéique : la diffusion de

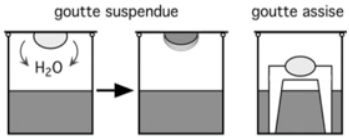
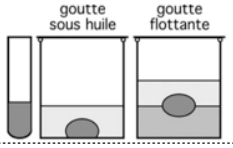
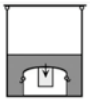
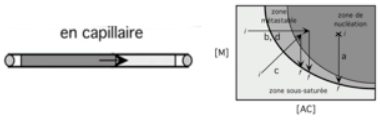
Diffusion de vapeur • Une gouttelette (solution protéique purifiée + tampon + précipitant) placée dans une cuve étanche ave un réservoir contenant le même tampon et le même précipitant en plus fortes concentrations • Diffusion du solvant afin d'équilibrer les concentrations du réservoir et de la gouttelette augmentation de la concentration protéique dans la goutte	
Microbatch • Solution protéique mélangée au précipitant dans une goutte d'huile pour la couper de toute interaction avec l'extérieur • La nucléation a lieu naturellement en présence des bonnes conditions de précipitation	
Microdialyse • Solution protéique placée dans une membrane semi-perméable qui permet la diffusion de petites molécules mais pas de protéine ou de polymères • Membrane placée dans un milieu concentré en précipitant • Equilibre atteint par diffusion membranaire	
Diffusion à l'interface libre • Mise en contact du précipitant et de la protéine sans les mélanger • Solution protéique et précipitant injectés de part et d'autre d'un capillaire • Etat d'équilibre atteint via un phénomène de diffusion	

Figure 21 Différentes méthodes utilisées pour faire cristalliser une solution protéique. D'après [215]

vapeur, le microbatch, la micro-dialyse ou encore la diffusion à l'interface libre (Figure 21)[221].

Une fois le cristal obtenu, des méthodes de cristallographie aux rayons X sont utilisées pour localiser la position des électrons qui gravitent autour de la protéine et éventuellement du ou des ligands du co-cristal. Un faisceau de rayons X est dirigé vers le cristal précédemment obtenu et est diffracté par les électrons des molécules ordonnées. Un détecteur permet d'enregistrer le motif de diffraction pour différentes orientations du cristal, ce qui permet de remonter à l'analyse de la densité électronique de la protéine (Figure 20[222]). *In fine*, une carte de densité électronique est obtenue. La protéine, qui consiste en un enchaînement connu d'acides aminés, va être informatiquement insérée dans le maillage de la carte de densité électronique et son positionnement va être optimisé. Selon la précision de la carte de densité électronique obtenue, les données tridimensionnelles de la protéine sont plus ou moins précises (Figure 22[221]).

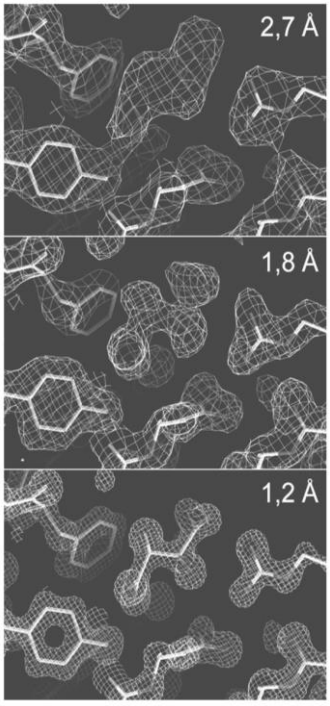


Figure 22 Illustration de différents niveaux de résolution obtenus par cristallisation aux rayons X. D'après [215]

La cristallographie aux rayons X est très utilisée pour élucider la structure tridimensionnelle d'une protéine. Il est cependant important d'étudier de manière critique le modèle final proposé et de ne pas se contenter des coordonnées 3D brutes extraites de la PDB pour étudier une protéine. Il est fréquent d'observer des anomalies dans les modèles générés liées à une mauvaise estimation de la

position d'atomes, de portion de protéines, ou de petites molécules. Ces anomalies sont dues à la basse résolution de la densité électronique obtenue qui ne permet pas de positionner avec certitude les atomes, et qui laisse place à une interprétation subjective. Par ailleurs, les régions flexibles des protéines, que ce soit un résidu adoptant plusieurs conformations ou une boucle mobile, sont difficiles à résoudre. Dans le cas d'un résidu flexible, plusieurs solutions sont parfois représentées dans le modèle 3D final lorsqu'une incertitude émerge ; les boucles flexibles sont quant à elles généralement tronquées. Il faut noter que les atomes d'hydrogène n'ont qu'un électron en gravitation, et ont donc un faible pouvoir de diffraction des rayons X. Par conséquent leur localisation est difficile à déterminer. Une récente étude conduite à l'institut de physique de l'académie Tchèque des Sciences et l'Université de Caen propose une méthode de cristallographie plus fine qui permettrait de résoudre le problème de la position des hydrogènes[223].

4.1.1.2 RMN

La résonnance magnétique nucléaire se base sur les propriétés magnétiques des atomes et notamment ceux possédant un spin nucléaire (ex : C, N, H). Placés dans un champ magnétique fort, les spins vont s'aligner avec le champ. L'interaction entre spins va permettre d'enregistrer des données sur les interactions scalaires (interaction entre électrons structurellement liés) et les interactions dipolaires (interaction entre atomes séparés de maximum 5 Å) sous forme de spectres. Ces spectres sont ensuite analysés et interprétés pour en déduire des contraintes de distance et de géométrie observées expérimentalement. L'objectif final est de proposer une conformation protéique qui réponde au mieux à l'ensemble des contraintes collectées. Des conformations aléatoires sont proposées, puis des contraintes leurs sont imposées pour conduire à un repliement plausible. Plusieurs points de départs sont considérés afin de vérifier la convergence vers une seule et unique configuration finale. De nombreux logiciels permettent d'automatiser cette phase d'un haut niveau de complexité (ex : NMRFAM[224], Mestre Nova[225], TopSpin[226,227]). L'avantage de la RMN est qu'elle est réalisée sur un échantillon en solution aqueuse contrairement à la cristallographie aux rayons X. La protéine est donc moins contrainte et plusieurs conformations d'une même protéine peuvent être capturées à partir d'un seul échantillon.

L'élucidation d'une structure tridimensionnelle par RMN est basée sur un processus de positionnement des atomes en fonction des contraintes physiques observées et de l'enchaînement connu des acides aminés de la protéine. Il faut noter que plus il y a d'atomes dans le système, plus

il est difficile de satisfaire toutes les contraintes. D'autant que les interactions dipolaires observées, autrement appelées effet Overhauser, ne permettent pas de distinguer les atomes proches liés des non liés, ce qui multiplie les effets observés dans le cas de grosses protéines et par conséquent le taux d'erreurs d'interprétation possibles par les outils informatique. On considère donc que cette méthode n'est applicable qu'à des protéines de protéines de masse moléculaire de 10 à 30 kDa.

4.1.1.3 Cryo-EM

La cryo-EM est basée sur des propriétés optiques. La solution protéique est déposée sur une grille de microscopie électronique recouverte d'une membrane à trou. L'excédent est absorbé afin de ne conserver qu'une fine couche d'échantillon (< 500 nm) puis la grille est rapidement plongée dans de l'éthane liquide à -160°C ce qui permet la vitrification de l'eau. Chaque trou de la membrane contient ainsi des protéines congelées dans un état natif. La grille est ensuite transférée à un

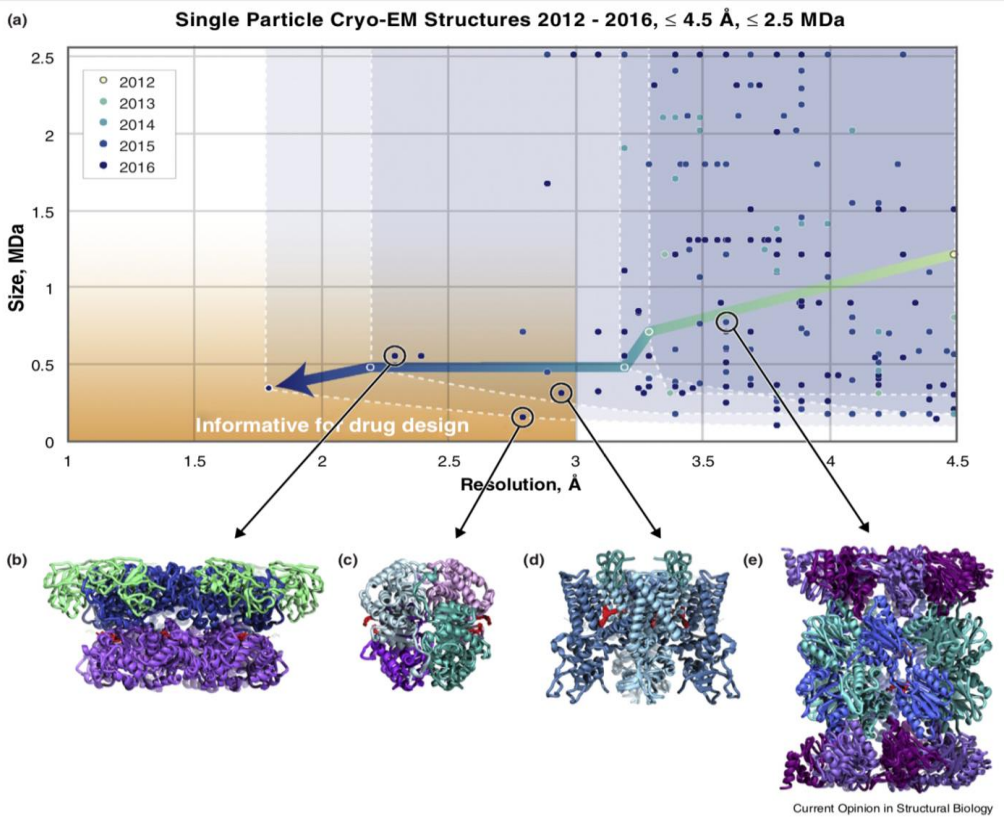


Figure 23 Illustration de l'évolution de la résolution et de la taille des structures protéiques résolues par cryo-EM (< 4,5 Å). D'après [223]

microscope optique dans de l'azote liquide à -185°C[228]. Le microscope électronique permet de recueillir des milliers d'images des protéines orientées de manière aléatoire. Ces images sont

ensuite analysées, débarrassées du bruit, puis combinées pour reconstituer la structure 3D de la protéine.

En comparaison à la cristallographie, la cryo-EM ne nécessite pas de cristal, et la solution protéique utilisée peut reproduire les conditions physiologiques en termes de pH et de concentration saline. La cryo-EM a très longtemps été utilisée pour résoudre la structure de gros systèmes biologiques telles que les capsides virales à des résolutions de l'ordre de 5 Å. L'optimisation de chacune des étapes du processus a permis d'accéder à des résolutions bien meilleures en deçà de 3 Å[229] (Figure 23). Par exemple la cryo-EM a permis de déterminer la structure de la glutamate déshydrogénase à une résolution de 1,8 Å[230]. De par sa capacité à déterminer la structure de protéines en milieu quasi-physiologique, sans contrainte de précipitation et sans contrainte de taille, elle représente une alternative prometteuse à la cristallographie et à la RMN

4.1.2 Élucidation *in silico* des structures 3D

4.1.2.1 Modélisation

Lorsqu'aucune structure n'est disponible pour une protéine donnée, il est possible d'utiliser des structures homologues issues d'une autre espèce, par exemple, ou bien de modéliser la structure de cette protéine. La modélisation de la structure 3D peut alors être guidée par 2 principes distincts : les lois de la physique (modélisation *ab initio*) et la théorie de l'évolution (modélisation comparative et modélisation par enfilage). La modélisation *ab initio* nécessite des calculs complexes utilisant différents niveaux d'information (modèles en treillis ou modèles atomistiques) afin d'échantillonner au mieux les repliements possibles d'une protéine sans a priori structuraux. Les méthodes basées sur l'évolution, moins coûteuses en termes de temps de calcul sont généralement privilégiées et plus accessibles pour les laboratoires ne disposant pas d'infrastructures informatiques suffisantes. Il est également possible d'échantillonner les conformations possibles d'une structure en réalisant des simulations de dynamique moléculaire. Cet aspect qui fait l'objet de nombreuses revues ne sera pas détaillé dans cette partie.

4.1.2.1.1 Ab Initio

Le mécanisme de repliement d'une protéine a fait l'objet de nombreuses études. En 1969, Levinthal a fait le calcul suivant (qui ne prenait pas en compte l'intervention des protéines chaperonnes) : si une protéine possède une séquence de 100 acides aminés, alors elle peut théoriquement adopter

environ 10^{48} conformations différentes. Or, s'il fallait 10^{-11}s pour passer d'une conformation à l'autre, il faudrait 10^{29} années pour explorer l'ensemble des conformations, ce qui est impossible en routine même avec les algorithmes les plus puissants. Plus tard, Anfisen a établi la théorie

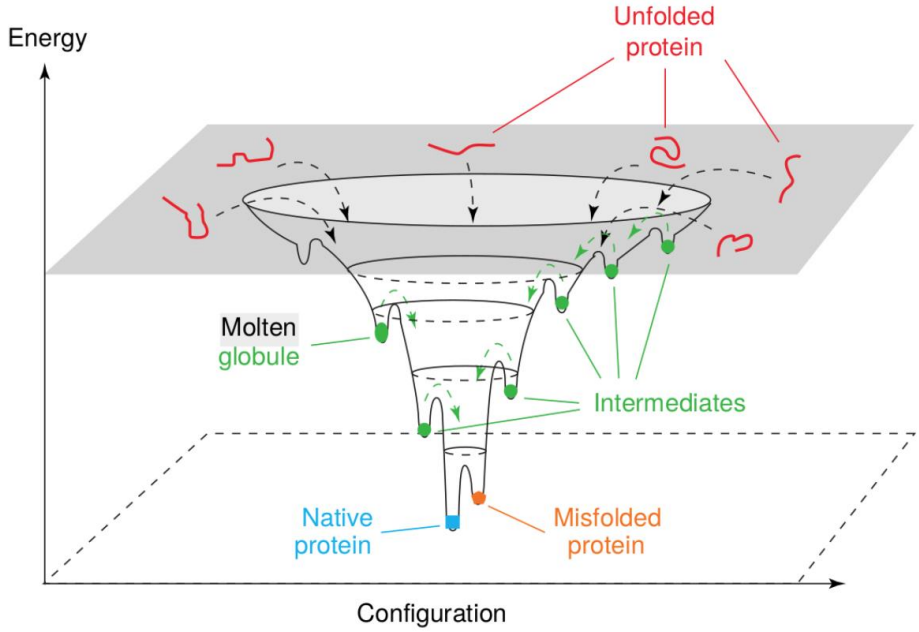


Figure 24 Exemple de paysage énergétique d'une protéine. D'après [225]

toujours largement acceptée, bien que challengée par les protéines intrinsèquement désordonnées, que les protéines, dans un environnement propice au repliement, adoptent une conformation dite « native », stable et cinétiquement accessible qui correspond au minimum global du paysage énergétique de la protéines[231] (Figure 24). Selon sa théorie, l'information sur le repliement d'une protéine serait contenue dans sa séquence seule, et aucun cofacteur ou aucune énergie externe ne serait nécessaire au repliement. La modélisation *ab initio* part de ce principe même. Elle est majoritairement utilisée sur de petites portions de protéines du fait de la complexité à atteindre ce minimum global sur des protéines entières, et elle fait appel à des méthodes de dynamique moléculaire. Grâce au développement de super-ordinateurs comme Anton, construit par D.E. Shaw Research, ou à des méthodes d'échantillonnages comme le *Replica Exchange*, des structures de protéines de plus de 100 acides aminées ont été retrouvées avec précision[232,233,234].

4.1.2.1.2 Modélisation comparative

La modélisation comparative, autrement appelée modélisation par homologie, part du principe que les structures sont généralement davantage conservées que les séquences au cours de l'évolution.

Ainsi, des molécules qui possèdent une forte identité de séquence ont de fortes chances d'être structurellement très similaires. Les méthodes de modélisation comparative utilisent un protocole en 3 temps : premièrement, les protéines partageant des similarités de séquence ou de profils avec la protéine d'étude sont identifiées, deuxièmement, elles sont alignées à la séquence de la protéine d'étude, puis, troisièmement, les coordonnées des atomes du squelette des modèles sont transférées à la séquence de la protéine d'étude (Figure 25) ; les boucles ainsi que les chaînes latérales sont ensuite modélisées

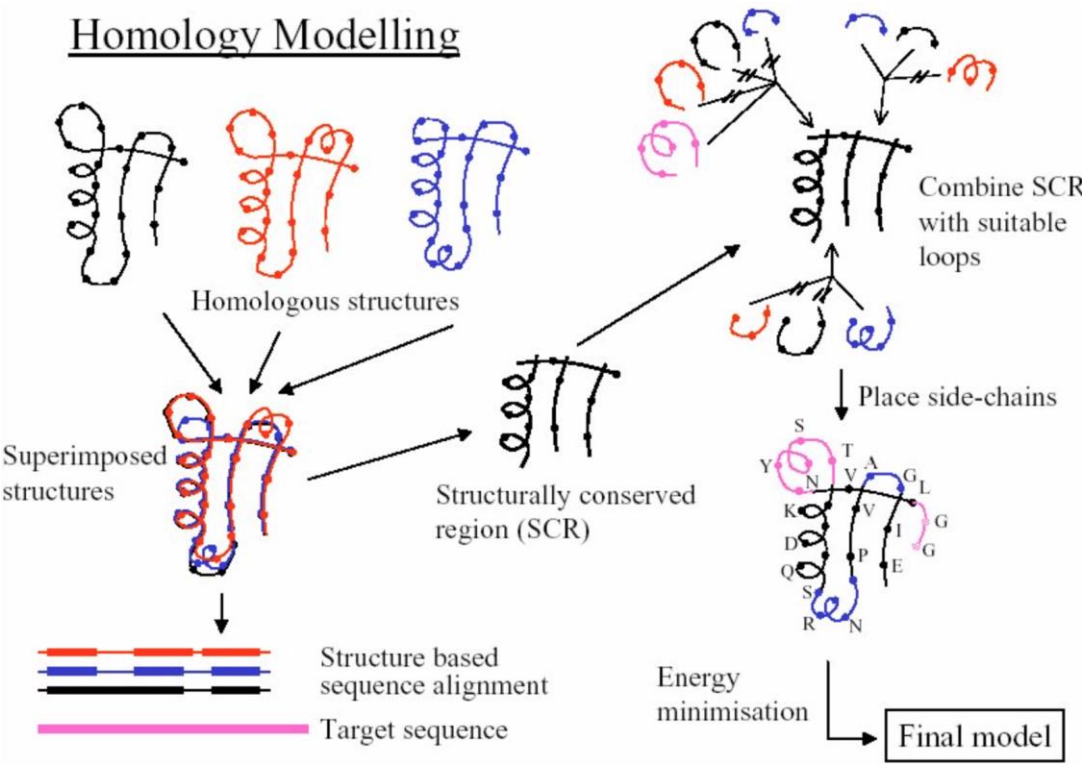


Figure 25 Représentation schématique de la modélisation par homologie. D'après le cours en ligne « Homology modelling and threading » de Dr. Peer Mittl

4.1.2.1.2.1 Identification de protéines similaires

Des protéines similaires sont identifiées via des approches séquence-séquence (BLAST), profil-séquence (PSI-BLAST[235]) ou profil-profil (HHSearch[236], ORION[237]). Un profil est une matrice (PSSM) qui associe à chaque position d'une séquence protéique un score obtenu à partir de l'alignement de séquences homologues et qui renseigne donc sur la conservation d'un résidu au cours de l'évolution[237]. Certaines méthodes combinent des informations de séquence et des

informations structurales en transformant les séquences en alphabets structuraux, chaque lettre correspondant à un motif d'acides aminés et ayant une probabilité de correspondre à un repliement ou à un autre. Le nombre de lettres encodées par chaque alphabet peut varier et apporter plus ou moins de précision quant à la structure locale probable[237,238]. Une protéine peut être acceptée comme modèle si elle partage au minimum 30% d'identité de séquence avec la protéine d'intérêt. En fonction de la valeur de l'identité de séquence entre la (ou les) protéine(s) modèle(s) et la protéine d'intérêt, un modèle de la structure 3D de la protéine peut être construit avec plus ou moins de confiance. En deçà de 30% d'identité de séquence, la similarité est trop faible pour assurer une conservation de structure au cours de l'évolution[239], c'est ce que qu'on appelle la *twilight zone*. Certaines protéines ne partagent que des portions de séquence communes avec la protéine d'intérêt. Pour parvenir à modéliser une protéine entière, la considération de plusieurs protéines localement similaires en terme de séquence a montré de meilleures performances[240,241,242,243]. Ceci permet notamment d'éviter que le modèle soit biaisé par les variations spécifiques à chaque protéine (boucles, conformations des chaînes latérales etc.). Selon le sujet d'étude, il est important de considérer plusieurs critères pouvant influencer et valider le choix des protéines modèles avant de passer à l'étape de la modélisation : les conditions expérimentales de résolutions des structures similaires, la qualité de la structure, la famille protéique à laquelle elle appartient ainsi que sa capacité ou non à se fixer à un ligand similaire à la protéine d'étude, etc.

4.1.2.1.2.2 Alignement et modélisation comparative

Après avoir sélectionné les protéines modèles, l'étape suivante consiste à trouver l'alignement optimal de la séquence d'étude et du(des) modèle(s) qui est ensuite soumis à un programme de modélisation comparative dont le rôle est de transférer les coordonnées des protéines modèles à celle à modéliser. Généralement, le « cœur » de la protéine est déterminé en premier ; il s'agit de la partie structurellement très conservée entre protéines modèles (RMSD < 4 Å).

4.1.2.1.2.3 Modélisation des boucles et des chaînes latérales

Les boucles sont ensuite modélisées grâce au(x) modèle(s), à des banques de données de boucles (WLOOP[244], MODLOOP[245]) ou à de la modélisation *de novo* par dynamique moléculaire et optimisation. Il est nécessaire de modéliser les chaînes latérales qui ont un rôle essentiel dans la stabilisation de la structure tridimensionnelle et dans la fonction de la protéine. Ces chaines latérales peuvent être modélisées par homologie ou bien à partir de banques de rotamères[246]. Les

modèles sont ensuite optimisés pour supprimer l'encombrement stérique des chaînes latérales et minimiser l'énergie globale.

Des outils comme Modeller[247] permettent d'effectuer de manière semi-automatique l'ensemble de ses étapes. D'autres, comme Robetta[248], combinent la modélisation comparative et *ab initio* pour proposer le modèle le plus complet possible.

4.1.2.1.3 Modélisation par enfilage

La modélisation par enfilage est plus communément appelée *threading*. Elle consiste à « enfiler » la séquence de la protéine d'intérêt dans des structures connues. En d'autres termes, il s'agit de sélectionner des structures issues de bases de données structurales (PDB[249], FSSP[250], SCOP[251] ou CATH[252]), d'aligner des portions de la protéine sur les structures sélectionnées et de calculer un score qui prend en compte différents éléments comme la compatibilité des structures secondaires et le potentiel de mutation notamment. L'objectif est d'identifier la ou les structures dans lesquelles les portions de protéines vont donner le meilleur score. De même que pour les méthodes précédentes, l'énergie des modèles finaux est minimisée pour obtenir un modèle de la plus basse énergie possible.

4.1.2.1.4 Validation des modèles

La validation des modèles est une étape cruciale de la modélisation comparative afin de s'assurer que les modèles construits sont biologiquement plausibles. Il est donc nécessaire de vérifier le repliement proposé (ex : disposition des acides aminées hydrophobes) et sa qualité (angles dièdres, longueurs des liaisons interatomiques, interactions électrostatiques etc.). Il faut ainsi s'assurer que le modèle minimise 1) les contraintes de torsion, 2) les cavités interstitielles, 3) les charges enfouies, 4) les clashs stériques et 5) l'énergie potentielle (énergie de Van der Waals et énergie électrostatique) et maximise 1) les liaisons hydrogènes, 2) l'enfouissement des groupements hydrophobes, et 3) l'exposition des groupements hydrophiles. Plusieurs serveurs automatisent cette vérification (PROCHECK[253], ProSA-web[254], VADAR[255], WHAT IF[256], DSSP[257]).

4.2 Outils de prédiction de site de liaison

Il existe deux manières de définir un site de liaison selon si la structure de la protéine a été résolue en complexe avec un ligand ou non. Si l'on dispose d'un complexe protéine-ligand, on utilise généralement le ligand comme référence et le site de liaison est défini autour de ce ligand[258]. Si

l'on cherche à cibler une zone de la protéine en dehors du site de liaison du ligand de référence, ou si l'on dispose uniquement d'une structure *apo*, il est possible d'utiliser des méthodes de prédiction des sites de liaisons. Ces méthodes sont divisées en trois catégories : les méthodes basées sur la connaissance, les méthodes basées sur la géométrie et les méthodes basées sur l'énergie d'interaction. Il existe des approches qui combinent plusieurs de ces méthodes. Par exemple, MetaPocket combine 3 outils basés sur la géométrie (LIGSITE, PASS[259], Surfnet) ainsi qu'un outil basé sur l'énergie (Q-SiteFinder[260]).

4.2.1 Outils de prédiction basés sur la connaissance

Les outils de prédiction basés sur la connaissance s'appuient essentiellement sur les informations apportées par des protéines partageant une similarité de séquence et/ou de structure avec la protéine d'étude.

Les méthodes de similarité de séquence reposent sur le fait que les séquences au niveau des sites de liaisons sont très conservées au cours de l'évolution, afin de conserver l'interaction avec leur(s) ligand(s) naturel(s)[261] ; des protéines avec des séquences similaires ont donc une forte probabilité de présenter des sites de liaison similaires. Le nombre de structures disponibles dans la Protein Data Bank augmentant chaque année, de plus en plus de site d'interactions protéine/ligand sont connus. Des bases de données telles que PROSITE[262] regroupant des informations sur des sites de liaison et la séquence associée sont généralement utilisées comme point de départ de la recherche de site(s) de liaison. Les méthodes de similarité de séquence tirent donc profit de cette abondance de données et recherchent les similarités de séquence avec des sites de liaisons connus (ex : ConfSurf[263]).

Les méthodes de similarité structurale s'appuient sur l'hypothèse que l'espace des sites de liaisons est restreint (il pourrait être représenté par 1000 formes de poches[264]), et des similarités structurales sont observées entre les protéines exerçant des fonctions similaires. Ces méthodes nécessitent une première étape d'alignement structural entre la protéine étudiée et des sites de liaisons connus. Pour cela, des banques de données de référence regroupant des données tridimensionnelles sur des sites de liaisons expérimentalement identifiés (CavBase[265], Patterns In Non-homologous Tertiary Structures (PINTS)[266], SiteEngine[267], eF-site[268], ProFunc[269]) peuvent être utilisées. L'un des outils d'alignement structural le plus utilisé dans cette optique est TM-align[270], qui peut également être

utilisé afin d'identifier des protéines *holo* similaires à la protéine d'étude, et en déduire un site de liaison potentiel.

Il existe des outils tel que COFACTOR[271], qui combinent des méthodes de prédiction basées sur la similarité de séquence et de structure.

4.2.2 Outils de prédiction basés sur la géométrie

Les méthodes de prédiction basés sur la géométrie consistent à parcourir la surface de la protéine afin d'identifier des cavités, autrement nommées poches, qui sont souvent associées aux sites de liaison. Pour cela, trois types d'approches sont utilisées : celles qui utilisent une grille tridimensionnelle entourant la protéine ou une description tridimensionnelle de la surface moléculaire (POCKET[272], LIGSITE[273] et son implémentation Pocket-Finder, VolSite[274]), celles qui utilisent des sondes sphériques représentant le solvant qui sont « roulées » à la surface de la protéines et celles qui utilisent les diagrammes de Voronoï.

Par exemple, POCKET utilise une méthode basée sur une grille tridimensionnelle. Une fois la protéine passée dans une grille, une sonde sphérique de 3A longe chaque ligne de la grille dans la direction x, y et z (Figure 26[275]). Une cavité est définie lorsque la sonde interagit avec la protéine, c'est à dire pénètre le rayon d'un atome de la protéine, rencontre une zone vide puis, interagit de nouveau avec la protéine. Cette approche est très dépendante de l'orientation initiale de la protéine dans la grille.

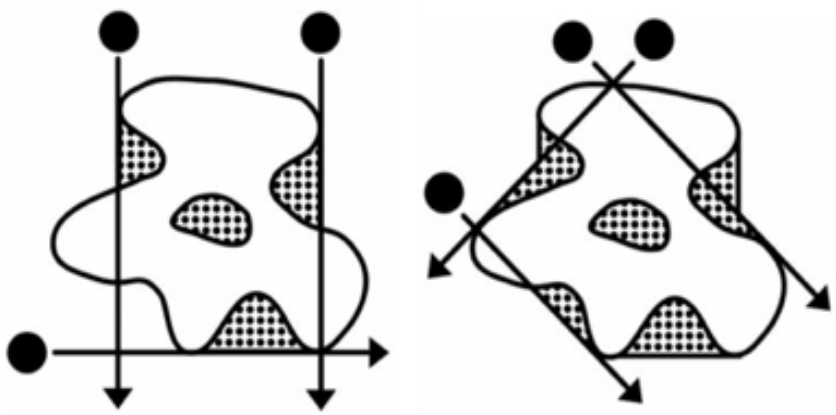


Figure 26 Illustration de l'algorithme de détection de cavités POCKET et LIGSITE. Les deux méthodes utilisent une grille tridimensionnelle dont chaque ligne est parcourue par une sonde ; LIGSITE explore également les lignes diagonales. D'après [269]

Une manière de s'affranchir des limites d'exploration de surface liées à l'utilisation d'une grille est d'utiliser des sondes sphériques bombardées à la surface de la protéine ou des méthodes basées sur les diagrammes de Voronoï. Les méthodes de détections de zones concaves via l'utilisation de sondes sphériques sont multiples : parmi les stratégies adoptées, SURFNET[276] identifie les clusters de sondes alors que PASS attribue un « facteur d'enfouissement » à chaque sphère – correspondant au nombre d'atomes de la protéine présents à une certaine distance de la sphère. Les méthodes basées sur les diagrammes de Voronoï peuvent être divisées en 2 catégories. Certaines méthodes

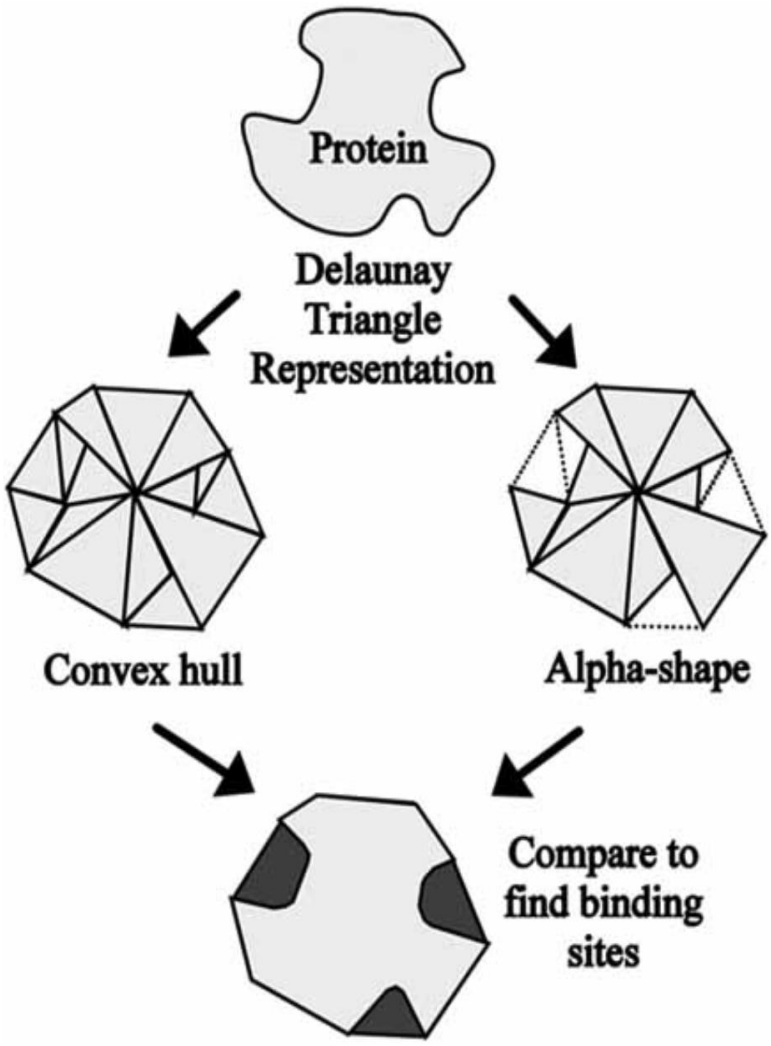


Figure 27 Illustration de l'algorithme de détection de cavités d'APROPOS. Les cavités sont définies comme la différence entre l'enveloppe convexe et l'alpha *shape*. D'après [269]

(CAST[277], APROPOS[278]) utilisent la tesselation de Voronoï et la triangulation de Delaunay pour définir une représentation α-*shape* de la protéine, de laquelle sont soustraits les sommets et les arrêtes de Voronoï qui se trouvent à l'extérieur de la protéine. Les zones soustraites correspondent alors aux cavités (Figure 27[275]). Sur le même principe, d'autres méthodes utilisent les sphères α (FPocket[279]), définies comme des sphères de diamètre variable au contact de 4 atomes mais ne recouvrant aucun atome, le positionnement de ces sphères α étant défini par la tesselation de Voronoï. Dans ce cas, les sphères α reflètent la courbure locale définie par les 4 atomes ; si la surface est concave, la sphère sera incluse dans un tétraèdre et sa taille sera relative à la distance interatomique ; si la surface est plane, la sphère sera de diamètre infini. Des critères de tailles sont ensuite appliqués afin d'éliminer les sphères de petites tailles qui correspondent aux zones inaccessibles au solvant (généralement fixé au radius de l'eau d'1.4Å), ainsi que les sphères de grandes tailles qui correspondent aux zones trop exposées.

4.2.3 Outils de prédiction basés sur les énergies

Les méthodes basées sur des critères d'énergies partent du principe que les sites de liaisons ont la capacité d'interagir avec des petites molécules en formant localement des liaisons polaires ou apolaires. De ce fait, les interactions entre la surface protéique et des sondes imitant les propriétés communes des petites molécules (hydrophobe, donneur/accepteur de liaison hydrogène) sont évaluées. Les outils appliquant cette méthode (AutoSite[280], FTMap[281], FTSite[282], VolSite[274]) bombardent des sondes de différentes natures à la surface de la protéine. Les sondes de même natures situées à proximité les unes des autres sont regroupées en cluster de sondes, et les zones de forte densité de clusters sont identifiées (Figure 28). Des différences dans le nombre et la nature des sondes utilisées sont observées parmi les outils de prédiction basés sur les énergies. Ainsi, dans la méthode de Jain et al.[283], 3 types de sondes sont utilisés (hydrophobe, donneur de liaison hydrogène (NH) et accepteur de liaison hydrogène C=O) contre 7 types pour VolSite (donneur/accepteur de liaison hydrogène, donneur de liaison hydrogène, accepteur de liaison hydrogène, chargé positivement, chargé négativement, aromatique, hydrophobe).

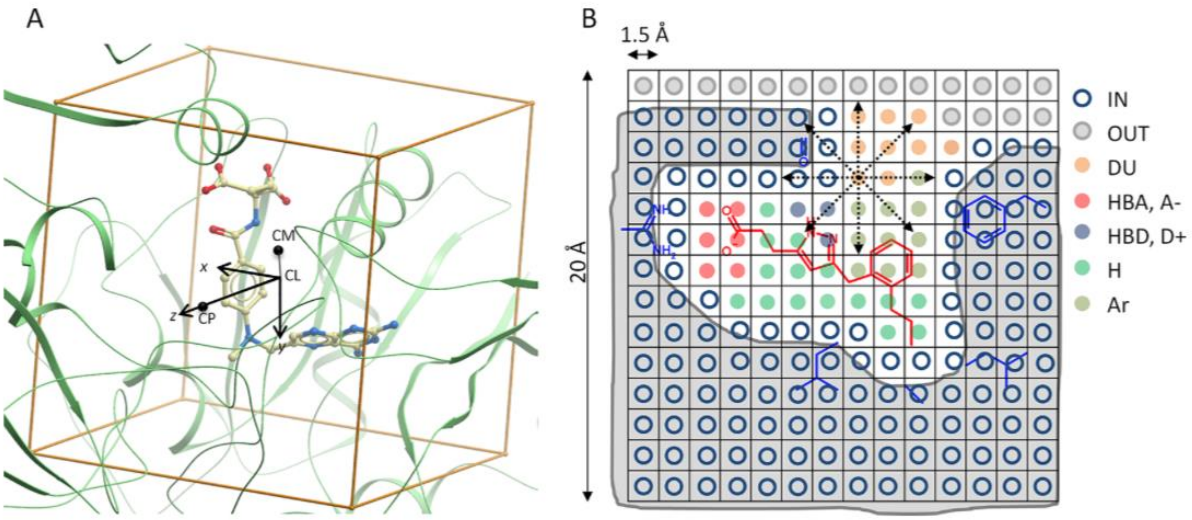


Figure 28 Illustration de l'algorithme de détection de cavités de VolSite. La protéine est placée dans une grille 3D, et une sonde est placée à chaque point de la grille. Si la sonde est à moins de 2.5Å d'un atome de la protéine, elle est considérée IN (à l'intérieur). 120 rayons de 8Å partant de chaque sonde "IN" sont ensuite générés, et le nombre de rayons croisant une autre sonde "IN" est calculé : s'il est en deçà d'un certain seuil, la sonde sera considérée « OUT » (à l'extérieur). Chaque sonde « IN » se voit ensuite attribuée une propriété qui est l'image négative de son environnement à la surface de la protéine. D'après [268]

4.2.4 Évaluation de la *druggabilité* d'une cavité

Avant de pouvoir développer un médicament pour une cible thérapeutique, il faut s'assurer que cette dernière possède des propriétés physicochimiques adéquates pour pouvoir être modulée par une petite molécule : c'est ce que l'on appelle l'étude de la *druggabilité*. La *druggabilité* d'une protéine est un terme devenu populaire en 2002 suite à la publication du « Druggable Genome » de Hopkins et Groom[284]. Elle est définie comme la capacité d'une protéine à « interagir avec une petite molécule ayant des propriétés physicochimiques appropriées et une affinité de liaison requise ». Les « propriétés physicochimique appropriées » se réfèrent aux propriétés *druglike* du ligand (Cf 2.2.3.1), et l' « affinité de liaison requise » correspond généralement à une affinité de l'ordre du nM[285]. Certains chercheurs jugent ce terme trop vaste et proposent d'évaluer le « *ligandability* » ou « *bindability* » qui correspond à la capacité d'une protéine à se lier à une petite molécule peu importe son caractère *druglike*. La plupart des outils de prédiction se base néanmoins sur des banques de données qui opposent les protéines ou cavités qui ont été ciblées avec succès à celles

qui ont échouées avec des molécules *druglike* et reflète donc plus la *druggabilité* que la *ligandabilité*.

Parmi les approches développées pour estimer la *druggabilité* d'une protéine, la première et également la plus simpliste, consiste à étudier la famille protéique à laquelle appartient la protéine cible. En effet, les enzymes, transporteurs et récepteurs interagissent naturellement avec des petites molécules et sont généralement *druggables* ; d'autres familles de protéines sont plus difficiles voire impossible à moduler avec des petites molécules[286]. Les propriétés *druglike* ou non du ligand naturel permettent de prédire la *druggabilité* d'une cible. Par exemple, les GPCRs de classe A et les kinases possèdent des ligands naturels quasi-*druglike* et sont plus *druggables* que les GPCR de classe C et les protéases qui interagissent avec des ligands naturels non *druglike* de type peptide[285]. Il existe également des méthodes de prédiction expérimentale (ex : HTS[287]) et *in silico*. La majorité des méthodes *in silico* vise à associer des descripteurs physico-chimiques de poches (polarité, ouverture, surface accessible au solvant, nombre d'acide aminés polaire/ apolaire etc.)[288] à leur *druggabilité*. Il a été remarqué, par exemple, que les cavités de plus grands volumes dans une protéine coïncident généralement avec le site de liaison de petites molécules. Ces approches *in silico* s'appuient sur des données d'affinité pour entraîner les modèles de prédiction : une cavité étant capable d'interagir avec une molécule avec une forte affinité est considérée *druggable,* une cavité pour laquelle un certain niveau d'affinité n'a jamais été observé est considérée *non druggable*. Cependant, il est extrêmement difficile de considérer une cavité comme non *druggable* du fait de la non-exhaustivité des tests réalisés expérimentalement et les jeux de données comportent donc de très faibles proportions de cavités *non druggables* en comparaison aux cavité *druggables*. Le jeu de données le plus fourni est le « *non-redundant set of druggable and less druggable binding sites* » (NRDLD)[289]. Il contient 115 cibles parmi lesquelles 71 sont *druggables* et 44 sont non *druggables*. Dans le NRDLD les cibles *non druggables* sont définies comme celles pour lesquels aucun ligand connu ne respecte les Ro5 (Cf 2.2.3.1), n'a un clogP ≥ 2 ou n'a une efficacité atomique ≥ 0.3 kcal/mol. Ce jeu de données a largement été utilisé (DrugPrep[289], VolSite[290], PockDrug[291,292]). Chaque modèle propose une combinaison de descripteurs plus ou moins nombreux pour prédire le caractère *druggable* d'un site de liaison, la *druggabilité* augmentant généralement avec l'augmentation de l'hydrophobicité et de l'enfouissement ; le volume d'hydrophilicité ayant tendance à la réduire[285]. De manière générale, de bonnes performances de prédiction de *druggabilité* sont atteintes avec un taux de succès d'environ 70%

pour la meilleure poche prédite et d'environ 90% en prenant en compte les 3 meilleures poches prédites par outils[293,294,295].

4.3 Préparation des protéines

Les structures des protéines (résolues expérimentalement ou modélisées *in silico*) ne peuvent généralement pas être utilisées telles quelles avec les méthodes de criblage basées sur la structure, il convient de la préparer correctement préalablement. La première étape consiste à vérifier si la structure est complète puisqu'il arrive que des atomes ou des résidus (pouvant constituer des portions entières de protéines) ne soient pas complètement résolus. Si le criblage est réalisé sur une zone réduite, i.e. un criblage focalisé, et que les atomes ou résidus manquants ne font pas partie de la zone ciblée, l'influence de leur absence sur les résultats sera nulle. En revanche, s'ils sont situés à des endroits clés, leur absence peut complètement modifier les résultats obtenus (volume de la

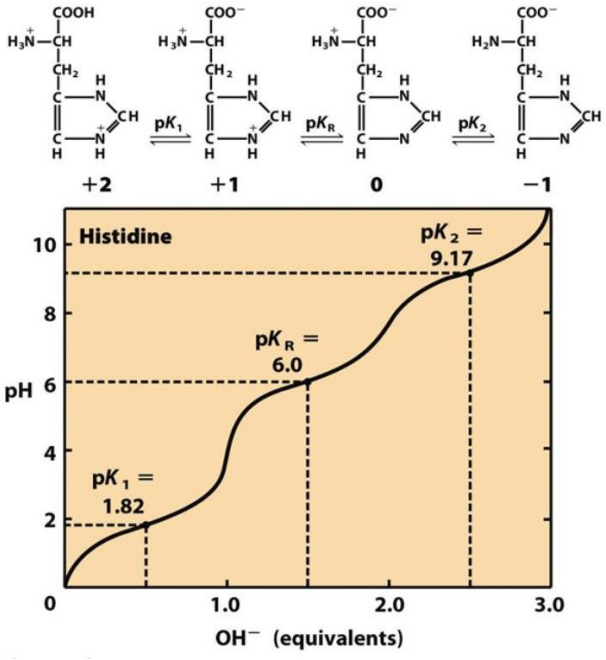


Figure 29 Courbe de titrage de l'histidine. L'histidine possède 3 pKa (pK1 = 1.82, pK2 = 6.0 et pK3 = 9.17). A pH < 6.0 l'histidine est chargée positivement, à pH > 6.0 l'histidine est neutre. D'après Dr Mohammed Saadeh

poche plus grand, perte d'interactions potentielles entre les résidus manquants et le ligand, etc.). Dans un tel cas de figure, il est possible d'utiliser une autre structure complète de la protéine

lorsqu'elle existe ou dans le cas contraire de modéliser les parties manquantes avec un logiciel adapté (ex : MODELLER[296,297], AMBER tleap[298]). La seconde étape de préparation de la structure de la protéine, également cruciale, est dédiée à sa protonation. En effet, en fonction du pH (7.4), de son pKa et de son environnement proche, un résidu peut adopter plusieurs états d'ionisation[208299]. C'est particulièrement le cas des histidines qui ont un pKa = 6, proche du pH physiologique qui peuvent être chargées ou neutres en fonction de leurs environnements (Figure 29). Des outils permettent d'estimer le pKa local des protéines et d'en déduire un état d'ionisation (AMBER, PARSE, MCCE[300,301], UHDB[302], PROPKA[303])[304].

4.4 Outils de criblage basés sur la structure

Le criblage virtuel basé sur la structure peut être appliqué une fois la structure de la protéine d'étude résolue, le site de liaison éventuellement identifié et la protéine correctement préparée. Parmi les méthodes utilisées, la modélisation de pharmacophores et le docking sont les deux plus populaires. Ils seront décrits dans cette partie.

4.4.1 Modélisation de pharmacophores

Comme mentionné précédemment (Cf 3.1.3), un modèle de pharmacophore est défini par l'IUPAC comme un « ensemble d'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 »[169]. Il peut être modélisé à partir de ligands connus de la cible (approches basées sur les ligands) ou bien déduit de la structure de la cible (approches basées sur la structure). Lorsqu'une ou plusieurs structures *holo* de la cible étudiée sont disponibles, un modèle de pharmacophore peut être déduit à partir des interactions détectées entre le ligand et la protéine (modèle de pharmacophore basé sur le couple ligand/protéine). Les modèles de pharmacophore résultant de l'approche basée sur la structure *holo* sont limités aux interactions formées par les ligands résolus en complexe avec la protéine, ce qui n'est probablement pas représentatif de toutes les interactions possibles. Ils ne peuvent être générés lorsqu'aucune structure *holo* n'est disponible. Pour pallier ces limites, des méthodes de génération de modèles de pharmacophores focalisées sur la structure *apo* (modèle de pharmacophore basé sur le site de liaison) ont été développées.

4.4.1.1 Pharmacophore basé sur le couple ligand/protéine

Les pharmacophores basés sur le couple ligand/protéine tirent des informations des interactions établies entre un ligand et une protéine. Généralement les complexes sont issus de la PDB et ont été résolus expérimentalement. Il est néanmoins possible d'utiliser des complexes issus d'études de docking, malgré leur manque de fiabilité. Le principe de la modélisation de pharmacophores basée sur le couple ligand/protéine est de convertir le profil d'interaction en points pharmacophoriques précisément localisés dans l'espace et déduire des contraintes. Ces méthodes reposent sur deux éléments : l'interprétation de la topologie du ligand puis l'identification et la classification d'interactions respectant des règles précises.

Par exemple, le logiciel LigandScout procède en 6 étapes pour générer un modèle de pharmacophore basé sur le couple ligand protéine. Premièrement, le ligand est detecté et les acides aminés situé à une distance maximale de 7Å autour du ligand sont protonnés. Les états d'hybridation et les types de liaisons du ligands sont ensuite interprétés. Dans un troisième temps, des points pharmacophoriques sont assignés au ligand (accepteur de liaison hydrogène, donneur de liaison hydrogène, charge positive, charge négative, cycle aromatique, ion metallique) puis à la protéine. A chaque type de point pharmacophorique implémenté (point hydrophobe, groupe chargé positivement, groupe chargé négativement, donneur de liaison hydrogène, accepteur de liaison hydrogène) est associé un ensemble de règles à respecter qui peuvent être customisées par l'utilisateur. Dans une 5ème étape, les points pharmacophoriques sont ajoutés au modèle si des propriétés complémentaires sont observées localement entre le ligand et la protéine. Par exemple, LigandScout considère un groupement chargé comme point pharmacophorique si une charge opposée est située entre 1.5 et 5.6 Å de son centre[305]. Enfin, des volumes d'exclusion sont ajouté au modèle en opposition aux points pharmacophoriques hydrophobes du ligand. Pour un complexe ligand/protéine, un seul modèle de pharmacophore est généré. Il faut noter que plusieurs modèles de pharmacophore peuvent être fusionnés pour combiner les informations de plusieurs complexes résolus[305]. Les modèles de pharmacophores combinés peuvent regrouper les informations

communes à chaque modèle pris en compte, ou bien les points pharmacophoriques fréquemment observés, la fréquence d'observation seuil pouvant être imposée par l'utilisateur.

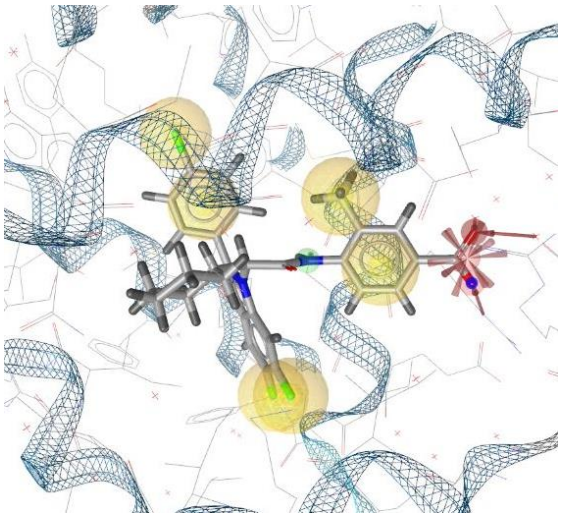


Figure 30 Pharmacophore du complexe ligand/protéines impliquant le ligand OLF et la protéine FXR (PDB : 3OLF) modélisé avec LigandScout 4.3

4.4.1.2 Modèles de pharmacophores basés sur le site de liaison seul

De plus en plus de structures protéiques sont résolues chaque année, la majorité étant en conformation *apo*, c'est-à-dire dépourvue de ligand. Afin de pouvoir appliquer les méthodes de modélisation de pharmacophore sur un large panel de structures, les méthodes focalisées sur la le site de liaison seul représentent une alternative nécessaire aux méthodes basées sur le couple ligand/protéine [306]. Trois grandes approches sont utilisées.

1) Les méthodes basées sur l'alignement de séquence permettent la modélisation d'un pharmacophore basée sur des résidus clés du site de liaison qui sont identifiés par alignement de séquences[307]. Ces résidus clés peuvent être des résidus très conservés au cours de l'évolution (qui peuvent donc avoir un rôle majeur dans la fixation d'un ligand), ou moins conservés mais apportant de la spécificité.

2) Les méthodes basées sur la dynamique moléculaire ont pour objectif d'inclure des données de flexibilité dans la création d'un modèle de pharmacophore. Pour cela, le comportement de sondes chimiques (eau ou solvant organique) sur la surface flexible de la protéine[308,309] est simulé. Cependant, l'utilisation de solvant organique pour la détection de région hydrophobe induit des changements de conformations locales qui sont peu probables *in vivo*[306].

3) Les méthodes basées sur les grilles tridimensionnelles sont les plus utilisées. Le principe de ces méthodes est de placer la zone d'intérêt de la protéine dans une grille tridimensionnelle afin d'identifier les points de la grille associés aux énergies les plus favorables[306]. Elles se divisent en 4 étapes : la définition d'une cavité et la génération d'une grille autour de cette cavité, le calcul d'énergie en chaque point de la grille, le filtrage des points afin d'éliminer ceux trop éloignés du site et ceux associés à des énergies jugées insuffisantes et enfin un clustering des points. Pour définir la position de la grille, certains outils autorisent l'utilisateur à définir des résidus faisant partie du site de liaison à étudier (Pocket v.2[310]) alors que d'autres possèdent un algorithme de détection de cavité (Ph4Dock[311], LigandScout[305], BioGPS[312]). Pour chaque type d'interaction, les

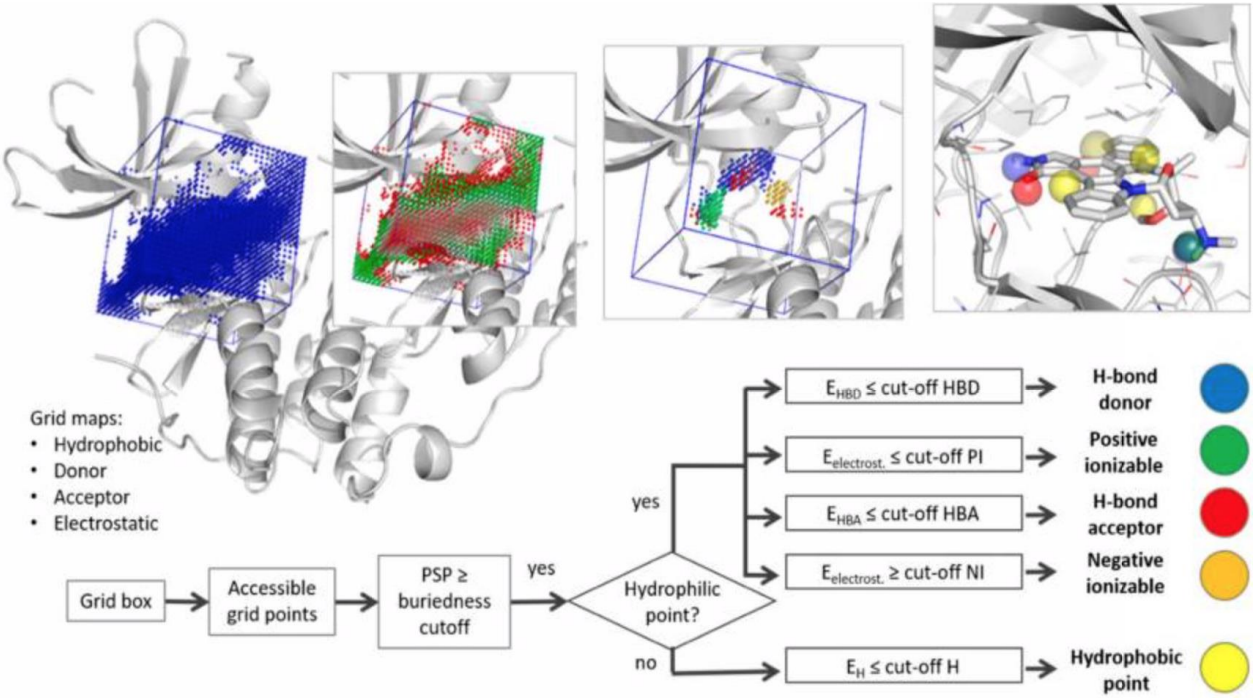


Figure 31 Schéma du protocole de modélisation de pharmacophore focalisé sur la structure implémenté dans T²F. D'après [300].

points d'énergie les plus favorables sont conservés puis clusterisés pour définir une région d'interaction qui sera traduite en point pharmacophorique (Tableau 7). Par exemple, T²F construit une grille à partir d'un centre de masse défini manuellement en se basant sur ses propres connaissances ou sur des cavités prédites par des outils de détection de poches (Cf 4.2) (un ligand connu de la protéine peut aussi être utilisé)[306]. Une grille d'énergie est ensuite générée pour chaque type de sonde (un carbone aliphatique pour identifier les contacts hydrophobes, un hydrogène pour identifier les accepteurs de liaison hydrogène, un oxygène pour identifier les donneurs de liaison

hydrogène et un groupe chargé pour calculer les énergies électrostatiques) (Figure 31). Un seuil d'énergie est défini pour chaque type de sonde, ce qui permet de ne conserver que les points associés à des énergies très favorables et d'éliminer les clashs stériques. Les points éloignés de la surface, qui correspondent aux points peu enfouis, sont éliminés. Les points restants sont clusterisés pour former les points pharmacophoriques puis un second clustering est appliqué pour les gros clusters de sorte à dissocier les points pharmacophoriques voisins. Un autre exemple est l'algorithme GRAIL qui repose sur une méthodologie similaire à la différence que les scores associés à chaque point de la grille ne sont pas des énergies mais un score directement lié au respect des contraintes géométriques en ce point selon le type de sonde. GRAIL présente l'avantage d'être applicable à des dynamiques moléculaires où chaque image (ou *frame*) est analysée a posteriori, ainsi les sondes n'influent pas sur le repliement local de la protéine.

Tableau 7 Liste de méthodes de modélisation de pharmacophores basées sur le site de liaison faisant appel à des méthodes basées sur des grilles. D'après [306]

Méthode	Détermination de la cavité	Approche	Méthode de clustering	Evaluation	Année
Ph4Dock[311]	Détection de cavité (Triangulation de Delaunay/ sphères α)	Interaction électrostatiques (MMFF94) d'atomes (leurre) chargés	Single-linkage	Jeu de validation CCDC/Astex	2004
Pocket V2[310]	Grille définie autour du ligand (ou à partir de résidus renseignés par l'utilisateur)	Grille (Score)	Non renseignée	CDK2, HIV1-PR, ER, 17b-HSD	2006
FLAP[313] **+ BioGPS**[312]	Grille définie autour du	Grille (GRID software)	Énergie minimum par	Jeu de données de Patel, DUD	2007

	ligand ou detectée avec FLAPsite		région		
Tintori et al.[314]	Grille définie autour du ligand	Grille (GRID software)	Pas de grille (GRID minima + interpolation)	TrxR (MTB), HIV1 IN, HIV-1 RT dimer	2008
Hydro-Pharm	Grille définie autour du ligand (3 Å)	Grille (ChemScore) + MD	k-means	HIV1-PR, DHFR, FXa	2012
PharmDock[316]	Grille définie autour du ligand (3 Å)	Grille (ChemScore)	k-means	PDB bind, DUD	2014
T²F-Pharm[306]	Grille définie autour du ligand ou à partir d'un centre défini par l'utilisateur	Grille (AutoDock)	CNN	Jeu de données de Patel + A2Areceptor	2018

4.4.2 Docking protéine-ligand

Le docking est la méthode de criblage basée sur la structure la plus utilisée depuis les années 1980[31759]. Il peut être utilisé pour modéliser les interactions entre une petite molécule et une protéine au niveau atomique et ainsi permettre de comprendre le mode de liaison de la petite molécule et d'élucider des processus biologiques fondamentaux[59]. Le processus de docking se divise en 2 étapes. La première étape appelée échantillonnage consiste à prédire la pose du ligand dans le site de liaison c'est-à-dire sa conformation ainsi que sa position, et la deuxième consiste à associer un score à chaque pose générée. Idéalement, l'échantillonnage doit permettre de reproduire le mode de liaison expérimental de la petite molécule, appelée pose native, et la fonction de score doit permettre de classer la pose la plus proche de la pose native parmi les meilleures. Le premier modèle d'interaction a été introduit par Fischer[318] ; il s'agit de la théorie serrure-clé, la petite

molécule s'insérant dans la protéine telle une clé dans une serrure. Cette théorie suggère un mode de liaison entre deux corps rigides et a donné naissance à des méthodes de docking dites « rigides ». Plus tard, la notion d'ajustement induit a été introduite par Koshland[319] pour décrire l'adaptation mutuelle (et donc des changements conformationnels) de la petite molécule et de la protéine lors de l'interaction. Les techniques de docking ont donc évolué pour tenter de prendre en compte la flexibilité du ligand et de la protéine[59]. La méthode la plus populaire et présentant le meilleur équilibre efficacité/temps de calcul est, à l'heure actuelle, le docking effectué avec un ligand flexible et une protéine rigide. De plus en plus de stratégies permettent néanmoins de prendre en compte la flexibilité locale ou globale de la protéine.

4.4.2.1 Echantillonnage

L'ensemble des modes de liaisons possibles d'une petite molécule dans un site de liaison devrait tenir compte des 6 degrés de liberté de translation et de rotation de la petite molécule ainsi que de ses nombreux degrés de liberté conformationnels liés aux liaisons rotatives. Générer l'ensemble des modes de liaison possibles devient donc extrêmement coûteux en temps de calcul dès lors que la taille du site de liaison et la flexibilité de la petite molécule sont importantes. Différents algorithmes d'échantillonnage sont utilisés pour pallier ce problème : les algorithmes stochastiques et les algorithmes de construction incrémentale.

4.4.2.1.1 Algorithmes stochastiques

Les méthodes stochastiques sondent l'espace conformationnel en modifiant de façon aléatoire la conformation, la position et l'orientation d'un ligand. Les méthodes les plus communément utilisées sont les algorithmes de Monte Carlo (MC) et les algorithmes génétiques (GA). La méthode de Monte Caro génère différentes poses d'une petite molécule en faisant subir des rotations aux liaisons rotatives des petites molécules ainsi que des rotations et translations à la molécule rigide. Les poses obtenues sont ensuite conservées ou non pour de nouvelles étapes d'optimisation en fonction d'un critère de sélection semi-aléatoire basé sur leur énergie. Si l'énergie est favorable, la conformation en question a de fortes chances d'être conservée, si l'énergie est défavorable, ses chances sont plus faibles. La petite molécule subit ainsi plusieurs itérations jusqu'à ce que 1) le nombre d'itérations imposé ou maximal soit atteint, ou 2) jusqu'à ce que les énergies obtenues convergent. ICM[320], QXP[321] et Affinity[322] utilisent cette méthode d'échantillonnage qui a l'avantage de sonder un espace conformationnel large et de permettre au ligand de franchir des

barrières énergétiques. Les algorithmes génétiques, quant à eux, se basent sur la théorie de l'évolution : les degrés de liberté sont encodés en tant que « gènes » et sont regroupés en ensemble ou « chromosome » correspondant à la pose de la petite molécule. L'algorithme génétique s'effectue en plusieurs itérations au cours desquelles les chromosomes peuvent subir des mutations au niveau de certains gènes ou bien des recombinaisons (ou « *crossover* ») qui correspondent à des échanges de gènes avec un autre chromosome. Les nouvelles conformations générées sont conservées pour les prochaines itérations si elles dépassent un seuil d'affinité imposé. AutoDock[323], AutoDockFR[324] et GOLD[325] utilisent cet algorithme. D'autres outils comme rDock combinent des algorithmes génétiques et de Monte Carlo.

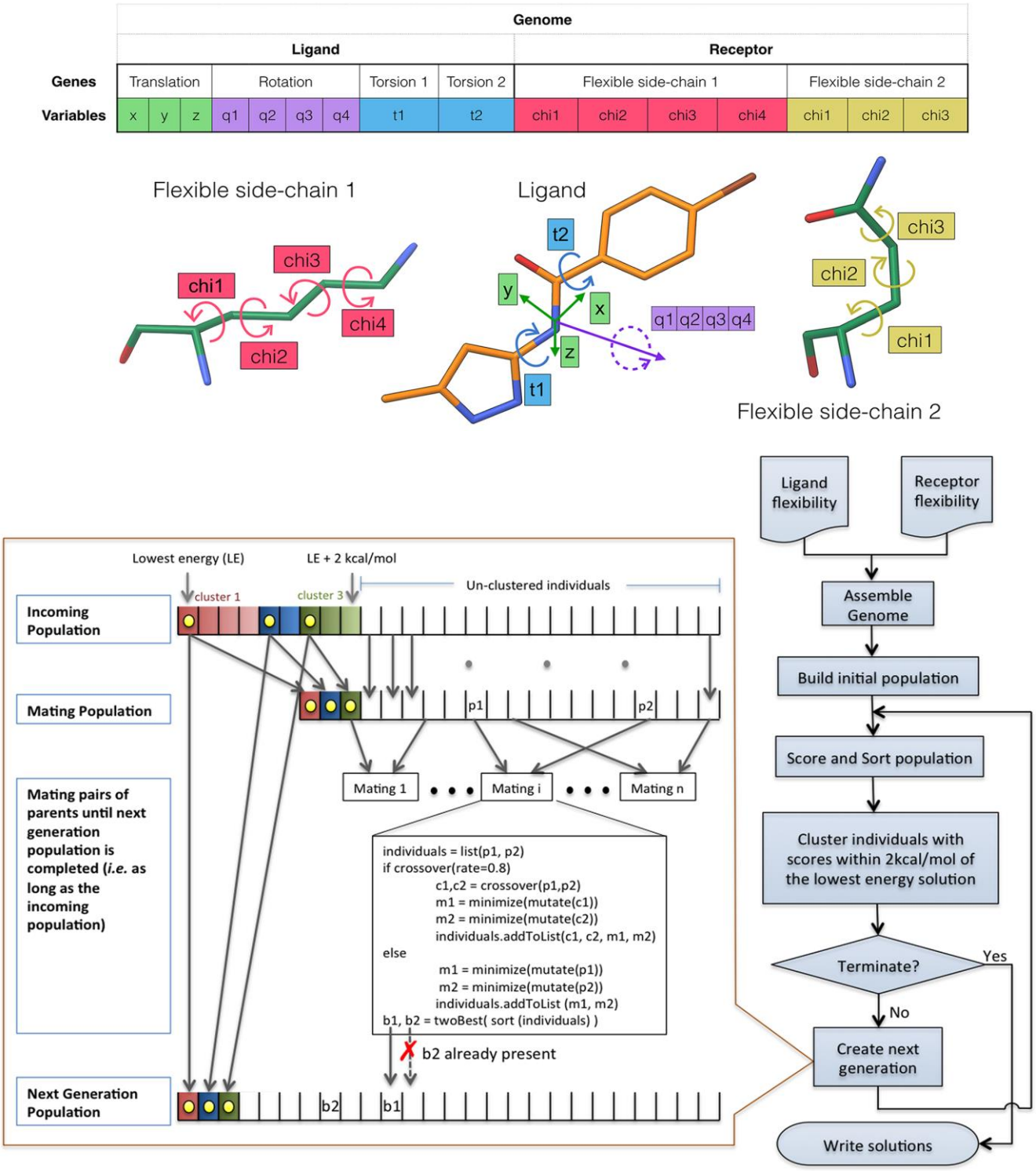


Figure 32 Représentation schématique de l'algorithme d'AutoDockFR. AutoDockFR effectue un docking avec la petite molécule et les chaînes latérales des résidus du site de liaison flexibles. La flexibilité de ces deux derniers est encodée dans le « chromosome » de l'algorithme génétique. Une fois la population construite aléatoirement, les conformations sont scorées et clusterisées. Des opérations génétiques (mutation et *crossovers*) sont appliquées et les nouvelles conformations de bonne énergie sont utilisées pour la génération suivante. D'après [318]

4.4.2.1.2 Algorithmes de construction incrémentale

Les méthodes de construction incrémentale reposent sur la déconstruction de la petite molécule en fragments et sa reconstruction progressive durant le docking[105]. La petite molécule est fragmentée par rupture de ses liaisons rotatives et un fragment est sélectionné selon des critères dépendant du protocole (ex : plus gros fragment, fragment ayant un rôle fonctionnel, ou bien sélection aléatoire).

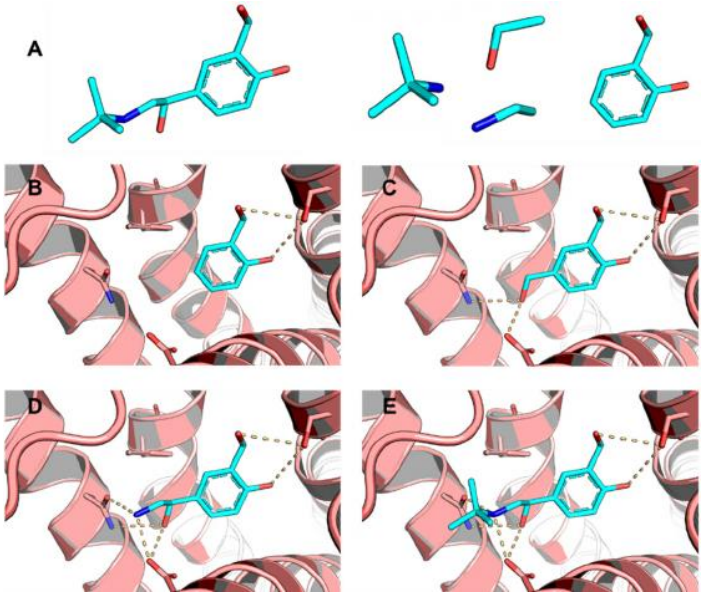


Figure 33 Génération de poses par construction incrémentale. (A) La petite molécule est coupée fragmentée, les cassures étant faites au niveau des liaisons rotatives. (B) Le fragment d'ancrage est docké dans le site de liaison, (C) un autre fragment docké avec les contraintes du fragment précédent et ainsi de suite jusqu'à la reconstruction complète du ligand (D et E). D'après [100]

Le fragment initial, dit fragment d'ancrage, est docké dans le site de liaison ; différentes poses sont alors proposées. Les fragments suivants sont dockés les uns après les autres en tenant compte des contraintes des fragments précédents. Un score est calculé pour les poses intermédiaires ainsi que les poses finales générées et permet de proposer différentes solutions. DOCK utilise cette approche : il propose une construction incrémentale de la molécule en partant de différents fragments d'ancrage pour optimiser l'échantillonage[326]. Les méthodes de Multiple Copy Simultaneously Search[308,327] (MCSS) et LUDI[328] proposent des algorithmes similaires. MCSS place des milliers de copies d'un groupement fonctionnel dans le site de liaison et y associe un

score. Des portions du site de liaison énergétiquement favorables pour le fragment sont identifiées. La démarche est répétée pour plusieurs fragments. Une molécule affine pour le site de liaison peut finalement être générée en reliant les fragments d'énergies favorables. LUDI se focalise sur les sites d'interactions capables de former des liaisons hydrogènes ou des interactions hydrophobes définis à partir de banques de données ou à partir de règles précises. Les fragments sont dockés dans ces sites d'interaction ; l'étape finale consiste à connecter les fragments dockés entre eux.

4.4.2.1.3 Autres méthodes

Il existe d'autres méthodes pour échantillonner les conformations possibles d'une petite molécule dans le site de liaison d'une protéine. Les méthodes de dynamique moléculaire en font partie. Ces méthodes reposent sur l'équation du mouvement décrite par Newton (Équation 1) pour simuler l'évolution d'un système récepteur-ligand au cours du temps à l'échelle atomique. Une vitesse initiale est attribuée à chaque atome afin de débuter la simulation[329,330]. La position de chaque atome est recalculée à intervalle de temps très court en fonction des forces qui lui sont appliquées. Ces forces sont dérivées des énergies potentielles et cinétiques.

$$\frac{d^2 r_i}{dt^2} = \frac{F_i}{m_i}$$

Équation 1 Formule de l'équation de Newton utilisée pour calculer la position d'un atome i dans un intervalle de temps $\frac{d^2 ri}{dt^2}$ en fonction de la masse de l'atome $\boldsymbol{m_i}$ et de l'ensemble des forces qui lui sont appliquées $\boldsymbol{F_i}$

Ces méthodes de dynamique moléculaires permettent de tenir compte de la flexibilité de la protéine et du ligand et d'étudier la stabilité des interactions. En revanche, elles sondent généralement des conformations correspondant à des minimums d'énergie locaux et n'échantillonnent pas suffisamment le paysage énergétique d'un complexe pour identifier le minimum d'énergie global. Leur faible capacité à franchir de hautes barrières énergétiques[59] les rendent inadéquates pour sonder des conformations variées du complexe. Les stratégies les plus fréquemment utilisées font appel aux méthodes stochastiques suivies d'une étape d'optimisation par dynamique moléculaire[59].

4.4.2.2 Fonctions de score

Les fonctions de score jouent un rôle majeur dans le docking en permettant le classement final des poses proposées lors de l'échantillonage. Une fonction de score idéale devrait permettre de calculer l'énergie libre de liaison d'une petite molécule sur une protéine. Or, l'énergie libre est très difficile à estimer notamment puisqu'elle prend en compte l'enthalpie et l'entropie du système et qu'elle est très sensible à la précision de la pose générée[331]. La plupart du temps des simplifications sont faites, l'entropie est notamment peu prise en compte[332].

Il existe des 3 types de fonctions de score (empiriques, basées sur la connaissance ou encore basées sur les champs de forces) qui permettent d'estimer l'affinité d'une molécule pour la cible étudiée (Tableau 8). Ces fonctions de score partagent un objectif commun : être capable de discriminer les bonnes poses des mauvaises et les molécules actives des inactives[59].

Tableau 8 Exemple de fonctions de score utilisées pour évalue une interaction petite molécule/protéine.

Fonction de scores empiriques	**Fonctions de score basées sur la connaissance**	**Fonction de score basée sur les champs de force**
ChemScore[333]	ITScore[334]	DOCK[335]
FlexX[336]	PMF[337]	AutoDock[323]
GlideScore[338]	DrugScore[339]	GoldScore[325]
LUDI[328]	KECSA[340]	COMBINE[341]
ICM[320]		MedusaScore[342]
Surflex[218]		

4.4.2.2.1 Fonctions de score empiriques

Les fonctions de score empiriques tentent de reproduire les affinités de liaisons expérimentales. Elles peuvent être décrites comme des modèles de régressions linéaires prenant en compte un ensemble de termes énergétiques importants dans l'établissement d'une interaction protéine-ligand[343]. Ces termes peuvent représenter par exemple, les liaisons hydrogènes, les interactions avec les ions métalliques et les contacts hydrophobes qui contribuent positivement au score final dans la fonction de ChemScore[333] implémentée dans GOLD, ou alors les contraintes imposées au

niveau des liaisons rotatives, les clash stériques et les contraintes internes qui sont pénalisants (Équation 2).

$$ChemScore = S_{H-bond} + S_{metal} + S_{hydrophobie} + P_{liaisons\ rotatives} + P_{contraintes\ internes} + P_{clash}$$

Équation 2 Formule de la fonction de score implémentée dans ChemScore. Les liaisons hydrogènes, métalliques ainsi que les contacts hydrophobes contribuent positivement au score final alors que les contraintes imposées aux liaisons rotatives, les contraintes internes et les clashs stériques sont pénalisants.

Certaines fonctions de scores empiriques sont plus sophistiquées que d'autres. Par exemple, GlideScore-XP différencie les liaisons hydrogènes en non chargée/non chargée, chargée/chargée et chargée/non chargée et considère les contacts hydrophobes ainsi que l'enfouissement hydrophobes comme deux termes distincts. Il faut noter que certains motifs d'interaction comme les interactions cation-π sont souvent omises malgré leur importance.

Afin de reproduire les affinités de liaisons expérimentales, les fonctions de score empiriques sont calibrées sur des jeux de données associant des complexes tridimensionnels à des données d'affinité expérimentales (PDB, Binding Mother Of All Databases[344] (Binding MOAD) et PDBbind[345,346]). A partir de ces données expérimentales, les facteurs impliqués dans l'établissement de l'interaction peuvent être déduits et un poids peut leur être associé. La diversité des complexes du jeu d'apprentissage assure l'applicabilité de la fonction de score à des familles de protéines et des chemotypes variés. En revanche, il est important de s'assurer de la qualité et de la fiabilité des données utilisées puisque de larges variations d'affinité peuvent être mesurées en fonction de la méthode expérimentale appliquée et de la précision des instruments de mesure[81].

Il faut noter que certaines fonctions de score peuvent être manuellement modifiées, notamment au niveau des poids attribués à chaque terme, afin d'être adaptées au sujet d'étude de l'utilisateur. C'est le cas de smina qui permet de pondérer les termes calculés par AutoDock VINA[347].

4.4.2.2.2 Fonctions de score basées sur les champs de force

Les champs de forces ont commencé à être utilisés pour calculer les interactions ligand/protéine depuis le travail fondateur de Martin Karplus dans les années 1970s[348,107]. Ces fonctions incluent des termes d'énergies intermoléculaires comme des forces de van der Waals et l'énergie

électrostatique et des termes d'énergies intramoléculaires (déformation des liaisons, déformation des angles de valences, et déformation des angles dièdres ou angles de torsion)[349,350] (Équation 3).

Équation 3 Formule générale de l'expression d'un champ de force. Les 3 premiers termes correspondent aux contributions intramoléculaires à l'énergie totale du système. Les deux autres décrivent les interactions intermoléculaires via l'énergie de van der Waals et l'énergie électrostatique. Ici l'énergie de van der Waals est décrite par la formule du potentiel de Lennard-Jones 12-6 et l'énergie électrostatique est décrite par la loi de Coulomb.

$$E = \sum_{Liaisons} K_L(r - r_O)^2 + \sum_{Valences} K_V(\theta - \theta_O)^2 + \sum_{Tosrions} K_T[1 + \cos(n\varphi - \varphi_0)] + \sum_{VDW} 4\varepsilon_{ij}\left(\frac{{\sigma_{ij}}^{12}}{{r_{ij}}^{12}} - \frac{{\sigma_{ij}}^{6}}{{r_{ij}}^{6}}\right) + \sum_{Coulomb} \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}}$$

$\boldsymbol{K_L, K_V}$ **et** $\boldsymbol{K_T}$ sont les facteurs de pénalités pour les écarts de distances de liaisons, d'angles de valences et des angles de torsions par rapport à une valeur idéale, r et $\boldsymbol{r_O}$ sont les distances de liaison mesurées et de référence, $\boldsymbol{\theta}$ et $\boldsymbol{\theta_O}$ sont les angles de valence mesurés et de référence, $\boldsymbol{\varphi}$ et $\boldsymbol{\varphi_0}$ et $\boldsymbol{\varphi_0}$ sont les angles de torsion mesurés et de référence, $\boldsymbol{\varepsilon}$ est la profondeur du puit de potentiel, $\boldsymbol{\sigma_{ij}}$ est le rayon de van der Waals[350].

Par exemple, DOCK[335] et AutoDock[323] utilisent des fonctions d'énergie basées sur le champ de force AMBER. Les fonctions de score basées sur les champs de force incluent généralement des termes de solvatation grâce à des modèles de solvatation implicite (Poisson-Boltzmann (PB) et Born (GB)). D'autres méthodes[351] permettent de prendre en compte le solvant de manière explicite (Linear Interaction Energy[352] (LIE) et Linear Response Approximation (LRA)) mais sont plus chronophages et sont privilégiées dans des phases d'optimisation de *lead* plutôt que de criblage haut débit[352]. Ces méthodes sont appliquées en aval de la phase de docking pour re-scorer des complexes ligand/protéine. Parmi elles, le MM-PBSA et MM-GBSA sont les plus connues et ont montré leur capacité à améliorer la fréquence de hits[353] et à mieux classer des ligands d'une protéine les uns par rapport aux autres[354]. Prédire l'énergie libre de liaison reste néanmoins complexe et la prise en compte de l'entropie est très souvent omise. Par ailleurs, la sensibilité de ces méthodes n'est pas toujours suffisante pour classer correctement des molécules d'énergie similaire de manière prédictive. Une solution pour réduire les marges d'erreurs dans l'estimation de l'affinité

est d'inclure de manière explicite la contribution de l'énergie de polarisation[355]. La polarisation correspond à la réorganisation de la distribution des charges d'une molécule lors de l'établissement d'interactions électrostatiques avec une ou plusieurs molécules environnantes, et est en partie responsable de la non-additivité de l'énergie globale[355]. Par exemple, lorsque 3 molécules non chargées interagissent entre elles, l'énergie d'interaction globale est supérieure à la somme des énergies 2 à 2[355].

4.4.2.2.3 Fonctions de score basées sur la connaissance

Les fonctions de score basées sur la connaissance sont souvent nommées « potentiels statistiques » puisqu'elles s'appuient sur une notion issue de statistique mécanique qui permet de transformer la distribution des distances interatomiques en potentiel de force moyenne. Dans un premier temps, les distances interatomiques entre les atomes du ligand et de la protéine d'étude sont mesurées. La densité de ses distances est comparée à la densité de distance observée dans une base de données de référence. L'équation de Boltzmann est utilisée pour transformer les distributions obtenues en score d'affinité (Équation 5).

Équation 4 Formule générale des fonctions de score basées sur la connaissance. Le score (A) est calculé comme la somme des potentiels statistiques $\boldsymbol{w_{ij}(r)}$ entre les atomes du ligand et de la protéine.

$$A = \sum_{i}^{ligand} \sum_{j}^{protéine} w_{ij}(r)$$

Équation 5 Formule du potentiel statistique $\boldsymbol{w_{ij}(r)}$. Ce potentiel dépend de la densité numérique des paires i-j à une distance r ($\boldsymbol{\rho_{ij}(r)}$), et de la densité numérique de la paire d'atome dans une banque de référence où les interactions interatomiques sont supposées nulles.

$$w_{ij}(r) = -K_B T ln\left[\frac{\rho_{ij}(r)}{\rho *_{ij}}\right]$$

De cette manière, si un contact est observé plus fréquemment que dans l'état de référence, il est considéré comme énergétiquement favorable, alors que lorsqu'il est moins observé, il est jugé défavorable. Ces fonctions sont apprises sur de larges jeux de données issus de la PDB. Contrairement aux fonctions de scores empiriques, elles ne considèrent jamais les affinités de liaisons expérimentales. Elles sont donc moins dépendantes des erreurs de mesure expérimentales,

mais très dépendantes des techniques de résolution de complexes tridimensionnels. L'abondance de données tend néanmoins à minimiser l'introduction d'erreur dans la fonction de score. PMF[337], DrugScore[339], ITScore[334] et KESCA[340] sont des exemples de fonctions de scores basées sur la connaissance.

4.4.2.2.4 Fonctions de score consensus

Chaque fonction de score possède ses spécificités propres ainsi que ses imperfections : aucune d'entre elle ne permet d'associer systématiquement le meilleur score à la conformation bioactive d'un complexe ligand-protéine. Des scores consensus combinant différentes fonctions de scores ont été proposés afin de tirer profit de l'information apportée par chacune d'entre elles. Ainsi, une pose peut être privilégiée si elle est associée à de bons scores avec différentes méthodes de calcul. L'utilisation de scores consensus a prouvé à plusieurs reprises son intérêt en améliorant les performances en terme d'enrichissement lors de criblage virtuel à haut débit et pour identifier la conformation générée par docking la plus proche de la pose bioactive[356357358]. Généralement, un score consensus est dérivé de la somme des rangs obtenus avec différents outils de docking ou de scoring (« sum-rank »), du rang minimal obtenu (« min-rank ») ou de la moyenne des rangs dépréciée du meilleur rend obtenu (« deprecated mean-rank »)[359]. Il faut néanmoins veiller à ce que les fonctions de score ne soient pas corrélées pour les utiliser conjointement.
D'autres fonctions de score se basent sur des analyses QSAR (Cf 3.2) et considèrent un ensemble pondéré de descripteurs dans le calcul du score d'affinité.

4.4.2.3 Docking rigide

Les premiers algorithmes de docking considéraient la petite molécule et la protéine d'étude comme deux corps rigides (ex : DOCK[317,360] et FLOG[361]). Ainsi, la recherche conformationnelle ne prenait en compte que 6 degrés de liberté (3 de rotation et 3 de translation). La flexibilité de la petite molécule était simulée par la génération de plusieurs conformères en amont du processus de docking ou en autorisant des chevauchements entre atomes de la petite molécule et de la protéine[59]. Par exemple, les premières versions de DOCK[317,360] définissaient la petite molécule et la protéine comme un ensemble de sphères. Un algorithme de détection de clique permettait d'associer les deux corps rigides en considérant la correspondance géométrique et chimique. Le complexe obtenu pouvait être évalué par calcul de la correspondance stérique, de complémentarité chimique ou de similarité pharmacophorique. Les méthodes de docking rigide simplifient énormément le processus

d'établissement de liaison *in vivo* puisque les systèmes biologiques ne sont jamais figés. La prise en compte de la flexibilité du ligand est un premier pas vers une meilleure modélisation de la liaison d'une molécule à une protéine et elle est notamment essentielle si l'on considère l'ajustement induit par l'interaction entre le ligand et la protéine. Un compromis entre la précision de la méthode et le temps de calcul associé est de considérer la petite molécule comme flexible et la protéine comme un corps rigide. C'est notamment le cas dans les versions les plus récentes de DOCK, pour lesquelles la flexibilité du ligand est prise en compte via l'algorithme d'échantillonnage par

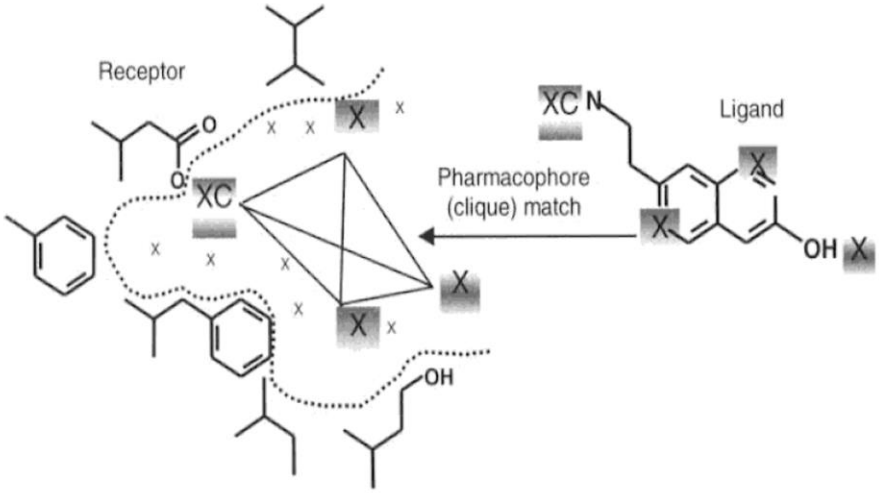


Figure 34 Schéma de l'algorithme de docking rigide implémenté dans l'une des premières versions de DOCK. La première version de l'algorithme de docking de DOCK se basait sur la détection de clique. Plus tard, des contraintes de correspondance pharmacophoriques ont été ajoutées (les points pharmacophoriques matchés sont représentés en gris). L'utilisateur pouvait imposer des contraintes d'interaction lorsqu'un site était caractérisé comme critique dans l'établissement d'une interaction et/ou le déclenchement d'une activité. D'après [354]

construction incrémentale et la recherche exhaustive de ses conformères, mais aussi dans d'autres logiciels très utilisés comme AutoDock[323], Autodock VINA[215] et FlexX[336].

4.4.2.4 Prise en compte de la flexibilité de la structure

Il a été montré que la mobilité intrinsèque de la protéine est intimement liée au comportement de son ligand lorsqu'une interaction est établie[362]. Deux grandes hypothèses prédominent sur la flexibilité lors de l'interaction protéine/ligand. Il est difficile de déterminer avec certitude si le

ligand interagit avec une conformation favorable de la protéine parmi l'ensemble de ses conformations possibles, ce que l'on appelle « sélection de conformère », ou si au contraire la protéine adopte une conformation qu'elle n'adopterait pas spontanément au contact du ligand, ce que l'on appelle « induction conformationnelle ». Afin de simuler cette flexibilité, des simulations de dynamique moléculaire devraient idéalement permettre de modéliser l'ensemble des degrés de liberté du complexe. Cependant, les méthodes actuelles de dynamique moléculaire présentent des faiblesses pour échantillonner les degrés de liberté du complexe, et elles sont extrêmement chronophages en comparaison des méthodes de criblage virtuel à haut débit[59]. Elles sont de ce fait inapplicables pour le criblage de grandes chimiothèques. Plusieurs stratégies ont été développées pour pallier ce problème : le docking d'ensemble, qui prend en compte plusieurs conformations de la protéine, le « soft-docking » qui minimise les termes de répulsion de l'énergie de van der Waals pour autoriser des chevauchements d'atomes entre la petite molécule et le site de liaison, ou le docking flexible, qui prend en compte la flexibilité du ligand et des résidus du site de liaison (Tableau 9).

