Développement d'outils statistiques d'évaluation de méthodes de criblage virtuel : courbes de prédictivité & Screening Explorer

Charly Empereur-Mot

le cnam

ÉCOLE DOCTORALE Sciences des Métiers de l'Ingénieur

Laboratoire Génomique, Bioinformatique et Applications

THÈSE présentée par :

Charly EMPEREUR-MOT

soutenue le : **30 juin 2017**

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

Discipline : Biochimie et biologie moléculaire / Spécialité : Bioinformatique

Développement d'outils statistiques d'évaluation de méthodes de criblage virtuel : courbes de prédictivité & Screening Explorer

THÈSE dirigée par :

M. MONTES Matthieu	Professeur, Directeur de thèse, Cnam
M. ZAGURY Jean-François	Professeur, Co-directeur de thèse, Cnam

RAPPORTEURS :

M. LANGER Thierry	Professeur, Université de Vienne
Mme. MUCCHIELLI Marie-Hélène	Maître de conférences, Institut de Biologie Intégrative de la Cellule

---

JURY :

Mme. GUINOT Christiane	Président du jury, Chercheur associé, Université de Tours
M. MOROY Gautier	Maître de conférences, Université Paris Diderot

« Choisissez un travail que vous aimez et vous n'aurez pas
à travailler un seul jour de votre vie. »
Confucius & Matthieu Montes

Remerciements

Je ne réalisais pas la portée de l'aventure dans laquelle je me lançai, il y a trois ans de cela, lorsque Matthieu Montes et Jean-François Zagury m'ont offert l'opportunité de réaliser une thèse avec eux. Je vous remercie, du fond du cœur, pour la confiance que vous m'avez accordée. Vous avez toujours été disponibles pour moi, vous m'avez soutenu dans les moments difficiles et vous m'avez poussé à donner le meilleur de moi-même. Votre bienveillance, votre excellence et votre patience resteront un souvenir fort et un modèle pour le reste de ma vie. Il faut du courage pour affronter l'inconnu, mais l'issue de cette période périlleuse est une élévation personnelle sans prix.

J'adresse ma sincère gratitude à mes rapporteurs Thierry Langer et Marie-Hélène Mucchielli pour leurs remarques constructives qui ont permis d'améliorer ce manuscript. Je remercie également Christiane Guinot et Gautier Moroy pour avoir accepté d'être membres de mon jury de thèse et pour leur participation à l'évaluation de mon travail. Un très grand merci également à Emmanuel Levy, Marc Baaden, Catherine Etchebest et Jean-Christophe Gelly, qui ont été des mentors scientifiques géniaux et m'ont permis de développer une démarche et un raisonnement scientifique solides.

Je souhaite également adresser mes plus profonds remerciements à Sigrid, Nathalie, Vincent, Taoufik, Josselin, Daniela, Anita, Jean-Louis, Cédric, Hervé, Lieng, Hadley, Lucille, Benjamin, Jérémy, Hélène, Christiane et Juanjo, avec qui j'ai partagé ces trois dernières années, pour rapidement être devenus des amis. Merci pour votre bonne humeur quotidienne, pour les éclats de rire et pour votre soutien, vous avez contribué à faire de ma thèse une expérience inoubliable. J'adresse également des remerciements spéciaux à Manon et Yassine, mes compères de cœur, pour leur soutien et leur coaching particulier.

Plus personnellement, un immense merci, Maman, Papa, Jean-Yves et Carole, pour votre soutien inconditionnel. Les moments familiaux ont été d'une grande importance pour repartir de plus belle. Merci, également, de m'avoir supporté quand les nerfs tendaient à lâcher (en réalité, c'est peut-être pour vous que ça a été le plus dur !).

Résumé

Les méthodes de criblage virtuel sont largement utilisées dans le processus de conception de médicaments afin de réduire le nombre de composés à tester expérimentalement. Cependant, les résultats obtenus par criblage virtuel ne sont que des prédictions et leur fiabilité n'est pas garantie. L'évaluation de ces méthodes est donc essentielle pour guider le bioinformaticien dans le choix de l'outil et du protocol adaptés dans les conditions de son expérience. Dans une première étude, nous avons développé une nouvelle métrique pour l'analyse des résultats de criblage : la Courbe de Prédictivité. Cette métrique permet une analyse fine de la pertinence des scores d'affinité pour la détection de composés actifs et complète les métriques existantes, permettant une meilleure compréhension des résultats de criblage. Lors de notre projet suivant, nous avons souhaité faciliter ce processus d'analyse en intégrant l'ensemble des métriques de criblage virtuel dans un outil web interactif : Screening Explorer. Une seconde partie de ma thèse a consisté en la recherche de nouveaux inhibiteurs du VIH (Virus de l'Immunodéficience Humaine). L'équipe génomique de notre laboratoire a identifié plusieurs gènes dont l'expression influence le développement du SIDA, révèlant ainsi de potentielles cibles thérapeutiques. Une étude bibliographique a permis d'identifier plusieurs composés inhibiteurs de ces cibles. La société Peptinov, associée à notre laboratoire, va prochainement estimer le potentiel thérapeutique de ces composés dans des essais *in vitro* (i) d'infection par le VIH, (ii) de prolifération virale et (iii) de réactivation virale.

Mots-clés : Criblage virtuel, Evaluation, Métriques, Courbe de prédictivité, Conception de médicaments, Petites molécules, VIH, SIDA.

Résumé en anglais

Virtual screening methods are widely used in drug discovery processes in order to reduce the number of compounds to test experimentally. However, virtual screening results are only predictions and their reliability is not guaranteed. Evaluating these methods is crucial to guide the bioinformatician in the choice of the right tool and protocol according to the conditions of his experiment. In a first study, we developed a new metric to analyze the results of virtual screening: the Predictiveness Curve. This metric allows to finely analyze the relevance of binding scores for the detection of active compounds and complete existing metrics, allowing a better comprehension of screening results. In a following project, we facilitated the analysis process by integrating all of the virtuel screening metrics in an interactive tool: Screening Explorer. The second part of my thesis consisted in the research of novel HIV inhibitors. The genomic team of our laboratory identified several genes whose expression influence the development of AIDS, therefore revealing potential therapeutic targets. A bibliographic study allowed to identify compounds that can inhibit those targets. The company Peptinov, associated to our laboratory, is currently estimating the therapeutic potential of the compounds *in vitro* in assays of (i) HIV infection, (ii) viral proliferation and (iii) viral reactivation.

Keywords: Virtual screening, Evaluation, Metrics, Predictiveness curve, Drug design, Small molecules, HIV, AIDS.

Table des matières

Liste des figures

Liste des tableaux

Liste des équations

Liste des abréviations

ACD : Advanced Chemical Directory

ADME-Tox : Absorption Distribution Métabolisme Elimination – Toxicologie

ADN : Acide DésoxyriboNucléique

AMM : Autorisation de Mise sur le Marché

AMPc : Adénosine MonoPhosphate cyclique

ANSM : Agence Nationale de Sécurité du Médicament

ARN : Acide RiboNucléique

ARNm : Acide RiboNucléique messager

AUC : Area Under the Curve

BEDROC : Botlzmann-Enhanced Discrimination of ROC

BOSS : Biochemical and Organic Simulation System

CATS : Chemical Advanced Template Search

CNS : Central Nervous System

COMBINE : COMparative BInding Energy

CODESSA : COmprehensive DEScriptors for Structural and Statistical Analysis

CoMFA : Comparative Molecular Field Analysis

CoMMA : Comparative Molecular Moment Analysis

CoMSA : Comparative Molecular Surface Analysis

CPP : Comité de Protection des Personnes

DISCO : DIStance COmparison

DUD : Directory of Useful Decoys

DUD-E : Directory of Useful Decoys Enhanced

DYLOMMS : DYnamic Lattice-Oriented Molecular Modeling System

EF : Enrichment Factor

EGFR : Epidermal Growth Factor Receptor

eQTL : expression Quantitative Trait Locus

ErG : Extended reduced Graph

FAERS : FDA Adverse Event Reporting System

FDA : Food and Drug Administration

FFN : Fraction des Faux Négatifs

FFP : Fraction des Faux Positifs

FN : Faux Négatif

FP : Faux Positif

FVN : Fraction des Vrais Négatifs

FVP : Fraction des Vrais Positifs

GALAHAD : Genetic Algorithm with Linear Assignment for the Hypermolecular Alignment of Datasets