Tableau 9 Liste de différentes méthodes permettant d'appréhender la flexibilité d'une protéine, leurs avantages et leurs inconvénients

Méthode	**Description**	**Avantages**	**Inconvénients**	**Logiciels**
Potentiel lissé	Modifie le terme de van der Waals pour autoriser le chevauchement entre atomes de la petite molécule et du récepteur	Méthode simple facile à mettre en place et à combiner avec d'autres approches	Prise en compte de la flexibilité simpliste (implicite et non quantitative) et inadéquate	AutoDock[323]
Librairie de rotamères	Cherche des conformations possibles de chaînes latérales dans des bibliothèques	Méthode simple qui ne nécessite pas d'étape d'optimisation	Dépend de la base de données utilisée, pas de flexibilité du squelette protéique	ICM[320]

	dédiées			
Flexibilité des chaînes latérales des récepteurs	Echantillonne la flexibilité du ligand et du site de liaison simultanément en utilisant un GA	Méthode facile à mettre en place, modélise l'effet du ligand sur le réarrangement structural local	Seules les chaînes latérales sélectionnées sont prises en compte, ne prends pas en la flexibilité du squelette protéique	GOLD[325], AutoDock 4[363]
Ensemble de conformations d'une protéine	Docke la petite molécule dans une série de structures du récepteur qui représentent différents états conformationnels	Prise en compte complète et explicite de la flexibilité	Computationnelement coûteux, et limité aux conformations inclues dans le jeu de données.	DOCK[317], FlexE[364]

4.4.2.4.1 Soft-docking

Au lieu de modéliser de manière explicite les changements conformationnels de la protéine lors du docking, le docking lissé, ou soft-docking, prend en compte la flexibilité du complexe en tolérant des clashs stériques mineurs entre les atomes de la petite molécule et du récepteur. La première méthode de soft-docking développée par Jiang représente la surface du ligand et de la protéine sous forme de cubes[365]. Lors de l'étape de docking réalisé par complémentarité de surface, un chevauchement mineur des deux surfaces est autorisé. Cependant, ce type de méthode dit géométrique est peu utilisé pour le docking ligand/protéine. Il est plus courant d'utiliser un potentiel lissé qui consiste à autoriser les chevauchements atomiques en atténuant le terme répulsif de l'énergie de van der Waals de la fonction de score. Généralement, les interactions de van der Waals sont calculées via le potentiel de Lennard Jones 12-6. Dans le cas du soft docking, le potentiel de Lennard Jones est souvent modifié pour être plus tolérant lorsque les atomes sont proches[366] (Glide[338], DOCK[317,360], AutoDock[323]).

4.4.2.4.2 Docking flexible

Certains logiciels de docking prennent en compte la flexibilité d'un ensemble de résidus situés au niveau du site de liaison et le plus souvent définis par l'utilisateur lors de l'étape d'échantillonnage du docking. C'est le cas d'AutoDock4[363] et d'AutoDockFR[324] par exemple (Figure 32). Les deux outils utilisent un algorithme génétique dans lequel le chromosome ne représente pas uniquement la conformation du ligand mais bien la conformation du ligand et de résidus du site de liaison. Ainsi, les conformations du ligand et des chaînes latérales des résidus sont échantillonnées simultanément. Dans AutoDock4[363], ICM[367], RosettaLigand[368], Fleksy[369] et FITTED[370], le mouvement des chaînes latérales est limité à un ensemble de conformations énergétiquement favorables issu de banques de données de rotamères[371,372,246]. Fréquemment, le docking est effectué avec une protéine rigide puis le complexe généré est optimisé grâce à une étape de minimisation. Dans certains cas, la liaison d'un ligand implique des modifications structurales à une échelle globale de la protéine. Par exemple, la famille des récepteurs nucléaires possède une hélice qui subit de larges changements de conformations pour passer d'un état agoniste à antagoniste qui vont permettre la liaison des co-activateurs et des co-represseurs respectivement[373]. La plupart des outils de docking flexible ne sont pas adaptés pour de tels réarrangements et un docking d'ensemble est privilégié dans de tels cas de figure.

4.4.2.4.3 Docking d'ensemble

Le docking d'ensemble consiste à sélectionner un ensemble de structures d'une protéine représentant différents états conformationnels. Le ligand est docké dans une série de structures et les résultats sont combinés de différentes manières en fonction du logiciel. Par exemple, DOCK[317] génère un potentiel d'énergie moyenné sur l'ensemble des conformations considérées. FlexE[364] fusionne quant à lui les parties similaires de la protéine et considère les parties dissimilaires comme des alternatives différentes. Il utilise un algorithme de construction incrémentale au cours duquel les interactions avec chacune des alternatives observées sont déterminées. Les conformations

conservées sont celles associées au meilleur score. Il est fréquent d'effectuer un docking sur un ensemble de structures et de considérer comme score final soit 1) la meilleure énergie obtenue

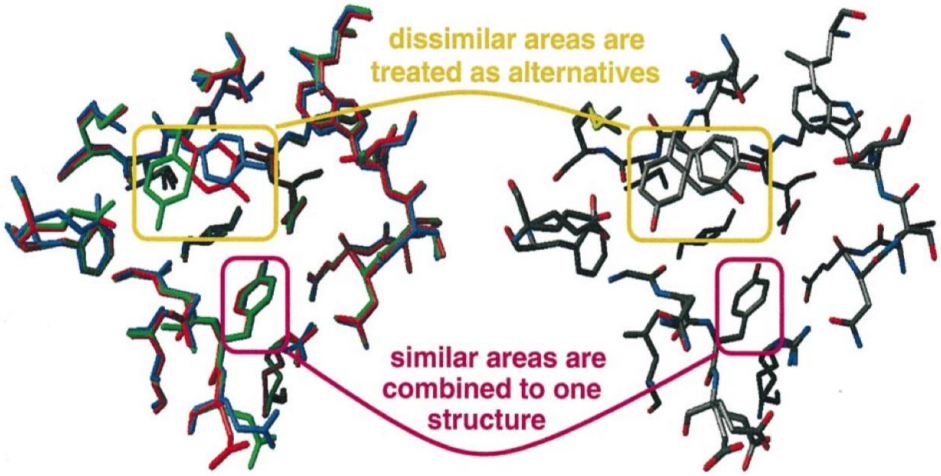


Figure 35 Schéma de la prise en compte de la flebilié avec le logiciel FlexE. L'algorithme de FlexE fusionne les parties similaires de protéines et considère les parties dissimilaires comme différentes alternatives lors du docking. D'après [399]

parmi toutes les structures, ou bien 2) l'énergie moyenne obtenue sur un ensemble de structures[374,211]. Dans ce cas, la prise en compte de la flexibilité est traitée *a posteriori* et ne nécessite pas un logiciel adapté. Il faut noter que l'espace conformationnel des structures inclues dans le jeu de données influe dramatiquement sur les résultats obtenus[59]. Les structures élucidées expérimentalement sont privilégiées mais elles peuvent être complétées par des structure issues de dynamique moléculaire ou bien d'analyse des modes normaux[371,59].

4.4.2.5 Importance du solvant

Une interaction ligand-protéine en milieu physiologique implique une multitude de paramètres environnementaux parmi lesquels le solvant (l'eau) se trouve être l'un des paramètres majeurs. Avant que le ligand n'atteigne sa cible protéique, ces deux entités sont entourées de molécules d'eau avec lesquelles ils établissent chacun des interactions. Pendant le processus de liaison, les interactions avec ces molécules d'eau sont rompues, ce qui conduit à une contribution enthalpique pénalisante et l'interaction entre le ligand et sa cible doit donc compenser cette perte d'énergie. Afin de modéliser ce phénomène, plusieurs logiciels incluent un terme de solvatation-désolvatation dans leur fonction de score (AMBER[320], HYDE[375]). De plus, des molécules d'eau, généralement

localisées dans des poches enfouies du site de liaison, jouent parfois un rôle stabilisateur de la structure protéique. Ces molécules d'eau peuvent alors être conservées lors de l'interaction et assurer des liaisons indirectes entre le ligand et la protéine, ou bien délogées, ce qui est entropiquement favorable mais enthalpiquement défavorable. La prise en compte de ces molécules d'eau peut aussi être incluse dans le docking.

4.4.2.5.1 Solvatation/ Désolvatation

L'énergie d'une liaison ligand/protéine dans un milieu aqueux peut être mesurée de différentes manières tant qu'elle respecte le cycle thermodynamique de liaison présentée en Figure 36. Les

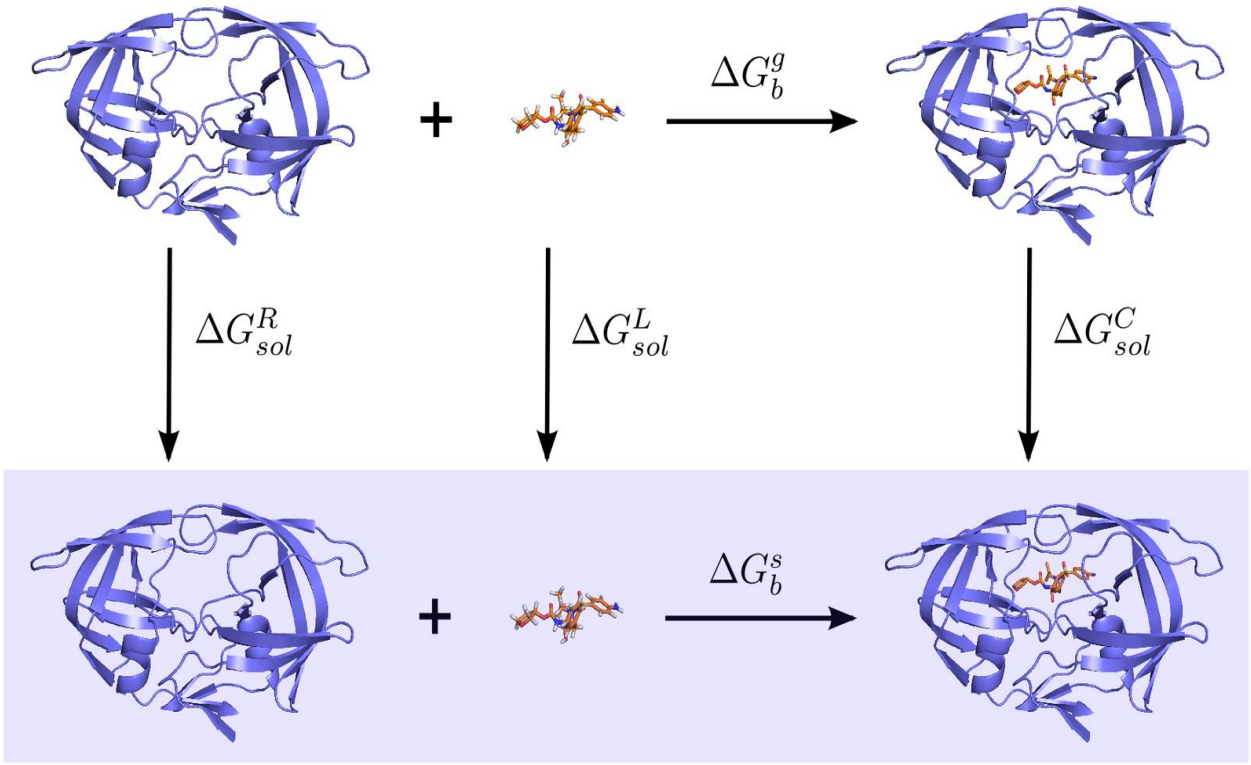


Figure 36 Schéma du cycle thermodynamique de liaison d'une petite molécule à une protéine. L'enthalpie libre de liaison dans le solvant $\Delta G_{solvent}^{bind}$ est équivalente à ΔG_{vacuum}^{bind} + ($\Delta G_{solv}^{\square}(C)$ - [$\Delta G_{solv}^{\square}(R) + \Delta G_{solv}^{\square}(L)$]). D'après Olivier Kuhn

approches MM-PBSA et MM-GBSA précédemment citées considèrent le solvant de manière implicite et estiment l'enthalpie libre de liaison dans le solvent comme $\Delta G_{solvent}^{bind} = \Delta G_{vacuum}^{bind}$ +

($\Delta G_{solv}(C)$ - [$\Delta G_{solv}(R) + \Delta G_{solv}(L)$]) (Cf Figure 36). Dans la méthode MM-GBSA, l'énergie électrostatique calculée est la somme de l'énergie électrostatique calculé via la loi de Coulomb et de l'énergie de Born Généralisée (GB) : $U_{elec} = U_{Coul} + U_{GB}$. L'énergie GB approxime l'interaction avec le solvant en incluant une constante diélectrique élevée (≈80) qui favorise le contact de résidus chargés avec le solvant. Les méthodes MM-GBSA et MM-PBSA sont utilisées pour re-scorer les poses issues de docking. Il en est de même pour la fonction de score HYDE, développée par l'équipe de Mathias Rarey, qui est utilisée *a posteriori* du docking. Dans cette approche, la déshydratation du ligand et de la protéine ainsi que l'énergie apportée par la formation de liaisons hydrogènes contribuent de façon majeure à l'énergie libre de liaison. La fonction de score HYDE contient deux termes distincts pour calculer l'énergie de déshydratation des atomes polaires et non polaires. Le terme d'énergie de déshydratation des atomes polaires prend en compte la probabilité d'un groupement hydrophile à établir une liaison avec une molécule d'eau dans sa position la plus favorable à l'établissement d'une liaison hydrogène. Ce terme permet de ne pas surestimer l'énergie nécessaire à la déshydratation du ligand. Par ailleurs, HYDE tient compte du fait que le réseau d'interaction des molécules d'eau impose une organisation qui ne permet pas de satisfaire l'ensemble des liaisons hydrogènes possibles lorsque plusieurs groupements polaires sont adjacents. Ainsi, le coût de la déshydratation et l'énergie nécessaire pour rompre ces liaisons hydrogènes sont plus faibles que dans un cas théorique idéal[213].

4.4.2.5.2 Molécules d'eau résiduelles

Les molécules d'eau résiduelles sont des molécules interagissant fréquemment avec les résidus du récepteur et assurant un rôle stabilisateur. Elles peuvent être délogées lors de l'interaction avec un ligand ou bien jouer un rôle d'intermédiaire d'interaction entre le ligand et le récepteur. En 2006, il était estimé que 65% des complexes ligand/protéine de la PDB impliquaient au moins une molécule d'eau résiduelle[376]. Plusieurs outils permettent d'identifier ce genre de molécules d'eau via des calculs énergétiques[377] ou géométriques[378]. Plusieurs outils de docking comme GOLD[325], SLIDE, FlexX ou rDock prennent en compte ces molécules d'eau. rDock[379] permet notamment à l'utilisateur d'imposer une localisation et une orientation de la molécule d'eau ou bien de lui laisser des degrés de liberté (rotation, translation ou les deux). Ces paramètres sont ensuite intégrés au chromosome utilisé dans l'algorithme génétique d'échantillonnage de pose implémenté dans rDock. SLIDE[380] utilise une approche basée sur la connaissance pour prédire la position des

molécules d'eau susceptibles d'être conservées dans un complexe ligand/protéine. FlexX[336] permet de prédire des positions probables des molécules d'eau résiduelles, plusieurs combinaisons de présence de molécules d'eau étant ensuite testées lors du processus de construction incrémental.

5 Évaluation des méthodes de criblage

L'objectif premier du criblage virtuel est d'identifier les molécules actives parmi une large collection de composés. Dans un cas de criblage idéal, l'outils maximise le nombre de vrais positifs retrouvés (sensibilité) et le nombre de vrais négatifs rejetés (spécificité). Ce cas de figure est très rare, et il incombe au chercheur de définir une balance sensibilité/spécificité acceptable. Lors d'étude de recherche de médicament, les faux positifs sont problématiques, et la spécificité est davantage prise en compte que la sensibilité. A l'inverse, lors d'études de prédiction de toxicité, ce sont les faux négatifs qui sont les plus problématiques et par conséquent, la sensibilité est privilégiée à la spécificité. Il est extrêmement important d'évaluer la performance d'un outil de criblage virtuel aussi bien lors du développement de cet outil que lors de son application. Lors du développement, son évaluation permet de paramétrer la méthode, de la comparer à des méthodes équivalentes et de la valider. L'évaluation en amont de l'application consiste généralement à évaluer et comparer plusieurs méthodes de criblage virtuel afin d'opter pour celle qui est la plus adaptée à l'application donnée. Ces évaluations dites rétrospectives sont faites sur des banques de données d'évaluation (ou banques de *benchmark*) qui regroupent des molécules actives et inactives[381] ; l'objectif est alors de tester si l'outil est capable d'identifier les molécules actives en tant que telles. La plupart du temps, du fait du manque de publication des données d'inactivité, les molécules inactives des banques d'évaluation sont remplacées par des molécules supposées inactives, appelées leurre ou « *decoys* »[382]. Le choix des *decoys* est un point très délicat qui va souvent de pair avec l'introduction de biais dans l'évaluation des outils de criblage[382]. Nous verrons que différents biais peuvent être introduits et mener soit à la surestimation soit à la sous-estimation des performances d'un outil. Les données d'inactivité étant plus abondantes pour certaines familles de protéines comme les tyrosines kinases, les GPCRs ou encore les récepteurs nucléaires[81], il est désormais possible d'intégrer des molécules inactives dans les banques d'évaluation. Par ailleurs, il existe une multitude de métriques d'évaluation des performances d'un outil, chacune mettant en avant des capacités différentes de l'outil[383]. Il est donc important de choisir une ou plusieurs métriques et une banque de données adaptées au système d'étude et aux résultats escomptés[381].

Lorsque des approches 3D basées sur le ligand ou sur la structure sont utilisées, il est important de s'assurer que la conformation ou la pose générée est en adéquation avec des conformations ou des pose observées expérimentalement. Pour cela, on évalue la distance moyenne (RMSD) entre les atomes de la conformation ou de la pose prédite et de conformation ou de poses recensées dans des banques de références comme la PDB[219] ou la Cambridge Structural Database[104]. Il existe également des métriques qui comparent les poses obtenues à la densité électronique issue de la cristallographie aux rayons X de sorte à s'affranchir des éventuelles erreurs d'interprétation de la position du ligand dans le maillage obtenu.

5.1 Bases de données d'évaluation

5.1.1 Premières banques d'évaluation

L'évaluation des méthodes de criblage virtuel est cruciale avant d'effectuer un criblage prospectif sur une large chimiothèque ; il est nécessaire de s'assurer que la méthodologie adoptée est appropriée pour le système étudié et est susceptible de générer des résultats fiables[381]. Les outils de criblage doivent donc être évalués rétrospectivement sur des banques de données d'évaluation avant d'être appliqués à une chimiothèque d'étude. Ces banques de données d'évaluation associent à une protéine cible un ensemble de molécules actives (issues de la littérature, de banques de données privées, académiques ou institutionnelles) et un ensemble de molécules inactives. L'objectif est alors de s'assurer que l'outil de criblage est capable de discriminer les molécules actives des molécules inactives. Le manque de publication des molécules inactives a conduit à remplacer celles-ci par des molécules supposées inactives dites « *decoys* »[384]. Par exemple, la première banque de benchmarck proposée par Bissantz et al. contient 10 molécules actives et 990 *decoys* sélectionnés aléatoirement parmi les composées du Advanced Chemical Directory (ACD) préalablement filtré pour éliminer les molécules réactives, inorganiques et de poids moléculaire extrême[385]. Cette banque a permis de comparer la performance de différents protocoles de docking et différentes fonctions de score. Afin de s'assurer que la capacité de discrimination des molécules actives n'était pas uniquement liée aux différences de poids moléculaire, Diller et al. ont ajouté des critères physicochimiques (poids moléculaire et polarité) de sélection des decoys[386]. Les *decoys* des premières banques d'évaluation consistaient donc en des molécules sélectionnées de manière aléatoire répondant éventuellement à des critères physico-chimiques spécifiques (molécules *drug-*

like, poids moléculaire, surface polaire etc.). Leur sélection pseudo-aléatoire justifiait leur faible probabilité d'exercer une activité sur la cible. Or, avec l'évolution des banques de données d'évaluation, de nombreux biais inhérents à la sélection des *decoys* ont été révélés[387,388,3,384], soulevant la nécessité d'évoluer vers une sélection rationnelle des *decoys* voire l'intégration de molécules inactives dans les banques d'évaluation (Tableau 10).

5.1.2 Biais inhérents à la sélection de *decoys*

Dès la publication des premières banques d'évaluation, la composition en molécules actives et en *decoys* a été désignée comme responsable de l'intégration de 3 biais majeurs : le biais d'analogie, le biais de complexité et l'intégration de faux négatifs. Les deux premiers peuvent mener à une surestimation artificielle des performances des outils de criblage, alors que la deuxième peut mener à une sous-estimation de ces performances.

5.1.2.1 Biais d'analogie

Les molécules actives référencées dans les banques de données et utilisées dans les banques d'évaluation sont généralement issues d'un nombre limité de séries chimiques. En conséquence, la variabilité structurale des molécules actives est très souvent réduite en comparaison aux *decoys* qui sont eux piochés dans de larges banques de données. La discrimination des molécules actives en est simplifiée et revient à identifier un ensemble restreint de séries chimiques surreprésenté parmi un ensemble de molécules de plus large variabilité structurale[389]. Le manque de diversité structurale des molécules actives ainsi que l'absence de critère de similarité imposé entre les molécules actives et les *decoys* constitue un cas simpliste d'évaluation d'un protocole de criblage qui peut conduire à une surestimation de ces performances[389].

5.1.2.2 Biais de complexité

Le biais de complexité est également lié au manque de cohérence entre les espaces chimiques couverts par les molécules actives et les *decoys* choisis[388]. Les molécules actives publiées ayant fait l'objet de plusieurs étapes d'optimisation, elles sont généralement plus complexes que les *decoys* piochés aléatoirement dans des banques de données n'imposant pas de contraintes physico-chimiques d'inclusion des molécules. Ce biais tend à simplifier la discrimination des molécules actives par rapport aux molécules inactives.

5.1.2.3 Faux négatifs

La sélection aléatoire des *decoys* expose au risque d'intégrer des molécules actives parmi les decoys. Or, la présence de molécules actives au sein des *decoys* peut avoir une influence dramatique sur l'évaluation de l'outil. Selon la proportion de molécules actives dans le jeu de decoys, le calcul de performance de l'outil de criblage est négativement impacté.

5.1.3 Sélection rationnelle des *decoys*

Plusieurs stratégies ont été développées pour minimiser le risque d'introduction de biais. Le biais de complexité est adressé en imposant une similarité physico-chimique entre les molécules actives et les *decoys*. Pour ce faire, la banque d'évaluation Maximum Unbiased Validation (MUV)[390] calcule l'enfouissement de molécules actives parmi les decoys en 4 étapes : 1) les molécules actives extraites d'essais de confirmation de HTS sont regroupées avec un ensemble B de *n* molécules extraites aléatoirement des banques de données DrugBank[391], Prous Drugs of the Future, le catalogue Sigma-Aldrich chemistry et la MDDR[392], 2) la distance entre chaque molécule active et la 500ème molécule la plus proche l'ensemble B est retenu, et l'opération est répétée 100 fois en tirant *n* molécules aléatoirement, 3) la distance d_{90} correspondant à l'intervalle de confiance à 90% est retenue comme distance de référence, 4) des decoys potentiels, qui dans la MUV sont des molécules n'ayant pas montré d'activité lors des essais primaires de HTS, sont sélectionnés ; toutes les molécules actives ayant une distance supérieure à d_{90} avec le 500ème *decoy* potentiel le plus proche sont exclues. Dans la DUD-E[1], les *decoys* sont sélectionnés en fonction de leur similarité avec 6 propriétés physico-chimiques des molécules actives (poids moléculaire, coefficient de partition octanol-eau, liaisons rotatives, accepteurs/donneurs de liaison hydrogène, charge nette). Dans la DEKOIS[393], les mêmes propriétés sont prises en compte auxquelles s'ajoutent le nombre de cycles aromatiques et une distinction de la charge en deux descripteurs comptabilisant le nombre de charges positives et le nombre de charge négative.

Le biais d'analogie peut être adressé en imposant une diversité structurale aux molécules actives en comparaison au jeu de *decoys*. La plupart du temps les molécules actives sont *clusterisées* et seul un représentant est considéré pour la suite de l'étude[394,395]. Dans la MUV[390], l'algorithme de Kennard Stone est utilisé afin d'assurer un recouvrement des espaces chimiques des molécules actives et des *decoys*.

Enfin, le risque d'inclure des molécules actives parmi le jeu de decoys est adressé en imposant une dissimilarité topologique entre les *decoys* et les molécules actives. Cette idée, amorcée dans la DUD[396], est aussi développée dans la DUD-E[1] à l'aide d'un filtre basé sur la distance de Tanimoto entre les empreintes ECFP4 des molécules actives et des *decoys* ce qui permet de sélectionner les *decoys* les plus dissimilaires. Dans la DEKOIS[3], le score Latent Actives in the Decoys Set (LADS) permet de quantifier ce risque. Pour chaque *decoy*, le LADS prend en compte l'ensemble des fragments communs aux molécules actives pondéré par leurs nombres d'atomes et leur fréquence dans la population de molécules actives, le tout par rapport au nombre de fragments total du *decoy* (Équation 6).

Équation 6 Formule du calcul du score LADS. $\boldsymbol{n}$ est le nombre de fragment encodés par l'empreinte FCPC6 partagé par le *decoy* et le jeu de molécules actives, $\boldsymbol{f_{i(FCPC6\ fragments)}}$ est la fréquence du fragment $\boldsymbol{i}$ dans le jeu de molécules actives, $\boldsymbol{N_{i(fragments)}}$ est le nombre d'atomes dans le fragment $\boldsymbol{i}$ et $\boldsymbol{N_{FCPC6\ fragments}}$ est le nombre total de fragment FCPC6 dans le decoy.

$$LADS = \frac{\sum_{i=1}^{n}(N_{i(fragments)} - f_{i(FCPC6\ fragments)})}{N_{FCPC6\ fragments}}$$

5.1.4 Points forts et points faibles des banques d'évaluation de référence

Outre les progrès effectués concernant la minimisation des biais, une deuxième amélioration remarquable est la diversification des cibles protéiques intégrées dans ces bases de données. Certaines banques de données comme la DUD-E couvrent une large variété de cibles (102 cibles : 26 kinases, 15 protéases, 11 récepteurs nucléaires, 5 récepteurs couplés aux protéines G, 2 canaux ioniques, 2 cytochromes P450s, 36 autres enzymes, 5 protéines diverses), alors que d'autres sont spécifiques de familles protéiques comme la GPCR Ligand Library (GLL)/ Decoys Database (GDD)[4] pour les récepteurs couplés aux protéines G, la Maximal Unbiased Benchmarking data sets for HDACs (MUB-HDAC) [6] pour les histones désacétylases, la Nuclear Receptors Ligands and Structures Benchmarking DataBase (NRLiSt BDB)[5] et la Nuclear Receptors DataBase Including Negative Data (NR-DBIND)[81] pour les récepteurs nucléaires. Il est également possible de générer un jeu de *decoys* de manière automatique en fonction des molécules actives fournies grâce à l'outil DecoyFinder[150] qui suit le protocole de la DUD-E.

De par la diversité de ses cibles et la sélection rationnelle des *decoys* qu'elle contient, la DUD-E fait office de référence[397]. Elle présente néanmoins des faiblesses : une seule structure protéique est associée à chacun des jeux de données, et les agonistes et antagonistes des récepteurs nucléaires sont confondus dans le même jeu de molécules actives. De récentes études ont souligné le besoin de considérer séparément les agonistes et les antagonistes particulièrement lorsque les structures adoptent différentes conformations lorsqu'elles sont liées à l'un ou à l'autre[398,211], comme c'est le cas pour les récepteurs nucléaires. Par ailleurs, l'ensemble des banques présentées ci-dessus impose une dissimilarité structurale entre les molécules actives et les *decoys,* ce qui permet d'éviter l'intégration de faux négatifs. Cependant, ceci ne permet jamais de confronter les outils de criblage à un cas de figure fréquent où de larges différences d'affinités sont observées entre deux molécules structuralement très proches. Ceci pourrait expliquer, au moins en partie, les différences de performances observées entre les études rétrospectives et prospectives. Pour évaluer les outils de criblage de manière plus robuste, une solution est l'intégration de molécules inactives validées expérimentalement.

Tableau 10 Liste de banque de données d'évaluation

Nom	Année	Lien de téléchargement	Origine des ligands	Origine des decoys	#cibles / #classes	Sélection des decoys
Jeu de decoys de Rognan (C. Bissantz, Folkers, et Rognan 2000)	2000	http://bioinfo-pharma.u-strasbg.fr/labwebsite/download.html	Littérature	ACD	2 / 2	Sélection aléatoire
Jeu de decoys de Shoichet [399]	2003	-	MDDR	MDDR	9 / 4	Retrait des composés avec des groupements fonctionnels indésirables
Jeu de decoys de Li [386]	2003	-	Littérature	MDDR	6 / 1	Molécules de polarité et poids moléculaire similaires aux inhibiteurs de kinases connus
Jeu de decoys de Jain [400]	2006	http://www.jainlab.org/downloads.html	PDBbind	ZINC "drug-like" and Rognan's decoy set	34 / 7	1000 molécules de la ZINC aléatoirement sélectionnées qui satisfont les critères : MW≤500, logP≤5, HBA≤10, HBD≤5 et RB≤12 ; s'y ajoutent les decoys de Rognan respectant le critère RB≤15
Directory of Useful Decoys (DUD) [396]	2006	http://dud.docking.org	Littérature et PDBbind	ZINC "drug-like"	40 / 6	Decoys satisfaisant les règles de Lipinski ; sélection basée sur la similarité physico-chimique, et dissimilarité topologique avec les molécules actives

DUD Clusters [401]	2008	http://dud.docking.org/clusters/	DUD	-	40 / 6	Données de la DUD clusterisées en fonction du chémotype des molécules actives
WOMBAT Data For Enrichment Studies [389]	2007	http://dud.docking.org/wombat/	WOMBAT	-	13 /4	Données de la DUD Clusters et données de la WOMBAT clusterisées en fonction des chémotypes des molécules actives via une approche de graphes réduits
Maxim Unbiased Validation (MUV) [2]	2009	-	PubChem	PubChem	18 / 7	Deux fonctions permettent de mesurer les distances actif-actif et decoy-actif à partir de descripteurs 2D ; les molécules actives les plus dispersées parmi le jeu de molécules actives sont conservées, les decoys avec une distribution spatiale similaire sont retenus
DUD LIB	2009	http://dud.docking.org/jahn/	DUD-cluster	DUD	13 / 4	Sous-jeu de données de la DUD satisfaisant les critères : MW ($\leq$ 450) et AlogP ($\leq$ 4,5)
Charge Matched DUD	2010	http://dud.docking.org/charge-matched/	DUD	ZINC	40 / 6	Sélection similaire à la DUD avec un critère supplémentaire sur la charge nette
Virtual Decoys Sets (VDS) [402]	2011	http://compbio.cs.toronto.edu/VDS	DUD	ZINC	40 / 6	Même que la DUD sans considération de la faisabilité synthétique

DEKOIS [3]	2011	http://dekois.com/dekois_orig.html	BindingDB	ZINC	40 / 6	Classe les molécules actives et les decoys dans des cellules caractérisées par 8 propriétés physico-chimiques et sélectionne les decoys selon (1) leur proximité physico-chimique avec les molécules actives et (2) le score *latent active in the decoys set* (LADS) basé sur la similarité des fragments partagés avec le jeu de molécules actives : plus un fragment est grand et fréquent chez les actifs, plus il augmente le score LADS
GPCR Ligand (GLL) / Decoys Database (GDD) [4]	2012	http://cavasotto-lab.net/Databases/GDD/	GLIDA, PDB, et Vilar et al. (10.1002/jcc.21346)	ZINC	147 / 1	Correspondance de propriétés physico-chimiques et filtre de similarité topologique ; sélection finale sur le poids moléculaire.
Decoy Finder [150]	2012	http://urvnutrigenomica-ctns.github.io/DecoyFinder/	Utilisateur	Utilisateur	-	
DUD Enhanced (DUD-E) [1]	2012	http://dud.docking.org/r2/	CHEMBL	ZINC	102 / 8	Correspondance de propriétés physico-chimiques et filtre de similarité topologique ; sélection sur le poids moléculaire ; sélection finale aléatoire

DEKOIS 2.0 (Ibrahim, Bauer, et Boeckler 2015a)	2013	http://www.dekois.com	BindingDB	ZINC	81 / 11	Pareil que pour la DEKOIS avec 3 propriétés physico-chimiques supplémentaires (nombre de charges négatives/ positives, cycle aromatiques), un filtre de PAINS et une version améliorée du score LADS
NRLiSt BDB [5]	2014	http://nrlist.drugdesign.fr	CHEMBL	ZINC et générateur de decoys DUD-E	27 / 1	Utilise le générateur de decoys de la DUD-E
REPROVIS-DB	2011	-	Littérature	Littérature	-	Extraits des screening fructueux
MUBD-HDACs [6]	2015	-	CHEMBL et littérature	ZINC	14 / 1	Sélection hierarchique : 1) filtre grossier sur la similarité physicochimique et topologique, 2) filtre affiné sur la similarité physico-chimique (*simp*) et 3) sélection des decoys partageant le plus de similarité topologique avec une molécule active (*simsdiff*)

5.1.5 Intégration de données d'inactivité

Plusieurs banques de données mélangent *decoys* et molécules inactives validées expérimentalement (DUD-E, MUV). Dans la DUD-E, les molécules n'ayant aucune activité mesurée à 30µM ou à plus basse concentration sont inclues dans le jeu de *decoys*. Dans la MUV, les molécules sont issues de données de HTS extraites de PubChem BioAssay, et celles n'ayant montré aucune activité lors de l'essai primaire sont considérées comme *decoys* potentiels et sont ensuite filtrées avant sélection. Pour les familles protéiques les plus étudiées (kinases, GPCRs, récepteurs nucléaires etc.), les données publiées dans la littérature ainsi que dans des banques de données publiques sont abondantes, et incluent généralement des données d'inactivité ou des mesures d'interaction faibles et insuffisantes pour utiliser la molécule dans un cadre thérapeutique. Plusieurs banques de donnée listées dans le Tableau 11 regroupent des données expérimentales qui peuvent être exploitées pour construire des jeux de données adaptés à un système d'étude. Une molécule peut être considérée inactive si elle ne présente aucune affinité pour la cible à une concentration supérieure ou égale à un seuil d'acceptabilité fixé[81]. Un seuil arbitraire peut être défini selon l'étude conduite et peut varier : de sorte à modifier la balance molécules actives/ inactives, ou de sorte modifier le seuil d'inclusion des molécules actives et d'exclusion des molécules considérées inactives. Il est raisonnable d'imposer une marge entre le seuil d'inclusion des molécules actives et d'exclusion des molécules inactives afin d'éviter les erreurs de classification dues aux marges d'erreurs des tests biologiques[81].
Il existe peu d'études permettant de comparer l'intérêt de l'utilisation de molécules inactives en comparaison aux *decoys*. Ce fut l'un des sujets d'étude de cette thèse (Cf Résultats 1).

Tableau 11 Banques de données contenant des données d'activité et d'inactivité

Banque de données	Contenu	Accès	Liens
ChEMBL[404]	Données de bioactivité	Gratuit	https://www.ebi.ac.uk/chembl/
Drugbank[391]	Annotations des médicaments approuvés par la FDA	Gratuit sous condition d'inscription	https://www.drugbank.ca
Binding MOAD[405,406]	Données d'affinité pour des complexes ligand/récepteur disponibles dans la PDB	Gratuit	http://bindingmoad.org

AffinDB[407]		Gratuit sous condition d'inscription	http://pc1664.pharmazie.uni-marburg.de/affinity/
PDSP Ki Database[408]	Données de criblage du National Institute of Mental Health's Psychoactive Drug Screening Program	Gratuit	https://pdsp.unc.edu/databases/kidb.php
BRENDA[409]	Constantes de liaison d'enzymes	Gratuit	https://www.brenda-enzymes.org/index.php
GPCRDB[410]	Donnée d'affinité pour les récepteurs couplés aux protéines G	Gratuit	https://gpcrdb.org
D3R[411]	Données d'affinité fournies par des laboratoires pharmaceutiques et académiques incluant des co-cristaux et des molécules inactives	Gratuit	https://drugdesigndata.org/about/datasets
Tox21[412]	Données de HTS de plus 10000 composés testés sur des récepteurs nucléaires et sur des protéines impliquées dans les voies de réponse au stress	Gratuit	https://www.ncbi.nlm.nih.gov/pcassay?term=tox21
NR-DBIND[81]	Données d'affinité et d'activité pour les récepteurs nucléaires	Gratuit	http://nr-dbind.drugdesign.fr

5.2 Les métriques génériques

5.2.1 Corrélations

Un coefficient de corrélation permet d'estimer une association statistique entre deux variables[413]. Le coefficient de corrélation de Pearson permet de mesurer la corrélation linéaire entre deux variables X et Y. Les coefficients de corrélation de Spearman et de Kendall, quant à eux, mesurent de corrélation linéaire entre les rangs des individus des variables X et Y. En conséquence, le coefficient de Pearson peut être calculé si le score associé aux résultats de criblage est censé être linéairement lié à l'affinité biologique mesurée. La plupart du temps ce n'est pas le cas et on favorise les corrélations de Spearman et de Kendall. Le coefficient de Kendall, τ (Équation 7), compare le rang de chaque paire X_i et X_j aux paires Y_i et Y_j. Si l'ordre

entre les rangs X_i et X_j est conservé entre Y_i et Y_j, les paires sont considérées concordantes. Inversement, si l'ordre n'est pas conservé, les paires sont dites discordantes. Lorsque $X_i = X_j$ ou $Y_i = Y_j$ en termes de rang, la paire n'est ni concordante, ni discordante, ce qui amplifie l'impact des paires non équivalentes sur le coefficient final obtenu. Contrairement au coefficient de Kendall, le coefficient de Spearman prend en compte les écarts de rang entre la variable X et la variable Y (Équation 8). Le coefficient de Spearman est donc plus sensible aux erreurs et sa valeur est généralement plus faible que le coefficient de Kendall.

Équation 7 Formule du coefficient de Kendall, $\boldsymbol{\tau}$. $\boldsymbol{\tau}$ dépend du nombre de paires concordantes et de paires discordantes entre les variables X et Y. Les paires de rang équivalent dans X ou dans Y ne sont ni considérées concordantes ni discordantes.

$$\tau = \frac{(nombre\ de\ paires\ concordantes) - (nombre\ de\ paires\ discordantes)}{\frac{1}{2}(n(n-1))}$$

Équation 8 Formule du coefficient de Spearman, Rs. Rs dépend de l'ensemble des distances de rang au carré $\boldsymbol{d_i^2}$ entre un variable X et une variable Y. n représente le nombre d'individus dans chaque variable.

$$Rs = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

5.2.2 ROC/ AUC

Une courbe ROC représente le taux de vrais positifs en fonction du taux de vrais négatifs[414]. Il s'agit de représenter visuellement la sensibilité en fonction de l'anti-spécificité (1 – spécificité) (Figure 37 et Figure 38). Dans le cas idéal, 100% des vrais positifs sont retrouvés aux meilleurs

		Valeurs expérimentales	
		Actifs	Inactifs
Prédiction	Actifs	Vrais positifs	Faux positifs
	Inactifs	Faux inactifs	Vrais inactifs

$$\textbf{Sensibilité} = \frac{VP}{VP + FN} \qquad \textbf{Spécificité} = \frac{VN}{VN + FP}$$

Figure 37 Calculs de la sensibilité et de la spécificité. La sensibilité correspond au nombre de Vrais Positifs retrouvés (VP) par rapport au nombre total de molécules actives (VP+FN). La spécificité correspond au nombre de Vrais Négatifs retrouvés par rapport au nombre total d'inactifs (VN+FP).

rangs du classificateur et la courbe ROC forme une courbe pleine stabilisée à 1 en ordonnées (Figure 38). Lors d'études de *benchmark*, on mesure la performance d'un classificateur binaire en calculant l'air sous la courbe ROC (Air Under the Curve ou AUC), comprise entre 0 et 1, 1 équivalent à une prédiction parfaite et 0.5 équivalent à une prédiction aléatoire (Figure 38). Cette métrique présente l'avantage d'être indépendante du ratio actifs/inactifs observé dans le jeu de données[383] et permet d'illustrer aussi bien visuellement que numériquement la capacité d'un outil à mieux classer les molécules actives en comparaison à l'aléatoire.

Lors de campagnes de criblage virtuel effectuées à des fins de recherche thérapeutique, il est important de s'assurer que le protocole utilisé permet de retrouver des molécules actives dans les premiers pourcentages du classement prédit, ce que l'on appelle reconnaissance précoce, puisque ce sont ces molécules qui seront étudiées de manière plus approfondie. Dans ce cas, les courbes de ROCs et le calcul de leur AUC ne sont pas utilisables puisque la valeur d'AUC n'informe que sur la capacité de prédiction globale. Ainsi dans le cas où 1) la moitié des actifs est retrouvée en début de classement et l'autre moitié en fin de classement, ou 2) les actifs sont distribués de manière aléatoire ou 3) tous les actifs sont retrouvés au milieu, la valeur d'AUC

sera la même. Il est néanmoins possible de calculer l'AUC partielle pour quantifier la capacité

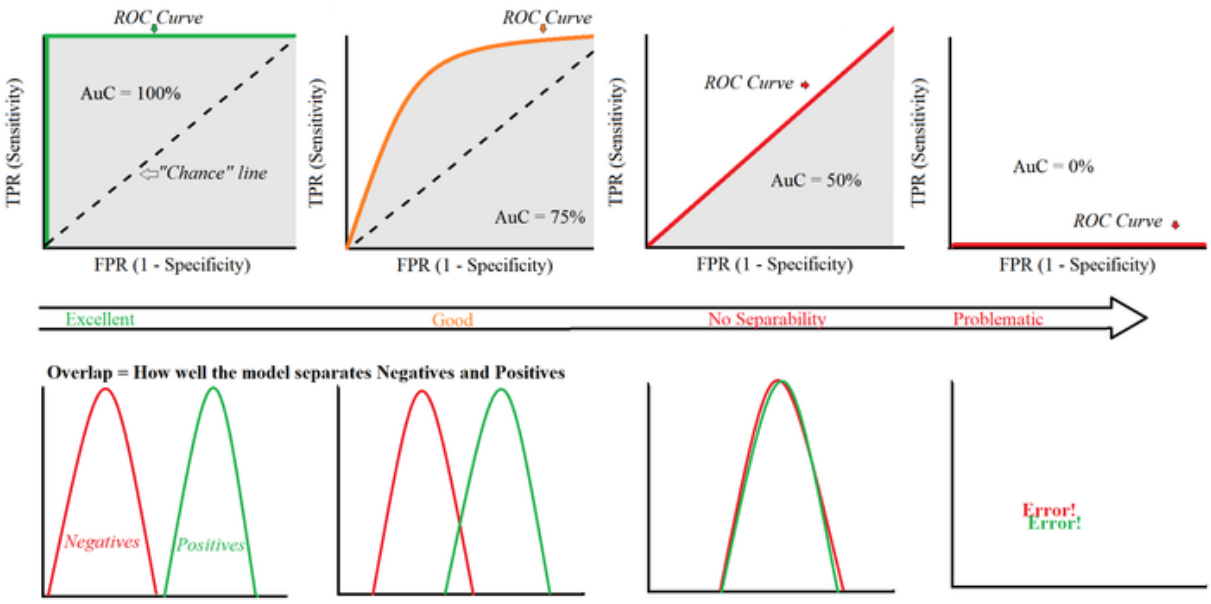


Figure 38 Interprétation des courbes ROCs. Dans un cas idéal, l'intégralité des composés actives sont identifiés dans la fraction précoce du classement : l'AUC correspondante est maximale (100%) et la séparation entre les molécules actives et inactives est nette. Dans un cas jugé bon, la séparation reste nette bien que chevauchante et il en résulte une courbe ROC au dessus de l'aléatoire et une valeur d'AUC comprise entre 50% et 100%. Dans un cas mauvais, les molécules actives et inactives ont des distributions de scores chevauchantes et ne sont pas distinguables : la courbe ROC est donc proche de l'aléatoire (diagonal), et l'AUC associée est proche de 50%. Lorsque l'AUC est inférieur à 50%, la méthode utilisée à un pouvoir de discrimination des molécules actives inférieur à l'aléatoire et n'a donc pas de pouvoir prédictif. D'après Stéphanie Glen

de discrimination d'une méthode sur une fraction de la classification étudiée.

Dans une récente étude, Ibrahim et al. proposent d'ajouter une information visuelle concernant la série chimique des ligands retrouvés en marge des courbes ROC[403]. Ceci permet d'informer sur 1) la diversité des molécules retrouvées dans les premiers pourcentages, 2) d'identifier les familles de molécules retrouvées plus loin dans la courbe ROC, 3) vérifier que l'outil distingue réellement les inactifs et ne se contente pas de séparer les molécules par chémotype. En effet, une courbe avec une bonne AUC perd de son intérêt dès lors que les actifs et inactifs partageant des similarités structurales sont confondus dans le classement.

5.2.3 Facteur d'enrichissement (EF)

Le facteur d'enrichissement (EF) est une métrique qui mesure la proportion de molécules dans une fraction précoce par rapport à leur proportion dans la chimiothèque (Équation 9). Elle permet d'estimer la capacité d'un outil à enrichir en molécules actives par rapport à une distribution aléatoire. L'EF dépend du nombre de molécules actives et inactives dans la chimiothèque ; de ce fait, deux EF sont comparables uniquement s'ils sont calculés sur la même chimiothèque.

Équation 9 Formule du calcul du facteur d'enrichissement. $\boldsymbol{a_{n\%}}$ et $\boldsymbol{t_{n\%}}$représentent respectivement la proportion de molécules actives et le nombre de molécule total dans les la fraction précoce correspondant à n% de la chimiothèque ordonnée, A correspond au nombre de molécules active total et T au nombre de molécules total.

$$EF = \frac{a_{n\%}/t_{n\%}}{A/T}$$

L'une des faiblesses de l'EF réside dans son incapacité à tenir compte du rang des molécules actives retrouvées dans la fraction précoce. Ainsi, une méthode A qui classe n molécules dans la première moitié de la fraction précoce aura le même EF qu'une méthode B classant le même nombre n de molécules actives dans la deuxième moitié de la fraction précoce. Une deuxième

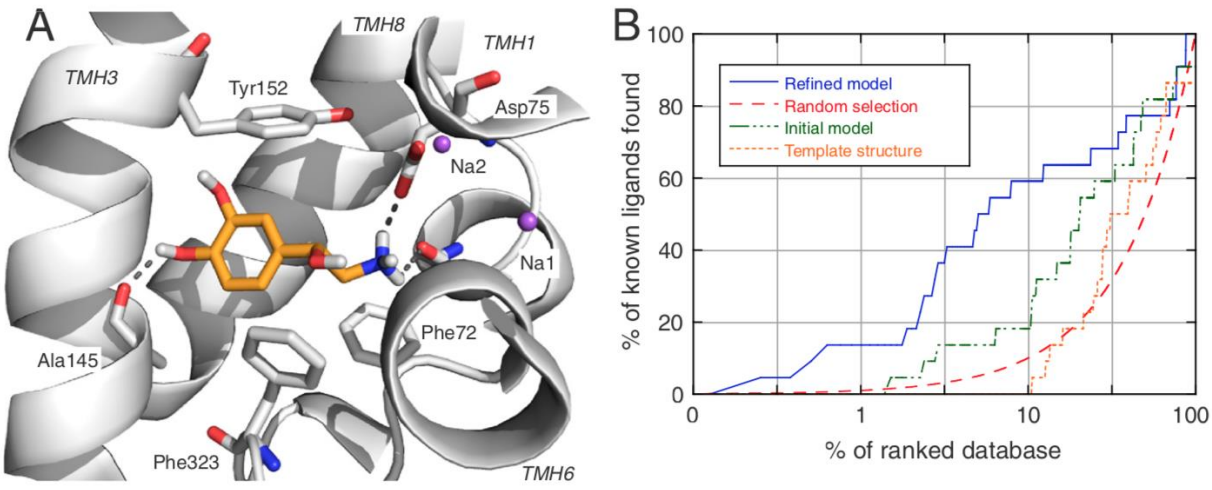


Figure 39 Exemple de courbe d'enrichissement. Les courbes sont tracées à partir des résultats de docking réalisés sur 3 modèles différents de noradrenaline transporter en suivant un protocole unique et à partir d'une même chimiothèque. L'axe des abscisses correspond au pourcentage seuil auquel est calculée la proportion de molécules actives retrouvée sur une échelle logarithmique. La courbe rouge représente un résultat aléatoire. La courbe bleue correspond au meilleur résultat obtenu. D'après [414]

faiblesse réside dans l'omission de toutes les informations concernant la fraction de molécules classée après le seuil imposé. Généralement ce seuil est déterminé en fonction de la composition de la chimiothèque. Plus la chimiothèque est fournie et le ratio nombre de molécules actives/ nombre de molécules totales est faibles, plus les calculs d'EF se calculent à des % faibles (1%, 2%, 5%). Afin de prendre en compte un maximum d'information apporté par l'EF, la fraction de molécules active à x% peut être calculé à plusieurs valeurs seuils x, ce qui permet de déduire une courbe d'enrichissement (Figure 39[415]).

5.2.4 Robust Initial Enhancement (RIE) et Boltzmann-Enhanced Discrimination of ROC (BEDROC)

Le Robust Initial Enhancement (RIE) est une métrique permettant de quantifier la reconnaissance précoce[416,417]. Elle utilise une exponentielle décroissante pour pondérer le classement de chacune des molécules criblées. Ainsi, les molécules identifiées en têtes de listes ont un poids proche de 1, et ce poids décroit de façon exponentielle avec le classement. Le RIE (Équation 10) est calculé comme le rapport entre la moyenne des poids associés aux n molécules actives du jeu de N données et la moyenne des poids associés à n molécules actives lorsqu'elles suivent une distribution uniforme dans une liste ordonnée de N composés. Le RIE renseigne donc sur le gain obtenu grâce au classificateur en comparaison à une classification aléatoire. Le RIE maximal atteignable dépend de la fraction de molécules actives dans la chimiothèque et du paramètre α qui contrôle le poids attribué à la fraction précoce des composés et peut être calculé via l'Équation 12. Plus la somme se rapproche du maximum atteignable, meilleur est le classificateur. Initialement, le paramètre α était calculé à partir de 1000 tests de classification aléatoire. Face à la nécessité d'augmenter ce nombre de test en fonction du nombre de données actives et inactives du jeu de données pour obtenir une valeur précise, Truchon et al. ont proposé de remplacer ce terme issus de tests aléatoires par une formule analytique (Équation 12)[383]. Ils ont également introduit la métrique BEDROC, qui correspond à une normalisation du RIE entre 0 et 1 selon l'Équation 13.

Équation 10 Formule du calcul du RIE. n est le nombre de molécules actives parmi une chimiothèque de N molécule, xi est le rang associé au ième composé actif, α est le paramètre qui contrôle le poids attribué à la fraction précoce des composés.

$$RIE = \frac{\sum_{i=1}^{n} e^{-\alpha x i}}{\frac{1}{N}\left(\frac{1 - e^{-\alpha}}{e^{\frac{\alpha}{N}} - 1}\right)}$$

Équation 11 Formule du calcul du $\boldsymbol{RIE_{min}}$. $\boldsymbol{\alpha}$ est le paramètre qui contrôle le poids attribué à la fraction précoce des composés, Ra est le taux d'actif dans la chimiothèque de composés.

$$RIE_{min} = \frac{1 - e^{\alpha R_a}}{R_a(1 - e^{\alpha})}$$

Équation 12 Formule du calcul du $\boldsymbol{RIE_{max}}$. $\boldsymbol{\alpha}$ est le paramètre qui contrôle le poids attribué à la fraction précoce des composés, Ra est le taux d'actif dans la chimiothèque de composés.

$$RIE_{max} = \frac{1 - e^{-\alpha R_a}}{R_a(1 - e^{\alpha})}$$

Équation 13 Formule du calcul de la valeur de BEDROC.

$$BEDROC = \frac{RIE - RIE_{min}}{RIE_{max} - RIE_{min}}$$

5.3 Évaluation des conformations générées et poses prédites

5.3.1 Écart quadratique moyen (RMSD)

La *Root Mean Square Deviation* (RMSD) est une mesure de la différence de position et/ou de conformation inter-moléculaire largement utilisée pour évaluer l'efficacité d'un outil de docking et de génération de conformère (Équation 14)[418]. Dans le cas du docking, on teste la capacité de l'outil à retrouver une pose proche de la pose bioactive du ligand dans le site de liaison du récepteur. La pose expérimentale et celle prédite par docking doivent impérativement être dans le même référentiel et ne doivent subir aucune transformation (translation et rotation) pour être comparée. Au contraire, lorsqu'on évalue la capacité d'un générateur de conformère à proposer une pose proche de la pose bioactive, une superposition optimale des molécules est nécessaire avant d'effectuer le calcul de RMSD. Généralement, le calcul de la RMSD ne prend en compte que les atomes lourds et deux conformations sont considérée proches lorsqu'elles ont une RMSD < 2 Å[109].

Équation 14 Formule du calcul de la RMSD. N est le nombre d'atome lourd dans la molécule, $\boldsymbol{a_i}$ et $\boldsymbol{b_i}$ sont les coordonnées cartésiennes de l'atome i dans les conformations A et B respectivement.

$$RMSD(A,B) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\|a_i - b_i\|^2}$$

Le calcul de la RMSD nécessite une étape d'identification des atomes correspondants entre deux conformations d'une molécule. Pour des raisons pratiques, ces correspondances sont généralement déduites de l'indice associé à chaque atome ou de l'ordre des atomes dans les fichiers d'entrée. Cette approche est très dangereuse lorsque les molécules présentent un axe de symétrie ; elle peut induire une surestimation de la distance. Allen et al. ont présenté un exemple frappant avec le 1,2-dichlorobenzene docké dans le lysozyme T4 (Figure 40) [418]. Bien que deux poses équivalentes aient été générées, l'une est associée à une RMSD < 2 Å et l'autre une RMSD > 2 Å du fait de son inversion par rapport à l'axe de symétrie (Figure 40). Ce cas est un

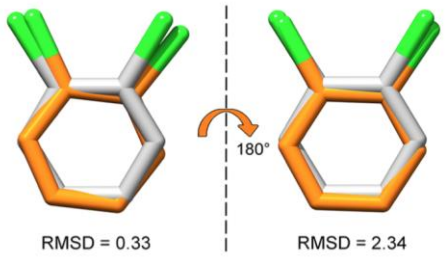


Figure 40 Illustration du problème de prise en compte de la symétrie dans le calcul de la RMSD. La pose cristallographique est représentée en grise et deux poses dockées inversées l'un par rapport à l'autre selon leur axe de symétrie sont représentées en orange. Les valeurs de RMSD sont indiquées en Å. D'après [417]

exemple typique de docking pouvant être à tort considéré comme un échec. Pour pallier ce problème, différentes variations de la RMSD ont été développées parmi lesquelles la RMSD de distance minimale et la RMSD de correspondance. Il faut noter que le calcul de RMSD ne fournit aucune information sur les interactions établies avec la cible et ne permet donc pas de quantifier la conservation de ces contacts. Par ailleurs, la valeur de RMSD est directement liée au nombre d'atome de la molécule étudiée ; en ce sens, le seuil de similarité fixé à 2 Å est discutable.

5.3.2 RMSD de distance minimale

Pour tenter de résoudre les problèmes de symétrie liés au calcul de RMSD par paire d'atomes, Trott et Olson[215] ont proposé une version modifiée prenant en compte la distance minimale entre deux atomes de même éléments de la conformation A et B de la molécule :

Équation 15 Formule générale de la RMSD de distance minimale. Elle correspond au maximum des RMSD' minimales observées entre A et B et entre B et A.

$$RMSD_{min}(A,B) = \max\{RMSD'_{min}(A,B), RMSD'_{min}(B,A)\}$$

Équation 16 Détail de la formule de la RMSD de distance minimale. Les atomes $\boldsymbol{a_i}$ de la pose A sont comparés itérativement à chaque atomes $\boldsymbol{b_j}$ de la pose B qui partagent le même type d'élément. La distance minimale calculée est conservée pour le calcul de la RMSD de distance minimal finale.

$$RMSD'_{min}(A,B) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(min_j\|a_i - b_j\|)^2}$$

Dans cette approche, les atomes $\boldsymbol{a_i}$ de la pose A sont comparés itérativement à chaque atomes b_i de la pose B qui partagent le même type d'élément. La distance minimale calculée est conservée pour le calcul de la RMSD de distance minimale finale. Le maximum entre les RMSD de distances minimales mesurées entre A et B et entre B et A est conservé comme valeur finale. Cette tentative de correction des erreurs de symétrie présente des faiblesses[418] : certains atomes peuvent être utilisés plus d'une fois dans le calcul de distance alors que d'autres peuvent être négligés, ce qui signifie que l'on perd l'information structurale de la molécule de départ et que la RMSD a tendance à surestimer la distance réelle entre les deux poses.

5.3.3 RMSD de correspondance optimale

Une alternative est le calcul de RMSD de correspondance optimale (Équation 17) qui fait appel à l'algorithme Hongrois, permettant de résoudre les problèmes d'assignement minimum[419].

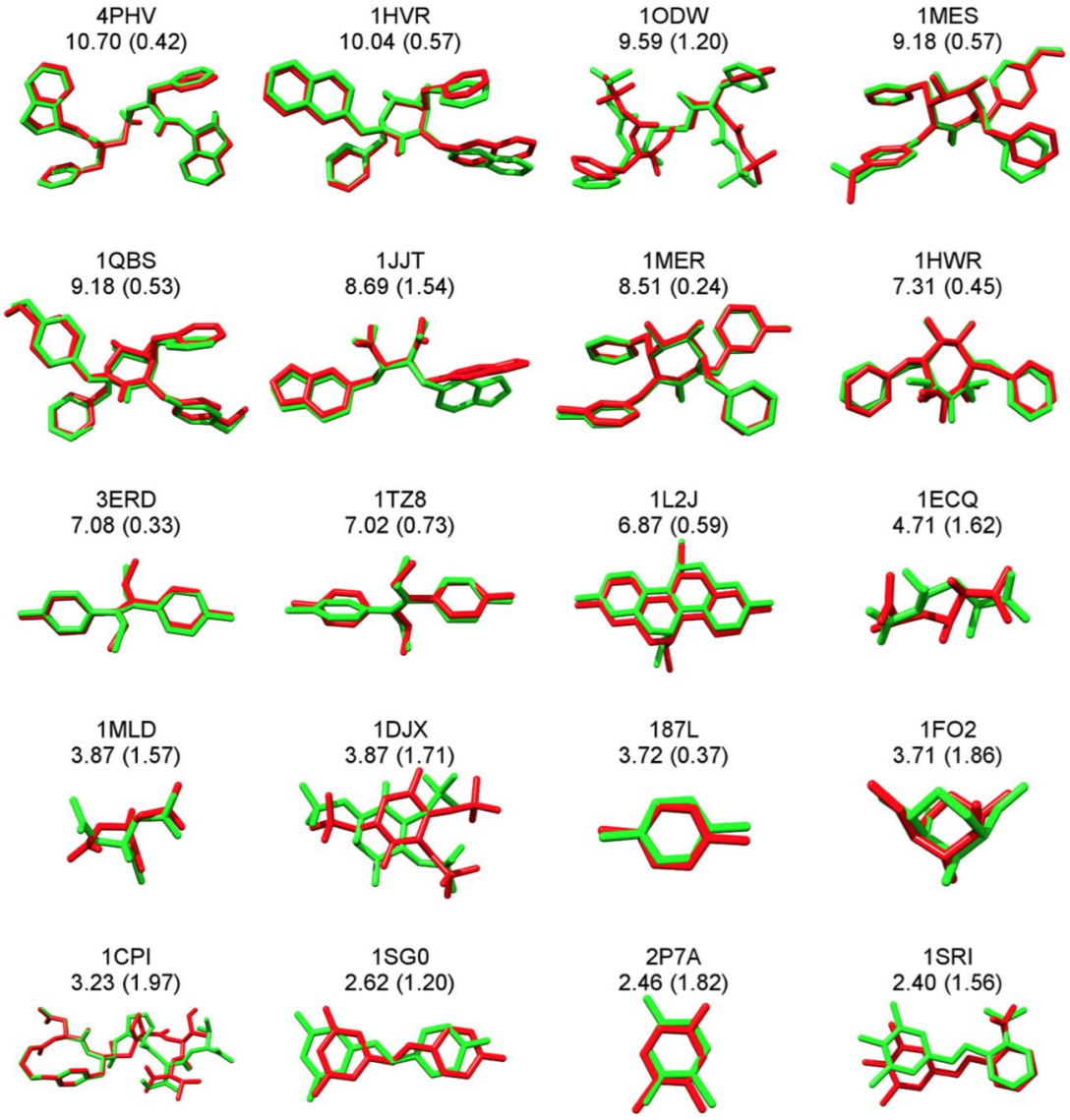


Figure 41 Illustration de la différence entre le calcul de RMSD classique et le calcul de RMSD de correspondance optimale sur un ensemble de molécules présentant un axe de symétrie. Les deux valeurs de RMSD sont présentées en Å, la RMSD de correspondance optimale est indiquée entre parenthèses. Pour chacun de ces exemples, la distance entre la pose de référence (rouge) et la pose prédite (vert) est optimisée grâce à la méthode de RMSD de correspondance optimale qui prend en compte la symétrie de composés. D'après [418]

Équation 17 Formule de la RMSD de distance optimale. Chaque atome $\boldsymbol{a_i}$ de la pose A est associé à un unique atome $\boldsymbol{b_j}$ de la pose B préalablement défini en résolvant le problème de correspondance maximale.

$$RMSD_{cor}(A,B) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(cor_j\|a_i - b_j\|)^2}$$

Dans cette approche, chaque atome de A et associé à un atome de B de sorte à minimiser la somme des distances entre toutes les paires d'atome. Allen et al. ont démontré l'efficacité de cette méthode[419] ; sur un total de 1043 ligands connus dockés sur leur cible avec DOCK6, 706 (67,7%) sont correctement dockés d'après un calcul de RMSD classique, et 52 (5%) supplémentaires sont correctement docké d'après un calcul de RMSD classique. La Figure 41 illustre des exemples de composés reclassés comme vrais positifs grâce à la correction de symétrie.

Outre la correction de symétrie apportée par cette méthode, Allen et al. proposent à l'utilisateur de définir eux même le seuil de RMSD de correspondance maximal et d'autoriser un nombre d'atomes non appariés. Ceci permet de comparer des molécules similaires mais non identiques comme illustré sur la Figure 42.

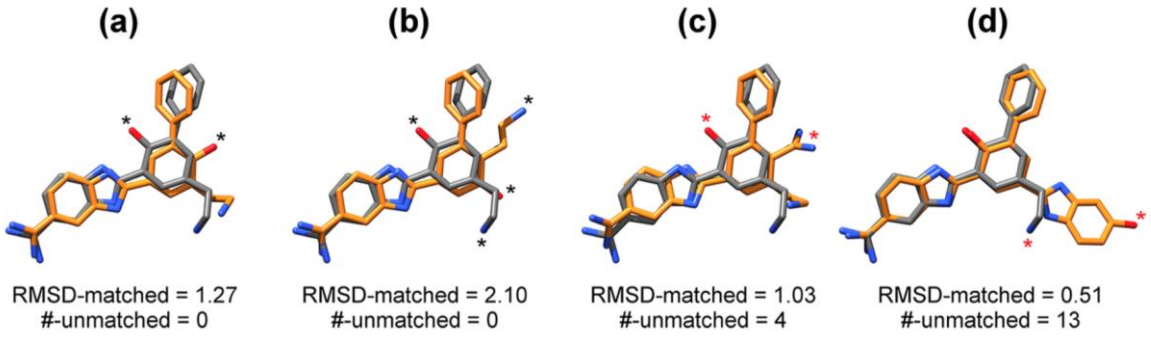


Figure 42 Illustration du calcule de RMSD de correspondance maximale réalisée entre des molécules similaires non identiques. Des différences de positionnement des groupements fonctionnels plus ou moins grandes (a et b) et des différences de composition des groupements fonctionnels (c et d) sont toléré selon les seuils de RMSD et d'atomes non appariés imposés. Les astérisques noirs indiquent un changement de position, les rouges un changement de composition. Les RMSD sont calculées sur les atomes en commun en Å. D'après [418]

5.3.4 *Real Space R-factor* (RSR)

En 2008, Yusuf et al. ont proposé de comparer directement les poses issus du docking à la densité électronique[420]. Ceci permet de s'affranchir des ambiguïtés de positionnement et d'orientation d'un ligand co-cristallisé dans le site de liaison et d'accepter les alternatives de placement ou de conformation satisfaisantes. La correspondance entre une pose de docking et la densité électronique est quantifiée par le *real space R-factor* (RSR) habituellement utilisé par les cristallographes pour évaluer la compatibilité entre le modèle généré et la densité électronique. La RSR compare la densité expérimentale ρ_{obs} avec la densité calculée ρ_{calc} du modèle généré[420].

Équation 18 Formule du RSR. $\boldsymbol{\rho_{obs}}$ et $\boldsymbol{\rho_{calc}}$ correspondent aux densité expérimentales et calculées.

$$RSR = \frac{\sum|\rho_{obs} - \rho_{cal}|}{\sum|\rho_{obs} + \rho_{cal}|}$$

La portion de densité électronique à prendre en compte est définie par proximité avec le ligand. Le RSR étant dépendant de la résolution de la cristallographie, Yusuf et al. proposent un calcul de RSR corrigé correspondant au ratio du RSR de la pose de docking et de celle du modèle de cristallographie (Équation 19 et Figure 44)[420].

Équation 19 Formule du RSR corrigé, calculé comme le rapport entre le RSR de la pose issue du docking et du modèle cristallographique.

$$RSR_n = \frac{RSR_d}{RSR_c}$$

Une RSR_n inférieure à 1 indique que la pose issue du docking correspond mieux à la densité électronique expérimentale que le modèle de cristallographie ; plus une RSR_n est supérieure à 1, plus la correspondance avec la densité électronique est mauvaise, une valeur acceptable de RSR_n étant fixée à 1.7. Cette approche s'avère particulièrement adaptée lorsque la carte de densité électronique est de faible résolution ou ne permet pas de positionner la totalité des atomes d'une molécule avec certitude (Figure 43). Lorsque des poses sont très similaires et la carte de densité associée est de haute résolution, la RSR_n tend à fortement pénaliser les faibles différences observées.

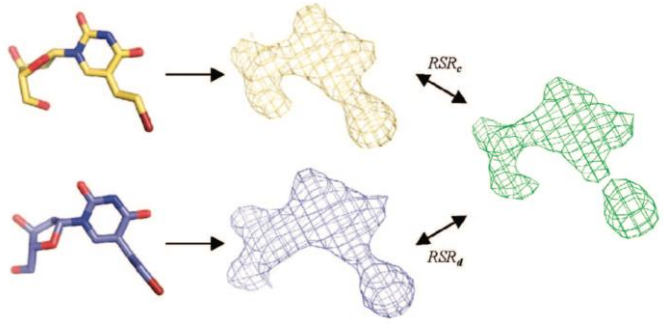


Figure 44 Les densités électroniques théoriques sont calculées à partir des coordonnées cartésiennes du modèle cristallographique (jaune) et de la pose issue du docking (bleu). Ces cartes de densité électroniques sont corrélées avec la carte de densité expérimentale, donnant $\boldsymbol{RSR_c}$ et $\boldsymbol{RSR_d}$. D'après [451].

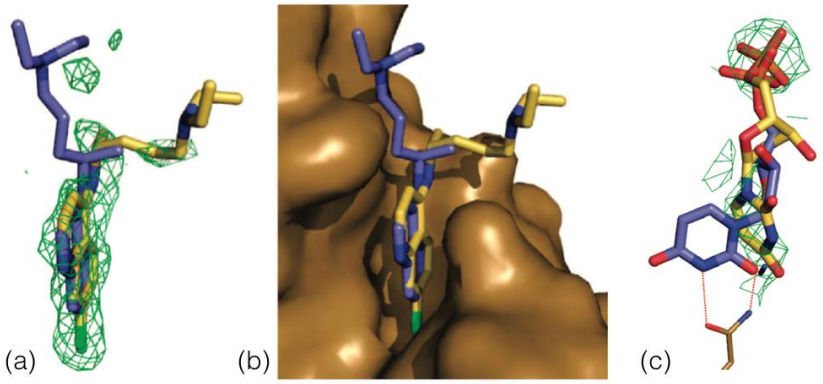


Figure 43 Exemple de cas ou la pose de docking serait considérée mauvaise selon un critère de RMSD. La structure cristallographique du ligand (jaune ; PDB = 1CET) et la pose la mieux classée selon GOLD t la fonction de score ChemScore (bleu) sont représentées (a et b) par rapport à la carte de densité électronique expérimentale. La valeur élevée de RMSD associée au ligand docké (3.7 Å) est due à un mouvement d'un groupement pour lequel la densité électronique n'est pas claire. La valeur de RSR_n (1.46) est principalement calculée sur la partie du ligand pour laquelle une densité électronique est observée. L'image(c) constitue un deuxième exemple ou la pose dockées possède une RMSD élevée par rapport au modèle de référence (3.4 Å) et un RSR_n acceptable (1.46) et propose deux liaison hydrogènes supplémentaires par rapport au modèle. D'après [451]

6 Les récepteurs nucléaires

6.1 Généralité sur les récepteurs nucléaires

Les récepteurs nucléaires (NRs) jouent un rôle majeur dans de nombreux processus physiologiques tels que le métabolisme, la prolifération cellulaire, la réponse immunitaire et inflammatoire, le développement et la reproduction[421,422,423,424]. Leur dérégulation est par conséquent critique puisqu'elle peut provoquer un dysfonctionnement de ces différentes fonctions et fait l'objet de nombreuses études. Leur implication dans de nombreuses maladies a conduit au développement de molécules thérapeutiques capables de moduler ces récepteurs nucléaires. Ainsi, les médicaments ciblant les récepteurs nucléaires représentaient en 2003 34 des 200 médicaments les plus prescrits[425] et en 2008, l'impact économique de ces médicaments était estimé à 10-15% d'un marché mondial de 400 milliards de dollars[425]. Outres les dérégulations internes à l'organisme, les NRs peuvent aussi être dérégulées par des molécules exogènes présentes dans l'environnement et auxquelles un individu peut être exposé contre son gré : ce sont des perturbateurs endocriniens. La famille des NRs présente donc un intérêt thérapeutique et de santé publique. De nombreuses données d'interactions impliquant des NRs sont disponibles dans des banques de données publiques, ce qui en fait une famille protéique appropriée pour des études d'évaluation d'outils de criblage et de construction de modèles fins.

6.2 Mode d'action des récepteurs nucléaires

La famille des NRs regroupe un ensemble de protéines capables d'interagir avec l'ADN pour promouvoir la transcription d'un gène : ce sont des facteurs de transcription. Parmi les 48 récepteurs humains comptabilisés, la plupart sont régulés via l'interaction avec des petites molécules hydrophobes. Il est à noter que pour certains NRs, aucun ligand naturel n'a été identifié, ce qui leur vaut la dénomination de NRs orphelins. Les NRs se divisent en 4 catégories[426] (Figure 45) :

- Type I : les récepteurs des stéroïdes (le récepteur aux androgènes (AR), les récepteurs aux estrogènes (ER α/β/γ), le récepteur aux glucocorticoïdes (GR), le récepteur aux minéralocorticoïdes (MR), et le récepteur à la progestérone (PR))
- Type II : les récepteurs formant des hétérodimères avec le récepteur X des rétinoïdes (RXR α/β/γ) (*constitutive androgen receptor* (CAR), *farnersoid X receptor* (FXR), récepteurs des

oxystérols (LXR α/β), récepteurs activés par les proliférateurs de peroxysomes (PPAR α/β/γ), *pregnane X receptor* (PXR), récepteurs à l'acide rétinoïque (RAR α/β/γ), récepteurs des hormones thyroïdiennes (TR α/β) et récepteur de la vitamine D (VDR))

- Type III : les récepteurs fonctionnant en homodimères (chicken ovalbumin upstream promoter transcription) factor (COUP-TF I/II), *germ cell nuclear factor* (GCNF), facteur nucléaire hépatocytaire 4 (HNF-4), récepteur X des rétinoïdes (RXR α/β/γ), récepteurs testiculaires (TR2/4))
- Type IV : les récepteurs orphelins fonctionnant en monomères (*estrogen related receptors* (ERR α/β/γ), nerve growth factor IB (NGFIB), *photoreceptor-specific nuclear receptor* (PNR), Rev-Erb, *retinoid related orphan receptors* (ROR α/β/γ), et *steroidogenic factor 1* (SF-1))

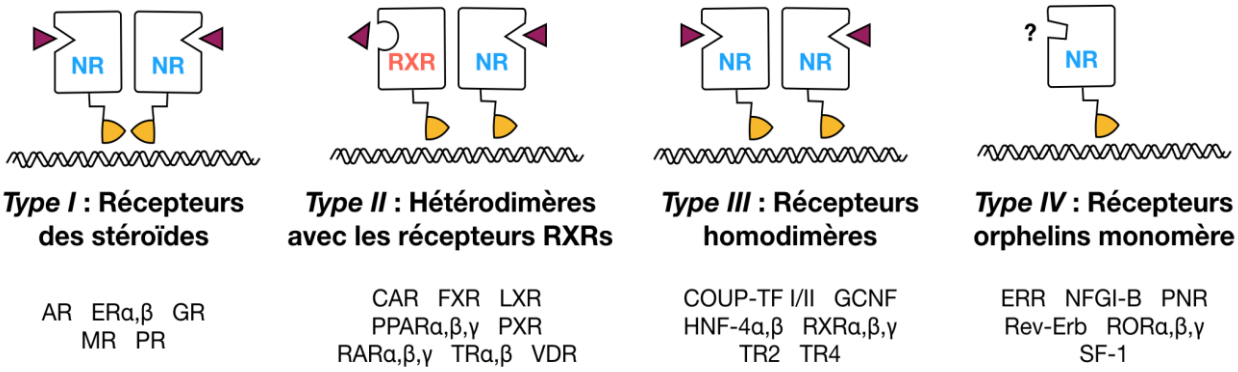


Figure 45 Classification des récepteurs nucléaires.

En absence de ligand agoniste, les récepteurs de type I (récepteurs des stéroïdes) et les récepteurs de type III sont complexés à des protéines chaperonnes de type *heat-shock protein* (HSP) qui les séquestrent dans le cytoplasmes[427]. Lors de la fixation d'un ligand agoniste, le récepteur subit un changement conformationnel qui permet de le dissocier des HSP, et de migrer dans le noyau cellulaire ou il va agir sous forme d'homodimère. C'est l'homodimère qui va pouvoir se fixer à l'ADN au niveau d'un élément de réponse aux hormones (HRE). Chaque type de récepteur (I ou III) reconnait des HRE spécifiques. Les récepteurs de type II fonctionnent généralement en hétérodimères et résident dans le noyau où ils sont constamment fixés à leur élément de réponse et maintenus inactifs grâce à des corépresseurs[428]. La fixation d'un ligand agoniste induit un changement de conformation qui entraîne la libération des corépresseurs (HDAC, SMRT/N-CoR) et le recrutement de co-activateurs qui vont permettre à la transcription des gènes cibles d'avoir lieu. Les HRE qu'ils reconnaissent sont à l'inverse des récepteurs de type I, organisés en séquence directe répétée[428]. Le domaine de liaison à l'ADN, qui s'avère être le plus conservé de la séquence polypeptidique des NRs possède une structure très conservée de 66 acides aminés. Parmi ces acides aminés, 8 cystéines interagissent avec

deux ions de zinc et assurent ainsi le maintien de la structure. Malgré la conservation de structure, différents groupes de NRs reconnaissent différents motifs HRE (Tableau 12). L'espacement entre les demi-sites constituent les HRE est crucial pour conférer sa sélectivité à chaque NR[426].

Tableau 12 HRE reconnus par les récepteurs nucléaires. Chaque récepteur nucléaire reconnaît un ou plusieurs HRE caractérisé(s) par la séquence consensus de ses demi-sites, le mode de dimérisation et la configuration des demi-sites (IR = séquences répétées inversées, DR = séquences directes répétées, P = séquences palindromes, IP = séquences de palindromes inversés, le numéro associé correspond au nombre d'acides nucléiques séparant chaque demi-site).

	NR	**HRE consensus**	**Dimérisation**	**Configuration**
Récepteurs stéroïdiens	AR, PR	5′-AGAACA-3′	Homodimère	IR3, DR3
	GR, MR	5′-AGAACA-3′	Homodimère	IR3
	ER	5′-AGGTCA-3′	Homodimère	IR3
Récepteurs non-stéroïdiens	RAR	5′-AGGTCA-3′	Homodimère	IR0
			Hétérodimère	DR1
				DR2
				DR5
	VDR	5′-AGGTCA-3′	Homodimère	DR3
			Hétérodimère	DR3
	PPAR	5′-AGGTCA-3′	Hétérodimère	DR1
	TR	5′-AGGTCA-3′	Monomère	Demi-site
			Homodimère	DR4, IP6, P0
			Hétérodimère	DR4
	RXR	5′-AGGTCA-3′	Homodimère	DR1
			Hétérodimère	DR1
				DR2
				DR3
				DR4
				DR5
Récepteurs orphelins	Nurr77	5′-AAA AGGTCA-3′	Monomère	Demi site allongé
	SF1, ERR2	5′-TCA AGGTCA-3′	Monomère	Demi site allongé

6.3 Structure générale

Les NRs partagent des séquences très conservées définies par 6 sous-régions (A à F sur la Figure 46)[429]. La région A/B, ou *activation function* 1 (AF-1) est la moins conservée parmi les NRs, et correspond à un domaine de transactivation hormone-indépendant chez de nombreux NRs.

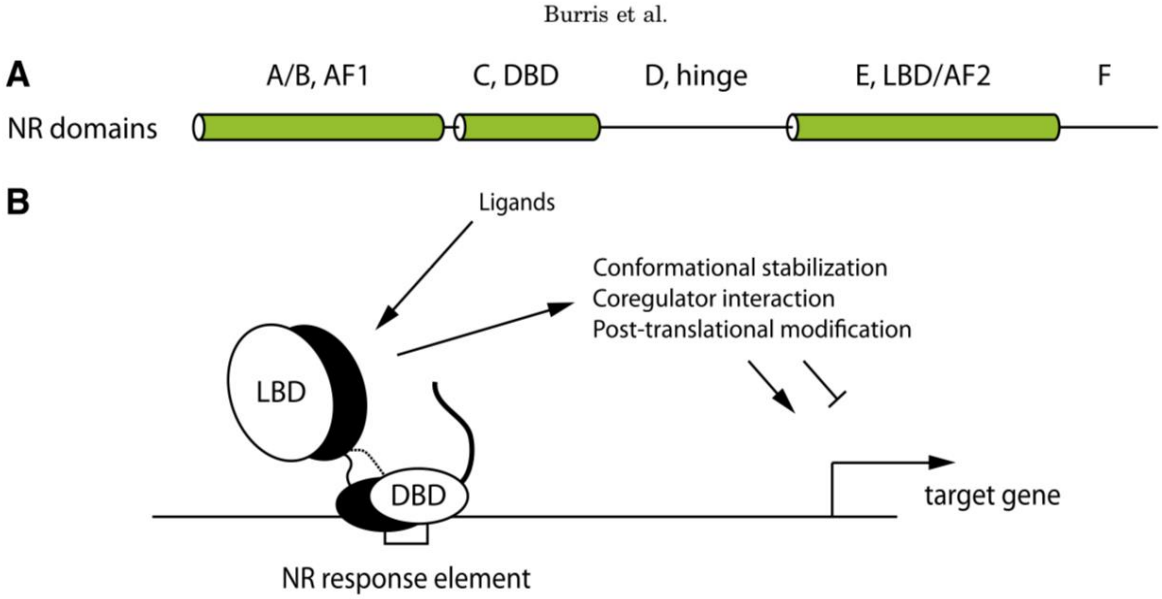


Figure 46 Schéma général des sous-domaines composant un récepteur nucléaire. D'après [428]

La région C est la plus conservée et elle constitue le domaine de fixation de l'ADN (*DNA binding domain*, DBD). Le DBD possède 8 cystéines dont la position est parfaitement conservée parmi les NRs et qui jouent un rôle primordial dans la conservation de la structure tridimensionnelle du domaine grâce à la formation de deux doigts de zinc. La région D est un domaine charnière peu conservé qui intervient dans la fixation à l'ADN pour certains récepteurs nucléaires. La région E correspond à un domaine de transactivation ligand-dépendant (*ligand binding domain*, LBD). C'est ce domaine qui interagit avec les petites molécules hydrophobes capables de moduler l'activité des NRs en induisant l'association ou la dissociation de co-activateurs ou de corépresseurs. Enfin la région F située en C-terminal n'est pas présente chez tous les NRs et son rôle biologique reste mal compris[429].

6.4 Structure globale du site de liaison des NRs

Les sites de liaison des NRs, ou LBDs, possèdent des conformations similaires qui consistent en une superposition de 3 couches d'hélices α. Les LBDs des NRs se caractérisent par une cavité hydrophobe unique en taille et en composition en acides aminés ce qui leur assure une sélectivité pour leur ligand endogène, excepté pour dans le cas des NRs isoformes. La structure

tridimensionnelle de la plupart des NRs a été résolue par cristallographie aux rayons X et a permis de déceler un changement structural entre les conformations *apo* et *holo* des NRs. En

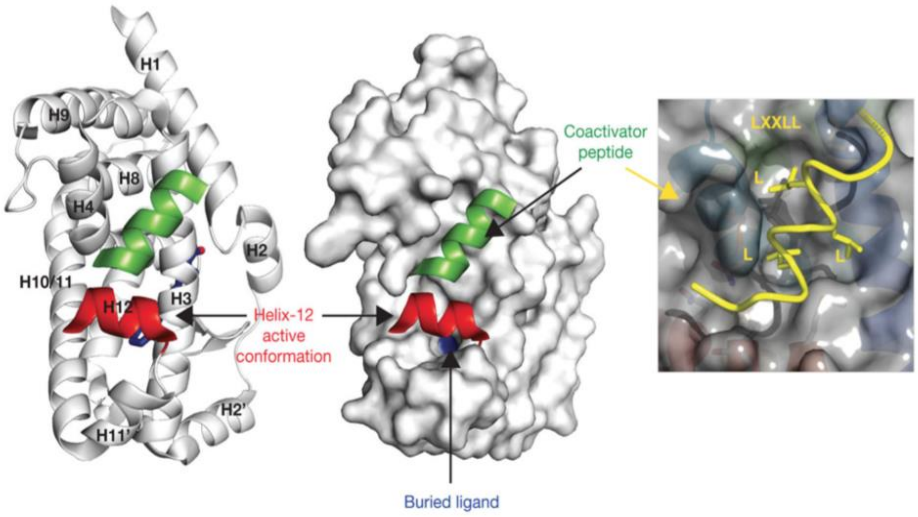


Figure 47 Structure cristallographique du récepteur ERα : l'hélice H12 est en conformation active, elle rend accessible une région capable de fixer des co-activateurs possédant un domaine riche en lysine de motif LXXLL. D'après [428]

absence de ligand, l'hélice H12, conservées dans tous les NRs, est éloignée du LBD. Lorsque qu'un ligand agoniste interagit avec le LBD, l'hélice H12 se referme sur ce dernier, piégeant le ligand dans une cavité fermée. Des études ont révélées que cette position « fermée » peut exister en l'absence de ligand et correspondre à l'état actif du NR[430]. Le ligand agoniste sert alors de stabilisateur et permet de maintenir la protéine dans sa conformation active. Lorsque l'hélice H12 est en conformation active, elle rend accessible une région capable de fixer des co-activateurs possédant un domaine riche en lysine de motif LXXLL (Figure 47). Lorsque l'hélice H12 est éloignée, cette région subit un réarrangement structural et devient apte à fixer un motif LXXXLXXX (I/L) appelé CoRNR (CoRepresseur des Récepteurs Nucléaires) propre aux corépresseurs des NRs. Des ligands des NRs, comme le rosiglitazone, interagissent directement avec des résidus de l'hélice H12, permettant de stabiliser la conformation active du NR. D'autres ligands ne forment que des interactions indirectes via le réarrangement de résidus du site de liaison.

6.5 Mécanismes de modulation des NRs

Les ligands naturels des NRs exercent une activité agoniste sur leur cible naturelle. Cependant, les multiples ligands de synthèse testés sur les NRs ont montré un plus large spectre d'activité, incluant des molécules agonistes, antagonistes, agonistes partielles, antagonistes partielles,

agoniste inverse ou encore des modulateurs spécifiques d'un ou plusieurs gène[81]. Il est désormais accepté que les molécules antagonistes agissent en perturbant le repliement de l'hélice H12. En revanche, le mécanisme moléculaire contrôlant cette activité reste opaque. Certaines études révèlent que la différence entre les agonistes et les antagonistes réside dans la présence d'extensions volumineuses chez les antagonistes qui empêchent l'hélice H12 de se positionner de manière optimale et d'exposer le site de fixation des co-activateurs[426]. Cependant, toutes les molécules antagonistes ne possèdent pas cette extension volumineuse. D'autres études proposent donc que certaines molécules jouent un rôle antagoniste en occupant le LBD sans provoquer de rapprochement de l'hélice H12 à cause d'un manque d'affinité avec cette dernière[431]. Les activités agonistes/antagonistes partielles correspondent aux molécules ayant des activités modérées en comparaison à un agoniste ou un antagoniste de référence. La frontière entre l'activité agoniste partielle et antagoniste partielle est ténue. Un antagoniste partiel peut raisonnablement être considéré agoniste partiel si l'on considère qu'il diminue l'effet d'un antagoniste complet. Les agonistes inverses sont les molécules capables d'exercer une activité opposée à l'activité agoniste dans le cas des récepteurs ayant une activité basale[432]. Il faut savoir que les NRs contrôlent la transcription de nombreux gènes et qu'il est important, dans un cadre thérapeutique, non seulement de cibler spécifiquement un NR pour éviter de multiples effets secondaires, mais aussi de développer autant que faire se peut des modulateurs gène-spécifiques[433]. Par ailleurs, la modulation des NRs ne dépend pas seulement du mode de fixation du ligand mais aussi du ratio co-activateurs/corépresseur disponible dans la cellule à l'état physiologique[426]. Certains modulateurs sont tissu-spécifiques. C'est le cas du tamoxifène, qui inhibe la prolifération induite par les estrogènes dans le sein par une action antagoniste mais qui joue un rôle agoniste dans l'os, le foie et le système cardio-vasculaire[434,435].

6.6 Intérêt thérapeutique

Du fait de leur rôle central dans une pléthore de processus physiologiques, les NRs sont impliqués dans un grand nombre de maladies. De nombreux modulateurs des NRs sont utilisés dans un cadre thérapeutique contre diverses maladies comme le diabète, les maladies cardiovasculaires, les maladies inflammatoires, ou encore le cancer du sein (Tableau 13)[429]. En 2003, il était estimé que 34/200 médicaments administrés ciblaient les NRs[425]. Par exemple le tamoxifène (Nolvadex®, Tamoxifene Mylan®), est administré contre le cancer du sein chez les femmes pré et post-ménopausées. Le raloxifène (Evista®, Optruma®), le lasofoxifène (Fablyn®) et le toremifène (Foreston®) sont des analogues du tamoxifène administrés contre l'ostéoporose chez les femmes menopausées pour les deux premiers, et contre le cancer du sein

hormono-sensible chez la femme ménopausée pour le dernier[436]. Plusieurs composés dérivés de la thiazolidinedione ciblant le récepteur PPARγ ont été mis sur le marché dans les années 1990 et administrés dans le cadre de diabètes de type 2. En 2007, une méta-analyse de données cliniques d'un dérivé de thiazolidinedione, le rosiglitazone, a révélé une augmentation de 43% du risque d'infarctus du myocarde chez les patients sous traitement, ce qui a conduit au retrait de marché en France de l'intégralité de cette famille de composé en 2011[437]. Des études américaines réfutent cependant les risques cardiovasculaires liés au rosiglitazone, et ont conduit la FDA à continuer à autoriser les dérivés de la thiazolidinedione et à lever la restriction précédemment appliquée en cas de maladie coronarienne[438]. La mifepristone, antagoniste des PR, (Mifegyne®, Korlym®) est quant à elle utilisée en tant que pilule d'avortement. Elle bloque l'action progestative nécessaire au maintien de la grossesse et entrave le développement embryonnaire. Les glucocorticoïdes, sont des molécules agonistes des PRs utilisées comme anti-inflammatoires et administrées dans le cadre de maladies inflammatoires, auto-immunes (polyarthrite rhumatoïde), contre les effets secondaires de la chimiothérapie ou encore contre les chocs anaphylactiques. En 1996, il était estimé que 0.5% de la population prenait de manière continue des glucocorticoïdes[439], et au début des années 2000s, 56 à 68% des patients souffrant de polyarthrite rhumatoïde étaient traités avec des glucocorticoïdes[440]. Ces exemples soulignent l'importance des NRs pour l'industrie pharmaceutique d'un point de vue thérapeutique et financier. De nombreuses études continuent d'être menées afin de développer des modulateurs sélectifs de NRs avec des effets secondaires réduits.

Récepteur nucléaire	**Abbréviation**	**Utilisation thérapeutique de modulateur du récepteur**
Récepteur aux androgènes	AR	Hypogonadisme, ostéoporose, fragilité, sarcopénie chronique, cachexie, anémie, problèmes de désir sexuel, contraception masculine
Récepteurs aux estrogènes	ER	Maladies cardiovasculaires chez les femmes ménauposées, ostéoporose, cancer du sein
Récepteurs aux glucocorticoïdes	GR	Maladies inflammatoires
Liver X Receptors	LXR	Athérosclérose, diabètes, maladies cardiovasculaires, maladies auto-immunes, maladie d'Alzheimer

Récepteurs à la progestérone	PR	Contraceptif d'urgence, traitement de l'endométriose (en développement clinique), avortement
Récepteurs activés par les proliférateurs de peroxysomes	PPAR	Diabète, athérosclérose
Récepteurs aux hormones Thyroïdiennes	TR	Hypothyroïdie, anti-cholestérol, maladie de Basedow, thyroïdectomie
Récepteurs à la vitamine D	VDR	Hyperparathyroïdie, carences en vitamine D

Tableau 13 Liste de maladies contre lesquelles des modulateurs de NRs sont administrés. Données issue de [429] et de la DrugBank [391]

D'un point de vue chémoinformatique, les études sur les récepteurs nucléaires sont nombreuses, mais aucune n'a conduit à la mise sur le marché d'une molécule.

6.7 Intérêt en santé publique

Outre(s) leur(s) ligand(s) naturel(s) et les médicaments développés pour les cibler, les NRs peuvent être régulés par des agents hydrophobes exogènes présents dans l'environnement et potentiellement toxiques : les perturbateurs endocriniens.

6.7.1 Perturbateurs endocriniens

La définition des perturbateurs endocriniens a été énoncée par l'OMS en 2002 : « Un perturbateur endocrinien est une substance ou un mélange de substances, qui altère les fonctions du système endocrinien et de ce fait induit des effets néfastes dans un organisme intact, chez sa progéniture ou au sein de (sous)- populations »[441]. Le mécanisme de toxicité des perturbateurs endocriniens n'est pas complétement élucidé mais leur liaison aux récepteurs nucléaires semble jouer un rôle important. Les pesticides ou d'autres composés chimiques issus de l'industrie peuvent interagir avec des récepteurs nucléaires et conduire à des effets non désirés incluant des anomalies congénitales, une détérioration de la capacité de reproduction ou encore une neuro-toxicité au stade du développement[442]. Par exemple les organostanniques comme le tributyl étain (TBT), le dibutyl étain (TPT) ou le monobutyl étain sont entre autres utilisés pour prévenir l'incrustation d'organismes aquatiques sur les coques de bateau ou encore dans les matières plastiques et les pesticides. En 1997, l'Ifremer a confirmé la pollution de l'ensemble du littoral français avec des concentrations de TBT pouvant aller jusqu'à 200 fois la dose toxique de 1ng/L[443]. Des phénomènes d'imposex, c'est à dire de développement d'organes génitaux du sexe opposé à celui d'un individu, ont été observé chez les gastropodes[444][445] et

directement liés à l'interaction d'organostanniques avec l'hétérodimère RXRα-PPARγ. Chez l'homme, le tributyl étain est reconnu comme obésogène. En interagissant avec les RXRs et PPARs, le tributyl étain perturbe l'homéostasie des lipides et stimule la différenciation de pré-adipocytes en adipocytes. Toujours chez l'Homme, l'exposition aux phthalates est responsable de retardement de la puberté et de la diminution de la distance ano-génitale chez les hommes. Il faut savoir que plus cette distance est faible, plus il y a de risque de sous-fertilité voire infertilité[446]. Par ailleurs, une étude réalisée sur 14,947 hommes de 61 nationalités différentes entre 1938 et 1992 a montré une baisse de la qualité du sperme entre les années 1940 (113 millions de spermatozoïdes/ml) et 1990 (66 millions de spermatozoïdes/ml). Si aucun lien direct avec les NRs n'a été établie, de nombreuses études épidémiologiques suggèrent un lien entre cette baisse de qualité du sperme et l'exposition aux perturbateurs endocriniens. D'autres altérations des fonction des NRs ont été imputées à des perturbateurs endocriniens[29] (Tableau 14). Tous ces éléments expliquent le grand intérêt qui est porté actuellement pour identifier les perturbateurs endocriniens et évaluer leur toxicité et les efforts menés en ce par la recherche académique. C'est le cas notamment d'une initiative américaine détaillée en 1.2.2 mais aussi d'une initiative française qui a mis en place de la 2ème Stratégie Nationale sur les Perturbateurs Endocriniens (SNPE2). La SNPE2 vise à élargir les connaissances sur les perturbateurs endocriniens, la communication auprès du grand publique et susciter des engagement auprès des professionnels de l'industrie pour substituer les perturbateurs endocriniens avérés[447].

Tableau 14 Exemple de récepteurs nucléaires humains dérégulés par des perturbateurs endocriniens. D'après [29]

Récepteur	**Abréviation**	**Fonction physiologique**	**Ligand endogène**	**Exemple de perturbateurs endocriniens**
Récepteur aux androgènes	AR	Développement de l'appareil sexuel masculin	Testostérone	Pesticides Phthalates Plastifiants Composés polyhalogénés
Récepteurs aux estrogènes	ER α, β	Développement de l'appareil sexuel féminin	Estradiol	Alklyphénoles Bisphénol A Dioxines Furanes

				Hydrocarbures halogénés Métaux lourds
Récepteurs aux hormones thyroïdiennes	TR α, β	Métabolisme Rythme cardiaque	Hormone thyroïdienne	Bisphénol A Dioxines Furanes Polybromodiphényléthers Polychlorobiphényles Perchlorates Pesticides Phalates Phytoestrogènes
Récepteurs à la progestérone	PR	Développement de l'appareil sexuel féminin	Progestérone	Bisphénol A Fongicides Herbicides Insecticides
Aryl Hydrocarbon Receptor	AhR	Rythme circadien Métabolisme Neurogénèse Développement d'organes Réponse au stress	Inconnu	Dioxines Flavonoïdes Herbicides Indoles Polychlorobiphényles Pesticides
Récepteurs activés par les proliférateurs de peroxysomes	PPAR α, β, λ	Homéostasie des lipides	Lipides/ Acides gras	Bisphénol A Organotines
Récepteurs aux glucocorticoïdes	GR α, β	Développement Métabolisme Réponse au stress	Cortisol	Arsenic Bisphénol A Phthalates

6.7.2 Évaluation du caractère perturbateur endocrinien

6.7.2.1 Méthodes expérimentales

D'un point de vue expérimental, l'étude la plus poussée sur l'évaluation de potentiels perturbateurs endocrinien ayant une action sur les NRs a été mise en place par plusieurs agences fédérales américaines (le *National Institutes of Health* (NIH), *U.S. Department of Health and Human Services* et l'*U.S. Environmental Protection Agency* (EPA)). Il s'agit du programme *Toxicology in the 21st Century*, ou Tox21, qui a permis d'évaluer la toxicité d'une collection de 10,000 composés. Cette étude inclut les résultats de 30 essais cellulaires réalisés par HTS quantitatif, et inclus des essais sur 10 récepteurs nucléaires (AR, ERα, FXR, GR, LXRβ, PPARγ, PPARδ, RXRα, TRβ, et VDR)[442]. L'ensemble des données générées par ce programme est accessible gratuitement et peut être utilisé pour la construction de modèles de prédiction de la toxicité. Des modèles relativement performant (ROC-AUC > 0.7) ont été créés par Huang et al. pour prédire le type de toxicité d'un composé, sans toutefois prédire le récepteur associé à la toxicité[448].

6.7.2.2 Méthodes in silico

Il existe plusieurs outils *in silico* pour prédire une affinité entre une petite molécule et un récepteur nucléaire et ainsi identifier de potentiels perturbateurs endocriniens. Par exemple VirtualToxLab permet de prédire le de toxicité de petites molécules en combinant docking flexible et analyses QSAR sur 16 cibles fréquemment associées à des effets secondaires[34]. Parmi les cibles étudiées 11 sont des NRs (AR, AhR, ERα, ERβ, GR, LXR, MR, PPARγ, PR, TRα, TRβ), 4 sont des cytochromes P (CYP1A2, CYP2C9, CYP2D6, CYP3A4), et la dernière est le canal potassique hERG. Les modèles de prédiction proposés par VirtualToxLab atteignent un r^2 de prédiction moyen de 0.747, avec un r^2 minimum à 0.652 (PR) et un r^2 maximum à 0.885 (ERα). Cette méthode ne permet cependant pas de distinguer les agonistes des antagonistes d'un récepteur nucléaire et, bien que l'outil soit disponible gratuitement sur demande, il n'offre pas la possibilité de charger plusieurs molécules à la fois et présente donc un usage non adapté au criblage à haut débit. Endocrine Disruptome est un autre outil développé par Kolsěk[449] entièrement basé sur du docking à l'aide d'Autodock VINA. Un docking avec les molécules actives sur les NRs issus de la DUD-E et les agonistes et antagonistes des NRs sélectionnés issus de la ChEMBL a été effectué et a permis d'identifier pour chaque NR, la structure protéique conduisant au meilleur enrichissement. Il faut noter qu'Endocrine Disruptome distingue les structures agoniste-liées et antagoniste-liées lors du docking et proposent une prédiction par récepteur nucléaire et par profil pharmacologique (agoniste et

antagoniste) pour 4 récepteurs : AR, ERα, Erβ et GR. La distribution des résultats de docking a permis d'établir pour chaque NR un seuil de faible, moyenne et forte probabilité d'agir comme perturbateur endocrinien. L'outil est disponible gratuitement en ligne mais ne permet de tester plusieurs composés simultanément. Il existe également de nombreux modèles QSAR qui ont été développés à cet effet.

7 Conclusion et objectifs de thèse

Les approches *in silico* offrent la possibilité de pouvoir prédire une interaction entre une petite molécule et une protéine. Le nombre grandissant de données disponibles gratuitement augmente le champ des possibles et permet aussi bien d'utiliser ces méthodes avec une visée thérapeutique à long terme ou bien une visée préventive. La nature et la diversité des données permet de mettre en place des méthodes variées, comme les méthodes basées sur les ligands ou sur la structure, capables d'apporter des informations complémentaires et sources de connaissance. Un bémol réside cependant dans le manque de publication des données négatives ; ce manque a longtemps contraint la communauté scientifique à utiliser des molécules supposées inactives lors de la génération et l'évaluation de modèles de prédiction. Ces molécules supposées inactives, les *decoys*, ont permis de prioriser des méthodes lors d'études de criblage et ont conduit à la découverte de *hits* dont certains ont évolué en médicaments[60]. Aujourd'hui, il est possible, pour les familles les plus étudiées comme les récepteurs nucléaires, de tirer profit des multiples études publiées pour apprécier l'impact de l'inclusion de molécule inactives sur la création et l'évaluation de modèles de criblage virtuel. C'est précisément le sujet qui a été exploré au cours de cette thèse et qui est présenté au travers des différentes publications dont elle a fait l'objet. Les récepteurs nucléaires, de par l'abondance de publications les concernant et l'intérêt majeur de leur dans un cadre aussi bien thérapeutique que de santé publique et environnementale, se sont imposés comme un cas d'étude idéal. La première étape de cette étude a été la collecte de données. Nous avons décidé d'utiliser la banque ChEMBL qui regroupe notamment des données issues de la littérature et qui permet de retracer l'origine de chaque donnée. Une étude réalisée au laboratoire en 2014 avait cependant quantifié à 30% les erreurs liées aux données concernant les récepteurs nucléaires. Dans le cadre de ma thèse, j'ai donc effectué un travail conséquent de vérification de chacune des données via la relecture des 1513 publications relatant des données d'affinité pour les récepteurs nucléaires. J'ai analysé et formaté la banque de donnée finale, appelée Nuclear Receptor DataBase Including Negative Data (NR-DBIND), puis j'ai créé le site http://nr-dbind.drugdesign.fr afin de la publier gratuitement et de la rendre exploitable par l'ensemble de la communauté scientifique. Une fois ce travail réalisé, nous avons exploité la base pour répondre aux interrogations suivantes :

- Sommes-nous capables de discriminer les molécules actives de la NR-DBIND des molécule inactives qu'elle recense avec des outils de docking libres d'accès ?

- Les performances obtenues diffèrent-elles de celles obtenues sur une banque de données constituée de molécules actives et de *decoys* générés par l'outil DUD-E, qui fait office de référence ?
- Peut-on construire des modèles de pharmacophores plus robustes en intégrant des données d'inactivité en complément aux données d'activité du jeu d'apprentissage ?

La partie Résultats de ce manuscrit apporte des éléments de réponse. Elle est organisée en 5 parties. La première partie est une revue sur l'évolution de la notion de *decoys* dans les banques de données de benchmark, la deuxième détaille le contenu de la NR-DBIND, la troisième présente l'étude de docking effectuée sur la NR-DBIND et la comparaison de l'utilisation de molécules inactives à l'utilisation de *decoys* lors de l'évaluation des performances des outils de docking, et la quatrième concerne la construction de modèles de pharmacophore tenant compte de l'information apportée par les molécules inactives et leur application sur une banque de données externe issue de la Tox21. Enfin, la 5ème partie regroupe deux études annexes d'application de protocole de criblage pour 1) classer des ligands du récepteur nucléaire FXR en fonction de leur affinité pour la cible, et 2) comprendre la différence d'activité déclenchée par l'interaction des molécules NRPa-47 et NRPa-48 avec la protéine NRP-1, la seule différence structurale entre ces deux molécules résidant dans l'absence d'un méthyl sur NRPa-48.

Résultats

1 Sélection des decoys dans les banques de données d'évaluation : historique et perspéctives

1.1 Introduction

La sélection de *decoys* est un élément clé de la construction d'une banque de données d'évaluation. De la sélection aléatoire de molécules, à la sélection rationnelle de petites molécules répondant à des critères de similarité physicochimique et de distance topologique, le choix des *decoys* a largement évolué depuis la création des premières banques de données. Le but de l'optimisation de la sélection des *decoys* est de minimiser les potentiels biais introduits et de resserrer l'écart entre les performances obtenues lors d'études rétrospectives et prospectives. La revue « Decoys Selection in Benchmarking Datasets: Overview and Perspectives » publiée dans le journal Frontiers in Pharmacology en janvier 2018 retrace l'évolution de la sélection des *decoys*, décrit la composition de banques de données de référence et propose des recommandations pour l'intégration de molécules inactives dans les banques de données.

1.2 Article

frontiers in Pharmacology

REVIEW
published: 24 January 2018
doi: 10.3389/fphar.2018.00011

Decoys Selection in Benchmarking Datasets: Overview and Perspectives

*Manon Réau†, Florent Langenfeld†, Jean-François Zagury, Nathalie Lagarde and Matthieu Montes**

Laboratoire GBA, EA4627, Conservatoire National des Arts et Métiers, Paris, France

Virtual Screening (VS) is designed to prospectively help identifying potential hits, i.e., compounds capable of interacting with a given target and potentially modulate its activity, out of large compound collections. Among the variety of methodologies, it is crucial to select the protocol that is the most adapted to the query/target system under study and that yields the most reliable output. To this aim, the performance of VS methods is commonly evaluated and compared by computing their ability to retrieve active compounds in benchmarking datasets. The benchmarking datasets contain a subset of known active compounds together with a subset of decoys, i.e., assumed non-active molecules. The composition of both the active and the decoy compounds subsets is critical to limit the biases in the evaluation of the VS methods. In this review, we focus on the selection of decoy compounds that has considerably changed over the years, from randomly selected compounds to highly customized or experimentally validated negative compounds. We first outline the evolution of decoys selection in benchmarking databases as well as current benchmarking databases that tend to minimize the introduction of biases, and secondly, we propose recommendations for the selection and the design of benchmarking datasets.

Keywords: virtual screening, benchmarking databases, benchmarking, decoy, structure-based drug design, ligand-based drug design

OPEN ACCESS

Edited by:
Adriano D. Andricopulo, University of São Paulo, Brazil

Reviewed by:
Katarina Nikolic, University of Belgrade, Serbia
Francesco Ortuso, Magna Græcia University, Italy

***Correspondence:**
Matthieu Montes
matthieu.montes@cnam.fr

†These authors have contributed equally to this work.

Specialty section:
This article was submitted to Experimental Pharmacology and Drug Discovery, a section of the journal Frontiers in Pharmacology

Received: *10 November 2017*
Accepted: *05 January 2018*
Published: *24 January 2018*

Citation:
Réau M, Langenfeld F, Zagury J-F, Lagarde N and Montes M (2018) Decoys Selection in Benchmarking Datasets: Overview and Perspectives. Front. Pharmacol. 9:11. doi: 10.3389/fphar.2018.00011

INTRODUCTION

Computer-aided drug design (CADD) is now a commonly integrated tool in drug discovery processes (Sliwoski et al., 2014). It represents a way to predict ligands bioactivity *in silico*, and help focusing the drug discovery efforts on a limited number of promising compounds, saving both time and money in this very competitive field. Among these computational methods, Virtual Screening (VS) is designed to prospectively help identifying potential hits, i.e., compounds able to interact with the target and to modulate its activity, out of large compound collections (Tanrikulu et al., 2013). VS approaches can be Ligand-Based (LBVS) when they rely only on the structure/properties of known active compounds to retrieve promising molecules from compound collections (using similarity search, QSAR or 2D/3D pharmacophore, etc.), or Structure-Based (SBVS) if the structural information of the target is used (like in molecular docking studies).

The evaluation of VS methods is crucial prior to large library prospective screening to select the appropriate methodology, and subsequently generate reliable outcome on real-life project. Thus, software and workflows must be thoroughly evaluated retrospectively using benchmarking datasets. Such datasets are composed of known active data together with inactive compounds referred to as "decoys" (Irwin, 2008). Ideally, both active and inactive compounds should be selected on the basis of experimental data. However, the documentation on inactive data is scarce,

and putative inactive compounds are generally used instead. Among the common metrics used to estimate the performance of VS methods we find receiver operating characteristics (ROC) curves, the area under the ROC curve (ROC AUC) (Triballeau et al., 2005), Enrichment Curves (EC), Enrichment Factors (EF) and predictiveness curves (Empereur-mot et al., 2015). While conceptually different, they all share the same objective: assess the ability of a given method to identify active compounds as such, and discriminate them from the decoy compounds.

However, since the publication of the first benchmarking database in the early 2000s, the composition in both active and decoy compounds have been pointed out to crucially impact VS methods evaluation; several biases have been shown to incline VS assessment outcomes positively or negatively. The difference between the two chemical spaces defined by the active compounds on the one hand and the decoy compounds on the other hand may lead to artificial overestimation of the enrichment (Bissantz et al., 2000). On the contrary, the possible presence of active compounds in the decoy compounds set may introduce an artificial underestimation of the enrichment (Verdonk et al., 2004; Good and Oprea, 2008) since decoys are usually assumed to be inactive rather than proved to be true inactive compounds (i.e., confirmed inactive through experimental bioassays). New databases were designed to minimize those biases (Rohrer and Baumann, 2009; Vogel et al., 2011; Mysinger et al., 2012; Ibrahim et al., 2015a). Finally, many studies pointed out that the VS performance depends on the target and its structural properties (structural flexibility, binding site physicochemical properties, etc.; Cummings et al., 2005). Taking this into consideration, and despite the growing number of protein families represented in databases, decoy datasets generation tools were made publicly available in order to allow any scientist to fine-tune target-dependant and reliable benchmarking datasets (Mysinger et al., 2012; Ibrahim et al., 2015a).

In this review, we first present how the notion of decoy compounds evolved from randomly selected putative inactive compounds to rationally selected putative inactive compounds and finally true negative compounds. We develop the successive benchmarking datasets that were published in the literature and their basic to highly refined decoys selection workflows together with the resulting positive or negative biases due to their design. We then detail 5 benchmarking databases or decoy sets generator tools along with their detailed decoy compounds selection that represent the current state-of-the-art as of 2017: their respective composition tend to minimize such biases. Finally, we propose recommendations to select minimally biased benchmarking datasets containing putative inactive compounds as decoy compounds and introduce guidelines to design true inactive compounds containing databases.

THE HISTORY OF DECOYS SELECTION

Randomly Selected Decoys

The first use of a benchmarking database to evaluate virtual screening tools dates back to 2000, with the pioneering work of Bissantz et al. (2000). The objective of their study was to evaluate the ligands enrichment, i.e., the ability of docking programs to associate active compounds with the best scores within a compound collection. Three docking programs [Dock (Kuntz et al., 1982), FlexX (Rarey et al., 1996), Gold (Jones et al., 1997)] combined with 7 scoring functions [ChemScore (Eldridge et al., 1997), Dock, FlexX, Fresno (Rognan et al., 1999), Gld, Pmf (Muegge and Martin, 1999), Score (Wang et al., 1998)] were evaluated on two different target proteins: Thymidine Kinase (TK) and the ligand binding domain of the Estrogen Receptor α subtype (ER α).

For each target, a dataset containing 10 known ligands and 990 molecules assumed to be inactive (decoy compounds) was created. The decoy compounds were selected following a two-step scheme: (1) the Advanced Chemical Directory (ACD v.2000-1, Molecular Design Limited, San Leandro) was filtered to eliminate undesired compounds (chemical reagents, inorganic compounds and molecules with unsuitable molecular weights), (2) 990 molecules were randomly selected out of the filtered dataset. The datasets were used to evaluate and compare several docking and scoring schemes. The authors eventually recommended a calibration of docking/consensus scoring schemes on reduced data sets prior to large dataset screens. Later on, Bissantz et al. (2003) applied the same protocol to three human GPCRs to investigate whether their homology models were suitable for virtual screening experiments.

A growing interest for virtual screening benchmarking databases soon emerged from the community (Kellenberger et al., 2004; Brozell et al., 2012; Neves et al., 2012; Repasky et al., 2012; Spitzer and Jain, 2012). New databases were designed with an increasing complexity in the decoys selection methodologies (see section Benchmarking Databases). Nowadays, benchmarking databases are widely used to evaluate various VS tools (Kellenberger et al., 2004; Warren et al., 2006; McGaughey et al., 2007; von Korff et al., 2009; Braga and Andrade, 2013; Ibrahim et al., 2015a; Pei et al., 2015) and to support the identification of hit/lead compounds using LBVS and SBVS (Allen et al., 2015; Ruggeri et al., 2015).

Integration of Physicochemical Filters to the Decoy Compounds Selection

In the early 2000s, Diller's group incorporated filters in the decoys selection to ensure that the discrimination they observed was not solely based on the size of the decoy compounds (Diller and Li, 2003). In addition to the 1,000 kinases inhibitors they retrieved from the literature for 6 kinases (EGFr, VEGFr1, PDGFrβ, FGFr1, SRC, and p38), 32,000 compounds were randomly selected from a filtered version of the MDL Drug Data Report (MDDR). The filters were designed to select decoy compounds displaying similar polarity and molecular weight. Similarly, in 2003, a benchmarking database derived from the MDDR was constructed by McGovern et al. (McGovern and Shoichet, 2003). Compounds with unwanted functional groups were removed, leading to 95,000 compounds. The targets of the MDDR for which at least 20 known ligands were available constituted a target dataset (CA II, MMP-3, NEP, PDF, and XO). The remaining compounds were used as decoy compounds. The addition of rational filters was a considerable step forward in the improvement of decoys selection, but due to the commercial licensing of the MDDR,

its use was limited (http://www.akosgmbh.de/accelrys/databases/mddr.htm[1]).

The first benchmarking databases were composed as follows: (1) true active compounds consisted in known ligands extracted from the literature while (2) decoy compounds consisted in putative inactive compounds randomly selected from large databases possibly filtered to be compliant to specific criteria (drug likeness, molecular weight, topological polar surface area…). Since the decoy compounds were pseudo-randomly selected, they were assumed to be inactive on the defined targets.

Despite the use of the MDDR and the filtering of the decoy compounds, these benchmarking databases displayed a major drawback: the significant differences occurring between the physicochemical properties of the active compounds and decoy compounds led to obvious discrimination and then artificially good enrichments (Verdonk et al., 2004; Huang et al., 2006).

In 2006, Irwin et al. proposed that the decoy compounds should be similar to the known ligands regarding their physicochemical properties to reduce the introduction of bias while being structurally dissimilar to the known ligands to reduce their probability to be active on the defined target. Following these recommendations, they created the DUD database (Huang et al., 2006) that was immediately considered as the gold standard for the evaluation of VS methods.

The DUD database is composed of 2,950 ligands and 95,326 decoys for a total of 40 proteins from 6 classes (nuclear hormone receptors, kinases, serine proteases, metalloenzymes, folate enzymes and others). The decoy compounds were extracted from the drug-like subset of the ZINC database (Irwin and Shoichet, 2005). The 2D-similarity between known ligands and decoy compounds was computed by calculating the Tanimoto distance based on the CACTVS type 2 substructure keys and 5 physicochemical properties. For each active compound, the 36 molecules sharing the most similar properties while being topologically dissimilar (Tanimoto < 0.9) were conserved. The evaluation of the performance of DOCK (Meng et al., 1992; Wei et al., 2002; Lorber and Shoichet, 2005; Huang et al., 2006) confirmed that uncorrected databases such as the MDDR led to over-optimistic enrichments compared to corrected databases such as the DUD.

Benchmarking Database Biases

Despite the precautions taken to build the DUD database, several remaining biases have been reported in the literature.

The "analogous bias" (Good and Oprea, 2008) lies in the limited chemical space of active compounds that is restricted to the chemical series that have been explored and referenced in databases. The discrimination of the active compounds from decoy compounds can be simplified since the decoy sets would display a larger structural variability that could induce an overestimation of the performance of VS methods. The lack of diversity in the structures of known active compounds limits the training and evaluation of LBVS methods to perform scaffold-hopping, i.e., the identification of active hit compounds that structurally differ from reference molecules while retaining similar activity.

The "complexity bias" (Stumpfe and Bajorath, 2011) or "artificial enrichment bias": active compounds and decoy compounds often display differences in their respective structural complexity since active compounds are often optimized compounds extracted from large series in the scientific and patent literature, which is not necessarily the case for the structures of pseudo-randomly selected decoy compounds.

The "false negative bias" (Vogel et al., 2011; Bauer et al., 2013) lies in the presence of active compounds in the decoy set. Unlike the analogous and complexity biases, it induces an underestimation of the performance of the VS methods that could be particularly dramatic for the evaluation of LBVS methods (Irwin, 2008).

The need for less biased benchmarking databases to objectively evaluate VS methods favored the emergence of new strategies to eradicate or at least minimize those biases. Two decoys selection strategies arose from benchmarking databases improvement attempts: (1) the use of highly refined decoys selection strategies and (2) the integration of true negative compounds in the decoy set.

Highly Refined Putative Inactive Compounds Selection

The reported biases pointed out that the composition of both active compounds and decoy compounds sets has a huge impact on the evaluation of the performance of VS methods (Verdonk et al., 2004; Good and Oprea, 2008). Therefore, particular efforts were performed in the selection strategies for active compounds and decoy compounds.

To address analogous bias, a strategy consists in modifying the receiver operating characteristics (ROC) curves (i.e., the fraction of actives among the top fraction x of the data set) (Triballeau et al., 2005) by weighting the rank of each active compound with the size of its corresponding lead series (Clark and Webster-Clark, 2008). This allows an equal contribution of each active chemotype to the ROC curve (rather than each active compound). Another widely used method is to fine-tune the active compounds dataset prior to screen to ensure an intrinsic structural diversity. To this aim, the MUV datasets (Rohrer and Baumann, 2009) were designed using the Kennard Jones algorithm to obtain an optimal spread of the active compounds in the decoy compounds chemical space while ensuring a balance between the active compounds self-similarity and separation from the decoy compounds. Despite these observations, the most used strategy in the literature still consists in clustering ligands based on 2D descriptors and retain only cluster representatives in the final dataset (Good and Oprea, 2008; Mysinger et al., 2012; Bauer et al., 2013).

To reduce artificial enrichment, efforts were made to match as much as possible the physicochemical properties of the decoys to the physicochemical properties of the active compounds. To this aim, the Maximum Unbiased Validation database (MUV) (Rohrer and Baumann, 2009) was designed to ensure embedding of active compounds in the decoy compounds

[1]MDDR licensed by Molecular Design, Ltd., San Leandro, CA.

chemical space based on an embedding confidence distance cut-off calibrated on multiple drug-like compounds banks' chemical space. Active compounds that were poorly embedded in the decoy set were discarded. A way to ensure the availability of potential decoy compounds for any ligand is to generate decoys that ignore synthetic feasibility (Wallach and Lilien, 2011). Other databases select decoys that match active compounds in a multiple physicochemical properties space. The DEKOIS 2.0 (Ibrahim et al., 2015a) proposed a workflow that used 8 physicochemical properties while the DUD-E added net charge to the 5 physicochemical properties already considered in the original DUD.

To address the risk of including false negatives in the decoy set, a common strategy is to select decoy compounds topologically different to any active compound. For this purpose, Bauer et al. introduced the LADS score to guide decoys selection (Vogel et al., 2011). In the DUD-E, potential false decoys are avoided by applying a stringent FCFP_6 fingerprints Tanimoto-based filter. It is important to note that since the evaluation of LBVS methods requires that decoy compounds should not be discriminated using basic 2D-based similarity tools, the use of 2D-based dissimilarity filters to avoid false negatives in the decoy set makes the concerned databases inappropriate for the evaluation of the performance of LBVS methods. Therefore, Xia et al. developed a method to select adequate decoys for both SBVS and LBVS (Xia et al., 2014) by favoring physicochemical similarity as well as topological similarity between active compounds and decoy compounds that passed a primary topological dissimilarity filter.

With these improvements, the notion of decoys remained the same—putative inactive compounds—but their selection critically evolved. Ever since, the main progress achieved in the literature lies in the diversification of the protein targets represented in benchmarking databases. The growing need for datasets dedicated to a given target led to (1) an increasing diversity of targets in benchmarking databases [the DUD-E (Mysinger et al., 2012) contains datasets against 102 targets while the previous DUD (Huang et al., 2006) contained datasets only for 40 targets] and (2) highly specialized benchmarking databases focused on a particular class of targets. Such specialized datasets exist for GPCRs [GPCR ligand library (GLL)/Decoy Database (GDD) (Gatica and Cavasotto, 2012)], histone deacetylases [maximal unbiased benchmarking data sets for HDACs—MUBD-HDACs (Xia et al., 2015)], or nuclear receptors [NRLiSt BDB (Lagarde et al., 2014a)]. As a notice, DUD-E or DecoyFinder (Cereto-Massagué et al., 2012) offer automated decoy set generation tools based on the properties of active compounds, enabling the community to easily design and tune their own dataset for a particular target.

Toward True Negative Compounds

A common issue about decoys is the lack of data regarding their potential bioactivity against the target. Most methods assume that the absence of data means an absence of activity, which may lead to include unknown active ligands into a decoy set. To eliminate such false negatives from decoy sets, one solution is to use referenced true negative compounds that can be either true inactive or compounds displaying an undesirable activity.

True inactive compounds, i.e., compounds that displayed no experimental binding affinity against the target of interest, can be used to identify binders. Inactive data is made available in several public activity and/or affinity annotated compound repositories and high throughput screening (HTS) initiatives such as: ChEMBL (Bento et al., 2014), Drugbank (Wishart et al., 2008) that provides annotations for approved drugs; PDBBind (Wang et al., 2004, 2005), Binding MOAD (Benson et al., 2008) and AffinDB (Block et al., 2006) that contain binding affinity data for protein–ligand complexes available in the Protein Data Bank (PDB) (Berman et al., 2000); PDSP Ki database (Roth et al., 2000) that stores screening data from the National Institute of Mental Health's Psychoactive Drug Screening Program; BRENDA (Placzek et al., 2017) that provides binding constants for enzymes; IUPHAR (Southan et al., 2016) that contains binding information for receptors and ion channels; GLIDA (Okuno et al., 2006) and GPCRDB (Munk et al., 2016) that contains binding data for G-protein-coupled receptors; D3R datasets (Drug Design Data Resource[2]) that have been provided by pharmaceutical companies and academia and contain affinity data for 7 proteins together with inactive compounds; ToxCast™/Tox21 (Kavlock et al., 2012) and PCBioAssay (Wang et al., 2017) that provide HTS data for various targets.

As an example, the DUD-Enhanced (Mysinger et al., 2012) (DUD-E) integrates some experimentally validated inactive compounds extracted from ChEMBL in the decoy set in addition to putative inactive compounds: an arbitrary 1 μM cutoff is used to classify ligands in the active set while molecules with no measurable activity at 30 μM or higher concentration were classified into the decoy set. Similarly, the Maximum Unbiased Validation (MUV) (Rohrer and Baumann, 2009) datasets are composed of both active and inactive compounds collected from the PubChem BioAssay annotated database.

Unwanted compounds, i.e., compounds that display unwanted activity or binding, can also be used as negatives. For instance, a recent study used ligands of the NRLiSt BDB (Lagarde et al., 2014a) either as active compounds or decoy compounds, depending on their activity for each nuclear receptor; antagonist (or agonist) ligands of a given nuclear receptor were used as decoys to evaluate agonistic (or antagonistic) pharmacophores (Lagarde et al., 2016, 2017). This strategy has shown successful results in the past: Guasch et al. (2012) focused on PPAR γ partial agonists to avoid side effects accompanying full receptor activation and built an anti-pharmacophore model with known full agonist compounds to remove all potential full agonist compounds from their initial set of 89,165 natural products and natural product derivatives. The authors screened the remaining compounds on a partial agonist pharmacophore model and identified 135 compounds as potential PPARγ partial agonists with good ADME properties among which 8 compounds with new chemical scaffolds for PPARγ partial agonistic activity. After

[2] Available at: drugdesigndata.org

BENCHMARKING DATABASES

DB name	Year	Download address	Origin of the ligands	Origin of the decoys	No. of targets / No. of classes	Decoy compounds selection	Remarks
Rognan's decoy set (Bissantz et al., 2000)	2000	http://bioinfo-pharma.u-strasbg.fr/labwebsite/download.html	Literature	ACD	2/2	Random selection	Design of decoy sets to evaluate the performance of 3 docking programs and 7 scoring functions
Shoichet's decoy set (McGovern and Shoichet, 2003)	2003	–	MDDR	MDDR	9/4	Remove compounds with unwanted functional groups	Compare VS performance depending on the binding site definition (apo, holo or homology modeled structures)
Li's decoy set (Diller and Li, 2003)	2003	–	Literature	MDDR	6/1	Fit polarity and MW to known kinases inhibitors	Compare decoys and ligands physicochemical properties to select decoys
Jain's decoy set (Jain and Nicholls, 2008)	2006	http://www.jainlab.org/downloads.html	PDBbind	ZINC "drug-like" and Rognan's decoy set	34/7	1,000 random molecules from the ZINC that comply to MW $\leq$ 500, logP $\leq$ 5, HBA $\leq$ 10, HBD $\leq$ 5 and RB $\leq$ 12 and Rognan's decoys with RB $\leq$ 15	Use of 5 physicochemical properties to match decoy sets to ligands sets
Directory of Useful Decoys (DUD) (Huang et al., 2006)	2006	http://dud.docking.org	Literature and PDBbind	ZINC "drug-like"	40/6	Decoys must be Lipinski-compliant. The selection is based on both the topologically dissimilarity to ligands and the fit of physicochemical properties	Largest decoy data set so far (40 proteins) and first attempt to select decoys topologically dissimilar decoys
DUD Clusters (Meyer, 2007)	2008	http://dud.docking.org/clusters/	DUD	–	40/6	–	DUD clusters more relevant for scaffold hopping
WOMBAT Datasets (Meyer, 2007)	2007	http://dud.docking.org/wombat/	WOMBAT	–	13/4	–	Design to decrease the analog bias on 13 of the 40 DUD targets, enrich DUD active data sets with compounds from WOMBAT database
Maximum Unbiased Validation (MUV) (Rohrer and Baumann, 2009)	2009	–	PubChem	PubChem	18/7	Two functions measure the active-active and decoy-active distances using 2D chemical descriptors. Actives with the maximum spread within the active set were chosen and decoys with similar spatial distribution were selected	Ligands and decoys are from biologically actives and inactive compounds, i.e., are true actives and inactives, respectively
DUD LIB	2009	http://dud.docking.org/jahn/	DUD-cluster	DUD	13/4	Subset of the DUD database, with more stringent criteria on MW ($\leq$450) and AlogP ($\leq$4,5), and a minimal number of chemotypes	Initially designed for "scaffold-hopping" studies

(Continued)

DB name	Year	Download address	Origin of the ligands	Origin of the decoys	No. of targets / No. of classes	Decoy compounds selection	Remarks
Charge Matched DUD	2010	http://dud.docking.org/charge-matched/	DUD	ZINC	40/6	Apply a net charge property match on DUD datasets	
REPROVIS-DB	2011	–	Literature	Literature	–	Extracted from previous successful studies	Designed for LBVS only
Virtual Decoy sets (VDS) (Wallach and Lilien, 2011)	2011	http://compbio.cs.toronto.edu/VDS	DUD	ZINC	40/6	Same as DUD, but does not consider synthetic feasibility	Purely virtual decoys, availability is not considered
DEKOIS (Vogel et al., 2011)	2011	http://dekois.com/dekois_orig.html	BindingDB	ZINC	40/6	Class decoys and ligands into "cells" based on 6 physicochemical properties and select the closest decoys based on (1) a weighted physicochemical similarity and (2) a LADS score based on functional fingerprints similarity elaborated from the active set	Original treatment of the physicochemical similarity, and introduce the concept of *Latent Active in Decoy set*, i.e., false false positives
GPCR Ligand (GLL)/Decoys Database (GDD) (Xia et al., 2014)	2012	http://cavasotto-lab.net/Databases/GDD/	GLIDA and PDB structures and Vilar et al., 2010	ZINC	147/1	Physico-chemical properties fit and topological dissimilarity filter. Final selection based on MW	First extensive database targeting a specific protein family
Decoy Finder (Cereto-Massagué et al., 2012)	2012	http://urvnutrigenomica-ctns.github.io/DecoyFinder/	User	User	–	Same as DUD	Graphical tool to generate decoy data sets with adaptable thresholds for physicochemical properties
DUD Enhanced (DUD-E) (Mysinger et al., 2012)	2012	http://dud.docking.org/r2/	CHEMBL	ZINC	102/8	Physico-chemical properties fit along with a topological dissimilarity filter. Random selection of decoys is then applied	Largest database so far (1,420,433 decoys and 66,695 actives)
DEKOIS 2.0 (Ibrahim et al., 2015a)	2013	http://www.dekois.com	BindingDB	ZINC	81/11	Same as DEKOIS with 3 additional physicochemical properties (nFC, nPC, Ar), a PAINS filter and an improved, weighted LADS score	
NRLiSt BDB (Lagarde et al., 2014a)	2014	http://nrlist.drugdesign.fr	CHEMBL	ZINC and DUD-E decoys generator	27/1	Use the DUD-E decoy generation tool	Ligands can be either agonists or antagonists (other actives are removed), depending on the purpose of the study
MUBD-HDACs (Xia et al., 2015)	2015	–	CHEMBL and literature	ZINC	14/1	Select decoys based a weighted physicochemical similarity (6 physicochemical properties are considered), and ensure a random spatial distribution of the decoys (i.e., decoys should be as distant to the other actives as a reference ligand)	Applicable both to SBVS and LBVS strategies, uses ligands with proved bioactivity

biological tests, 5 compounds were confirmed to be PPAR γ partial agonists.

SELECTED DATABASES

Maximum Unbiased Validation (MUV)

The MUV was designed to propose unbiased datasets in regard to both artificial enrichment and analogous bias by proposing a new approach gleaned from spatial statistics (Rohrer and Baumann, 2009). The authors ensured homogeneity in actives-actives similarity and actives-decoys dispersion in order to reach a random-like distribution of active compounds and decoy compounds in a physicochemical descriptors chemical space. This implies that the molecular properties contained no information about the bioactivities of active and decoy compounds. Datasets were designed for 18 targets with a total of 30 actives and 15,000 decoys for each target.

Initial Compounds Database

Potential active and decoy compounds were extracted from HTS experiments available in PCBioAssay (June 2008) (PubChem BioAssay[3]). In these assays, a primary screen was performed in a large number of compounds (>50,000) and was followed by a low throughput confirmatory screen. Compounds with an experimental EC50 in the confirmatory screen were selected as potential active compounds while inactive compounds from the primary screen were selected as potential decoys.

Actives Selection

A two-step process was applied to rationally select final active compounds for the MUV data sets. (1) Potential active compounds were filtered to eliminate artifacts caused by organic chemicals aggregation in aqueous buffers ("Hill slope filter"), as well as off-targets, cytotoxic effects or interference with optical detection methods ["frequency of hits filter" and "autofluorescence (Simeonov et al., 2008) and luciferase inhibition (Auld et al., 2008) filters"]. (2) A "chemical space embedding filter" was applied to ensure that actives located in regions of the chemical space devoid of decoys were eliminated from the dataset (**Figure 1**). Subsets of 30 actives with the maximum spread per target were generated using a Kennard-Jones algorithm. Selected active compounds were exchanged with remaining potential active compounds until all datasets were adjusted to a common level of spread.

Decoys Selection

To carefully match active and decoys physicochemical properties, Rohrer et al. proposed that the level of self-similarity within the active compounds set [measured using the "nearest neighbor function" $G(t)$] should be equal to the degree of separation between the active compounds set and the decoy compounds set [evaluated with the "empty space function" $F(t)$] (**Figure 1**). Following guideline, the data clumping should be null, ensuring a random-like distribution of decoy and active compounds in the overall chemical space. The distances were computed based on 1D molecular properties (counts of all atoms, heavy atoms, boron, bromine, carbon, chlorine, fluorine, iodine, nitrogen, oxygen, phosphorus, and sulfur atoms in each molecule as well as the number of H-bond acceptors, H-bond donors, the logP, the number of chiral centers, and the number of ring systems). The level of separation between the decoy compounds and the active compounds was adjusted to the same level of spread so that the data clumping is null. In total, 500 decoys were selected per selected active, resulting in 15,000 decoys per dataset.

The minimization of analog bias and artificial enrichment makes the MUV datasets fitted for LBVS. The availability of structures in the PDB (2008) for seven of the MUV targets makes it suitable for SBVS as well (Löwer et al., 2011). Thus, the MUV constituted the first dataset that enabled comparative evaluations of SB and LBVS methods and protocols.

Demanding Evaluation Kits for Objective *in Silico* Screening (DEKOIS)

In 2011, Vogel et al. proposed a new generator of decoy compounds sets called *Demanding Evaluation Kits for Objective In Silico Screening* (DEKOIS) (Vogel et al., 2011). The authors designed their tool to avoid the introduction of well-known and described biases into the decoy sets, i.e., analog bias and artificial enrichment. A first step in their workflow is subsequently to closely match physicochemical properties of both ligand and decoys to limit the analog bias. Then, to deal with the risk of including false negative compounds in the decoy compounds set, a new concept is applied to the decoys selection process: the *latent actives in the decoy set* (LADS). Finally, the structural diversity of the active and decoy compounds structures into the sets is evaluated and maximized, and the embedding of the actives into the decoys chemical is assessed. The whole workflow was further improved in 2013 to produce the current version of this tool, DEKOIS 2.0 (Bauer et al., 2013), and 81 ready-to-use (active and decoys) benchmarking datasets for 11 target classes are currently available through the DEKOIS website (www.dekois.com/, accessed 10/23/2017).

Initial Compounds Database

Decoy compounds from the DEKOIS 2.0 benchmarking datasets are selected from a subset of the ZINC database of 15 million molecules. Eight physicochemical properties are evaluated: molecular weight, octanol–water partition coefficient, hydrogen bonds acceptor and/or donor, number of rotatable bonds, positive and negative charges, and the number of aromatic rings. For each physicochemical property, bins are defined, and all possible combinations of bins are used to split the database compounds into cells. The initial bins are defined so that each bin is equally populated, and each final cell is characterized by a set of 8 physicochemical properties. Each user-provided active compound is associated with the closest cell (in terms of physicochemical properties), and 1,500 decoys are randomly preselected from this parent cell, or from the direct neighbor cells if the parent cell is not populated enough to provide 1,500 decoy compounds (**Figure 1**).

Decoys Selection

The two criteria for the refinement steps are the structural diversity and the low rate of *latent active in decoy set* (LADS). A physicochemical similarity score (PSS) and a LADS score are

[3]Available online at: http://pubchem.ncbi.nlm.nih.gov/sources#assay

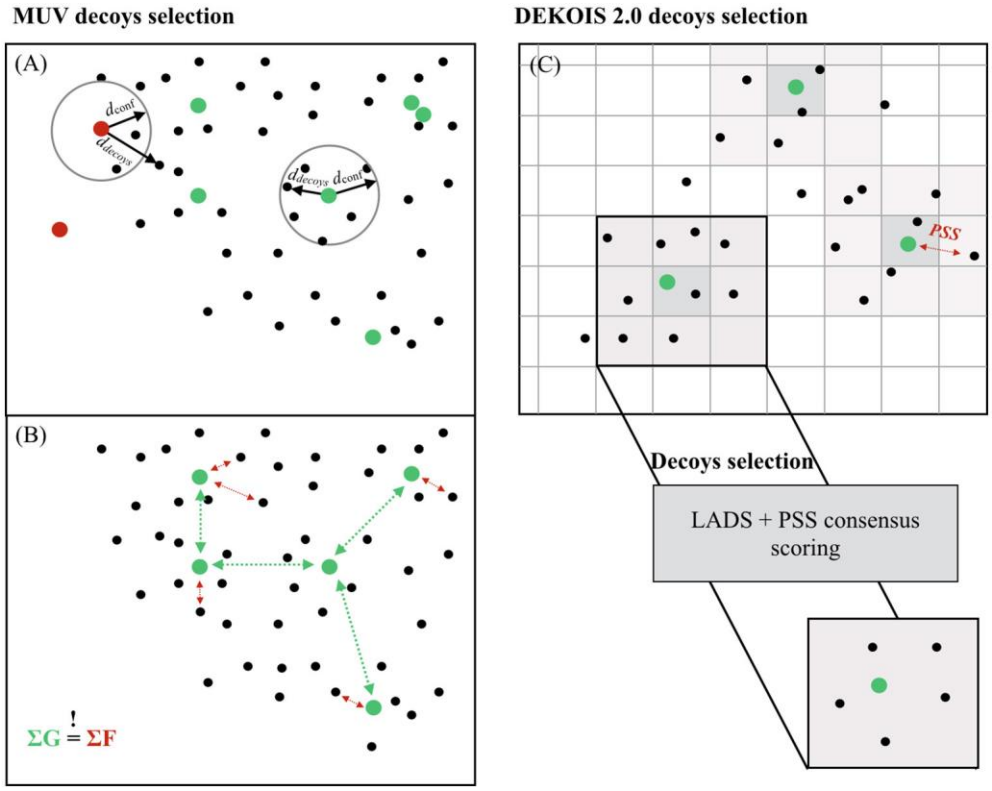


FIGURE 1 | Decoys selection in MUV and DEKOIS 2.0. **(A)** For each active of the MUV, a distance to the 500th nearest neighbor from 100 random samples from multiple drug-like compounds collections was computed. The 90th percentile was recorded as the confidence distance for a good embedding (d_{conf}). Active compounds were accepted only if the 500th nearest neighbor from the decoy compounds (d_{decoys}) set was within the d_{conf}. **(B)** Selected active compounds datasets from the MUV were adjusted to the same level of spread ($\Sigma G \approx$ constant), and decoy compounds sets were, in turn, adjusted to this level of spread ($\Sigma F \approx \Sigma G$). **(C)** The chemical space of both active and decoy compounds was divided into cells characterized by a set of 8 physicochemical properties. Each user-provided compound is associated with its property matching cell, and 1,500 decoys are selected from the same cell, or direct neighboring cells if the parent cell is not populated enough.

computed, normalized and combined to select the final 30 decoys associated with each active ligand:

(1) The PSS score is the arithmetic mean of the normalized distance between a decoy and the reference ligand, for each physicochemical property.
(2) The avoidance of LADS relies on the fingerprints bit strings shared by the active compounds: the fingerprint bit strings of each preselected decoy compound is matched to the fingerprint bit strings of all active compounds using the following:

$$LADS\ score = \frac{\sum_{i=1}^{n}\left(N_{i(HeavyAtoms)} \cdot f_{i(FCFP_6 fragment)}\right)}{N_{FCPC_6 fragments}},$$

with n the number of fingerprint bit strings shared by the decoy and the active set, f_i the frequency of fragment i in the active set, N_i the number of heavy atoms into fragment i, and N the total number of FCPC_6 fragments into the decoy.

The weighting of the LADS score by the frequency of the bit string and the size of the corresponding fragment was added in the second version of DEKOIS (Bauer et al., 2013) to ensure that large bioactive substructures and substructures frequently found exert a greater influence on LADS score compared to smaller and rare functional groups.
(3) The LADS and PSS scores are normalized and combined into a consensus score to sort decoy compounds. The subsequently best 100 decoys are selected. Finally, the fingerprints are used to select the 30 most dissimilar decoys for each active.

Using this enhanced protocol, Bauer et al. showed an improvement of the "deviation from optimal embedding score" (DOE score) (Vogel et al., 2011; Bauer et al., 2013) for DEKOIS 2.0 compared to DEKOIS, and found a good (<0.2) DOE score for 89% out of the 81 targets considered.

Dud-Enhanced (DUD-E)

Despite the extensive use of the DUD, several studies pointed out that some scaffolds were over-represented in the active sets, that the charge was not considered in property-matching for ligand selection, and that true ligands could be found in the decoy sets (Good and Oprea, 2008; Hawkins et al.,

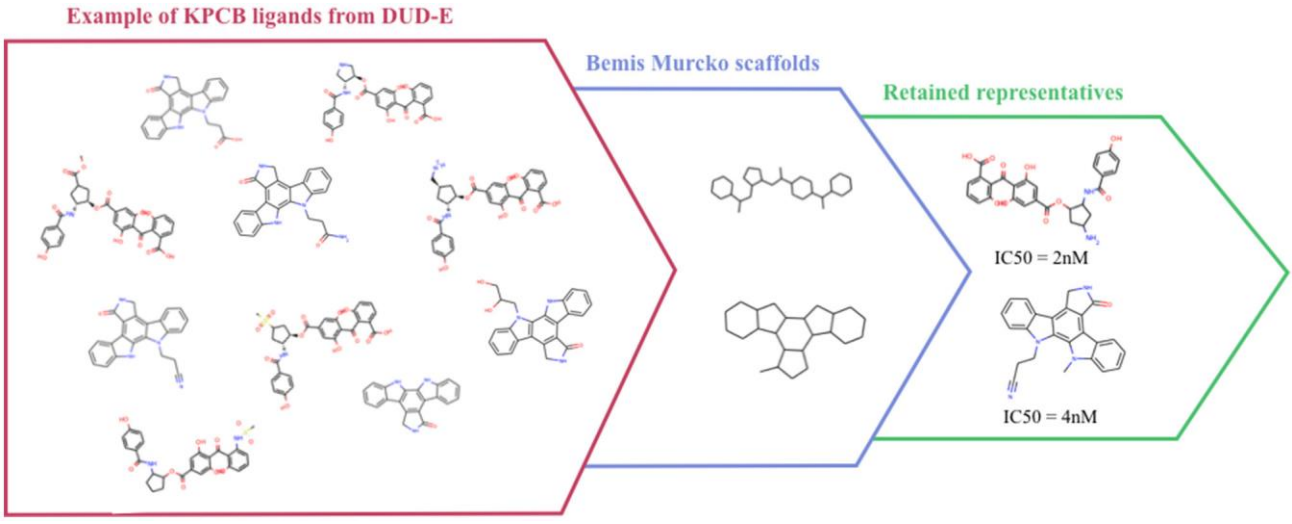


FIGURE 2 | Example of Bemis-Murcko atomic frameworks clustering of Protein kinase C beta type (KPCB) ligands from the DUD-E.

2008; Irwin, 2008; Mysinger and Shoichet, 2010). Shoichet et al. proposed the DUD-E (DUD-Enhanced) to address these weaknesses in both the active and the decoy sets design in the DUD, and extended the number of represented protein families in the database. The DUD-E contains 102 proteins that span diverse target classes. To address analogous bias, ligands were clustered by their Bemis-Murcko atomic frameworks (Bemis and Murcko, 1996) (**Figure 2**), and a topological dissimilarity filter was applied to avoid active compounds in the decoy sets.

Initial Compounds Database

Active compounds assigned to each target of the DUD-E were collected from the ChEMBL09 database if their activity/affinity (Ki, Kd, IC50, EC50, or associated logP) was $\leq 1\mu M$ (Gaulton et al., 2012). Additionally, 9,219 experimental decoys displaying no measurable affinity up to $30\,\mu M$ were included in the decoy sets.

Active Set Preparation

Active compounds were clustered based on their Bemis-Murcko atomic frameworks. When more than 100 frameworks were represented, the highest energy ligand from each cluster is considered, while when less than 100 frameworks are represented, the numbered of considered ligands was raised to obtain more than 100 molecules. Even if this selection protocol could have been optimized for sets with low frameworks diversity, it ensures sufficient diversity and quantity of compounds for the other sets.

Decoys Selection

The decoy compounds were extracted from the ZINC database (Irwin and Shoichet, 2005) and selected by narrowing or widening windows around 6 physicochemical properties: molecular weight, octanol-water partition coefficient, rotatable bonds, hydrogen bonds acceptors, hydrogen bonds donors, and the net charge. To avoid active compounds in the decoy sets, a topological dissimilarity filter was applied. Molecules were sorted according to their Tanimoto distance to any ligands using CACTVS fingerprints, and the 25% most dissimilar decoy molecules were retained. Finally, up to 50 decoys were randomly selected for each ligand and pooled with the 9,219 experimental decoys.

An automated tool was made available online to generate decoys from user-supplied ligands using the same protocol (http://decoys.docking.org). The possibility to generate decoy sets for any target has been revealed successful and is now widely used by the scientific community (Lacroix et al., 2016; Nunes et al., 2016; Allen et al., 2017; Meirson et al., 2017).

Despite the success of the DUD-E, some weaknesses should be corrected in the DUD-E benchmarking database. The 102 targets are defined as a UniProt gene prefix (such as DRD3) and not a full gene_species (such as DRD3_HUMAN or P35462), which can bias the actives selection when the binding site composition differs between species. Additionally, only one single structure was considered for each protein while many docking studies pointed out that the structure selection is crucial for screening and docking, particularly for proteins that accommodate ligands with different binding modes (May and Zacharias, 2005; Ben Nasr et al., 2013; Lionta et al., 2014). A recent study has shown that the ligand pharmacological profile should be considered for both the active set design and the structure selection (Lagarde et al., 2017). For instance, nuclear receptors (NR) can be inhibited by antagonists or activated by agonists that differ in their structure and properties: agonists should be considered in the active set if the screening is performed on an agonist-bound structure while antagonists should be used in the active set if the screening is performed on an antagonist-bound structure.

Nuclear Receptors Ligands and Structures Benchmarking Database (NRLiSt BDB)

The NRLiSt BDB (Nuclear Receptors Ligands and Structures Benchmarking DataBase) was created to address the lack of annotation information and pharmacological profile consideration in existing NR databases.

Ligands Preparation

The NRLiSt BDB is composed of 9,905 active molecules targeting 27 nuclear receptors (NRs). Active compounds are divided into 2 datasets per target according to their agonist or antagonist profile. All active compounds were extracted from the ChEMBL database and included in the NRLiSt after a manual inspection of the corresponding ligands bioactivity data in the original papers. All inverse-agonists, modulators, agonists/antagonists, weak to partial agonists, weak to partial antagonists and ligands with unknown pharmacological profile were discarded.

In addition 339 human holo structures extracted from the PDB are provided, among which 266 are agonists-bound, 17 are antagonists-bound and 56 are others-bound. Valid active compounds extracted from literature were clustered using chemical fingerprints, and a Tanimoto cut-off of 0.5.

Decoys Selection

In total 458,981 decoys generated with the DUD-E online tool were provided, with a mean ratio of 1/51 for each dataset.

In further studies, Lagarde et al. integrated the anti-pharmacological profile ligands in the decoy set to orient the screening toward the desired pharmacological profile (Lagarde et al., 2014b). For instance, antagonists were considered as the decoy compounds set for agonists screening research, while agonists were considered as the decoy compounds set for antagonists screening research. In agreement, the corresponding agonist- and antagonist-bound structures were used for SBVS, when available. Results showed that the enrichment is better when the pharmacological profile is considered prior to screening and should therefore be systematically considered to avoid artificially bad ligands enrichment.

Maximal Unbiased Benchmarking Data Sets for HDACs (MUBD-HDACs)

So far, most of the decoy datasets [such as DUD-E (Mysinger et al., 2012) and DEKOIS (Vogel et al., 2011; Bauer et al., 2013)] or decoys generator [such as DecoyFinder (Cereto-Massagué et al., 2012) or the DUD-E generator server] are designed for SBVS purpose. Few databases [i.e., MUV (Rohrer and Baumann, 2009), NRLiSt BDB] are intended to propose benchmarking datasets for LBVS. Xia et al. thus proposed a workflow to fulfill this need, and built up decoy datasets for LBVS targeting the histone deacetylases protein family (HDACs).

Ligands Preparation

Active compounds were retrieved from the ChEMBL18 database (Gaulton et al., 2012), among molecules annotated with quantitative data (i.e., IC50), manually checked, and filtered (exclusion of salts, molecules with more than 20 rotatable bonds or with a MW of 600 or more). Finally, ligands displaying a Tanimoto coefficient greater than 0.75 based on MACCS fingerprints were removed to exclude analog molecules, and 6 physicochemical properties (MW, logP, HBAs, HBDs, RBs and net Formal Charge–nFC) were computed for all HDACs inhibitors (HDACIs).

Decoys Selection

The "All-Purchasable Molecules" subset of the ZINC database was used as the initial set of molecules before a two-step filtering:

(1) Compounds outside of the bounds of the HDACIs physicochemical properties were removed, as well as molecules with a Tanimoto coefficient ("similarity in structure" or *sims*) greater than 0.75 to any active compounds to circumvent the introduction of potential active structures (false negatives) into the decoy set.

(2) To retain only 39 decoys per HDACI, compounds were further filtered to ensure similar physicochemical properties and a random spatial distribution of the decoys around the ligands. A specific metric was assigned to each step, specifically the *simp* ("similarity in properties") and the *simsdiff* ("*sims* difference"). The *simp* is the Euclidian distance of the physicochemical properties between a target compound and a reference compound. The *simsdiff* between a potential decoy and a query ligand is the average difference between (a) the topological similarity *sims* between the potential decoy and the remaining ligands and (b) the topological similarity *sims* between the query ligands and the remaining ligands. First, a cut-off is applied on the *simp* to ensure properties similarity between ligands and decoy compounds and second, the 39 lowest *simsdiff* decoys for each ligand are selected.

Last, for each ligand, the PDB (Berman et al., 2000) structures of the targeted HDAC isoform were prepared and provided for SBVS data sets. Unlike DUD-E (Mysinger et al., 2012), only Homo sapiens 3D-data were considered.

The MUBD-HDAC datasets for HDAC2 and HDAC8 isoforms were compared to DUD-E (Mysinger et al., 2012) and DEKOIS 2.0 (Ibrahim et al., 2015a) datasets, in terms of structural diversity [Bemis-Murcko atomic frameworks (Bemis and Murcko, 1996)], property matching and ligand enrichment in SB- and LB-VS approaches. The MUBD-HDAC displayed similar to better results in terms of structural diversity and property matching and was more challenging as measured by ligand enrichment using GOLD (Jones et al., 1997) or fingerprints similarity search, in agreement with a higher structural similarity. Finally, the MUBD-HDACs sets displayed small to great improvement in terms of nearer ligands bias (i.e., ligands that are more similar structurally to a ligand than to any decoy), compared to DUD-E and DEKOIS 2.0, respectively. This bias is known to produce artificially positive LBVS evaluation outcomes (Cleves and Jain, 2008) and thus, should be minimized.

Of note, a similar work was done (Xia et al., 2014) on GPCRs using the GLL/GDD database (Gatica and Cavasotto, 2012) as ligands set, and also resulted in reduced artificial enrichment and analog bias compared to the original GLL/GDD sets.

DISCUSSION AND RECOMMENDATIONS

Ideal Benchmarking Database

The ideal VS benchmarking datasets composition should mimic real-life cases, where a small number of diverse active ligands is embedded into a much larger fraction of inactive compounds. Moreover, both sets of molecules are usually indistinguishable using simple descriptors like their physicochemical properties and share common fragments or functional chemical groups; such features should therefore be transposed to benchmarking datasets design, so that the putative inactive compounds constitute good "decoy" compounds in line with the active compounds and ensure a robust evaluation of the VS methods(Good and Oprea, 2008; Lagarde et al., 2015; Xia et al., 2015).

Comparison of Decoys Selection Methods for SBVS

Among the recent tools to help create benchmarking sets (MUV, DEKOIS, DUD-E, and MUBD), the main difference resides in the strategy used to achieve their respective objectives: the DUD-E and DEKOIS data sets are designed for evaluating SBVS methods while MUV and MUBD are conceived for benchmarking LBVS approaches. Following this basic distinction, the respective algorithms to generate decoy datasets differ significantly. In the former case, the topological dissimilarity between ligand compounds and decoy compounds is maximized to avoid inclusion of active compounds into decoy datasets. In the latter case, the proper embedding of decoy compounds into the ligands chemical space is of primary importance.

For the DUD-E, the final decoys were randomly selected from the 25% most topologically dissimilar molecules compared to the ligands to ensure unbiased selection of decoy compounds. However, several studies pointed out that bias are still present into DUD-E data sets. For instance, Chaput et al. recently evidenced that the performance of four VS programs (Glide, Gold, FlexX and Surflex) is biased (over-estimated) using the DUD-E. Good performance (as measure by BEDROC curves) could be achieved for all programs when original DUD-E datasets were used, while only Glide was considered successful when chemical library biases (i.e., datasets whose decoys and active compounds differ for nine physicochemical properties) were removed. While the DUD-E was successfully used for numerous studies, this observation clearly showed that there is still place for improvements.

Boeckler's group proposed a similar workflow in DEKOIS and DEKOIS 2.0. A physicochemical similarity over eight properties (and represented by the physicochemical similarity score PSS) is used and the topological dissimilarity between the active compounds and the future decoy compounds is computed as in the DUD-E. However, two main differences have to be noted: (1) the topological dissimilarity was computed using the more elaborated weighted LADS score rather than a 2D fingerprint based Tanimoto coefficient filter and (2) the LADS score was combined with the PSS prior to final selection of the decoys. Therefore, the final decoys selection was balanced by both parameters (physicochemical similarity and topological dissimilarity) rather than using successive arbitrary (even if widely used) thresholds, and was successfully used by Hamza et al. (2014) for drug repurposing. This balance may come at a cost, as evidenced by Xia et al.: DEKOIS datasets for HDAC2 and HDAC8 were shown to be less efficient in terms of property matching between the active compounds and the decoy compounds (Xia et al., 2015). However, the DUD-E and DEKOIS sets perform similarly in enrichment using Gold and DEKOIS perform significantly worse than DUD-E using 2D based similarity search approaches.

Comparison of Decoys Selection Methods for LBVS

Both DUD-E and DEKOIS databases share the same overall decoy selection procedure by combining topological dissimilarity and physicochemical properties similarity. While adapted to SBVS, this approach may hinder the objective evaluation of LBVS that is very sensitive to topological difference between active and decoy compounds. The MUV datasets (Rohrer and Baumann, 2009) was designed to overcome this specific weakness of the benchmarking datasets. The authors introduced the notion that decoy compounds and active compounds should be homogeneously spread in the chemical space rather than decoy compounds should be topologically dissimilar to the active compounds (as in the DUD-E for instance). The authors tested 18 datasets and claimed that MUV benchmarking datasets displayed neither analogous bias nor artificial enrichment. Furthermore, they noticed that their data sets were SBVS compliant and compared advantageously to the biased DUD sets, leading to a potential broader use of their sets. MUV sets were applied to the evaluation of VS tools (Tiikkainen et al., 2009; Abdo et al., 2010), the training of new QSAR models (Marchese Robinson et al., 2017) or molecular graph convolutions (Kearnes et al., 2016).

As highlighted by Xia et al. "MUV is restricted by the sufficient experimental decoys (chemical space of decoys)" (Xia et al., 2015). Indeed, MUV relies on the availability of experimental data and is restricted to well-studied targets. The authors subsequently proposed the Maximum Unbiased Benchmarking Data sets (MUBD, see section Benchmarking Databases) that was applied to GPCRs (Xia et al., 2014), HDACs (Xia et al., 2015; Hu et al., 2017) and Toll-like receptor 8 (Pei et al., 2015). The MUBD-DecoyMaker algorithm relies on both a minimal and required topological dissimilarity (*sims*) between decoy and active compounds, but makes use of an additional criterion that minimizes the *simsdiff* parameter, i.e., ensures that decoy and active compounds are as similar as possible.

One should note that this additional step (the decoy-actives similarity check) yield datasets also suitable for SBVS; they seemed even more challenging in SBVS (for HDAC2 and HDAC8) as they provided datasets with higher structural similarity (Xia et al., 2015). Thus, these approaches are particularly appealing as they provide benchmarking datasets that (1) are adapted to LB and SB-VS approaches, (2) subsequently allow comparative evaluations of the performance of LB and SB-VS approaches, and (3) may be more challenging for SBVS.

196

Fine-Tuned Benchmarking Datasets

The quality of an evaluation lies in the consistency between the retrospectively screened benchmarking datasets and the prospectively screened compound collections as well as the target binding site properties (Ben Nasr et al., 2013). The recent trend to publish protein family-specific datasets or user-provided active compounds dependent decoys generation tools paves the way for a valuable and systematic use of benchmarking datasets prior to prospective VS of large compound collections.

In SBVS, tuned datasets should be used to identify the protocol, conformational sampling, and/or scoring methods that induces the best enrichment in active compounds (Allen et al., 2015, 2017; Lacroix et al., 2016; Li et al., 2016; Nunes et al., 2016; Meirson et al., 2017). For instance, Allen et al. (2015, 2017) evaluated different scoring schemes using DUD-E generated decoys and successfully identified dual EFGR/BRD4 inhibitors. In LBVS, the choice of the dataset is crucial to build a reliable model that can be used to distinguish active compounds from decoy compounds. For example, Ruggeri et al. (2015) used DUD-E generated decoys to define and optimize pharmacophore models that led to the identification of 2 dual competitive inhibitors of *P. Falciparum* M1 (*Pf* A-M1) and M17 (*Pf* A-M17) aminopeptidases.

Of note, when using automatic decoy datasets generation tools, the provided active compounds should be carefully selected to avoid the previously detailed biases.

Integration of True Inactive Compounds

Despite the open-data initiatives that should ease the access to data in the near future, the low documentation about negative data (inactive and/or non-binding) is still an open issue. The inclusion of experimental data in a dataset requires great attention since (1) publicly available databases may present annotation errors that should be manually corrected (Lagarde et al., 2014a), and (2) diversity in the type of value and experimental conditions make some data barely comparable. The selection and the use of negative compounds (inactive and/or non-binding) in the evaluation/development of methods is a delicate step that strongly influences the quality of the resulting model. In agreement with Lagarde et al. (2014a) and Kaserer et al. (2015), we recommend that:

(1) Interaction data should be extracted from receptor binding or enzymatic activity assays on isolated or recombinant protein; cell-based assays should be avoided because of the many factors that can influence the outcome of the assay (non-specific binding…).
(2) Low binders or high IC50/EC50 should not be included in the active set and could be either classified as "inactive," as negative data or discarded.
(3) Experimental bias should be minimized by (a) considering the measured affinity/activity confidence based on the number of documented repeated assays and/or convergent values in different studies and (b) filtering compounds which measured activity/affinity may be an artifact caused by organic chemicals aggregation in aqueous buffers, off-targets effects, cytotoxic effects or interference with optical detection methods (auto-fluorescence and luciferase inhibition).
(4) The origin of the protein used in the assay should be considered, favoring 100% identity with the reference.
(5) Attention should be paid to the ligand binding-site, particularly for proteins that possess more than one binding site, and for multiple conformation binding sites.

One should note that the integration of inactive/non-binding compounds comes with new basics for datasets design. This case is particularly challenging since the inactive/non-binding compounds are usually extracted from the same chemical series as the active compounds. In this case, small fragments modification can induce important bioactivity loss or gain, thus, clustering active compounds to guarantee diversity and minimize analogous bias would have no meaning. Since the final objective of using such data is to harshly evaluate ability of VS methods to discriminate active from inactive compounds based on small signals, the proximity between active and inactive compounds within a chemotype should be conserved, as well as the similarity within the active compounds of a chemotype. However the over representation of a given chemotype could hinder the evaluation of VS method by masking the enrichment of low populated chemotypes. We suggest that a work should be made to equally represent chemotypes and/or to weight the resulting ROC curve (Ibrahim et al., 2015b).

CONCLUSION

Benchmarking databases are widely used to evaluate virtual screening methods. They are particularly important to compare performance of virtual screening methods and therefore to select appropriate protocol prior to large compounds collections screening, and to estimate the reliability of the results of a screening. The characterization of the weaknesses of the first published databases helped designing improved benchmarking datasets with minimized bias. The rational selection of decoy compounds is particularly important to avoid artificial enrichment in the evaluation of the different methods. The diversification of public datasets gathering both active and decoy compounds for a given protein family, and the publication of online decoys generation tools contributed to the democratization of the use of benchmarking studies to help identifying protocols adapted for the query/target system under study. Nowadays, experimental data are being integrated in the decoy compounds set to look for a specific activity or to identify methods fitted for highly similar binders/non binders discrimination. Experimentally validated decoys selection requires careful attention to minimize experimental biases that may arise.

AUTHOR CONTRIBUTIONS

All authors listed have made substantial, direct and intellectual contribution to the work, and approved it for publication.

REFERENCES

Abdo, A., Chen, B., Mueller, C., Salim, N., and Willett, P. (2010). Ligand-based virtual screening using Bayesian networks. *J. Chem. Inf. Model.* 50, 1012–1020. doi: 10.1021/ci100090p

Allen, B. K., Mehta, S., Ember, S. W., Schonbrunn, E., Ayad, N., and Schürer, S. C. (2015). Large-scale computational screening identifies first in class multitarget inhibitor of EGFR kinase and BRD4. *Sci. Rep.* 5:srep16924. doi: 10.1038/srep16924

Allen, B. K., Mehta, S., Ember, S. W. J., Zhu, J.-Y., Schönbrunn, E., Ayad, N. G., et al. (2017). Identification of a novel class of BRD4 inhibitors by computational screening and binding simulations. *ACS Omega* 2, 4760–4771. doi: 10.1021/acsomega.7b00553

Auld, D. S., Southall, N. T., Jadhav, A., Johnson, R. L., Diller, D. J., Simeonov, A., et al. (2008). Characterization of chemical libraries for luciferase inhibitory activity. *J. Med. Chem.* 51, 2372–2386. doi: 10.1021/jm701302v

Bauer, M. R., Ibrahim, T. M., Vogel, S. M., and Boeckler, F. M. (2013). Evaluation and optimization of virtual screening workflows with DEKOIS 2.0 – a public library of challenging docking benchmark sets. *J. Chem. Inf. Model.* 53, 1447–1462. doi: 10.1021/ci400115b

Bemis, G. W., and Murcko, M. A. (1996). The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* 39, 2887–2893. doi: 10.1021/jm9602928

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

Benson, M. L., Smith, R. D., Khazanov, N. A., Dimcheff, B., Beaver, J., Dresslar, P., et al. (2008). Binding MOAD, a high-quality protein-ligand database. *Nucleic Acids Res.* 36, D674–D678. doi: 10.1093/nar/gkm911

Bento, A. P., Gaulton, A., Hersey, A., Bellis, L. J., Chambers, J., Davies, M., et al. (2014). The ChEMBL bioactivity database: an update. *Nucleic Acids Res.* 42, D1083–D1090. doi: 10.1093/nar/gkt1031

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., et al. (2000). The protein data bank. *Nucleic Acids Res.* 28, 235–242. doi: 10.1093/nar/28.1.235

Bissantz, C., Bernard, P., Hibert, M., and Rognan, D. (2003). Protein-based virtual screening of chemical databases. II. Are homology models of G-Protein Coupled Receptors suitable targets? *Proteins* 50, 5–25. doi: 10.1002/prot.10237

Bissantz, C., Folkers, G., and Rognan, D. (2000). Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. *J. Med. Chem.* 43, 4759–4767. doi: 10.1021/jm001044l

Block, P., Sotriffer, C. A., Dramburg, I., and Klebe, G. (2006). AffinDB: a freely accessible database of affinities for protein–ligand complexes from the PDB. *Nucleic Acids Res.* 34, D522–D526. doi: 10.1093/nar/gkj039

Braga, R. C., and Andrade, C. H. (2013). Assessing the performance of 3D pharmacophore models in virtual screening: how good are they? *Curr. Top. Med. Chem.* 13, 1127–1138. doi: 10.2174/1568026611313090010

Brozell, S. R., Mukherjee, S., Balius, T. E., Roe, D. R., Case, D. A., and Rizzo, R. C. (2012). Evaluation of DOCK 6 as a pose generation and database enrichment tool. *J. Comput. Aided Mol. Des.* 26, 749–773. doi: 10.1007/s10822-012-9565-y

Cereto-Massagué, A., Guasch, L., Valls, C., Mulero, M., Pujadas, G., and Garcia-Vallvé, S. (2012). DecoyFinder: an easy-to-use python GUI application for building target-specific decoy sets. *Bioinformatics* 28, 1661–1662. doi: 10.1093/bioinformatics/bts249

Clark, R. D., and Webster-Clark, D. J. (2008). Managing bias in ROC curves. *J. Comput. Aided Mol. Des.* 22, 141–146. doi: 10.1007/s10822-008-9181-z

Cleves, A. E., and Jain, A. N. (2008). Effects of inductive bias on computational evaluations of ligand-based modeling and on drug discovery. *J. Comput. Aided Mol. Des.* 22, 147–159. doi: 10.1007/s10822-007-9150-y

Cummings, M. D., DesJarlais, R. L., Gibbs, A. C., Mohan, V., and Jaeger, E. P. (2005). Comparison of automated docking programs as virtual screening tools. *J. Med. Chem.* 48, 962–976. doi: 10.1021/jm049798d

Diller, D. J., and Li, R. (2003). Kinases, homology models, and high throughput docking. *J. Med. Chem.* 46, 4638–4647. doi: 10.1021/jm020503a

Eldridge, M. D., Murray, C. W., Auton, T. R., Paolini, G. V., and Mee, R. P. (1997). 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.* 11, 425–445. doi: 10.1023/A:1007996124545

Empereur-mot, C., Guillemain, H., Latouche, A., Zagury, J.-F., Viallon, V., and Montes, M. (2015). Predictiveness curves in virtual screening. *J. Cheminformatics* 7:52. doi: 10.1186/s13321-015-0100-8

Gatica, E. A., and Cavasotto, C. N. (2012). Ligand and decoy sets for docking to G protein-coupled receptors. *J. Chem. Inf. Model.* 52, 1–6. doi: 10.1021/ci200412p

Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., et al. (2012). ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* 40, D1100–D1107. doi: 10.1093/nar/gkr777

Good, A. C., and Oprea, T. I. (2008). Optimization of CAMD techniques 3. Virtual screening enrichment studies: a help or hindrance in tool selection? *J. Comput. Aided Mol. Des.* 22, 169–178. doi: 10.1007/s10822-007-9167-2

Guasch, L., Sala, E., Castell-Auví, A., Cedó, L., Liedl, K. R., Wolber, G., et al. (2012). Identification of PPARgamma partial agonists of natural origin (I): development of a virtual screening procedure and *in vitro* validation. *PLoS ONE* 7:e50816. doi: 10.1371/journal.pone.0050816

Hamza, A., Wagner, J. M., Wei, N.-N., Kwiatkowski, S., Zhan, C.-G., Watt, D. S., et al. (2014). Application of the 4D fingerprint method with a robust scoring function for scaffold-hopping and drug repurposing strategies. *J. Chem. Inf. Model.* 54, 2834–2845. doi: 10.1021/ci5003872

Hawkins, P. C., Warren, G. L., Skillman, A. G., and Nicholls, A. (2008). How to do an evaluation: pitfalls and traps. *J. Comput. Aided Mol. Des.* 22, 179–190. doi: 10.1007/s10822-007-9166-3

Hu, H., Xia, J., Wang, D., Wang, X. S., and Wu, S. (2017). A thoroughly validated virtual screening strategy for discovery of novel HDAC3 inhibitors. *Int. J. Mol. Sci.* 18, 137. doi: 10.3390/ijms18010137

Huang, N., Shoichet, B. K., and Irwin, J. J. (2006). Benchmarking sets for molecular docking. *J. Med. Chem.* 49, 6789–6801. doi: 10.1021/jm0608356

Ibrahim, T. M., Bauer, M. R., and Boeckler, F. M. (2015a). Applying DEKOIS 2.0 in structure-based virtual screening to probe the impact of preparation procedures and score normalization. *J. Cheminformatics* 7:21. doi: 10.1186/s13321-015-0074-6

Ibrahim, T. M., Bauer, M. R., Dörr, A., Veyisoglu, E., and Boeckler, F. M. (2015b). pROC-Chemotype plots enhance the interpretability of benchmarking results in structure-based virtual screening. *J. Chem. Inf. Model.* 55, 2297–2307. doi: 10.1021/acs.jcim.5b00475

Irwin, J. J. (2008). Community benchmarks for virtual screening. *J. Comput. Aided Mol. Des.* 22, 193–199. doi: 10.1007/s10822-008-9189-4

Irwin, J. J., and Shoichet, B. K. (2005). ZINC–a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* 45, 177–182. doi: 10.1021/ci049714+

Jain, A. N., and Nicholls, A. (2008). Recommendations for evaluation of computational methods. *J. Comput. Aided Mol. Des.* 22, 133–139. doi: 10.1007/s10822-008-9196-5

Jones, G., Willett, P., Glen, R. C., Leach, A. R., and Taylor, R. (1997). Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* 267, 727–748. doi: 10.1006/jmbi.1996.0897

Kaserer, T., Beck, K. R., Akram, M., Odermatt, A., and Schuster, D. (2015). Pharmacophore models and pharmacophore-based virtual screening: concepts and applications exemplified on hydroxysteroid dehydrogenases. *Mol. Basel Switz.* 20, 22799–22832. doi: 10.3390/molecules201219880

Kavlock, R., Chandler, K., Houck, K., Hunter, S., Judson, R., Kleinstreuer, N., et al. (2012). Update on EPA's ToxCast program: providing high throughput decision support tools for chemical risk management. *Chem. Res. Toxicol.* 25, 1287–1302. doi: 10.1021/tx3000939

Kearnes, S., McCloskey, K., Berndl, M., Pande, V., and Riley, P. (2016). Molecular graph convolutions: moving beyond fingerprints. *J. Comput. Aided Mol. Des.* 30, 595–608. doi: 10.1007/s10822-016-9938-8

Kellenberger, E., Rodrigo, J., Muller, P., and Rognan, D. (2004). Comparative evaluation of eight docking tools for docking and virtual screening accuracy. *Proteins* 57, 225–242. doi: 10.1002/prot.20149

Kuntz, I. D., Blaney, J. M., Oatley, S. J., Langridge, R., and Ferrin, T. E. (1982). A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.* 161, 269–288. doi: 10.1016/0022-2836(82)90153-X

Lacroix, C., Fish, I., Torosyan, H., Parathaman, P., Irwin, J. J., Shoichet, B. K., et al. (2016). Identification of novel smoothened ligands using structure-based docking. *PLoS ONE* 11:e0160365. doi: 10.1371/journal.pone.0160365

Lagarde, N., Ben Nasr, N., Jérémie, A., Guillemain, H., Laville, V., Labib, T., et al. (2014a). NRLiSt BDB, the manually curated nuclear receptors

ligands and structures benchmarking database. *J. Med. Chem.* 57, 3117–3125. doi: 10.1021/jm500132p

Lagarde, N., Delahaye, S., Jérémie, A., Ben Nasr, N., Guillemain, H., Empereur-Mot, C., et al. (2017). Discriminating agonist from antagonist ligands of the nuclear receptors using different chemoinformatics approaches. *Mol. Inform.* 36:1700020. doi: 10.1002/minf.201700020

Lagarde, N., Delahaye, S., Zagury, J.-F., and Montes, M. (2016). Discriminating agonist and antagonist ligands of the nuclear receptors using 3D-pharmacophores. *J. Cheminformatics* 8:43. doi: 10.1186/s13321-016-0154-2

Lagarde, N., Zagury, J.-F., and Montes, M. (2014b). Importance of the pharmacological profile of the bound ligand in enrichment on nuclear receptors: toward the use of experimentally validated decoy ligands. *J. Chem. Inf. Model.* 54, 2915–2944. doi: 10.1021/ci500305c

Lagarde, N., Zagury, J.-F., and Montes, M. (2015). Benchmarking data sets for the evaluation of virtual ligand screening methods: review and perspectives. *J. Chem. Inf. Model.* 55, 1297–1307. doi: 10.1021/acs.jcim.5b00090

Li, J., Wang, H., Li, J., Bao, J., and Wu, C. (2016). Discovery of a potential HER2 inhibitor from natural products for the treatment of HER2-positive breast cancer. *Int. J. Mol. Sci.* 17:1055. doi: 10.3390/ijms17071055

Lionta, E., Spyrou, G., Vassilatis, D. K., and Cournia, Z. (2014). Structure-based virtual screening for drug discovery: principles, applications and recent advances. *Curr. Top. Med. Chem.* 14, 1923–1938. doi: 10.2174/1568026614666140929124445

Lorber, D. M., and Shoichet, B. K. (2005). hierarchical docking of databases of multiple ligand conformations. *Curr. Top. Med. Chem.* 5, 739–749. doi: 10.2174/1568026054637683

Löwer, M., Geppert, T., Schneider, P., Hoy, B., Wessler, S., and Schneider, G. (2011). Inhibitors of helicobacter pylori protease HtrA found by 'virtual ligand' screening combat bacterial invasion of epithelia. *PLoS ONE* 6:e17986. doi: 10.1371/journal.pone.0017986

Marchese Robinson, R. L., Palczewska, A., Palczewski, J., and Kidley, N. (2017). Comparison of the predictive performance and interpretability of random forest and linear models on benchmark data sets. *J. Chem. Inf. Model.* 57, 1773–1792. doi: 10.1021/acs.jcim.6b00753

May, A., and Zacharias, M. (2005). Accounting for global protein deformability during protein-protein and protein-ligand docking. *Biochim. Biophys. Acta* 1754, 225–231. doi: 10.1016/j.bbapap.2005.07.045

McGaughey, G. B., Sheridan, R. P., Bayly, C. I., Culberson, J. C., Kreatsoulas, C., Lindsley, S., et al. (2007). Comparison of topological, shape, and docking methods in virtual screening. *J. Chem. Inf. Model.* 47, 1504–1519. doi: 10.1021/ci700052x

McGovern, S. L., and Shoichet, B. K. (2003). Information decay in molecular docking screens against holo, apo, and modeled conformations of enzymes. *J. Med. Chem.* 46, 2895–2907. doi: 10.1021/jm0300330

Meirson, T., Samson, A. O., and Gil-Henn, H. (2017). An *in silico* high-throughput screen identifies potential selective inhibitors for the non-receptor tyrosine kinase Pyk2. *Drug Des. Devel. Ther.* 11, 1535–1557. doi: 10.2147/DDDT.S136150

Meng, E. C., Shoichet, B. K., and Kuntz, I. D. (1992). Automated docking with grid-based energy evaluation. *J. Comput. Chem.* 13, 505–524. doi: 10.1002/jcc.540130412

Meyer, K. (2007). WOMBAT—A tool for mixed model analyses in quantitative genetics by restricted maximum likelihood (REML). *J. Zhejiang Univ. Sci. B* 8, 815–821. doi: 10.1631/jzus.2007.B0815

Muegge, I., and Martin, Y. C. (1999). A general and fast scoring function for protein-ligand interactions: a simplified potential approach. *J. Med. Chem.* 42, 791–804. doi: 10.1021/jm980536j

Munk, C., Isberg, V., Mordalski, S., Harpsøe, K., Rataj, K., Hauser, A. S., et al. (2016). GPCRdb: the G protein-coupled receptor database – an introduction. *Br. J. Pharmacol.* 173, 2195–2207. doi: 10.1111/bph.13509

Mysinger, M. M., Carchia, M., Irwin, J. J., and Shoichet, B. K. (2012). Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *J. Med. Chem.* 55, 6582–6594. doi: 10.1021/jm300687e

Mysinger, M. M., and Shoichet, B. K. (2010). Rapid context-dependent ligand desolvation in molecular docking. *J. Chem. Inf. Model.* 50, 1561–1573. doi: 10.1021/ci100214a

Neves, M. A., Totrov, M., and Abagyan, R. (2012). Docking and scoring with ICM: the benchmarking results and strategies for improvement. *J. Comput. Aided Mol. Des.* 26, 675–686. doi: 10.1007/s10822-012-9547-0

Nunes, R. R., Costa, M. D., Santos, B. D., Fonseca, A. L., Ferreira, L. S., Chagas, R. C., et al. (2016). Successful application of virtual screening and molecular dynamics simulations against antimalarial molecular targets. *Mem. Inst. Oswaldo Cruz* 111, 721–730. doi: 10.1590/0074-02760160207

Okuno, Y., Yang, J., Taneishi, K., Yabuuchi, H., and Tsujimoto, G. (2006). GLIDA: GPCR-ligand database for chemical genomic drug discovery. *Nucleic Acids Res.* 34, D673–D677. doi: 10.1093/nar/gkj028

Pei, F., Jin, H., Zhou, X., Xia, J., Sun, L., Liu, Z., et al. (2015). Enrichment assessment of multiple virtual screening strategies for Toll-like receptor 8 agonists based on a maximal unbiased benchmarking data set. *Chem. Biol. Drug Des.* 86, 1226–1241. doi: 10.1111/cbdd.12590

Placzek, S., Schomburg, I., Chang, A., Jeske, L., Ulbrich, M., Tillack, J., et al. (2017). BRENDA in 2017: new perspectives and new tools in BRENDA. *Nucleic Acids Res.* 45, D380–D388. doi: 10.1093/nar/gkw952

Rarey, M., Kramer, B., Lengauer, T., and Klebe, G. (1996). A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* 261, 470–489. doi: 10.1006/jmbi.1996.0477

Repasky, M. P., Murphy, R. B., Banks, J. L., Greenwood, J. R., Tubert-Brohman, I., Bhat, S., et al. (2012). Docking performance of the glide program as evaluated on the Astex and DUD datasets: a complete set of glide SP results and selected results for a new scoring function integrating WaterMap and glide. *J. Comput. Aided Mol. Des.* 26, 787–799. doi: 10.1007/s10822-012-9575-9

Rognan, D., Lauemoller, S. L., Holm, A., Buus, S., and Tschinke, V. (1999). Predicting binding affinities of protein ligands from three-dimensional models: application to peptide binding to class I major histocompatibility proteins. *J. Med. Chem.* 42, 4650–4658. doi: 10.1021/jm9910775

Rohrer, S. G., and Baumann, K. (2009). Maximum unbiased validation (MUV) data sets for virtual screening based on PubChem bioactivity data. *J. Chem. Inf. Model.* 49, 169–184. doi: 10.1021/ci8002649

Roth, B. L., Lopez, E., Patel, S., and Kroeze, W. K. (2000). The multiplicity of serotonin receptors: Uselessly diverse molecules or an embarrassment of riches? *Neuroscientist* 6, 252–262. doi: 10.1177/107385840000600408

Ruggeri, C., Drinkwater, N., Sivaraman, K. K., Bamert, R. S., McGowan, S., and Paiardini, A. (2015). Identification and validation of a potent dual inhibitor of the *P. falciparum* M1 and M17 aminopeptidases using virtual screening. *PLoS ONE* 10:e0138957. doi: 10.1371/journal.pone.0138957

Simeonov, A., Jadhav, A., Thomas, C. J., Wang, Y., Huang, R., Southall, N. T., et al. (2008). Fluorescence spectroscopic profiling of compound libraries. *J. Med. Chem.* 51, 2363–2371. doi: 10.1021/jm701301m

Sliwoski, G., Kothiwale, S., Meiler, J., and Lowe, E. W. (2014). Computational methods in drug discovery. *Pharmacol. Rev.* 66, 334–395. doi: 10.1124/pr.112.007336

Southan, C., Sharman, J. L., Benson, H. E., Faccenda, E., Pawson, A. J., Alexander, S. P., et al. (2016). The IUPHAR/BPS guide to PHARMACOLOGY in 2016: towards curated quantitative interactions between 1300 protein targets and 6000 ligands. *Nucleic Acids Res.* 44, D1054–D1068. doi: 10.1093/nar/gkv1037

Spitzer, R., and Jain, A. N. (2012). Surflex-dock: docking benchmarks and real-world application. *J. Comput. Aided Mol. Des.* 26, 687–699. doi: 10.1007/s10822-011-9533-y

Stumpfe, D., and Bajorath, J. (2011). "Applied virtual screening: strategies, recommendations, and caveats," in *Virtual Screening: Principles, Challenges, and Practical Guidelines*, ed C. Sotriffer (Weinheim: Wiley-VCH Verlag GmbH and Co. KGaA), 291–318. doi: 10.1002/9783527633326.ch11

Tanrikulu, Y., Krüger, B., and Proschak, E. (2013). The holistic integration of virtual screening in drug discovery. *Drug Discov. Today* 18, 358–364. doi: 10.1016/j.drudis.2013.01.007

Tiikkainen, P., Markt, P., Wolber, G., Kirchmair, J., Distinto, S., Poso, A., et al. (2009). Critical comparison of virtual screening methods against the MUV data set. *J. Chem. Inf. Model.* 49, 2168–2178. doi: 10.1021/ci900249b

Triballeau, N., Acher, F., Brabet, I., Pin, J.-P., and Bertrand, H.-O. (2005). Virtual screening workflow development guided by the "receiver operating characteristic" curve approach. Application to high-throughput docking on metabotropic glutamate receptor subtype 4. *J. Med. Chem.* 48, 2534–2547. doi: 10.1021/jm049092j

Verdonk, M. L., Berdini, V., Hartshorn, M. J., Mooij, W. T., Murray, C. W., Taylor, R. D., et al. (2004). Virtual screening using protein-ligand docking: avoiding artificial enrichment. *J. Chem. Inf. Comput. Sci.* 44, 793–806. doi: 10.1021/ci034289q

Vilar, S., Karpiak, J., and Costanzi, S. (2010). Ligand and structure-based models for the prediction of ligand-receptor affinities and virtual screenings: development and application to the β2-adrenergic receptor. *J. Comput. Chem.* 31, 707–720. doi: 10.1002/jcc.21346

Vogel, S. M., Bauer, M. R., and Boeckler, F. M. (2011). DEKOIS: Demanding evaluation kits for objective *in silico* screening—a versatile tool for benchmarking docking programs and scoring functions. *J. Chem. Inf. Model.* 51, 2650–2665. doi: 10.1021/ci2001549

von Korff, M., Freyss, J., and Sander, T. (2009). Comparison of ligand- and structure-based virtual screening on the DUD data set. *J. Chem. Inf. Model.* 49, 209–231. doi: 10.1021/ci800303k

Wallach, I., and Lilien, R. (2011). Virtual decoy sets for molecular docking benchmarks. *J. Chem. Inf. Model.* 51, 196–202. doi: 10.1021/ci100374f

Wang, R., Fang, X., Lu, Y., and Wang, S. (2004). The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *J. Med. Chem.* 47, 2977–2980. doi: 10.1021/jm030580l

Wang, R., Fang, X., Lu, Y., Yang, C.-Y., and Wang, S. (2005). The PDBbind database: methodologies and updates. *J. Med. Chem.* 48, 4111–4119. doi: 10.1021/jm048957q

Wang, R., Liu, L., Lai, L., and Tang, Y. (1998). SCORE: a new empirical method for estimating the binding affinity of a protein-ligand complex. *Mol. Model. Annu.* 4, 379–394. doi: 10.1007/s008940050096

Wang, Y., Bryant, S. H., Cheng, T., Wang, J., Gindulyte, A., Shoemaker, B. A., et al. (2017). PubChem BioAssay: 2017 update. *Nucleic Acids Res.* 45, D955–D963. doi: 10.1093/nar/gkw1118

Warren, G. L., Andrews, C. W., Capelli, A.-M., Clarke, B., LaLonde, J., Lambert, M. H., et al. (2006). A critical assessment of docking programs and scoring functions. *J. Med. Chem.* 49, 5912–5931. doi: 10.1021/jm050362n

Wei, B. Q., Baase, W. A., Weaver, L. H., Matthews, B. W., and Shoichet, B. K. (2002). A model binding site for testing scoring functions in molecular docking. *J. Mol. Biol.* 322, 339–355. doi: 10.1016/S0022-2836(02)00777-5

Wishart, D. S., Knox, C., Guo, A. C., Cheng, D., Shrivastava, S., Tzur, D., et al. (2008). DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.* 36, D901–D906. doi: 10.1093/nar/gkm958

Xia, J., Jin, H., Liu, Z., Zhang, L., and Wang, X. S. (2014). An unbiased method to build benchmarking sets for ligand-based virtual screening and its application to GPCRs. *J. Chem. Inf. Model.* 54, 1433–1450. doi: 10.1021/ci500062f

Xia, J., Tilahun, E. L., Kebede, E. H., Reid, T.-E., Zhang, L., and Wang, X. S. (2015). Comparative modeling and benchmarking data sets for human histone deacetylases and sirtuin families. *J. Chem. Inf. Model.* 55, 374–388. doi: 10.1021/ci5005515

Conflict of Interest Statement: 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.