GAMMA : Genetic Algorithm for Multiple Molecule Alignment

GASP : Genetic Algorithm Superposition Program

GERM : Genetically Evolved Receptor Models

GLM : Generalized Linear Model

GPb : Glycogène Phosphorylase b

GPCR : G-Protein-Coupled Receptors

GWAS : Genome Wide Association Study

HSP : Heat Shock Protein

HTS : High Throughput Screening

ICM : Internal Coordinate Mechanics

KO : Knock-Out

MCS : Maximum Common Substructure

MDDR : MDL Drug Data Report

MLR : Multiple Linear Regression

MOE : Molecular Operating Environment

MPHIL : Mapping PHarmacophore In Ligands

PC : Predictiveness Curve

PCA : Principal Component Analysis

PDB : Protein Data Bank

PHASE : PHarmacophore Alignment and Scoring Engine

PLS : Partial Least Squares

PMF : Potential of Mean Force

QSAR : Quantitative Structure-Activity Relationship

RAPID : RAndomized Pharmacophore Identification for Drug design

RD : Receptor Dependent

RDE : Relative Displacement Error

RI : Receptor Independent

RIE : Robust Initial Enhancement

RMN : Résonance Magnétique Nucléaire

RMSD : Root Mean Square Deviation

ROC : Receiver Operating Characteristics

RSR : Real Space R-factor

SCAMPI : Statistical Classification of Activities of Molecules for Pharmacophore Identification

SIDA : Syndrome d'ImmunoDéficience Acquise

SOMFA : Self-Organizing Molecular Field Analysis

SMARTS : SMiles ARbitrary Target Specification

SMILES : Simplified Molecular-Input Line-Entry System

SNP : Single Nucleotide Polymorphism

T_c : Coefficient de Tanimoto

USR : Ultrafast Shape Recognition

VEGFR : Vascular Epidermal Growth Factor Receptor

VHTS : Virtual High Throughput Screening

VIH : Virus de l'Immunodéficience Humaine

VN : Vrai négatif

VP : Vrai Positif

WDI : World Drug Index

ZINC : ZINC Is Not Commercial

Première partie : Introduction

1. Recherche de nouveaux médicaments

1.1. Historique & Premières découvertes

Si l'on peut considérer que les maladies et les « essais thérapeutiques » remontent à l'existence de l'Homme sur Terre, les premières traces de préparations médicinales dateraient de -1600 avant JC, inscrites sur des papyrus et tablettes d'argile. Pendant des milliers d'années, des substances naturelles d'origine végétale, animale, ou minérale furent utilisées et sélectionnées suite à l'étude empirique de leurs effets sur l'évolution des maladies. La thérapeutique primitive relevait alors d'un caractère « magico-religieux », d'abord transmise oralement à travers les générations, puis conservée grâce à l'utilisation de l'écriture. La transmission des savoirs médicinaux, rendue possible par l'écriture, accélérée par l'imprimerie et aujourd'hui aboutie grâce à la publication quasi-systématique et mondiale des résultats scientifiques, a largement contribué aux succès de la recherche moderne de médicaments.

Cependant, la fin du XVIII^{e} et le début du XIX^{e} siècle marquent un tournant historique dans la recherche de nouveaux médicaments : les chimistes et pharmaciens vont découvrir et isoler les principes actifs des préparations utilisées dans la médecine primitive et traditionnelle. En 1763, Edward Stone décrit les effets de l'acide salicylique, isolé à partir du saule blanc, dans une lettre au président de la Royal Society of London.[1] Son étude ouvrira la voie à la conception de l'aspirine. En 1803, Friedrich Wilhelm Adam Sertürner isole la morphine à partir du pavot, qui sera commercialisée par les laboratoires Merck à partir des années 1827. D'autres découvertes vont conduire à l'utilisation de nouveaux médicaments d'origine minérale : la découverte de l'iode par Louis Joseph Gay-Lussac en 1813, la découverte du brome par Antoine Jérôme Balard en 1826, l'usage du fer réduit préparé par Miquelard et Quevenne en 1840.

L'essor des médicaments de synthèse commence au milieu du XIX^{e} siècle, avec la découverte de la synthèse de l'acide acétylsalicylique par Charles Frédéric Gerhardt en 1853. Le composé, moins toxique que l'acide salicylique et conservant les mêmes vertues thérapeutiques, sera commercialisé par les laboratoires Bayer sous le nom « Aspirin » en 1899. Aujourd'hui, la production d'aspirine mondiale annuelle est estimée à 40.000 tonnes (soit 50 à 120 milliards de doses).[2] La production de morphine annuelle mondiale dépasse 600 tonnes.[3]

Plusieurs grandes disciplines ont permis la révolution de la recherche de nouveaux médicaments et ont abouti aux processus de R&D mis en œuvre actuellement. Le développement de la

médecine moderne, grâce au diagnostic physique, physiologique ou psychologique, permet d'identifier les symptômes et syndromes caractérisant et définissant une maladie. Les progrès en génétique, liées aux séquençages d'ADN et aux techniques statistiques, permettent de rationaliser le choix de cibles d'intérêt thérapeutique, sur lesquelles il sera possible d'agir en cherchant à modifier leur activité biologique. La pharmacologie, les techniques de biologie moléculaire et la robotique ont permis le développement de techniques de criblage à haut débit (High Throughput Screening ou HTS) pour une identification rapide de composés prometteurs, efficaces pour la modification de l'activité d'une cible d'intérêt thérapeutique. La bioinformatique, couplée à la pharmacologie et la chimie, permet la construction de modèles expliquant l'action d'un composé sur une cible d'intérêt thérapeutique, la pratique de criblages virtuels à haut débit (Virtual High Throughput Screening ou VHTS) pour l'optimisation des résultats de campagnes HTS suivantes, ou le guidage des synthèses chimiques dans l'optimisation des composés.[4,5]

1.2. Processus général de R&D

La conception d'un nouveau médicament suit un processus de recherche long et coûteux, principalement engagé par l'industrie pharmaceutique avec un support académique important. Des premières phases de recherche à l'obtention d'une Autorisation de Mise sur le Marché (AMM), le temps nécessaire à la découverte d'un nouveau médicament est estimé, en moyenne, entre 12 et 14 ans pour un coût total de 800 millions à 1 milliard de dollars (dollars de 2010).[6] Le financement des phases d'essai clinique représente à lui seul environ 63% de ce coût (53% de la phase II à l'AMM), tandis que 32% du coût total est associé aux phases de recherche et pré-cliniques.[7]

Suivant l'ensemble des estimations réalisées entre 2000 et 2010, le financement total nécessaire à la conception d'un nouveau médicament varie entre 160 et 1800 millions de dollars (dollars de 2009).[6] Ainsi, s'il est difficile d'en obtenir une estimation fiable, le coût important de ce processus est établi. Afin d'accélérer le développement de nouveaux médicaments et de réduire les coûts associés à chaque étape, les différentes phases du processus (***Figure 1***) doivent donc être optimisées. L'usage des méthodes bioinformatiques et chémoinformatiques est ainsi devenu un standard, celles-ci étant peu coûteuses et devenant de plus en plus efficaces.[8,9]