1.3 Conclusion et perspectives

Les banques de données d'évaluation ont évolué selon 2 axes principaux : 1) l'intégration de *decoys* rationnels et l'intégration de molécules expérimentalement validées comme inactives, et 2) la multiplication des cibles couvertes par les banques de référence et la possibilité de générer des *decoys* adaptés à un jeu de molécules actives fournis par l'utilisateur. Par exemple, la DUD-E possède des jeux de données pour 102 protéines issues de 8 familles protéiques différentes et propose un outil de génération de *decoys* à partir de molécules actives fournies par l'utilisateur. La DUD-E et la MUV intègrent des molécules inactives extraites respectivement de la ChEMBL[79] et de données de HTS issus de la PCBioassay[80] en complément des *decoys*. La NRLiSt BDB offre une possibilité nouvelle dans l'évaluation de méthodes de criblage : l'utilisation de ligands présentant un profil pharmacologique non désirés comme molécules inactives. Lorsqu'une molécule agoniste est recherchée, les molécules antagonistes peuvent être introduites dans le jeu de données inactives. Des études de la NRLiSt BDB ont montré l'importance de l'annotation des profils pharmacologiques des ligands des récepteurs nucléaires pour conduire des études robustes de criblages basés sur la structure ou le ligand. En effet, les antagonistes et les agonistes présentent des modes de liaisons différents qui impliquent des changements structuraux du récepteur nucléaire ciblé. Les molécules antagonistes ne peuvent donc pas servir pour construire un modèle permettant d'identifier des molécules agonistes, tout comme les structures antagonistes-liées ne peuvent pas être utilisées pour l'identification de molécules agonistes.

La sélection des *decoys* a donc évolué au cours du temps vers une sélection rationnelle et vise à être partiellement substituée par l'intégration de molécules inactives (en termes d'affinité et/ ou d'activité) qui apportent des informations complémentaires.

2 Construction d'une banque de données incluant des données d'inactivité : Nuclear Receptors Database Including Negative Data (NR-DBIND)

2.1 Introduction

Les récepteurs nucléaires sont une famille de protéines impliquées dans de nombreux processus physiologiques humains (Cf 6.1). Leur implication dans des maladies et leur modulation non désirée par des petites molécules expliquent le large intérêt porté par la communauté scientifique aux récepteurs nucléaires. L'objectif de notre laboratoire étant d'étudier l'importance de l'intégration de données d'inactivité dans les modèles de criblage vitruel, nous avons choisi de concentrer notre étude sur cette famille de protéine. Les récepteurs nucléaires ayant été l'objet de multiples études, de nombreuses données sont disponibles dans les banques de données publiques comme la ChEMBL qui contient des données d'inactivité. Une étude de 2014 a cependant révélé qu'environ 30% des entrées associées aux récepteurs nucléaires dans la ChEMBL contiennent des erreurs[5]. Nous avons donc construit la Nuclear Receptor DataBase Including Negative Data (NR-DBIND) pour collecter des données d'interaction fiables incluant des données d'inactivité recensées dans la littérature. La ChEMBL a été utilisée pour recenser les publications mentionnant des interactions entre une petite molécule et un récepteur nucléaire, et chaque information a été collectée et corrigée manuellement depuis la publication d'origine de sorte à s'assurer de la qualité des données. Une interaction a été inclue dans la NR-DBIND dès lors qu'une affinité était mentionnée dans la littérature. Les molécules pour lesquelles une affinité faible ou nulle pour la cible étudiée est mentionnée ont également été inclues dans la NR-DBIND. Lorsque des données d'activité étaient disponibles, elles ont été ajoutées aux données d'affinité de sorte à fournir une annotation des petites molécules selon leur profil pharmacologique (ex : agoniste, antagoniste, etc.). La NR-DBIND possède également un volet « protéine » dans lequel sont listées des structures PDB *holo* et *apo* de 28 récepteurs nucléaires ne possédant pas de mutation. Les structures *holo* sont annotées en fonction du profil pharmacologique du ligand associé. La NR-DBIND est disponible gratuitement à l'ensemble de la communauté scientifique (http://nr-dbind.drugdesign.fr).

2.2 Article

Journal of Medicinal Chemistry

Perspective

Cite This: J. Med. Chem. 2019, 62, 2894−2904

pubs.acs.org/jmc

Nuclear Receptors Database Including Negative Data (NR-DBIND): A Database Dedicated to Nuclear Receptors Binding Data Including Negative Data and Pharmacological Profile

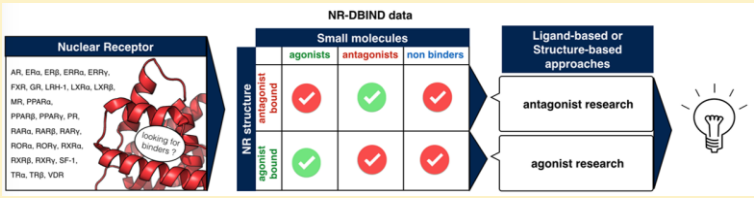
Miniperspective

Manon Réau,[†] Nathalie Lagarde,[†,‡] Jean-François Zagury,[†] and Matthieu Montes*[,†]

[†]Laboratoire GBA, EA4627, Conservatoire National des Arts et Métiers, 2 Rue Conté, 75003 Paris, France

[‡]Université Paris Diderot, Sorbonne Paris Cité, Molécules Thérapeutiques in Silico, INSERM UMR-S 973, 75205 Paris, France

Supporting Information



ABSTRACT: Nuclear receptors (NRs) are transcription factors that regulate gene expression in various physiological processes through their interactions with small hydrophobic molecules. They constitute an important class of targets for drugs and endocrine disruptors and are widely studied for both health and environment concerns. Since the integration of negative data can be critical for accurate modeling of ligand activity profiles, we manually collected and annotated NRs interaction data (positive and negative) through a sharp review of the corresponding literature. 15 116 positive and negative interactions data are provided for 28 NRs together with 593 PDB structures in the freely available Nuclear Receptors Database Including Negative Data (http://nr-dbind.drugdesign.fr). The NR-DBIND contains the most extensive information about interaction data on NRs, which should bring valuable information to chemists, biologists, pharmacologists and toxicologists.

INTRODUCTION

Nuclear receptors (NRs) are transcription factors that regulate the expression of genes involved in key physiological processes such as reproduction, development, and metabolism[1] through their interactions with hydrophobic molecules. They have been proven to be key therapeutic targets[2,3] that "may be important on−off switches for specific physiological states".[1]

Most of the 48 identified human members of the NR family (Table 1) are regulated by endogenous ligands including steroids, vitamins, bile acids, fatty acids, and hormones.[4] The remaining ones have not yet been associated with any ligands and are classified as orphan receptors.[5]

NRs use distinct strategies in achieving their complex control of gene regulation:[6] in the absence of an agonist ligand, (1) type I/III NRs (e.g., androgen receptor (AR), estrogen receptors (ERs), glucocorticoid receptor (GR), mineralocorticoid receptor (MR), and progesterone receptor (PR)) are maintained inactive by interacting with chaperone proteins in the cytosol, while (2) type II NRs (e.g., thyroid hormone receptors (TRs), retinoic acid receptors (RARs)) are located in the nucleus where they are bound to DNA response elements along with co-repressors.

Upon binding of an agonist ligand, NRs ligand binding domains (LBDs) are subjected to conformational changes that induce dissociation of either (1) the chaperone protein, which leads to transduction into the nucleus and recruitment of coactivators for type I/III NRs, or (2) the co-repressors, to promote coactivator recruitment for type II NRs.

Depending on the pharmacological profile of the bound ligand, different types of activities are elicited. Antagonist compounds block NRs in their inactive conformation by impeding the recruitment of coactivators and favoring interactions with co-repressors; inverse agonist compounds trigger the opposite physiological response compared to agonist compound binding; i.e., they inhibit NRs basal unbound transcriptional activity; partial agonist or partial antagonist compounds are associated with an incomplete physiological response.

Many epidemiological and experimental studies have highlighted that NRs are also targeted by endocrine disrupting chemicals (EDCs) that mimic and replace the endogenous ligands.[7−12] Inappropriate exposures to EDCs can impair key physiological functions and lead to dramatic impact on human and wildlife health.[13]

A major focus in the current studies of NRs is to identify selective modulators for specific NRs[14−18] and to evaluate EDCs' potential.[19−22] In this context, it is important to

Received: July 13, 2018
Published: October 24, 2018

Table 1. Summary of the 48 Human NRs[a]

	abbreviation	name	NRNC symbol	UniProt	representative ligand	*holo* structures	*apo* structures
Thyroid hormone receptor	**TRα**	**Thyroid hormone receptor α**	**NR1A1**	P10827	**thyroid hormones**	**4**	**0**
	TRβ	**Thyroid hormone receptor β**	**NR1A2**	P10828		**13**	**0**
Retinoic acid receptor	RARα	Retinoic acid receptor α	NR1B1	P10276	vitamin A	5	0
	RARβ	Retinoic acid receptor β	NR1B2	P10826		3	0
	RARγ	Retinoic acid receptor γ	NR1B3	P13631		9	0
Peroxisome proliferator-activated receptor	**PPARα**	**Peroxisome proliferator-activated receptor α**	**NR1C1**	Q07869	**fatty acids, prostaglandins**	**17**	**0**
	PPARβ/δ	**Peroxisome proliferator-activated receptor-β/δ**	**NR1C2**	Q03181		**34**	**1**
	PPARγ	**Peroxisome proliferator-activated receptor γ**	**NR1C3**	P37231		**131**	**4**
Rev-ErbA	Rev-ErbAα	Rev-ErbAα	NR1D1	P20393	heme		
	Rev-ErbAβ	Rev-ErbAα	NR1D2	Q14995			
RAR-related orphan receptor	**RORα**	**RAR-related orphan receptor α**	**NR1F1**	P35398	**cholesterol ATRA**	**3**	**0**
	RORβ	**RAR-related orphan receptor β**	**NR1F2**	Q92753			
	RORγ	**RAR-related orphan receptor γ**	**NR1F3**	P51449		**41**	**2**
Liver X receptor-like	LXRα	Liver X receptor α	NR1H3	Q13133	oxysterols	7	0
	LXRβ	Liver X receptor β	NR1H2	P55055		12	0
	FXR	Farnesoid X receptor	NR1H4	Q96RI1		49	0
Vitamin D receptor-like	**VDR**	**Vitamin D receptor**	**NR1I1**	P11473	**vitamin D**	**23**	**0**
	PXR	**Pregnane X receptor**	**NR1I2**	O75469	**xenobiotics**	**15**	**4**
	CAR	**Constitutive androstane receptor**	**NR113**	Q14994	**androstane**		
Hepatocyte nuclear factor 4	HNF4α	Hepatocyte nuclear factor 4α	NR2A1	P41235	fatty acids		
	HNF4γ	Hepatocyte nuclear factor 4γ	NR2A2	Q14541			
Retinoid X receptor	**RXRα**	**Retinoid X receptor α**	**NR2B1**	P19793	**retinoids**	**54**	**5**
	RXRβ	**Retinoid X receptor β**	**NR2B2**	P28702		**2**	**0**
	RXRγ	**Retinoid X receptor γ**	**NR2B3**	P48443		**0**	**0**
Testicular receptor	TR2	Testicular receptor 2	NR2C1	P13056			
	TR4	Testicular receptor 4	NR2C2	P49116			
TLX/PNR	**TLX**	**Homolog of the Drosophila tailless gene**	**NR2E1**	Q9Y466			
	PNR	**Photoreceptor cell-specific nuclear receptor**	**NR2E3**	Q9Y5X4			
COUP/EAR	COUP-TFI	Chicken ovalbumin upstream promoter-transcription factor I	NR2F1	P10589			
	COUP-TFII	Chicken ovalbumin upstream promoter-transcription factor II	NR2F2	P24468			
	EAR-2	V-erbA-related	NR2F6	P10588			
Estrogen receptor	**ERα**	**Estrogen receptor α**	**NR3A1**	P03372	**estrogens**	**32**	**0**
	ERβ	**Estrogen receptor β**	**NR3A2**	Q92731		**28**	**0**
Estrogen related receptor	ERRα	Estrogen-related receptor α	NR3B1	P11474		1	1
	ERRβ	Estrogen-related receptor β	NR3B2	O95718			
	ERRγ	Estrogen-related receptor γ	NR3B3	P66258		10	5
3-Ketosteroid receptor	**GR**	**Glucocorticoid receptor**	**NR3C1**	P04150	**cortisol**	**0**	**0**
	MR	**Mineralocorticoid receptor**	**NR3C2**	P08235	**aldosterone**	**3**	**0**
	PR	**Progesterone receptor**	**NR3C3**	P06401	**progesterone**	**18**	**0**
	AR	**Androgen receptor**	**NR3C4**	P10275	**testosterone**	**44**	**0**
NGFIB/NURR1/NOR1	NGFIB	Nerve growth factor IB	NR4A1	P22736			
	NURR1	Nuclear receptor related 1	NR4A2	P43354			
	NOR1	Neuron-derived orphan receptor 1	NR4A3	Q92570			
SF1/LRH1	**SF1**	**Steroidogenic factor 1**	**NR5A1**	Q13285	**phosphatidylinositols**	**1**	**0**
	LRH-1	**Liver receptor homolog 1**	**NR5A2**	O00482		**9**	**3**
GCNF	GCNF	Germ cell nuclear factor	NR6A1	Q15406			
DAX/SHP	**DAX1**	**Dosage-sensitive sex reversal, adrenal hypoplasia critical region, on chromosome X, gene 1**	**NR0B1**	P51843			
	SHP	**Small heterodimer partner**	**NR0B2**	Q15466			

[a]The NRs described in the NR-DBIND are highlighted in bold, and the number of PDB structures included in the NR-DBIND is indicated.

understand and predict how molecules selectively bind NRs and how their binding triggers a specific activity. To supplement existing molecular and cellular technologies, one may extract experimental data from the literature or NRs-specific databases. To the best of our knowledge, six available databases are devoted to NRs: Androgen Receptor gene mutations DataBase[23] (ARDB) dedicated to AR mutations; Nuclear Receptor Resource[24] (NRR) provides information on the nuclear receptor superfamily; Hormone Receptor Target Binding Loci Database[25] (HRTBLDb) contains information

DOI: 10.1021/acs.jmedchem.8b01105
J. Med. Chem. 2019, 62, 2894−2904

about the NR binding sites on mammalian genomes; Hmrbase[26] provides curated information about hormones and their receptors; Orphan Nuclear Receptor Ligand DataBase[27] (ONRLDB) gathers information on small molecules targeting orphan receptors; International Union of Basic and Clinical Pharmacology DataBase[28,29] (IUPHAR-DB) is dedicated to pharmacological data on major therapeutic target families including a specific section for NRs. The gold standard for small molecules/protein interaction data is the ChEMBL database. However, 30% of annotation errors were quantified[30] in the NR-related ChEMBL entries. Since data quality is required to generate reliable models, the Nuclear Receptors Ligands and Structures Benchmarking DataBase[30] (NRLiSt BDB) was designed. This database mainly contains activity values (IC_{50} and EC_{50}) collected through a manual literature review, including their corresponding "agonist" or "antagonist" pharmacological profile annotation.

This database constitutes a robust basis for in silico NRs research but can still be improved to provide a NR-dedicated database more suitable for compound profiling, rational chemical library design prior to screening campaign, model calibration, and/or selection in computer aided drug design protocols. In the present work, we thus propose the Nuclear Receptors Database Including Negative Data (NR-DBIND), a new database including (1) negative data that are crucial to understand what impedes ligand binding and (2) consistent pharmacological profile annotations that rely on activity data. This database is an exhaustive compilation of binding affinity data for the compounds described in the literature that have been experimentally tested on NRs (positive and negative results) and of the Protein Data Bank (PDB) structures that are fit for structure-based studies.

All details about the collected data are summarized in a database freely accessible at http://nr-dbind.drugdesign.fr. The objective of the NR-DBIND is to provide a robust raw basis for the development of NRs ligand- and structure-based models for (1) structure−activity relationship and structure−property relationship studies (SAR and SPR) and (2) machine learning methods training and evaluation. Here, we detail the composition of the NR-DBIND and discuss how it is annotated.

METHODS

Protein Collection and Annotation. The *apo* and *holo* PDB structures associated with the UNIPROT identifiers of the 48 human nuclear receptors (Table 1) were retrieved. Structures with mutations or associated with unpublished articles were discarded. Each *holo* structure was manually annotated based on its bound-ligand activity. Note that an unusual "inverse antagonist" bound annotation is provided for several liver receptor homolog 1 (LRH-1) PDB structures which corresponds to compounds capable of inhibiting inverse agonist compounds according to the authors.[31]

Ligand Collection. On the basis of NRs UNIPROT identifiers, all interactions data from ChEMBL23[32] involving one of the listed NRs were collected, and all studies related to these interactions were reviewed to correct, complete, and extend the interaction annotations. We focused on compounds associated with binding affinity data: each reported interaction is documented with the corresponding references from the literature, the ligand identification key, SMILES, IUPAC name, and additional identifiers (ZINC, CHEMBL), the protein UNIPROT ID, and when available, its associated activity and affinity assay types and values, the percentage of activity and the associated pharmacological profile. A logarithmic transformation was applied to all binding affinity data.

Binding Annotations. Ligands were annotated as binders ("B") if their associated pIC_{50} or pK_i is >7, as nonbinders ("NB") if their associated pIC_{50} or pK_i is ≤5. Other ligands are considered in the 2 log margin ($5 < pK_i/pIC_{50} \leq 7$) that is commonly used to discriminate binders from non/low binders[33,34] and that is used here to avoid classification errors (see Figures S1 and S2). Of note, pIC_{50} and pK_i values are provided, and the annotation cutoffs can be customized for convenience.

Ligand Pharmacological Profile Annotation. Two pharmacological profile annotations were provided according to the nature of the related biological assay.

A first classification referred to as the IEC50 classification is based on IC_{50} and/or EC_{50} activity data. A compound is classified as (1) "agonist" if a finite EC_{50} has been experimentally measured in an agonistic activity evaluation assay, (2) "antagonist" if it is a finite IC_{50} in an antagonistic activity evaluation assay, or (3) "agonist/antagonist" if both have been measured. This classification is indicative and displays different limits that are further developed in the Discussion section.

A second classification, the maximum activity set, accounts for the experimentally measured percentage activity as compared to a reference compound (p.a.r.) when it reaches a plateau on the dose−response curve. In agonist mode, ligands displaying ≥75 p.a.r. are classified as "agonists"; those displaying between ≥25 and <75 p.a.r. are classified as "partial agonists", and those displaying <−25 p.a.r. are classified as "inverse agonists". Similarly, in antagonist mode, ligands displaying ≥75 p.a.r. are classified as "antagonists", and those displaying between ≥25 and <75 p.a.r. are classified as "partial antagonists". Data displaying between −25 and 25 p.a.r. were not annotated because of their weak activity. When several activities are observed for a single ligand/NR pair, both are conserved. Again, raw data are provided to allow customization.

Ligands and Proteins Structures Preparation. For each ligand, the majority of protonated microspecies at pH 7.4 was computed using Marvin 17.22.0, 2017, ChemAxon (http://www.chemaxon.com). The corresponding 3D conformation was generated using iCon as implemented in LigandScout[35] (version 4.2). Basic physicochemical descriptors (molecular weight, number of hydrogen bonds donor/acceptors, number of rotatable bonds, number of aromatic rings, octanol−water partition coefficient, and topological polar surface area) and Bemis−Murcko scaffold-based clusters were computed using RDKIT (version 2016.09.1).[36]

Protein structures were protonated at pH 7 using PDB 2PQR (version 2.1.1).[37] PDB files were stripped to conserve only the chain interacting with the bound ligand.

RESULTS

NR-DBIND Overall Content. The NR-DBIND gathers data for 28 NRs and contains 593 protein structures (568 *holo* and 25 *apo*) and 7593 small molecules, with a total of 15116 affinity data for 13 566 unique interactions. Among the 13 566 small molecule/NR unique pairs, we provide up to 47 binding values extracted from different studies for a single pair. Included proteins have at least one resolved structure associated except for GR and retinoid X receptor γ (RXRγ) for which only mutated structures are available on the PDB. They were nonetheless included in the NR-DBIND due to the abundance of available binding data.

Pharmacological profile is defined for 545 *holo* structures (378 agonist-bound, 33 antagonist-bound, 47 partial agonist-bound, 48 modulator-bound, 26 inverse agonist-bound, 4 inverse antagonist-bound, and 4 weak agonist-bound structures), and 5362 and 2734 small molecules/NR interactions are annotated with IEC50 and maximum activity pharmacological profile annotations, respectively.

NR-DBIND Subsets. Since the IC_{50} or EC_{50} measurements are assay-dependent and the K_i values are assay-independent,

DOI: 10.1021/acs.jmedchem.8b01105
J. Med. Chem. 2019, 62, 2894−2904

we propose to split up those data into pIC_{50} and pK_i subsets. A logarithmic transformation was applied to all binding affinity data. A total of 9866 binding data are reported in the pIC_{50} subset, among which 8995 pairs are unique. In the pK_i subset, 4061 binding data are reported (Figure 1), among which 3693 pairs are unique.

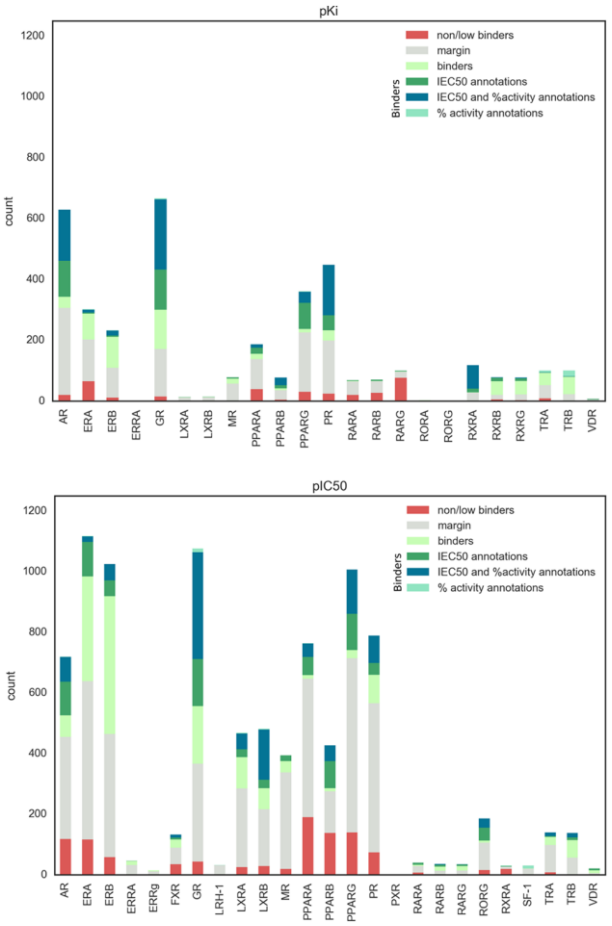


Figure 1. Number of unique ligands per subset and per NR.

Different distributions of pIC_{50} and pK_i values are observed among NRs and can be subdivided according to the type of experiment that was performed to measure the affinities (Figure S3 and Figure S4).

NR-DBIND Binding and Nonbinding Molecules. Binding annotations were assigned to interaction pairs (B or NB for binders and nonbinders, while no annotation was provided for ambiguous ligands as detailed in the Methods section). Figure 1 summarizes the number of unique B and NB ligands collected per NR (numeric details are found in Table S1). In total, 1544 NBs are reported, representing a tenth of the database content, with an average of 34.8 NBs per protein data set with at least 1 NB and a maximum of 190 NBs for PPARα in the pIC_{50} set.

NR-DBIND Annotations per Subset. In the pIC_{50} subset, 3870 and 1991 interaction data (3569 and 1749 unique interaction pairs) are annotated with IEC50 and maximum activity pharmacological profile respectively, with 1958 and 1139 unique interactions involving binders.

In the pK_i subset, 1855 and 1124 interaction data (1735 and 991 unique interaction pairs) are annotated with IEC50 and maximum activity pharmacological profile, respectively, with 1224 and 749 unique interactions involving binders.

For convenience, we provide Figure 2 and Figure 3 that give a quick overview of the different pharmacological profile ratio within the portion of annotated content per protein.

More details about the number of ligands associated with each pharmacological profile per protein are provided in Supporting Information (Tables S2 and S3). Of note, when more than one pharmacological profile was associated with a unique ligand/NR interaction pair, all profiles were considered.

Bemis–Murcko Representatives. Bemis–Murcko (BM) scaffold clustering was performed to have an overview of the number of representative chemotypes contained in the data sets (Table 2, Table S4).

In the pK_i subset, populated data sets have a varying number of representative BM scaffolds: less than 10 $_{BM}$ clusters are found in 6 protein data sets (estrogen-related receptor α (ERRα), liver X receptor α (LXRα), LXRβ, RAR-related orphan receptor α (RORα), RORγ, and vitamin D receptor (VDR)), while 10–122 $_{BM}$ clusters are found in the 17 other protein data sets (AR, ERα, ERβ, GR, MR, peroxisome proliferator-activated receptor α (PPARα), PPARβ, PPARγ, PR, RARα, RARβ, RARγ, RXRα, RXRβ, RXRγ, TRα, and TRβ). Globally, more representative scaffolds are observed in the pIC_{50} subset: 6 protein data sets (ERRγ, LRH-1, pregnate X receptor (PXR), RXRα, steroidogenic factor 1 (SF-1), VDR) are described by less than 10 $_{BM}$ scaffolds, and 19 protein data sets (AR, ERα, ERβ, ERRα, farnesoid X receptor (FXR), GR, LXRα, LXRβ, MR, PPARα, PPARβ, PPARγ, PR, RARα, RARβ, RARγ, RORγ, TRα, TRβ) are described by 10–216 $_{BM}$ clusters.

Low diversity is observed in the pIC_{50} and pK_i maximum activity sets in terms of BM scaffolds (Table S4). Among all the BM scaffolds associated with agonist and antagonist annotation in all sets, 14% are common to both activities.

NR-DBIND Web Site. Our manually curated and annotated database is freely available at http://nr-dbind.drugdesign.fr. Different subsets can be selected and retrieved depending on the protein or affinity/activity annotation as described previously. Dynamic tables allow custom selection, download, and visualization of data (Table S5).

DISCUSSION

In the present study, we describe an exhaustive NR-focused database that includes curated and comprehensive binding and affinity data with detailed pharmacological profile. This database is, to our knowledge, the largest database dedicated to NRs. It is to note that the volume of data described in the literature for each NR is variable. As expected, we observe a large amount of data available for the PPAR family[38,39] for which several drugs are commercialized with indications for diabetes (rosiglitazone)[40−42] and hyperlipidemia (bezafibrate, gemfibrozil)[43,44] and for the GR receptor that is highly studied for the treatment of inflammatory diseases and targeted by a plethora of synthetic glucocorticoids (i.e., dexamethasone, cortisol).[45,46] At the opposite very few data were collected for SF-1 that is an orphan receptor.

Integration of Negative Data. A key feature of our database is the integration of inactive compounds. Focusing on affinity data allows the collection of low- to no-affinity data that carry as much information as binding molecules, particularly

DOI: 10.1021/acs.jmedchem.8b01105
J. Med. Chem. 2019, 62, 2894−2904

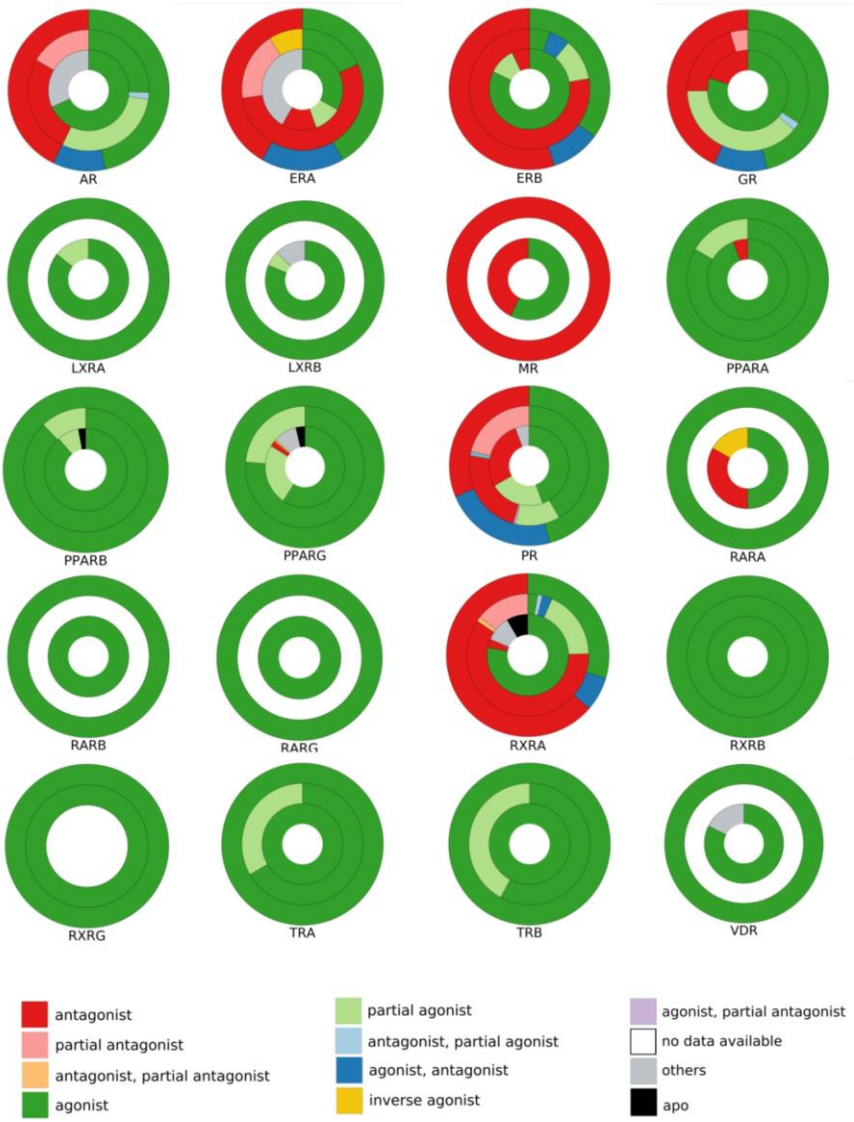


Figure 2. Content of the NR-DBIND pK_i data set in term of PDB structures (internal disk), maximum activity (middle disk), and IEC50 (external disk) pharmacological annotations when at least one molecule is annotated. The "others" annotation corresponds to "weak agonists", "inverse antagonists", "synergistic agonists", "covalent agonist/binders", and undefined pharmacological profiles.

when both share common scaffolds. In the NR-DBIND, we thus found that 30 of the 1311 $_{BM}$ scaffolds are common to both NB and B sets (Table S4).

Negative data are published in two forms in the literature: either no binding is detected, without any information on the highest tested concentration, or the highest tested concentration is provided. If both data can provide knowledge, the second one should be privileged. It is reasonable to fix a threshold value to delineate molecules considered as nonbinders to those considered as binders. The highest tested concentration should exceed this threshold to limit the introduction of bias (i.e., false negatives). Of note, affinity values can be influenced by experimental conditions and human or material errors and measurements might differ from one lab to another. Therefore, a margin should be established between the "binders" inclusive threshold and the "nonbinders" exclusive threshold. In the present work, we established a 2 log margin[34,33] between our pK_i/IC_{50} < 7 inclusive threshold and our pK_i/IC_{50} ≥ 5 exclusive threshold that should minimize the classification errors. Modifying these thresholds would rebalance the risk of misclassification and the quantity of excluded data. We recommend optimizing the threshold for specific NR studies, taking into account the binding data distribution. For instance, when a bimodal distribution is observed (as in pIC_{50} values for PPARβ) (Figures S3 and S4), the thresholds should be defined as the intersection of the two distributions. Also, very high affinity ligands have been identified for TR receptors (pIC_{50} and/or pK_i > 9) as compared to other receptors for which the maximal affinities barely reach pK_i/pIC_{50} values of 9 (RXRα, RORγ); adapted margin thresholds should therefore be defined if the aim of the study is to look for a specific range of affinity values or to reach affinities comparable to the best ones identified so far.

Negative data are informative in understanding what makes a ligand unable to bind a studied protein. In the NR-DBIND,

DOI: 10.1021/acs.jmedchem.8b01105
J. Med. Chem. 2019, 62, 2894−2904

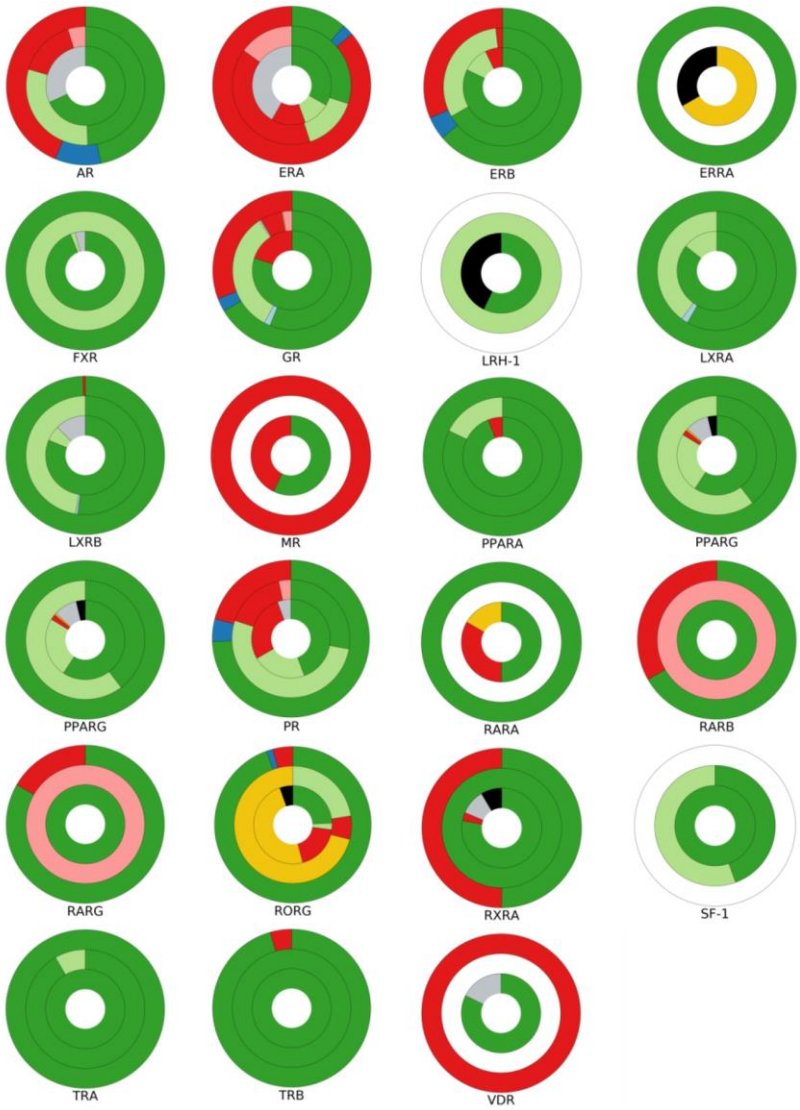


Figure 3. Content of the NR-DBIND pIC_{50} data set in term of PDB structures (internal disk), maximum activity (middle disk), and IEC50 (external disk) pharmacological annotations when at least one molecule is annotated. The "others" annotation corresponds to "weak agonists", "inverse antagonists", "synergistic agonists", "covalent agonist/binders", and undefined pharmacological profiles. See color legend in Figure 2.

we show that large affinity variations (up to 4 logs in both the pK_i and pIC_{50} sets) can be observed within a chemotype (Figure 4 and Figure 5). These particular cases, which could be referred as activity cliffs (ACs), are particularly informative to reveal chemical changes that determine structure−activity relationships (SARs).[49,50]

Such discontinuity within the affinity or activity values associated with a single chemotype is particularly valuable at the early stage of drug design to calibrate nonlinear activity prediction methods and to refine active compound profiling.[49]

Pharmacological Profile Annotation. When studying a specific activity on a given biomolecular target, it is important to rely on accurate ligand annotations.[51] A difficulty lies in the definition of the pharmacological profiles. Activity values such as EC_{50} or IC_{50} may indicate that an agonist or an antagonist activity is detected, but it may not be sufficient to determine the pharmacological profile of a ligand. A molecule may display a low EC_{50} but have only a partial efficacy. This case is particularly complex since the molecule could be considered as a partial agonist if it competes with an antagonist compound or as a partial antagonist if it competes with an agonist compound. Typical examples to illustrate the complexity in compounds annotations are the selective androgen receptor modulators (SARMs), the selective estrogen receptor modulators (SERMs), and the selective progesterone receptor modulators (SPRMs) that can elicit different tissue-selective pharmacological activities.[52−54]

Another information in the profile assignment is brought by percentage activity measurement as compared to a reference compound (p.a.r.). This value measures the biological response triggered by a query molecule as a percentage of the biological response triggered by a reference molecule when a plateau is reached in the dose−response curve. Percentage activity values must be carefully interpreted since it is fully relative to the reference compound used. Ideally, only percentage activity values obtained with the same reference compound should be

DOI: 10.1021/acs.jmedchem.8b01105
J. Med. Chem. 2019, 62, 2894−2904

Table 2. Number of Representative BM Scaffolds Per Subsets and "Binding" Categories[a]

		AR	ERα	ERβ	ERRα	ERRγ	FXR	GR	LRH-1	LXRα	LXRβ	MR	PPARα	PPARβ	PPARγ
pIC_{50}	binder	56/265	142/477	114/559	6/14	1/4	14/44	147/709	1/1	56/184	59/266	22/56	40/117	25/152	79/293
pK_i		61/323	33/100	38/124	0	0	0	95/496	0	2/2	2/2	7/21	15/49	9/41	39/135
pIC_{50}	nonbinder	31/118	47/116	36/58	0	0	14/35	19/43	1/1	15/25	18/29	11/19	75/190	44/138	52/140
pK_i		11/20	27/65	10/11	0	0	0	9/15	0	1/1	1/1	1/1	20/39	4/4	13/30
pIC_{50}	margin	82/337	104/523	126/405	9/32	1/11	18/54	83/323	8/30	75/260	68/187	64/319	133/457	46/137	185/574
pK_i		65/286	49/137	37/98	1/1	0	0	48/156	0	5/10	4/11	22/56	36/99	16/32	69/195
pIC_{50}	total BM	124	216	203	11	1	34	200	9	100	110	78	179	84	236
pK_i		101	92	71	1	0	0	122	0	7	6	26	55	23	92

		PR	PXR	RARα	RARβ	RARγ	R0Rα	RORγ	RXRα	RXRβ	RXRγ	SF-1	TRα	TRβ	VDR
pIC_{50}	binder	51/223	0	7/11	13/23	10/21	0	25/80	2/3	0	0	4/10	11/42	21/82	2/16
pK_i		38/249	0	2/5	3/5	1/4	1/1	1/1	10/90	9/58	9/55	0	22/49	23/78	2/3
pIC_{50}	nonbinder	26/74	1/1	3/7	1/2	1/2	0	5/16	2/19	0	0	0	5/8	0	1/2
pK_i		15/24	0	6/20	6/27	8/75	0	0	0	1/5	1/3	0	4/9	1/1	0
pIC_{50}	margin	103/492	0	10/22	7/11	6/12	0	26/90	2/8	0	0	4/20	28/90	17/56	2/3
pK_i		56/175	0	14/44	14/40	12/21	2/2	0	7/28	5/15	6/19	0	14/43	11/22	2/4
pIC_{50}	total BM	140	1	17	14	13	0	43	4	0	0	6	34	31	4
pK_i		72	0	15	15	14	3	1	12	11	11	0	30	30	3

[a]The data are given as a ratio (number of representative BM scaffolds)/(total number of ligand) except for the "total" raw that informs the number of representative BM scaffolds irrespective of the "binding" category.

DOI: 10.1021/acs.jmedchem.8b01105
J. Med. Chem. 2019, 62, 2894−2904

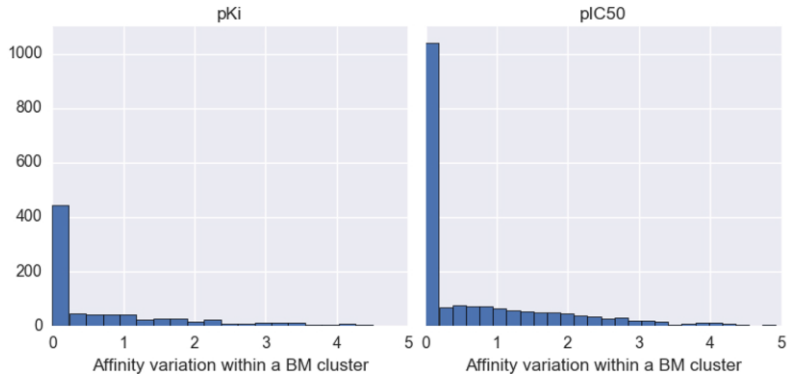


Figure 4. Distribution of the pK_i and the pIC_{50} variations (logarithmic scale) observed within a BM cluster.

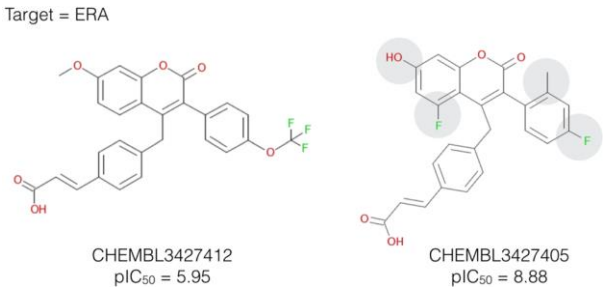


Figure 5. Example of large target affinity variation within a BM cluster.

compared, particularly when distinction between partial and full (ant)agonism is intended.

In the NR-DBIND, we propose a curated set in which all annotations are supported by percentage activity data (maximum activity data set). When available, the reference compound and its concentration are displayed so that the user can select homogeneous subsets.

Applications of the NR-DBIND. The NR-DBIND can be used as a reliable source of data for researchers from different fields trying to understand and modulate NRs functions. The NR-DBIND data sets are notably suited for in silico studies using structure- and ligand-based methods.

Since X-ray structures are provided for most of the proteins considered in the database, structural studies can be conducted; hypothetical compounds binding modes can be generated and used to identify physicochemical and conformational details that might be critical for compound potency and SARs.[47] For NRs such as AR for which no antagonist structure is available, molecular dynamics simulations can be used to generate structures of the human AR prior to apply docking methods to select the structure that enrich antagonist molecules the most.[48]

An example of ligand-based application is for NR selective pharmacophore modeling: to look for ERα agonists, agonist molecules should be used to build pharmacophore models, and compounds with a different activity annotation (i.e., inverse agonists, antagonists, etc.) plus compounds displaying low to no affinity for ERα should be used as negative compounds to evaluate and refine the model. The recent work of Vogt[55,56] that proposes a protocol to predict compound profile from machine learning methods and the Métivier et al.[57] pharmacophore network analysis that helps identify relations between chemical series and understanding of the influence of a chemical feature on activity or affinity are perfect examples of what could be performed using the annotations provided in the NR-DBIND.

CONCLUSION

To the best of our knowledge, the NR-DBIND is the most exhaustive database that contains small molecules affinity and activity data for NRs that have been extracted from thorough manual literature reviewing. The NR-DBIND contains both positive and negative affinity data and therefore represents a robust basis to conduct SAR/SPR studies, to evaluate both ligand- and structure-based methods, and to refine ligand-based models. The NR-DBIND can be used to assist scientists from various research fields in selecting ligands for NR in drug design campaign or in understanding EDCs mode of action.

The pharmacological profile annotation provided is solely based on collected activity values (IC_{50}, EC_{50} for one classification or percentage activity compared to reference compounds for the other) and the associated type of assay, allowing the application of a unique definition for each activity profile (agonist, antagonist, partial agonist, partial antagonist, inverse agonist, inverse antagonist) regardless of the studies activity profile definition. Despite the difficulty of homogenizing the experimental values due to the many external parameters they rely on, the use of a unique definition increases the reliability of the annotation. The NR-DBIND provides manually curated and annotated, transparent and customizable data sets that are freely available online at http://nr-dbind.drugdesign.fr.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jmedchem.8b01105.

Figure S1 showing observed uncertainty in binding values from the pIC_{50}; Figure S2 showing observed uncertainty in binding values from the pK_i; Figure S3 showing distribution of the pIC_{50} affinity values; Figure S4 showing distribution of the pK_i affinity values; Table S1 listing overall content of the NR-DBIND database; Table S2 listing content of the IEC50 annotations data set; Table S3 listing content of the maximum activity annotations data sets; Table S4 listing number of BM scaffold in the entire data set and the maximum activity

DOI: 10.1021/acs.jmedchem.8b01105
J. Med. Chem. 2019, 62, 2894−2904

annotated subsets; Table S5 listing NR-DBIND column names and descriptors (PDF)

AUTHOR INFORMATION

Corresponding Author

*E-mail: matthieu.montes@cnam.fr.

ORCID

Matthieu Montes: 0000-0001-5921-460X

Notes

The authors declare no competing financial interest.

Biographies

Manon Réau attained a M.Sc. in Biotechnology at Sup'Biotech Paris and a M.Sc. in in silico drug design at the University of Paris Diderot. She joined the Conservatoire National des Arts et Métiers (Cnam) in 2016 as a Ph.D. student. Most of her Ph.D. work consists in studying the integration of negative data in benchmarking database and computational models that aim to enhance both methods evaluation and prediction of small molecules affinities for nuclear receptors. She authored three publications in peer-reviewed journals.

Nathalie Lagarde obtained her PharmD at the University of Caen (France) and completed a Ph.D. in Structural Bioinformatics in 2014 at Le Conservatoire National des Arts et Métiers (France). Her Ph.D. was focused on the evaluation of virtual screening methods and their application to the identification of cytokines inhibitors. After receiving her Ph.D., Nathalie Lagarde studied, as a postdoctoral researcher, nuclear receptors agonist and antagonist ligands prediction, protein−protein interactions prediction, and drug repositioning. In 2018, she returned to Le Conservatoire National des Arts et Métiers as an Assistant Professor. Nathalie Lagarde has published 13 publications in peer-reviewed journals.

Jean-François Zagury has obtained his Ph.D. in immunovirology in 1991 and his M.D. degree in 1992 from Université Pierre and Marie Curie (Paris). He is chairman of the Bioinformatics Department in Conservatoire National des Arts et Métiers since 2004 and founded the Genomics, Bioinformatics, and Applications Laboratory, gathering 15 researchers in genomics, structural bioinformatics, and drug design. His research has always been oriented towards medical applications. He has been a precursor in the genomics field in AIDS. He has created the phasing software Shape-IT which has become an international reference. He has patented several candidate drugs, some of which are under clinical development. He has published 144 publications in international scientific peer-reviewed journals, has founded three start-up companies, and is inventor of 22 patents.

Matthieu Montes is the head of the molecular modeling and drug design team of the GBA lab at Conservatoire National des Arts et Métiers, Paris, France. He holds a M.Sc. in Biochemistry and Bioinformatics from Paris Diderot University (2004), a Ph.D. in Pharmaceutical Sciences from Paris Descartes University (2007), and a Habilitation in Structural Biochemistry from Paris Sud University (2014). He published more than 40 papers in peer-reviewed journals and is co-inventor of 6 patent applications. Since 2014, he is a Fellow of the European Research Council. His research interests include molecular modeling, drug discovery and design, interactive visualization and simulation methods, and computational geometry.

ACKNOWLEDGMENTS

M.R. is supported by a MESR fellowship. The authors thank Dr. Florent Langenfeld for his comments on the manuscript.

ABBREVIATIONS USED

AC, activity cliff; AR, androgen receptor; BM, Bemis−Murcko; B, binder; CADD, computer aided drug design; EDC, endocrine disrupting chemical; ERα, estrogen receptor α; ERβ, estrogen receptor β; ERRα, estrogen-related receptor α; ERRγ, estrogen-related receptor γ; FXR, farnesoid X receptor; GR, glucocorticoid receptor; IEC50, combination of IC_{50} and EC_{50} information; LBD, ligand binding domain; LRH-1, liver receptor homolog; LXRα, liver X receptor α; LXRβ, liver X receptor β; MR, mineralocorticoid receptor; NB, nonbinder; NR, nuclear receptor; NR-DBIND, Nuclear Receptors Database Including Negative Data; p.a.r., percentage activity compared to a reference; PPARα, peroxisome proliferator-activated receptor α; PPARβ, peroxisome proliferator-activated receptor β; PPARγ, peroxisome proliferator-activated receptor γ; PDB, Protein Data Bank; PR, progesterone receptor; PXR, pregnate X receptor; RARα, retinoic acid receptor α; RARβ, retinoic acid receptor β; RARγ, retinoic acid receptor γ; RORα, RAR-related orphan receptor α; RORβ, RAR-related orphan receptor β; RORγ, RAR-related orphan receptor γ; RXRα, retinoid X receptor α; RXRβ, retinoid X receptor β; RXRγ, retinoid X receptor γ; SAR, structure−activity relationship; SARM, selective androgen receptor modulator; SERM, selective estrogen receptor modulator; SF-1, steroidogenic factor; SPR, structure−property relationship; TRα, thyroid hormone receptor α; TRβ, thyroid hormone receptor β; VDR, vitamin D receptor

REFERENCES

(1) Liu, S.; Downes, M.; Evans, R. M. Metabolic Regulation by Nuclear Receptors. In *Innovative Medicine*; Springer: Tokyo, 2015; pp 25−37, DOI: 10.1007/978-4-431-55651-0_2.

(2) Landry, Y.; Gies, J.-P. Drugs and Their Molecular Targets: An Updated Overview. *Fundam. Clin. Pharmacol.* **2008**, *22* (1), 1−18.

(3) Schulman, I. G.; Heyman, R. A. The Flip Side: Identifying Small Molecule Regulators of Nuclear Receptors. *Chem. Biol.* **2004**, *11* (5), 639−646.

(4) Zhao, Y.; Zhang, K.; Giesy, J. P.; Hu, J. Families of Nuclear Receptors in Vertebrate Models: Characteristic and Comparative Toxicological Perspective. *Sci. Rep.* **2015**, *5*, 8554.

(5) Burris, T. P.; Busby, S. A.; Griffin, P. R. Targeting Orphan Nuclear Receptors for Treatment of Metabolic Diseases and Autoimmunity. *Chem. Biol.* **2012**, *19* (1), 51−59.

(6) Sever, R.; Glass, C. K. Signaling by Nuclear Receptors. *Cold Spring Harbor Perspect. Biol.* **2013**, *5* (3), a016709.

(7) 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.

(8) 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.

(9) Santos-Silva, A. P.; Andrade, M. N.; Pereira-Rodrigues, P.; Paiva-Melo, F. D.; Soares, P.; Graceli, J. B.; Dias, G. R. M.; Ferreira, A. C. F.; de Carvalho, D. P.; Miranda-Alves, L. Frontiers in Endocrine Disruption: Impacts of Organotin on the Hypothalamus-Pituitary-Thyroid Axis. *Mol. Cell. Endocrinol.* **2018**, *460*, 246−257.

(10) Engel, A.; Buhrke, T.; Imber, F.; Jessel, S.; Seidel, A.; Völkel, W.; Lampen, A. Agonistic and Antagonistic Effects of Phthalates and Their Urinary Metabolites on the Steroid Hormone Receptors ERα, ERβ, and AR. *Toxicol. Lett.* **2017**, *277*, 54−63.

(11) Fisher, J. S. Environmental Anti-Androgens and Male Reproductive Health: Focus on Phthalates and Testicular Dysgenesis Syndrome. *Reproduction* **2004**, *127* (3), 305−315.

(12) Rouiller-Fabre, V.; Guerquin, M. J.; N'Tumba-Byn, T.; Muczynski, V.; Moison, D.; Tourpin, S.; Messiaen, S.; Habert, R.; Livera, G. Nuclear Receptors and Endocrine Disruptors in Fetal and

Neonatal Testes: A Gapped Landscape. *Front. Endocrinol.*, **2015**, *6*, DOI: 10.3389/fendo.2015.00058.
(13) 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.
(14) Ripa, L.; Edman, K.; Dearman, M.; Edenro, G.; Hendrickx, R.; Ullah, V.; Chang, H.-F.; Lepistö, M.; Chapman, D.; Geschwindner, S.; Wissler, L.; Svanberg, P.; Lawitz, K.; Malmberg, J.; Nikitidis, A.; Olsson, R. I.; Bird, J.; Llinas, A.; Hegelund-Myrbäck, T.; Berger, M.; Thorne, P.; Harrison, R.; Köhler, C.; Drmota, T. Discovery of a Novel Oral Glucocorticoid Receptor Modulator (AZD9567) with Improved Side Effect Profile. *J. Med. Chem.* **2018**, *61* (5), 1785−1799.
(15) Benod, C.; Carlsson, J.; Uthayaruban, R.; Hwang, P.; Irwin, J. J.; Doak, A. K.; Shoichet, B. K.; Sablin, E. P.; Fletterick, R. J. Structure-Based Discovery of Antagonists of Nuclear Receptor LRH-1. *J. Biol. Chem.* **2013**, *288* (27), 19830−19844.
(16) Taub, R.; Chiang, E.; Chabot-Blanchet, M.; Kelly, M. J.; Reeves, R. A.; Guertin, M.-C.; Tardif, J.-C. Lipid Lowering in Healthy Volunteers Treated with Multiple Doses of MGL-3196, a Liver-Targeted Thyroid Hormone Receptor-β Agonist. *Atherosclerosis* **2013**, *230* (2), 373−380.
(17) Kono, M.; Ochida, A.; Oda, T.; Imada, T.; Banno, Y.; Taya, N.; Masada, S.; Kawamoto, T.; Yonemori, K.; Nara, Y.; Fukase, Y.; Yukawa, T.; Tokuhara, H.; Skene, R.; Sang, B. C.; Hoffman, I. D.; Snell, G. P.; Uga, K.; Shibata, A.; Igaki, K.; Nakamura, Y.; Nakagawa, H.; Tsuchimori, N.; Yamasaki, M.; Shirai, J.; Yamamoto, S. Discovery of [*cis*-3-({(5*R*)-5-[(7-Fluoro-1,1-dimethyl-2,3-dihydro-1*H*-inden-5-yl)carbamoyl]-2-methoxy-7,8-dihydro-1,6-naphthyridin-6(5*H*)-yl}-carbonyl)cyclobutyl]acetic Acid (TAK-828F) as a Potent, Selective, and Orally Available Novel Retinoic Acid Receptor-Related Orphan Receptor Γt Inverse Agonist. *J. Med. Chem.* **2018**, *61* (7), 2973−2988.
(18) Tria, G. S.; Abrams, T.; Baird, J.; Burks, H. E.; Firestone, B.; Gaither, L. A.; Hamann, L. G.; He, G.; Kirby, C. A.; Kim, S.; Lombardo, F.; Macchi, K. L.; McDonnell, D. P.; Mishina, Y.; Norris, J. D.; Nunez, J.; Springer, C.; Sun, Y.; Thomsen, N. M.; Wang, C.; Wang, J.; Yu, B.; Tiong-Yip, C. L.; Peukert, S. Discovery of LSZ102, a Potent, Orally Bioavailable Selective Estrogen Receptor Degrader (SERD) for the Treatment of Estrogen Receptor Positive Breast Cancer. *J. Med. Chem.* **2018**, *61* (7), 2837−2864.
(19) Fan, F.; Hu, R.; Munzli, A.; Chen, Y.; Dunn, R. T.; Weikl, K.; Strauch, S.; Schwandner, R.; Afshari, C. A.; Hamadeh, H.; Nioi, P. Utilization of Human Nuclear Receptors as an Early Counter Screen for off-Target Activity: A Case Study with a Compendium of 615 Known Drugs. *Toxicol. Sci.* **2015**, *145* (2), 283−295.
(20) 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. *C. R. Biol.* **2017**, *340* (9−10), 414−420.
(21) Hunt, J. P.; Schinn, S.-M.; Jones, M. D.; Bundy, B. C. Rapid, Portable Detection of Endocrine Disrupting Chemicals through Ligand-Nuclear Hormone Receptor Interactions. *Analyst* **2017**, *142* (24), 4595−4600.
(22) 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.
(23) Gottlieb, B.; Beitel, L. K.; Wu, J. H.; Trifiro, M. The Androgen Receptor Gene Mutations Database (ARDB): 2004 Update. *Hum. Mutat.* **2004**, *23* (6), 527−533.
(24) Martinez, E.; Moore, D. D.; Keller, E.; Pearce, D.; Robinson, V.; MacDonald, P. N.; Simons, S. S.; Sanchez, E.; Danielsen, M. The Nuclear Receptor Resource Project. *Nucleic Acids Res.* **1997**, *25* (1), 163−165.
(25) Kennedy, B. A.; Gao, W.; Huang, T. H.-M.; Jin, V. X. HRTBLDb: An Informative Data Resource for Hormone Receptors Target Binding Loci. *Nucleic Acids Res.* **2010**, *38* (Suppl.1, Database), D676−D681.
(26) Rashid, M.; Singla, D.; Sharma, A.; Kumar, M.; Raghava, G. P. Hmrbase: A Database of Hormones and Their Receptors. *BMC Genomics* **2009**, *10*, 307.
(27) Nanduri, R.; Bhutani, I.; Somavarapu, A. K.; Mahajan, S.; Parkesh, R.; Gupta, P. ONRLDB—Manually Curated Database of Experimentally Validated Ligands for Orphan Nuclear Receptors: Insights into New Drug Discovery. *Database* **2015**, *2015*, bav112.
(28) Harmar, A. J.; Hills, R. A.; Rosser, E. M.; Jones, M.; Buneman, O. P.; Dunbar, D. R.; Greenhill, S. D.; Hale, V. A.; Sharman, J. L.; Bonner, T. I.; Catterall, W. A.; Davenport, A. P.; Delagrange, P.; Dollery, C. T.; Foord, S. M.; Gutman, G. A.; Laudet, V.; Neubig, R. R.; Ohlstein, E. H.; Olsen, R. W.; Peters, J.; Pin, J. P.; Ruffolo, R. R.; Searls, D. B.; Wright, M. W.; Spedding, M. IUPHAR-DB: The IUPHAR Database of G Protein-Coupled Receptors and Ion Channels. *Nucleic Acids Res.* **2009**, *37* (Database), D680−D685.
(29) Sharman, J. L.; Mpamhanga, C. P.; Spedding, M.; Germain, P.; Staels, B.; Dacquet, C.; Laudet, V.; Harmar, A. J. IUPHAR-DB: New Receptors and Tools for Easy Searching and Visualization of Pharmacological Data. *Nucleic Acids Res.* **2011**, *39* (Suppl. 1, Database), D534−D538.
(30) Lagarde, N.; Ben Nasr, N.; Jérémie, A.; Guillemain, H.; Laville, V.; Labib, T.; Zagury, J.-F.; Montes, M. NRLiSt BDB, the Manually Curated Nuclear Receptors Ligands and Structures Benchmarking Database. *J. Med. Chem.* **2014**, *57* (7), 3117−3125.
(31) Matsushima, A.; Kakuta, Y.; Teramoto, T.; Koshiba, T.; Liu, X.; Okada, H.; Tokunaga, T.; Kawabata, S.-I.; Kimura, M.; Shimohigashi, Y. Structural Evidence for Endocrine Disruptor Bisphenol A Binding to Human Nuclear Receptor ERR Gamma. *J. Biochem.* **2007**, *142* (4), 517−524.
(32) ChEMBL23, 10.6019/CHEMBL.database.23, May, 2017.
(33) Husby, J.; Bottegoni, G.; Kufareva, I.; Abagyan, R.; Cavalli, A. Structure-Based Predictions of Activity Cliffs. *J. Chem. Inf. Model.* **2015**, *55* (5), 1062−1076.
(34) Stumpfe, D.; Hu, Y.; Dimova, D.; Bajorath, J. Recent Progress in Understanding Activity Cliffs and Their Utility in Medicinal Chemistry. *J. Med. Chem.* **2014**, *57* (1), 18−28.
(35) 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.
(36) RDKit: Open-Source Cheminformatics. http://www.rdkit.org, version 2016.09.1.
(37) Dolinsky, T. J.; Nielsen, J. E.; McCammon, J. A.; Baker, N. A. PDB2PQR: An Automated Pipeline for the Setup of Poisson−Boltzmann Electrostatics Calculations. *Nucleic Acids Res.* **2004**, *32* (Suppl. 2, Web Server), W665−W667.
(38) Desvergne, B.; Wahli, W. Peroxisome Proliferator-Activated Receptors: Nuclear Control of Metabolism. *Endocr. Rev.* **1999**, *20* (5), 649−688.
(39) Straus, D. S.; Glass, C. K. Anti-Inflammatory Actions of PPAR Ligands: New Insights on Cellular and Molecular Mechanisms. *Trends Immunol.* **2007**, *28* (12), 551−558.
(40) Edvardsson, U.; Bergström, M.; Alexandersson, M.; Bamberg, K.; Ljung, B.; Dahllöf, B. Rosiglitazone (BRL49653), a PPARgamma-Selective Agonist, Causes Peroxisome Proliferator-like Liver Effects in Obese Mice. *J. Lipid Res.* **1999**, *40* (7), 1177−1184.
(41) Kahn, B. B.; McGraw, T. E. Rosiglitazone, PPARγ, and Type 2 Diabetes. *N. Engl. J. Med.* **2010**, *363* (27), 2667−2669.
(42) Berger, J.; Wagner, J. A. Physiological and Therapeutic Roles of Peroxisome Proliferator-Activated Receptors. *Diabetes Technol. Ther.* **2002**, *4* (2), 163−174.
(43) Ho, L. T.; Chiang, H. L.; Tsai, H.; Chou, T. Y.; Kwok, C. F. A Therapeutic Trial of Bezafibrate on Patients with Hyperlipidemia with or without Diabetes Mellitus. *Proc. Natl. Sci. Counc., Repub. China B* **1984**, *8* (3), 240−245.
(44) Tenenbaum, A.; Motro, M.; Fisman, E. Z. Dual and Pan-Peroxisome Proliferator-Activated Receptors (PPAR) Co-Agonism: The Bezafibrate Lessons. *Cardiovasc. Diabetol.* **2005**, *4*, 14.
(45) Gulliver, L. S. M. Xenobiotics and the Glucocorticoid Receptor. *Toxicol. Appl. Pharmacol.* **2017**, *319*, 69−79.

DOI: 10.1021/acs.jmedchem.8b01105
J. Med. Chem. 2019, 62, 2894−2904

(46) McMaster, A.; Ray, D. W. Drug Insight: Selective Agonists and Antagonists of the Glucocorticoid Receptor. *Nat. Clin. Pract. Endocrinol. Metab.* **2008**, *4* (2), 91−101.
(47) Mellor, C. L.; Steinmetz, F. P.; Cronin, M. T. D. Using Molecular Initiating Events to Develop a Structural Alert Based Screening Workflow for Nuclear Receptor Ligands Associated with Hepatic Steatosis. *Chem. Res. Toxicol.* **2016**, *29* (2), 203−212.
(48) 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.
(49) Bajorath, J. Representation and Identification of Activity Cliffs. *Expert Opin. Drug Discovery* **2017**, *12* (9), 879−883.
(50) Hu, X.; Hu, Y.; Vogt, M.; Stumpfe, D.; Bajorath, J. MMP-Cliffs: Systematic Identification of Activity Cliffs on the Basis of Matched Molecular Pairs. *J. Chem. Inf. Model.* **2012**, *52* (5), 1138−1145.
(51) Lagarde, N.; Zagury, J.-F.; Montes, M. Importance of the Pharmacological Profile of the Bound Ligand in Enrichment on Nuclear Receptors: Toward the Use of Experimentally Validated Decoy Ligands. *J. Chem. Inf. Model.* **2014**, *54* (10), 2915−2944.
(52) Chen, T. Nuclear Receptor Drug Discovery. *Curr. Opin. Chem. Biol.* **2008**, *12* (4), 418−426.
(53) Martinkovich, S.; Shah, D.; Planey, S. L.; Arnott, J. A. Selective Estrogen Receptor Modulators: Tissue Specificity and Clinical Utility. *Clin. Interventions Aging* **2014**, *9*, 1437−1452.
(54) Feng, Q.; O'Malley, B. W. Nuclear Receptor Modulation - Role of Coregulators in Selective Estrogen Receptor Modulator (SERM) Actions. *Steroids* **2014**, *90*, 39−43.
(55) Vogt, M.; Jasial, S.; Bajorath, J. Extracting Compound Profiling Matrices from Screening Data. *ACS Omega* **2018**, *3* (4), 4706−4712.
(56) Rodríguez-Pérez, R.; Miyao, T.; Jasial, S.; Vogt, M.; Bajorath, J. Prediction of Compound Profiling Matrices Using Machine Learning. *ACS Omega* **2018**, *3* (4), 4713−4723.
(57) Métivier, J.-P.; Cuissart, B.; Bureau, R.; Lepailleur, A. The Pharmacophore Network: A Computational Method for Exploring Structure-Activity Relationships from a Large Chemical Data Set. *J. Med. Chem.* **2018**, *61* (8), 3551−3564.

DOI: 10.1021/acs.jmedchem.8b01105
J. Med. Chem. 2019, 62, 2894−2904

2.3 Conclusion et perspectives

2.3.1 Contenu de la NR-DBIND

Au total, la banque de données NR-DBIND recense 15116 données d'affinités, parmi lesquelles 9866 et 4061 sont respectivement renseignées en pIC50 et en pKi. Environ 10% des données collectées (1544) sont des données d'inactivités. 5364 interactions, correspondant à 3192 couples molécule/protéine uniques, sont annotées via des données d'IC50 et d'EC50 extraites de test d'activité et 2732 interactions, correspondant à 1840 couples molécule/protéine uniques, sont annotées via des données de pourcentage d'activité mesurés par rapport à un ligand de référence.

2.3.2 Informations regroupées dans la NR-DBIND

La NR-DBIND recense des informations brutes directement extraites de la littérature, ainsi que leur version formatée, de sorte à simplifier l'exploration de la banque de données (Tableau 15). Des descripteurs simples (le poids moléculaire des atomes lourds (HAMW), le poids moléculaire total (MW), le nombre de donneurs et d'accepteurs de liaison hydrogène (HBA/HBD), le nombre de liaisons rotatives (rotabonds), le nombre de cycle aromatiques (ArRings) le coefficient de partition octanol/eau (cLogP) et l'aire de la surface polaire (TPSA)) ont été pré-calculés avec RDKIT. Des annotations déduites des données récoltées sont proposées (*binder*/*non-binder* ; agoniste/antagoniste/agoniste partiel/antagoniste partiel etc.). Elles peuvent cependant être modifiées selon les critères de l'utilisateur.

Tableau 15 Ensemble des informations répertoriées dans la NR-DBIND

Identifiers	ID	Unique ligand identifier
	pubmed_ID	PubMed identifier
	cpd_key	Compound key as referred in the reference paper
	Protein	Protein abbreviation
	Organism	Organism
	IUPAC	International Union of Pure and Applied Chemistry chemical nomenclature
	CHEMBLID	CHEMBL identifier
Raw data	Agonist_assay	Agonist assay performed
	Antagonist_assay	Antagonist assay performed

	Binding_assay	Binding assay performed
	Binding_assay_value	Raw binding assay value
	Activity	Percentage activity as compared to a reference
	type_activity_test	Nature of the test performed for percetage activity measurement
Formatted data	ag_type	Type of agonist activity measured (pIC50, pKi)
	ag_eq	>, = or <
	ag_val	Agonist activity value
	ag_unit	Agonist activity unit
	antag_type	Type of antagonist activity measured (pIC50, pKi)
	antag_eq	>, = or <
	antag_val	Antagonist activity value
	antag_unit	Antagonist activity unit
	binding_eq	>, = or <
	p_binding_value	Binding value in logarythmic scale
	p_binding_type	Binding value type (pKi, pIC50 or RBA (Relative Binding Affinity))
	binder_nonbinder	Binder (pValue < 7), margin (7 <= pValue < 5), non binder (5 < = pValue)
	activity_val	Percentage activity if measured
	binding_assay_type	Type of binding assay performed (homogenised)
Topological information	canonical_smile	Canonical smile
	BM	Bemis Murcko scaffold
	Bmcluster	Bemis Murcko cluster
Computed descriptors	HAMW	Heavy Atom Molecular Weight
	MW	All Atom Molecular Weight
	numHdon	Number of hydrogen donor(s)
	numHacc	Number of hydrogen acceptor(s)
	Rotabonds	Number of rotatable bonds
	ArRings	Number of aromatic rings
	logP	logP
	TPSA	Topological Polar Surface Area
Annotations	IEC50_based_annotation	Activity annotation deduced from pIC50 or pEC50 values in the corresponding paper
	perc_based_annotation	Activity annotation deduced from percentage activity

		values in the corresponding paper
	transfered_IEC50_annot	All activity annotation(s) deduced from pIC50 or pEC50 for a unique protein/ligand interaction pair
	transfered_perc_annot	All activity annotation(s) deduced from percentage activity for a unique protein/ligand interaction pair

2.3.3 Mise en ligne de la NR-DBIND

La NR-DBIND est accessible gratuitement en ligne (nr-dbind.drugdesign.fr)(Figure 48). La page internet offre la possibilité de télécharger la base de données dans son intégralité ou sous forme de sous-jeux de données découpés par :

- Récepteur nucléaire
- Type de d'affinité mesurée (pIC50/pKi)
- Type de données utilisées pour l'annotation du profil pharmacologique (données d'IC50 ou d'EC50/ pourcentage d'activité par rapport à une molécule de référence).

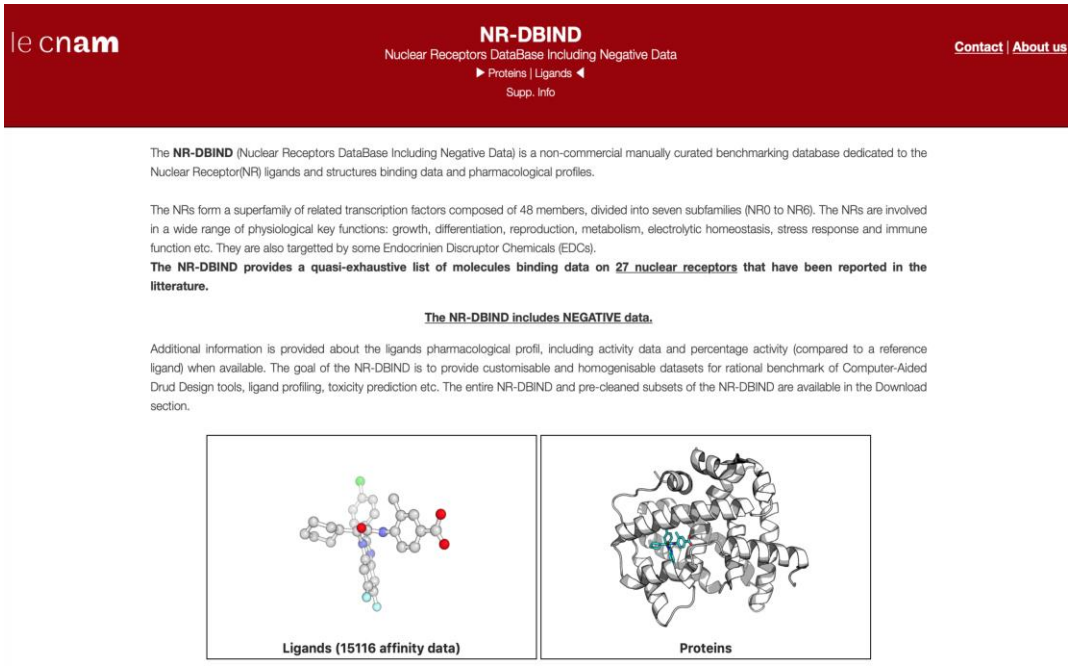


Figure 48 Page d'accueil du site nr-dbind.drugdesign.fr

Des onglets « Protéines » et « Ligands » permettent d'accéder rapidement aux données extraites des structures protéiques recensées et des interactions ligand/protéine extraites de la littérature. Dans chacun des onglets, un tableau interactif permet à l'utilisateur de filtrer les données et de les exporter. La NR-DBIND se veut modulable de sorte à ce que les données collectées puissent être facilement filtrées, téléchargées et exploitées par l'utilisateur.

2.3.4 Discussion : Biais de publication

Le pourcentage de molécules inactives recensées dans la NR-DBIND est moindre en comparaison des taux observées lors d'études de HTS[450451] (Figure 49). Idéalement, le ratio de composés actifs/inactifs d'une banque de données devrait se rapprocher de ceux observés dans les expériences de HTS, qui présentent des taux de succès faibles (0,1 à 4%[44]). Ce ratio est généralement respecté dans les banques de données incluant des *decoys*. Dans la NR-DBIND, l'inclusion de molécules inactives expérimentalement validées bouleverse ces ratios : des ratios entre 0.05 et 2,90 sont observés selon le récepteur nucléaire (Tableau 16). Ces ratios élevés illustrent le biais de publication : le système de publication actuel tend à favoriser la publication de résultats positifs et à sous-estimer l'apport des résultats négatifs. Un des effets néfastes du traitement préférentiel des données positives est l'absence de publication de séries chimiques n'ayant aucune affinité pour la cible et la négligence de l'information apportée par ces résultats négatifs. Cette tendance, largement observée en recherche clinique, limite l'élaboration de modèles de prédiction robustes. Des initiatives comme la PubChem Bioassay, qui regroupe environ 230 millions de données de bio-activités issues de HTS fournies par des laboratoires

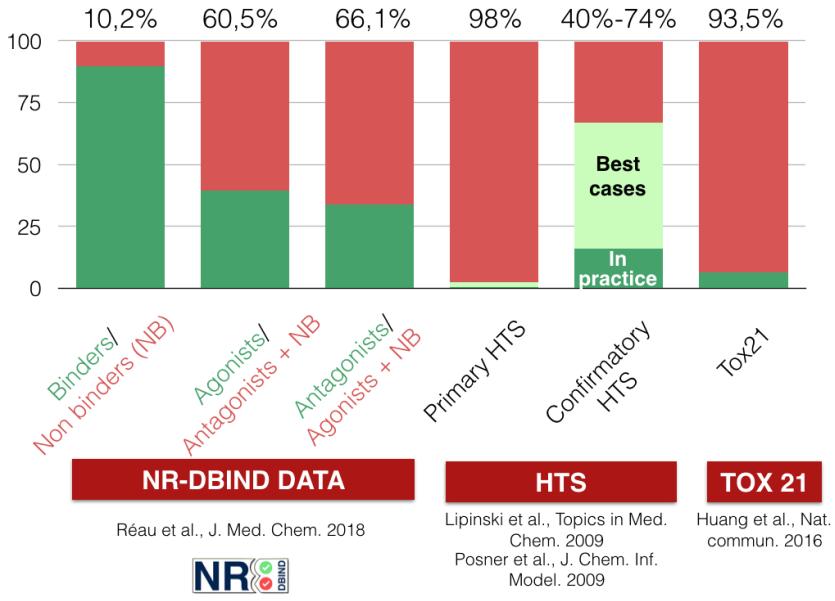


Figure 49 Illustration du biais de publication : seulement 10.2% des molécules collectées dans la banque de données NR-DBIND et extraites de la littérature ne présentent pas ou peu d'affinité pour la cible. Lorsque le profil pharmacologique est pris en compte, ce pourcentage augmente principalement à cause de la diminution de la quantité de données considérée. Ces valeurs sont nettement inférieures aux pourcentages de molécules inactives observés dans les résultats de HTS primaires et confirmatoires.

académiques et privés et qui inclut des données d'inactivité, doivent être encouragées. La publication massive de données d'inactivité fiables devrait permettre d'améliorer les performances des modèles de criblages.

3 Analyse de la capacité de discrimination des molécules actives et inactives de différents protocoles de docking, comparaison avec l'utilisation de *decoys* générés à partir de la DUD-E

3.1 Introduction

Les approches de criblage *in silico* sont largement utilisées en phases précoces de recherche de médicament afin d'identifier, à partir de chiomiothèques virtuelles, ceux qui sont capable d'interagir avec la cible d'étude et de moduler son activité. Il n'existe pas de méthode consensus pour identifier ces molécules appelées « hits » et les performances des outils de *drug design* assisté par ordinateur (CADD) présentent des performances variables selon 1) les données d'interaction protéine/ligand et les structures protéiques disponibles ainsi que leur qualité, 2) l'espace chimique étudié et 3) les propriétés physicochimiques et géométriques du site de liaison de la protéine d'étude. Pour ces raisons, les études de criblage virtuel débutent avec une étape d'évaluation, dite rétrospective, des différents protocoles de CADD sur une banque de données de référence (banque d'évaluation) appropriée avant d'appliquer le protocole optimal sur une plus large chiomiothèques[452,453,454,372]. Afin de reproduire les conditions des études prospectives, les banques de données d'évaluation sont composées d'un ensemble de données actives et d'un ensemble de données inactives. Idéalement, toutes ces données devraient être sélectionnées à partir de résultats expérimentaux confirmant l'activité ou l'inactivité d'une molécule. Cependant, du fait du manque de publication des molécules expérimentalement prouvées comme inactives, des molécules supposées inactives, appelées *decoys*, sont utilisées[381,454,455,393,1,4,456]. Le mode de sélection des *decoys*, les biais inhérents à leur sélection et les stratégies adoptées pour minimiser ces biais sont décrits en partie 5.1.2 et Résultats 1. Les performances observées sur les banques de données d'évaluation sont cependant peu reproduites lors des protocoles de criblages prospectifs de nouvelles chimiothèques. Une explication possible réside dans les méthodes de sélection des *decoys* qui, pour minimiser l'intégration de biais dans les modèles, imposent une similarité physicochimique et une distance

topologique avec les molécules actives. Or d'autres molécules actives peuvent répondre à ces critères tout comme une molécule inactive peut partager une forte similarité structurale avec les molécules actives. Pour mieux appréhender les études prospectives, et grâce à la multiplication de données inactives publiées dans la littérature, plusieurs banques de données comme la MUV [455] et la NR-DBIND[81] (Cf Résultats 2) imposent la validation expérimentale comme critère d'inclusion des molécules actives et inactives. Les composés supposés inactifs, ou *decoys*, sont donc substitués par des molécules expérimentalement validées comme étant inactives. Cependant, l'utilsation de decoys reste encore une pratique courante puisqu'il peuvent être générés rapidement à l'aide d'outils de génération (semi-) automatiques de decoys[1,150]. L'introduction de molécules inactives nécessite, quant à elle, une étape minutieuse et chronophage de revue de la littérature scientifique (Cf Résultats 1 et 2). Par ailleurs, le biais de publication, c'est-à-dire tendance à favoriser la publication de résultats positifs par rapport aux résultats négatifs[457,458], se traduit par des ratios de molécules actives/inactives recensés qui restent très élevés en comparaison des ratios molécules actives/*decoys* communément utilisés et des résultats observés en chimie médicinale (Cf Résultats 2.3.4) et qui nécessitent une interprétation adaptée.

Il existe peu d'études s'intéressant à l'impact de l'intégration de données d'inactivité expérimentalement validées dans les jeux données sur l'évaluation des performances des protocoles de CADD[459,170,211]. Dans la présente étude, nous tentons de fournir des réponses aux interrogations suivantes :

1) Sommes-nous capables de discriminer les molécules actives des molécules inactives avec des outils de docking disponibles gratuitement ?
2) Existe-t-il des conditions de docking qui favorisent cette discrimination ?
3) Les performances observées en utilisant des molécules expérimentalement validées comme inactives reflètent-elles celles obtenues via l'utilisation de *decoys* ?

Pour cette étude, nous avons choisi de comparer deux outils de criblage basés respectivement sur des algorithmes d'optimisation locale et d'optimisation de colonies de fourmis et libres d'accès : VINA[215] et PLANTS[460]. Les études ont été conduites sur 10 récepteurs nucléaires en considérant l'ensemble des structures issues de la NR-DBIND ne présentant pas de résidus manquant à proximité du site de liaison. Les résultats par structure, et combinés par ensemble de 2 et 3 structures ont été comparés. Il faut noter que l'intégration de molécules inactives nécessite d'établir une définition claire. Dans cette étude, les molécules ne présentant pas d'affinité pour la cible étudiée (« non-binders ») et celles capables de provoquer une activité contraire à celle recherchée sont considérées inactives. Par exemple, dans le cas de recherche

de molécules agonistes, les molécules *non-binders* et les molécules antagonistes sont inclues dans le jeu de données inactives. Dans un second temps, des *decoys* générés avec l'outil de génération de *decoys* de la DUD-E ont été dockés afin de comparer les performances obtenues selon la composition du jeu de données « inactives ».

3.2 Matériel et méthodes

3.2.1 Sélection et préparation des petites molécules

Les petites molécules et les protéines ont été extraites de la NR-DBIND. La NR-DBIND fournit des données d'affinité pour des petites molécules expérimentalement testées contre au moins un récepteur nucléaire, incluant des résultats négatifs et des annotations en termes de profil pharmacologique de l'interaction testée lorsque des données d'activité sont disponibles.

Dans cette étude, nous avons utilisé les molécules classées « binders » ($pIC50 \geq 7$ ou $pKi \geq 7$) et « non –binders » ($pIC50 \leq 5$ ou $pKi \leq 5$) selon la NR-DBIND. Il faut noter que la NR-DBIND recense des molécules pour lesquelles les données d'affinité sont fournies sous forme de pIC50 ou de pKi ; dans cette étude, nous avons combiné l'ensemble des molécules pour étoffer au maximum le jeu de données. Parmi les « binders », 1) les molécules dépourvues d'annotation quant à leur profil pharmacologique ont été rejetées, 2) les molécules annotées « agonistes » et jamais annotées « antagonistes » ont été considérées dans le jeu de molécules agonistes, et 3) les molécules annotées « antagonistes » et jamais « agonistes » ont été classées dans le jeu de molécules antagonistes. Pour chaque récepteur nucléaire, deux jeux de données ont été construits :

- Le jeu Ag/AntNB : les molécules agonistes constituent le jeu de données actives et les antagonistes et *non-binders* constituent le jeu de données inactives
- Le jeu Ant/AgNB : les molécules antagonistes constituent le jeu de données actives et les agonistes et *non-binders* constituent le jeu de données inactives

Pour chaque jeu de données, des *decoys* ont été générés via le générateur de *decoys* de la DUD-E : 50 *decoys* ont ainsi été générés pour chaque molécule du jeu de données actives.

L'ensemble des petites molécules a été extrait de la NR-DBIND et de la DUD-E au format SMILES. L'état de protonation majoritaire à pH7.4 a été calculé avec Marvin 17.22.0, 2017, ChemAxon (http://www.chemaxon.com). La conformation de plus basse énergie selon iCon (implémenté dans LigandSCout 4.2) a été générée et exportée au format SDF. MGTools a été utilisé pour convertir les molécules du format SDF au format PDBQT requis comme format

d'entrée d'AutoDock Vina en assignant les charges partielles de Gasteiger et les types d'atomes adéquats.

L'empreinte moléculaire MACCS a été générée pour chaque molécule et les distances de Tanimoto entre les molécules actives et molécules inactives ont été calculées. Cette même procédure a été appliquée aux molécules définies par leur squelette de Bemis Murcko fourni par la NR-DBIND.

3.2.2 Sélection et préparation des structures

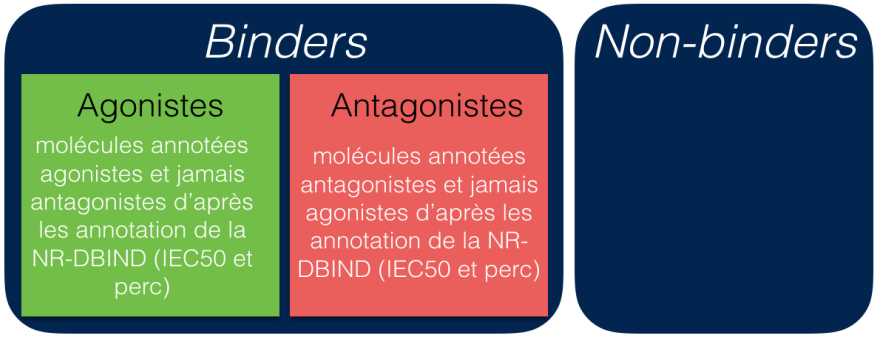


Figure 50 Organisation des données extraites de la NR-DBIND et conservées pour l'étude. Parmi les molécules capables d'interagir avec les NRs ("binders", pKi ≥ 7 ou pIC50 ≥ 7), seules les molécules strictement annotées « agonistes » ou « antagonistes » sont considérées. Les non-binders (pKi ≤ 5 ou pIC50 ≤ 5) sont systématiquement considérés inclues dans le jeu de données inactives.

Les protéines ont été sélectionnées selon les recommandations proposée par Lagarde et. al[374] : les structures co-cristallisées des NRs AR, ERα, ERβ, FXR, PPARα, PPARβ, PPARγ, PR, RXRα et TRβ répertoriées dans la NR-DBIND ont été extraites, et seules les structures agonistes-liées et antagonistes-liées ont été conservées et assignées respectivement aux jeux de données Ag/AntNB et Ant/AgNB. Les sites de liaison des petites molécules ont été estimés avec FPocket et CASTP qui calculent également un ensemble de descripteurs des cavités estimées (Cf Annexe 6). Les deux logiciels se basent sur une tesselation de Voronoï pour estimer les volumes des cavités (Cf 4.2.2). Les paramètres par défaut ont été utilisés pour ces deux logiciels, excepté le rayon maximal des sphères alpha de FPocket qui a été élevé à 6.5Å afin de mieux estimer les larges poches enfouies et localement plates. Les structures présentant des atomes manquants au niveau du site de liaison estimé par FPocket et CASTP ont été écartées. Les structures restantes ont été protonées à pH 7.4 avec PDB2PQR (Version 2.1.1). Les molécules d'eau et les hétéroatomes ont été omis.

3.2.3 Docking

Le docking a été réalisé avec AutoDock VINA 1.1.2[215] et PLANTS[460]. Dans les deux cas, le solvant a été omis.

AutoDock VINA

AutoDock VINA génère des poses de docking via un algorithme qui consiste en une succession d'étapes de mutations stochastiques et d'optimisations locales qui tendent à minimiser la somme des énergies intra- et intermoléculaires et une sélection de pose basée sur un critère de Metropolis. A chaque étape, l'algorithme Broyden-Fletcher-Goldfarb-Shanno (BFGS) est utilisé pour l'optimisation locale. AutoDock VINA évalue les poses avec sa propre fonction de score apprise sur des données d'interactions préférentielles entre les couples ligand/protéine (données basées sur la connaissance) et des mesures d'affinité expérimentales (données empiriques). Cette fonction de score simplifiée ne possède pas de terme électrostatique, assimile les potentiels de liaisons hydrogènes à des sphères symétriques et considère les hydrogènes de façon implicite. Les paramètres par défaut ont été utilisés (*exhaustiveness* = 8) et 20 poses par molécule ont été générées.

PLANTS

PLANTS génère les poses de docking en utilisant un algorithme de colonies de fourmis (ACO). Pour chaque degré de liberté i du ligand (translations (3), rotations (3), et torsions (x)), un vecteur discrétisé Ti comprenant autant de points que de valeurs j adoptables est construit (360 valeurs pour les angles de rotation et de torsion et un nombre de valeurs variable et dépendant de la taille du site de liaison pour les libertés de translation). La fourmi (virtuelle) construit un ensemble de possibilités en choisissant une valeur j par degré de liberté avec une probabilité initialement dépendante d'informations heuristiques, ou piste de phéromone. A noter que les informations heuristiques initiales ont été retirées des dernières versions de PLANTS du fait de leur faible influence sur les performances finales d'échantillonage. Chaque solution (pose) est optimisée avec l'algorithme de recherche locale simplexe, et la meilleure solution à l'issue de chaque itération sont acceptée si un critère de diversité et de score parmi les meilleures solutions proposées est rencontré. La solution de chaque itération permet d'augmenter la valeur de phéromone des composentes de la solution (degrès de liberté i choisis par la fourmi) et de diminuer les valeurs associées aux autres solutions non retenues. L'algorithme converge vers différentes solutions après plusieurs itérations. PLANTS évalue les poses avec sa propre

fonction de score empirique inspirée de PLP[461] pour modéliser les interactions stériques intramoléculaires, de CHEMSCORE (GOLD) pour les interactions hydrogènes, ainsi que d'une fonction d'énergie intramoléculaire prenant en compte les clashs stériques et les potentiels de torsion.

Le site de liaison est défini par une sphère de rayon et de centre arbitrairement imposé par l'utilisateur. Nous avons imposé un rayon de 20Å et la génération de 20 poses distantes de minimum 2Å les unes des autres. Les paramètres par défaut ont été utilisés (fonction de score : chemplp).

Approches structure unique (*single docking*) vs ensemble de structures (*ensemble docking*)

Nous avons étudié l'importance de la considération de la flexibilité des NRs en comparant les résultats de docking obtenus sur une structure seule aux résultats obtenus sur un ensemble de 2 à 3 structures. Pour ce faire, toutes les combinaisons de 2 à 3 structures de chaque jeu de données ont été considérées. Pour chaque combinaison, le meilleur score observé entre une molécule et l'une des structures de l'ensemble a été considéré. L'aire sous la courbe ROC (AUC) a été utilisée comme métrique de performance.

3.2.4 Comparaison de l'utilisation de molécules inactives validées expérimentalement et de decoys générés par la DUD-E

Les résultats obtenus en docking sur structure seule en utilisant les jeux de données issus de la NR-DBIND et ne comportant que des molécules inactives expérimentalement validées ont été comparés aux résultats obtenus en suivant le même protocole mais en remplaçant les molécules inactives par des *decoys* issus de la DUD-E. Les ratios molécules actives/inactives observés dans les données de la NR-DBIND étant très éloignés de ceux observés dans la DUD-E, deux stratégies ont été adoptées :

1) Les ratios des jeux de données de la NR-DBIND ont été conservés : les AUCs ont été calculées en remplaçant chaque molécule inactive validée expérimentalement par un de *decoy* de la DUD-E. Pour cela, nous avons généré *n decoys* par molécules active qui ont été regroupés en éliminant les redondants. Un protocole de tirage aléatoire a ensuite été réalisé afin d'obtenir exactement le même nombre de *decoys* que de molécules inactives : 25 tirages ont été effectués par *bootstrap* et l'AUC moyenne sur ces 25 tirages a été retenue de sorte à éviter l'introduction de biais lié au faible nombre de molécules choisies.
2) Les ratios de la DUD-E ont été conservés : une AUCs a été calculée en considérant le jeu de données d'actifs et le jeu de données d'inactifs constitué de decoys générés avec l'outils de

génération de la DUD-E (50 *decoys* par molécule active), puis une autre AUCs a été calculée en remplaçant *m decoys* dans le jeu de données d'inactifs par les *m* inactifs expérimentalement validés issus de la NR-DBIND.

Pour chacune des méthodes, une ΔAUC a été calculé, correspondant à la différence d'AUC obtenu par structure sur le jeu de données contenant exclusivement des *decoys* et celui contenant des molécules inactives validées expérimentalement dans le jeu d'inactifs. Une valeur de ΔAUC positive signifie que l'utilisation de *decoys* conduit à de meilleures performances que l'utilisation de molécule inactives, alors qu'une valeur négative signifie que l'utilisation de *decoys* conduit à de moins bonnes performances que l'utilisation de molécule inactives. Une $\overline{\Delta AUC}$ a été calculée, représentant la moyenne des ΔAUC obtenues sur l'ensemble des structures d'un NR.

3.2.5 Comparaison des espaces chimiques des jeux de données issus de la NR-DBIND et des *decoys* générés par la DUD-E

Une analyse en composantes principales (ACP) a été réalisée sur les molécules des jeux de données issus de la NR-DBIND et les *decoys* générés par la DUD-E, tous décrits par 11 descripteurs (poids moléculaire, nombre de donneur de liaison hydrogène, nombre d'accepteur de liaison hydrogène, cLogP, surface accessible au solvant, indice de flexibilité moléculaire, indice de complexité moléculaire, nombre de liaison rotatives, nombre de cycles aromatiques, nombre d'atomes impliqués dans un cycle aromatique, surface polaire accessible au solvant). Les descripteurs et l'ACP ont été calculés avec DataWarrior.(Version 4.7.2).

3.3 Résultats

3.3.1 Contenu de la banque de données traitée

Tableau 16 Contenu des jeux de données extraits de la NR-DBIND.

	Molécules				Structures PDBs		Ratios molécules actives/inactives	
	Agonistes	**Antagonistes**	**Non-binders**	**Total**	**Agonistes-liées**	**Antagonistes-liées**	**Ag/ AntNB**	**Ant/ AgNB**
AR	224	451	137	812	29	0	0,38	1,25
ERα	18	160	180	358	4	2	0,05	0,81

ERβ	73	101	69	243	18	0	0,43	0,72
PPARα	135	0	226	361	4	1	0,60	-
PPARβ	175	8	141	324	13	0	1,17	-
PPARγ	383	0	170	553	24	0	2,25	-
PR	195	156	98	449	5	5	0,77	0,53
FXR	17	1	35	53	22	0	0,47	-
TRβ	32	10	1	43	10	0	2,91	0,30
RXRα	27	63	19	109	32	1	0,33	1,37
VDR	2	7	1	10	14	0	-	-

La NR-DBIND a été filtrée de sorte à en extraire des molécules agonistes, antagonistes et *non-binders*. Les annotations proposées par la NR-DBIND et extraites de tests de bio-activité ont permis d'inclure les molécules capables de se fixer au récepteur étudié (pIC50 ≥ 7 ou pKi ≥7) et estimées agonistes ou antagonistes. Les modulateurs des NRs pouvant exercer une activité tissus-dépendante, certaines molécules possèdent la double annotation « agoniste » et « antagoniste » dans la NR-DBIND. Ces molécules jugées ambigües ont été retirées de l'étude. Pour chaque NR étudié, les structures répertoriées dans la NR-DBIND ont été téléchargées depuis la PDB et nettoyées de sorte à exclure les structures présentant des résidus manquant à proximité du site de liaison. Le nombre de molécules et de structures conservées dans cette étude est détaillé dans le Tableau 16. Les ratios molécules actives/inactives ont été calculés pour chaque jeu de données Ag/AntNB et Ant/AgNB comportant plus de 10 molécules actives. Dans les banques de données d'évaluation classiquement utilisées (DUD-E, MUV et DEKOIS), les ratios varient de 0.002 à 0.03. Dans cette étude, des ratios plus élevés compris entre 0.05 et 2.91 et entre 0.30 et 2.33 sont observés respectivement dans les jeux Ag/AntNB et Ant/AgNB. Les ratios utilisés étant censés être représentatifs des ratios observés dans des cas réels (de l'ordre de 0,1 à 4%[44]), seul le jeu de données Ag/AntNB de ERα semble être adapté à une étude robuste. Afin d'étoffer l'étude nous avons arbitrairement choisi de considérer tous les jeux de données présentant des ratios inférieurs à 1. Ce premier critère combiné à la nécessité d'avoir au minimum 1 structure pour effectuer le docking réduit le jeu de données global à 7 jeux Ag/AntNB (AR, ERα, ERβ, FXR, PPARα, PR, RXRα) et 2 jeux Ant/AgNB (ERα, PR).

3.3.2 Comparaison des performances de VINA et PLANTS

3.3.2.1 Distribution des scores en fonction du profil pharmacologique du ligand co-cristallisé

Selon la petite molécule avec laquelle ils interagissent, les NRs adoptent des conformations différentes (Cf 6.5). Par conséquent les sites de liaison observés dans les structures agoniste-liées et antagoniste-liées diffèrent en termes de volume, d'ouverture et de propriétés physicochimiques et il est important de les distinguer lors d'études de docking. Dans un premier temps, nous avons souhaité étudier la capacité de AutoDock VINA et PLANTS à associer de

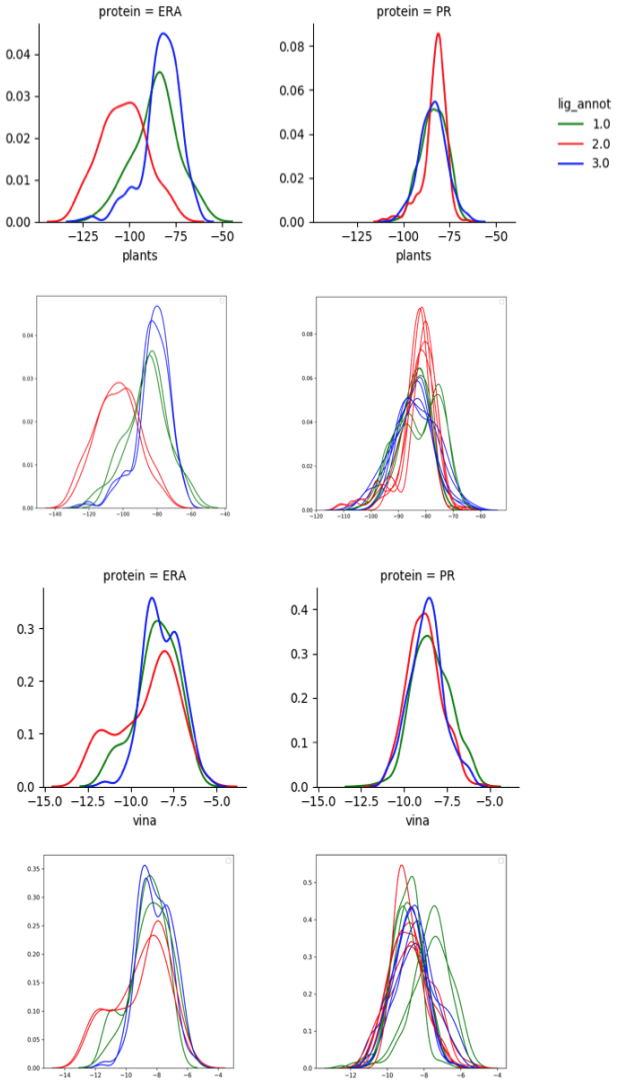


Figure 51 Distributions des scores de docking obtenus pour les structures antagonistes-liées en utilisant PLANTS (4 graphiques du haut) et AutoDock VINA (4 graphiques du bas). Chaque courbe représente la distribution des scores de docking pour un type de molécule (antagonistes : rouge, agonistes : vert, et *non-binders* : bleu) et pour l'ensemble des structure (haut) ou pour chaque structure (bas).

meilleurs scores aux petites molécules de même profil pharmacologique que le ligand co-cristallisé dans la structure utilisée. Les objectifs étaient 1) de s'assurer que les molécules agonistes et antagonistes sont correctement discriminées et 2) d'observer si les outils sont capables de discriminer les *non-binders*.

Les distributions des scores par structure et par protéine montrent que ni AutoDock VINA ni PLANTS ne discriminent de manière stricte les molécules agonistes des molécules antagonistes aussi bien dans les structures agonistes-liées qu'antagonistes-liées (Figure 51 et Figure 52). Néanmoins, les scores moyens attribués aux agonistes en utilisant des structures agonistes-liées sont meilleurs que ceux attribués aux antagonistes pour 3 NRs/7 (AR, ERβ et FXR) avec AutoDock VINA. Les scores moyens attribués aux antagonistes en utilisant des structures antagonistes-liées sont meilleurs que ceux attribués aux agonistes pour 1 NR/2 (ERα) avec PLANTS et AutoDock VINA. Dans les structures agonistes-liées, les *non-binders* sont associés à de moins bons scores moyens que les agonistes dans 2 cas avec AutoDock VINA (AR et FXR) et 1 seul cas avec PLANTS (RXRα). Dans les structures antagonistes-liées, les *non-binders* sont associés à de moins bons scores moyens que les antagonistes avec AutoDock VINA et PLANTS dans les structures de ERα seulement. Ces résultats montrent une difficulté à discriminer les molécules de profil pharmacologique identique au ligand co-cristallisé des autres molécules. La prise en compte du profil pharmacologique du ligand co-cristallisé n'est donc pas suffisante pour discriminer les molécules actives des molécules inactives dans les jeux Ag/AntNB et Ant/AgNB.

3.3.2.2 Performances globales obtenues en docking sur structure seule

Le docking a été réalisé sur chacune des structures retenues dans l'étude. Les performances de PLANTS et de AutoDock VINA ont été évaluées en terme d'AUC maximale et d'AUC moyenne obtenue pour chaque protéine et chaque jeu de données (Ag/AntNB et Ang/AgNB) (Tableau 17). Concernant les jeux Ag/AntNB, dans 5/7 cas (AR, ERα, ERβ, FXR et RXRα) AutoDock VINA permet d'obtenir de meilleures AUCs maximales et moyennes que PLANTS. Dans les 2/7 cas restants, PLANTS présente de meilleures performances en terme d'AUCs maximales et moyennes (PPARα, PR). Concernant les jeux Ant/AgNB, PLANTS présente de meilleures performances pour ERα alors que AutoDock VINA présente de meilleures performances pour PR. En considérant les meilleures performances atteintes pour chacune des protéines avec AutoDock VINA et PLANTS, des AUCs comprises entre 0.59 et 0.95 sont obtenues pour les jeux de Ag/AntNB avec une moyenne de 0.74, et des AUCs comprises entre 0.70 et 0.90 sont obtenues pour les jeux de Ant/AgNB avec une moyenne de 0.8. Au total des

AUCs > 0.7 sont obtenues pour 5/7 protéines sur les jeux Ag/AntNB et pour 2/2 protéines sur les jeux Ant/AgNB. Dans deux cas (PR-Ag/AntNB et PR-Ant/AgNB), les structures associées à l'AUC maximale selon AutoDock VINA et PLANTS sont les mêmes.

Tableau 17 Performances de criblage obtenues avec PLANTS et AutoDock VINA en docking sur structure seule en termes d'AUC. La structure pdb associée à l'AUC maximale obtenue est indiquée (max : AUC maximale, min : AUC minimale, mean : AUC moyenne, std : écart-type)

		PLANTS					AutoDock VINA				
		max	pdb	min	mean	std	max	pdb	min	mean	std
Ag/AntNB	AR	0,69	1xow	0,43	0,56	0,06	**0,72**	2pir	0,52	**0,66**	0,05
	ERα	0,56	2yja	0,42	0,49	0,06	**0,59**	1x7e	0,45	**0,53**	0,07
	ERβ	0,46	1x76	0,27	0,38	0,07	**0,76**	4j24	0,32	**0,63**	0,11
	FXR	0,87	5q1d	0,08	0,62	0,26	**0,95**	5q0o	0,38	**0,77**	0,21
	PPARα	**0,75**	2p54	0,71	**0,73**	0,02	0,67	4bcr	0,6	0,63	0,03
	PR	**0,63**	1sr7	0,36	**0,42**	0,12	0,43	1sr7	0,26	0,32	0,06
	RXRα	0,66	2zxz	0,31	0,41	0,08	**0,79**	1fby	0,46	**0,67**	0,09
Ant/AgNB	ERα	**0,90**	1xp1	0,89	**0,89**	0,01	0,63	1xp1	0,61	0,62	0,01
	PR	0,56	3zra	0,35	0,46	0,08	**0,70**	3zra	0,49	**0,59**	0,10

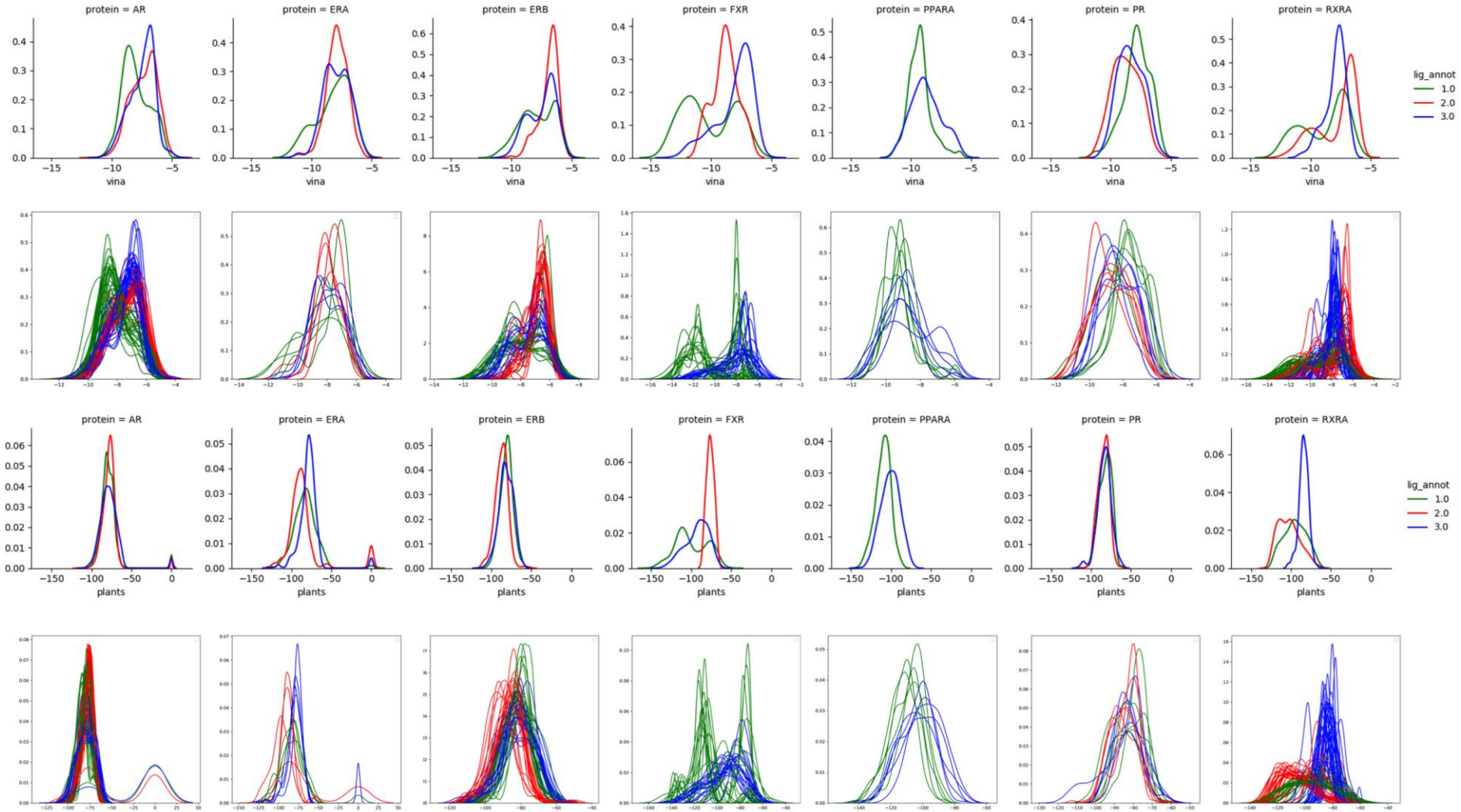


Figure 52 Distributions des scores de docking obtenus pour les structures agonistes-liées, en utilisant les logiciels AutoDock VINA (14 graphiques du haut) et PLANTS (14 graphiques du bas). Chaque courbe représente la distribution des scores de docking pour un type de molécules (antagonistes : rouge, agonistes : vert, et non-binders : bleu) et pour l'ensemble des structure (haut) ou pour chaque structure (bas).

3.3.2.3 Comparaison du docking sur structure seule avec le docking d'ensemble

Les résultats du docking sur structure seule ont été comparés aux résultats obtenus sur les ensembles de 2 à 3 structures (Tableau 18, Figure 53 et Figure 54). Le docking d'ensemble avec PLANTS ne permet d'améliorer les performances de docking que dans 1/9 cas (FXR-Ag/AntNB). Dans ce cas précis, l'AUC est améliorée de +0.02 avec le docking d'ensemble sur 2 et 3 structures par rapport au docking sur structure seule. Avec AutoDockVINA, le docking d'ensemble permet d'améliorer les performances dans 5/9 (AR/ERβ/FXR/RXRα-Ag/AntNB ; ERα-Ant/AgNB). Dans ces 5 cas, l'AUC est améliorée de +0.03 à +0.11 avec une moyenne de +0.07. Dans 2/5 cas seulement le docking d'ensemble sur 3 structures améliore les performances obtenues en docking d'ensemble sur 2 structures (AR-Ag/AntNB : +0.02 et RXRα-Ag/AntNB : +0.03). Les moyennes des AUCs obtenues avec AutoDock VINA augmentent entre le docking sur 1 structure et le docking sur 3 structures dans 7/9 cas (AR/ERα/ERβ/FXR/PR/RXRα-Ag/AntBN : +0.02 à +0.12 ; ERα-Ant/AgNB : +0.05) et diminuent faiblement dans 2/9 cas (PPARα-Ag/AntBN : -0.01 et PR-Ant/AgNB : -0.03) alors que l'écart type diminue dans 8/9 cas (excepté pour PR-Ag/AntNB : +0.00). Si le docking sur structure seule permet d'obtenir de meilleures AUCs que le docking d'ensemble dans 2/9 cas, le docking d'ensemble apparaît légèrement plus performant en utilisant AutoDock VINA. Avec PLANTS, le docking d'ensemble améliore rarement la moyenne des AUCs obtenues (3/9 cas) ; l'écart type diminue en revanche avec le nombre de structures utilisées.

Nous avons ensuite cherché si quelques descripteurs simples permettaient d'orienter vers un choix de structures assurant de bonnes performances. Pour cela, un ensemble de descripteurs des sites de liaison de chaque protéine a été calculé avec les outils FPocket et CASTP (Cf Annexe 6). Les structures ont été classées en 3 groupes selon la valeur du descripteur étudié (groupe 1 : x < valeur moyenne – écart type, groupe 2 : valeur moyenne – écart type < x < valeur moyenne + écart type, groupe 3 : x > valeur moyenne – écart type). Nous avons comparé les distributions des AUCs de chaque groupe. Les résultats (Cf Annexe 7) n'ont pas permis d'identifier des descripteurs communs aux différents NRs et associés à de bonnes performances. Néanmoins, les structures d'AR ayant les plus grandes surfaces accessibles au solvant selon CASTP (Vol_sa) sont associées à de meilleure AUC avec AutoDock VINA ; les structures d'ERA associées à de gros volumes (Volume/Volume moyen des sphères alpha/Densité des sphères alpha – Fpocket, Vol_sa – CASTP) ont de meilleures AUCs avec AutoDock VINA et celles avec de nombreuses sphères alpha et chargées (Nombre de sphère alpha/ score de charge – FPocket) ont de meilleures AUCs avec PLANTS ; les structures de FXR associées à peu de

sphères alpha exposées au solvant et à une faible flexibilité ont de meilleures AUCs avec AutoDock VINA et PLANTS.

Tableau 18 Performances de criblage obtenues avec PLANTS et AutoDock VINA en docking sur structure seule et en docking d'ensemble (2 et 3 structures) et données en terme d'AUC.

			PLANTS					AutoDock VINA				
			max	pdb	min	mean	std	max	pdb	min	mean	std
Ag/AntNB	AR	single	**0,69**	**1xow**	0,43	0,56	0,06	0,72	2pir	0,52	0,66	0,05
		ensemble_2	0,69	1xow, 2ama	0,47	0,56	0,06	0,76	1t5z, 2piw	0,6	0,71	0,03
		ensemble_3	0,68	1xow, 1xj7, 2ama	0,46	0,56	0,05	**0,78**	**2piq, 2pip, 2amb**	0,64	0,73	0,02
	ERα	single	**0,56**	**2yja**	0,42	0,49	0,06	**0,59**	**1x7e**	0,45	0,53	0,07
		ensemble_2	0,52	2yja, 1x7e	0,42	0,46	0,04	0,59	1a52, 1x7e	0,48	0,55	0,04
		ensemble_3	0,49	1g50, 2yja, 1x7e	0,42	0,45	0,03	0,58	1a52, 2yja, 1x7e	0,55	0,56	0,01
	ERβ	single	**0,46**	**1x76**	0,27	0,38	0,07	0,76	4j24	0,32	0,63	0,11
		ensemble_2	0,45	1x76, 1zaf	0,25	0,35	0,05	**0,81**	**3oll, 4j24**	0,43	0,7	0,07
		ensemble_3	0,44	1x76, 1u9e, 1zaf	0,24	0,34	0,05	0,81	1zaf, 3oll, 4j24	0,46	0,72	0,06
	FXR	single	0,87	5q1d	0,08	0,62	0,26	0,95	5q0o	0,38	0,77	0,21
		ensemble_2	**0,89**	**5q12, 5q1d**	0,1	0,72	0,17	**0,98**	**5q1a, 5q1h**	0,29	0,85	0,16
		ensemble_3	0,89	5q12, 3ruu, 5q1d	0,12	0,75	0,12	0,98	5q1a, 5q12, 5q1h	0,29	0,89	0,1
	PPARα	single	**0,75**	**2p54**	0,71	0,73	0,02	**0,67**	**4bcr**	0,6	0,63	0,03
		ensemble_2	0,75	4bcr, 2p54	0,71	0,73	0,02	0,66	4bcr, 2p54	0,6	0,63	0,02
		ensemble_3	0,74	4bcr, 3et1, 2p54	0,71	0,73	0,01	0,63	4bcr, 1k7l, 2p54	0,61	0,62	0,01
	PR	single	**0,63**	**1sr7**	0,36	0,42	0,12	**0,43**	**1sr7**	0,26	0,32	0,06
		ensemble_2	0,6	1a28, 1sr7	0,34	0,44	0,1	0,4	1sqn, 1sr7	0,25	0,32	0,07

		ensemble_3	0,54	1zuc, 1a28, 1sr7	0,34	0,45	0,08	0,39	1e3k, 1sqn, 1sr7	0,26	0,34	0,06
	RXRα	single	0,66	2zxz	0,31	0,41	0,08	0,79	1fby	0,46	0,67	0,09
		ensemble_2	**0,67**	**3e94, 2zxz**	0,27	0,38	0,05	0,87	4pp5, 1mv9	0,5	0,69	0,08
		ensemble_3	0,65	3e94, 4oc7, 2zxz	0,26	0,36	0,04	**0,9**	**1fby, 3oap, 1mv9**	0,5	0,7	0,07
Ant/AgNB	ERα	single	**0,90**	**1xp1**	0,89	0,89	0,01	0,63	1xp1	0,61	0,62	0,01
		ensemble_2	0,90	1xp9, 1xp1	0,90	0,90	-	**0,67**	**1xp9, 1xp1**	0,67	0,67	0,00
		ensemble_3	-	-	-	-	-	-	-	-	-	-
	PR	single	**0,56**	**3zra**	0,35	0,46	0,08	**0,70**	**3zra**	0,49	0,59	0,10
		ensemble_2	0,54	3zra, 3zrb	0,38	0,45	0,05	0,67	3zra, 3zrb	0,49	0,55	0,05
		ensemble_3	0,49	3zra, 3zrb, 2ovh	0,40	0,45	0,03	0,58	3zra, 3zrb, 2ovm	0,49	0,53	0,03

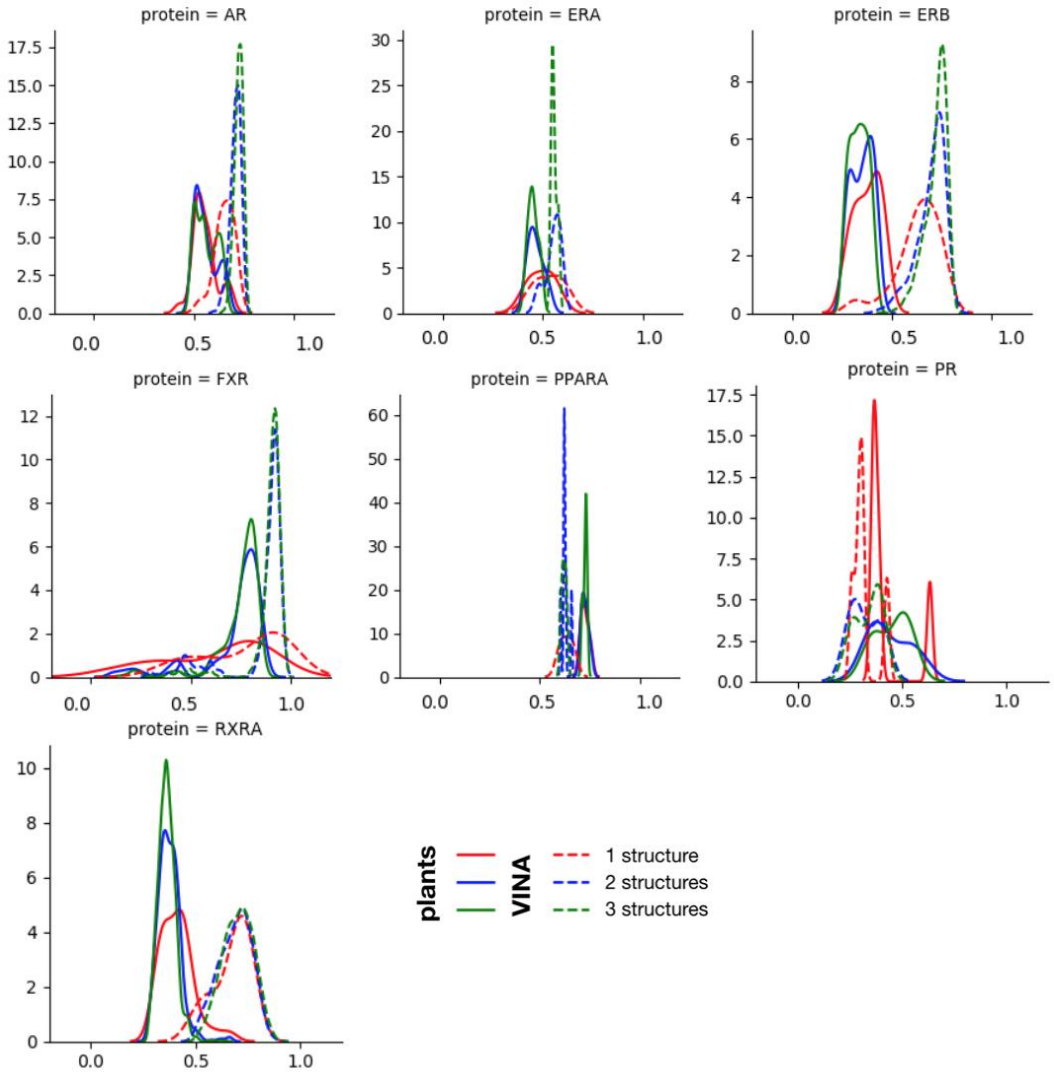


Figure 53 Distributions des AUCs obtenues avec PLANTS et AutoDock VINA sur les différents jeux Ag/AntNB.

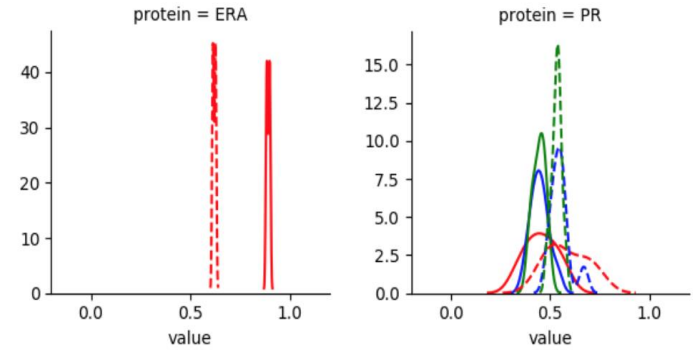


Figure 54 Distributions des AUCs obtenues avec PLANTS et AutoDock VINA sur les différents jeux Ant/AgNB.

3.3.3 Comparaison de l'utilisation d'inactifs validés et de decoys générés par la DUD-E

Afin de comparer les performances obtenues en évaluant les performances sur les jeux Ag/AntNB et Ant/AgNB à celles obtenues en utilisant des *decoys* de la DUD-E, nous avons adopté deux stratégies. La première consiste à remplacer les inactifs issus de la NR-DBIND par des *decoys* (Ag/Decoys et Ant/Decoys) en respectant les ratios des données issues de la NR-DBIND (Ag/AntNB et Ant/AgNB). La seconde stratégie consiste à conserver les ratios actifs/decoys de la DUD-E (Ag/Decoys-dude et Ant/Decoys-dude). Faute de temps, cette étude n'a été conduite qu'avec AutoDock VINA qui a montré de meilleures capacités discriminatoires sur les jeux de données issus de la NR-DBIND.

3.3.3.1 Méthode 1

Les AUC_{moy} obtenues en utilisant des *decoys* issus de la DUD-E (Ag/Decoys et Ant/Decoys) ont été comparées à celles obtenues avec les jeux de données Ag/AntNB et Ant/AgNB (Figure 56 et Figure 55). Les AUCs obtenues dans les structures de 5/7 jeux Ag/AntNB et 1/2 jeu Ant/AgNB sont comparables avec les AUC moyennes obtenues avec les jeux Ag/Decoys et

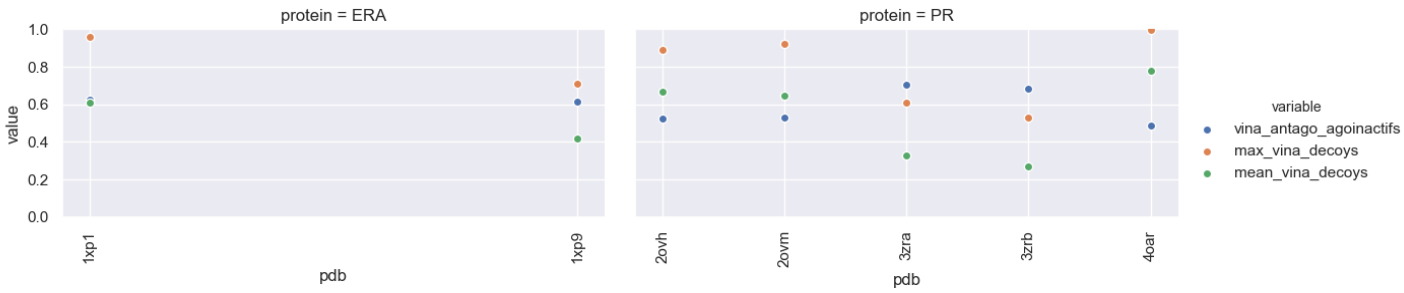


Figure 55 Comparaison des AUC moyennes obtenues en utilisant les 25 jeux de *decoys* générés par la DUD-E par NR (correspondant aux 25 tirages) en respectant les ratios molécules actives/inactives observés dans la NR-DBIND (vert) à l'AUC maximale obtenue parmi les 25 tirages (orange) et à l'AUC obtenue en utilisant les jeux Ant/AgNB issus de la NR-DBIND.

Ant/Decoys. Pour chaque NR, la moyenne $\overline{\Delta AUC}$ des ΔAUC obtenues sur chacune des structures est toujours comprise entre -0.1 et < 0.1 ; une valeur de $\overline{\Delta AUC}$ positive signifiant que l'utilisation de *decoys* conduit à de meilleures performances que l'utilisation des molécules issues de la NR-DBIND, et des valeurs de $\overline{\Delta AUC}$ négatives signifiant que l'utilisation de *decoys*

conduit à de moins bonnes performances que l'utilisation des molécules issues de la NR-DBIND ($\overline{\Delta AUC}$ - AR(Ag) = -0.02, $\overline{\Delta AUC}$ - ERα(Ag) = +0.06, $\overline{\Delta AUC}$ - ERβ(Ag) = +0.05, $\overline{\Delta AUC}$ - FXR(Ag) = -0.03, $\overline{\Delta AUC}$- PR(Ant) = -0.05). Dans 2/9 cas, de meilleures performances sont atteintes en utilisant les *decoys* ($\overline{\Delta AUC}$ - PPARα(Ag) = +0.16, MoyΔAUC - PR(Ag) = +0.12), alors que dans 2/9 de meilleures performances sont atteintes en utilisant les molécules

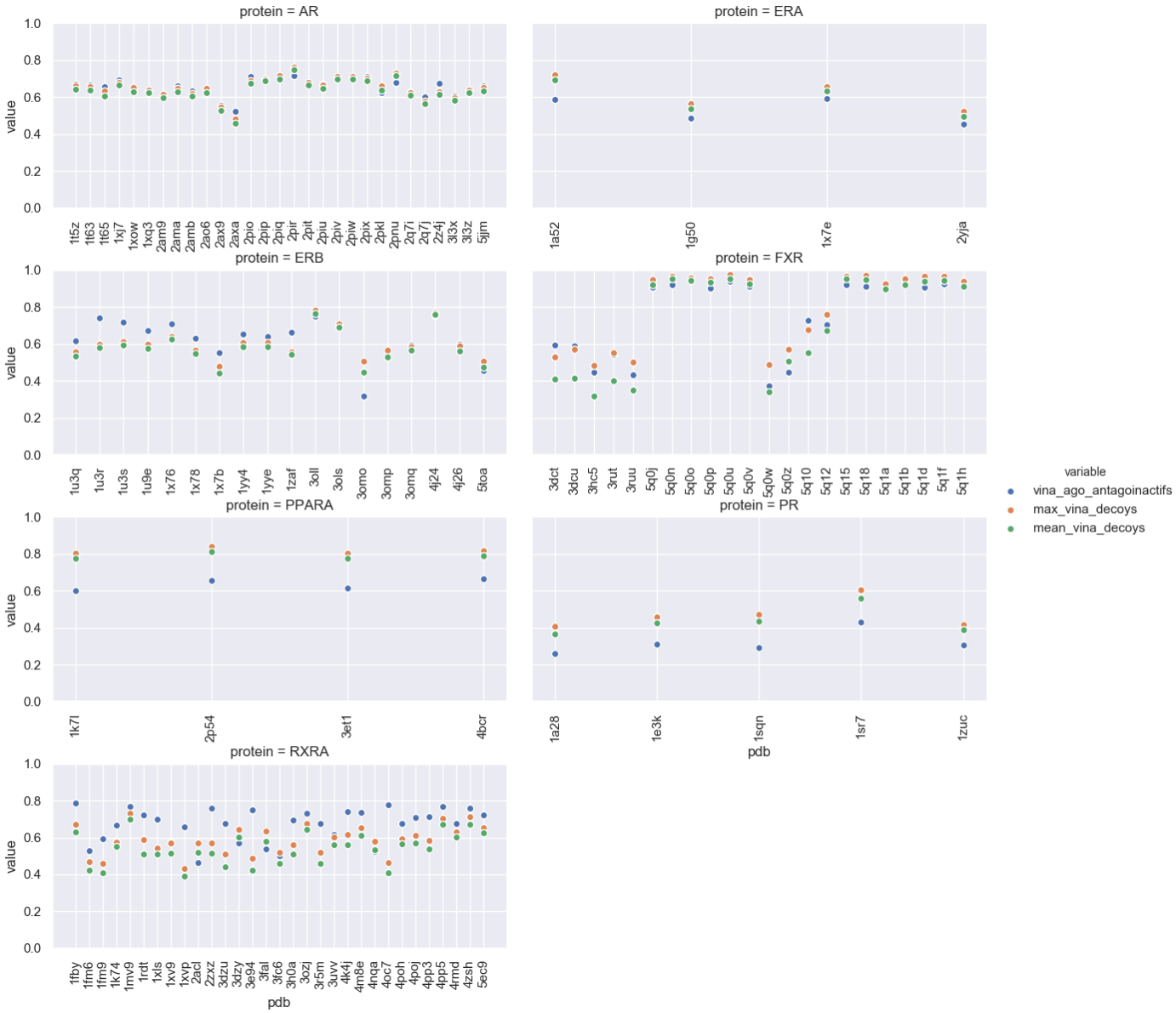


Figure 56 Comparaison des AUC moyennes obtenues en utilisant les 25 jeux de *decoys* générés par la DUD-E par NR (correspondant aux 25 tirages) en respectant les ratios molécules actives/inactives observés dans la NR-DBIND (vert) à l'AUC maximale obtenue parmi les 25 tirages (orange) et à l'AUC obtenue en utilisant les jeux Ag/AntNB issus de la NR-DBIND.

issues de la NR-DBIND ($\overline{\Delta AUC}$ - RXRα(Ag) = -0.13, $\overline{\Delta AUC}$ - ERα(Ant) = -0.11). Si les $\overline{\Delta AUC}$ sont systématiquement > -0.13 et < 0.16, des ΔAUC plus importants sont observées par structure (ex. ΔAUC - ERβ(Ag-3omo) = -0.16, ΔAUC - PPARAα(Ag-1k7l) = +0.18, ΔAUC - RXRα(Ag-4oc7) = -0.36, ΔAUC - PR(Ant-3zrb) = +0.29).

3.3.3.2 Méthode 2

Les AUC obtenues en utilisant l'ensemble des *decoys* issus de la DUD-E (Ag/Decoys-dude et Ant/Decoys dude) ont été comparées à celles obtenues avec les jeux de données Ag/AntNB-Decoys-dude et Ant/AgNB- Decoys-dude (Figure 58 et Figure 57). Des différences d'AUC négligeables sont observées pour chaque structure de chaque NR. Il est à noter que les jeux de données Ag/AntNB- Decoys-dude et Ant/AgNB- Decoys-dude possèdent des ratios AgNB/decoys et AntNB/decoys très faibles (< 0.7, excepté pour ERA(Ag) = 0.61).

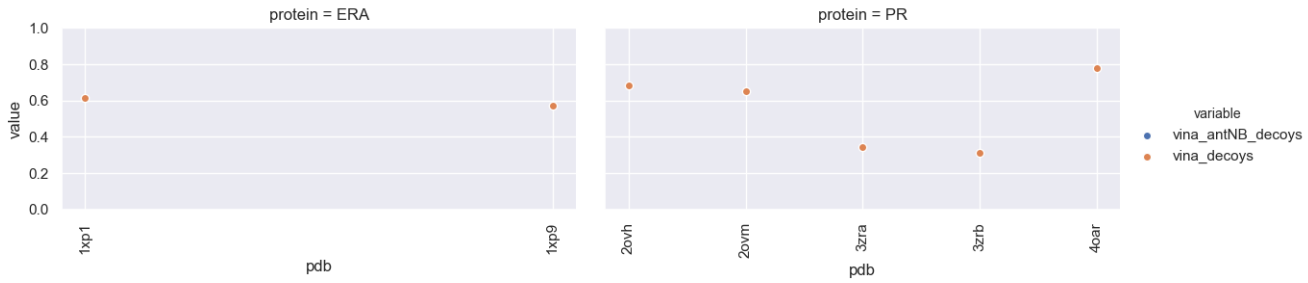


Figure 57 Comparaison des AUC obtenues en utilisant des *decoys* générés par la DUD-E par NR en respectant les ratios molécules actives/inactives observés dans la DUD-E (orange) à l'AUC obtenue en utilisant les jeux Ant/ AgNB-Decoys (bleu).

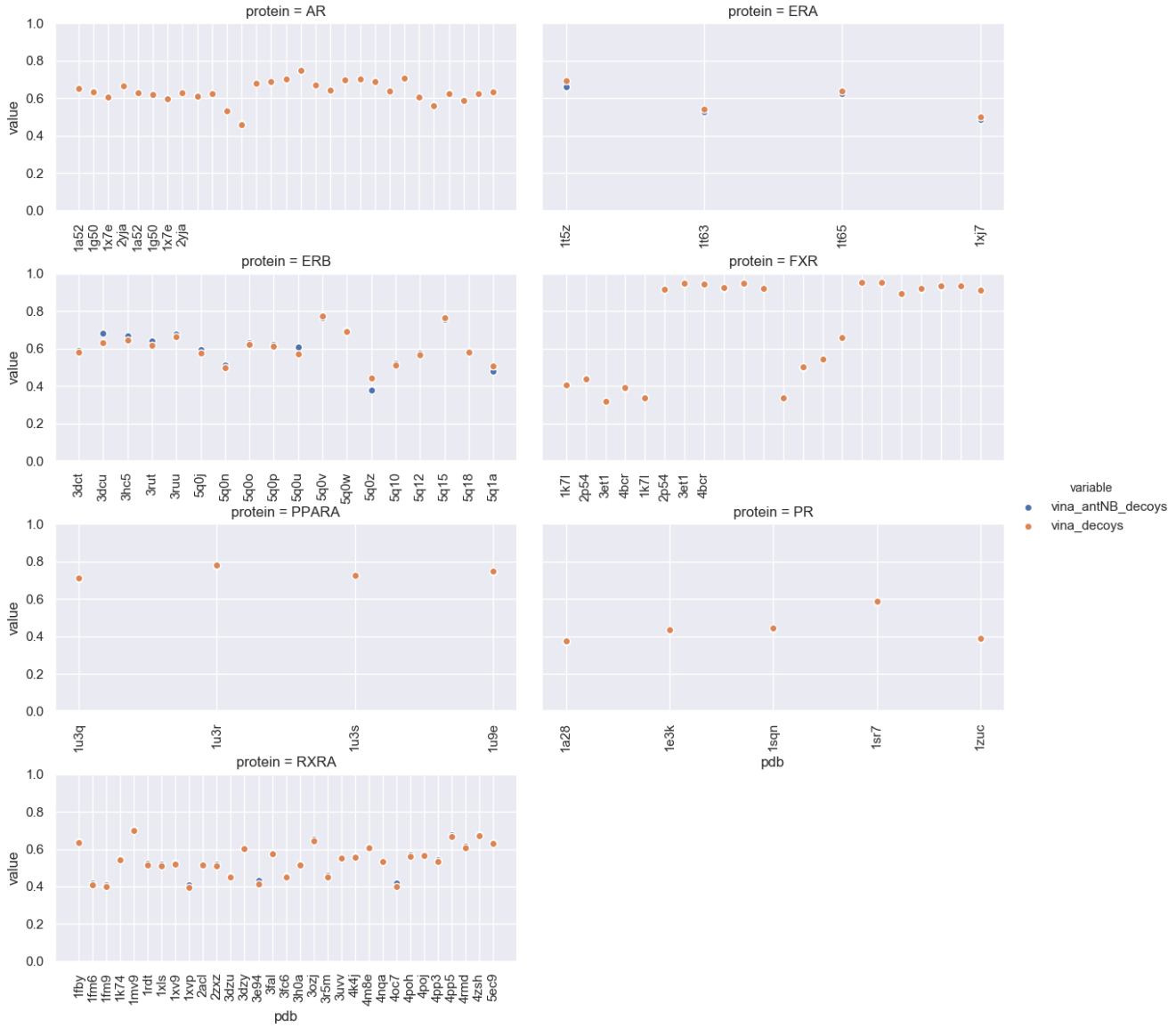


Figure 58 Comparaison des AUC obtenues en utilisant des *decoys* générés par la DUD-E par NR en respectant les ratios molécules actives/inactives observés dans la DUD-E (orange) à l'AUC obtenue en utilisant les jeux Ag/ AntNB-Decoys (bleu).

3.4 Discussion

3.4.1 Capacité de discrimination des molécules actives et inactives par PLANTS et AutoDock VINA

Les informations d'affinité, d'activité et de structure sur 28 récepteurs nucléaires contenus dans la NR-DBIND nous ont permis d'étudier la capacité de discrimination Ag/AntNB pour 7 récepteurs nucléaires et de discrimination Ant/AgNB pour 2 récepteurs nucléaires, les données

étant peu abondantes pour les autres récepteurs. Les résultats de docking montrent que AutoDock VINA et PLANTS ne sont pas plus capables de discriminer les molécules agonistes des molécules antagonistes que les molécules agonistes des molécules inactives dans les jeux Ag/AntNB. Il en est de même dans les jeux Ant/AgNB. Dans 8/9 cas (6/7 Ag/AntNB ; 2/2 Ant/AgNB), les meilleurs résultats obtenus avec PLANTS présentent des AUCs > 0.56 ; l'AUC moyenne n'étant > 0.5 que dans 4/9 cas (3/7 Ag/AntNB ; 1/2 Ant/AgNB). Avec AutoDock VINA, dans 8/9 cas, les meilleurs résultats présentent des AUCs > 0.59 (6/7 Ag/AntNB ; 2/2 Ant/AgNB) ; l'AUC moyenne étant > 0.5 dans 8/9 cas (6/7 Ag/AntNB ; 2/2 Ant/AgNB). AutoDock VINA présente donc de meilleures performances globales que PLANTS dans la discrimination de molécules actives/inactives. Le docking d'ensemble améliore peu les performances obtenues aves les deux outils ; avec AutoDock VINA, il permet toutefois d'obtenir des AUCs moyennes légèrement supérieures au docking sur structure seule ainsi que des écart types plus faibles. L'utilisation de plusieurs structures n'est donc pas systématiquement optimale, mais dans un cas de docking à l'aveugle sur un nombre restreint de structures, le docking d'ensemble permet d'avoir des résultats plus fiables que le docking sur structure seule.

3.4.2 Recommandations de docking par protéine

Aucune tendance aussi bien dans le choix de l'outil et du protocole de docking que dans les propriétés des sites de liaison ne permet d'orienter vers un protocole et un choix de structure qui favorise systématiquement la discrimination de molécules actives et inactives pour l'ensemble des NRs. Si le docking d'ensemble avec AutoDock VINA semble globalement plus adapté dans la majorité des cas, il présente de faibles performances dans 2 cas (ERα - Ag/AntNB : AUCs ≤ 0.59, PR - Ag/AntNB : AUCs ≤ 0.43). Dans 1/2 cas, PLANTS permet d'améliorer ces performances (ERα - Ag/AntNB : AUC = 0.63 (+0.20)). Il faut noter que le faible gain de performance apporté par le docking d'ensemble dans certains cas ne reflète pas une nécessité formelle de prendre en compte la flexibilité des NRs lors du docking des molécules issues de la NR-DBIND. Les conditions pour obtenir les meilleures AUCs par protéine sont regroupées dans le Tableau 19. On observe des AUCs > 0.70 dans 7/9 cas, ce qui révèle que les outils dont nous disposons sont globalement capables de discriminer les molécules de actives des molécules inactives issues de la NR-DBIND.

Tableau 19 Conditions de docking associées aux meilleures performances obtenues par jeu de données.

	Protéine	Outils	AUC maximale	PDB(s)
Agonistes vs. Antagonistes + inactifs	AR	VINA	0,78	2piq, 2pip, 2amb
	ERα	VINA	0,59	1a52, 1x7e
	ERβ	VINA	0,81	3oll, 4j24
	FXR	VINA	0,98	5q1a, 5q1h
	PPARα	Plants	0,75	4bcr, 2p54
	PR	Plants	0,63	1sr7
	RXRα	VINA	0,90	1fby, 3oap, 1mv9
Antagonistes vs. Agonistes + inactifs	ERα	Plants	0,90	1xp9, 1xp1
	PR	VINA	0,70	3zra

3.4.3 Similarité des molécules actives/inactives

Dans cette étude, nous modifions le schéma classique qui oppose des molécules actives à des molécules inactives, et qui ne prends en compte que des données d'affinité sans considération de l'activité biologique de la molécule. Ici, le jeu de molécules actives est composé de molécule agonistes (Ag/AntNB) ou de molécules antagonistes (Ant/AgNB) et les molécules de profil pharmacologique opposées sont considérées dans le jeu de données inactives. Ce dernier contient aussi les molécules n'ayant montré aucune capacité à interagir avec la cible lors de tests d'affinité expérimentaux (les *non-binders*, NB). Par ailleurs, contrairement à l'utilisation de *decoys*, L'inactivité dans cette étude n'est plus hypothétique ; elle est définie comme l'incapacité à stimuler une activité donnée (agonisme (Ag/AntNB) ou antagonisme (Ant/AgNB)). Cette définition peut être appliquée à d'autres familles de protéines comme les récepteurs couplés aux protéines G (GPCRs) qui peuvent aussi être modulés par des molécules agonistes et des molécules antagonistes[462].

Nous observons qu'au-delà de l'importance du choix de l'outils de docking adapté aux récepteurs nucléaires, et au choix de structure(s), une explication à la difficulté d'atteindre de bonnes performances de docking en utilisant les données de la NR-DBIND réside dans les similarités observées entre les molécules agonistes et antagonistes ainsi qu'entre les molécules actives (agonistes et antagonistes) et les non-binders (Tableau 20).

Tableau 20 Tableau de similarité entre sous-jeux de données. Pour chaque récepteur, le nombre de molécules du jeu de données A ayant un coefficient de Tanimoto > 0.8 avec une molécule du jeu de données B ainsi que le nombre de squelettes de Bemis Murcko du jeu de données A possédant un coefficient de Tanimoto de 1 avec un squelette de Bemis Murcko du jeu de données B sont donnés en pourcentage.

		AR		**Erα**		**ERβ**		**FXR**		**PPARα**		**PR**		**RXRα**	
		ind. Mol. Tc > 0.8	Tc BM = 1	ind. Mol. Tc > 0.8	Tc BM = 1	ind. Mol. Tc > 0.8	Tc BM = 1	ind. Mol. Tc > 0.8	Tc BM = 1	ind. Mol. Tc > 0.8	Tc BM = 1	ind. Mol. Tc > 0.8	Tc BM = 1	ind. Mol. Tc > 0.8	Tc BM = 1
Agonist	Antagonists	48,2	34,0	33,3	21,4	8,2	12,5	-	-	-	-	58,0	21,4	33,3	41,7
	Non-binders	0.0	8,0	44,4	21,4	35,6	8,3	94,1	16,7	77,8	22,8	6,7	7,1	-	-
Antagonist	Agonists	21,5	13,4	19,9	6,5	20,6	10,0	-	-	-	-	41,7	17,3	66,2	40,0
	Non-binders	18,0	7,9	30,4	6,5	39,2	0.0	-	-	-	-	32,1	19,2	-	-
Non-binders	Agonists	0.0	5,7	6,1	3,5	24,6	6,4	11,4	5,0	30,1	16,5	4,1	5,7	-	-
Ag	AntNB	48,2	34,0	72,2	42,8	42,5	20,8	94,1	16,7	**77,8**	**22,8**	**58,5**	**26,2**	**33,3**	**41,7**
AntNB	Ag	16,5	9,9	12,6	4,6	22,2	8,0	11,1	4,8	**30,1**	**16,5**	**27,2**	**11,3**	**53,1**	**35,3**
Ant	AgNB	-	-	49,1	13,0	-	-	-	-	-	-	**62,8**	**34,6**	-	-
AgNB	Ant	-	-	9,6	6,2	-	-	-	-	-	-	**42,3**	**17,4**	-	-

3.4.3.1 Similarités entre molécules agonistes et antagonistes

L'analyse des structures cristallographiques des NRs a révélé que le déclenchement d'une activité agoniste ou antagoniste est intimement lié à la position de l'hélice H12 du récepteur. Dans un état non lié, l'hélice H12 est déplacée du site de liaison et expose ainsi le site de liaison des co-represseurs[463464373465]. L'interaction d'un ligand agoniste avec le site de liaison induit un repliement de l'hélice H12 vers le site de liaison, formant une cavité enfouie obstruant le site de liaison des corépresseurs et structurant le site d'interaction des co-activateurs du NR. Les ligands antagonistes empêchent la fixation de co-activateurs en perturbant le repliement de l'hélice H12 vers le site de liaison. Le mécanisme moléculaire entrant en jeu dans cette perturbation reste peu compris (Cf 6.5). Jusqu'à maintenant, malgré les tentatives pour tenter

de prédire le caractère agoniste ou antagoniste d'un ligand[466467468], la frontière entre les molécules agonistes et antagonistes des NRs reste opaque. Lagarde et al. ont généré des pharmacophores 3D sélectifs des composés agonistes des NRs ou sélectifs des composés antagonistes des NRs[170]. Afin d'atteindre cette sélectivité pour les agonistes et pour les antagonistes, jusqu'à respectivement 52 et 64 pharmacophores basés sur le ligand et sur la structure ont été nécessaires, ce qui souligne la difficulté à comprendre leur mode de liaison. Des différences de propriétés physicochimiques ont néanmoins été observées entre les pharmacophores sélectifs des agonistes et ceux sélectifs des antagonistes ; les pharmacophores sélectifs des agonistes possèdent significativement moins d'accepteurs de liaisons hydrogènes, de groupements hydrophobes, de cycles aromatiques, de charges positives et de charges négatives que les modèles de pharmacophores sélectifs des antagonistes pour respectivement 9, 5 4, 2 et 1 NR(s).

En accord avec la littérature, nous observons des similarités structurales entre les agonistes et les antagonistes (Tableau 20) : en moyenne 37.2% des molécules agonistes partagent un Tc > 0.8 avec les molécules antagonistes, et 46.7% des molécules antagonistes partagent un Tc > 0.8 avec les molécules agonistes. On note notamment des squelettes très proches entre les deux jeux de données (21.8% de squelettes de Bemis Murcko du jeu de molécules agonistes sont communs aux squelettes de Bemis Murcko du jeu de molécules antagonistes ; 17.4% de squelettes de Bemis Murcko du jeu de molécules antagonistes sont communs aux squelettes de Bemis Murcko du jeu de molécules agonistes). Il faut noter que les distances structurales imposées dans les sélections de *decoys* pour éviter l'intégration de faux négatifs tendent à exclure les molécules similaires.

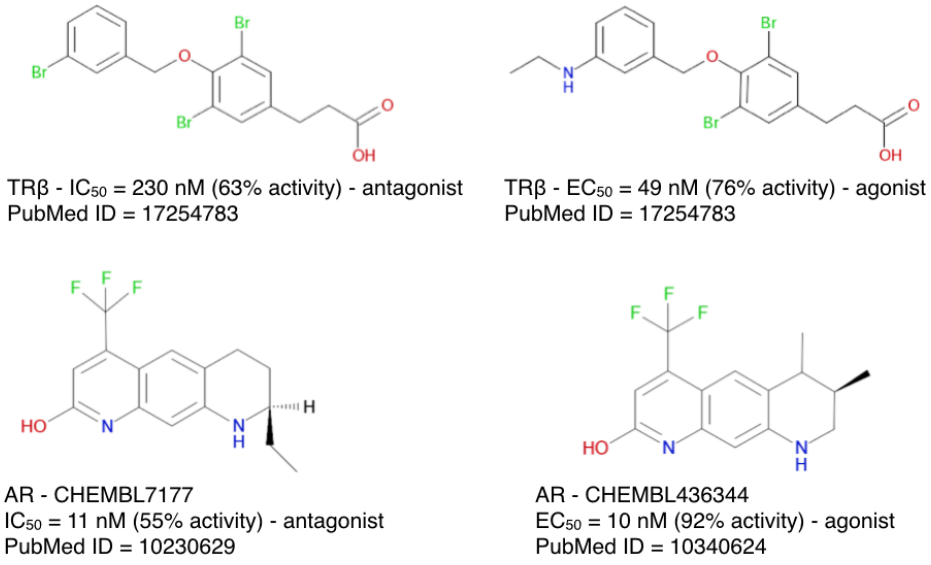


Figure 59 Exemple de molécules topologiquement similaires présentant des profils pharmacologiques différents.

3.4.3.2 Similarité entre molécules agonistes et non-binders/ antagonistes et non-binders

Des similarités structurales sont également observées entre les molécules actives (agonistes et antagonistes) et les *non-binders* (Tableau 20) ; en moyenne 43.2% et 35.1% des agonistes et antagonistes partagent des coefficients de Tanimoto > 0.80 avec des *non-binders*, et 17.2% et 13.3% des *non-binders* partagent des coefficients de Tanimoto > 0.80 avec les molécules agonistes et antagonistes respectivement.

3.4.4 Comparaison à l'utilisation de decoys

Dans cette étude, AutoDock VINA permet d'obtenir des résultats comparables avec l'utilisation de *decoys* et de molécules expérimentalement validées comme inactives dans la majorité des cas. Dans certains cas, les *decoys* sont plus facilement discriminés que les molécules inactives (PPARα (Ag) et PR(Ag)). Il faut noter que 77.8% des agonistes (Ag) de PPARα partagent un Tc > 0.80 avec au moins une molécule du jeu AntNB et que 30.1% des de ces des molécules AntNB partagent un Tc > 0.80 avec au moins un agoniste. De fortes similarités sont également observées entre les molécules agonistes et AntNB de PR, avec 58.5% des molécules AntNB partageant un Tc > 0.80 avec au moins un agoniste et 27.7% des agonistes partageant un Tc > 0.80 avec au moins une molécule. Ces fortes similarités peuvent expliquer une plus grande difficulté à discriminer les Ag/AntNB que les Ag/Decoys. Par ailleurs, ces jeux de données

comptent parmi les 3 (PPARα (Ag), PR(Ag) et RXRα (Ag)) partageant le plus de similarité structurales Ag/AntNB. Dans d'autres cas, les *decoys* sont moins facilement discriminés que les molécules inactives (RXRα (Ag) et ERα (Ant)). Dans le cas d'ERα (Ant), les molécules du jeu de molécules inactives (AgNB) présentent le taux de similarité avec les molécules actives (Ant) le plus bas (9,6% des AgNB possèdent un Tc > 0.80 avec au moins un antagoniste), ce

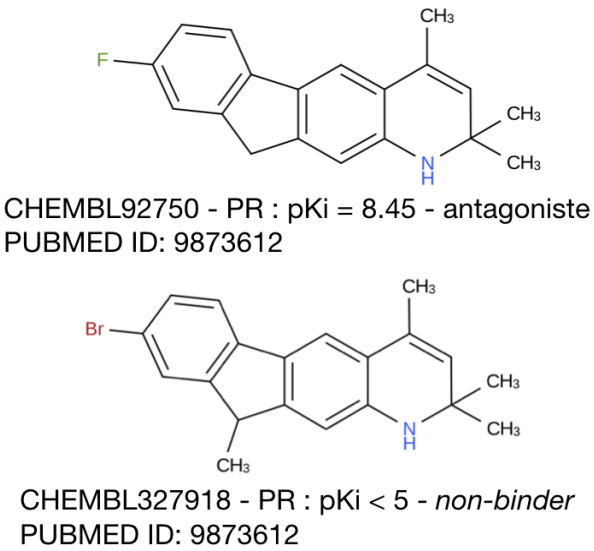


Figure 60 Exemple de molécules topologiquement similaires présentant des affinités différentes pour le récepteur PR

qui peut expliquer la plus grande capacité à discriminer les Ant/AgNB que les Ant/Decoys. Les résultats obtenus sur RXRα s'expliquent partiellement par le faible recouvrement de l'espace chimique des molécules actives et des molécules inactives, malgré les similarités structurales observées entre les molécules du jeu Ag/AntNB de RXRα (Cf annexe 8).

Les différences de performances observées entre l'utilisation de *decoys* et de molécules inactives dépendent majoritairement de la similarité entre les molécules actives et inactives. L'utilisation de molécules inactives facilite la discrimination lorsque les molécules sont structurellement différentes des molécules actives et la complexifie lorsque les molécules partagent des similarités avec les molécules actives. De la même façon que pour les *decoys*, il est important de s'assurer de l'espace chimique couvert par les molécules actives et inactives pour interpréter les résultats obtenus. Les molécules inactives restant rares, l'information qu'elles apportent couvrent souvent une faible espace chimique en comparaison aux *decoys* (Cf Annexe 8). Les deux approches peuvent apporter des informations complémentaires dès lors

que des molécules similaires sont observées entre le jeu de molécules active et inactives. Dans ce cas précis, la sélection d'une ou plusieurs structures pour des études de criblage devrait découler d'un score dérivé des AUCs obtenues en utilisant des molécules inactives et des *decoys* de façon indépendante. Il reste cependant difficile de définir un seuil de recouvrement des espaces chimiques nécessaire pour juger les informations complémentaires et dans le cas où les molécules inactives sont très différentes des molécules actives, il convient d'étudier davantage la structure des molécules pour comprendre ce qui confère l'activité ou l'inactivité aux molécules.

3.5 Conclusion

A notre connaissance, cette étude est la première à ce jour à évaluer la capacité d'outils de docking à discriminer des molécules actives de molécules inactives validées expérimentalement. Dans cette étude, nous proposons de définir l'inactivité comme l'absence de l'activité recherchée, incluant ainsi les non-binders et les molécules affines pour la cible déclenchant une activité non souhaitée. Les performances de deux outils de docking, PLANTS et AutoDock VINA en docking sur structure seule et sur des ensembles de 2 et 3 structures ont été comparées.

Les résultats de cette étude montrent que les outils de docking PLANTS et AutoDock VINA discriminent avec difficulté les molécules actives des molécules inactives, mais permettent toutefois d'obtenir des AUCs maximales ≥ 0.7 dans 7/9 cas, AutoDock VINA montrant de meilleures performances sur l'ensemble des données. L'étude des conditions de docking favorisant ces discriminations n'a pas permis d'identifier une règle commune à l'ensemble des NRs. L'étude révèle qu'avec AutoDock VINA, le docking d'ensemble reproduit ou améliore les performances observées en docking sur structure seule. Les faibles gains d'AUCs observés soulignent que la prise en compte de la flexibilité n'est pas essentielle pour les NRs étudiés, en revanche, l'utilisation d'au moins 2 structures améliore la moyenne des AUCs observées avec AutoDock VINA, et minimise ainsi le risque d'obtenir de faibles performances dû à un mauvais choix de structure. Enfin, nous observons que les performances obtenues avec AutoDock VINA en utilisant des molécules inactives et des *decoys* sont dans la majorité des cas comparables. Les différences d'AUC les plus importantes sont observées dans les jeux de données partageant le plus de similarité entre les molécules constituant le jeu de données d'actifs et le jeu de données d'inactifs, ou à l'inverse, partageant le moins de similarité. Dans le premier cas, les meilleures performances sont associées aux jeux de données contenant des *decoys* dans la majorité des cas (2NRs/3). Dans le second cas, les meilleures performances sont associées aux

jeux de données contenant des molécules inactives pour 1 NRs/2). Nous suggérons que l'information apportée par l'utilisation des molécules inactives n'est pas suffisante du fait des ratios de molécule actives/ molécule inactives observés et que, par conséquent, la sélection d'une ou plusieurs structures pour des études de criblage devrait découler d'un score dérivé des AUCs obtenues en utilisant des molécules inactives et des *decoys* de façon indépendante. Nous notons finalement que le réel défi dans l'évaluation de la capacité des méthodes à discriminer les molécules actives des molécules inactives validées expérimentalement réside dans la disponibilité et la diversité des données. Nous suggérons que l'intégration de davantage de molécules inactives partageant ou non des structures similaires avec les molécules actives devrait faciliter et rationnaliser l'interprétation des résultats de docking, et devrait améliorer la construction de modèles fiables dans de nombreux domaines de recherche. Des initiatives favorisant l'accès libre à des données d'inactivité fiables devraient être activement supportées.

4 Construction de modèles de pharmacophores sélectifs du récepteur nucléaire AR

4.1 Introduction

Un pharmacophore est défini selon l'IUPAC comme un « ensemble d'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 »[169]. L'identification d'un pharmacophore est donc très complexe : il est très difficile d'identifier l'ensemble des interactions essentielles entre une petite molécule et une protéine ainsi que leur arrangement spatial. Ceci revient à identifier les éléments de mode(s) de liaison de petites molécules qui sont à l'origine de l'activité déclenchée. Comme détaillée en partie 3.1.3 et en partie 4.4.1, les pharmacophores peuvent être modélisés à partir de données structurales des ligands d'une protéine ou de la structure tridimensionnelle de la protéine complexée ou non à un ligand. Les modèles de pharmacophores basés sur les ligands sont généralement construits à partir de molécules actives, et éventuellement optimisés à l'aide d'informations sur les molécules inactives. Dans cette étude, nous avons tiré profit des molécules actives et inactives sur AR avec l'objectif de construire des modèles de pharmacophores adaptés à la prédiction de modulateur potentiellement non-désirés d'AR. Nous nous sommes, pour cela, focalisés sur la modélisation de pharmacophores basés sur les ligands sélectifs des molécules agonistes du récepteur aux androgènes. La prédiction de modulateurs potentiellement non-désirés nécessite une maximisation de la sensibilité de sorte à correctement identifier un maximum de vrai positif. Nous avons donc construit des modèles de pharmacophores à partir de molécules actives avec LigandScout, puis les avons optimisés de sorte à les rendre plus sensibles, tout en contrôlant leur spécificité grâce aux informations apportées par les molécules inactives.

L'objectif de ce projet est d'étudier la faisabilité de la modélisation de pharmacophores dédiés à la prédiction de modulateurs potentiellement non-désirés. Les modèles de pharmacophores ont donc été construits sur un jeu de données propre extrait de la NR-DBIND, puis ils ont été testés sur des données issues de la Tox21.

4.2 Matériel et méthodes

4.2.1 Sélection et préparation des données

NR-DBIND

Dans cette étude, nous avons utilisés les molécules classées « binders » (pIC50 ≥ 7 ou pKi ≥ 7) et « non –binders » (pIC50 ≤ 5 ou pKi ≤ 5) selon la NR-DBIND. La NR-DBIND recense des molécules pour lesquelles les données d'affinité sont fournies sous forme de pIC50 ou de pKi ; dans cette étude, ils ont été séparés en deux jeu de données indépendants. Parmi les « binders », 1) les molécules dépourvues d'annotation quant à leur profil pharmacologique ont été rejetées, 2) les molécules annotées « agonistes » et jamais annotées « antagonistes » ont été considérées dans le jeu de molécules agonistes, et 3) les molécules annotés « antagonistes » et jamais « agonistes » ont été classés dans le jeu de molécules antagonistes. Les molécules agonistes constituent le jeu de données actives tandis que les antagonistes et les *non-binders* constituent le jeu de données inactives. Pour chaque molécule, l'état de protonation majoritaire à pH 7.4 a été calculé avec Marvin 17.22.0, 2017, ChemAxon (http://www.chemaxon.com). 50 conformations ont été générées avec iCon (implémenté dans LigandSCout 4.2) et exportées aux formats SDF et LDB.

Tox21

Un autre jeu de molécules a été constitué à partir de résultats de HTS effectué sur le récepteur AR dans le cadre du projet Tox21 (AID 743053, données modifiées le 2014-08-20, https://pubchem.ncbi.nlm.nih.gov/bioassay/743053). Les molécules ont été testées dans un test d'activité agoniste effectué sur des cellules AR-bla, possédant la protéine AR humaine, et des tests d'auto-fluorescence ont permis d'éliminer les faux positifs. Les molécules ont été extraites au format SMILES et ont été préparées suivant le même protocole que celles de la NR-DBIND. Les molécules agonistes ont été considérées dans le jeu de données actives, et les antagonistes et molécules ne présentant pas d'activité pour AR ont été considérées dans le jeu de données inactives. Au total, les structures 3D de 7087/7125 molécules inactives et 232/262 molécules actives ont pu être générées, faute de structures disponibles pour les autres.

4.2.2 Comparaison des espaces chimiques des jeux de données issus de la NR-DBIND et de la Tox21

Une analyse en composantes principales (ACP) a été réalisée sur les molécules des jeux de données issus de la NR-DBIND et de la Tox21, toutes décrites par 11 descripteurs (poids

moléculaire, nombre de donneur de liaison hydrogène, nombre d'accepteur de liaison hydrogène, cLogP, surface accessible au solvant, indice de flexibilité moléculaire, indice de complexité moléculaire, nombre de liaison rotatives, nombre de cycles aromatiques, nombre d'atomes impliqués dans un cycle aromatique, surface polaire accessible au solvant). Les descripteurs et l'ACP ont été calculés avec DataWarrior.(Version 4.7.2).

4.2.3 Protocole de génération des pharmacophores

Les jeux de données de pKi et de pIC50 ont été considérés séparément pour construire des pharmacophores indépendants. Pour chaque jeu de données (pKi et de pIC50), les molécules actives et les molécules inactives ont été divisées en jeu d'apprentissage (~2/3) et jeu de test (~1/3) ; 25 tirages aléatoires avec remise ont été réalisés et ont servi de points de départ à un protocole de génération de pharmacophores. Nous avons ensuite appliqué un protocole en 4 étapes (Figure 61) :

1) Les molécules actives du jeu d'apprentissage ont été clusterisées avec LigandScout (paramètres par défaut) ;
2) Des modèles de pharmacophore dits « combinés », c'est-à dire qui possèdent des points pharmacophoriques communs à minimum 10% des molécules actives, ont été générés ; 10 modèles de pharmacophore ont été générés par cluster de molécules actives ;
3) Les molécules actives et inactives du jeu d'apprentissage ont été criblées en autorisant un maximum de points pharmacophoriques omis ;
4) L'AUC et le facteur d'enrichissement à 25% (EF25) ont été calculés

4.2.3.1 Sélection des modèles de pharmacophores

Afin de sélectionner un ensemble de modèles de pharmacophores à optimiser, nous avons sélectionné les modèles de pharmacophores présentant les meilleurs EF25. Les molécules issues de la NR-DBIND étant clusterisées en fonction de leur squelette de Bemis Murcko, nous avons choisi, parmi cette première sélection, le modèle de pharmacophore criblant le plus de squelettes de Bemis Murcko dans les premiers % de la chimiothèque. Pour ceci, nous avons généré des représentations graphiques des courbes ROC informant sur le type de molécules retrouvées (agonistes/antagonistes/non-binders ; squelette de Bemis Murcko ; molécule appartenant ou non au cluster sur lequel a été construit le pharmacophore) par un modèle de pharmacophore et classées selon le score attribué (Figure 62). Parmi les 25 jeux de données réalisés, seul le jeu de données correspondant aux meilleures performances (plus grand nombre de pharmacophores associés à des bons EF25) est conservé pour les prochaines étapes, de sorte à conserver un jeu d'apprentissage et un jeu de test indépendants. Le modèle de pharmacophore

retenu a été optimisé puis, le second modèle de pharmacophore permettant de retrouver le plus de molécules supplémentaires parmi les 25% les mieux classées à été à son tour sélectionné et optimisé. Ce cycle a été répété jusqu'à ce que l'ajout d'un modèle de pharmacophore n'induise aucune amélioration de sensibilité et de spécificité.

4.2.3.2 Optimisation des modèles de pharmacophores

L'optimisation de chaque modèle de pharmacophore a été faite en 3 étapes :

1) Les molécules actives et inactives du jeu de données d'apprentissage ont été de nouveau criblées en diminuant le nombre de points pharmacophoriques omis autorisé. Cette étape a été reproduite plusieurs fois et ce, tant que le modèle de pharmacophore criblait des molécules actives présentant un minimum 2 squelettes de Bemis Murcko différents et jusqu'à ce que le modèle de pharmacophore ne crible plus de molécules inactives ou que le nombre de molécules inactives criblé ne diminue plus.
2) Le modèle de pharmacophore a ensuite été modifié en déplaçant légèrement la position d'un point pharmacophorique, sa sphère de tolérance ainsi que leur poids et en ajoutant des volumes d'exclusion, de sorte à améliorer le nombre de molécules actives criblées, tout en limitant le nombre de molécules inactives criblées ; le nombre de points pharmacophoriques omis autorisé a été augmenté lorsqu'il n'influait pas sur le nombre de molécules inactives criblées ; une modification permettant de cribler un squelette jusqu'alors jamais criblé a été acceptée malgré l'inclusion de nouvelles molécules inactives
3) L'optimisation prend fin lorsque les nouvelles modifications proposées ne permettent de cribler aucune molécule active supplémentaire sans inclure davantage de molécules inactives ; il est a noté que plusieurs modèles de pharmacophores peuvent être dérivés d'un même modèle de pharmacophore initial

Ces étapes ont nécessité des modifications manuelles qui limitent la reproduction à l'identique et l'automatisation du protocole.

4.2.4 Calcul des performances

La sensibilité et la spécificité de l'ensemble des modèles de pharmacophores retenus ont finalement été calculées sur le jeu d'apprentissage et le jeu de test. Les modèles de pharmacophores issus des données de pKi ont également été testés sur le jeu de pIC50, et inversement, les modèles de pharmacophores issus du jeu de pIC50 ont été testées sur le jeu de pKi. Afin d'évaluer les performances des pharmacophores modélisés, nous les avons testés sur les données extraites de la Tox21.

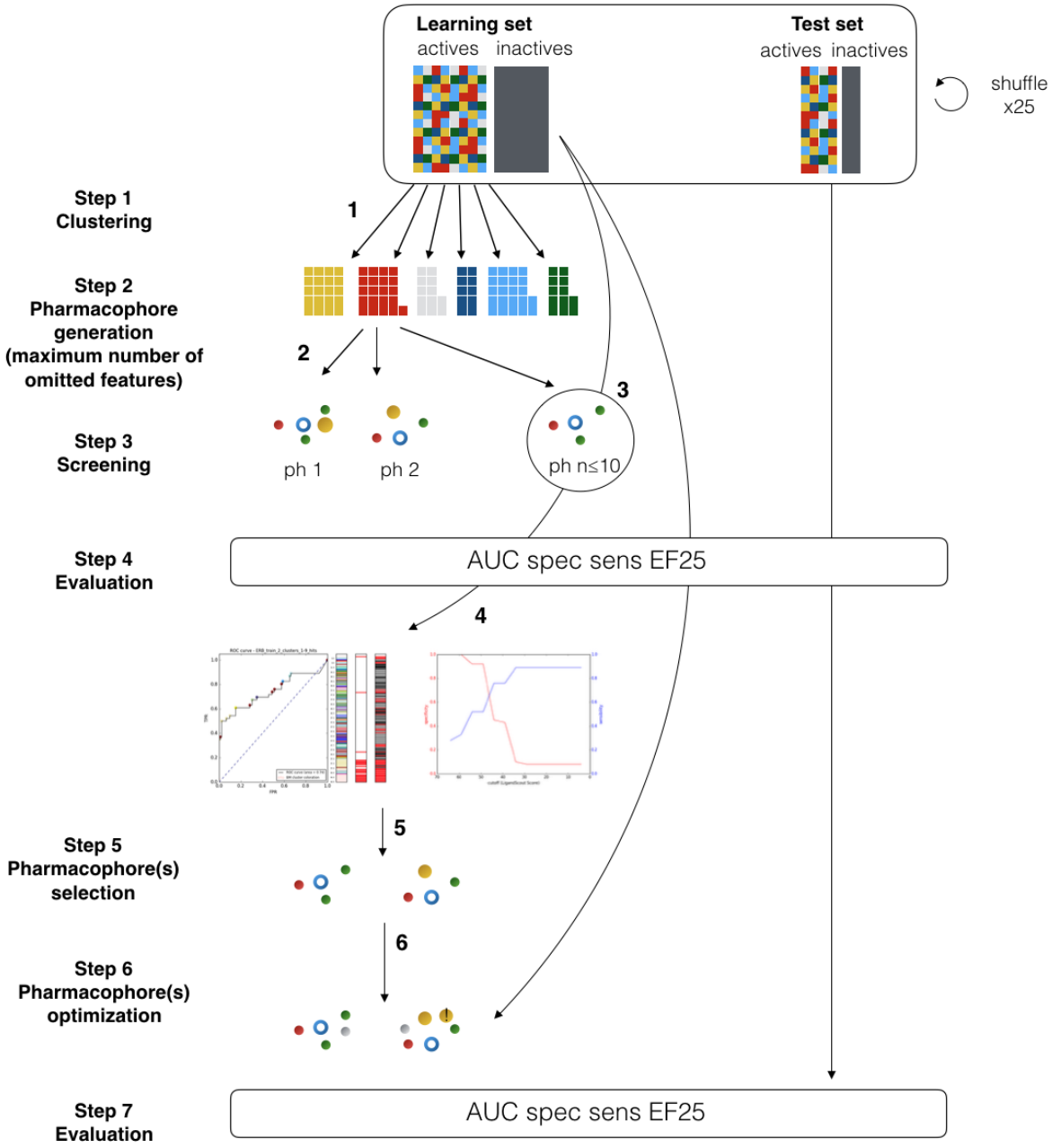


Figure 61 Schéma du protocole de génération et d'optimisation de pharmacophore.1 1) les molécules sont divisées en jeu d'apprentissage et de test, 2) les molécules actives du jeu d'apprentissage sont clusterisées avec LigandScout 4.2, 3) pour chaque cluster, 10 modèles d pharmacophores sont générés, 4) les performances en terme d'EF25, de sensibilité et de spécificité sont calculés sur le jeu d'apprentissage, 5) les pharmacophores sont classés en fonction de leur EF25, et parmi les meilleurs, celui enrichissant plusieurs squelettes de Bemis Murcko est sélectionné pour 5) être optimisé à partir des molécules actives et inactives du jeu d'apprentissage. Enfin, 6) les performances du pharmacophore optimisé sont évaluées sur le jeu d'apprentissage et le jeu de test

4.3 Résultats

4.3.1 Performances des modèles de pharmacophores optimisés

Suivant le protocole explicité ci-dessus, 9 modèles de pharmacophores optimisés ont été dérivés de 3 modèles de pharmacophores générés à partir du jeu de pIC50, et 9 modèles de pharmacophores optimisés ont été dérivés de 4 modèles de pharmacophores générés à partir du jeu de pKi (Tableau 21). La Figure 62 illustre les résultats du criblage primaire effectué avec un nombre de points pharmacophoriques omis autorisé maximal.