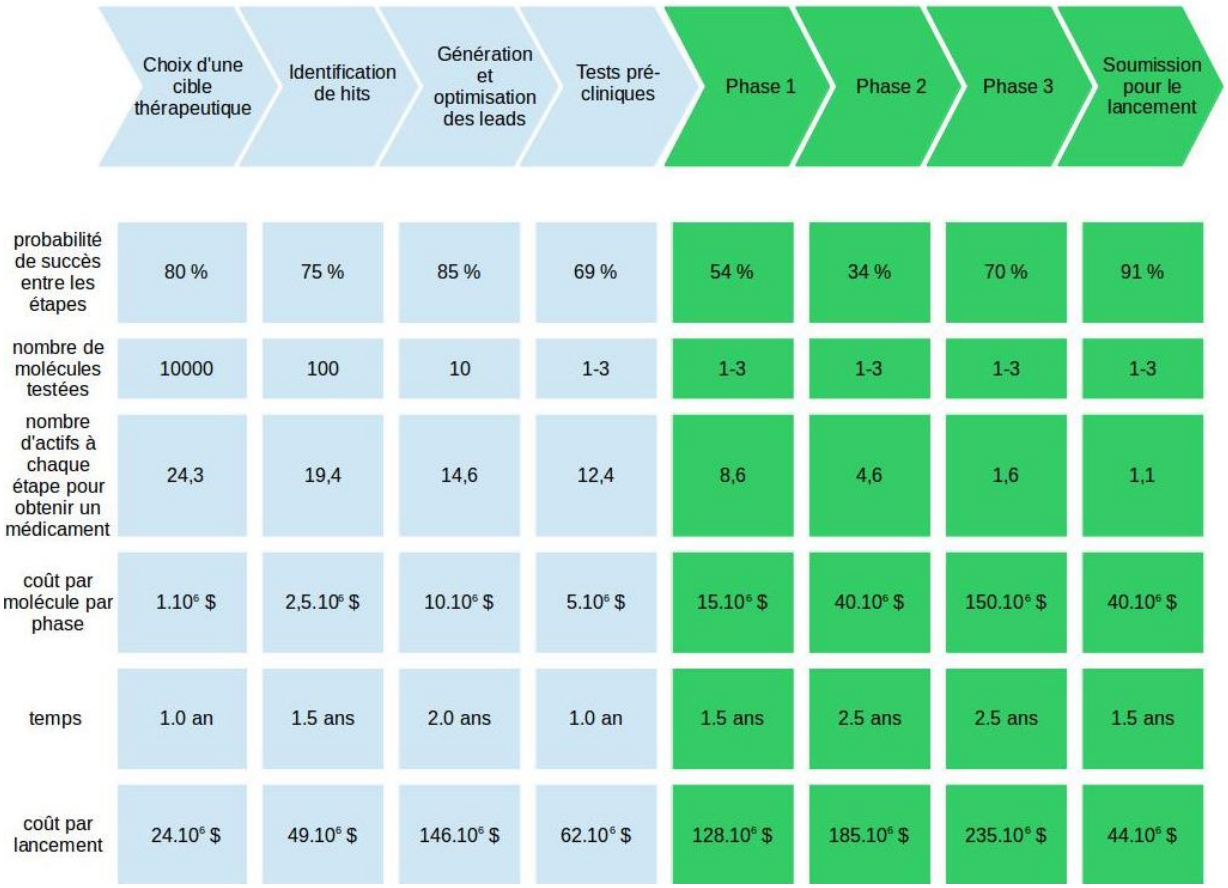


Figure 1. *Représentation des statistiques de succès, durées et coûts financiers des phases successives du processus de recherche de nouveaux médicaments (dollars de 2010). (Bleu) Phases de recherche et pré-cliniques. (Vert) Phases d'essai clinique.*[7]

1.2.1. Sélection d'une cible thérapeutique

Le processus de R&D s'amorce suite à l'émergence d'une maladie dans les populations. Les pathologies les plus répandues, touchant une large partie des populations, sont largement étudiées puisqu'elles représentent un besoin important en termes de réponse thérapeutique, associé également à un fort potentiel commercial. Les maladies neurodégénératives, le Syndrome d'ImmunoDéficience Acquise (SIDA) et les différents types de cancers font l'objet d'un effort de recherche important, largement financé par les gouvernements et l'industrie pharmaceutique. Inversement, les maladies dites « orphelines » touchent une faible partie des populations et représentent un besoin thérapeutique moindre, associé à un faible potentiel commercial, et tendent à être délaissées. Cependant, une pathologie pour laquelle il n'existe aucune réponse thérapeutique représente, économiquement, l'accès à un monopole de marché lorsque le processus de recherche aboutit à un traitement.

Une fois le besoin thérapeutique défini, l'objectif est d'identifier une cible biologique pertinente dont on pourra moduler l'activité afin d'enrayer le processus pathologique. Les cibles thérapeutiques incluent principalement les protéines et plus rarement les gènes et les ARNs.[10] D'après Overington et al.,[10] la pharmacopée actuelle comprend plus de 21.000 produits médicaux. Cependant, lorsque l'on ignore les différentes formulations, formes salines, suppléments, adjuvants et vitamines, le nombre de composés actifs uniques est réduit à 1357, dont 1204 sont de type « petite molécule » et 153 sont de type « biologique ».[10] Les médicaments de type « petite molécule » sont des composés chimiques ou des peptides, tandis que les médicaments de type « biologique » regroupent les anticorps, vaccins et interleukines. D'après cette même étude, le nombre de cibles thérapeutiques uniques de notre pharmacopée serait de 324, en ne considérant qu'un nombre très restreint de cibles principales par médicament (266 seraient des protéines et gènes humains, les 58 autres regroupent des cibles de pathogènes bactériens, viraux ou fongiques). Parmi les 1204 composés actifs uniques de type « petite molécule », 1065 devraient leur activité à une interaction avec une protéine.[10] L'arsenal thérapeutique actuel serait donc constitué à 78% de composés actifs de type « petite molécule » ayant pour cible une protéine. Au cours de cette thèse, nous nous sommes concentrés majoritairement sur ce dernier cas.

Trois critères sont capitaux dans la définition du potentiel d'une protéine à devenir une cible thérapeutique : l'efficacité dans l'altération du processus pathologique lorsque l'on agit sur celle-ci, la sécurité du patient (l'altération du processus pathologique ne doit pas induire d'effets secondaires importants) et la « druggabilité » de la cible.[5] Une cible est définie comme « druggable » lorsqu'il est possible de moduler son activité avec d'autres partenaires biologiques, en y liant de petites molécules ou d'autres agents à vocation thérapeutique.[11] Ainsi, dans le cas des cancers, le paradigme thérapeutique est l'altération du processus de division cellulaire afin de prévenir la prolifération des cellules cancéreuses et la formation de tumeurs. Sur ce même exemple, la sécurité serait définie par le succès dans l'altération de la division cellulaire des cellules cancéreuses, sans affecter la division des cellules saines.

Différentes approches permettent d'identifier une cible thérapeutique. Le processus pathologique comporte généralement plusieurs cascades d'évènements, impliquant chacune plusieurs protéines et autres agents biologiques.[10,12] L'objectif est donc de comprendre ces cascades de réactions, de la manière la plus complète possible, afin d'identifier des cibles thérapeutiques potentielles suivant les trois principes précédemment évoqués.

La recherche bibliographique et la compilation des résultats scientifiques peuvent aboutir à l'identification de processus pathologiques et de cibles potentielles. L'exploration de la littérature, grâce aux approches bioinformatiques, permet le tri et l'exploitation de différentes sources de données par « text-mining » (publications, brevets) ou « data-mining » (données d'expression génétiques et protéomiques).[12,13] L'étude de l'expression des ARNm ou des protéines permet également l'identification de cibles. Dans ces cas, des cohortes de patients peuvent être constituées. Il est possible, par exemple, de constituer un premier groupe de patients atteints par la pathologie étudiée afin de le comparer à un second groupe de sujets sains.[14] Il est également possible d'établir des échelles de gravité, vitesse de progression, ou résistance à une pathologie, conduisant ainsi à la définition d'un plus grand nombre de groupes de patients et permettant la mise en œuvre de différentes méthodes statistiques.[15–17] L'étude quantitative de l'expression des entités biologiques permettra de les corréler, ou non, à l'évolution de la pathologie étudiée. Si une protéine est identifiée comme surexprimée dans un groupe de patients atteints, par rapport au groupe contrôle, l'objectif pourra être de rétablir un niveau d'expression normal de cette protéine en agissant sur son processus de synthèse.