Les modèles de pharmacophores optimisés générés à partir du jeu de pKi permettent d'atteindre des sensibilités de 0.95 et 0.80 et des spécificités de 0.88 et 0.89 sur le jeu d'apprentissage et de test respectivement (Tableau 22, A). En comparaison avec le criblage effectué sur les modèles de pharmacophores primaires dans les mêmes conditions, les optimisations ont permis de réduire considérablement le nombre de molécules inactives criblées et ainsi d'améliorer la spécificité du modèle (+0.70) tout en conservant une bonne sensibilité (-0.10) (85/89 molécules actives et 21/182 molécules inactives criblées dans le jeu d'apprentissage ; 43/48 molécules actives et 18/90 molécules inactives criblées dans le jeu de test). Les modèles de pharmacophores optimisés générés à partir du jeu de pIC50 permettent quant à eux d'atteindre des sensibilités de 0.95 et 0.74 et des spécificités de 0.94 et 0.91 sur le jeu d'apprentissage et de test respectivement. Les modèles de pharmacophores optimisés ont cette fois-ci permis de cribler davantage de molécules actives et ainsi d'améliorer la sensibilité (+0.38) tout en conservant une bonne spécificité (+0.00).

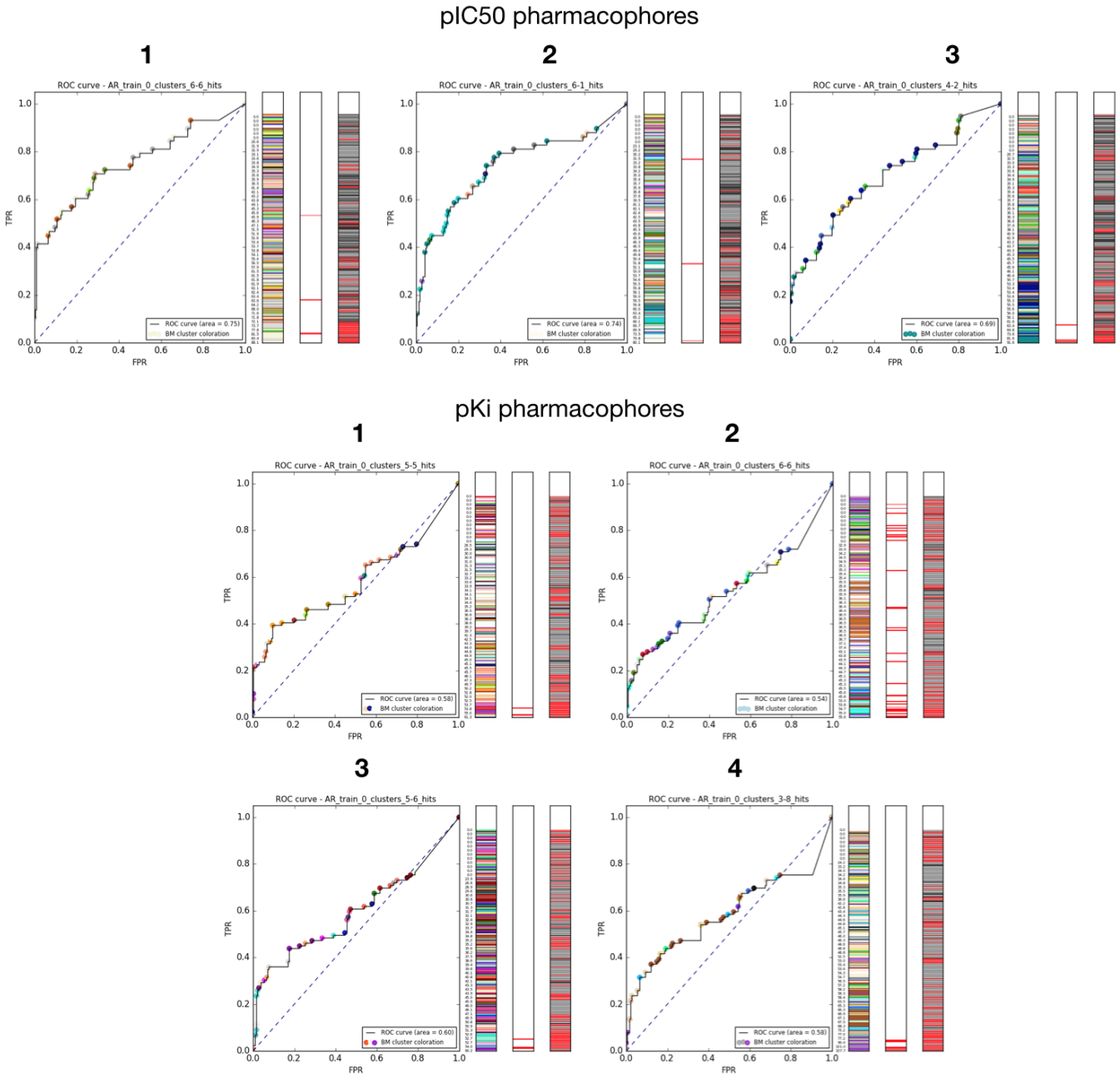


Figure 62 Courbes ROC du criblage du jeu de données d'apprentissage avec les modèles de pharmocophores initiaux (non optimisés) construits à partir du jeu de données de pIC50 (1,2,3) et du jeu de données de pKi (4,5,6,7) changer la numérotation sur le graphe). Les 3 colonnes à droite de la courbe de ROC permettent de représenter les molécules du jeu de données triées selon leurs scores de criblage (la molécule associée au meilleure score correspond à la ligne la plus basse) avec 3 codes couleur : dans la colonne de gauche chaque squelette de Bemis Murcko des molécules criblées est affiché avec une couleur différente, dans la colonne du centre, les molécules du cluster à partir duquel le pharmacophore a été généré sont indiquées en rouge et dans la colonne de droite les molécules agonistes sont indiquées en rouge, les molécules antagonistes sont indiquées en gris et les molécules inactives sont indiquées en noir.

Tableau 21 Liste des modèles de pharmacophores optimisés retenus, du modèle de pharmacophore primaire duquel ils ont été dérivés, et des résultats des criblages réalisés sur les jeux d'apprentissage et de test

				Avant optimisation		**Après optimisation**			
		nombre de point(s) pharmacophorique(s) omis autorisé	**pharmacophore initial**	**Jeu d'apprentissage**		**Jeu d'apprentissage**		**Jeu de test**	
	Pharmacophore		**nom**	**Agonistes** (pKi: 89 ; pIC50: 58)	**Autres** (pKi: 182 ; pIC50: 216)	**Agonistes** (pKi: 89 ; pIC50: 58)	**Autres** (pKi: 182 ; pIC50: 216)	**Agonistes** (pKi: 48; pIC50: 34)	**Autres** (pKi: 90; pIC50: 107)
pharmacophores pKi	1	1	train_0_cluster-5_5	81	142	15	1	10	2
	2		train_0_cluster-5_5			33	3	16	3
	3		train_0_cluster-5_5			32	4	16	4
	4	2	train_0_cluster-6_6	83	160	40	3	21	4
	5	1	train_0_cluster-6_6	83	156	31	3	19	2
	6	1	train_0_cluster-5_6	82	136	22	3	12	1
	7		train_0_cluster-5_6			21	0	10	0
	8		train_0_cluster-5_6			39	13	21	12
	9	0	train_0_cluster-3_8	7	3	7	1	1	0
	Total			**88**	**165**	**85**	**21**	**43**	**18**
pharmacophores pIC50	1	0	Train_0_cluster-6_6	20	2	26	4	9	0
	2	0	Train_0_cluster-6_1	14	1	33	5	10	2
	3	0	Train_0_cluster-6_1			7	1	7	0

	4	0	Train_0_cluster-6_1			19	1	9	2
	5	1	Train_0_cluster-4_2	16	1	24	1	5	4
	6	1	Train_0_cluster-4_2			31	0	6	0
	7	1	Train_0_cluster-4_2			5	0	1	0
	8	0	Train_0_cluster-4_2	30	11	6	1	6	1
	9	1	Train_0_cluster-4_2	16	1	7	0	3	1
	Total			**33**	**11**	**57**	**13**	**26**	**9**

Tableau 22 Performances obtenues en termes de sensibilité, de spécificité et de facteur d'enrichissement, lors du criblage des jeux de données d'apprentissage (train) et de test (test) pKi et pIC50, et des jeux de données de validation externe pKi-pIC50 et Tox21 avec les modèles de pharmacophores pKi, pIC50 et la combinaison des deux.

A

Jeu de données	**pharmacophores pKi**			**pharmacophores pIC50**			**Combinaison des deux**		
	Sensibilité	spécificité	EF	Sensibilité	spécificité	EF	Sensibilité	spécificité	EF
pKi	Train: 0.95 Test: 0.80	Train: 0.88 Test: 0.89	Train: 2.44 Test: 2.03	0.23	0.87	1.22	-	-	-
pIC50	0.61	0.62	1.43	Train: 0.98 Test: 0.74	Train: 0.94 Test: 0.91	Train: 3.45 Test: 3.08	-	-	-
pKi-pIC50	0.78	0.73	1.90	0.53	0.89	2.33	0.92	0.68	1.88

B

	Application à la NR-DBIND								
Tox21 (<1uM)	0.07	0.95	1,37	0.19	0.90	1,94	**0.21**	**0.88**	**1,72**
Tox21 (<100nM)	0.09	0.95	1,87	0.32	0.90	3,28	**0.33**	**0.88**	**2,74**
Tox21 (<10nM)	0.14	0.95	2,71	0.38	0.90	4,01	**0.38**	**0.88**	**3,24**

4.3.2 Performances des modèles de pharmacophores optimisés en criblages croisés

Le jeu de données de pKi a été criblé avec les modèles de pharmacophores pIC50 et le jeu de données de pIC50 a été criblé avec les modèles de pharmacophores pKi. Les performances obtenues avec ce criblage croisé sont inférieures à celles obtenues sur le jeu de données initial : des sensibilités de 0.23 et de 0.61 et des spécificités de 0.87 et de 0.62 sont obtenues avec les modèles de pharmacophores pKi et pIC50 respectivement (Tableau 22, A). Le criblage de l'ensemble des données (pKi-pIC50) avec les modèles de pharmacophore pKi-pIC50 permet d'atteindre une forte sensibilité (0.92) et une spécificité de 0.68.

4.3.3 Criblage de la Tox21

Les données issues de la Tox21 (Cf annexe 9) ont été criblées avec les modèles de pharmacophore pKi et pIC50 ; 3 jeux de données ont été considérés, chacun ayant un jeu de données actives respectivement constitué de :

1) Toutes les molécules annotées agonistes d'AR et ayant une AC50 < 1uM (232 molécules) ;
2) Les molécules annotées agonistes d'AR et ayant une AC50 < 100nM (106 molécules) ;
3) Les molécules annotées agonistes d'AR et ayant une AC50 <1nM (44 molécules) ;

les autres molécules (antagonistes et molécules ne déclenchant aucune activité biologique) sont considérées dans le jeu de données inactives (7087). Les modèles de pharmacophores pKi permettent de retrouver peu de molécules agonistes (16/232, 10/106 et 6/44), et présentent de faibles sensibilités (de 0.07 à 0.14) (Tableau 22, B). De meilleures sensibilités (0.21 à 0.38) sont retrouvées avec les modèles de pharmacophores pIC50 (44/232, 34/106 et 17/44), mais restent néanmoins très faibles par rapport aux résultats obtenus sur les données issues de la NR-DBIND. Les modèles de pharmacophores pKi et pIC50 permettent tous les deux d'obtenir de fortes spécificités (0.95 et 90 respectivement). Les performances obtenues avec l'ensemble des modèles de pharmacophore combinés (pKi-pIC50) montrent une sensibilité et une spécificité comparables.

4.4 Conclusion et perspectives

Les modèles de pharmacophores générés sur les jeux de données pKi et pIC50 extraits de la NR-DBIND présentent de très bonnes sensibilités et spécificités sur leur jeu d'apprentissage respectif. Les performances sont légèrement inférieures sur le jeu de test. Ces résultats montrent que l'utilisation de molécules actives et inactives a permis de générer des pharmacophores

capables de retrouver certains agonistes d'AR et d'éliminer de nombreux antagonistes et *non-binders* sur les données extraites de la NR-DBIND. Appliqués sur des jeux de données indépendant extraits de la Tox21, ces modèles de pharmacophores montrent de faibles sensibilités, mais conservent de bonnes spécificités. La comparaison des espaces chimiques des molécules actives et inactives issues de la NR-DBIND et de la Tox21 montre que les molécules de la Tox21 (actives et inactives) couvrent un espace chimique plus large, avec peu de chevauchement (Cf Annexe 9). Cependant, de manière intéressante, les molécules criblées de la Tox21 ne coïncident pas systématiquement avec l'espace couvert par les molécules de la NR-DBIND, ce qui montre que l'approche utilisée ne se limite pas à l'espace appris. L'applicabilité de ces modèles de pharmacophore à des fins de prédiction d'effet potentiellement non désirés dépend donc des molécules sur lesquelles les modèles ont été appris mais ne se limitent pas à cet espace chimique. Les spécificités atteintes sur les jeux de données issus de la Tox21 montrent une bonne capacité à rejeter les molécules antagonistes ou inactives. Les pharmacophores semblent donc plus adaptés pour rejeter des vrais négatifs que pour accepter l'ensemble des vrais positifs. Malgré les optimisations apportées pour augmenter la sensibilité des modèles de pharmacophore tout en contrôlant la spécificité, les modèles construits ne semblent pas suffisants pour être utilisés seuls dans un objectif de prédiction de modulation potentiellement non désirée d'AR. Les résultats sont en revanche très encourageants quant à l'utilisation de ces modèles dans un objectif de recherche de molécules thérapeutiques.

5 Application de protocoles de criblage virtuel

Parallèlement aux travaux de thèse présentés ci-avant, des études de criblage virtuel ont été conduites sur 3 cibles protéiques différentes (FXR, NRP-1 et TNFα). Ces travaux, issus de collaborations et d'une participation à un challenge de *drug design in silico,* sont présentés dans cette partie, exceptés les travaux concernant le TNFα qui fait l'objet d'un accord de confidentialité.

5.1 Predicting the affinity of Farnesoid X Receptor ligands through a hierarchical ranking protocol: a D3R Grand Challenge 2 case study

5.1.1 Introduction

Le D3R Grand Challenge est un défi annuel hébergé par l'université de La Jolla en Californie et organisé par le Pr Mickael K. Gilson. L'objectif est de prédire le mode de liaison de petites molécules dans le site de liaison d'une protéine donnée, puis de classer un nombre plus grand de petites molécules en fonction de leur affinité pour la cible. La participation de nombreux laboratoires privés et académiques permet de comparer les performances de différents protocoles et différents outils sur le système d'étude. En 2016, nous avons participé au D3R Grand Challenge 2 (D3R GC2), dont la cible d'intérêt était le récepteur nucléaire *Farnesoid X receptor* (FXR). FXR est une protéine impliquée dans le métabolisme des lipides et fait parties des cibles étudiées pour leur potentiel thérapeutique dans les diabètes et les hyperlipidémies. Le challenge GC2 s'est divisé en 2 étapes : la première consistait à prédire le mode de liaison de 36 ligands sur FXR et de classer 102 ligands (66+36) en fonction de leur affinité pour FXR, la seconde consistait à affiner le classement suite à la publication des structures co-cristallisées de FXR avec les 36 ligands de la première étape. Nous nous sommes focalisés sur la seconde étape du challenge et nous avons proposé un protocole hiérarchique de classement des molécules: 1) dans un premier temps, les structures des ligands proches des structures co-cristallisées publiées ont été générées manuellement par édition des co-cristaux, et dockées en parallèle avec AutoDock 4 ; 2) les poses obtenues par ces deux méthodes ont été scorées avec HYDE ; 3) pour chaque molécule, la pose associée à la meilleure affinité prédite par HYDE a été retenue ; 4) les molécules ont été classées selon ce score, puis 5) les molécules ayant le

même score HYDE ont été départagées par l'énergie libre de liaison estimée par MM/GBSA. Ce protocole nous a permis d'obtenir un coefficient de kendall τ de 0.41 nous classant 5/77 dans la catégorie des protocoles basés sur la structure. Les résultats ont été valorisés dans une publication qui détaille le protocole appliqué, les résultats obtenus et présente une approche alternative automatisée dépourvue d'étape manuelle qui a permis d'atteindre des performances similaires (kendall $\tau = 0.44$).

5.1.2 Article

J Comput Aided Mol Des (2018) 32:231–238
DOI 10.1007/s10822-017-0063-0

Predicting the affinity of Farnesoid X Receptor ligands through a hierarchical ranking protocol: a D3R Grand Challenge 2 case study

Manon Réau[1] · **Florent Langenfeld**[1] · **Jean-François Zagury**[1] · **Matthieu Montes**[1]

Received: 2 June 2017 / Accepted: 8 September 2017 / Published online: 14 September 2017

Abstract The Drug Design Data Resource (D3R) Grand Challenges are blind contests organized to assess the state-of-the-art methods accuracy in predicting binding modes and relative binding free energies of experimentally validated ligands for a given target. The second stage of the D3R Grand Challenge 2 (GC2) was focused on ranking 102 compounds according to their predicted affinity for Farnesoid X Receptor. In this task, our workflow was ranked 5th out of the 77 submissions in the structure-based category. Our strategy consisted in (1) a combination of molecular docking using AutoDock 4.2 and manual edition of available structures for binding poses generation using SeeSAR, (2) the use of HYDE scoring for pose selection, and (3) a hierarchical ranking using HYDE and MM/GBSA. In this report, we detail our pose generation and ligands ranking protocols and provide guidelines to be used in a prospective computer aided drug design program.

Keywords D3R GC2 · FXR · Docking · SeeSAR · Hyde · Autodock · MM/GBSA

Introduction

The Drug Design Data Resource (D3R) organizes each year a blind prediction challenge that represents a unique opportunity to evaluate and validate computer-aided drug design workflows [1]. The D3R community provides a set of ligands as input and asks participants to blindly predict their relative binding affinities for a given target, and finally compares predicted ranking to experimental data.

For its second edition, the Grand Challenge 2 (GC2), the D3R community focused on the Farnesoïd X receptor (FXR) and provided a set of 102 ligands [2]. For Stage 1, participants were asked to classify the 102 ligands based on their predicted affinity for FXR, and to predict the binding pose of a subset of 36 ligands (S1 set). By the end of Stage 1, co-crystallized X-ray structures of the S1 set were released, allowing participants to refine their pose prediction protocol and to focus on the ligands relative affinities ranking for the GC2 Stage 2. The 66 ligands for which no co-crystal structures were available are subsequently referred to as the S2 set. In the present report, we describe our participation to Stage 2 using the S1 set structural data provided at the end of Stage 1.

In GC2, a key issue was to account for the flexibility of the FXR ligand binding site (LBS) [3]. Several methods are used to apprehend flexibility in docking procedures, starting either from a single or from an ensemble of protein conformations. The former case comprises induced-fit docking that typically allows only restricted motions of LBS side chains. The ensemble approach uses PDB structures or molecular dynamics (MD) simulation snapshots to either perform docking on each receptor conformation, or merge the energetic contribution of individual protein conformations into a potential grid used as receptor for the docking. While the former method could explore

Manon Réau and Florent Langenfeld have contributed equally to this study.

Electronic supplementary material The online version of this article (doi:10.1007/s10822-017-0063-0) contains supplementary material, which is available to authorized users.

✉ Matthieu Montes
matthieu.montes@cnam.fr

[1] Laboratoire GBA, EA4627, Conservatoire National des Arts et Métiers, 2 rue Conté, 75003 Paris, France

extensively LBS local motions, the ensemble approach is likely to better account for large ligand-induced conformational changes in the LBS, such as loop motion. In the present report, we took advantage of the availability of diverse FXR structures co-crystallized with ligands of the same chemical series as the D3R ligands to combine two approaches to account for FXR LBS flexibility: (1) using the available FXR X-ray structures, we manually generated poses for the S2 set ligands based on S1 set co-crystallized ligands from the same chemical series; (2) using an ensemble docking approach with Autodock 4.2 [4] based on six FXR structures sampling the flexibility of the FXR LBS. We merged the poses generated by both protocols, and retained a single pose for each ligand of the S2 set based on HYDE [5] scores. The selected poses for S2 set ligands and X-ray structures of S1 set ligands were pooled and ranked according to HYDE. The ex-aequo compounds were re-ranked according to their binding free energy estimated by MM/GBSA (Fig. 1).

The classification we submitted to the GC2 Stage 2 ranked our protocol 5th out of the 77 submissions in the structure-based category. In the present report, we describe our poses generation and ligands ranking protocol, and discuss potential improvements to be routinely applied in a computer aided drug design pipeline.

Methods

Ligand preparation

The 102 small molecules provided by the D3R community in 2D SDF format were converted to 3D and protonated at pH 7.4 using Open Babel 2.3.2 [6]. Gasteiger partial charges were attributed using AutoDockTools 1.5.6 [4].

Manual edition using SeeSAR

Among the 102 ligands of the GC2, 96 are derived from four series (benzimidazoles, isoxazoles, spiros and sulfonamides) and six are miscellaneous. Most of the ligands from S2 set are very similar to the ligands from S1 set, whose co-crystals were released by the end of Stage 1. We selected the structurally closest compound from the S1 set for each compound of the S2 set, and used the SeeSAR software [7] to modify the corresponding ligand structure in the co-crystal. The energy of the generated poses was minimized using a two-step scheme: (1) the H-bond network within the protein and at the protein–ligand interface was optimized using Protoss [8]; (2) the ligand geometry was optimized using HYDE [5] as implemented in SeeSAR [7].

Structures selection

All FXR structures available in the PDB [9], consisting of 25 agonist-bound and 1 antagonist-bound structures (PDB Set), were retrieved and added to the 36 FXR structures provided after Stage 1 (S1 Set). We performed a hierarchical clustering of the PDB and the S1 Sets based on their pairwise LBS RMSD to select a subset of structures that samples experimentally observed FXR LBS flexibility. The LBS was defined as the residues with at least one atom within 5 Å of a co-crystallized ligand. Pairwise RMSD were computed using Pymol [10]. The structures were clustered according to hclust in R 3.2.3 [11]. We selected the representative structures based on three criteria: (1) we discarded isoxazoles and miscellaneous bound structures since they were not represented in the S2 set, (2) we selected at least one structure for each cluster and (3) the co-crystalized ligand shared a common scaffold with at least one of the 66 Stage 2 ligands. Selected FXR structures were protonated at pH 7.4 with Chimera [12] using the ProPKA [13] module and Gasteiger partial charges were attributed using AutoDockTools 1.5.6 [4]. Water molecules were not considered.

AutoDock 4.2

We ran AutoDock 4.2 [4] docking with default parameters (250 GA runs) to sample near-native poses and clustered the generated poses based on a 2 Å RMSD criteria. We considered cluster representatives, i.e. the lowest score pose of each cluster, for further scoring. S1 ligands predicted poses *versus* S1 ligands binding mode symmetry-corrected RMSD were computed using RDKit [14]. Ligand #33 was not included into the dataset for the docking protocol since the co-crystal showed an inconsistency in the ligand structure. A latter crystal structure provided by Roche allowed the completion of the dataset and was further included for the affinity ranking step.

Pose selection

For the ligands of the S1 set, we retained the crystallographic binding poses. For the ligands of the S2 set, refined poses of manually edited structures and cluster representatives of the predicted binding modes using AutoDock 4.2 were scored with HYDE; top score pose of each ligand was retained for further analysis.

HYDE

The HYdrogen bond and DEhydration (HYDE) empirical scoring function subsequently describes the energy balance between unfavorable hydrophilic dehydration and favorable hydrogen bonding during the binding process [5].

The essential feature of HYDE is the integrated use of log P-derived atomic increments for the prediction of free dehydration energy and hydrogen bonding energy. Taking the dehydration of atoms within the interface into account shows that some atoms contribute favorably to the overall score, while others contribute unfavorably [15].

MM/GBSA approach

MM/GBSA estimates the binding-free energy of a ligand to a target as the sum of the classical enthalpic contributions (bound, Van der Waals and electrostatic energies), the solvation free energies, and the entropic contribution [16]. Ligands parameters were generated with Antechamber [17, 18]. Up to 100 steps of conjugate gradient minimization of the complex were performed using NAMD [19] and the GAFF/ff99SB force field [18, 20]. The OBC1 Generalized Born (GB) model parameters [21] and the LCPO method [22] were used to compute the polar contribution to the solvation energy and the solvent-accessible surface area (SASA), respectively. The entropic term was omitted.

Results

Pose generation

Manual edition

Among the 102 of the challenge, the co-crystals of the 36 S1 set molecules were made available after GC2 Stage 1. Since the 66 S2 set molecules share high structural similarity with the S1 set ligands (average MACCS fingerprint Tanimoto = 0.94), we generated poses for the S2 ligands by editing closely related co-crystallized ligands from the S1 set. We edited the poses with the SeeSAR ligand editor. Poses were then submitted to a two-step minimization protocol using ProToss and HYDE as implemented in SeeSAR. The S2 ligands and their corresponding S1 co-crystallized ligands are presented in Supplementary Table S1.

This pose generation procedure was not possible for the compounds FXR #45 and #90 since steric clashes occurred in the binding site. Hence, only the docking procedure was used for these two ligands.

Docking

In order to cover the experimentally observed conformational variability of the FXR structures available in the PDB and the S1 set, we performed a hierarchical clustering based on their pairwise LBS RMSD. Overall, 1 structure from the PDB (PDB ID: 3OLF), and 5 structures from the S1 set (HQMF, YFJN, SJPR, KJYP and HVIH) were retained (Fig. 2).

Predicted binding modes of the S1 and S2 ligands into the six selected conformations of FXR were generated using AutoDock 4.2. All generated poses were clustered according to their relative RMSD and only the cluster representatives, i.e. the lowest score member of each cluster, were retained. For each ligand, up to 55 poses were retained. Since experimental binding modes were available for the ligands of the S1 set, the docking accuracy of our protocol was evaluated. As presented in Table 1 and Fig. 3, for single structure docking, the best-predicted poses of the S1 ligands displayed average RMSDs from 2.63 to 3.84 Å. For the ensemble docking, the best predicted binding modes of the S1 ligands displayed an average RMSD of 1.62 Å. Near native poses (RMSD <2 Å) could be retrieved for single structure and

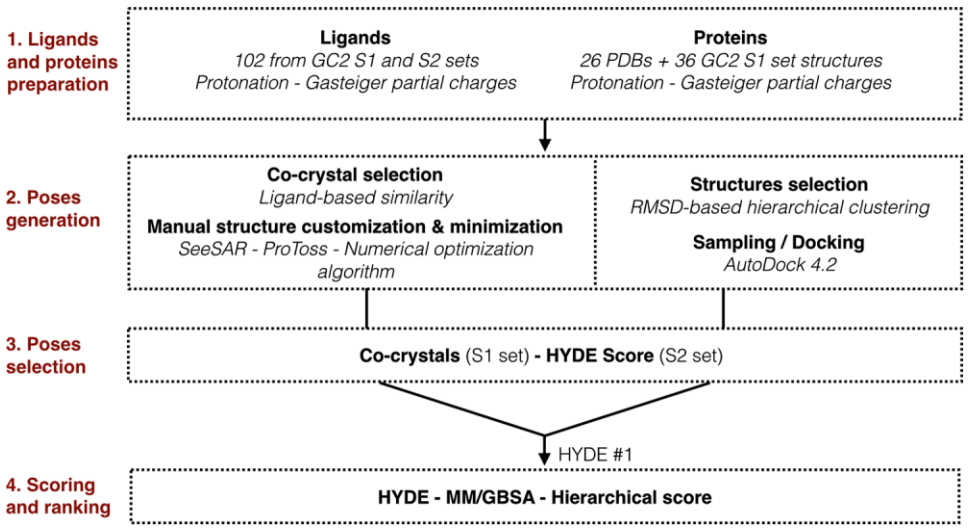


Fig. 1 Schematic representation of the protocol used to predict ligands relative affinities

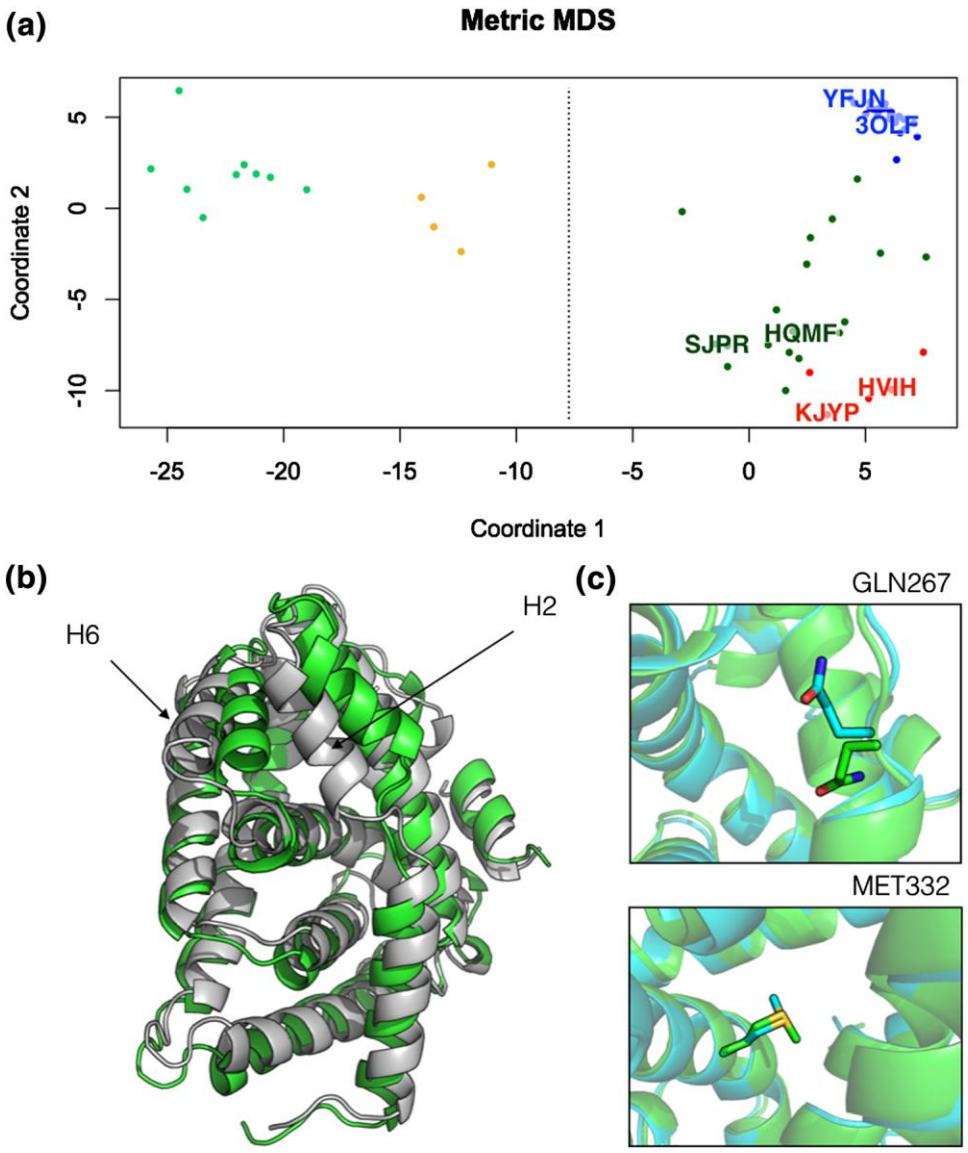


Fig. 2 **a** LBS RMSD based multidimensional scaling of FXR available structures with hclust clusters coloration. No protein on the left of the dashed lines is bound to D3R like ligands. **b** Superimposition of 3OLF (grey) and SJPR (green) and **c** local structural changes between JSPR (green) and HVIH (blue)

Table 1 Mean RMSD values and number of near native poses retrieved by the single structure docking approach and the ensemble docking approach

Structures	3OLF	SJPR	KJYP	HVIH	HQMF	YFJN	Ensemble
AutoDock 4.2 mean RMSD (Å)	2.63	3.16	3.21	3.05	3.84	2.85	1.62
Retrieved near native poses (RMSD <2 Å)	19	7	4	7	3	18	28

ensemble docking for respectively 19 and 28 out of the 36 S1 ligands (Fig. 4).

According to the chemotype classification of the ligands of the S1 set provided by the D3R, the ensemble docking approach displayed a better performance on the benzimidazoles, spiros and sulfonamides compared to the isoxazoles and miscellaneous compounds (Fig. 4).

Pose selection

S1 set

Since the experimental binding modes were provided by the D3R for the S1 ligands, the performance of the HYDE scoring function as implemented in SeeSAR in retrieving their

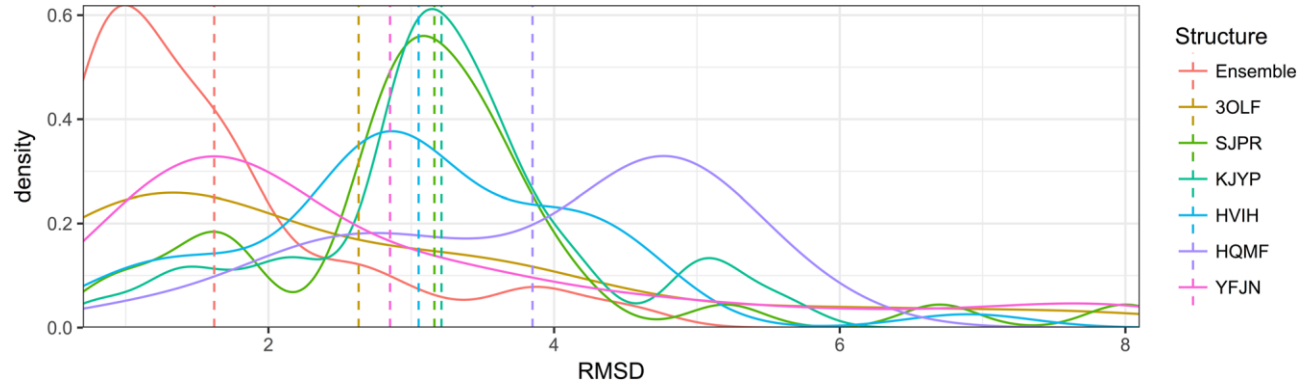


Fig. 3 RMSD distribution for ensemble docking (red), and single structure docking. Dashed lines represent the mean RMSD for each distribution

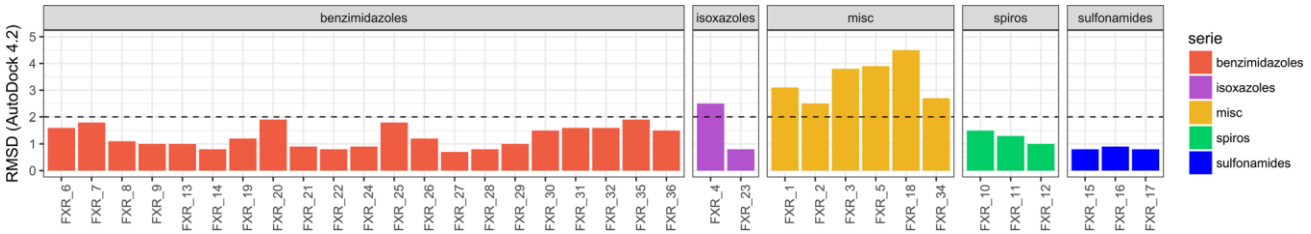


Fig. 4 Ensemble docking—best RMSD per ligand and per chemotype. The dashed line represents the near-native threshold (2 Å RMSD)

experimental binding mode among the retained docking poses generated with AutoDock 4.2 was evaluated.

For 27 out of the 35 S1 ligands, the top score provided by the HYDE scoring function corresponded to the experimental binding mode. For the remaining S1 ligands, one near native pose (RMSD <2 Å) was retrieved as a top score.

For the next step, experimental binding modes were retained.

S2 set

For the ligands of the S2 set, the poses generated using the manual edition and the docking protocols were scored using HYDE; the top score poses were selected. The poses generated using the docking approach (FXR ligands #39, 40, 41, 43–45, 50, 53, 55, 56, 60–64, 66, 68, 72, 73, 80, 90, 93 and 98) and using the manual edition approach (remaining ligands) were selected for the next step.

Relative affinity prediction

The S1 and S2 retained ligands were pooled and ranked according to HYDE. In case of similar scores, notably for the predicted high affinity ligands, the ex-aequo compounds were re-ranked according to their binding free energy estimated by MM/GBSA.

The final ranking of the pooled S1 and S2 compounds was submitted as our prediction for the D3R GC2 under the alias dh2du.

Discussion

Chemical series dependent sampling variations

As presented in Fig. 4, AutoDock 4.2. allowed to retrieve near-native poses for the benzimidazoles, sulfonamides and spiros compounds using the ensemble docking approach. This observation is probably due to the presence of these chemotypes in the structures selected in the ensemble docking approach. For the miscellaneous class of compounds (for which no co-crystal was selected in the ensemble), no near-native poses could be retrieved. Since the FXR LBS conformation depends on the bound ligand (Fig. 2, helix H2 and H6), it should be taken into consideration in the selection of structures for docking studies.

Scoring

Relative affinity prediction

By the end of GC2 Stage 2, the D3R disclosed the experimental IC_{50} of the ligands of the S1 and the S2 sets. The Kendall correlation factor τ between the ranking of the S1 and the S2 ligands set according to (1) their experimental IC_{50} (experimental ranking) and (2) their predicted relative affinities was computed to evaluate the performance of the different GC2 participants protocols in terms of ranking accuracy. The ranking we submitted to the D3R GC2 led to Kendall τ coefficients of 0.35 and 0.41, respectively, on the S1 ligands set and the S1 + S2 ligands set. Our ranking was classified as the 5th best predictive result out of the 77 submissions in the structure-based scoring category; the best Kendall τ coefficient of this category for the S1 + S2 ligands set being 0.46, and the best Kendall τ of the ligand-based category being 0.38.

To complete our observations, we evaluated the individual performance of the HYDE, MM/GBSA and the combinaison of HYDE and MM/GBSA ranking schemes by assessing the Kendall τ correlation coefficients between their associated rankings on either the S1 set or the S1 + S2 set, and the experimental ranking (Table 2). On the S1 set, the MM/GBSA and HYDE ranking schemes displayed similar performance (τ = 0.38). On the S1 + S2 set, the HYDE and HYDE–MM/GBSA ranking schemes displayed a better correlation with the experimental data than the MM/GBSA ranking scheme with respective τ of 0.42, 0.41 and 0.35 (Table 2).

It is to note that the maximum predicted affinity for a given ligand with HYDE is $<10^{-3}$ nM, which led to ex-aequo ranks for several ligands of the S1 and the S2 set that strongly impacted the Kendall τ correlation coefficients.

We then plotted the predicted ranking against the experimental IC_{50} values for the compounds of the S1 + S2 sets (Fig. 5). According to the binding free energies obtained with MM/GBSA, the spiros compounds were ranked before the benzimidazole compounds (Fig. 5a) whereas the scores obtained with HYDE resulted in the opposite trend (Fig. 5b), which is in better accordance with experimental data.

Considering the ranking of the most populated chemotypes, benzimidazoles and spiros were predicted with respective Kendall factors τ of 0.38 and 0.41 by MM/GBSA (Table 2). MM/GBSA outperformed HYDE score for the spiros family (τ = 0.41 vs. τ = 0.16, Table 2). Benzimidazoles rankings were better predicted by HYDE (τ = 0.46). Sulfonamides rankings were predicted by both methods with τ of −0.1 for HYDE and −0.09 for MM/GBSA. Other chemotypes (isoxazoles and miscellaneous compounds) were not sufficiently populated for a significant interpretation.

The MM/GBSA approach is widely used to approximate ligand–protein binding free energy [16]. In the present work, we used an unique MM/GBSA calculation on each

Table 2 Kendall correlation factor τ for the S1, S1 + S2 sets and their most populated chemotypes according to HYDE, MM/GBSA and the hierarchical ranking

	S1 + S2 sets	S1 set	Benzimidazoles	Spiros	Sulfonamides
Compounds	102	36	47	22	23
HYDE	0.42	0.38	0.46	0.16	−0.10
MM/GBSA	0.35	0.38	0.38	0.41	−0.09
Hierarchical ranking	0.41	0.35	0.37	0.17	−0.11

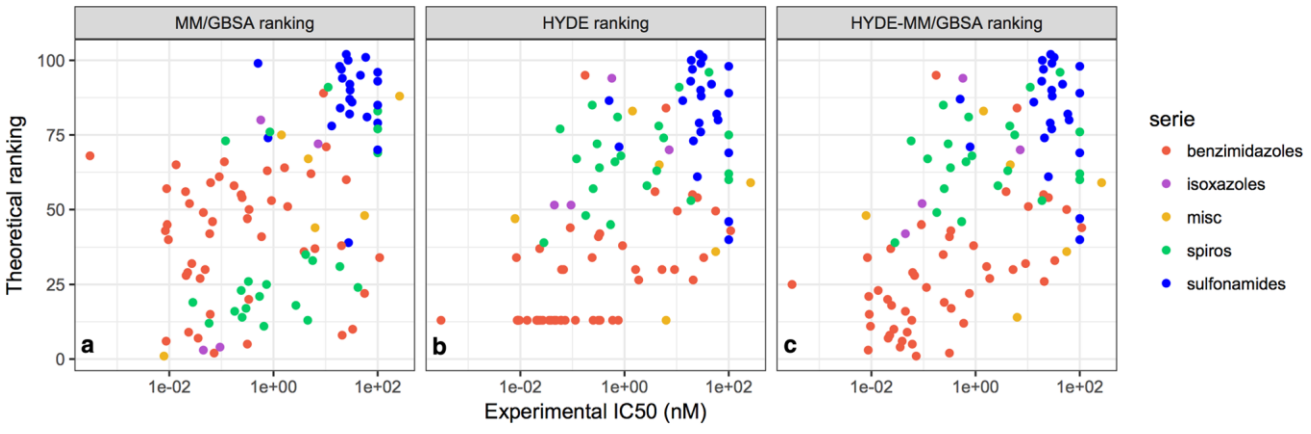


Fig. 5 Theoretical rankings against the experimental IC_{50} values as predicted by MM/GBSA (**a**), HYDE (**b**), and HYDE-MM/GBSA (**c**)

individual pose. Better results could be obtained using MM/GBSA free energies averaged over a set of MD simulations snapshots or an ensemble of X-ray crystal structures since (1) the dynamics of the complex influences the binding free energy and (2) the accuracy of the results can be statistically assessed [16]. However, these benefits come with a significantly higher computational cost hardly compatible with a high-throughput pipeline. We only applied few minimization steps to reduce the steric clashes and computed the binding free energies from a single ligand–protein structure, implying a strong dependence between the predicted binding free energies and the predicted binding modes.

Protocol automation

For the D3R GC2, a lot of structural data were available since many X-ray structures were (1) provided by the organizers and (2) available in the PDB. Since structural data were not always available for the chemotypes that are structurally close to the compounds studied in a drug discovery program, we performed a retrospective evaluation of a fully automated procedure using AutoDock 4.2 [4] and AutoDock VINA [23] for the binding mode prediction step (Fig. 6) and HYDE for the pose selection and ranking steps.

We used the ensemble docking protocol we applied in the D3R GC2 with the 36 ligands from the S1 set. AutoDock VINA displayed a slightly better performance than AutoDock 4.2 in retrieving near-native poses (RMSD <2 Å) (30 and 28 out of 36 ligands, respectively) with respective average RMSD of 1.32 and 1.62 Å (Fig. 6). HYDE score allowed to retrieve most of the AutoDock VINA generated near-native poses (23 out of 30) which was not the case with the AutoDock 4.2 generated near-native poses (14 out of 28).

In terms of ranking, the automated protocol using AutoDock VINA and HYDE score displayed similar results compared to the strategy we used in the D3R GC2, with respective τ values of 0.44 and 0.40 on the S1 + S2 set and the S1 set, respectively. This automated protocol should be favored in a prospective CADD program.

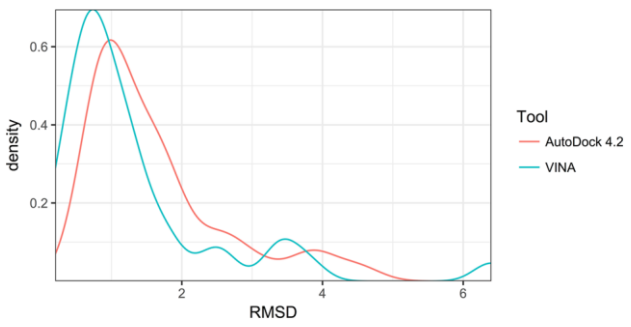


Fig. 6 RMSD distribution for ensemble docking with AutoDock 4.2. and AutoDock VINA

Conclusion

In the present report, we described our structure-based hierarchical ranking protocol that was ranked 5th out of the 77 submissions in the structure-based category of the D3R GC2. This protocol strongly relied on experimental data since we used a manual edition of co-crystallized ligands and numerous target structures through an ensemble docking approach. However, we also described a fully automated approach to overcome the manual edition step that displays similar results (Kendall τ of 0.41 and 0.44, respectively). Scoring and ranking using the HYDE scoring function is adapted for high to low affinity ligands and can be routinely applied for medium size compounds collections. HYDE displayed a good performance in estimating the relative affinity of the FXR ligands in the datasets provided by the D3R. However, since ties can be obtained with ligands displaying similarly theoretical high affinities, a strict ranking cannot always be performed. Advanced rescoring at the cost of computational time using methods such as polarizable force field and molecular mechanics could bring a higher degree of accuracy for binding free energy prediction.

Acknowledgements MR is recipient of a MNESR fellowship. We thank Dr. Marcus Gastreich and BioSolveIT GmBH for providing SeeSAR.

References

1. Gathiaka S, Liu S, Chiu M, Yang H, Stuckey JA, Kang YN et al (2016) D3R Grand Challenge 2015: evaluation of protein–ligand pose and affinity predictions. J Comput Aided Mol Des 30:651–668
2. Gaieb Z, Liu S, Chui M, Yang H, Shao C, Gathiaka S et al (2016) Drug Design Data Resource Grand Challenge 2 Dataset: Farnesoid X Receptor. Drug Design Data Resource. University of California, San Diego
3. Jin L, Feng X, Rong H, Pan Z, Inaba Y, Qiu L et al (2013) The antiparasitic drug ivermectin is a novel FXR ligand that regulates metabolism. Nat Commun 4:1937
4. Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, Goodsell DS et al (2009) AutoDock4 and AutoDockTools4: automated docking with selective receptor flexibility. J Comput Chem 30:2785–2791
5. Schneider N, Lange G, Hindle S, Klein R, Rarey M (2013) A consistent description of HYdrogen bond and DEhydration energies in protein–ligand complexes: methods behind the HYDE scoring function. J Comput Aided Mol Des 27:15–29
6. O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR (2011) Open Babel: an open chemical toolbox. J Cheminf 3:33
7. SeeSAR v5.5, BioSolveIT GmbH, St.Augustin, Germany. http://www.biosolveit.de/SeeSAR
8. Bietz S, Urbaczek S, Schulz B, Rarey M (2014) Protoss: a holistic approach to predict tautomers and protonation states in protein–ligand complexes. J Cheminf 6:12

9. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H et al (2000) The Protein Data Bank. Nucleic Acids Res 28:235–242
10. De Lano W (2002) The PyMOL molecular graphics system. DeLano Scientific, Palo Alto. http://www.pymol.org
11. Dessau RB, Pipper CB (2008) [‘‘R"--project for statistical computing]. Ugeskr Laeger 170:328–330
12. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC et al (2004) UCSF Chimera—a visualization system for exploratory research and analysis. J Comput Chem 25:1605–1612
13. Olsson MH, Sondergaard CR, Rostkowski M, Jensen JH (2011) PROPKA3: Consistent treatment of internal and surface residues in empirical pKa predictions. J Chem Theory Comput 7:525–537
14. RDKit, Open-source cheminformatics. http://www.rdkit.org
15. Reulecke I, Lange G, Albrecht J, Klein R, Rarey M (2008) Towards an integrated description of hydrogen bonding and dehydration: decreasing false positives in virtual screening with the HYDE scoring function. ChemMedChem 3:885–897
16. Genheden S, Ryde U (2015) The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. Expert Opin Drug Discov 10:449–461
17. Wang J, Wang W, Kollman PA, Case DA (2006) Automatic atom type and bond type perception in molecular mechanical calculations. J Mol Graph Model 25:247–260
18. Wang J, Wolf RM, Caldwell JW, Kollman PA, Case DA (2004) Development and testing of a general amber force field. J Comput Chem 25:1157–1174
19. Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E, Villa E et al (2005) Scalable molecular dynamics with NAMD. J Comput Chem 26:1781–1802
20. Hornak V, Abel R, Okur A, Strockbine B, Roitberg A, Simmerling C (2006) Comparison of multiple Amber force fields and development of improved protein backbone parameters. Proteins 65:712–725
21. Onufriev A, Bashford D, Case DA (2004) Exploring protein native states and large-scale conformational changes with a modified generalized born model. Proteins 55:383–394
22. Weiser J, Shenkin P, Still WC (1999) Approximate atomic surfaces from linear combinations of pairwise overlaps (LCPO). J Comput Chem 20:217–230
23. Trott O, Olson AJ (2010) AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem 31:455–461

5.1.3 Conclusion et discussion

Le protocole proposé lors du D3R GC2 pour classer les molécules selon leur affinité pour FXR a montré de bonnes performances en comparaison aux autres protocoles proposés par la communauté scientifique (Figure 63). Cependant, la génération de poses repose partiellement sur une édition manuelle, ce protocole n'est pas automatisable et nécessite une connaissance du mode de liaison des molécules étudiées. Nous avons donc travaillé sur l'automatisation du protocole en générant des poses par docking avec les outils AutoDock 4.2 et AutoDock VINA. L'échantillonnage proposé par AutoDock VINA a permis de retrouver des poses proches des co-cristaux (<2Å) pour davantage de molécules que AutoDock 4.2 (30/36 contre 28/36). Par ailleurs, HYDE a classé ces poses en top 1 dans 23/30 cas pour les poses AutoDock VINA et 14/28 cas pour les poses AutoDock 4.2. Nous avons donc opté pour une automatisation du protocole avec AutoDock VINA, ce qui nous a permis d'obtenir des résultats légèrement meilleurs que le protocole initial (kendall $\tau = 0.44$ contre kendall $\tau = 0.41$). Les performances atteintes nous ont permis de valider le protocole de docking communément utilisé au sein du laboratoire avec les outils dont nous disposons.

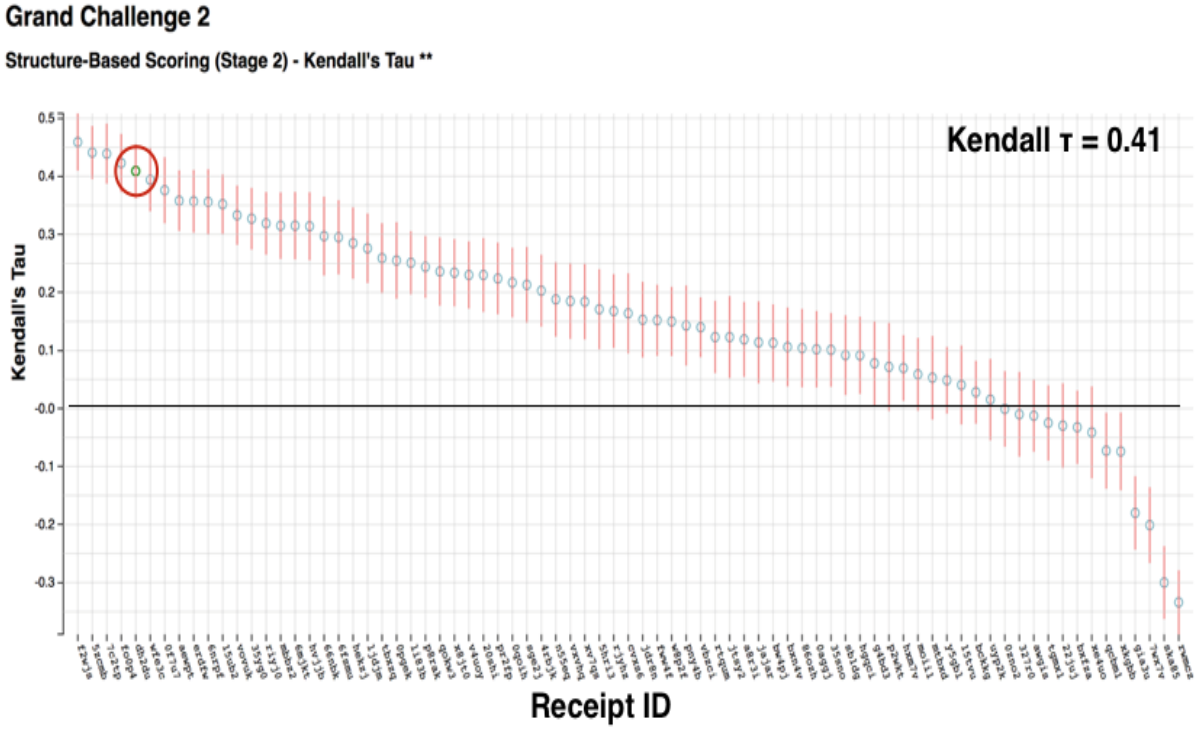


Figure 63 Graphique des corrélations de Kendall obtenues entre les classements des molécules soumis lors du stage 2 du a Grand Challenge 2 et les données expérimentales. Le protocole adopté au sein de notre laboratoire nous a permis d'obtenir une corrélation de Kendall $\tau = 0.41$ (cercle rouge).

5.2 Influence or Neuropilin-1 species on VEGF-A_{165}/NRP-1 platform screening of small inhibitory molecules exerting $-CH_3$ variation

5.2.1 Introduction

La Neuropiline 1 (NRP1) est un co-récepteur de l'isoforme a du facteur de croissance de l'endothélium vasculaire 165 ($VEFG_{165}a$) avec le récepteur VEGF-R2 (Figure 64)[469]. Elle représente une cible thérapeutique importante puisqu'elle est surexprimée dans divers cancers et est directement impliquée dans la migration et la survie des cellules cancéreuses, ainsi que dans l'angiogenèse[470,471,472,473]. Sa surexpression est notamment corrélée avec le potentiel métastatique de la tumeur. Biologiquement, l'interaction de NRP1 avec $VEFG_{165}a$ permet de renforcer l'interaction de cette dernière avec VEGF-R2, de promouvoir la transduction du signal et d'induire la vasodilatation, la survie cellulaire ainsi que la prolifération et la migration des cellules. Si NRP1 n'est pas essentielle à l'interaction de $VEFG_{165}a$ avec VEGF-R2, sa

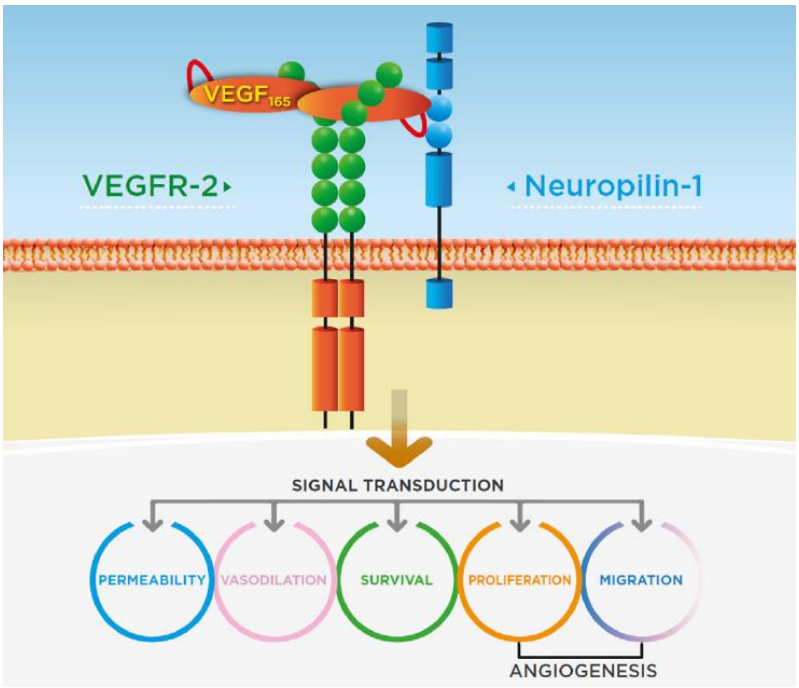


Figure 64 Schéma du complexe formé entre NRP-1, VEGF-R2 et VEGF. L'activation de VEGF-R2 VEGF-dépendante est amplifiée en présence de VEGF. Le signal transduit favorise la perméabilité de la cellule, la vasodilatation, la survie, la prolifération et la migration cellulaire. D'après [468]

présence multiplie par 4 voire 6 les interactions $VEFG_{165}a$/VEGF-R2 en comparaison aux

cellules ne possédant que VEGF-R2[474]. Plusieurs biothérapies ont été développées dans le but d'inhiber l'interaction protéine-protéine NRP1/ $VEFG_{165}a$ (Tableau 23), parmi lesquels des anticorps, des peptides et des pseudo-peptides ont été développés[475,476,477,478,479]. Cependant, la production d'anticorps reste coûteuse et les peptides et pseudo-peptides présentent des problèmes de stabilité in vivo. Le développement de nouvelles alternatives thérapeutiques revêt donc un intérêt majeur.

Tableau 23 Exemples de thérapie anti-VEGF approuvées par la FDA

Nom	**Type de molécule**	**Cible biologique**	**Stade de développement**	**Compagnie**
Ranibizumab	Fragment d'anticorps monoclonal humanisé	Tous les isoformes de VEGF-a	Approuvé	Novartis
Pegaptanib	Aptamer ARN	VEGF165-a	Approuvé	Bausch& Lomb
Bevacizumab	Anticorps monoclonal humanisé	Tous les isoformes de VEGF-a	Approuvé	Genentech
Aflibercept	Anticorps monoclonal humanisé	Tous les isoformes de VEGF-a	Approuvé	Regeneron Pharmaceuticals

Les laboratoires GBCM (Cnam), CiTCoM (CNRS), PSB, INSERM UMR 1163, CNRS ERL 8254, la faculté de chimie de Varsovie et l'institut Imagine travaillent conjointement à l'identification de petites molécules capable d'inhiber l'interaction $VEFG_{165}a$/VEGF-R2. En

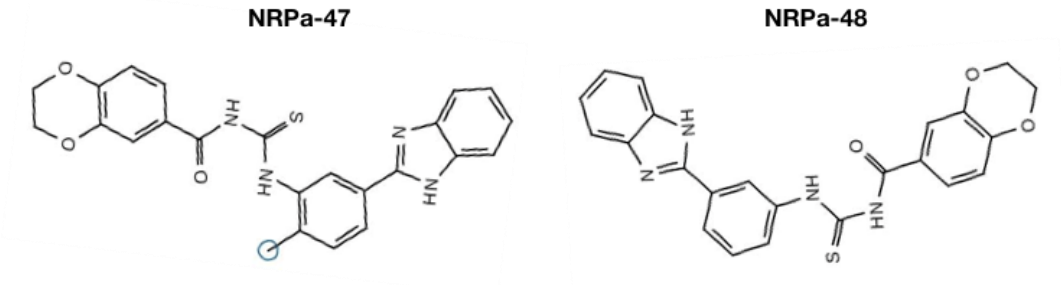


Figure 65 Structure des molécules NRPa-47 et NRPa-48. La NRPa-47 possède un groupement méthyle en position 2 du benzène qui est absent dans NRPa-48.

2014, la collaboration mise en place a permis d'identifier deux hits, NRPa-47 et NRPa-48 (Figure 65) à partir de la combinaison de criblage *in silico* et *in vitro*, via un test d'inhibition du complexe NRP1/$VEFG_{165}$a-biotinylé[480]. L'unique différence entre les molécules NRPa-47 et NRPa-48 réside dans la présence d'un groupement méthyle en position 2 du benzène sur le NRPa-47. La faible qualité des $VEFG_{165}$a-biotinylés ne permettait cependant pas d'obtenir des résultats reproductibles. Une nouvelle plateforme de criblage été mise en place depuis, et le $VEFG_{165}$a acheté chez Bio-Techne[481] a permis d'obtenir des résultats reproductibles. L'affinité de NRP-47 et NRP-48 pour des NRP1 issues de l'homme (h-NRP1), du rat (r-NRP1) et de la souris (mNRP1) a été évaluée. Les résultats montrent que NRP-47 et NRP-48 présentent des affinités du même ordre de grandeur pour h-NRP1 (IC50 = 21.5+/-3.9 µM et IC50 = 11.4+/-1.7 µM respectivement). Des affinités du même ordre sont retrouvées lorsque NRPa-47 est testé sur le r-NRP1 et m-NRP1, cependant des variations de 4 à 6.2 fois l'IC50 sont observées lorsque NRPa-48 est testée sur des NRP1 non humains. Nous avons réalisé des études de docking avec les molécules NRPa-47 et NRPa-48 pour comprendre ces résultats et nous avons effectué une analyse de population des poses obtenues pour étudier la stabilité des modes de liaison prédits. L'étude a été soumise à la publication sous le format d'un article court dans lequel les sections classiques (introduction, matériel et méthode, résultats, discussion et conclusion) sont confondues.

5.2.2 Article

Title: Influence of Neuropilin-1 species on VEGF-A_{165}/NRP-1 platform screening of small inhibitory molecules exerting $-CH_3$ variation

Short title: NRP-1 Platform screening depends of species and molecule structures

Authors:

Anna K. Puszko [a**], Manon Reau [b**], Luc Demange [c], Nicolas Lopez [d], Olivier Hermine [e,f,g], Françoise Raynaud [e,f,g], Matthieu Montes [b*], Yves Lepelletier [e,f,g*]

Author's affiliation:

[a] Faculty of Chemistry, University of Warsaw, Pasteura 1, 02-093 Warsaw, Poland.

[b] Laboratoire Génomique, Bioinformatique et Chimie Moléculaire, EA 7528, Conservatoire National des Arts et Métiers, 292 rue Saint Martin, 75003 Paris, France.

[c] CiTCoM, Université de Paris, CNRS UMR 8038, Faculté de Pharmacie, 4 avenue de l'Observatoire, 75006 Paris, France.

[d] PSB, 59 Rue Nationale, 75013 Paris.

[e] INSERM UMR 1163, Laboratory of cellular and molecular basis of normal hematopoiesis and hematological disorders: therapeutical implications, 24 boulevard Montparnasse 75015 Paris, France.

[f] Université de Paris, *Imagine* Institute, 24 boulevard Montparnasse 75015 Paris, France.

[g] CNRS ERL 8254, 24 boulevard Montparnasse 75015 Paris, France.

** These authors contributed equally to this work,

* These authors directed this work.

Correspondence should be addressed to:

-Yves Lepelletier, PhD: Phone:+33.1.42.75.42.83; E mail:y.lepelletier@gmail.com for biology

-Matthieu Montès, PhD; Phone:+33140272809; Email:matthieu.montes@cnam.fr for docking

Subject category: Analytical techniques for biological molecules

List of abbreviations: VEGF-A_{165}, Vascular endothelial growth factor-A_{165}; VEGF-R, Vascular endothelial growth factor receptor; NRP-1, Neuropilin-1; PBS, Phosphate buffered saline; DMSO, Dimethylsulfoxyde; BSA, Bovine serum albumin; (bt)-VEGF-A_{165}, Biotinylated-VEGF-A_{165}; NRPa, Neuropilin antagonist.

Abstract

Neuropilin-1 proved to be a relevant target to inhibit angiogenesis and tumor growth; we identified its firsts small-sized organic antagonists (NRPas) active *in vivo* thanks to an *in vitro* screening procedure. We report herein a more sensitive and efficient screening platform, allowing the evaluation of these NRPas on NRP-1 issued form different species, which is mandatory for further *in vivo* assay. This assay validates unambiguously NRPa-47 on NRP-1 issued from human, mouse and rat, conversely to NRPa-48 which appears sensitive to species change. This is reliable to small structural changes between the NRPas, as illustrated by a complementary docking study.

In this decade, the VEGF-A_{165} receptor Neuropilin-1 has emerged as a promising pharmacological target since it is over-expressed in several cancers and promotes tumour cell migration and survival [1]. More importantly, its over-expression also correlates with metastatic potential and poor prognosis in different cancer types [2]-[3]. In this way, antibodies, peptides and pseudo-peptides have been developed to antagonize VEGF-A_{165}/NRP-1 protein-protein interaction [4-11].

However, humanized antibody development is very costly and peptidic and/or pseudo-peptidic compounds rarely provide pharmacological agents due to their *in vivo* instability. For these raisons, we have initiated the development of small-sized and fully organic inhibitors targeting the VEGF-A_{165}/NRP-1 protein-protein interaction. In 2014, we characterized for the first time [12, 13] a family of Neuropilin-antagonist (NRPa), among them the two hit compounds, so-called NRPa-47 and NRPa-48, exerting *in vivo* anti-tumour efficiency on breast tumour cells xenografted mice. These molecules have been identified using a structure-based virtual ligand screening performed on 429623 molecules of the ChemBridge database followed by a screening of 1317 selected compounds on a VEGF-A_{165}/NRP-1 protein-protein interaction platform, as previously described. These molecules were evaluated as specific inhibitors of biotinylated (bt)-VEGF-A_{165}/NRP-1 binding. However, in that time, commercial (bt)-VEGF-A_{165} showed a poor quality and was unstable, since it was provided at low concentration and as a part of a kit. Consequently, the biological activities seemed to be "batch dependent" (R&D-Systems®, NFVEO kit). Nowadays, high quality and amount of (bt)-VEGF-A_{165} can be purchased from Bio-Techne® with reproducible batch activity. Thus, we have performed new tests using this (bt)-VEGF-A_{165} reference to improve our precedent results obtained with our synthesized and characterized NRPa-47 and NRPa-48 inhibitors [12, 13]. As reproductive, sensitive and rapid screening platform in this field is crucial, we have optimized our precedent protocol and proposed this following one.

In the first step, we have sequentially changed time exposures to develop our new high sensitive and reproducible screening platform assay as depicted below. A 96-well plate surface was coated overnight at 4°C with 100µL PBS containing 2 µg/mL of recombinant human-NRP-1 (h-NRP-1) (Bio-Techne®, France). Then, 3 washes with PBS 0.5% Tween 20 (buffer A), saturation with PBS 0.5% BSA (Sigma-Aldrich®, France) 1h at room temperature (R.T.) were done. Plates were subsequently incubated with different concentrations of each inhibitor in solution in DMSO (final DMSO concentrations between 0.1 to 0.5%, vol/vol, Sigma-Aldrich®) in presence of 50µL of (bt)-VEGF-A_{165} (400 ng/mL) (Bio-Techne®, France) supplemented with 4 µg/mL heparin (Sigma-Aldrich®) during 1h30 at R.T. Then, 3 washes with buffer A were done before the addition of 100µL of AMDEX streptavidin horseradish peroxidase (Amersham®, USA) diluted at 1:8000 in buffer A into each well. After 45 min. in darkness at R.T., plates were washed 3-times with buffer A and 100µL of SuperSignal West Pico Chemiluminescent Substrate (Pierce®, USA) were added. Chemiluminescence was immediately quantified with an EnVision™ 2101 Multilabel reader (Perkin Elmer®, USA). Data were analysed using the nonlinear regression function in Prism (GraphPad Software, USA). In summary, the new protocol technic might be performed in 3h15 in contrast to 5h before. The results obtained with this new version of our screening platform are closed similar to these obtained previously with the NRPa-47 and most sensitive for NRPa-48. Briefly, NRPa-47 IC_{50} is 21.5+/-3.9 *vs* 34+/- 3 µM and NRPa-48 IC_{50} is 11.4+/-1.7 *vs* 38+/-2 µM (Table 1 and sup Table 1).

In a second step, the study of NRPas in the cancer field requires the development of *in vivo* cancer model xenografted on mice and or rat. More importantly, the toxicity study of druggable compounds needs also *in vivo* tests to investigate potential side effects. Thus, taking together, we have tested NRPas on the new NRP-1 platform using different commercial recombinant NRP-1 specie origins as referenced in sup Table 2 to answer to these

crucial questions. To compare results with our platform described above, we didn't change the protocol excepted for the recombinant NRP-1 origins i.e. mouse (m-NRP-1) or rat (r-NRP-1) produced using NSO or SF21 cell models provided by Bio-Techne® (France), which are coated at the same amount concentration of proteins (sup Table 2). The validation of this NRP-1 species platform has been performed using Tuftsin binding inhibition. Independently of NRP-1 origins, Tuftsin similarly inhibits VEGF-A_{165}/NRP-1 binding (h-NRP-1, 53.3+/-1.2 µM; r-NRP-1/SF21, 45.1+/-2.3 µM; r-NRP-1/NSO, 42.5+/-2.8 µM; m-NRP-1, 48.7+/-1.4 µM) (sup Table 3). As shown in Table 1, efficiency of NRPa-47 is closed similar for all NRP-1 species (IC_{50} values: h-NRP-1, 21.5+/-3.9 µM; r-NRP-1/SF21, 18.2+/-4.0 µM; r-NRP-1/NSO, 8.8+/-1.4 µM; m-NRP-1, 28+/-1.0 µM). In details, a barely variation of IC_{50} value may be observed for NRPa-47 comprise between 0.4- to 0.85-fold increase for r-NRP-1 inhibition and 1.3-fold decrease for m-NRP-1 inhibition. In contrast, significant changes in IC_{50} values may be highlighted with NRPa-48, depending on the NRP-1 species (IC_{50} values: h-NRP-1, 11.4+/-1.7 µM; r-NRP-1/SF21, 71.0+/-3.4 µM; r-NRP-1/NSO, 56.2+/-4.3 µM; m-NRP-1, 46.6+/-1.4 µM). Thus, NRPa-48 showed an IC_{50} variation comprise between 4- to 6.2-fold decrease for non h-NRP-1. Taken together, both NRPa-47 and NRPa-48 may be used in human model, however only NRPa-47 may be used in mouse and/rat model to compare investigation in human biology. The difference of efficiency of these NRPas is exclusively reliable to the difference in their chemical structures, i.e. the presence (NRPa-47) or the absence (NRPa-48) of a methyl group on the central aromatic ring [12, 13]. To explain this result, we performed a new and substantial docking analysis of both antagonists.

We prepared NRPa-47 and NRPa-48 structures for docking as follow. The majority protonated micro specie at pH=7.4 (physiological) was computed with Marvin 17.22.0, 2017, ChemAxon (http://www.chemaxon.com), the lowest energy 3D conformation generated by iCon as implemented in LigandScout 4.3 was retained, and Gasteiger partial charges were

assigned using MGTools. The PDB structures of h-NRP-1 (2qqm), r-NRP-1 (2orz) and m-NRP-1 (4gz9) were selected and protonated at physiological pH with PDB2PQR (Version 2.11). Water molecules and heteroatoms were removed. The docking was performed using the gold standard AutoDock VINA within the known binding site of the Tuftsin and generated poses were re-scored with HYDE as implemented in SeeSAR 5.4 that consistently accounts for the hydrogen bonding, the hydrophobic effect and desolvation. Since VINA has proven good performances in ranking correctly native-like poses and HYDE scoring function brings more accuracy in affinity estimation, a consensus ranking of docked poses was performed giving the same weight to each method. We performed a population analysis of the top 5% and top 10% ranking poses: the similar poses (< 2 Å) were clustered using the Butina algorithm from RDKIT and the number of poses as well as the minimum HYDE score per cluster were retained. Of note, similar poses were obtained using the 3 structures, mostly because they share high structural similarity nearby the binding site (Figure 1A). The results on the h-NRP-1 structure (2qqm) illustrated in Figure 1B reveal that the best affinity score obtained is associated with a cluster of NRPa-47 poses that is also the most populated in the top 5%. Interestingly, the most similar pose for NRPa-48 (RMSD = 1,13 Å) also belongs to the highest populated cluster at 5% and 10%, nonetheless it is associated with lower affinity scores. Looking at the representative poses for each cluster at 5% (Figure 1C), we observe more variations in the best-ranked poses for NRPa-48 than NRPa-47. We suggest that the presence of a methyl group in NRPa-47 as compared to NRPa-48 could help driving the compound in a favourable position in the h-NRP-1 (Figure 1C, 1D), either because of its favourable interaction within the hydrophobic cavity of the NRP-1 binding site (Figure 1D), or because of its inherent steric constraints that could be responsible for a lower negative entropic contribution to the binding event, or both. In addition, the only residue laying the binding side that differs among species across species, S298 in the h-NRP-1 (G298 in the rat

and mouse), seems to stabilize the interaction with the top 1 pose of NRPa-47 (Figure 1D). The absence of both the methyl group plus the absence of this stabilizing interaction could explain the IC_{50} loss observed with NRPa-48 on the r-NRP-1 and m-NRP-1 as compared to the human and to NRPa-47 on all species.

Taking together, it is of utmost interest to validate the small-sized organic NRPas on NRP-1 issued from different specie, since we report herein that minor structural changes in the NRPas, such as a methyl group, modify in a relevant manner their binding with NRP-1. Thus, in future, study of this influence should be investigated on cell signalling regulation.

REFERENCES

[1] D.R. Bielenberg, C.A. Pettaway, S. Takashima and M. Klagsbrun, Neuropilins in neoplasms: expression, regulation, and function. Exp Cell Res 312 (2006) 584-93.

[2] G.J. Prud'homme and Y. Glinka, Neuropilins Are Multifunctional Coreceptors Involved in Tumor Initiation, Growth, Metastasis and Immunity. Oncotarget 3 (2012) 921-39.

[3] J.R. Wild, C.A. Staton, K. Chapple and B.M. Corfe, Neuropilins: expression and roles in the epithelium. Int J Exp Pathol 93 (2012) 81-103.

[4] Q. Pan, Y. Chanthery, W.C. Liang, S. Stawicki, J. Mak, et al., Blocking neuropilin-1 function has an additive effect with anti-VEGF to inhibit tumor growth, Cancer Cell 11 (2007) 53-67.

[5] D.G. Nowak, J. Woolard, E.M. Amin, O. Konopatskaya, M.A. et al., Expression of pro- and anti-angiogenic isoforms of VEGF is differentially regulated by splicing and growth factors. J Cell Sci 121 (2008) 3487-95.

[6] Y. Xin, S. Bai, L.A. Damico-Beyer, D. Jin, et al., Anti-neuropilin-1 (MNRP1685A): unexpected pharmacokinetic differences across species, from preclinical models to humans, Pharm. Res. 29 (2012) 2512-21.

[7] M. Caunt, J. Mak, W.C. Liang, S. Stawicki, et al., Blocking neuropilin-2 function inhibits tumor cell metastasis, Cancer Cell 13 (2008) 331-42.

[8] A. Starzec, P. Ladam, R. Vassy, S. Badache, et al., Structure-function analysis of the antiangiogenic ATWLPPR peptide inhibiting VEGF(165) binding to neuropilin-1 and molecular dynamics simulations of the ATWLPPR/neuropilin-1 complex, Peptides 28 (2007) 2397-402.

[9] A. Novoa, N. Pellegrini-Moise, D. Bechet, M. Barberi-Heyob and Y. Chapleur, Sugar-based peptidomimetics as potential inhibitors of the vascular endothelium growth factor binding to neuropilin-1, Bioorg. Med. Chem 18. (2010) 3285-98.

[10] C. Nasarre, M. Roth, L. Jacob, L. Roth, et al., Peptide-based interference of the transmembrane domain of neuropilin-1 inhibits glioma growth in vivo, Oncogene 29 (2010) 2381-92.

[11] A. Jarvis, C.K. Allerston, H. Jia, B. Herzog, et al., Small molecule inhibitors of the neuropilin-1 vascular endothelial growth factor A (VEGF-A) interaction, J. Med. Chem. 53 (2010) 2215-26.

[12] L. Borriello, M. Montès, Y. Lepelletier, B. Leforban, et al., Structure-based discovery of a small non-peptidic Neuropilins antagonist exerting in vitro and in vivo anti-tumor activity on breast cancer model, Cancer Lett. 349 (2014) 120-7.

[13] WQ. Liu, V. Megale, L. Borriello, B. Leforban, et al., Synthesis and structure-activity relationship of non-peptidic antagonists of neuropilin-1 receptor, Bioorg. Med. Chem. Lett. 24 (2014) 4254-9.

Compounds	NRP-1 origins (host)	NRP-1/VEGF-A_{165} binding inhibition [%]					IC_{50} [µM]
		100 µM	50 µM	25 µM	10 µM	1 µM	
NRPa-47	Human (NSO)	nd	59.9 ± 1.6	52.9 ± 0.8	28.3 ± 1.5	10.6 ± 1.2	**21.5 ± 3.9**
	Rat (NSO)	nd	71.9 ± 3.2	78.5 ± 1.1	56.3 ± 3.3	14.7 ± 3.3	**8.8 ± 1.4**
	Rat (SF21)	nd	59.9 ± 3.9	56.4 ± 1.7	28.0 ± 4.9	5.4 ± 3.8	**18.2 ± 4.0**
	Mouse (NSO)	80.7 ± 0.5	76.2 ± 1.4	43.0 ± 0.9	38.7 ± 1.0	0.0 ± 1.2	**28.0 ± 1.0**
NRPa-48	Human (NSO)	85.5 ± 1.4	68.5 ± 2.3	57.5 ± 1.6	53.5 ± 3.3	16.8 ± 2.7	**11.4 ± 1.7**
	Rat (NSO)	75.4 ± 2.3	45.5 ± 0.1	39.5 ± 2.4	27.1 ± 3.8	24.4 ± 2.8	**56.2 ± 4.3**
	Rat (SF21)	76.6 ± 0.4	43.0 ± 0.6	18.7 ± 0.8	19.7 ± 0.6	8.0 ± 4.6	**71.0 ± 3.4**
	Mouse (NSO)	76.2 ± 3.1	67.4 ± 0.5	25.5 ± 1.7	20.1 ± 1.2	14.9 ± 0.9	**46.6 ± 1.4**

Table 1: NRP-1/VEGF-A_{165} binding inhibition sensitivity.

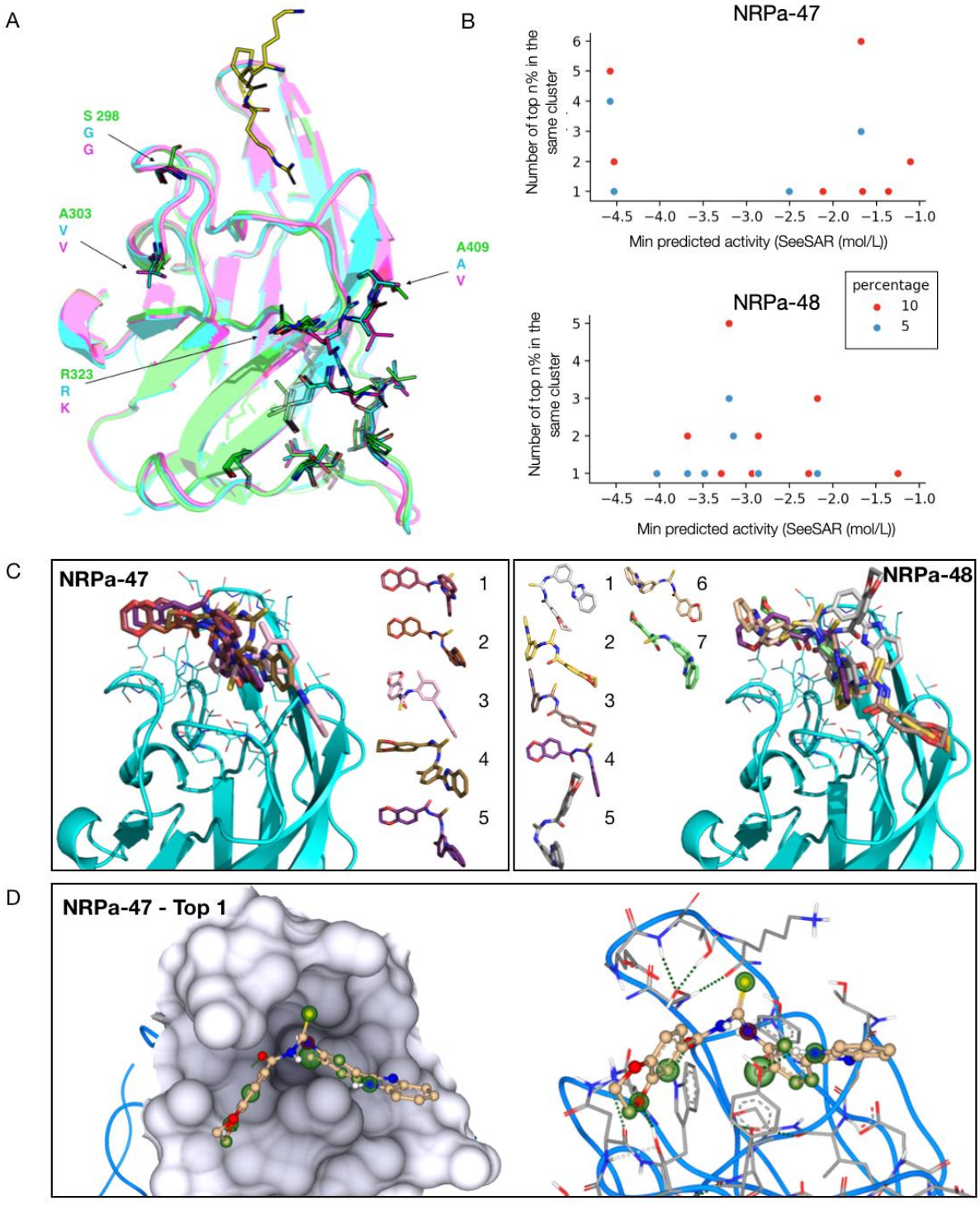


Figure 1. A- Superimposition of the human (2qqm - green), rat (2orz - blue) and mouse (4gz9 - magenta) NRP-1 structures show that the S298 residue from the human NRP-1 is the only that differs among species nearby the Tuftsin binding site. Results of the docking of NRPa-47 and NRPa-48 performed on the h-NRP-1 structure (2qqm). **B-** Each cluster is represented as the number of poses it represents in function of the minimum SeeSAR affinity score associated to the cluster for the first 5% and 10% ranked poses. **C-** The representative poses

of each cluster display more variations for the ligands NRPa-48 than NRPa-47. **D-** The atomic contribution for top 1 ranking pose of NRPa-47 computed with SeeSAR displays favorable contribution of the methyl group within the hydrophobic NRP-1 binding site and of the 1-4 dioxane with the S298 of h-NRP-1.

Acknowledgments:

We thank Prof. Jain for generously providing the Surflex package, and Molsoft LLC for providing academic licenses for the ICM suite. We thank COST Action CA15135: Multi-target paradigm for innovative ligand identification in the drug discovery process (MuTaLig), supported by COST (European Cooperation in Science and Technology) for a STSM grant that enabled AKP to work for one month at Imagine Institute in Paris.

Conflict of interest

All the authors have read the manuscript, concur with its content, and state that its content has not been submitted elsewhere. The authors declare no competing financial interests and no conflict of interest.

Sup Table 1: old platform technic

Technic used	Compounds	NRP-1 origin	NRP-1/VEGF-A_{165} binding inhibition [%]				IC_{50} [µM]
			100 µM	60 µM	40 µM	20 µM	
Old version	**NRPa-47**	Human	72 ± 4	61 ± 1	56 ± 1	48 ± 4	**34 ± 3**
	NRPa-48	Human	91 ± 7	75 ± 5	56 ± 3	43 ± 3	**38 ± 2**

Sup Table 2: Commercial NRP-1 characteristics

Origin species	Host cells	Accession numbers	Sequences	Tags	Forms	RhVEGF165 (K_D)	Ref
Human	NS0	NP_001019799	Phe22-Lys644	C-term 6-His	Mix of mono and homodimers	< 1 nM	3870-N1
Rat	NS0	Q9QWJ9	Phe22 - Ala810 /Arg/ Ser829 - Asp854	$HFcIgG_1$ (Pro100 - Lys330) /C-term 6-His	Disulfide-linked homodimer	< 1 nM	566-NNS
	SF21		Phe22 - Asp854 (Lys811Arg and Pro812 - Gly828 del)	$HFcIgG_1$ (Pro100 - Lys330) /C-term 6-His	Disulfide-linked homodimer	< 1 nM	566-N1
Mouse	NS0	NP_032763	Phe22-Pro856	C-term 6-His	Noncovalently-linked homodimer	< 1 nM	5994-N1

Sup Table 3: Tuftsin peptide efficiency on NRP-1 from different species

NRP-1 Origins (host) / Compound	Human (NSO)	Rat (NSO)	Rat (SF21)	Mouse (NSO)
Tuftsin (100 µM)	53.3 ± 1.2	42.5 ± 2.8	45.1 ± 2.3	48.7 ± 1.4

5.2.3 Conclusion et discussion

200 poses de NRPa-47 et NRPa-48 ont été générées par docking dans le site de liaison de la Tuftsin sur la h-NRP1 avec AutoDock VINA. L'étude de population de poses a permis d'identifier les poses redondantes parmi les 5% et 10% les mieux scorées selon un score concensus AutoDock VINA/HYDE. Les poses similaires (<2Å) ont été clusterisées et la pose associée au meilleur score du cluster a été retenue comme pose « représentative » du cluster ; seul le score de la pose représentative est retenu pour chaque cluster. Cette stratégie, en accord avec les travaux de Chang[482], repose sur l'hypothèse que la fréquence des poses retrouvées donne une information sur le paysage énergétique de l'interaction, et qu'une haute fréquence est associée à une entropie de liaison favorable. Le cluster de poses le plus peuplé de NRPa-47 à 5% correspond également à celui associé au meilleur score. Le cluster le plus peuplé pour NRPa-48 à 5% correspond à une pose proche (RMSD : 1.3Å), mais en revanche il est associé à un score inférieur. La superposition des poses représentatives de NRPa-47 et NRPa-48 montre une certaine constance dans les poses prédites de NRPa-47, ce qui n'est pas le cas pour NRPa-48 pour laquelle on observe d'importantes rotations de la molécule. Cette observation tend à conforter le mode de liaison associé au meilleur score de la NRPa-47. La superposition des structures h-NRP1 (PDB ID : 2qqm), r-NRP1 (PDB ID : 2orz) et m-NRP1 (PDB ID : 2orz) révèle par ailleurs un résidu qui diffère selon les espèces à proximité du site de liaison de la Tuftsin (h-NRP1 : S298, r-NRP1 et m-NRP1 : G298), et qui semble stabiliser la meilleure pose de NRPa-47 chez l'Homme. Une hypothèse est que la présence du méthyl sur NRPa-47 aide à guider la molécule dans le site de liaison grâce à son interaction favorable avec la cavité formée par la Y297, le W301 et la Y353. La molécule est aussi stabilisée une liaison hydrogène établie entre le 1,4-benzodioxine de NRPa-47 et la chaîne latérale de la S298 de la h-NRP1. L'absence combinée de l'interaction avec la S298 qui est substituée par une G298 sur r-NRP1 et m-NRP1 et du groupement méthyle semble responsable d'une baisse d'affinité pour la cible.

Cette étude a permis d'établir des hypothèses pour expliquer le changement d'affinité des molécules NRPa-47 et NRPa-48 sur les r-NRP1 et m-NRP1, malgré leur forte similarité structurale. L'étude sera poursuivie avec des test de signalisation cellulaire afin d'étudier le comportement *in vitro* de ces deux molécules.

Conclusion

Au cours de cette étude, nous avons étudié l'importance de l'intégration de molécules inactives dans la construction et l'évaluation de modèles de criblages virtuels. Nous avons focalisé notre étude sur la famille des récepteurs nucléaires, pour laquelle nous disposons de nombreuses références dans la littérature, et qui présente un intérêt thérapeutique et de santé publique. En effet, les récepteurs nucléaires sont impliqués dans de nombreuses maladies et sont la cible de composés capables de (dé)réguler le système endocrinien. Nous avons créé la banque de données NR-DBIND qui recense des données d'affinité et d'activité entre des petites molécules et des récepteurs nucléaires et qui inclut des données d'inactivité (nr-dbind.drugdesign.fr). Les données recensées dans la NR-DBIND présentent l'avantage d'avoir fait l'objet d'un contrôle manuel minutieux via une relecture de chacun des articles associés à la banque, ce qui assure la qualité du contenu. Si la publication massive de données devrait permettre, à moyen terme, de gommer le bruit inhérent à la fiabilité des données (marge d'erreurs des outils de mesure, erreur humaine, faux positif, faux négatif etc.), la qualité des données se révèle critique dès lors qu'elles ne sont pas nombreuses. Le travail de vérification et de nettoyage est néanmoins très couteux en temps ; c'est la raison pour laquelle nous avons souhaité mettre la NR-DBIND à la disposition de la communauté scientifique gratuitement. Son exploitation aussi bien par des chémoinformaticiens que des toxicologues ou biostatisticiens devrait constituer une source importante d'information dans la recherche thérapeutique et de santé publique.