Plus récemment, les études d'association entre les polymorphismes génétiques et le développement de pathologies de type GWAS (Genome Wide Association Studies)[18] et eQTL (expression Quantitative Trait Locus)[19] ont permis d'explorer de nouveaux processus pathologiques et d'identifier de nouvelles cibles thérapeutiques.[5,13,19] L'importance des polymorphismes génétiques (Single Nucleotide Polymorphism ou SNP) dans le développement de pathologies peut être illustrée par le cas extrême de l'anémie falciforme, une pathologie entraînant des troubles de la circulation sanguine associés à un fort taux de mortalité.[20] L'anémie falciforme se caractérise par une forme et une rigidité anormales des érythrocytes dues à une forme mutée de l'hémoglobine impliquant un unique SNP.[21] Ce SNP induit la substitution d'une Glutamine en Valine en 6^{e} position de la séquence de la β-globine et provoque la polymérisation de l'hémoglobine mutée, déformant ainsi les érythrocytes (***Figure 2***).[21]

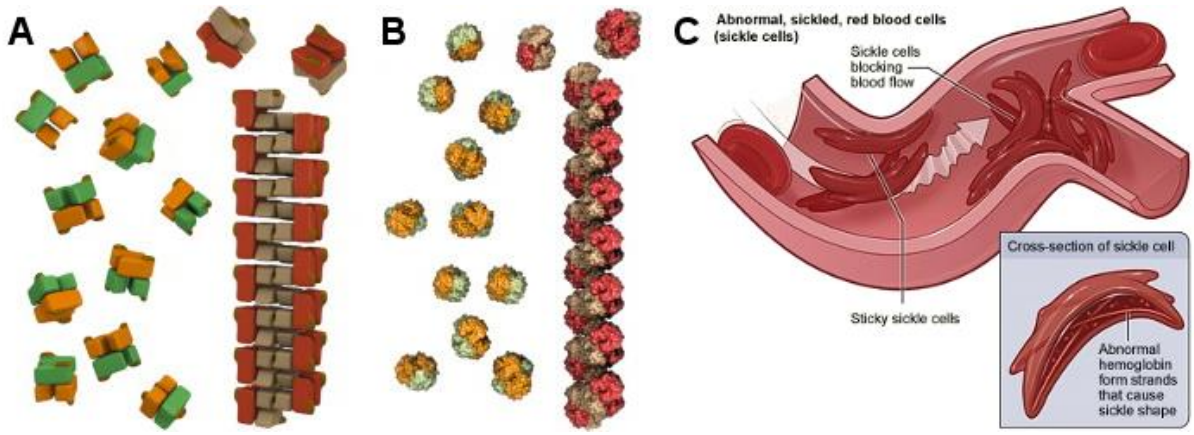


Figure 2. *Une unique mutation est nécessaire pour provoquer le phénomène d'agrégation observé dans l'anémie falciforme. (A) Représentation schématique des homomères d'hémoglobine mutée et de leurs agrégats en filaments. Les couleurs différencient α- et β-globines. (B) Représentation des structures. (C) Illustration de l'obstruction des vaisseaux sanguins par les érythrocytes déformés.*[22]

Après avoir sélectionné une cible thérapeutique potentielle, il convient de procéder à sa validation pour s'assurer d'obtenir des effets bénéfiques lors de sa modulation. L'étape de validation de la cible peut s'effectuer par de nombreux outils *in vitro* et *in vivo* et être assistée par la bioinformatique.[14] Par exemple, lorsque la protéine ciblée existe également chez la souris avec une bonne homologie, il sera possible de générer des animaux transgéniques pour lesquels le gène associé à la protéine cible est absent (Knock-Out d'un gène ou KO).[23,24] On pourra alors observer les conséquences d'une modulation de l'activité de la cible, sur modèle animal, dans le cas extrême de son absence totale d'activité. Les animaux transgéniques permettent également d'observer les conséquences d'une surexpression de la protéine cible.[25] Si une souris KO présente des signes d'amélioration dans l'évolution de la pathologie étudiée et que les effets délétères sont considérés comme acceptables, ou absents, la cible pourra être considérée comme valide.

Plusieurs études GWAS[18] et eQTL[19] conduites dans notre laboratoire ont permis d'identifier des protéines dont l'expression favorise la non-progression[15,17] ou la progression rapide vers le SIDA.[16] Au cours de cette thèse, nous avons exploré ces résultats à la recherche de cibles « druggables » et de « hits » potentiels (voir partie Résultats).

1.2.2. Identification de « hits »

Après avoir identifié et choisi la cible thérapeutique, l'objectif est de moduler son activité afin d'altérer le processus pathologique. On appelle touches ou « hits » les composés capables de

tels effets. L'approche la plus couramment utilisée dans la recherche de « hits » est le criblage (ou « screening ») de banques de petites molécules (ou chimiothèques). Un criblage peut être réalisé *in vitro* grâce aux techniques HTS permettant d'estimer l'affinité et activité de milliers de composés sur la cible d'intérêt. Cependant, la mise en œuvre de HTS n'est pas toujours possible du fait de leurs coûts importants et du temps nécessaire à leur exécution. Les criblages de type HTS et la validation pré-clinique des caractéristiques pharmacologiques des composés représenteraient 14% du budget de R&D de l'industrie pharmaceutique.[26]

Les techniques de criblage *in silico* (VHTS), analogues virtuels du HTS, tendent à réduire ces problèmes par trois avantages principaux : une relative facilité dans leur mise en œuvre, un coût très réduit et l'obtention assez rapide des résultats, en fonction des moyens de calcul disponibles.[27,28] Généralement utilisé en amont du HTS, le criblage virtuel permet un premier filtrage des chimiothèques en réduisant le nombre de composés à tester expérimentalement et, supposément, en optimisant le potentiel des programmes HTS à découvrir des « hits » réels.[8,29] Cependant, les résultats de VHTS restent des prédictions dont la fiabilité peut être très variable en fonction de la cible thérapeutique choisie, des caractéristiques des chimiothèques, des méthodes VHTS choisies et de nombreux autres facteurs.[8,28,30–32] Au cours de cette thèse, nous nous sommes particulièrement attachés à l'évaluation des méthodes de criblage virtuelles, notamment « structure-based » (voir partie Résultats).

Les « hits » validés par HTS sont sélectionnés principalement en fonction de leur affinité avec la cible. Toutefois, il est également important de sélectionner des composés variés, aux caractéristiques différentes (poids moléculaire, nombre de groupements donneurs ou accepteurs de liaisons hydrogènes, diversité des châssis moléculaires, etc.), puisqu'ils ne constituent à cette étape que des pistes d'étude et rentreront ensuite dans une phase d'optimisation.[5,33,34]

1.2.3. Recherche et optimisation de « leads »

A partir des « hits » précédemment identifiés, l'étape d'optimisation a pour objectif d'obtenir des composés plus sélectifs, plus actifs et présentant des propriétés pharmacocinétiques optimales.[5,35] Cette étape fait intervenir, en étroite collaboration, des chimistes médicinaux et des chémoinformaticiens afin de moduler la structure chimique des « hits » et aboutit à l'obtention de composés qualifiés de « têtes de séries » ou « leads ». Il est possible de procéder à l'étude des relations structure-activité en modulant les structures chimiques des « hits » par l'ajout ou le retrait de groupements fonctionnels, tout en conservant leurs châssis moléculaires.[9]

Les composés nouvellement synthétisés seront comparés aux « hits » initiaux et l'on évoluera, par étapes successives, vers la sélection de composés plus actifs, plus sélectifs et aux propriétés pharmacocinétiques désirables dans la sélection d'un candidat médicament (composés dits « drug-like »). Cette approche, bien que très efficace, reste fastidieuse du fait des nombreuses synthèses chimiques qu'elle requiert. La chémoinformatique pourra être employée pour orienter les synthèses chimiques grâce à des modèles quantifiant les relations structure-activité (Quantitative Structure-Activity Relationship ou QSAR), qui permettront également de tirer profit de la connaissance des structures des cibles lorsque celles-ci sont disponibles (Receptor Dependant-QSAR ou RD-QSAR).[36,37]

Durant la phase d'optimisation des « leads », il convient également de s'assurer que l'activité observée est toujours due à une interaction avec la cible thérapeutique.[37,38] La propriété intellectuelle est également à prendre en compte dans la diversification des « hits » et la définition des composés « leads », il conviendra alors de restreindre l'espace de recherche hors des brevets existants grâce aux approches chémoinformatiques et de « text-mining ».[5,9]

1.2.4. Essais pré-cliniques

Après avoir obtenu des « leads » optimisés, les essais pré-cliniques consistent en de nombreuse études dont l'objectif est de qualifier le candidat médicament sur les plans pharmacologique, pharmacocinétique et toxicologique. L'usage d'expérimentations animales rationnelles permettra d'envisager l'administration du candidat médicament chez l'homme au cours des essais cliniques suivants. Les études pharmacologiques ont pour objectif de valider le mécanisme d'action du candidat médicament et de mesurer son activité dans des modèles expérimentaux de la pathologie, *in vitro* et *in vivo* chez l'animal. Les études pharmacocinétiques permettent de décrire le devenir d'un composé, sa distribution, son absorption et son métabolisme, puis son élimination par l'organisme. Les études toxicologiques servent à déterminer les doses toxiques du candidat médicament sur l'organisme étudié (principalement des souris ou rats, plus rarement des chats, chiens, porcs, ou primates).[5,39]

Ces données permettront de déterminer les doses à administrer à l'homme lors des essais cliniques et constituent une première approche dans l'étude de potentiels effets indésirables du candidat médicament, permettant de suivre ces effets de manière proactive. Une partie annexe du développement pré-clinique consiste également en l'évaluation du risque environnemental lié à la mise sur le marché du candidat médicament. Toutes les informations recueillies lors des

essais pré-cliniques seront compilées dans un dossier de demande d'autorisation de mise sur le marché du candidat médicament. Celui-ci sera étroitement étudié par les autorités de santé compétentes (en France, il s'agit de l'Agence Nationale de Sécurité du Médicament (ANSM) et du Comité de Protection des Personnes (CPP)), avant d'autoriser, ou non, l'entrée du candidat médicament en phase d'essais cliniques.[5,39]

1.2.5. Essais cliniques

Les essais cliniques représentent l'étape la plus critique du processus de conception de médicaments, validant ou non plusieurs années de recherche pré-clinique et extrêmement onéreuse. Ils se divisent en quatre phases. Les trois premières permettent d'établir l'efficacité et la sécurité du candidat médicament afin d'obtenir une AMM, tandis que la quatrième consiste en la surveillance des effets secondaires durant toute la durée de commercialisation et d'utilisation du médicament (phase de pharmacovigilance).[40] L'encadrement des essais cliniques par les autorités de santé est très strict. En France, le consentement éclairé des volontaires est requis et un registre national tenu par l'ANSM recense tous les sujets, le montant de leurs indemnités (plafonnées à 4500 euros par an pour éviter d'éventuelles dérives), la date et la durée de leurs protocoles. Au cours de chaque phase, si des effets indésirables sont détectés, l'étude clinique peut prendre fin et le candidat médicament peut être abandonné de manière définitive.

La phase I est préliminaire à l'étude de l'efficacité d'un candidat médicament. Elle se déroule sur un faible nombre de volontaires sains (20 à 80) et a pour seul objectif d'évaluer la tolérance ou l'absence d'éventuels effets secondaires liés à l'administration du candidat médicament. Ces essais peuvent toutefois être proposés à des patients en échec thérapeutique, pour lesquels le traitement étudié représente la seule chance de survie. Environ 54% des composés testés en phase I accèderont à la phase suivante.[7] La phase II vise à déterminer la posologie optimale, l'efficacité du candidat médicament à ce dosage et sa tolérance à court terme. Elle est généralement réalisée sur un groupe homogène de 100 à 200 malades. Seuls 18% des essais cliniques démarrés accèdent à la phase III.[7] Ces essais, de plus grande envergure, sont conduits sur plusieurs milliers de patients représentatifs de la population de malades à laquelle le traitement est destiné. Il s'agit d'essais comparatifs au cours desquels le médicament en développement est comparé soit à un traitement efficace déjà commercialisé, soit à un placebo. Les essais de phase III sont le plus souvent réalisés en double aveugle et avec tirage au sort : les traitements ou placebo sont attribués de manière aléatoire aux patients et aux médecins

chargés du suivi, sans qu'ils ne soient informés de quelle attribution ils ont fait l'objet. Cette méthode permet d'éviter les biais liés au processus de prise en charge du patient, connus communément comme « l'effet placebo ». Le processus d'essai clinique complet donne lieu à l'obtention d'une AMM pour 11% des candidats médicaments.[7,41]

Lorsque l'AMM est accordée, la commercialisation et l'application du traitement peut commencer. La phase IV des essais cliniques, aussi appelée pharmacovigilance, consiste alors à suivre les effets secondaires potentiels du traitement sur l'ensemble des patients qui en bénéficient (population large et hétérogène). Dans ce cadre, les médecins ont l'obligation de reporter les effets indésirables décrits par leurs patients, ce qui permet d'identifier rapidement l'émergence de nouveaux effets qui n'auraient pas été détectés lors des essais cliniques et garanti une plus grande sécurité d'utilisation aux patients.

1.3. Limitations et échecs

La recherche pharmaceutique fait donc face à défis majeurs, principalement scientifiques, mais aussi financiers et politiques. Le taux de succès des phases d'essai clinique est d'environ 11%, tous types de pathologies confondues.[7,41] Celui-ci varie en fonction des types de pathologies, d'environ 5% pour celles impliquant le système nerveux central à environ 20% pour les maladies cardiovasculaires (***Figure 3***).[41]

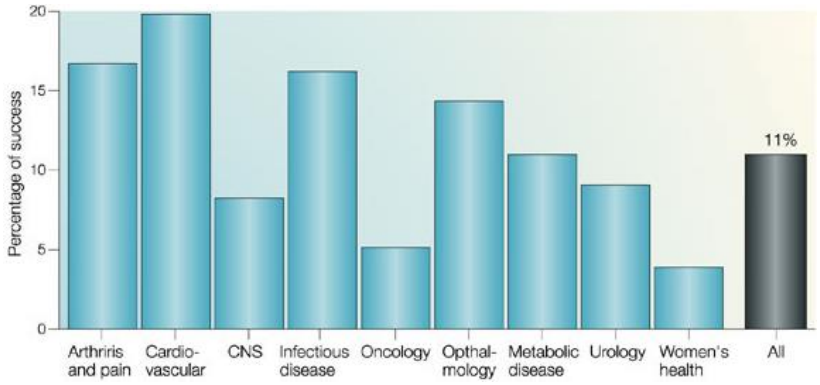


Figure 3. *Taux de succès des essais cliniques de la phase I à l'obtention d'une AMM pour les dix plus grandes compagnies pharmaceutiques entre 1991 et 2000. Le taux de succès varie en fonction des pathologies ciblées.*[41]

Ces faibles taux de succès peuvent s'expliquer de plusieurs manières : la recherche pharmaceutique se concentre actuellement sur des pathologies d'une grande complexité (cancers, SIDA, etc.), la compétition entre les différentes compagnies pharmaceutiques augmente parallèlement aux standards de soins et les autorités de santé deviennent plus exigeantes. En 1991, les mauvaises caractéristiques pharmacocinétiques des candidats

médicaments représentaient la première cause d'échec des essais cliniques (40%). En 2000, les problèmes de pharmacocinétique ne représentaient plus que 10% des échecs en essai clinique, principalement grâce à l'application des filtres ADME-Tox (voir point 2.2.2.2). Aujourd'hui, les premières causes d'échec lors des phases d'essai clinique sont le manque d'efficacité des composés (30%) et leur toxicité (30%).[41] Les échecs liés au manque d'efficacité des composés sont plus fréquents lorsque les modèles animaux utilisés sont peu prédictifs, notamment pour les pathologies du système nerveux central et les cancers, induisant de forts taux d'échec en phases II et III (***Figure 4***).[42]