Dans le cadre de mes travaux de thèse, la NR-DBIND a été exploitée pour répondre à la problématique initiale, qui pose la question de l'importance de l'intégration de données négatives dans la construction et l'évaluation de modèles.

En premier lieu, nous avons constaté que le pourcentage de molécules actives recensées dans la NR-DBIND est très élevé (5 à 74%) en comparaison aux études de HTS[450451] (0,1 à 4%[44]) et illustre le biais de publication qui tend à promouvoir la publication de résultats positifs et à sous-estimer l'information apportée par des résultats négatifs. Cette première observation présente une limite majeure aux travaux réalisés puisque le faible espace chimique couvert par les molécules inactives n'est pas représentatif d'un cas d'étude prospective et biaise donc inéluctablement l'interprétation des résultats. Seule la publication massive de données inactives devrait partiellement pallier ce problème. Nous avons décidé de porter ce discours auprès de la communauté de chimistes lors des journées Young Research Fellows Meeting organisés par la Société de Chimie Thérapeutique.

Afin d'amoindrir le biais intégré dans l'interprétation de nos résultats, nous avons conservé 9 jeux de données issus de la NR-DBIND pour lesquels au moins une molécule inactive est disponible par molécule active et une structure disponible. Nous avons effectué une étude comparative des outils de docking AutoDock VINA et PLANTS. Ces outils sont gratuits, applicables à haut débit et reposent sur des algorithmes de génération de poses très différents : le premier est basé sur un algorithme consistant en une succession de mutations stochastiques et d'optimisation locale alors que PLANTS utilise un algorithme d'optimisation de colonie de fourmis. L'évaluation des outils de docking a montré que PLANTS et AutoDock VINA ont des difficultés globales à discriminer les molécules actives des molécules inactives, mais présentent néanmoins des résultats satisfaisants avec des AUCs maximales et moyennes ≥ 0.7 dans 7/9 cas dans les conditions de docking optimales (docking sur structure seul/socking d'ensemble ; PLANTS/AutoDock VINA). Aucune condition de docking n'a été associée à un gain de performance pour l'ensemble des récepteurs, ce qui souligne la singularité de chacun des systèmes et la nécessité de les étudier au cas par cas. En revanche, le docking d'ensemble avec AutoDock VINA tend à minimiser le risque de s'orienter vers des structures non adaptées. La comparaison de l'utilisation de molécules inactives à l'utilisation de decoys dans le jeu de données d'inactifs révèle que des performances différentes sont obtenues, particulièrement lorsque les molécules inactives partagent peu de similarité structurale avec les molécules actives, ou au contraire, en partagent beaucoup. Si les molécules inactives permettent de s'affranchir du risque d'intégration de molécules actives dans le jeu de données d'inactifs, leur faible ratio n'empêche pas l'introduction de biais d'analogie. Nous en avons conclu que l'information qu'elles apportent peut être couplé à l'information apportée par les decoys qui couvrent un espace chimique plus large afin de sélectionner un modèle présentant de bonnes performances avec les deux jeux de données lors d'étude prospective, ce qui tend à minimiser l'impact des biais inhérents à chaque jeu de données. A terme, il conviendra d'étudier quelles combinaisons de métriques d'évaluation de la performance basée sur l'utilisation de molécules actives/inactives et actives/decoys permettent d'atteindre des performances optimales sur un jeu de données externe.

Par ailleurs, l'étude ici conduite sur la NR-DBIND gagnerait à être validée sur des jeux de données externes comportant idéalement de nouvelles molécules actives accompagnées de molécules inactives. Des données de HTS sont déjà disponibles pour certains récepteurs nucléaires (ex : AR, GR, PR, ERα/β) et pourront être exploitées.

L'intégration de molécules inactives dans notre protocole de modélisation de pharmacophores des agonistes d'AR nous a permis d'augmenter la sensibilité des modèles en les simplifiant, tout en contrôlant leur spécificité en limitant le criblage de molécules inactives. Les modèles de pharmacophore générés montrent de bonnes performances sur leurs jeux de données d'apprentissage et de test issus de la NR-DBIND, cependant ces performances ne sont pas reproductibles sur des jeux de données externes pour lesquelles la spécificité est conservée mais la sensibilité chute. Les modèles générés sont donc plus adaptés à la recherche de molécules thérapeutiques qu'à la prédiction d'effets indésirables pour laquelle une haute sensibilité est cruciale. Cette étude souligne que l'utilisation de molécules inactives est porteuse d'information mais n'est à ce jour pas suffisante du fait du peu de données dont nous disposons dans le cas des récepteurs nucléaires. L'intégration de molécules inactives dans l'évaluation et la construction de modèle de criblage gagnera en impact avec la publication de données d'inactivité plus nombreuses et plus diverses.

En ce sens, la NR-DBIND devrait être enrichie par l'apport de données issues de HTS publiées sur des sites comme PubChem Bioassay pour lesquelles un nettoyage des données en vue d'une utilisation en criblage virtuel présente un atout majeur. La quantité et la diversité des données testées lors de campagnes de HTS devraient permettre d'améliorer les performances des modèles entrainés sur un jeu d'apprentissage comme les modèles de pharmacophore utilisés au cours de cette thèse, ou le développement de modèles QSAR. L'étude du récepteur AR a notamment révélé que l'espace couvert par les molécules actives issues de données de HTS (Tox21) est plus large que celui couvert par la NR-DBIND. Nous avons initié des recherches prenant en compte l'ensemble de ces données dans le but de développer des modèles de prédiction des modulateurs des récepteurs nucléaires plus entraînés.

Ce projet de thèse conduit sur les récepteurs nucléaires est une ouverture à l'utilisation de molécules inactives dans l'évaluation et la construction de modèles de criblage virtuel qui devrait s'étendre à d'autres familles protéiques. Bien qu'il s'agisse d'une famille fortement étudiée, nous déplorons un manque de publication des données les concernant et particulièrement celles associés à des résultats dit « négatifs ». Cette tendance à publier les résultats positifs et par conséquent jugés plus valorisables constitue un biais de publication qui limite le progrès dans le développement de modèles de prédiction. Pour contrer ce biais, la valorisation d'une étude devrait non plus se baser uniquement sur le caractère positif de ses résultats mais bien sur la fiabilité du protocole et l'innovation liée aux recherches effectuées,

indépendamment du résultat. Si certains laboratoires conservent précieusement ces données pour des questions de confidentialité, de propriété intellectuelle et économiques, il existe des initiatives pour aider à la publication de données négatives pour ceux qui cherchent à les valoriser. Par exemple, le Journal of Pharmaceutical Negative Results, encourage la soumission d'études ayant conduit à un échec dans le domaine de la recherche pharmaceutique. Des sites de dépôt de données libres sont également disponibles, mais les données répertoriées ne font pas l'objet d'une relecture par les pairs. Une autre solution serait d'ajouter les tests ayant conduit à des échecs dans les informations supplémentaires liées à des publications relatant des résultats positifs ou encourageants.

Bibliographie

(1) 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.

(2) Rohrer, S. G.; Baumann, K. Maximum Unbiased Validation (MUV) Data Sets for Virtual Screening Based on PubChem Bioactivity Data. *J Chem Inf Model* **2009**, *49* (2), 169–184. https://doi.org/10.1021/ci8002649.

(3) Vogel, S. M.; Bauer, M. R.; Boeckler, F. M. DEKOIS: Demanding Evaluation Kits for Objective in Silico Screening--a Versatile Tool for Benchmarking Docking Programs and Scoring Functions. *J Chem Inf Model* **2011**, *51* (10), 2650–2665. https://doi.org/10.1021/ci2001549.

(4) Xia, J.; Jin, H.; Liu, Z.; Zhang, L.; Wang, X. S. An Unbiased Method to Build Benchmarking Sets for Ligand-Based Virtual Screening and Its Application to GPCRs. *J Chem Inf Model* **2014**, *54* (5), 1433–1450. https://doi.org/10.1021/ci500062f.

(5) Lagarde, N.; Ben Nasr, N.; Jérémie, A.; Guillemain, H.; Laville, V.; Labib, T.; Zagury, J.-F.; Montes, M. NRLiSt BDB, the Manually Curated Nuclear Receptors Ligands and Structures Benchmarking Database. *J. Med. Chem.* **2014**, *57* (7), 3117–3125. https://doi.org/10.1021/jm500132p.

(6) Xia, J.; Tilahun, E. L.; Kebede, E. H.; Reid, T.-E.; Zhang, L.; Wang, X. S. Comparative Modeling and Benchmarking Data Sets for Human Histone Deacetylases and Sirtuin Families. *J Chem Inf Model* **2015**, *55* (2), 374–388. https://doi.org/10.1021/ci5005515.

(7) Petrovska, B. B. Historical Review of Medicinal Plants' Usage. *Pharmacogn Rev* **2012**, *6* (11), 1–5. https://doi.org/10.4103/0973-7847.95849.

(8) Dev, S. Ancient-Modern Concordance in Ayurvedic Plants: Some Examples. *Environ. Health Perspect.* **1999**, *107* (10), 783–789. https://doi.org/10.1289/ehp.99107783.

(9) Borchardt, J. K. The Beginnings of Drug Therapy: Ancient Mesopotamian Medicine. *Drug News Perspect.* **2002**, *15* (3), 187–192.

(10) Wan-Loy, C.; Siew-Moi, P. Marine Algae as a Potential Source for Anti-Obesity Agents. *Mar Drugs* **2016**, *14* (12). https://doi.org/10.3390/md14120222.

(11) Wallace, M. S. Ziconotide: A New Nonopioid Intrathecal Analgesic for the Treatment of Chronic Pain. *Expert Rev Neurother* **2006**, *6* (10), 1423–1428. https://doi.org/10.1586/14737175.6.10.1423.

(12) Stierle, A.; Strobel, G.; Stierle, D. Taxol and Taxane Production by Taxomyces Andreanae, an Endophytic Fungus of Pacific Yew. *Science* **1993**, *260* (5105), 214–216.

(13) Li, J.-Y.; Sidhu, R. S.; Bollon, A.; Strobel, G. A. Stimulation of Taxol Production in

Liquid Cultures of Pestalotiopsis Microspora. *Mycological Research* **1998**, *102* (4), 461–464. https://doi.org/10.1017/S0953756297005078.

(14) Rates, S. M. K. Plants as Source of Drugs. *Toxicon* **2001**, *39* (5), 603–613. https://doi.org/10.1016/S0041-0101(00)00154-9.

(15) 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.; et al. An Analysis of the Attrition of Drug Candidates from Four Major Pharmaceutical Companies. *Nature Reviews Drug Discovery* **2015**, *14* (7), 475–486. https://doi.org/10.1038/nrd4609.

(16) Loo, J. A.; DeJohn, D. E.; Du, P.; Stevenson, T. I.; Ogorzalek Loo, R. R. Application of Mass Spectrometry for Target Identification and Characterization. *Med Res Rev* **1999**, *19* (4), 307–319.

(17) Borrel, A.; Regad, L.; Xhaard, H.; Petitjean, M.; Camproux, A.-C. PockDrug: A Model for Predicting Pocket Druggability That Overcomes Pocket Estimation Uncertainties. *Journal of Chemical Information and Modeling* **2015**, *55* (4), 882–895. https://doi.org/10.1021/ci5006004.

(18) Le Guilloux, V.; Schmidtke, P.; Tuffery, P. Fpocket: An Open Source Platform for Ligand Pocket Detection. *BMC Bioinformatics* **2009**, *10* (1), 168. https://doi.org/10.1186/1471-2105-10-168.

(19) Jorgensen, W. L. The Many Roles of Computation in Drug Discovery. *Science* **2004**, *303* (5665), 1813–1818. https://doi.org/10.1126/science.1096361.

(20) Van Norman, G. A. Drugs, Devices, and the FDA: Part 1: An Overview of Approval Processes for Drugs. *JACC: Basic to Translational Science* **2016**, *1* (3), 170–179. https://doi.org/10.1016/j.jacbts.2016.03.002.

(21) Prasad, V.; Mailankody, S. Research and Development Spending to Bring a Single Cancer Drug to Market and Revenues After Approval. *JAMA Intern Med* **2017**, *177* (11), 1569–1575. https://doi.org/10.1001/jamainternmed.2017.3601.

(22) Mestres, J.; Gregori-Puigjané, E.; Valverde, S.; Solé, R. V. The Topology of Drug-Target Interaction Networks: Implicit Dependence on Drug Properties and Target Families. *Mol Biosyst* **2009**, *5* (9), 1051–1057. https://doi.org/10.1039/b905821b.

(23) Zhang, W.; Liu, F.; Luo, L.; Zhang, J. Predicting Drug Side Effects by Multi-Label Learning and Ensemble Learning. *BMC Bioinformatics* **2015**, *16* (1). https://doi.org/10.1186/s12859-015-0774-y.

(24) Pauwels, E.; Stoven, V.; Yamanishi, Y. Predicting Drug Side-Effect Profiles: A Chemical Fragment-Based Approach. *BMC Bioinformatics* **2011**, *12*, 169.

https://doi.org/10.1186/1471-2105-12-169.

(25) Lauschke, V. M.; Hendriks, D. F. G.; Bell, C. C.; Andersson, T. B.; Ingelman-Sundberg, M. Novel 3D Culture Systems for Studies of Human Liver Function and Assessments of the Hepatotoxicity of Drugs and Drug Candidates. *Chem. Res. Toxicol.* **2016**, *29* (12), 1936–1955. https://doi.org/10.1021/acs.chemrestox.6b00150.

(26) Whitebread, S.; Hamon, J.; Bojanic, D.; Urban, L. Keynote Review: In Vitro Safety Pharmacology Profiling: An Essential Tool for Successful Drug Development. *Drug Discov. Today* **2005**, *10* (21), 1421–1433. https://doi.org/10.1016/S1359-6446(05)03632-9.

(27) Leedale, J.; Sharkey, K. J.; Colley, H. E.; Norton, Á. M.; Peeney, D.; Mason, C. L.; Sathish, J. G.; Murdoch, C.; Sharma, P.; Webb, S. D. A Combined In Vitro/In Silico Approach to Identifying Off-Target Receptor Toxicity. *iScience* **2018**, *4*, 84–96. https://doi.org/10.1016/j.isci.2018.05.012.

(28) Fisher, J. S. Environmental Anti-Androgens and Male Reproductive Health: Focus on Phthalates and Testicular Dysgenesis Syndrome. *Reproduction* **2004**, *127* (3), 305–315. https://doi.org/10.1530/rep.1.00025.

(29) Schug, T. T.; Janesick, A.; Blumberg, B.; Heindel, J. J. Endocrine Disrupting Chemicals and Disease Susceptibility. *J Steroid Biochem Mol Biol* **2011**, *127* (3–5), 204–215. https://doi.org/10.1016/j.jsbmb.2011.08.007.

(30) Zoeller, R. T.; Vandenberg, L. N. Assessing Dose–Response Relationships for Endocrine Disrupting Chemicals (EDCs): A Focus on Non-Monotonicity. *Environ Health* **2015**, *14*. https://doi.org/10.1186/s12940-015-0029-4.

(31) Braun, J. M. Early-Life Exposure to EDCs: Role in Childhood Obesity and Neurodevelopment. *Nat Rev Endocrinol* **2017**, *13* (3), 161–173. https://doi.org/10.1038/nrendo.2016.186.

(32) Thomas, R. S.; Paules, R. S.; Simeonov, A.; Fitzpatrick, S. C.; Crofton, K. M.; Casey, W. M.; Mendrick, D. L. The US Federal Tox21 Program: A Strategic and Operational Plan for Continued Leadership. *ALTEX* **2018**, *35* (2), 163–168. https://doi.org/10.14573/altex.1803011.

(33) 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.

(34) Vedani, A.; Dobler, M.; Smieško, M. VirtualToxLab — A Platform for Estimating the Toxic Potential of Drugs, Chemicals and Natural Products. *Toxicology and Applied Pharmacology* **2012**, *261* (2), 142–153. https://doi.org/10.1016/j.taap.2012.03.018.

(35) Van Drie, J. H. Computer-Aided Drug Design: The next 20 Years. *J. Comput. Aided Mol. Des.* **2007**, *21* (10–11), 591–601. https://doi.org/10.1007/s10822-007-9142-y.

(36) Trends in the cost of computing https://aiimpacts.org/trends-in-the-cost-of-computing/ (accessed Mar 19, 2019).

(37) Lolli, M.; Narramore, S.; Fishwick, C. W. G.; Pors, K. Refining the Chemical Toolbox to Be Fit for Educational and Practical Purpose for Drug Discovery in the 21st Century. *Drug Discov. Today* **2015**, *20* (8), 1018–1026. https://doi.org/10.1016/j.drudis.2015.04.010.

(38) Sliwoski, G.; Kothiwale, S.; Meiler, J.; Lowe, E. W. Computational Methods in Drug Discovery. *Pharmacol Rev* **2014**, *66* (1), 334–395. https://doi.org/10.1124/pr.112.007336.

(39) Volochnyuk, D. M.; Ryabukhin, S. V.; Moroz, Y. S.; Savych, O.; Chuprina, A.; Horvath, D.; Zabolotna, Y.; Varnek, A.; Judd, D. B. Evolution of Commercially Available Compounds for HTS. *Drug Discovery Today* **2019**, *24* (2), 390–402. https://doi.org/10.1016/j.drudis.2018.10.016.

(40) Boisclair, M. D.; Egan, D. A.; Huberman, K.; Infantino, R. High-Throughput Screening in Industry. In *Anticancer Drug Development Guide*; Teicher, B. A., Andrews, P. A., Eds.; Humana Press: Totowa, NJ, 2004; pp 23–39. https://doi.org/10.1007/978-1-59259-739-0_2.

(41) Dorr, P.; Westby, M.; Dobbs, S.; Griffin, P.; Irvine, B.; Macartney, M.; Mori, J.; Rickett, G.; Smith-Burchnell, C.; Napier, C.; et al. Maraviroc (UK-427,857), a Potent, Orally Bioavailable, and Selective Small-Molecule Inhibitor of Chemokine Receptor CCR5 with Broad-Spectrum Anti-Human Immunodeficiency Virus Type 1 Activity. *Antimicrob. Agents Chemother.* **2005**, *49* (11), 4721–4732. https://doi.org/10.1128/AAC.49.11.4721-4732.2005.

(42) Erickson-Miller, C. L.; DeLorme, E.; Tian, S.-S.; Hopson, C. B.; Stark, K.; Giampa, L.; Valoret, E. I.; Duffy, K. J.; Luengo, J. L.; Rosen, J.; et al. Discovery and Characterization of a Selective, Nonpeptidyl Thrombopoietin Receptor Agonist. *Exp. Hematol.* **2005**, *33* (1), 85–93. https://doi.org/10.1016/j.exphem.2004.09.006.

(43) Gao, M.; Nettles, R. E.; Belema, M.; Snyder, L. B.; Nguyen, V. N.; Fridell, R. A.; Serrano-Wu, M. H.; Langley, D. R.; Sun, J.-H.; O'Boyle, D. R.; et al. Chemical Genetics Strategy Identifies an HCV NS5A Inhibitor with a Potent Clinical Effect. *Nature* **2010**, *465* (7294), 96–100. https://doi.org/10.1038/nature08960.

(44) Roy, A. Early Probe and Drug Discovery in Academia: A Minireview. *High Throughput* **2018**, *7* (1). https://doi.org/10.3390/ht7010004.

(45) Furger, C. *Les tests cellulaires: De la recherche aux applications industrielles en toxicité et santé*; ISTE Editions, 2016.

(46) Doman, T. N.; McGovern, S. L.; Witherbee, B. J.; Kasten, T. P.; Kurumbail, R.;

Stallings, W. C.; Connolly, D. T.; Shoichet, B. K. Molecular Docking and High-Throughput Screening for Novel Inhibitors of Protein Tyrosine Phosphatase-1B. *J. Med. Chem.* **2002**, *45* (11), 2213–2221.

(47) Bissantz, C.; Folkers, G.; Rognan, D. Protein-Based Virtual Screening of Chemical Databases. 1. Evaluation of Different Docking/Scoring Combinations. *J. Med. Chem.* **2000**, *43* (25), 4759–4767.

(48) Sawyer, J. S.; Anderson, B. D.; Beight, D. W.; Campbell, R. M.; Jones, M. L.; Herron, D. K.; Lampe, J. W.; McCowan, J. R.; McMillen, W. T.; Mort, N.; et al. Synthesis and Activity of New Aryl- and Heteroaryl-Substituted Pyrazole Inhibitors of the Transforming Growth Factor-Beta Type I Receptor Kinase Domain. *J. Med. Chem.* **2003**, *46* (19), 3953–3956. https://doi.org/10.1021/jm0205705.

(49) Singh, J.; Chuaqui, C. E.; Boriack-Sjodin, P. A.; Lee, W. C.; Pontz, T.; Corbley, M. J.; Cheung, H.-K.; Arduini, R. M.; Mead, J. N.; Newman, M. N.; et al. Successful Shape-Based Virtual Screening: The Discovery of a Potent Inhibitor of the Type I TGFbeta Receptor Kinase (TbetaRI). *Bioorg. Med. Chem. Lett.* **2003**, *13* (24), 4355–4359.

(50) Ruddigkeit, L.; van Deursen, R.; Blum, L. C.; Reymond, J.-L. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *J. Chem. Inf. Model.* **2012**, *52* (11), 2864–2875. https://doi.org/10.1021/ci300415d.

(51) Irwin, J. J.; Shoichet, B. K. ZINC--a Free Database of Commercially Available Compounds for Virtual Screening. *J Chem Inf Model* **2005**, *45* (1), 177–182. https://doi.org/10.1021/ci049714+.

(52) Sterling, T.; Irwin, J. J. ZINC 15 – Ligand Discovery for Everyone. *J. Chem. Inf. Model.* **2015**, *55* (11), 2324–2337. https://doi.org/10.1021/acs.jcim.5b00559.

(53) Kamisetty, H.; Ovchinnikov, S.; Baker, D. Assessing the Utility of Coevolution-Based Residue-Residue Contact Predictions in a Sequence- and Structure-Rich Era. *Proc. Natl. Acad. Sci. U.S.A.* **2013**, *110* (39), 15674–15679. https://doi.org/10.1073/pnas.1314045110.

(54) Levinson, N. M.; Boxer, S. G. A Conserved Water-Mediated Hydrogen Bond Network Defines Bosutinib's Kinase Selectivity. *Nat. Chem. Biol.* **2014**, *10* (2), 127–132. https://doi.org/10.1038/nchembio.1404.

(55) Laurent, B.; Murail, S.; Shahsavar, A.; Sauguet, L.; Delarue, M.; Baaden, M. Sites of Anesthetic Inhibitory Action on a Cationic Ligand-Gated Ion Channel. *Structure* **2016**, *24* (4), 595–605. https://doi.org/10.1016/j.str.2016.02.014.

(56) Filippakopoulos, P.; Qi, J.; Picaud, S.; Shen, Y.; Smith, W. B.; Fedorov, O.; Morse, E. M.; Keates, T.; Hickman, T. T.; Felletar, I.; et al. Selective Inhibition of BET Bromodomains.

Nature **2010**, *468* (7327), 1067–1073. https://doi.org/10.1038/nature09504.

(57) Volkamer, A.; Kuhn, D.; Rippmann, F.; Rarey, M. DoGSiteScorer: A Web Server for Automatic Binding Site Prediction, Analysis and Druggability Assessment. *Bioinformatics* **2012**, *28* (15), 2074–2075.

(58) Le Guilloux, V.; Schmidtke, P.; Tuffery, P. Fpocket: An Open Source Platform for Ligand Pocket Detection. *BMC Bioinformatics* **2009**, *10*, 168. https://doi.org/10.1186/1471-2105-10-168.

(59) Meng, X.-Y.; Zhang, H.-X.; Mezei, M.; Cui, M. Molecular Docking: A Powerful Approach for Structure-Based Drug Discovery. *Curr Comput Aided Drug Des* **2011**, *7* (2), 146–157.

(60) Success Stories of Computer-Aided Design https://www.researchgate.net/publication/228050687_Success_Stories_of_Computer-Aided_Design (accessed Mar 20, 2019).

(61) Mouhsine, H.; Guillemain, H.; Moreau, G.; Fourati, N.; Zerrouki, C.; Baron, B.; Desallais, L.; Gizzi, P.; Ben Nasr, N.; Perrier, J.; et al. Identification of an in Vivo Orally Active Dual-Binding Protein-Protein Interaction Inhibitor Targeting TNFα through Combined in Silico/in Vitro/in Vivo Screening. *Sci Rep* **2017**, *7*. https://doi.org/10.1038/s41598-017-03427-z.

(62) Andaloussi, M.; Henriksson, L. M.; Więckowska, A.; Lindh, M.; Björkelid, C.; Larsson, A. M.; Suresh, S.; Iyer, H.; Srinivasa, B. R.; Bergfors, T.; et al. Design, Synthesis, and X-Ray Crystallographic Studies of α-Aryl Substituted Fosmidomycin Analogues as Inhibitors of Mycobacterium Tuberculosis 1-Deoxy-D-Xylulose 5-Phosphate Reductoisomerase. *J. Med. Chem.* **2011**, *54* (14), 4964–4976. https://doi.org/10.1021/jm2000085.

(63) Umeda, T.; Tanaka, N.; Kusakabe, Y.; Nakanishi, M.; Kitade, Y.; Nakamura, K. T. Crystallization and Preliminary X-Ray Crystallographic Study of 1-Deoxy-D-Xylulose 5-Phosphate Reductoisomerase from Plasmodium Falciparum. *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun.* **2010**, *66* (Pt 3), 330–332. https://doi.org/10.1107/S1744309110001739.

(64) Chaudhary, K. K.; Prasad, C. V. S. S. Virtual Screening of Compounds to 1-Deoxy-Dxylulose 5-Phosphate Reductoisomerase (DXR) from Plasmodium Falciparum. *Bioinformation* **2014**, *10* (6), 358–364. https://doi.org/10.6026/97320630010358.

(65) 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 Bioinformatics* **2014**, *15* (Suppl 11), S4. https://doi.org/10.1186/1471-2105-15-S11-S4.

(66) Nose, T.; Tokunaga, T.; Shimohigashi, Y. Exploration of Endocrine-Disrupting Chemicals on Estrogen Receptor Alpha by the Agonist/Antagonist Differential-Docking Screening (AADS) Method: 4-(1-Adamantyl)Phenol as a Potent Endocrine Disruptor Candidate. *Toxicol. Lett.* **2009**, *191* (1), 33–39. https://doi.org/10.1016/j.toxlet.2009.08.001.

(67) Celik, L.; Davey, J.; Lund, D.; Schiøtt, B. Exploring Interactions of Endocrine-Disrupting Compounds with Different Conformations of the Human Estrogen Receptor Alpha Ligand Binding Domain: A Molecular Docking Study. *Chem. Res. Toxicol.* **2008**, *21* (11), 2195–2206. https://doi.org/10.1021/tx800278d.

(68) Rella, M.; Rushworth, C. A.; Guy, J. L.; Turner, A. J.; Langer, T.; Jackson, R. M. Structure-Based Pharmacophore Design and Virtual Screening for Novel Angiotensin Converting Enzyme 2 Inhibitors. *J. Chem. Inf. Model.* **2006**, *46* (2), 708–716. https://doi.org/10.1021/ci0503614.

(69) Chen, C.; Wang, T.; Wu, F.; Huang, W.; He, G.; Ouyang, L.; Xiang, M.; Peng, C.; Jiang, Q. Combining Structure-Based Pharmacophore Modeling, Virtual Screening, and in Silico ADMET Analysis to Discover Novel Tetrahydro-Quinoline Based Pyruvate Kinase Isozyme M2 Activators with Antitumor Activity. *Drug Des Devel Ther* **2014**, *8*, 1195–1210. https://doi.org/10.2147/DDDT.S62921.

(70) Liao, H.-S.; Liu, H.-L.; Chen, W.-H.; Ho, Y. Structure-Based Pharmacophore Modeling and Virtual Screening to Identify Novel Inhibitors for Anthrax Lethal Factor. *Med Chem Res* **2014**, *23* (8), 3725–3732. https://doi.org/10.1007/s00044-014-0947-7.

(71) Schneider, G.; Fechner, U. Computer-Based *de Novo* Design of Drug-like Molecules. *Nature Reviews Drug Discovery* **2005**, *4* (8), 649–663. https://doi.org/10.1038/nrd1799.

(72) Niu, Y.; Shi, D.; Li, L.; Guo, J.; Liu, H.; Yao, X. Revealing Inhibition Difference between PFI-2 Enantiomers against SETD7 by Molecular Dynamics Simulations, Binding Free Energy Calculations and Unbinding Pathway Analysis. *Sci Rep* **2017**, *7*. https://doi.org/10.1038/srep46547.

(73) Wakui, N.; Yoshino, R.; Yasuo, N.; Ohue, M.; Sekijima, M. Exploring the Selectivity of Inhibitor Complexes with Bcl-2 and Bcl-XL: A Molecular Dynamics Simulation Approach. *Journal of Molecular Graphics and Modelling* **2018**, *79*, 166–174. https://doi.org/10.1016/j.jmgm.2017.11.011.

(74) Kumar, A.; Zhang, K. Y. J. Advances in the Development of Shape Similarity Methods and Their Application in Drug Discovery. *Front Chem* **2018**, *6*. https://doi.org/10.3389/fchem.2018.00315.

(75) Kunimoto, R.; Bajorath, J. Combining Similarity Searching and Network Analysis for

the Identification of Active Compounds. *ACS Omega* **2018**, *3* (4), 3768–3777. https://doi.org/10.1021/acsomega.8b00344.

(76) Vuorinen, A.; Engeli, R.; Meyer, A.; Bachmann, F.; Griesser, U. J.; Schuster, D.; Odermatt, A. Ligand-Based Pharmacophore Modeling and Virtual Screening for the Discovery of Novel 17β-Hydroxysteroid Dehydrogenase 2 Inhibitors. *J Med Chem* **2014**, *57* (14), 5995–6007. https://doi.org/10.1021/jm5004914.

(77) Che Jinxin; Wang Zhilong; Sheng Haichao; Huang Feng; Dong Xiaowu; Hu Youhong; Xie Xin; Hu Yongzhou. Ligand-Based Pharmacophore Model for the Discovery of Novel CXCR2 Antagonists as Anti-Cancer Metastatic Agents. *Royal Society Open Science 5* (7), 180176. https://doi.org/10.1098/rsos.180176.

(78) Marriott, D. P.; Dougall, I. G.; Meghani, P.; Liu, Y. J.; Flower, D. R. Lead Generation Using Pharmacophore Mapping and Three-Dimensional Database Searching: Application to Muscarinic M(3) Receptor Antagonists. *J. Med. Chem.* **1999**, *42* (17), 3210–3216. https://doi.org/10.1021/jm980409n.

(79) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; et al. ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. *Nucleic Acids Res* **2012**, *40* (Database issue), D1100–D1107. https://doi.org/10.1093/nar/gkr777.

(80) Wang, Y.; Bryant, S. H.; Cheng, T.; Wang, J.; Gindulyte, A.; Shoemaker, B. A.; Thiessen, P. A.; He, S.; Zhang, J. PubChem BioAssay: 2017 Update. *Nucleic Acids Res.* **2017**, *45* (D1), D955–D963. https://doi.org/10.1093/nar/gkw1118.

(81) 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.* **2018**. https://doi.org/10.1021/acs.jmedchem.8b01105.

(82) Polishchuk, P. G.; Madzhidov, T. I.; Varnek, A. Estimation of the Size of Drug-like Chemical Space Based on GDB-17 Data. *J Comput Aided Mol Des* **2013**, *27* (8), 675–679. https://doi.org/10.1007/s10822-013-9672-4.

(83) 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. *Advanced Drug Delivery Reviews* **1997**, *23* (1), 3–25. https://doi.org/10.1016/S0169-409X(96)00423-1.

(84) Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrián-Uhalte, E.; et al. The ChEMBL Database in

2017. *Nucleic Acids Res* **2017**, *45* (Database issue), D945–D954. https://doi.org/10.1093/nar/gkw1074.

(85) Wu, J. J. Comparison of SPA, FRET, and FP for Kinase Assays. In *High Throughput Screening: Methods and Protocols*; Janzen, W. P., Ed.; Methods in Molecular Biology™; Humana Press: Totowa, NJ, 2002; pp 65–85. https://doi.org/10.1385/1-59259-180-9:065.

(86) Bienstock, R. J. Overview: Fragment-Based Drug Design. In *Library Design, Search Methods, and Applications of Fragment-Based Drug Design*; ACS Symposium Series; American Chemical Society, 2011; Vol. 1076, pp 1–26. https://doi.org/10.1021/bk-2011-1076.ch001.

(87) Zoete, V.; Grosdidier, A.; Michielin, O. Docking, Virtual High Throughput Screening and in Silico Fragment-Based Drug Design. *J. Cell. Mol. Med.* **2009**, *13* (2), 238–248. https://doi.org/10.1111/j.1582-4934.2008.00665.x.

(88) Murray, C. W.; Rees, D. C. The Rise of Fragment-Based Drug Discovery. *Nature Chemistry* **2009**, *1* (3), 187–192. https://doi.org/10.1038/nchem.217.

(89) Visini, R.; Awale, M.; Reymond, J.-L. Fragment Database FDB-17. *J. Chem. Inf. Model.* **2017**, *57* (4), 700–709. https://doi.org/10.1021/acs.jcim.7b00020.

(90) Cayley, E. Ueber Die Analytischen Figuren, Welche in Der Mathematik Bäume Genannt Werden Und Ihre Anwendung Auf Die Theorie Chemischer Verbindungen. *Berichte der deutschen chemischen Gesellschaft* **1875**, *8* (2), 1056–1059. https://doi.org/10.1002/cber.18750080252.

(91) Chevillard, F.; Kolb, P. SCUBIDOO: A Large yet Screenable and Easily Searchable Database of Computationally Created Chemical Compounds Optimized toward High Likelihood of Synthetic Tractability. *J Chem Inf Model* **2015**, *55* (9), 1824–1835. https://doi.org/10.1021/acs.jcim.5b00203.

(92) Bax, B.; Chung, C.; Edge, C. Getting the Chemistry Right: Protonation, Tautomers and the Importance of H Atoms in Biological Chemistry. *Acta Crystallogr D Struct Biol* **2017**, *73* (Pt 2), 131–140. https://doi.org/10.1107/S2059798316020283.

(93) Software, O. S. QUACPAC | Database Preparation https://www.eyesopen.com/quacpac (accessed Mar 25, 2019).

(94) Certara https://www.certara.com/ (accessed Mar 25, 2019).

(95) BIOVIA - Scientific Enterprise Software for Chemical Research, Material Science R&D http://www.3dsbiovia.com/ (accessed Mar 25, 2019).

(96) Welcome to MN-AM | MN-AM https://www.mn-am.com/ (accessed Mar 25, 2019).

(97) Aki-Sener, E.; Yalcin, I. 15th European Symposium on Quantitative Structure-Activity

Relationships and Molecular Modelling (Euro-QSAR 2004). *QSAR & Combinatorial Science* **2005**, *24* (4), 441–441. https://doi.org/10.1002/qsar.200590024.

(98) Schrödinger | Schrödinger is the scientific leader in developing state-of-the-art chemical simulation software for use in pharmaceutical, biotechnology, and materials research. https://www.schrodinger.com/ (accessed Mar 25, 2019).

(99) Bietz, S.; Urbaczek, S.; Schulz, B.; Rarey, M. Protoss: A Holistic Approach to Predict Tautomers and Protonation States in Protein-Ligand Complexes. *Journal of Cheminformatics* **2014**, *6* (1), 12. https://doi.org/10.1186/1758-2946-6-12.

(100) Epik: A software program for pKa prediction and protonation state generation for drug-like molecules | Request PDF https://www.researchgate.net/publication/5946120_Epik_A_software_program_for_pKa_prediction_and_protonation_state_generation_for_drug-like_molecules (accessed Mar 25, 2019). http://dx.doi.org/10.1007/s10822-007-9133-z.

(101) Hawkins, P. C. D. Conformation Generation: The State of the Art. *J. Chem. Inf. Model.* **2017**, *57* (8), 1747–1756. https://doi.org/10.1021/acs.jcim.7b00221.

(102) Rose, P. W.; Prlić, A.; Altunkaya, A.; Bi, C.; Bradley, A. R.; Christie, C. H.; Costanzo, L. D.; Duarte, J. M.; Dutta, S.; Feng, Z.; et al. The RCSB Protein Data Bank: Integrative View of Protein, Gene and 3D Structural Information. *Nucleic Acids Res* **2017**, *45* (Database issue), D271–D281. https://doi.org/10.1093/nar/gkw1000.

(103) Watson, D. G. The Cambridge Structural Database (CSD): Current Activities and Future Plans. *J Res Natl Inst Stand Technol* **1996**, *101* (3), 227–229. https://doi.org/10.6028/jres.101.024.

(104) Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C. The Cambridge Structural Database. *Acta Crystallogr B Struct Sci Cryst Eng Mater* **2016**, *72* (Pt 2), 171–179. https://doi.org/10.1107/S2052520616003954.

(105) Halgren, T. A. Merck Molecular Force Field. I. Basis, Form, Scope, Parameterization, and Performance of MMFF94. *Journal of Computational Chemistry* **1996**, *17* (5-6), 490–519. https://doi.org/10.1002/(SICI)1096-987X(199604)17:5/6<490::AID-JCC1>3.0.CO;2-P.

(106) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.* **1996**, *118* (45), 11225–11236. https://doi.org/10.1021/ja9621760.

(107) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *Journal of Computational Chemistry* **1983**, *4* (2), 187–217.

https://doi.org/10.1002/jcc.540040211.

(108) Clark, M.; Cramer, R. D.; Opdenbosch, N. V. Validation of the General Purpose Tripos 5.2 Force Field. *Journal of Computational Chemistry* **1989**, *10* (8), 982–1012. https://doi.org/10.1002/jcc.540100804.

(109) 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.

(110) Friedrich, N.-O.; de Bruyn Kops, C.; Flachsenberg, F.; Sommer, K.; Rarey, M.; Kirchmair, J. Benchmarking Commercial Conformer Ensemble Generators. *J. Chem. Inf. Model.* **2017**, *57* (11), 2719–2728. https://doi.org/10.1021/acs.jcim.7b00505.

(111) Friedrich, N.-O.; Meyder, A.; de Bruyn Kops, C.; Sommer, K.; Flachsenberg, F.; Rarey, M.; Kirchmair, J. High-Quality Dataset of Protein-Bound Ligand Conformations and Its Application to Benchmarking Conformer Ensemble Generators. *J. Chem. Inf. Model.* **2017**, *57* (3), 529–539. https://doi.org/10.1021/acs.jcim.6b00613.

(112) Vainio, M. J.; Johnson, M. S. Generating Conformer Ensembles Using a Multiobjective Genetic Algorithm. *J. Chem. Inf. Model.* **2007**, *47* (6), 2462–2474. https://doi.org/10.1021/ci6005646.

(113) Tosco, P.; Stiefl, N.; Landrum, G. Bringing the MMFF Force Field to the RDKit: Implementation and Validation. *J Cheminform* **2014**, *6*, 37. https://doi.org/10.1186/s13321-014-0037-3.

(114) Riniker, S.; Landrum, G. A. Better Informed Distance Geometry: Using What We Know To Improve Conformation Generation. *J. Chem. Inf. Model.* **2015**, *55* (12), 2562–2574. https://doi.org/10.1021/acs.jcim.5b00654.

(115) Poli, G.; Seidel, T.; Langer, T. Conformational Sampling of Small Molecules With ICon: Performance Assessment in Comparison With OMEGA. *Front. Chem.* **2018**, *6*. https://doi.org/10.3389/fchem.2018.00229.

(116) 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.

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

(118) Leeson, P. Drug Discovery: Chemical Beauty Contest. *Nature* **2012**, *481* (7382), 455–456. https://doi.org/10.1038/481455a.

(119) Oprea, T. I.; Allu, T. K.; Fara, D. C.; Rad, R. F.; Ostopovici, L.; Bologa, C. G. Lead-like, Drug-like or "Pub-like": How Different Are They? *J Comput Aided Mol Des* **2007**, *21* (1–3), 113–119. https://doi.org/10.1007/s10822-007-9105-3.

(120) Lumley, J. A. Compound Selection and Filtering in Library Design. *QSAR & Combinatorial Science* **2005**, *24* (9), 1066–1075. https://doi.org/10.1002/qsar.200520136.

(121) Davis, A. M.; Keeling, D. J.; Steele, J.; Tinker, N. P. T. and A. C. Components of Successful Lead Generation http://www.eurekaselect.com/78951/article (accessed Mar 24, 2019).

(122) Verheij, H. J. Leadlikeness and Structural Diversity of Synthetic Screening Libraries. *Mol Divers* **2006**, *10* (3), 377–388. https://doi.org/10.1007/s11030-006-9040-6.

(123) Collins, I.; Workman, P. New Approaches to Molecular Cancer Therapeutics. *Nature Chemical Biology* **2006**, *2* (12), 689–700. https://doi.org/10.1038/nchembio840.

(124) Oprea, T. I.; Davis, A. M.; Teague, S. J.; Leeson, P. D. Is There a Difference between Leads and Drugs? A Historical Perspective. *J. Chem. Inf. Comput. Sci.* **2001**, *41* (5), 1308–1315. https://doi.org/10.1021/ci010366a.

(125) Hann, M. M.; Oprea, T. I. Pursuing the Leadlikeness Concept in Pharmaceutical Research. *Current Opinion in Chemical Biology* **2004**, *8* (3), 255–263. https://doi.org/10.1016/j.cbpa.2004.04.003.

(126) Mellor, C. L.; Steinmetz, F. P.; Cronin, M. T. D. Using Molecular Initiating Events to Develop a Structural Alert Based Screening Workflow for Nuclear Receptor Ligands Associated with Hepatic Steatosis. *Chem. Res. Toxicol.* **2016**, *29* (2), 203–212. https://doi.org/10.1021/acs.chemrestox.5b00480.

(127) Tcheremenskaia, O.; Benigni, R.; Nikolova, I.; Jeliazkova, N.; Escher, S. E.; Batke, M.; Baier, T.; Poroikov, V.; Lagunin, A.; Rautenberg, M.; et al. OpenTox Predictive Toxicology Framework: Toxicological Ontology and Semantic Media Wiki-Based OpenToxipedia. *J Biomed Semantics* **2012**, *3* (Suppl 1), S7. https://doi.org/10.1186/2041-1480-3-S1-S7.

(128) Toropov, A. A.; Toropova, A. P.; Raska, I.; Leszczynska, D.; Leszczynski, J. Comprehension of Drug Toxicity: Software and Databases. *Comput. Biol. Med.* **2014**, *45*, 20–25. https://doi.org/10.1016/j.compbiomed.2013.11.013.

(129) Zhang, L.; Zhang, H.; Ai, H.; Hu, H.; Li, S.; Zhao, J.; Liu, H. Applications of Machine Learning Methods in Drug Toxicity Prediction. *Curr Top Med Chem* **2018**, *18* (12), 987–997. https://doi.org/10.2174/1568026618666180727152557.

(130) Mishra, N. K.; Singla, D.; Agarwal, S.; Raghava, G. P. S. ToxiPred: A Server for Prediction of Aqueous Toxicity of Small Chemical Molecules in T. Pyriformis https://www.ingentaconnect.com/content/asp/jtt/2014/00000001/00000001/art00004 (accessed Mar 25, 2019). https://doi.org/info:doi/10.1166/jtt.2014.1005.

(131) Cheng, F.; Li, W.; Zhou, Y.; Shen, J.; Wu, Z.; Liu, G.; Lee, P. W.; Tang, Y. AdmetSAR: A Comprehensive Source and Free Tool for Assessment of Chemical ADMET Properties. *J Chem Inf Model* **2012**, *52* (11), 3099–3105. https://doi.org/10.1021/ci300367a.

(132) Sharma, A. K.; Srivastava, G. N.; Roy, A.; Sharma, V. K. ToxiM: A Toxicity Prediction Tool for Small Molecules Developed Using Machine Learning and Chemoinformatics Approaches. *Front Pharmacol* **2017**, *8*. https://doi.org/10.3389/fphar.2017.00880.

(133) Lagorce, D.; Bouslama, L.; Becot, J.; Miteva, M. A.; Villoutreix, B. O. FAF-Drugs4: Free ADME-Tox Filtering Computations for Chemical Biology and Early Stages Drug Discovery. *Bioinformatics* **2017**, *33* (22), 3658–3660. https://doi.org/10.1093/bioinformatics/btx491.

(134) Lahl, U.; Gundert-Remy, U. The Use of (Q)SAR Methods in the Context of REACH. *Toxicol. Mech. Methods* **2008**, *18* (2–3), 149–158. https://doi.org/10.1080/15376510701857288.

(135) Baell, J. B.; Nissink, J. W. M. Seven Year Itch: Pan-Assay Interference Compounds (PAINS) in 2017—Utility and Limitations. *ACS Chem Biol* **2018**, *13* (1), 36–44. https://doi.org/10.1021/acschembio.7b00903.

(136) Baell, J. B.; Holloway, G. A. New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays. *J. Med. Chem.* **2010**, *53* (7), 2719–2740. https://doi.org/10.1021/jm901137j.

(137) Huth, J. R.; Song, D.; Mendoza, R. R.; Black-Schaefer, C. L.; Mack, J. C.; Dorwin, S. A.; Ladror, U. S.; Severin, J. M.; Walter, K. A.; Bartley, D. M.; et al. Toxicological Evaluation of Thiol-Reactive Compounds Identified Using a La Assay to Detect Reactive Molecules by Nuclear Magnetic Resonance. *Chem. Res. Toxicol.* **2007**, *20* (12), 1752–1759. https://doi.org/10.1021/tx700319t.

(138) Metz, J. T.; Huth, J. R.; Hajduk, P. J. Enhancement of Chemical Rules for Predicting Compound Reactivity towards Protein Thiol Groups. *J. Comput. Aided Mol. Des.* **2007**, *21* (1–3), 139–144. https://doi.org/10.1007/s10822-007-9109-z.

(139) Huth, J. R.; Mendoza, R.; Olejniczak, E. T.; Johnson, R. W.; Cothron, D. A.; Liu, Y.; Lerner, C. G.; Chen, J.; Hajduk, P. J. ALARM NMR: A Rapid and Robust Experimental Method to Detect Reactive False Positives in Biochemical Screens. *J. Am. Chem. Soc.* **2005**,

127 (1), 217–224. https://doi.org/10.1021/ja0455547.

(140) McGovern, S. L.; Caselli, E.; Grigorieff, N.; Shoichet, B. K. A Common Mechanism Underlying Promiscuous Inhibitors from Virtual and High-Throughput Screening. *J. Med. Chem.* **2002**, *45* (8), 1712–1722.

(141) Lagorce, D.; Sperandio, O.; Baell, J. B.; Miteva, M. A.; Villoutreix, B. O. FAF-Drugs3: A Web Server for Compound Property Calculation and Chemical Library Design. *Nucleic Acids Research* **2015**, *43* (W1), W200–W207. https://doi.org/10.1093/nar/gkv353.

(142) Saubern, S.; Guha, R.; Baell, J. B. KNIME Workflow to Assess PAINS Filters in SMARTS Format. Comparison of RDKit and Indigo Cheminformatics Libraries. *Mol Inform* **2011**, *30* (10), 847–850. https://doi.org/10.1002/minf.201100076.

(143) Lagorce, D.; Maupetit, J.; Baell, J.; Sperandio, O.; Tufféry, P.; Miteva, M. A.; Galons, H.; Villoutreix, B. O. The FAF-Drugs2 Server: A Multistep Engine to Prepare Electronic Chemical Compound Collections. *Bioinformatics* **2011**, *27* (14), 2018–2020. https://doi.org/10.1093/bioinformatics/btr333.

(144) M Nissink, J. W.; Blackburn, S. Quantification of Frequent-Hitter Behavior Based on Historical High-Throughput Screening Data. *Future Medicinal Chemistry* **2014**, *6* (10), 1113–1126. https://doi.org/10.4155/fmc.14.72.

(145) Cereto-Massagué, A.; Ojeda, M. J.; Valls, C.; Mulero, M.; Garcia-Vallvé, S.; Pujadas, G. Molecular Fingerprint Similarity Search in Virtual Screening. *Methods* **2015**, *71*, 58–63. https://doi.org/10.1016/j.ymeth.2014.08.005.

(146) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *Journal of Chemical Information and Computer Sciences* **2002**, *42* (6), 1273–1280. https://doi.org/10.1021/ci010132r.

(147) Bolton, E. E.; Wang, Y.; Thiessen, P. A.; Bryant, S. H. Chapter 12 - PubChem: Integrated Platform of Small Molecules and Biological Activities. In *Annual Reports in Computational Chemistry*; Wheeler, R. A., Spellmeyer, D. C., Eds.; Elsevier, 2008; Vol. 4, pp 217–241. https://doi.org/10.1016/S1574-1400(08)00012-1.

(148) John M. Barnard*, † and; Downs‡, G. M. Chemical Fragment Generation and Clustering Software§ https://pubs.acs.org/doi/abs/10.1021/ci960090k (accessed Mar 27, 2019). https://doi.org/10.1021/ci960090k.

(149) Robert P. Sheridan, *; Michael D. Miller, §; Dennis J. Underwood, § and; Kearsley†, S. K. Chemical Similarity Using Geometric Atom Pair Descriptors† https://pubs.acs.org/doi/abs/10.1021/ci950275b (accessed Mar 27, 2019). https://doi.org/10.1021/ci950275b.

(150) Cereto-Massagué, A.; Guasch, L.; Valls, C.; Mulero, M.; Pujadas, G.; Garcia-Vallvé, S. DecoyFinder: An Easy-to-Use Python GUI Application for Building Target-Specific Decoy Sets. *Bioinformatics* **2012**, *28* (12), 1661–1662. https://doi.org/10.1093/bioinformatics/bts249.
(151) Daylight http://www.daylight.com/ (accessed Mar 27, 2019).
(152) Andreas Bender, *; Hamse Y. Mussa, and; Glen, R. C.; Reiling, S. Molecular Similarity Searching Using Atom Environments, Information-Based Feature Selection, and a Naïve Bayesian Classifier https://pubs.acs.org/doi/abs/10.1021/ci034207y (accessed Mar 27, 2019). https://doi.org/10.1021/ci034207y.
(153) Andreas Bender; Hamse Y. Mussa, and; Glen*, R. C.; Reiling, S. Similarity Searching of Chemical Databases Using Atom Environment Descriptors (MOLPRINT 2D): Evaluation of Performance https://pubs.acs.org/doi/abs/10.1021/ci0498719 (accessed Mar 27, 2019). https://doi.org/10.1021/ci0498719.
(154) Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. https://pubs.acs.org/doi/abs/10.1021/c160017a018 (accessed Mar 27, 2019). https://doi.org/10.1021/c160017a018.
(155) Riniker, S.; Landrum, G. A. Open-Source Platform to Benchmark Fingerprints for Ligand-Based Virtual Screening. *Journal of Cheminformatics* **2013**, *5* (1), 26. https://doi.org/10.1186/1758-2946-5-26.
(156) Venkatraman, V.; Pérez-Nueno, V. I.; Mavridis, L.; Ritchie, D. W. Comprehensive Comparison of Ligand-Based Virtual Screening Tools Against the DUD Data set Reveals Limitations of Current 3D Methods https://pubs.acs.org/doi/abs/10.1021/ci100263p (accessed Mar 27, 2019). https://doi.org/10.1021/ci100263p.
(157) Brown, N.; Jacoby, E. On Scaffolds and Hopping in Medicinal Chemistry. *Mini Rev Med Chem* **2006**, *6* (11), 1217–1229.
(158) Kumar, A.; Zhang, K. Y. J. Advances in the Development of Shape Similarity Methods and Their Application in Drug Discovery. *Front Chem* **2018**, *6*, 315. https://doi.org/10.3389/fchem.2018.00315.
(159) Nilakantan, R.; Bauman, N.; Venkataraghavan, R. New Method for Rapid Characterization of Molecular Shapes: Applications in Drug Design. *J Chem Inf Comput Sci* **1993**, *33* (1), 79–85.
(160) Ballester Pedro J; Richards W. Graham. Ultrafast Shape Recognition for Similarity Search in Molecular Databases. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* **2007**, *463* (2081), 1307–1321.

https://doi.org/10.1098/rspa.2007.1823.

(161) Ballester, P. J.; Westwood, I.; Laurieri, N.; Sim, E.; Richards, W. G. Prospective Virtual Screening with Ultrafast Shape Recognition: The Identification of Novel Inhibitors of Arylamine N-Acetyltransferases. *J R Soc Interface* **2010**, *7* (43), 335–342. https://doi.org/10.1098/rsif.2009.0170.

(162) Liu, Y.-S.; Wang, M.; Paul, J.-C.; Ramani, K. 3DMolNavi: A Web-Based Retrieval and Navigation Tool for Flexible Molecular Shape Comparison. *BMC Bioinformatics* **2012**, *13*, 95. https://doi.org/10.1186/1471-2105-13-95.

(163) Liu, Y.-S.; Fang, Y.; Ramani, K. IDSS: Deformation Invariant Signatures for Molecular Shape Comparison. *BMC Bioinformatics* **2009**, *10*, 157. https://doi.org/10.1186/1471-2105-10-157.

(164) Reutlinger, M.; Koch, C. P.; Reker, D.; Todoroff, N.; Schneider, P.; Rodrigues, T.; Schneider, G. Chemically Advanced Template Search (CATS) for Scaffold-Hopping and Prospective Target Prediction for 'Orphan' Molecules. *Mol Inform* **2013**, *32* (2), 133–138. https://doi.org/10.1002/minf.201200141.

(165) Stiefl, N.; Watson, I. A.; Baumann, K.; Zaliani, A. ErG: 2D Pharmacophore Descriptions for Scaffold Hopping /paper/ErG%3A-2D-Pharmacophore-Descriptions-for-Scaffold-Stiefl-Watson/88023c222503ae77df88d9c3d8ae365755973c42 (accessed Mar 31, 2019).

(166) Rarey, M.; Dixon, J. S. Feature Trees: A New Molecular Similarity Measure Based on Tree Matching. *J Comput Aided Mol Des* **1998**, *12* (5), 471–490. https://doi.org/10.1023/A:1008068904628.

(167) Leach, A. R.; Gillet, V. J. *An Introduction to Chemoinformatics*; Springer Netherlands, 2007.

(168) Stepniewska-Dziubinska, M. M.; Zielenkiewicz, P.; Siedlecki, P. DeCAF—Discrimination, Comparison, Alignment Tool for 2D PHarmacophores. *Molecules* **2017**, *22* (7). https://doi.org/10.3390/molecules22071128.

(169) Wermuth, C. G.; Ganellin, C. R.; Lindberg, P.; Mitscher, L. A. Glossary of Terms Used in Medicinal Chemistry (IUPAC Recommendations 1998). *Pure and Applied Chemistry* **1998**, *70* (5), 1129–1143. https://doi.org/10.1351/pac199870051129.

(170) Lagarde, N.; Delahaye, S.; Zagury, J.-F.; Montes, M. Discriminating Agonist and Antagonist Ligands of the Nuclear Receptors Using 3D-Pharmacophores. *J Cheminform* **2016**, *8* (1). https://doi.org/10.1186/s13321-016-0154-2.

(171) 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.

(172) Discovery Studio Predictive Science Application | Dassault Systèmes BIOVIA https://www.3dsbiovia.com/products/collaborative-science/biovia-discovery-studio/ (accessed Apr 2, 2019).

(173) 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.

(174) Wolber, G.; Seidel, T.; Bendix, F.; Langer, T. Molecule-Pharmacophore Superpositioning and Pattern Matching in Computational Drug Design. *Drug Discovery Today* **2008**, *13* (1), 23–29. https://doi.org/10.1016/j.drudis.2007.09.007.

(175) 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.

(176) Chemical Computing Group (CCG) | Computer-Aided Molecular Design https://www.chemcomp.com/ (accessed Apr 2, 2019).

(177) Phase | Schrödinger https://www.schrodinger.com/phase (accessed Apr 2, 2019).

(178) Dixon, S. L.; Smondyrev, A. M.; Rao, S. N. PHASE: A Novel Approach to Pharmacophore Modeling and 3D Database Searching. *Chemical Biology & Drug Design* **2006**, *67* (5), 370–372. https://doi.org/10.1111/j.1747-0285.2006.00384.x.

(179) Pharmer: Open-Source Software for Efficient and Exact Pharmacophore Search http://smoothdock.ccbb.pitt.edu/pharmer/ (accessed Apr 2, 2019).

(180) Schneidman-Duhovny, D.; Dror, O.; Inbar, Y.; Nussinov, R.; Wolfson, H. J. PharmaGist: A Webserver for Ligand-Based Pharmacophore Detection. *Nucleic Acids Res* **2008**, *36* (suppl_2), W223–W228. https://doi.org/10.1093/nar/gkn187.

(181) 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.

(182) Dror, O.; Shulman-Peleg, A.; Nussinov, R.; Wolfson, H. J. Predicting Molecular Interactions in Silico: I. A Guide to Pharmacophore Identification and Its Applications to Drug Design. *Curr. Med. Chem.* **2004**, *11* (1), 71–90.

(183) Chopra, M.; Gupta, R.; Gupta, S.; Saluja, D. Molecular Modeling Study on Chemically Diverse Series of Cyclooxygenase-2 Selective Inhibitors: Generation of Predictive

Pharmacophore Model Using Catalyst. *J Mol Model* **2008**, *14* (11), 1087–1099. https://doi.org/10.1007/s00894-008-0350-8.

(184) Leach, A. R.; Gillet, V. J.; Lewis, R. A.; Taylor, R. Three-Dimensional Pharmacophore Methods in Drug Discovery. *J. Med. Chem.* **2010**, *53* (2), 539–558. https://doi.org/10.1021/jm900817u.

(185) Tropsha, A. 4.07 - Predictive Quantitative Structure–Activity Relationship Modeling. In *Comprehensive Medicinal Chemistry II*; Taylor, J. B., Triggle, D. J., Eds.; Elsevier: Oxford, 2007; pp 149–165. https://doi.org/10.1016/B0-08-045044-X/00248-0.

(186) Andrade, C. H.; Pasqualoto, K. F. M.; Ferreira, E. I.; Hopfinger, A. J. 4D-QSAR: Perspectives in Drug Design. *Molecules* **2010**, *15* (5), 3281–3294. https://doi.org/10.3390/molecules15053281.

(187) Lill, M. A. Multi-Dimensional QSAR in Drug Discovery. *Drug Discov. Today* **2007**, *12* (23–24), 1013–1017. https://doi.org/10.1016/j.drudis.2007.08.004.

(188) Todeschini, R.; Consonni, V.; Mannhold, R.; Kubinyi, H.; Timmerman, H. *Handbook of Molecular Descriptors, Volume 11*; Wiley VCH: Weinheim ; New York, 2000.

(189) Cramer, R. D.; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110* (18), 5959–5967. https://doi.org/10.1021/ja00226a005.

(190) Klebe, G.; Abraham, U.; Mietzner, T. Molecular Similarity Indices in a Comparative Analysis (CoMSIA) of Drug Molecules to Correlate and Predict Their Biological Activity. *J. Med. Chem.* **1994**, *37* (24), 4130–4146. https://doi.org/10.1021/jm00050a010.

(191) Todeschini, R.; Lasagni, M.; Marengo, E. New Molecular Descriptors for 2D and 3D Structures. Theory. *Journal of Chemometrics* **1994**, *8* (4), 263–272. https://doi.org/10.1002/cem.1180080405.

(192) Vedani, A.; Briem, H.; Dobler, M.; Dollinger, H.; McMasters, D. R. Multiple-Conformation and Protonation-State Representation in 4D-QSAR: The Neurokinin-1 Receptor System. *J. Med. Chem.* **2000**, *43* (23), 4416–4427.

(193) Ekins, S.; Bravi, G.; Binkley, S.; Gillespie, J. S.; Ring, B. J.; Wikel, J. H.; Wrighton, S. A. Three- and Four-Dimensional-Quantitative Structure Activity Relationship (3D/4D-QSAR) Analyses of CYP2C9 Inhibitors. *Drug Metab. Dispos.* **2000**, *28* (8), 994–1002.

(194) Hopfinger, A. J.; Wang, S.; Tokarski, J. S.; Jin, B.; Albuquerque, M.; Madhav, P. J.; Duraiswami, C. Construction of 3D-QSAR Models Using the 4D-QSAR Analysis Formalism. *J. Am. Chem. Soc.* **1997**, *119* (43), 10509–10524. https://doi.org/10.1021/ja9718937.

(195) Pan, D.; Tseng, Y.; Hopfinger, A. J. Quantitative Structure-Based Design: Formalism

and Application of Receptor-Dependent RD-4D-QSAR Analysis to a Set of Glucose Analogue Inhibitors of Glycogen Phosphorylase. *J Chem Inf Comput Sci* **2003**, *43* (5), 1591–1607. https://doi.org/10.1021/ci0340714.

(196) Vedani, A.; Dobler, M. 5D-QSAR: The Key for Simulating Induced Fit? *J. Med. Chem.* **2002**, *45* (11), 2139–2149.

(197) Vedani, A.; Dobler, M.; Lill, M. A. Combining Protein Modeling and 6D-QSAR. Simulating the Binding of Structurally Diverse Ligands to the Estrogen Receptor. *J. Med. Chem.* **2005**, *48* (11), 3700–3703. https://doi.org/10.1021/jm050185q.

(198) Gini, G. QSAR Methods. *Methods Mol. Biol.* **2016**, *1425*, 1–20. https://doi.org/10.1007/978-1-4939-3609-0_1.

(199) Chemistry Development Kit https://cdk.github.io/ (accessed Apr 7, 2019).

(200) AFGen: Fragment-based Descriptors for Chemical Compounds | Karypis Lab http://glaros.dtc.umn.edu/gkhome/afgen/overview (accessed Apr 7, 2019).

(201) Computer aided molecular design, drug design, QSAR, Molecular Modeling Software http://www.vlifesciences.com/ (accessed Apr 7, 2019).

(202) Molecular descriptors calculation - Dragon - Talete srl http://www.talete.mi.it/products/dragon_description.htm (accessed Apr 7, 2019).

(203) Höltje, H.-D.; Sippl, W.; Rognan, D.; Folkers, G. *Molecular Modeling: Basic Principles and Applications*, 3 edition.; Wiley-VCH: Weinheim, 2008.

(204) Gramatica, P. On the Development and Validation of QSAR Models. *Methods Mol. Biol.* **2013**, *930*, 499–526. https://doi.org/10.1007/978-1-62703-059-5_21.

(205) Roy, K.; Mitra, I. On Various Metrics Used for Validation of Predictive QSAR Models with Applications in Virtual Screening and Focused Library Design. *Comb. Chem. High Throughput Screen.* **2011**, *14* (6), 450–474.

(206) Chan-Yao-Chong, M.; Durand, D.; Ha-Duong, T. Molecular Dynamics Simulations Combined with Nuclear Magnetic Resonance and/or Small-Angle X-Ray Scattering Data for Characterizing Intrinsically Disordered Protein Conformational Ensembles. *J Chem Inf Model* **2019**. https://doi.org/10.1021/acs.jcim.8b00928.

(207) Brown, D.; Superti-Furga, G. Rediscovering the Sweet Spot in Drug Discovery. *Drug Discov. Today* **2003**, *8* (23), 1067–1077.

(208) O'Brien, E. P.; Brooks, B. R.; Thirumalai, D. Effects of PH on Proteins: Predictions for Ensemble and Single Molecule Pulling Experiments. *J Am Chem Soc* **2012**, *134* (2), 979–987. https://doi.org/10.1021/ja206557y.

(209) Davies, M. J. The Oxidative Environment and Protein Damage. *Biochimica et*

Biophysica Acta (BBA) - Proteins and Proteomics **2005**, *1703* (2), 93–109. https://doi.org/10.1016/j.bbapap.2004.08.007.

(210) Huang, P.; Chandra, V.; Rastinejad, F. Structural Overview of the Nuclear Receptor Superfamily: Insights into Physiology and Therapeutics. *Annu Rev Physiol* **2010**, *72*, 247–272. https://doi.org/10.1146/annurev-physiol-021909-135917.

(211) 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**. https://doi.org/10.1002/minf.201700020.

(212) Yoshida, N. Role of Solvation in Drug Design as Revealed by the Statistical Mechanics Integral Equation Theory of Liquids. *J. Chem. Inf. Model.* **2017**, *57* (11), 2646–2656. https://doi.org/10.1021/acs.jcim.7b00389.

(213) 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.

(214) Warren, G. L.; Andrews, C. W.; Capelli, A.-M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; et al. A Critical Assessment of Docking Programs and Scoring Functions. *J. Med. Chem.* **2006**, *49* (20), 5912–5931. https://doi.org/10.1021/jm050362n.

(215) Trott, O.; Olson, A. J. AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading. *Journal of Computational Chemistry* **2010**, *31* (2), 455–461. https://doi.org/10.1002/jcc.21334.

(216) DOCK 6: Impact of New Features and Current Docking Performance - 5788482d08aedc252a937326.Pdf.

(217) Elokely, K. M.; Doerksen, R. J. Docking Challenge: Protein Sampling and Molecular Docking Performance. *Journal of Chemical Information and Modeling* **2013**, *53* (8), 1934–1945. https://doi.org/10.1021/ci400040d.

(218) Spitzer, R.; Jain, A. N. Surflex-Dock: Docking Benchmarks and Real-World Application. *J. Comput. Aided Mol. Des.* **2012**, *26* (6), 687–699. https://doi.org/10.1007/s10822-011-9533-y.

(219) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28* (1), 235–242.

(220) The Nobel Prize in Chemistry 2017 https://www.nobelprize.org/prizes/chemistry/2017/summary/ (accessed Apr 8, 2019).

(221) Richard Giegé, C. S. La Cristallogenèse Des Macromolécules Biologiques. *REGARD sur la BIOCHIMIE* **2001**, No. n°3, 21–31.

(222) X-ray Crystallography Platform - Creative Biostructure https://www.creative-biostructure.com/x-ray-crystallography-platform_60.htm (accessed Apr 10, 2019).

(223) Palatinus, L.; Brázda, P.; Boullay, P.; Perez, O.; Klementová, M.; Petit, S.; Eigner, V.; Zaarour, M.; Mintova, S. Hydrogen Positions in Single Nanocrystals Revealed by Electron Diffraction. *Science* **2017**, *355* (6321), 166–169. https://doi.org/10.1126/science.aak9652.

(224) Lee, W.; Tonelli, M.; Markley, J. L. NMRFAM-SPARKY: Enhanced Software for Biomolecular NMR Spectroscopy. *Bioinformatics* **2015**, *31* (8), 1325–1327. https://doi.org/10.1093/bioinformatics/btu830.

(225) mestrelab. Mnova NMR to Visualize, Process, Analyze & Report 1D and 2D NMR Data. *Mestrelab*.

(226) Software for NMR Data Analysis and NMR Spectra Data Procession https://www.bruker.com/products/mr/nmr/nmr-software/software/topspin/overview.html (accessed Apr 10, 2019).

(227) Clos, L. J.; Jofre, M. F.; Ellinger, J. J.; Westler, W. M.; Markley, J. L. NMRbot: Python Scripts Enable High-Throughput Data Collection on Current Bruker BioSpin NMR Spectrometers. *Metabolomics* **2013**, *9* (3), 558–563. https://doi.org/10.1007/s11306-012-0490-9.

(228) Milne, J. L. S.; Borgnia, M. J.; Bartesaghi, A.; Tran, E. E. H.; Earl, L. A.; Schauder, D. M.; Lengyel, J.; Pierson, J.; Patwardhan, A.; Subramaniam, S. Cryo-Electron Microscopy: A Primer for the Non-Microscopist. *FEBS J* **2013**, *280* (1), 28–45. https://doi.org/10.1111/febs.12078.

(229) Subramaniam, S.; Earl, L. A.; Falconieri, V.; Milne, J. L.; Egelman, E. H. Resolution Advances in Cryo-EM Enable Application to Drug Discovery. *Current Opinion in Structural Biology* **2016**, *41*, 194–202. https://doi.org/10.1016/j.sbi.2016.07.009.

(230) Breaking Cryo-EM Resolution Barriers to Facilitate Drug Discovery: Cell https://www.cell.com/cell/fulltext/S0092-8674(16)30591-8?_returnURL=https%3A%2F%2Flinkinghub.elsevier.com%2Fretrieve%2Fpii%2FS0092867416305918%3Fshowall%3Dtrue (accessed Apr 10, 2019).

(231) Anfinsen, C. B. Principles That Govern the Folding of Protein Chains. *Science* **1973**, *181* (4096), 223–230.

(232) 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.

(233) Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Eastwood, M. P.; Bank, J. A.; Jumper, J. M.; Salmon, J. K.; Shan, Y.; et al. Atomic-Level Characterization of the Structural Dynamics of Proteins. *Science* **2010**, *330* (6002), 341–346. https://doi.org/10.1126/science.1187409.

(234) Perez, A.; MacCallum, J. L.; Coutsias, E. A.; Dill, K. A. Constraint Methods That Accelerate Free-Energy Simulations of Biomolecules. *J. Chem. Phys.* **2015**, *143* (24), 243143. https://doi.org/10.1063/1.4936911.

(235) Altschul, S. F.; Madden, T. L.; Schäffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs. *Nucleic Acids Res.* **1997**, *25* (17), 3389–3402.

(236) 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* **2016**, *17* (11), 1214–1226. https://doi.org/10.1111/tra.12432.

(237) 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. *Scientific Reports* **2016**, *6*, 28268. https://doi.org/10.1038/srep28268.

(238) Camproux, A. C.; Gautier, R.; Tufféry, P. A Hidden Markov Model Derived Structural Alphabet for Proteins. *J. Mol. Biol.* **2004**, *339* (3), 591–605. https://doi.org/10.1016/j.jmb.2004.04.005.

(239) Baker, D.; Sali, A. Protein Structure Prediction and Structural Genomics. *Science* **2001**, *294* (5540), 93–96. https://doi.org/10.1126/science.1065659.

(240) Cheng, J. A Multi-Template Combination Algorithm for Protein Comparative Modeling. *BMC Struct. Biol.* **2008**, *8*, 18. https://doi.org/10.1186/1472-6807-8-18.

(241) Meier, A.; Söding, J. Automatic Prediction of Protein 3D Structures by Probabilistic Multi-Template Homology Modeling. *PLoS Comput. Biol.* **2015**, *11* (10), e1004343. https://doi.org/10.1371/journal.pcbi.1004343.

(242) Larsson, P.; Wallner, B.; Lindahl, E.; Elofsson, A. Using Multiple Templates to Improve Quality of Homology Models in Automated Homology Modeling. *Protein Sci.* **2008**, *17* (6), 990–1002. https://doi.org/10.1110/ps.073344908.

(243) Chakravarty, S.; Godbole, S.; Zhang, B.; Berger, S.; Sanchez, R. Systematic Analysis of the Effect of Multiple Templates on the Accuracy of Comparative Models of Protein Structure. *BMC Struct Biol* **2008**, *8*, 31. https://doi.org/10.1186/1472-6807-8-31.

(244) Wojcik, J.; Mornon, J. P.; Chomilier, J. New Efficient Statistical Sequence-Dependent Structure Prediction of Short to Medium-Sized Protein Loops Based on an Exhaustive Loop Classification. *J. Mol. Biol.* **1999**, *289* (5), 1469–1490. https://doi.org/10.1006/jmbi.1999.2826.

(245) Fiser, A.; Sali, A. ModLoop: Automated Modeling of Loops in Protein Structures. *Bioinformatics* **2003**, *19* (18), 2500–2501.

(246) Dunbrack, R. L. Rotamer Libraries in the 21st Century. *Curr. Opin. Struct. Biol.* **2002**, *12* (4), 431–440.

(247) About MODELLER https://salilab.org/modeller/ (accessed Apr 12, 2019).

(248) Robetta: full-chain protein structure prediction server http://robetta.bakerlab.org/ (accessed Apr 12, 2019).

(249) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res* **2000**, *28* (1), 235–242.

(250) Holm, L.; Sander, C. The FSSP Database of Structurally Aligned Protein Fold Families. *Nucleic Acids Res* **1994**, *22* (17), 3600–3609.

(251) Lo Conte, L.; Ailey, B.; Hubbard, T. J. P.; Brenner, S. E.; Murzin, A. G.; Chothia, C. SCOP: A Structural Classification of Proteins Database. *Nucleic Acids Res* **2000**, *28* (1), 257–259.

(252) Knudsen, M.; Wiuf, C. The CATH Database. *Hum Genomics* **2010**, *4* (3), 207–212. https://doi.org/10.1186/1479-7364-4-3-207.

(253) Morris, A. L.; MacArthur, M. W.; Hutchinson, E. G.; Thornton, J. M. Stereochemical Quality of Protein Structure Coordinates. *Proteins* **1992**, *12* (4), 345–364. https://doi.org/10.1002/prot.340120407.

(254) Wiederstein, M.; Sippl, M. J. ProSA-Web: Interactive Web Service for the Recognition of Errors in Three-Dimensional Structures of Proteins. *Nucleic Acids Res* **2007**, *35* (Web Server issue), W407–W410. https://doi.org/10.1093/nar/gkm290.

(255) Willard, L.; Ranjan, A.; Zhang, H.; Monzavi, H.; Boyko, R. F.; Sykes, B. D.; Wishart, D. S. VADAR: A Web Server for Quantitative Evaluation of Protein Structure Quality. *Nucleic Acids Res* **2003**, *31* (13), 3316–3319.

(256) Vriend, G. WHAT IF: A Molecular Modeling and Drug Design Program. *J Mol Graph* **1990**, *8* (1), 52–56, 29.

(257) Kabsch, W.; Sander, C. Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features. *Biopolymers* **1983**, *22* (12), 2577–2637. https://doi.org/10.1002/bip.360221211.

(258) Nayal, M.; Honig, B. On the Nature of Cavities on Protein Surfaces: Application to the Identification of Drug-Binding Sites. *Proteins* **2006**, *63* (4), 892–906. https://doi.org/10.1002/prot.20897.

(259) Brady, G. P.; Stouten, P. F. Fast Prediction and Visualization of Protein Binding Pockets with PASS. *J. Comput. Aided Mol. Des.* **2000**, *14* (4), 383–401.

(260) Laurie, A. T. R.; Jackson, R. M. Q-SiteFinder: An Energy-Based Method for the Prediction of Protein–Ligand Binding Sites. *Bioinformatics* **2005**, *21* (9), 1908–1916. https://doi.org/10.1093/bioinformatics/bti315.

(261) Emile Zuckerkandl, Linus Pauling, and the molecular evolutionary clock, 1959-1965. - Semantic Scholar /paper/Emile-Zuckerkandl%2C-Linus-Pauling%2C-and-the-molecular-Morgan/12d849c221466cd01b0ba288f60f5ff04e6fc927 (accessed May 29, 2018).

(262) 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.

(263) 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. *Israel Journal of Chemistry* **2013**, *53* (3–4), 199–206. https://doi.org/10.1002/ijch.201200096.

(264) Gao, M.; Skolnick, J. A Comprehensive Survey of Small-Molecule Binding Pockets in Proteins. *PLOS Computational Biology* **2013**, *9* (10), e1003302. https://doi.org/10.1371/journal.pcbi.1003302.

(265) Hendlich, M. Databases for Protein-Ligand Complexes. *Acta Crystallogr. D Biol. Crystallogr.* **1998**, *54* (Pt 6 Pt 1), 1178–1182.

(266) 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.

(267) 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.

(268) 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.

(269) Laskowski, R. A.; Watson, J. D.; Thornton, J. M. ProFunc: A Server for Predicting Protein Function from 3D Structure. *Nucleic Acids Res.* **2005**, *33* (Web Server issue), W89-93. https://doi.org/10.1093/nar/gki414.

(270) 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.

(271) Zhang, C.; Freddolino, P. L.; Zhang, Y. COFACTOR: Improved Protein Function Prediction by Combining Structure, Sequence and Protein–Protein Interaction Information. *Nucleic Acids Res* **2017**, *45* (Web Server issue), W291–W299. https://doi.org/10.1093/nar/gkx366.

(272) 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.

(273) Hendlich, M.; Rippmann, F.; Barnickel, G. LIGSITE: Automatic and Efficient Detection of Potential Small Molecule-Binding Sites in Proteins. *J. Mol. Graph. Model.* **1997**, *15* (6), 359–363, 389.

(274) Desaphy, J.; Azdimousa, K.; Kellenberger, E.; Rognan, D. Comparison and Druggability Prediction of Protein-Ligand Binding Sites from Pharmacophore-Annotated Cavity Shapes. *J Chem Inf Model* **2012**, *52* (8), 2287–2299. https://doi.org/10.1021/ci300184x.

(275) Laurie, A. T. R.; Jackson, R. M. Methods for the Prediction of Proteinligand Binding Sites for Structure-Based Drug Design and Virtual Ligand Screening. *Curr. Protein Pept. Sci* **2006**, 395–406.

(276) Laskowski, R. A. SURFNET: A Program for Visualizing Molecular Surfaces, Cavities, and Intermolecular Interactions. *J Mol Graph* **1995**, *13* (5), 323–330, 307–308.

(277) Liang, J.; Woodward, C.; Edelsbrunner, H. Anatomy of protein pockets and cavities: Measurement of binding site geometry and implications for ligand design. *Protein Science 7* (9), 1884–1897. https://doi.org/10.1002/pro.5560070905.

(278) Peters, K. P.; Fauck, J.; Frömmel, C. The Automatic Search for Ligand Binding Sites in Proteins of Known Three-Dimensional Structure Using Only Geometric Criteria. *J. Mol. Biol.* **1996**, *256* (1), 201–213. https://doi.org/10.1006/jmbi.1996.0077.

(279) Le Guilloux, V.; Schmidtke, P.; Tuffery, P. Fpocket: An Open Source Platform for Ligand Pocket Detection. *BMC Bioinformatics* **2009**, *10*, 168. https://doi.org/10.1186/1471-2105-10-168.

(280) Ravindranath, P. A.; Sanner, M. F. AutoSite: An Automated Approach for Pseudo-Ligands Prediction—from Ligand-Binding Sites Identification to Predicting Key Ligand Atoms. *Bioinformatics* **2016**, *32* (20), 3142–3149. https://doi.org/10.1093/bioinformatics/btw367.

(281) Ngan, C. H.; Bohnuud, T.; Mottarella, S. E.; Beglov, D.; Villar, E. A.; Hall, D. R.; Kozakov, D.; Vajda, S. FTMAP: Extended Protein Mapping with User-Selected Probe

Molecules. *Nucleic Acids Res* **2012**, *40* (Web Server issue), W271–W275. https://doi.org/10.1093/nar/gks441.

(282) Ngan, C.-H.; Hall, D. R.; Zerbe, B.; Grove, L. E.; Kozakov, D.; Vajda, S. FTSite: High Accuracy Detection of Ligand Binding Sites on Unbound Protein Structures. *Bioinformatics* **2012**, *28* (2), 286–287. https://doi.org/10.1093/bioinformatics/btr651.

(283) Jain, A. N. Scoring Noncovalent Protein-Ligand Interactions: A Continuous Differentiable Function Tuned to Compute Binding Affinities. *J. Comput. Aided Mol. Des.* **1996**, *10* (5), 427–440.

(284) Hopkins, A. L.; Groom, C. R. The Druggable Genome. *Nature Reviews Drug Discovery* **2002**, *1* (9), 727–730. https://doi.org/10.1038/nrd892.

(285) Barril, X. Druggability Predictions: Methods, Limitations, and Applications. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2013**, *3* (4), 327–338. https://doi.org/10.1002/wcms.1134.

(286) Rask-Andersen, M.; Almén, M. S.; Schiöth, H. B. Trends in the Exploitation of Novel Drug Targets. *Nat Rev Drug Discov* **2011**, *10* (8), 579–590. https://doi.org/10.1038/nrd3478.

(287) Gupta, A.; Gupta, A. K.; Seshadri, K. Structural Models in the Assessment of Protein Druggability Based on HTS Data. *J. Comput. Aided Mol. Des.* **2009**, *23* (8), 583–592. https://doi.org/10.1007/s10822-009-9279-y.

(288) Henrich, S.; Salo-Ahen, O. M. H.; Huang, B.; Rippmann, F. F.; Cruciani, G.; Wade, R. C. Computational Approaches to Identifying and Characterizing Protein Binding Sites for Ligand Design. *J. Mol. Recognit.* **2010**, *23* (2), 209–219. https://doi.org/10.1002/jmr.984.

(289) Krasowski, A.; Muthas, D.; Sarkar, A.; Schmitt, S.; Brenk, R. DrugPred: A Structure-Based Approach to Predict Protein Druggability Developed Using an Extensive Nonredundant Data Set. *J Chem Inf Model* **2011**, *51* (11), 2829–2842. https://doi.org/10.1021/ci200266d.

(290) Desaphy, J.; Azdimousa, K.; Kellenberger, E.; Rognan, D. Comparison and Druggability Prediction of Protein–Ligand Binding Sites from Pharmacophore-Annotated Cavity Shapes. *Journal of Chemical Information and Modeling* **2012**, *52* (8), 2287–2299. https://doi.org/10.1021/ci300184x.

(291) Borrel, A.; Regad, L.; Xhaard, H.; Petitjean, M.; Camproux, A.-C. PockDrug: A Model for Predicting Pocket Druggability That Overcomes Pocket Estimation Uncertainties. *J. Chem. Inf. Model.* **2015**, *55* (4), 882–895. https://doi.org/10.1021/ci5006004.

(292) Hussein, H. A.; Borrel, A.; Geneix, C.; Petitjean, M.; Regad, L.; Camproux, A.-C. PockDrug-Server: A New Web Server for Predicting Pocket Druggability on Holo and Apo Proteins. *Nucleic Acids Research* **2015**, *43* (W1), W436–W442.

https://doi.org/10.1093/nar/gkv462.

(293) Schmidtke, P.; Souaille, C.; Estienne, F.; Baurin, N.; Kroemer, R. T. Large-Scale Comparison of Four Binding Site Detection Algorithms. *J Chem Inf Model* **2010**, *50* (12), 2191–2200. https://doi.org/10.1021/ci1000289.

(294) Volkamer, A.; Griewel, A.; Grombacher, T.; Rarey, M. Analyzing the Topology of Active Sites: On the Prediction of Pockets and Subpockets. *J Chem Inf Model* **2010**, *50* (11), 2041–2052. https://doi.org/10.1021/ci100241y.

(295) Le Guilloux, V.; Schmidtke, P.; Tuffery, P. Fpocket: An Open Source Platform for Ligand Pocket Detection. *BMC Bioinformatics* **2009**, *10*, 168. https://doi.org/10.1186/1471-2105-10-168.

(296) Webb, B.; Sali, A. Comparative Protein Structure Modeling Using MODELLER. *Current Protocols in Protein Science* **2016**, *86*, 2.9.1–2.9.37. https://doi.org/10.1002/cpps.20.

(297) About MODELLER https://salilab.org/modeller/ (accessed Apr 16, 2019).

(298) CASE, D. A.; CHEATHAM, T. E.; DARDEN, T.; GOHLKE, H.; LUO, R.; MERZ, K. M.; ONUFRIEV, A.; SIMMERLING, C.; WANG, B.; WOODS, R. J. The Amber Biomolecular Simulation Programs. *J Comput Chem* **2005**, *26* (16), 1668–1688. https://doi.org/10.1002/jcc.20290.

(299) Bombarda, E.; Ullmann, G. M. PH-Dependent PKa Values in Proteins--a Theoretical Analysis of Protonation Energies with Practical Consequences for Enzymatic Reactions. *J Phys Chem B* **2010**, *114* (5), 1994–2003. https://doi.org/10.1021/jp908926w.

(300) 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.

(301) Alexov, E. G.; Gunner, M. R. Calculated Protein and Proton Motions Coupled to Electron Transfer: Electron Transfer from QA- to QB in Bacterial Photosynthetic Reaction Centers. *Biochemistry* **1999**, *38* (26), 8253–8270. https://doi.org/10.1021/bi982700a.

(302) Madura, J. D.; Briggs, J. M.; Wade, R. C.; Davis, M. E.; Luty, B. A.; Ilin, A.; Antosiewicz, J.; Gilson, M. K.; Bagheri, B.; Scott, L. R.; et al. Electrostatics and Diffusion of Molecules in Solution: Simulations with the University of Houston Brownian Dynamics Program. *Computer Physics Communications* **1995**, *91* (1), 57–95. https://doi.org/10.1016/0010-4655(95)00043-F.

(303) Li, H.; Robertson, A. D.; Jensen, J. H. Very Fast Empirical Prediction and Rationalization of Protein PKa Values. *Proteins* **2005**, *61* (4), 704–721. https://doi.org/10.1002/prot.20660.

(304) Davies, M. N.; Toseland, C. P.; Moss, D. S.; Flower, D. R. Benchmarking PKa Prediction. *BMC Biochem* **2006**, *7*, 18. https://doi.org/10.1186/1471-2091-7-18.

(305) Wolber, G.; Langer, T. LigandScout: 3-D Pharmacophores Derived from Protein-Bound Ligands and Their Use as Virtual Screening Filters. *Journal of Chemical Information and Modeling* **2005**, *45* (1), 160–169. https://doi.org/10.1021/ci049885e.

(306) Mortier, J.; Dhakal, P.; Volkamer, A. Truly Target-Focused Pharmacophore Modeling: A Novel Tool for Mapping Intermolecular Surfaces. *Molecules* **2018**, *23* (8). https://doi.org/10.3390/molecules23081959.

(307) 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.

(308) 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.

(309) Meagher, K. L.; Carlson, H. A. Incorporating Protein Flexibility in Structure-Based Drug Discovery: Using HIV-1 Protease as a Test Case. *J. Am. Chem. Soc.* **2004**, *126* (41), 13276–13281. https://doi.org/10.1021/ja0469378.

(310) 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.

(311) 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.

(312) Siragusa, L.; Cross, S.; Baroni, M.; Goracci, L.; Cruciani, G. BioGPS: Navigating Biological Space to Predict Polypharmacology, off-Targeting, and Selectivity. *Proteins* **2015**, *83* (3), 517–532. https://doi.org/10.1002/prot.24753.

(313) Baroni, M.; Cruciani, G.; Sciabola, S.; Perruccio, F.; Mason, J. S. A Common Reference Framework for Analyzing/Comparing Proteins and Ligands. Fingerprints for Ligands and Proteins (FLAP): Theory and Application. *J Chem Inf Model* **2007**, *47* (2), 279–294. https://doi.org/10.1021/ci600253e.

(314) Tintori, C.; Corradi, V.; Magnani, M.; Manetti, F.; Botta, M. Targets Looking for Drugs: A Multistep Computational Protocol for the Development of Structure-Based Pharmacophores and Their Applications for Hit Discovery. *J Chem Inf Model* **2008**, *48* (11), 2166–2179. https://doi.org/10.1021/ci800105p.

(315) 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.

(316) Hu, B.; Lill, M. A. PharmDock: A Pharmacophore-Based Docking Program. *Journal of Cheminformatics* **2014**, *6* (1), 14. https://doi.org/10.1186/1758-2946-6-14.

(317) 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.

(318) Fischer, E. Einfluss Der Configuration Auf Die Wirkung Der Enzyme. *Berichte der deutschen chemischen Gesellschaft* **1894**, *27* (3), 2985–2993. https://doi.org/10.1002/cber.18940270364.