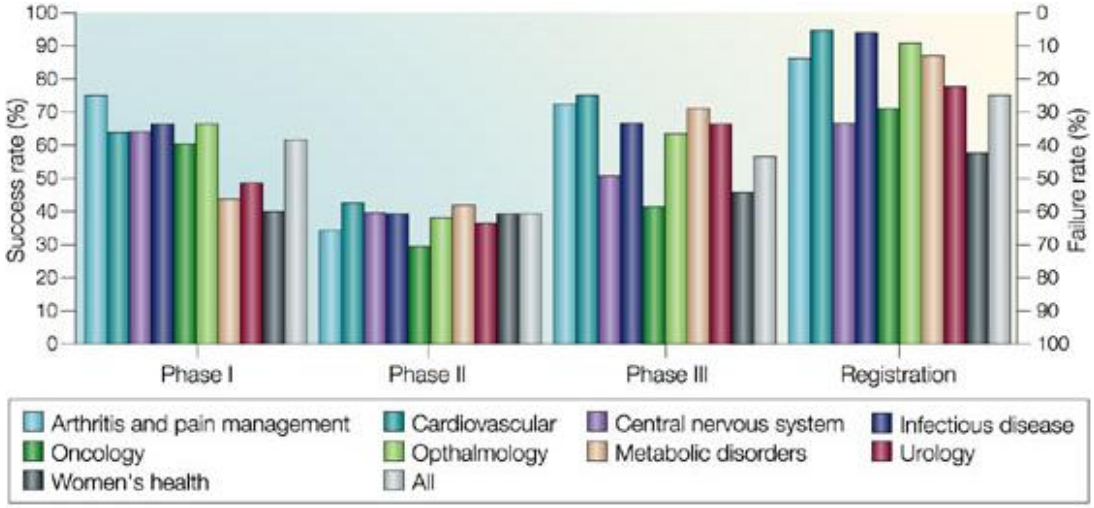


Figure 4. *Taux de succès par type de pathologie et suivant les différentes phases des essais cliniques pour les dix plus grandes compagnies pharmaceutiques entre 1991 et 2000.*[41]

2. Méthodes de criblage virtuel

2.1. Objectifs du criblage virtuel

Ces 20 dernières années, le criblage virtuel a pris une importance croissante dans les processus de conception de médicaments, largement appuyé par les progrès en informatique et bioinformatique (architecture et logiciels).[8,9] Analogue *in silico* du HTS, le criblage virtuel a pour objectif de filtrer de larges chimiothèques de composés (10^5 à 10^7) afin de procéder à des HTS optimaux, comportant moins de composés (10^0 à 10^4) et supposément plus aptes à générer des « hits ». Les protocoles établis permettent d'éliminer les composés supposés inactifs ou indésirables du fait de leurs propriétés pharmacologiques et sélectionneront les composés supposés actifs en leur attribuant de hauts scores d'affinité à la cible.[5,8]

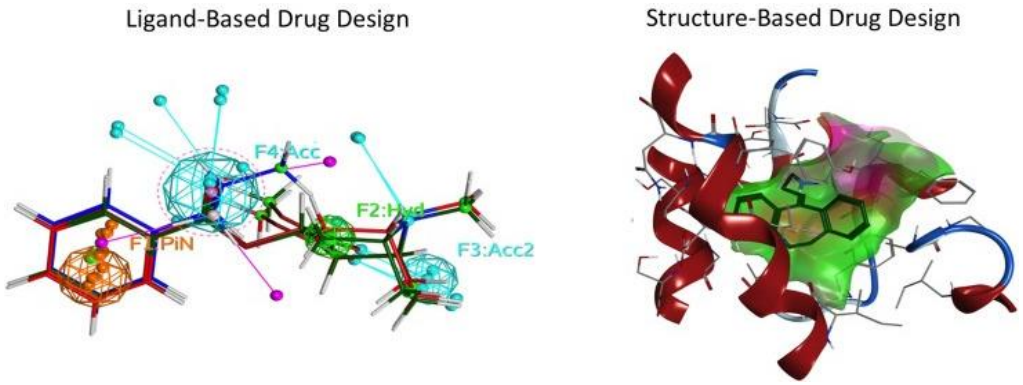


Figure 5. *Illustration schématique des approches « ligand-based » et « structure-based ». (Gauche) Représentation d'un pharmacophore « ligand-based ». Les sphères indiquent les caractéristiques d'interaction des ligands (bleu : accepteurs de liaison hydrogène, vert : point hydrophobe, orange : interaction π-π). (Droite) Représentation d'un composé amarré dans son site de liaison (vert) par des méthodes « structure-based ».*[43]

Les méthodes de criblage sont séparées en deux grandes familles : « ligand-based » et « structure-based », en fonction du type de données sur lesquelles elles reposent (***Figure 5***).[44] Lorsqu'une structure 3D de la protéine ciblée est disponible, les méthodes « structure-based » permettent d'évaluer le potentiel d'interaction entre les composés criblés et le site d'interaction sélectionné sur la structure. Elles comprennent les méthodes de docking, les approches pharmacophoriques, les QSAR « structure-based » (RD-QSAR) et le design *de novo*.[33,45,46] Les approches « ligand-based » tirent profit de la connaissance d'un ou plusieurs ligands interagissant avec la cible et des données d'affinité déterminées expérimentalement lors d'études précédentes. Il s'agira alors d'analyser les relations structure-activité de ces ligands

afin de guider la recherche et la synthèse de nouveaux composés potentiellement actifs. Les approches « ligand-based » peuvent être : la recherche de similarité, les approches pharmacophoriques et les QSAR « ligand-based ».[9,33,36] Il sera également possible de combiner les approches « ligand-based » et « structure-based » lorsque les deux types de données sont disponibles.

2.2. Utilisation des chimiothèques

Afin de procéder aux criblages, il convient en premier lieu de sélectionner une chimiothèque de composés. Le succès des différents types de criblages est directement conditionné par le choix de la chimiothèque utilisée, sa taille, composition, diversité, les caractéristiques des molécules qu'elle regroupe et de nombreux autres facteurs. Suivant les architectures informatiques et les logiciels utilisés, les temps de calcul nécessaires à la mise en œuvre des méthodes de criblage restreignent actuellement le nombre de composés criblés. L'espace chimique exploré par ces méthodes dépasse donc rarement un million de composés.[33]

2.2.1. Spécificités des chimiothèques

La première spécificité des chimiothèques est l'espace chimique dans lequel elles s'inscrivent. L'espace chimique réel, regroupant l'ensemble des composés uniques découverts et synthétisés par le passé, est identifié par le registre mondial CAS (Chemical Abstracts Service) qui comporte aujourd'hui plus de 66 millions de composés.[47] Inversement, l'espace chimique théorique est défini comme l'ensemble des composés potentiellement synthétisables et stables selon les règles de chimie établies. Le projet Chemical Universe Database explore cet espace par génération *in silico* systématique de tels composés.[48–51] Suivant Lorenz et al., l'espace chimique théorique des composés comportant 13 atomes ou moins (hors hydrogènes et parmi les suivants : C, N, O, S, ou Cl) représenterait plus de 977 millions de composés (GDB-13).[49]