(319) Koshland, D. E. Application of a Theory of Enzyme Specificity to Protein Synthesis. *Proc. Natl. Acad. Sci. U.S.A.* **1958**, *44* (2), 98–104.

(320) Schapira, M.; Abagyan, R.; Totrov, M. Nuclear Hormone Receptor Targeted Virtual Screening. *J. Med. Chem.* **2003**, *46* (14), 3045–3059. https://doi.org/10.1021/jm0300173.

(321) Pellegrini, M.; Doniach, S. Computer Simulation of Antibody Binding Specificity. *Proteins: Structure, Function, and Bioinformatics* **1993**, *15* (4), 436–444. https://doi.org/10.1002/prot.340150410.

(322) Ring, C. S.; Sun, E.; McKerrow, J. H.; Lee, G. K.; Rosenthal, P. J.; Kuntz, I. D.; Cohen, F. E. Structure-Based Inhibitor Design by Using Protein Models for the Development of Antiparasitic Agents. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90* (8), 3583–3587.

(323) 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.

(324) 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 Computational Biology* **2015**, *11* (12), e1004586. https://doi.org/10.1371/journal.pcbi.1004586.

(325) 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.

(326) Allen, W. J.; Balius, T. E.; Mukherjee, S.; Brozell, S. R.; Moustakas, D. T.; Lang, P. T.; Case, D. A.; Kuntz, I. D.; Rizzo, R. C. DOCK 6: Impact of New Features and Current Docking Performance. *J Comput Chem* **2015**, *36* (15), 1132–1156. https://doi.org/10.1002/jcc.23905.

(327) Caflisch, A.; Miranker, A.; Karplus, M. Multiple Copy Simultaneous Search and Construction of Ligands in Binding Sites: Application to Inhibitors of HIV-1 Aspartic

Proteinase. *J. Med. Chem.* **1993**, *36* (15), 2142–2167.

(328) Böhm, H. J. The Computer Program LUDI: A New Method for the de Novo Design of Enzyme Inhibitors. *J. Comput. Aided Mol. Des.* **1992**, *6* (1), 61–78.

(329) Paquet, E.; Viktor, H. L. Molecular Dynamics, Monte Carlo Simulations, and Langevin Dynamics: A Computational Review https://www.hindawi.com/journals/bmri/2015/183918/ (accessed Apr 30, 2019). https://doi.org/10.1155/2015/183918.

(330) Nichols, S. E.; Baron, R.; Ivetac, A.; McCammon, J. A. Predictive Power of Molecular Dynamics Receptor Structures in Virtual Screening. *J Chem Inf Model* **2011**, *51* (6), 1439–1446. https://doi.org/10.1021/ci200117n.

(331) Chodera, J. D.; Mobley, D. L. Entropy-Enthalpy Compensation: Role and Ramifications in Biomolecular Ligand Recognition and Design. *Annu Rev Biophys* **2013**, *42*, 121–142. https://doi.org/10.1146/annurev-biophys-083012-130318.

(332) Ruvinsky, A. M. Role of Binding Entropy in the Refinement of Protein-Ligand Docking Predictions: Analysis Based on the Use of 11 Scoring Functions. *J Comput Chem* **2007**, *28* (8), 1364–1372. https://doi.org/10.1002/jcc.20580.

(333) Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D. Improved Protein-Ligand Docking Using GOLD. *Proteins* **2003**, *52* (4), 609–623. https://doi.org/10.1002/prot.10465.

(334) 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.

(335) Meng, E. C.; Shoichet, B. K.; Kuntz, I. D. Automated Docking with Grid-Based Energy Evaluation. *J. Comput. Chem.* **1992**, *13* (4), 505–524. https://doi.org/10.1002/jcc.540130412.

(336) 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.

(337) Muegge, I.; Martin, Y. C. A General and Fast Scoring Function for Protein-Ligand Interactions: A Simplified Potential Approach. *J. Med. Chem.* **1999**, *42* (5), 791–804. https://doi.org/10.1021/jm980536j.

(338) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; et al. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* **2004**, *47* (7), 1739–1749. https://doi.org/10.1021/jm0306430.

(339) 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.

(340) Zheng, Z.; Merz, K. M. Development of the Knowledge-Based and Empirical Combined Scoring Algorithm (KECSA) To Score Protein–Ligand Interactions. *J. Chem. Inf. Model.* **2013**, *53* (5), 1073–1083. https://doi.org/10.1021/ci300619x.

(341) Ortiz, A. R.; Pisabarro, M. T.; Gago, F.; Wade, R. C. Prediction of Drug Binding Affinities by Comparative Binding Energy Analysis. *J. Med. Chem.* **1995**, *38* (14), 2681–2691. https://doi.org/10.1021/jm00014a020.

(342) Yin, S.; Biedermannova, L.; Vondrasek, J.; Dokholyan, N. V. MedusaScore: An Accurate Force Field-Based Scoring Function for Virtual Drug Screening https://pubs.acs.org/doi/full/10.1021/ci8001167 (accessed Apr 24, 2019). https://doi.org/10.1021/ci8001167.

(343) Liu, J.; Wang, R. Classification of Current Scoring Functions. *J Chem Inf Model* **2015**, *55* (3), 475–482. https://doi.org/10.1021/ci500731a.

(344) Benson, M. L.; Smith, R. D.; Khazanov, N. A.; Dimcheff, B.; Beaver, J.; Dresslar, P.; Nerothin, J.; Carlson, H. A. Binding MOAD, a High-Quality Protein Ligand Database. *Nucleic Acids Research* **2007**, *36* (Database), D674–D678. https://doi.org/10.1093/nar/gkm911.

(345) Wang, R.; Fang, X.; Lu, Y.; Wang, S. The PDBbind Database: Collection of Binding Affinities for Protein-Ligand Complexes with Known Three-Dimensional Structures. *J. Med. Chem.* **2004**, *47* (12), 2977–2980. https://doi.org/10.1021/jm030580l.

(346) Wang, R.; Fang, X.; Lu, Y.; Yang, C.-Y.; Wang, S. The PDBbind Database: Methodologies and Updates. *J. Med. Chem.* **2005**, *48* (12), 4111–4119. https://doi.org/10.1021/jm048957q.

(347) 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.

(348) McCammon, J. A.; Gelin, B. R.; Karplus, M. Dynamics of Folded Proteins. *Nature* **1977**, *267* (5612), 585–590.

(349) Ferreira, L.; Santos, R.; Oliva, G.; Andricopulo, A. Molecular Docking and Structure-Based Drug Design Strategies. *Molecules* **2015**, *20*, 13384–13421. https://doi.org/10.3390/molecules200713384.

(350) González, M. A. Force Fields and Molecular Dynamics Simulations. *École thématique de la Société Française de la Neutronique* **2011**, *12*, 169–200. https://doi.org/10.1051/sfn/201112009.

(351) Decherchi, S.; Masetti, M.; Vyalov, I.; Rocchia, W. Implicit Solvent Methods for Free Energy Estimation. *Eur J Med Chem* **2015**, *0*, 27–42. https://doi.org/10.1016/j.ejmech.2014.08.064.

(352) Gutiérrez-de-Terán, H.; Aqvist, J. Linear Interaction Energy: Method and Applications in Drug Design. *Methods Mol. Biol.* **2012**, *819*, 305–323. https://doi.org/10.1007/978-1-61779-465-0_20.

(353) Kuhn, B.; Gerber, P.; Schulz-Gasch, T.; Stahl, M. Validation and Use of the MM-PBSA Approach for Drug Discovery. *J. Med. Chem.* **2005**, *48* (12), 4040–4048. https://doi.org/10.1021/jm049081q.

(354) Réau, M.; Langenfeld, F.; Zagury, J.-F.; Montes, M. Predicting the Affinity of Farnesoid X Receptor Ligands through a Hierarchical Ranking Protocol: A D3R Grand Challenge 2 Case Study. *J Comput Aided Mol Des* **2018**, *32* (1), 231–238. https://doi.org/10.1007/s10822-017-0063-0.

(355) Demachy, I.; Piquemal, J.-P. La Surface d'énergie Potentielle Vue Par Les Champs de Forces. *L'Actualité Chimique* **2014**, *388–389*, 37–42.

(356) 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.

(357) Li, D.-D.; Meng, X.-F.; Wang, Q.; Yu, P.; Zhao, L.-G.; Zhang, Z.-P.; Wang, Z.-Z.; Xiao, W. Consensus Scoring Model for the Molecular Docking Study of MTOR Kinase Inhibitor. *Journal of Molecular Graphics and Modelling* **2018**, *79*, 81–87. https://doi.org/10.1016/j.jmgm.2017.11.003.

(358) 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.

(359) 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.

(360) Leach, A. R.; Kuntz, I. D. Conformational analysis of flexible ligands in macromolecular receptor sites. *Journal of Computational Chemistry* **1992**, *13* (6), 730–748. https://doi.org/10.1002/jcc.540130608.

(361) 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.

(362) Teague, S. J. Implications of Protein Flexibility for Drug Discovery. *Nat Rev Drug Discov* **2003**, *2* (7), 527–541. https://doi.org/10.1038/nrd1129.

(363) Morris, G. M.; Huey, R.; Lindstrom, W.; Sanner, M. F.; Belew, R. K.; Goodsell, D. S.; Olson, A. J. AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility. *J Comput Chem* **2009**, *30* (16), 2785–2791. https://doi.org/10.1002/jcc.21256.

(364) 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.

(365) Jiang, F.; Kim, S. H. "Soft Docking": Matching of Molecular Surface Cubes. *J. Mol. Biol.* **1991**, *219* (1), 79–102.

(366) *Comprehensive Medicinal Chemistry III*; Elsevier, 2017.

(367) Totrov, M.; Abagyan, R. Flexible Protein-Ligand Docking by Global Energy Optimization in Internal Coordinates. *Proteins* **1997**, *Suppl 1*, 215–220.

(368) Meiler, J.; Baker, D. ROSETTALIGAND: Protein-Small Molecule Docking with Full Side-Chain Flexibility. *Proteins* **2006**, *65* (3), 538–548. https://doi.org/10.1002/prot.21086.

(369) Nabuurs, S. B.; Wagener, M.; de Vlieg, J. A Flexible Approach to Induced Fit Docking. *J. Med. Chem.* **2007**, *50* (26), 6507–6518. https://doi.org/10.1021/jm070593p.

(370) Corbeil, C. R.; Englebienne, P.; Yannopoulos, C. G.; Chan, L.; Das, S. K.; Bilimoria, D.; L'heureux, L.; Moitessier, N. Docking Ligands into Flexible and Solvated Macromolecules. 2. Development and Application of Fitted 1.5 to the Virtual Screening of Potential HCV Polymerase Inhibitors. *J Chem Inf Model* **2008**, *48* (4), 902–909. https://doi.org/10.1021/ci700398h.

(371) Kokh, D. B.; Wade, R. C.; Wenzel, W. Receptor Flexibility in Small-Molecule Docking Calculations. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2011**, *1* (2), 298–314. https://doi.org/10.1002/wcms.29.

(372) Neves, M. A. C.; Totrov, M.; Abagyan, R. Docking and Scoring with ICM: The Benchmarking Results and Strategies for Improvement. *J. Comput. Aided Mol. Des.* **2012**, *26* (6), 675–686. https://doi.org/10.1007/s10822-012-9547-0.

(373) Bourguet, W.; Germain, P.; Gronemeyer, H. Nuclear Receptor Ligand-Binding Domains: Three-Dimensional Structures, Molecular Interactions and Pharmacological Implications. *Trends Pharmacol. Sci.* **2000**, *21* (10), 381–388.

(374) Lagarde, N.; Zagury, J.-F.; Montes, M. Importance of the Pharmacological Profile of the Bound Ligand in Enrichment on Nuclear Receptors: Toward the Use of Experimentally Validated Decoy Ligands. *J Chem Inf Model* **2014**, *54* (10), 2915–2944. https://doi.org/10.1021/ci500305c.

(375) 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.

(376) Klebe, G. Virtual Ligand Screening: Strategies, Perspectives and Limitations. *Drug Discov. Today* **2006**, *11* (13–14), 580–594. https://doi.org/10.1016/j.drudis.2006.05.012.

(377) Michel, J.; Tirado-Rives, J.; Jorgensen, W. L. Energetics of Displacing Water Molecules from Protein Binding Sites: Consequences for Ligand Optimization. *J. Am. Chem. Soc.* **2009**, *131* (42), 15403–15411. https://doi.org/10.1021/ja906058w.

(378) Amadasi, A.; Surface, J. A.; Spyrakis, F.; Cozzini, P.; Mozzarelli, A.; Kellogg, G. E. Robust Classification of "Relevant" Water Molecules in Putative Protein Binding Sites. *J. Med. Chem.* **2008**, *51* (4), 1063–1067. https://doi.org/10.1021/jm701023h.

(379) Ruiz-Carmona, S.; Alvarez-Garcia, D.; Foloppe, N.; Garmendia-Doval, A. B.; Juhos, S.; Schmidtke, P.; Barril, X.; Hubbard, R. E.; Morley, S. D. RDock: A Fast, Versatile and Open Source Program for Docking Ligands to Proteins and Nucleic Acids. *PLoS Comput. Biol.* **2014**, *10* (4), e1003571. https://doi.org/10.1371/journal.pcbi.1003571.

(380) Schnecke, V.; Kuhn, L. A. Virtual Screening with Solvation and Ligand-Induced Complementarity. *Perspectives in Drug Discovery and Design* **2000**, *20* (1), 171–190. https://doi.org/10.1023/A:1008737207775.

(381) Lagarde, N.; Zagury, J.-F.; Montes, M. Benchmarking Data Sets for the Evaluation of Virtual Ligand Screening Methods: Review and Perspectives. *J. Chem. Inf. Model.* **2015**, *55* (7), 1297–1307. https://doi.org/10.1021/acs.jcim.5b00090.

(382) 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.

(383) Truchon, J.-F.; Bayly, C. I. Evaluating Virtual Screening Methods: Good and Bad Metrics for the "Early Recognition" Problem. *J Chem Inf Model* **2007**, *47* (2), 488–508. https://doi.org/10.1021/ci600426e.

(384) Irwin, J. J. Community Benchmarks for Virtual Screening. *J Comput Aided Mol Des* **2008**, *22* (3), 193–199. https://doi.org/10.1007/s10822-008-9189-4.

(385) Bissantz, C.; Folkers, G.; Rognan, D. Protein-Based Virtual Screening of Chemical Databases. 1. Evaluation of Different Docking/Scoring Combinations. *J. Med. Chem.* **2000**, *43* (25), 4759–4767.

(386) Diller, D. J.; Li, R. Kinases, Homology Models, and High Throughput Docking. *J. Med. Chem.* **2003**, *46* (22), 4638–4647. https://doi.org/10.1021/jm020503a.

(387) Good, A. C.; Oprea, T. I. Optimization of CAMD Techniques 3. Virtual Screening Enrichment Studies: A Help or Hindrance in Tool Selection? *J Comput Aided Mol Des* **2008**, *22* (3–4), 169–178. https://doi.org/10.1007/s10822-007-9167-2.

(388) Stumpfe, D.; Bajorath, J. Applied Virtual Screening: Strategies, Recommendations, and Caveats. In *Virtual Screening*; Sotriffer, C., Ed.; Wiley-VCH Verlag GmbH & Co. KGaA, 2011; pp 291–318. https://doi.org/10.1002/9783527633326.ch11.

(389) Good, A. C.; Oprea, T. I. Optimization of CAMD Techniques 3. Virtual Screening Enrichment Studies: A Help or Hindrance in Tool Selection? *J. Comput. Aided Mol. Des.* **2008**, *22* (3–4), 169–178. https://doi.org/10.1007/s10822-007-9167-2.

(390) Rohrer, S. G.; Baumann, K. Maximum Unbiased Validation (MUV) Data Sets for Virtual Screening Based on PubChem Bioactivity Data. *J Chem Inf Model* **2009**, *49* (2), 169–184. https://doi.org/10.1021/ci8002649.

(391) Wishart, D. S.; Knox, C.; Guo, A. C.; Cheng, D.; Shrivastava, S.; Tzur, D.; Gautam, B.; Hassanali, M. DrugBank: A Knowledgebase for Drugs, Drug Actions and Drug Targets. *Nucleic Acids Res.* **2008**, *36* (Database issue), D901-906. https://doi.org/10.1093/nar/gkm958.

(392) www.mdli.com. *MDDR Licensed by Molecular Design, Ltd., San Leandro, CA.*

(393) Vogel, S. M.; Bauer, M. R.; Boeckler, F. M. DEKOIS: Demanding Evaluation Kits for Objective in Silico Screening--a Versatile Tool for Benchmarking Docking Programs and Scoring Functions. *J Chem Inf Model* **2011**, *51* (10), 2650–2665. https://doi.org/10.1021/ci2001549.

(394) Mysinger, M. M.; Shoichet, B. K. Rapid Context-Dependent Ligand Desolvation in Molecular Docking. *J Chem Inf Model* **2010**, *50* (9), 1561–1573. https://doi.org/10.1021/ci100214a.

(395) Bauer, M. R.; Ibrahim, T. M.; Vogel, S. M.; Boeckler, F. M. Evaluation and Optimization of Virtual Screening Workflows with DEKOIS 2.0--a Public Library of Challenging Docking Benchmark Sets. *J Chem Inf Model* **2013**, *53* (6), 1447–1462. https://doi.org/10.1021/ci400115b.

(396) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking Sets for Molecular Docking. *J. Med. Chem.* **2006**, *49* (23), 6789–6801. https://doi.org/10.1021/jm0608356.

(397) Chaput, L.; Martinez-Sanz, J.; Saettel, N.; Mouawad, L. Benchmark of Four Popular Virtual Screening Programs: Construction of the Active/Decoy Dataset Remains a Major Determinant of Measured Performance. *Journal of Cheminformatics* **2016**, *8* (1), 56. https://doi.org/10.1186/s13321-016-0167-x.

(398) Lagarde, N.; Zagury, J.-F.; Montes, M. Importance of the Pharmacological Profile of

the Bound Ligand in Enrichment on Nuclear Receptors: Toward the Use of Experimentally Validated Decoy Ligands. *J Chem Inf Model* **2014**, *54* (10), 2915–2944. https://doi.org/10.1021/ci500305c.

(399) McGovern, S. L.; Shoichet, B. K. Information Decay in Molecular Docking Screens against Holo, Apo, and Modeled Conformations of Enzymes. *J. Med. Chem.* **2003**, *46* (14), 2895–2907. https://doi.org/10.1021/jm0300330.

(400) Jain, A. N.; Nicholls, A. Recommendations for Evaluation of Computational Methods. *J Comput Aided Mol Des* **2008**, *22* (3–4), 133–139. https://doi.org/10.1007/s10822-008-9196-5.

(401) Meyer, K. WOMBAT—A Tool for Mixed Model Analyses in Quantitative Genetics by Restricted Maximum Likelihood (REML). *J Zhejiang Univ Sci B* **2007**, *8* (11), 815–821. https://doi.org/10.1631/jzus.2007.B0815.

(402) Wallach, I.; Lilien, R. Virtual Decoy Sets for Molecular Docking Benchmarks. *J. Chem. Inf. Model.* **2011**, *51* (2), 196–202. https://doi.org/10.1021/ci100374f.

(403) Ibrahim, T. M.; Bauer, M. R.; Boeckler, F. M. Applying DEKOIS 2.0 in Structure-Based Virtual Screening to Probe the Impact of Preparation Procedures and Score Normalization. *J Cheminform* **2015**, *7*, 21. https://doi.org/10.1186/s13321-015-0074-6.

(404) Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrián-Uhalte, E.; et al. The ChEMBL Database in 2017. *Nucleic Acids Res* **2017**, *45* (Database issue), D945–D954. https://doi.org/10.1093/nar/gkw1074.

(405) Benson, M. L.; Smith, R. D.; Khazanov, N. A.; Dimcheff, B.; Beaver, J.; Dresslar, P.; Nerothin, J.; Carlson, H. A. Binding MOAD, a High-Quality Protein–Ligand Database. *Nucleic Acids Res* **2008**, *36* (Database issue), D674–D678. https://doi.org/10.1093/nar/gkm911.

(406) Ahmed, A.; Smith, R. D.; Clark, J. J.; Dunbar, J. B.; Carlson, H. A. Recent Improvements to Binding MOAD: A Resource for Protein-Ligand Binding Affinities and Structures. *Nucleic Acids Research* **2015**, *43* (D1), D465–D469. https://doi.org/10.1093/nar/gku1088.

(407) Block, P.; Sotriffer, C. A.; Dramburg, I.; Klebe, G. AffinDB: A Freely Accessible Database of Affinities for Protein–Ligand Complexes from the PDB. *Nucleic Acids Res* **2006**, *34* (Database issue), D522–D526. https://doi.org/10.1093/nar/gkj039.

(408) Roth, B. L.; Lopez, E.; Patel, S.; Kroeze, W. K. The Multiplicity of Serotonin Receptors: Uselessly Diverse Molecules or an Embarrassment of Riches? *Neuroscientist* **2000**, *6* (4), 252–262. https://doi.org/10.1177/107385840000600408.

(409) Placzek, S.; Schomburg, I.; Chang, A.; Jeske, L.; Ulbrich, M.; Tillack, J.; Schomburg, D. BRENDA in 2017: New Perspectives and New Tools in BRENDA. *Nucleic Acids Res.* **2017**, *45* (D1), D380–D388. https://doi.org/10.1093/nar/gkw952.

(410) Munk, C.; Isberg, V.; Mordalski, S.; Harpsøe, K.; Rataj, K.; Hauser, A. S.; Kolb, P.; Bojarski, A. J.; Vriend, G.; Gloriam, D. E. GPCRdb: The G Protein-coupled Receptor Database – an Introduction. *Br J Pharmacol* **2016**, *173* (14), 2195–2207. https://doi.org/10.1111/bph.13509.

(411) D3R | About D3R https://drugdesigndata.org/about/about-d3r (accessed May 6, 2019).

(412) Attene-Ramos, M. S.; Miller, N.; Huang, R.; Michael, S.; Itkin, M.; Kavlock, R. J.; Austin, C. P.; Shinn, P.; Simeonov, A.; Tice, R. R.; et al. The Tox21 Robotic Platform for Assessment of Environmental Chemicals - from Vision to Reality. *Drug Discov Today* **2013**, *18* (0), 716–723. https://doi.org/10.1016/j.drudis.2013.05.015.

(413) LIU, J.; TANG, W.; CHEN, G.; LU, Y.; FENG, C.; TU, X. M. Correlation and Agreement: Overview and Clarification of Competing Concepts and Measures. *Shanghai Arch Psychiatry 28* (2), 115–120. https://doi.org/10.11919/j.issn.1002-0829.216045.

(414) Triballeau, N.; Acher, F.; Brabet, I.; Pin, J.-P.; Bertrand, H.-O. Virtual Screening Workflow Development Guided by the "Receiver Operating Characteristic" Curve Approach. Application to High-Throughput Docking on Metabotropic Glutamate Receptor Subtype 4. *J. Med. Chem.* **2005**, *48* (7), 2534–2547. https://doi.org/10.1021/jm049092j.

(415) Schlessinger, A.; Geier, E.; Fan, H.; Irwin, J. J.; Shoichet, B. K.; Giacomini, K. M.; Sali, A. Structure-Based Discovery of Prescription Drugs That Interact with the Norepinephrine Transporter, NET. *Proceedings of the National Academy of Sciences* **2011**, *108* (38), 15810–15815. https://doi.org/10.1073/pnas.1106030108.

(416) McGaughey, G. B.; Sheridan, R. P.; Bayly, C. I.; Culberson, J. C.; Kreatsoulas, C.; Lindsley, S.; Maiorov, V.; Truchon, J.-F.; Cornell, W. D. Comparison of Topological, Shape, and Docking Methods in Virtual Screening. *J. Chem. Inf. Model.* **2007**, *47* (4), 1504–1519. https://doi.org/10.1021/ci700052x.

(417) Sheridan, R. P.; Singh, S. B.; Fluder, E. M.; Kearsley, S. K. Protocols for Bridging the Peptide to Nonpeptide Gap in Topological Similarity Searches. *J Chem Inf Comput Sci* **2001**, *41* (5), 1395–1406.

(418) Allen, W. J.; Rizzo, R. C. Implementation of the Hungarian Algorithm to Account for Ligand Symmetry and Similarity in Structure-Based Design. *Journal of Chemical Information and Modeling* **2014**, *54* (2), 518–529. https://doi.org/10.1021/ci400534h.

(419) Kuhn, H. W. The Hungarian Method for the Assignment Problem. *Naval Research*

Logistics Quarterly **1955**, *2* (1–2), 83–97. https://doi.org/10.1002/nav.3800020109.

(420) Yusuf, D.; Davis, A. M.; Kleywegt, G. J.; Schmitt, S. An Alternative Method for the Evaluation of Docking Performance: RSR vs RMSD. *Journal of Chemical Information and Modeling* **2008**, *48* (7), 1411–1422. https://doi.org/10.1021/ci800084x.

(421) Hollman, D. A. A.; Milona, A.; van Erpecum, K. J.; van Mil, S. W. C. Anti-Inflammatory and Metabolic Actions of FXR: Insights into Molecular Mechanisms. *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids* **2012**, *1821* (11), 1443–1452. https://doi.org/10.1016/j.bbalip.2012.07.004.

(422) Lamers, C.; Schubert-Zsilavecz, M.; Merk, D. Therapeutic Modulators of Peroxisome Proliferator-Activated Receptors (PPAR): A Patent Review (2008–Present). *Expert Opinion on Therapeutic Patents* **2012**, *22* (7), 803–841. https://doi.org/10.1517/13543776.2012.699042.

(423) Pascual-García, M.; Valledor, A. F. Biological Roles of Liver X Receptors in Immune Cells. *Arch. Immunol. Ther. Exp.* **2012**, *60* (4), 235–249. https://doi.org/10.1007/s00005-012-0179-9.

(424) Verhoeven Guido; Willems Ariane; Denolet Evi; Swinnen Johannes V.; De Gendt Karel. Androgens and Spermatogenesis: Lessons from Transgenic Mouse Models. *Philosophical Transactions of the Royal Society B: Biological Sciences* **2010**, *365* (1546), 1537–1556. https://doi.org/10.1098/rstb.2009.0117.

(425) Ottow, E.; Weinmann, H. *Nuclear Receptors as Drug Targets*; John Wiley & Sons, 2008.

(426) 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.

(427) Volle, D. H. Nuclear Receptors as Pharmacological Targets, Where Are We Now? *Cell. Mol. Life Sci.* **2016**, *73* (20), 3777–3780. https://doi.org/10.1007/s00018-016-2327-6.

(428) Sever, R.; Glass, C. K. Signaling by Nuclear Receptors. *Cold Spring Harb Perspect Biol* **2013**, *5* (3). https://doi.org/10.1101/cshperspect.a016709.

(429) Burris, T. P.; Solt, L. A.; Wang, Y.; Crumbley, C.; Banerjee, S.; Griffett, K.; Lundasen, T.; Hughes, T.; Kojetin, D. J. Nuclear Receptors and Their Selective Pharmacologic Modulators. *Pharmacol. Rev.* **2013**, *65* (2), 710–778. https://doi.org/10.1124/pr.112.006833.

(430) 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.

(431) 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. *The Journal of Steroid Biochemistry and Molecular Biology* **2005**, *93* (2), 127–137. https://doi.org/10.1016/j.jsbmb.2005.01.004.

(432) Kojetin, D. J.; Burris, T. P. Small Molecule Modulation of Nuclear Receptor Conformational Dynamics: Implications for Function and Drug Discovery. *Mol Pharmacol* **2013**, *83* (1), 1–8. https://doi.org/10.1124/mol.112.079285.

(433) Huang, P.; Chandra, V.; Rastinejad, F. Structural Overview of the Nuclear Receptor Superfamily: Insights into Physiology and Therapeutics. *Annu. Rev. Physiol.* **2010**, *72*, 247–272. https://doi.org/10.1146/annurev-physiol-021909-135917.

(434) Grey, A. B.; Stapleton, J. P.; Evans, M. C.; Reid, I. R. The Effect of the Anti-Estrogen Tamoxifen on Cardiovascular Risk Factors in Normal Postmenopausal Women. *J. Clin. Endocrinol. Metab.* **1995**, *80* (11), 3191–3195. https://doi.org/10.1210/jcem.80.11.7593425.

(435) Love, R. R.; Mazess, R. B.; Barden, H. S.; Epstein, S.; Newcomb, P. A.; Jordan, V. C.; Carbone, P. P.; DeMets, D. L. Effects of Tamoxifen on Bone Mineral Density in Postmenopausal Women with Breast Cancer. *N. Engl. J. Med.* **1992**, *326* (13), 852–856. https://doi.org/10.1056/NEJM199203263261302.

(436) European Medicines Agency https://www.ema.europa.eu/en (accessed May 9, 2019).

(437) Nissen, S. E.; Wolski, K. Effect of Rosiglitazone on the Risk of Myocardial Infarction and Death from Cardiovascular Causes. *New England Journal of Medicine* **2007**, *356* (24), 2457–2471. https://doi.org/10.1056/NEJMoa072761.

(438) Mahaffey, K. W.; Hafley, G.; Dickerson, S.; Burns, S.; Tourt-Uhlig, S.; White, J.; Newby, L. K.; Komajda, M.; McMurray, J.; Bigelow, R.; et al. Results of a Reevaluation of Cardiovascular Outcomes in the RECORD Trial. *American Heart Journal* **2013**, *166* (2), 240-249.e1. https://doi.org/10.1016/j.ahj.2013.05.004.

(439) Walsh, L. J.; Wong, C. A.; Pringle, M.; Tattersfield, A. E. Use of Oral Corticosteroids in the Community and the Prevention of Secondary Osteoporosis: A Cross Sectional Study. *BMJ* **1996**, *313* (7053), 344–346.

(440) Buttgereit, F.; Straub, R. H.; Wehling, M.; Burmester, G.-R. Glucocorticoids in the Treatment of Rheumatic Diseases: An Update on the Mechanisms of Action. *Arthritis Rheum.* **2004**, *50* (11), 3408–3417. https://doi.org/10.1002/art.20583.

(441) WHO | State of the science of endocrine disrupting chemicals - 2012 http://www.who.int/ceh/publications/endocrine/en/ (accessed Jul 1, 2019).

(442) Huang, R.; Xia, M.; Cho, M.-H.; Sakamuru, S.; Shinn, P.; Houck, K. A.; Dix, D. J.; Judson, R. S.; Witt, K. L.; Kavlock, R. J.; et al. Chemical Genomics Profiling of Environmental Chemical Modulation of Human Nuclear Receptors. *Environ Health Perspect* **2011**, *119* (8),

1142–1148. https://doi.org/10.1289/ehp.1002952.

(443) Michel, P.; Averty, B. Bilan 1997 de la contamination des eaux côtières françaises par les composés organostanniques. **1998**.

(444) Horiguchi, T. *Biological Effects by Organotins*; Springer, 2016.

(445) le Maire, A.; Grimaldi, M.; Roecklin, D.; Dagnino, S.; Vivat-Hannah, V.; Balaguer, P.; Bourguet, W. Activation of RXR-PPAR Heterodimers by Organotin Environmental Endocrine Disruptors. *EMBO Rep.* **2009**, *10* (4), 367–373. https://doi.org/10.1038/embor.2009.8.

(446) Luccio-Camelo, D. C.; Prins, G. S. Disruption of Androgen Receptor Signaling in Males by Environmental Chemicals. *J Steroid Biochem Mol Biol* **2011**, *127* (1–2), 74–82. https://doi.org/10.1016/j.jsbmb.2011.04.004.

(447) DICOM_Jocelyne.M; DICOM_Jocelyne.M. 4ème Plan national santé environnement : « Mon environnement, ma santé » et consultation publique sur le projet de nouvelle stratégie nationale sur les perturbateurs endocriniens https://solidarites-sante.gouv.fr/actualites/presse/communiques-de-presse/article/4eme-plan-national-sante-environnement-intitule-mon-environnement-ma-sante-et (accessed May 9, 2019).

(448) Huang, R.; Xia, M.; Sakamuru, S.; Zhao, J.; Shahane, S. A.; Attene-Ramos, M.; Zhao, T.; Austin, C. P.; Simeonov, A. Modelling the Tox21 10 K Chemical Profiles for in Vivo Toxicity Prediction and Mechanism Characterization. *Nat Commun* **2016**, *7*. https://doi.org/10.1038/ncomms10425.

(449) 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 https://pubs.acs.org/doi/abs/10.1021/ci400649p (accessed May 9, 2019). https://doi.org/10.1021/ci400649p.

(450) Posner, B. A.; Xi, H.; Mills, J. E. J. Enhanced HTS Hit Selection via a Local Hit Rate Analysis. *J. Chem. Inf. Model.* **2009**, *49* (10), 2202–2210. https://doi.org/10.1021/ci900113d.

(451) Huang, R.; Xia, M.; Sakamuru, S.; Zhao, J.; Shahane, S. A.; Attene-Ramos, M.; Zhao, T.; Austin, C. P.; Simeonov, A. Modelling the Tox21 10 K Chemical Profiles for *in Vivo* Toxicity Prediction and Mechanism Characterization. *Nature Communications* **2016**, *7*, 10425. https://doi.org/10.1038/ncomms10425.

(452) Cummings, M. D.; DesJarlais, R. L.; Gibbs, A. C.; Mohan, V.; Jaeger, E. P. Comparison of Automated Docking Programs as Virtual Screening Tools. *J. Med. Chem.* **2005**, *48* (4), 962–976. https://doi.org/10.1021/jm049798d.

(453) Irwin, J. J. Community Benchmarks for Virtual Screening. *J. Comput. Aided Mol. Des.* **2008**, *22* (3–4), 193–199. https://doi.org/10.1007/s10822-008-9189-4.

(454) 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.

(455) Rohrer, S. G.; Baumann, K. Maximum Unbiased Validation (MUV) Data Sets for Virtual Screening Based on PubChem Bioactivity Data. *J. Chem. Inf. Model.* **2009**, *49* (2), 169–184. https://doi.org/10.1021/ci8002649.

(456) Xia, J.; Tilahun, E. L.; Reid, T.-E.; Zhang, L.; Wang, X. S. Benchmarking Methods and Data Sets for Ligand Enrichment Assessment in Virtual Screening. *Methods* **2015**, *0*, 146–157. https://doi.org/10.1016/j.ymeth.2014.11.015.

(457) Dickersin, K. The Existence of Publication Bias and Risk Factors for Its Occurrence. *JAMA* **1990**, *263* (10), 1385–1389.

(458) Nassir Ghaemi, S.; Shirzadi, A. A.; Filkowski, M. Publication Bias and the Pharmaceutical Industry: The Case of Lamotrigine in Bipolar Disorder. *Medscape J Med* **2008**, *10* (9), 211.

(459) Guasch, L.; Sala, E.; Castell-Auví, A.; Cedó, L.; Liedl, K. R.; Wolber, G.; Muehlbacher, M.; Mulero, M.; Pinent, M.; Ardévol, A.; et al. Identification of PPARgamma Partial Agonists of Natural Origin (I): Development of a Virtual Screening Procedure and in Vitro Validation. *PLoS ONE* **2012**, *7* (11), e50816. https://doi.org/10.1371/journal.pone.0050816.

(460) 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.

(461) Verkhivker, G. M.; Bouzida, D.; Gehlhaar, D. K.; Rejto, P. A.; Arthurs, S.; Colson, A. B.; Freer, S. T.; Larson, V.; Luty, B. A.; Marrone, T.; et al. Deciphering Common Failures in Molecular Docking of Ligand-Protein Complexes. *J. Comput. Aided Mol. Des.* **2000**, *14* (8), 731–751.

(462) Wacker, D.; Stevens, R. C.; Roth, B. L. How Ligands Illuminate GPCR Molecular Pharmacology. *Cell* **2017**, *170* (3), 414–427. https://doi.org/10.1016/j.cell.2017.07.009.

(463) Brzozowski, A. M.; Pike, A. C.; Dauter, Z.; Hubbard, R. E.; Bonn, T.; Engström, O.; Ohman, L.; Greene, G. L.; Gustafsson, J. A.; Carlquist, M. Molecular Basis of Agonism and Antagonism in the Oestrogen Receptor. *Nature* **1997**, *389* (6652), 753–758. https://doi.org/10.1038/39645.

(464) Glass, C. K.; Rosenfeld, M. G. The Coregulator Exchange in Transcriptional Functions of Nuclear Receptors. *Genes Dev.* **2000**, *14* (2), 121–141.

(465) 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* (2–5), 127–137. https://doi.org/10.1016/j.jsbmb.2005.01.004.

(466) Spencer, T. A.; Li, D.; Russel, J. S.; Collins, J. L.; Bledsoe, R. K.; Consler, T. G.; Moore, L. B.; Galardi, C. M.; McKee, D. D.; Moore, J. T.; et al. Pharmacophore Analysis of the Nuclear Oxysterol Receptor LXRα. *J. Med. Chem.* **2001**, *44* (6), 886–897. https://doi.org/10.1021/jm0004749.

(467) Schuster, D.; Langer, T. The Identification of Ligand Features Essential for PXR Activation by Pharmacophore Modeling. *J Chem Inf Model* **2005**, *45* (2), 431–439. https://doi.org/10.1021/ci049722q.

(468) Lewis, S. N.; Garcia, Z.; Hontecillas, R.; Bassaganya-Riera, J.; Bevan, D. R. Pharmacophore Modeling Improves Virtual Screening for Novel Peroxisome Proliferator-Activated Receptor-Gamma Ligands. *J. Comput. Aided Mol. Des.* **2015**, *29* (5), 421–439. https://doi.org/10.1007/s10822-015-9831-x.

(469) Amadio, M.; Govoni, S.; Pascale, A. Targeting VEGF in Eye Neovascularization: What's New?: A Comprehensive Review on Current Therapies and Oligonucleotide-Based Interventions under Development. *Pharmacol. Res.* **2016**, *103*, 253–269. https://doi.org/10.1016/j.phrs.2015.11.027.

(470) Plein, A.; Fantin, A.; Ruhrberg, C. Neuropilin Regulation of Angiogenesis, Arteriogenesis, and Vascular Permeability. *Microcirculation* **2014**, *21* (4), 315–323. https://doi.org/10.1111/micc.12124.

(471) Bielenberg, D. R.; Pettaway, C. A.; Takashima, S.; Klagsbrun, M. Neuropilins in Neoplasms: Expression, Regulation, and Function. *Exp. Cell Res.* **2006**, *312* (5), 584–593. https://doi.org/10.1016/j.yexcr.2005.11.024.

(472) Pan, Q.; Chathery, Y.; Wu, Y.; Rathore, N.; Tong, R. K.; Peale, F.; Bagri, A.; Tessier-Lavigne, M.; Koch, A. W.; Watts, R. J. Neuropilin-1 Binds to VEGF121 and Regulates Endothelial Cell Migration and Sprouting. *J. Biol. Chem.* **2007**, *282* (33), 24049–24056. https://doi.org/10.1074/jbc.M703554200.

(473) Prud'homme, G. J.; Glinka, Y. Neuropilins Are Multifunctional Coreceptors Involved in Tumor Initiation, Growth, Metastasis and Immunity. *Oncotarget* **2012**, *3* (9), 921–939.

(474) Soker, S.; Takashima, S.; Miao, H. Q.; Neufeld, G.; Klagsbrun, M. Neuropilin-1 Is Expressed by Endothelial and Tumor Cells as an Isoform-Specific Receptor for Vascular Endothelial Growth Factor. *Cell* **1998**, *92* (6), 735–745. https://doi.org/10.1016/S0092-8674(00)81402-6.

(475) Pan, Q.; Chanthery, Y.; Liang, W.-C.; Stawicki, S.; Mak, J.; Rathore, N.; Tong, R. K.; Kowalski, J.; Yee, S. F.; Pacheco, G.; et al. Blocking Neuropilin-1 Function Has an Additive Effect with Anti-VEGF to Inhibit Tumor Growth. *Cancer Cell* **2007**, *11* (1), 53–67. https://doi.org/10.1016/j.ccr.2006.10.018.

(476) Starzec, A.; Ladam, P.; Vassy, R.; Badache, S.; Bouchemal, N.; Navaza, A.; du Penhoat, C. H.; Perret, G. Y. Structure-Function Analysis of the Antiangiogenic ATWLPPR Peptide Inhibiting VEGF(165) Binding to Neuropilin-1 and Molecular Dynamics Simulations of the ATWLPPR/Neuropilin-1 Complex. *Peptides* **2007**, *28* (12), 2397–2402. https://doi.org/10.1016/j.peptides.2007.09.013.

(477) Novoa, A.; Pellegrini-Moïse, N.; Bechet, D.; Barberi-Heyob, M.; Chapleur, Y. Sugar-Based Peptidomimetics as Potential Inhibitors of the Vascular Endothelium Growth Factor Binding to Neuropilin-1. *Bioorg. Med. Chem.* **2010**, *18* (9), 3285–3298. https://doi.org/10.1016/j.bmc.2010.03.012.

(478) Nasarre, C.; Roth, M.; Jacob, L.; Roth, L.; Koncina, E.; Thien, A.; Labourdette, G.; Poulet, P.; Hubert, P.; Crémel, G.; et al. Peptide-Based Interference of the Transmembrane Domain of Neuropilin-1 Inhibits Glioma Growth in Vivo. *Oncogene* **2010**, *29* (16), 2381–2392. https://doi.org/10.1038/onc.2010.9.

(479) Jarvis, A.; Allerston, C. K.; Jia, H.; Herzog, B.; Garza-Garcia, A.; Winfield, N.; Ellard, K.; Aqil, R.; Lynch, R.; Chapman, C.; et al. Small Molecule Inhibitors of the Neuropilin-1 Vascular Endothelial Growth Factor A (VEGF-A) Interaction. *J Med Chem* **2010**, *53* (5), 2215–2226. https://doi.org/10.1021/jm901755g.

(480) Borriello, L.; Montès, M.; Lepelletier, Y.; Leforban, B.; Liu, W.-Q.; Demange, L.; Delhomme, B.; Pavoni, S.; Jarray, R.; Boucher, J. L.; et al. Structure-Based Discovery of a Small Non-Peptidic Neuropilins Antagonist Exerting in Vitro and in Vivo Anti-Tumor Activity on Breast Cancer Model. *Cancer Lett.* **2014**, *349* (2), 120–127. https://doi.org/10.1016/j.canlet.2014.04.004.

(481) Life Science & Diagnostic Brands: Bio-Techne https://www.bio-techne.com/ (accessed Jun 13, 2019).

(482) Chang, M.; Belew, R. K.; Carroll, K. S.; Olson, A. J.; Goodsell, D. S. Empirical Entropic Contributions in Computational Docking: Evaluation in APS Reductase Complexes. *J Comput Chem* **2008**, *29* (11), 1753–1761. https://doi.org/10.1002/jcc.20936.

(483) Wu, S.; Liu, B. Application of Scintillation Proximity Assay in Drug Discovery. *BioDrugs* **2005**, *19* (6), 383–392.

(484) Zeng, H.; Xu, W. Chapter 16 - Enzymatic Assays of Histone Methyltransferase

Enzymes. In *Epigenetic Technological Applications*; Zheng, Y. G., Ed.; Academic Press: Boston, 2015; pp 333–361. https://doi.org/10.1016/B978-0-12-801080-8.00016-8.

Annexes

Annexe 1 Tests dépendants de ligands labellisés

Transfert d'énergie entre molécules fluorescentes (FRET)

La méthode de transfert d'énergie entre molécules fluorescente s'appuie sur les propriétés optiques des fluorophores : chaque fluorophore possède son propre spectre d'excitation et d'absorption. Si le spectre d'émission d'un fluorophore donneur (A) recouvre le spectre d'excitation d'un fluorophore accepteur (B), alors un transfert d'énergie de fluorescence aura lieu entre les fluorophores A et B à proximité l'un de l'autre (< 10nM), permettant à B d'émettre. Dans le cas des interactions ligand/protéine, chaque entité biologique est marquée par un fluorophore soit donneur, soit accepteur. Suivant le principe expliqué ci-dessus, la mesure de l'émission du fluorophore accepteur permet d'évaluer les interactions ligand/protéine ayant lieu.

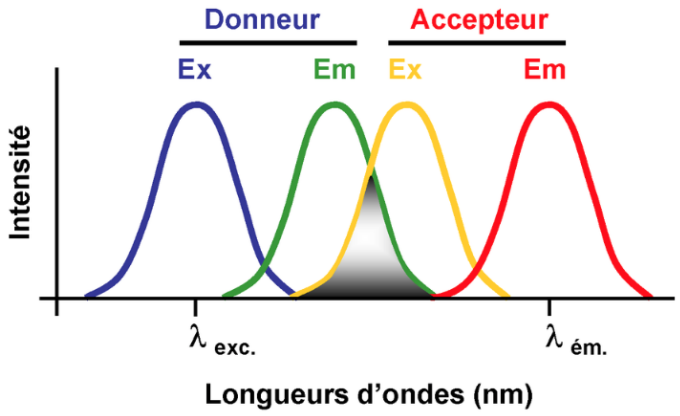


Figure 66 Représentation du phénomène de transfert énergétique exploité par la méthode FRET. Un fluorophore donneur est excité à une longueur d'onde λ_{exc} et émet à une longueur d'onde recouvrant le spectre d'excitation du fluorophore accepteur, qui émet à une longueur d'onde λ_{em}. D'après Damien Maurel

Analyse de la scintillation par proximité

L'analyse de la scintillation par proximité, plus communément appelé Scintillation Proximity Assay (SPA), est utilisée aussi bien pour l'analyse d'interactions petite molécule/protéine, protéine/peptide, ou ADN/protéine.

La mesure de l'affinité d'un ligand pour une protéine par SPA se fait par compétition[483]. Dans un premier temps un ligand de référence est radio-marqué en utilisant des radio-isotopes (ex : 3H, 14C, 32P, 35S ou 125I) et des billes de métal scintillant sont recouvertes de protéines. Les protéines sont fixées sur les

billes par l'intermédiaire de couple substrat/enzyme tels que le couple glutathion/glutathion S-transférase, ou streptavidin/biotine, le premier membre ayant une affinité pour le métal, l'autre étant attaché aux protéines étudiées. Lorsque les ligands radio-marqués se lient aux protéines, le métal scintillant est excité et émet un photon qui est capté par un détecteur de scintillation. L'ajout d'une molécule compétitrice non radio-marquée vient perturber ces interactions et diminuer le signal lumineux. Cette différence de signal permet de mesurer l'IC_{50} de la molécule compétitrice.

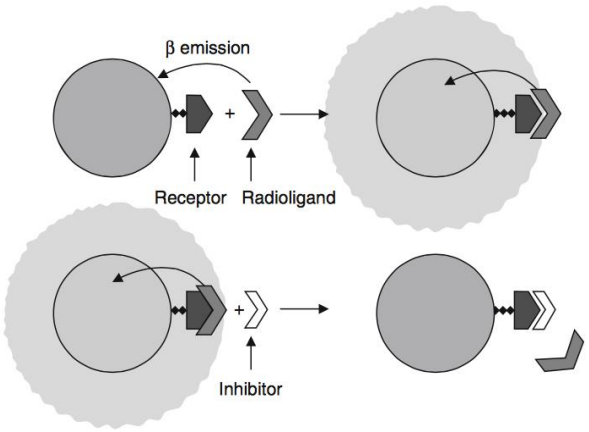


Figure 67 Illustration du principe du test de scintillation par proximité par compétition d'après [522] Lorsque le ligand de référence radio-marqué interagit avec son récepteur, le rayonnement gamma qu'il émet excite la bille en métal scintillant et entraine l'émission d'un signal lumineux. Une molécule d'une molécule compétitrice va déplacer tout ou partie des ligands radio-marqués et engendrer une diminution du signal lumineux.

Polarisation de Fluorescence (FP)

La polarisation de fluorescence part du principe que le degré de polarisation d'un fluorophore est directement lié à son mouvement de rotation[484](Figure 74). Un fluorophore en rotation rapide dévie la lumière et engendre une forte dépolarisation, alors qu'une rotation lente entraine une plus faible dépolarisation.

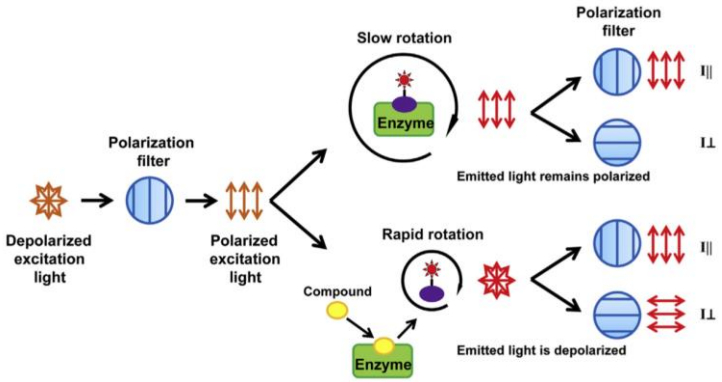


Figure 68 Illustration du principe de la polarisation de fluorescence. Lorsqu'un ligand libre est excité par une lumière polarisée, sa rotation rapide engendre une forte dépolarisation. S'il se fixe sur la protéine d'intérêt, sa vitesse de rotation diminue et on observe une plus faible dépolarisation. D'après [524]

Dans le cas de l'interaction ligand/protéine, c'est le ligand qui est marqué avec un marqueur fluorescent et placé dans un milieu homogène où il est excité par une lumière polarisée et dans lequel on ajoute la protéine d'intérêt. Si le ligand reste libre dans le solvant, sa rotation rapide engendre une forte dépolarisation. S'il se fixe sur la protéine d'intérêt, sa vitesse de rotation diminue et on observe une faible dépolarisation.

Annexe 2 Tests indépendants de ligands labellisés

Résonance des plasmons de surface (SPR)

La résonance des plasmons de surface se base sur les propriétés des matériaux : lorsque des métaux comme l'or ou l'argent sont en contact avec l'eau, l'air, ou un milieu biologique, une onde de surface (plasmon de surface) se forme à l'interface. Cette onde peut être excitée par un faisceau lumineux. Lorsque la surface métallique subit une variation de propriétés physiques, l'onde de surface va subir des changements d'amplitude et de phase qui vont se répercuter sur l'indice de réfraction de la lumière réfléchie.

Dans le cadre de la mesure des liaisons ligand/protéine, la protéine d'intérêt est adsorbée sur la surface métallique. La fixation d'un ligand sur la protéine va entrainer une modification de la phase et de l'amplitude de l'onde, et par conséquent l'angle de réfraction. Le faisceau réfléchi est capté par un détecteur qui permet de quantifier le taux d'association (k_{on}) et de dissociation (k_{off}), ainsi que la constante de dissociation (kd).

Annexe 3 Liste des jeux de données disponibles sur la ChEMBL.

D'après [84]

Nom	Source	Nombre de composés	Nombre d'essais	Nombre de données d'activité
LITERATURE	Scientific Literature	967 242	963 186	5 635 084
PUBCHEM_BIOASSAY	PubChem BioAssays	489 575	2937	7 559 601
GATES_LIBRARY	Gates Library compound collection	68 490	2	69 444
BINDINGDB	BindingDB Database	68 149	1317	99 061
GSK_TCMDC	GSK Malaria Screening	13 467	6	81 198
ST_JUDE_LEISH	St Jude Leishmania Screening	13 422	6	42 105
USP/USAN	USP Dictionary of USAN and International Drug Names	11 356	0	0
DNDI	Drugs for Neglected Diseases Initiative (DNDi)	7053	233	14 452
ASTRAZENECA	AstraZeneca Deposited Data	5799	15	11 687
NOVARTIS	Novartis Malaria Screening	5614	6	27 888
ORANGE_BOOK	Orange Book	2016	0	0
SUPPLEMENTARY	Deposited Supplementary Bioactivity Data	1786	13	4817
CANDIDATES	Clinical Candidates	1633	0	0

ST_JUDE	St Jude Malaria Screening	1524	16	5456
TP_TRANSPORTER	TP-search Transporter Database	1434	3592	6765
DRUGMATRIX	DrugMatrix	930	113 678	350 929
METABOLISM	Curated Drug Metabolism Pathways	828	0	0
GSK_TB	GSK Tuberculosis Screening	826	15	1814
WHO_TDR	WHO-TDR Malaria Screening	740	16	5853
GSK_TCAKS	GSK Kinetoplastid Screening	592	13	7235
MMV_MBOX	MMV Malaria Box	400	138	45 158
MMV_PBOX	MMV Pathogen Box	400	0	0
ATLAS	Gene Expression Atlas Compounds	398	0	0
DRUGS	Manually Added Drugs	378	0	0
GSK_PKIS	GSK Published Kinase Inhibitor Set	366	456	169 451
OSM	Open Source Malaria Screening	211	22	344
WITHDRAWN	Withdrawn Drugs	192	0	0
TG_GATES	Open TG-GATEs	160	158 199	158 199
SANGER	Sanger Institute Genomics of Drug Sensitivity in Cancer	137	714	73 169
FDA_APPROVAL	FDA Approval Packages	43	1386	1387
HARVARD	Harvard Malaria Screening	37	4	111

Annexe 4 Différentes chimiothèques disponibles

Database	Type	No. Cpds (20/03/2019)	Website		Type de chimiothèque
ZINC15[5]	Publique	750 millions	http://zinc15.docking.org/	En vente	Chimiothèque commerciale
eMolecules	Semi-publique	7 millions	http://www.emolecules.com	En vente	
ChemSpider	Publique	71 millions	http://www.chemspider.com	En vente	
ChemBridge	Commerciale	1.3 millions	http://www.chembridge.com	En vente	
Asinex	Commerciale	286 342	http://www.asinex.com	En vente	
Enamine	Publique	> 700 millions	https://www.enaminestore.com/search	En vente	
Maybridge	Commerciale	53 000	http://www.maybridge.com	En vente	
Pubchem	Publique	252 300 934	http://pubchem.ncbi.nlm.nih.gov	Synthétisables (biologiquement testées)	Chimiothèque de molécules bioactives
DrugBank (version 5.1.2, released 2018-12-20)	Publique	12 108	http://www.drugbank.ca	Synthétisables (biologiquement testées)	

ChEMBL	Publique	1 828 820 (15 207 914 activités)	http://www.ebi.ac.uk/chembldb/index.php	Synthétisables (biologiquement testées)	
SCUBIDOO	Publique	~21 millions 7,805 « building-blocks » commercialement disponible combinant 58 réactions.	http://kolblab.org/scubidoo/index.php	Synthétisables	Chimiothèques statiques de composés virtuels
SAVI	Publique	~283 millions de molécules synthétisables à partir de « building-blocks » disponibles chez Sigma Aldrich	https://cactus.nci.nih.gov/download/savi_download/	Synthétisables	
CH*I*PMUNK	Publique	~95 millions de molécules synthétisables à partir de « building-blocks » issus de la ZINC, de Molecules et de MolPort	http://www.ccb.tu-dortmund.de/ag-koch/chipmunk/	Synthétisables	

Annexe 5 Bases de données de toxicité

D'après [128,129]

Database	Website	Comments
FAERS	http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/default.htm	Adverse Effects Reporting system (FAERS) of post-market safety surveillance for all approved drug and therapeutic biologic products
ACToR	http://actor.epa.gov/actor/faces/ACToRHome.jsp	ACToR (Aggregated Computational Toxicology Resource) is EPA's online warehouse of publicly available chemical toxicity data. ACToR provides the numerical data on over 50000 environmental chemicals searchable by chemical name and/or by chemical structure
Cal/EPA	http://www.oehha.ca.gov/risk/ChemicalDB/index.asp	State of California EPA Toxicity. User can obtain information about CAS number, use, list of synonyms, and a group of criteria for risk assessment of a wanted substance
CCRIS	http://toxnet.nlm.nih.gov/cgi-bin/sis/htmlgen?CCRIS	Chemical carcinogenesis research information system (CCRIS). The numerical data on various carcinogenic endpoints (mice, rats, ames salmonella typhimurium, and human) for over 8000 compounds
CPDB	http://potency.berkeley.edu/	University of California Berkeley carcinogenic potency database contains long-term animal cancer tests on 1547 chemicals
Drugs@FDA	http://www.fda.gov/Drugs/InformationOnDrugs/ ucm135821.htm	Information about brand name and generic prescription and over-the-counter human drugs and biological therapeutic products
DSSTox	http://www.epa.gov/ncct/dsstox/index.html	User can obtain data on toxicity represented by PDF SDF or XLS files

ECOTOX	http://cfpub.epa.gov/ecotox/	User can use quick and/or advanced database query. There is an user guide. There are links to other databases on toxicity
EXTOXNET	http://extoxnet.orst.edu/ghindex.html	University-based database of issues related to pesticide toxicology
FDA Poisonous Plant Database	http://www.accessdata.fda.gov/scripts/plantox/index.cfm	US FDA/CFSAN database with references to scientific literature describing studies of the toxic properties of plants
Gene-Tox	http://toxnet.nlm.nih.gov/cgi-bin/sis/htmlgen?GENETOX	US NLM Peer-reviewed genetic toxicology test data for over 3000 chemicals
HERA	http://www.heraproject.com/RiskAssessment.cfm	Human and Environmental Risk Assessment on Ingredients and Household Cleaning Products; toxicity and risk data on ingredients supplied and formulated by European manufacturers
Household Products	http://hpd.nlm.nih.gov/	This database contains over 12000 consumer brands with numerical criteria their health effects
IRIS	http://cfpub.epa.gov/ncea/iris/index.cfm	Integrated Risk Information System; a compilation of electronic reports on environmental substances and their potential to cause human health effects. User can obtain PDF file with description of toxicological review of a substance in detail
ITER	http://www.tera.org/iter/	Database of human health risk values and cancer classifications for over 600 environmental chemicals
JECDB	http://dra4.nihs.go.jp/mhlw_data/jsp/SearchPageENG.jsp	Japanese Ministry of Health Labour and Welfare Chemical Toxicity Database; toxicity data for 369 chemicals
LAZAR	http://www.in-silico.de/	Lazy structure–activity relationships database; provides QSAR predictions for liver toxicity mutagenicity and carcinogenicity

MR	http://www.atsdr.cdc.gov/mrls/index.html	User can obtain data on minimal risk (MR) levels for hazardous substances represented in this database
N-Class database, KemI	http://apps.kemi.se/nclass/	Using special interface user can define a wanted compound. The system provides information on large list of different endpoints related to this compound
NPIC	http://npic.orst.edu	National Pesticide Information Center through Oregon State University and US EPA provides science-based information about pesticides including toxicity. In fact the database is online encyclopedia for pesticides
NTP	http://ntp.niehs.nih.gov/	US NIH/NIEHS National Toxicology Program testing status and information of agents registered in the US of public health interest. User can obtain documents related to different substances and protocols of definition of different toxic endpoints as well as information on other aspects of toxicology in general
PAN Pesticide	http://www.pesticideinfo.org/	Pesticide Action Network North America; data on 6500 pesticides insecticides and herbicides including toxicity, water pollution, ecological toxicity uses and regulatory status. In fact the database is a digest of pesticides
Riskline, kemi	http://apps.kemi.se/riskline/	Contains information on both environment and health Useful for classification and labelling. Provides links to references associated with a chemical. User can obtain a set of abstracts related to indicated substance
STITCH	http://stitch.embl.de/	Search Tool for Interactions of Chemicals (STITCH). Knowledge database to explore known and predicted interactions between proteins and small-molecule chemicals for understanding of

		molecular and cellular functions. Over 68000 chemicals are represented
TEXTRATOX	http://www.vet.utk.edu/TETRATOX/index.php	The University of Tennessee Institute of Agriculture. A collection of aquatic toxic potency data for more than 2400 industrial organic compounds
TOXNET	http://toxnet.nlm.nih.gov/	Databases on toxicology hazardous chemicals environmental health and toxic releases. User can obtain data on toxic endpoint related to different animals and human as well as data on physicochemical endpoints such as boiling points, water solubility, logP (octanol–water) etc.
ToxRefDB	http://www.epa.gov/ncct/toxrefdb/	US EPA relational database of standard toxicity test results for pesticides and other environmental chemicals including acute, subchronic, chronic, reproductive, and developmental toxicity in support of the ToxCast program. User can obtain data represented by XLS files
ToxCast	https://www.epa.gov/chemical-research/toxicity-forecaster-toxcasttm-data	ToxCast has data for approximately 1,800 chemicals from a broad range of sources including industrial and consumer products, food additives, and potentially green chemicals that could be safer alternatives to existing chemicals. They are screen in more than 700 HTS assays
admetSAR	http://lmmd.ecust.edu.cn/admetsar2	200,000 ADMET annotated data points for about 96 thousands of unique compounds have been manually curated from large literatures
ISSTox	http://www.iss.it/ampp/dati/cont.php?id=233&lang=1&tipo=7	Chemical Toxicity databases from Istituto Superiore di Sanitá (ISS), Italy, designed to be usable for SARs studies for toxicity prediction: (i) long-term carcinogenicity bio- assay on rodents (rat and mouse) (ISSCAN), (ii) *in vitro S. typhimurium* mutagenesis (Ames test) (ISSSTY),

		(iii) *in vivo* mutagenesis (micronucleus test) (ISSMIC), (iv) Cell Transformation Assays (ISSCTA) and (v) Mutagenicity and Carcinogenicity of Biocides (ISSBIOC)
Ames Mutagenicity Benchmark Data Set	https://doc.ml.tu-berlin.de/toxbenchmark	6500 nonconfidential compounds (available as SMILES strings and SDF) together with their biological activity
T3DB	www.t3db.ca	>3600 common toxic substances along with detailed information on their chemical properties, descriptions, targets (>2000), toxic effects, toxicity thresholds, sequences (for both targets and toxins), mechanisms and references
Tox21	https://ntp.niehs.nih.gov/results/tox21/index.html	The Tox21 Phase II library contains approximately 10,000 (10K) compounds and assays focused initially focused on nuclear receptor and stress response pathway

Annexe 6 Descripteurs de poches calculés par les logiciels FPocket et CASTP

Descriptor	Definition	Logiciel
SASA	Total SASA	Fpocket
PolarSASA	Polar SASA	Fpocket
ApolaSASA	Apolar SASA	Fpocket
Volume	Volume	Fpocket
MLHD	Mean Local Hydrophobic density	Fpocket
HydScore	Hydrophobicity score	Fpocket
VolScore	Volume score	Fpocket
PolarityScore	Polarity score	Fpocket
ChargeScore	Charge score	Fpocket
PPA	Proportion polar atoms	Fpocket
Flexibility	Flexibility	Fpocket
N_mth	Number of mouths	CASTP
Area_sa	Area of the solvent accessible surface (Richards' surface)	CASTP
Area_ms	Area of the molecular surface model (Connolly's model)	CASTP
Vol_sa	Volume of the solvent accessible surface (Richards' surface)	CASTP
Vol_ms	Volume of the solvent accessible surface (Richards' surface)	CASTP
Area_sa_mouth	Area of the solvent accessible surface (Richards' surface)	CASTP
Area_ms_mouth	Area of the molecular surface model (Connolly's model)	CASTP
Len_sa_mouth	length of the solvent accessible surface (Richards' surface)	CASTP
Len_ms_mouth	lenght of the solvent accessible surface (Richards' surface)	CASTP

Figure 69 Liste des descripteurs de poches calculés par FPocket et CASTP pris en compte lors de l'étude présentée en Résultats 1.3.

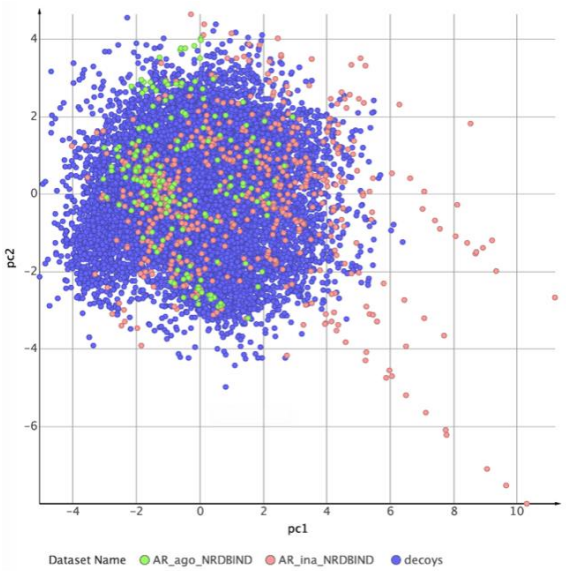
Annexe 7 Étude de la capacité des descripteurs FPocket et CASTP à discriminer les structures d'un NR associé aux meilleures AUCs de celles associées aux moins bonnes AUCs

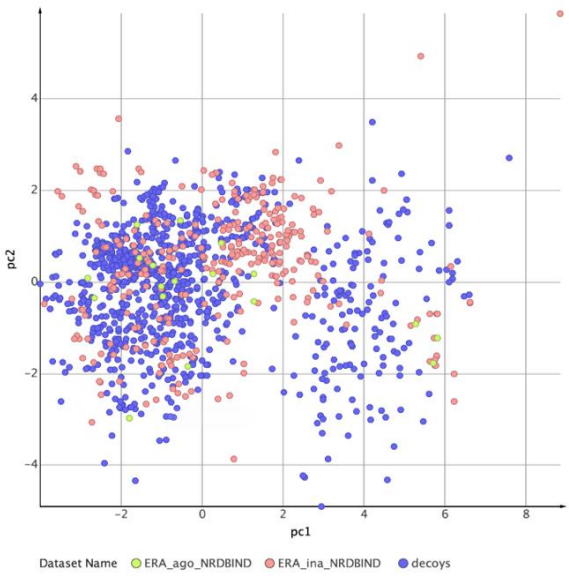
Descripteur	Définition	logiciel	AR	ERA	ERB	FXR	PPARA	PPARA	PR	RXRA
NAS	Number of alpha sphere	Fpocket			Plants +					
SASA	Total SASA				VINA +					
PolarSASA	Polar SASA									
ApolaSASA	Apolar SASA									
Volume	Volume				VINA +					
MLHD	Mean Local Hydrophobic density				Plants +					
MASR	Mean alpha sphere radius				VINA +					
MASSA	Mean alpha sphere solvent access					Plant -, VINA -				
AASP	Apolar apha sphere proportion									
HydScore	Hydrofobicity score									
VolScore	Volume score				VINA +					
PolarityScore	Polarity score									
ChargeScore	Charge score				Plants +, VINA +	Vina +				
PPA	Proportion polar atoms									
ASD	Alpha sphere density				VINA +					
Flexibility	Flexibility					Plant -, VINA -				
N_mth	Number of mouths	CASTP								
Area_sa	Area of the solvent accessible surface (Richards' surface)									
Area_ms	Area of the molecular surface model (Connolly's model)									
Vol_sa	volume of the solvent accessible surface (Richards' surface)		VINA +		VINA +					
Vol_ms	volume of the solvent accessible surface (Richards' surface)									
Length										
Area_sa_mouth	Area of the solvent accessible surface (Richards' surface)									
Area_ms_mouth	Area of the molecular surface model (Connolly's model)									
Len_sa_mouth	length of the solvent accessible surface (Richards' surface)									
Len_ms_mouth	lenght of the solvent accessible surface (Richards' surface)									
Ntri_mouth	Number_triangle									

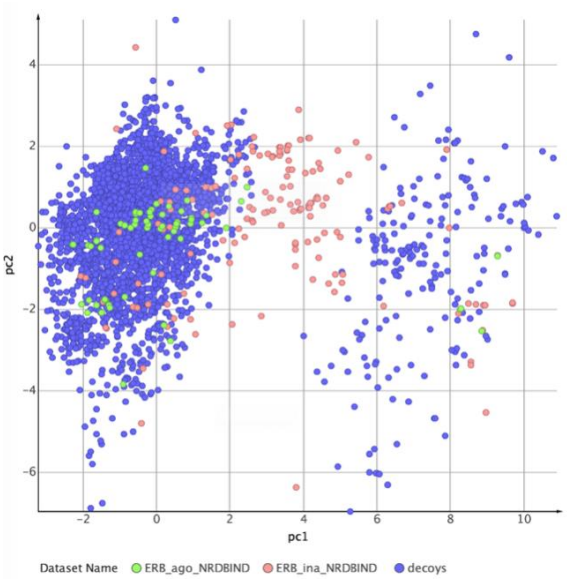
Étude de la capacité des descripteurs FPocket et CASTP à discriminer les structures d'un NR associé aux meilleures AUCs de celles associées aux moins bonnes AUCs. Les structures de chaque NR ont été classées en 3 groupes selon la valeur du descripteur étudié (groupe 1 : x < valeur moyenne – écart type, groupe 2 : valeur moyenne – écart type < x < valeur moyenne + écart type, groupe 3 : x > valeur moyenne – écart type). La comparaison des distributions des

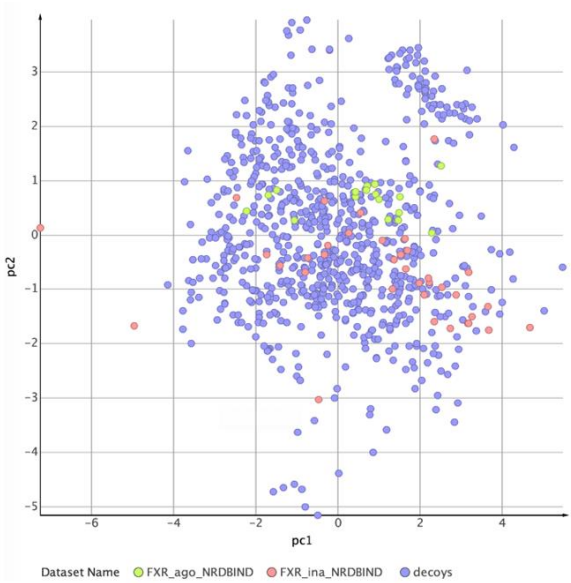
AUCs de chaque groupe a permis d'identifier les descripteurs discriminant les meilleures AUCs des moins bonnes avec VINA et PLANTS. Un signe « - » indique que les plus faibles valeurs du descripteur sont associées au meilleures AUCs, un signe « + » indique que les plus hautes valeurs du descripteur sont associées aux meilleures AUCs

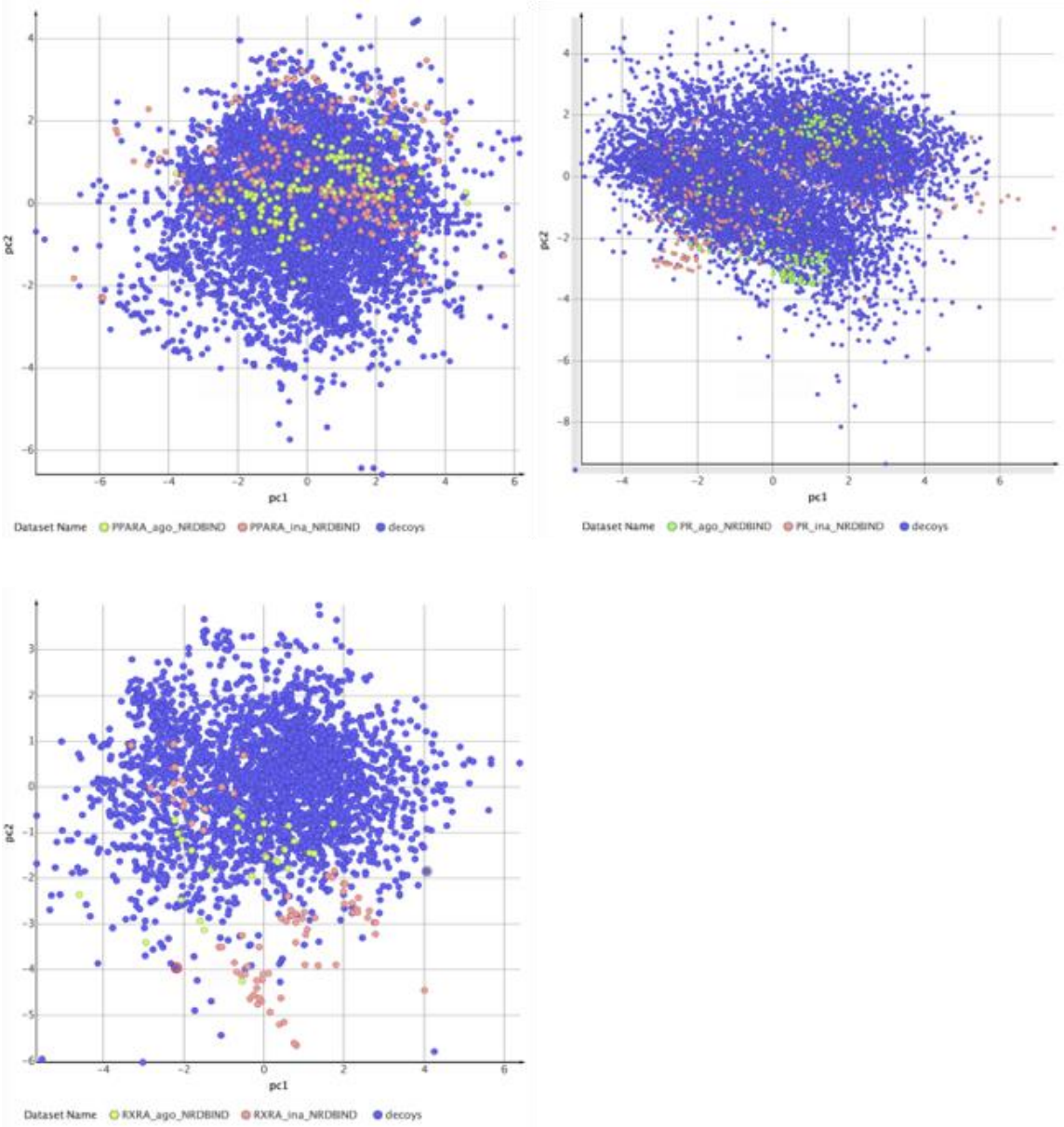
Annexe 8 Analyses en composantes principales des jeux de données de molécules actives et inactives de la NR-DBIND et des *decoys* générés par la DUD-E



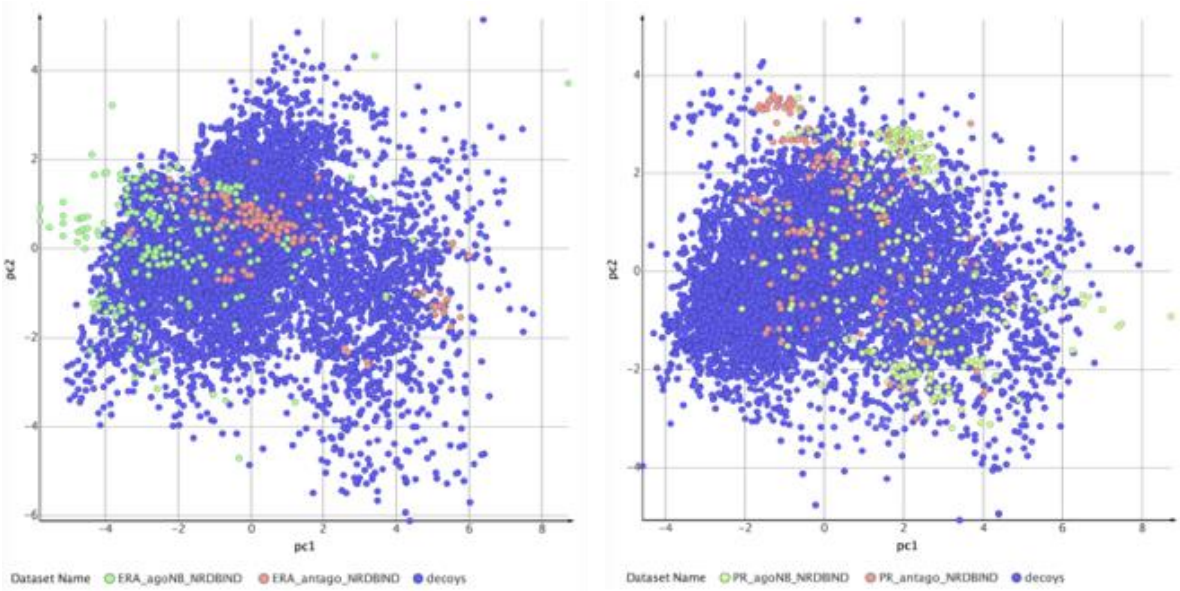






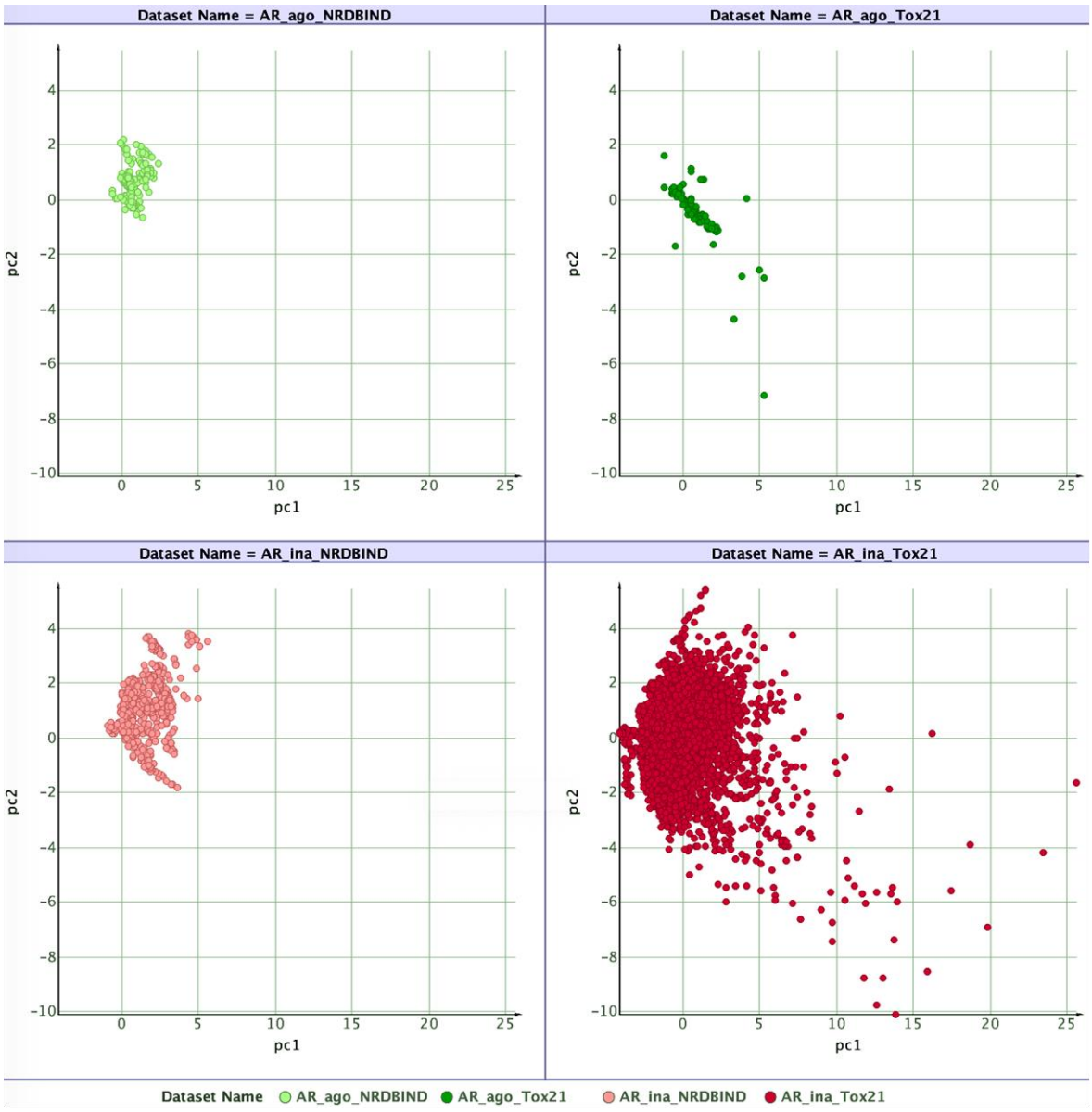


Analyse en composantes principales des jeux de données issus de la NR-DBIND (molécules actives (agonistes (vert)) vs. molécules inactives (antagonistes et *non-binders* (rouge))) et des *decoys* générés par la DUD-E (bleu). La variabilité représentée par les deux axes principaux de l'ACP varie selon la cible étudiée (AR : PC1 = 30%, PC2 = 20% ; ERα : PC1 = 48%, PC2 = 16% ; ERβ : PC1 = 44%, PC2 = 18%; FXR : PC1 = 29%, PC2 = 21%; PPARα : PC1 = 31%, PC2 = 19%; PR : PC1 = 39%, PC2 = 21%). Dans les cas de ERα, ERβ et RXRα, les espaces chimiques des molécules inactives et des *decoys* sont peu recouvrants.

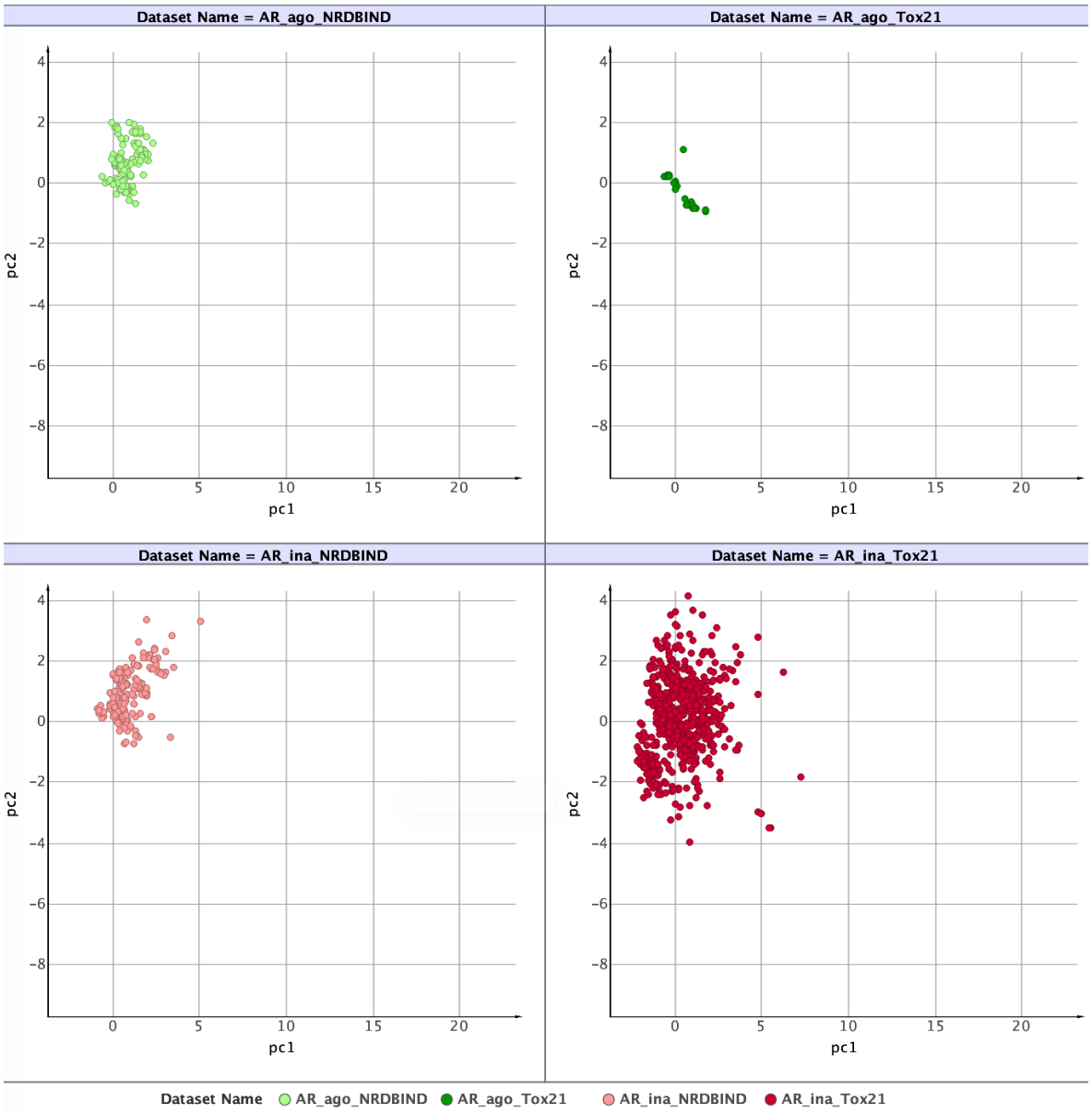


Analyse en composantes principales des jeux de données issus de la NR-DBIND (molécules actives (antagonistes (vert)) vs. molécules inactives (agonistes et *non-binders* (rouge))) et des *decoys* générés par la DUD-E (bleu). La variabilité représentée par les deux axes principaux de l'ACP varie selon la cible étudiée (ERα : PC1 = 39%, PC2 = 19% ; PR : PC1 = 43%, PC2 = 15%). Les *decoys* couvrent un espace chimique plus large que les molécules inactives des jeux de données.

Annexe 9 Analyses en composantes principales des jeux de données issus de la NR-DBIND et de la Tox21 pour le récepteur AR



Analyse en composantes principales des données d'activité (vert) et d'inactivité (rouge) issues de la NR-DBIND (gauche, couleurs pâles) et de la Tox21 (droite, couleurs vives). Les axes principaux de la PCA représentent respectivement 41% et 21% de la variabilité. Les molécules actives des deux jeux de données sont partiellement recouvrantes. Les molécules inactives sont chavauchantes, les données issues de la Tox21 englobant l'espace chimique représenté par celles de la NR-DBIND.



L'analyse en composante principale est ci-dessus représentée en n'affichant que les molécules criblées par l'ensembles des modèles de pharmacophore pKi-pIC50. L'espace chimique des molécules actives de la Tox21 criblées est partiellement chevauchant avec celui des molécucles actives issues de la NR-DBIND, soulignant que le criblage de pharmacophore est dépendant du jeu d'apprentissage mais que son application ne se limite pas strictement à ce dernier. Les molécules inactives criblées partagent globalement un espace chimique proche ; cependant la majorité des molécules partageant le même espace chimique que les molécules actives du jeu de données issus de la NR-DBIND est éliminée.

le cnam

Manon REAU

Importance des données inactives dans les modèles

Application aux méthodes de criblage virtuel en santé humaine et environnementale

Résumé

Le criblage virtuel est utilisé dans la recherche de médicaments et la construction de modèle de prédiction de toxicité. L'application d'un protocole de criblage est précédée par une étape d'évaluation sur une banque de données de référence. La composition des banques d'évaluation est un point critique ; celles-ci opposent généralement des molécules actives à des molécules supposées inactives, faute de publication des données d'inactivité. Les molécules inactives sont néanmoins porteuses d'information. Nous avons donc créé la banque NR-DBIND composée uniquement de molécules actives et inactives expérimentalement validées et dédiées aux récepteurs nucléaires. L'exploitation de la NR-DBIND nous a permis d'étudier l'importance des molécules inactives dans l'évaluation de modèles de docking et dans la construction de modèles de pharmacophores. L'application de protocoles de criblage a permis d'élucider des modes de liaison potentiels de petites molécules sur FXR et NRP-1.

Mots clés : Molécules inactives, modèles, criblage virtuel, docking, pharmacophores, récepteurs nucléaires, banque de données, benchmark, FXR, NRP-1.

Résumé en anglais

Virtual screening is widely used in early stages of drug discovery and to build toxicity prediction models. Commonly used protocols include an evaluation of the performances of different tools on benchmarking databases before applying them for prospective studies. The content of benchmarking tools is a critical point; most benchmarking databases oppose active data to putative inactive due to the scarcity of published inactive data in the literature. Nonetheless, experimentally validated inactive data also bring information. Therefore, we constructed the NR-DBIND, a database dedicated to nuclear receptors that contains solely experimentally validated active and inactive data. The importance of the integration of inactive data in docking and pharmacophore models construction was evaluated using the NR-DBIND data. Virtual screening protocols were used to resolve the potential binding mode of small molecules on FXR and NRP-1.

Keywords: Inactive molecules, models, virtual screening, docking, pharmacophores, nuclear receptors, database, benchmarking, FXR, NRP-1.