Les chimiothèques utilisées pour les criblages virtuels sont principalement définies dans l'espace chimique réel, pour permettre l'évaluation expérimentale des « hits » obtenus *in silico*. Ces chimiothèques, principalement commerciales, sont mises à disposition en format digital et l'ensemble des composés est directement commandable au fournisseur, conditionné en microplaques (***Tableau 1***).[33] Cependant, devant les difficultés croissantes dans l'identification de nouvelles petites molécules thérapeutiques, l'exploration de l'espace chimique théorique tend à se développer (***Figure 6***).[51–54]

Chimiothèque	Disponibilité	Nombre de composés	Données expérimentales	Accès
DrugBank	Publique	6.825	oui	http://www.drugbank.ca
Maybridge	Commerciale	53.000	non	http://www.maybridge.com
Chimiothèque nationale	Commerciale	53.430	non	http://chimiotheque-nationale.enscm.fr
MDDR	Commerciale	150.000	oui	http://accelrys.com/products/databases/bioactivity/mddr.html
NCI Open database	Publique	250.250	non	http://cactus.nci.nih.gov/ncidb2.2
WOMBAT	Commerciale	270.918	oui	http://www.sunsetmolecular.com
Drug Discovery Center Collection	Commerciale	340.000	non	http://drugdiscovery.uc.edu
ChemBridge	Commerciale	950.000	non	http://www.chembridge.com
ChemBank	Publique	1.200.000	oui	http://chembank.broadinstitute.org
ChEMBL	Publique	1.324.941	oui	https://www.ebi.ac.uk/chembl
ChemDiv	Commerciale	1.500.000	non	http://eu.chemdiv.com
Enamine	Commerciale	1.800.000	non	http://www.enamine.net
ChemDB	Publique	5.000.000	non	http://cdb.ics.uci.edu
Cococo	Publique	6.981.556	non	http://cococo.unibo.it/cococo
eMolecules	Commerciale	6.000.000	non	http://www.emolecules.com
ACD	Commerciale	7.000.000	non	http://accelrys.com/products/databases/sourcing/available-chemicals-directory.html
ZINC	Publique	21.000.000	oui	http://zinc.docking.org
ChemSpider	Publique	29.000.000	oui	http://www.chemspider.com
PubChem	Publique	30.000.000	oui	http://pubchem.ncbi.nlm.nih.gov

Tableau 1. *Principales chimiothèques dans l'espace réel. Les types de données expérimentales fournies peuvent varier (K_i, IC_{50}, pIC_{50}, EC_{50}, ou autres) et celles-ci ne sont pas toujours disponibles.*[33]

D'autres types de chimiothèques, dites focalisées, sont dédiées à la mise en œuvre de criblages sur une famille de cibles spécifique. La connaissance de ligands d'une famille de cibles spécifique permet de constituer des chimiothèques restreintes, grâce à la sélection de composés aux caractéristiques physico-chimiques similaires. Plusieurs chimiothèques commerciales sont disponibles et/ou déclinées dans des versions focalisées, particulièrement pour le ciblage des GPCR, kinases, récepteurs nucléaires, ou canaux ioniques (ChemBridge,[55] Asinex,[56] Life Chemicals,[57] Timtec,[58] Maybridge[59]).

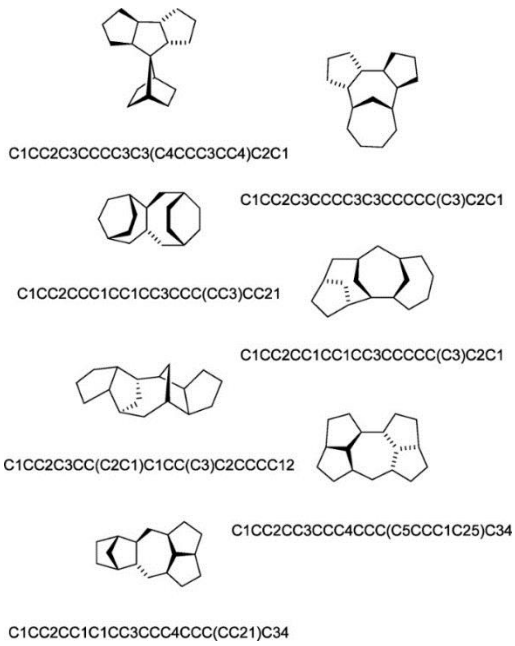


Figure 6. *Structures de 7 systèmes cycliques aujourd'hui inconnus identifiés par le projet Chemical Universe Database GDB-17, avec leurs codes SMILES. Ces systèmes sont introuvables dans les bases de données CAS et SciFinder, également lorsque la stéréochimie et les types d'atomes sont ignorés. Les systèmes sont représentés ici par l'un de leurs stéréoisomères.*[51]

Inversement, les chimiothèques de fragments sont développées dans l'objectif d'augmenter la diversité de l'espace chimique exploré.[60–62] Ainsi, par assemblage et combinaison des différents fragments moléculaires (100 à 250 Da) dans le site de liaison de la cible choisie, il est possible d'obtenir des composés variés, généralement de plus faible poids moléculaire et de plus forte affinité en comparaison aux « hits » obtenus par HTS (voir point 2.4.5). Cette technique présente de bons taux de succès dans l'identification de « hits » (3-5%), supérieurs à ceux obtenus par les HTS (1%).[60] Cependant, les « hits » identifiés par assemblage de fragments sont rarement des molécules disponibles auprès des fournisseurs. L'étude expérimentale des « hits » ainsi obtenus requière presque systématiquement la mise en œuvre de synthèses chimiques.[60,62,63]

Enfin, il existe des chimiothèques de produits naturels, également mises à disposition par les principaux fournisseurs.[33] Suite à l'émergence des nouvelles techniques de criblage, la recherche de médicaments par l'étude de produits naturels tend à être délaissée. Cependant, ces produits ont généralement une sélectivité importante pour leurs cibles, présentent de bonnes caractéristiques ADME-Tox (voir point 2.2.2.2) et leurs structures diffèrent de celles des produits de synthèse.[64,65] L'utilisation des chimiothèques de produits naturels reste donc pertinente.

2.2.2. Préparation d'une chimiothèque

Après avoir sélectionné une chimiothèque et avant de procéder aux criblages, il convient de préparer les molécules dans le format adapté aux méthodes mises en œuvre, à partir des données brutes informatisées. Différents types de filtres pourront également être appliqués afin de réduire la taille de la chimiothèque et d'optimiser les résultats.

2.2.2.1. Etats d'ionisation, mésomérie et tautomérie

Selon les conditions physiologiques, une molécule peut subir des phénomènes de tautomérie, de mésomérie ou d'ionisation. La tautomérie se caractérise par un transfert intramoléculaire de proton, la mésomérie par une délocalisation électronique dans le cas de molécules conjuguées et l'ionisation par le gain ou la perte d'atome(s) d'hydrogène (***Figure 7***).

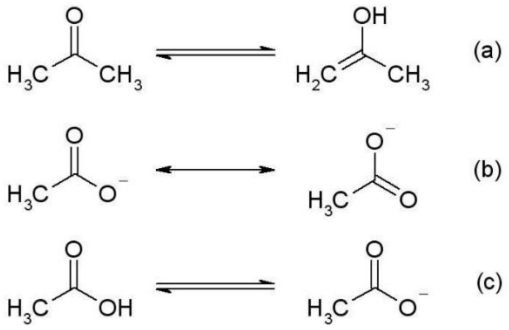


Figure 7. *Exemple de formes chimiques tautomères (a), mésomère (b) ou états d'ionisation différents (c).*

Il est important de considérer ces différentes formes lors de la préparation des chimiothèques. Dans chacun des cas, les modifications de la molécule conduisent à un nouvel arrangement des groupements fonctionnels, impactant directement l'établissement de liaisons avec la protéine cible. Généralement, seule la forme la plus stable sera conservée dans les chimiothèques dédiées au criblage virtuel. Dans certains cas, on peut caractériser cette forme sans difficulté (par exemple, l'acide carboxylique qui n'existe que sous sa forme anionique à pH neutre). Dans d'autres cas, la forme chimique peut s'avérer difficile à identifier, notamment en ce qui concerne l'état d'ionisation lorsque l'un des pK_a de la molécule est proche du pH physiologique. Il est important de noter que 30% des composés issus des bases de données commerciales sont potentiellement tautomériques et 10% des composés ne sont pas représentés dans leur forme aqueuse la plus stable.[66] Plusieurs outils permettent d'identifier les formes tautomères (Ambit-Tautomer,[67] ChemAxon Marvin,[68] SPARC[69]), mésomères ou les états d'ionisation (LigPrep[70], Epik[71]) les plus stables. Lorsqu'il subsiste une incertitude sur la forme chimique optimale, il convient d'inclure l'ensemble des états pertinents dans la chimiothèque.[72]

2.2.2.2. Filtrage ADME-Tox

Dans les années 1990, les échecs dans le développement de médicaments étaient principalement la conséquence de mauvaises performances pharmacocinétiques. Ce problème a été en grande partie résolu grâce à l'utilisation d'un filtrage précédant l'utilisation des chimiothèques, réduisant les taux d'échec dans les phases de développement.[41] Les filtres de type ADME-Tox (Absorption, Distribution, Métabolisme, Elimination et Toxicité) sont rapidement devenus populaires. Ils reposent sur plusieurs critères déterminant les propriétés pharmacocinétiques potentielles des molécules et sont désormais largement utilisés pour réduire le nombre de composés d'une chimiothèque en sélectionnant les plus aptes à devenir des candidats médicaments, avant tout processus de criblage. L'utilisation de ces filtres a donné de très bons résultats, réduisant le taux d'échec imputable à une cause pharmacocinétique de 40% à moins de 10% en 10 ans.[41] Désormais, la toxicité et le manque d'efficacité des candidats médicaments sont les deux plus grandes causes d'échecs dans le développement d'un médicament.[41]

2.2.2.1.1. Règles de Lipinski

Ainsi, Lipinski et al. définirent dès 1997 des règles simples permettant d'identifier rapidement et à grande échelle des molécules à caractère « drug-like », plus susceptibles de présenter les caractéristiques de biodisponibilité nécessaires au développement d'un candidat médicament.[73] Ces règles, communément appelées « règles de Lipinski » ou « la règle de 5 », comportent quatre critères physico-chimiques qui décrivent la molécule : poids moléculaire $\leq$ 500 Da, logP $\leq$ 5, nombre d'accepteurs de liaisons hydrogènes $\leq$ 10 et nombre de donneurs de liaisons hydrogènes $\leq$ 5 (***Tableau 2***). Particulièrement, la mesure du logP caractérise la polarité du composé (estimée par le coefficient de partition octanol/eau), permettant ainsi d'estimer la distribution du composé dans l'organisme. Les molécules hydrophobes (hautes valeurs du logP) sont principalement distribuées dans les régions hydrophobes, comme la bicouche lipidique des cellules. Inversement, les molécules hydrophiles sont retrouvées principalement dans des régions aqueuses, comme le sérum sanguin. La 5^{e} règle de Lipinski stipule que les adjuvants et assimilés font exception aux quatre autres règles.

D'après ces règles, déterminées à partir de 2245 molécules extraites du World Drug Index (WDI) et ayant passé avec succès la phase II des tests cliniques, les composés ne validant pas au moins deux des critères suivants auraient de très fortes chances d'avoir des problèmes d'absorption ou de perméabilité intestinale.[73] Une mauvaise biodisponibilité orale serait également détectée dès lors qu'une de ces règles est violée. Néanmoins, il n'est pas garanti que

les composés adhérant à cette règle possèdent une excellente absorption, perméabilité et/ou biodisponibilité. La morphine, par exemple, satisfait toutes les règles de Lipinski mais présente une biodisponibilité orale modérée (***Figure 8***). De plus, un médicament qui enfreint une ou plusieurs règles peut tout de même avoir une biodisponibilité satisfaisante.

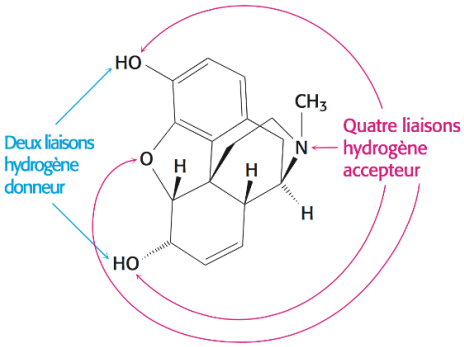


Figure 8. *Structure moléculaire de la morphine avec ses sites d'interaction donneurs et accepteurs d'hydrogène.*

Les filtres de ce type guident aujourd'hui l'évaluation de nouveaux candidats médicaments de manière efficace. Cependant, ils sont régulièrement contestés et remis en cause puisque lorsqu'ils sont appliqués trop fermement, ceux-ci restreignent de manière non négligeable l'espace chimique des composés évalués lors de criblages virtuels.

2.2.2.1.2. Autres critères de sélection ADME-Tox

Par la suite, d'autres critères ont été mis en place pour compléter et ajuster les règles de Lipinski dans la sélection de composés « drug-like ». Ainsi, Veber et al. choisissent d'utiliser les quatre critères suivants : nombre de liaisons rotatives, aire de la surface polaire et nombre de donneurs et d'accepteurs de liaisons hydrogènes (***Tableau 2***).[74] Ces critères ont été établis par l'étude de la biodisponibilité orale chez le rat de 1100 candidats médicaments (données de GlaxoSmithKline). D'autres approches ont été envisagées pour la sélection ADME-Tox, notamment basées sur les types d'atomes et de liaisons,[75] ainsi que sur le comptage de points pharmacophoriques spécifiques.[76] Les méthodes d'apprentissage ont également été explorées pour l'identification de composés « drug-like », notamment les machines à vecteurs de support[77,78] (Support Vector Machines ou SVM) et les réseaux de neurones,[77,79,80] avec des résultats comparables aux méthodes de Lipinski et Veber. Cependant, ces dernières fonctionnent comme des « boites noires » et ne permettent pas de mettre en évidence de manière claire l'importance des différents descripteurs utilisés dans la caractérisation des composés.

2.2.2.1.3. Critères de sélection « lead-like »

Après avoir identifié un nombre satisfaisant de composés « drug-like » suite aux criblages virtuels et HTS, un nombre réduit de composés « leads » entre en phase d'optimisation. De manière similaire au filtrage « drug-like », l'optimisation des « leads » pourra être restreinte par des critères « lead-like », puisque celle-ci conduit généralement à une augmentation de leurs masses molaires, du logP et de la complexité générale des composés (affectant par la suite les possibilités de synthèse chimique). Hann & Oprea ont étudié les caractéristiques physico-chimiques de 532 médicaments et de 176 « leads » à la recherche de critères « lead-like ».[81] Ils définissent, de manière plus exigeante, une combinaison des règles de Lipinski et Veber (***Tableau 2***).

Propriétés physico-chimiques	**« Drug-likeness »**		**« Lead-likeness »**	
	Règles de Lipinski[73]	Veber et al.[74]	Hann & Oprea[81]	« Règle de 3 »
Poids moléculaire (en Da)	≤ 500	-	≤ 460	< 300
Lipophilie (logP)	≤ 5	-	[-4, 4.2]	≤ 3
Nombre de donneurs de liaisons hydrogènes (DLH)	≤ 5	DLH + ALH ≤ 12 *	≤ 5	≤ 3
Nombre d'accepteurs de liaisons hydrogènes (ALH)	≤ 10		≤ 9	≤ 3
Aire de la surface polaire	-	≤ 140 Å² *	-	≤ 60 Å² **
Nombre de liaisons rotatives	-	≤ 10	≤ 10	≤ 3 **
Nombre de cycles aromatiques	-	-	≤ 4	-
Solubilité dans l'eau ($logS_w$)	-	-	≥ -5	-

Tableau 2. *Valeurs utilisées dans les différents filtres permettant la sélection de molécules ou fragments aux propriétés pharmacocinétiques « drug-like » ou « lead-like ». (*) Pour les critères de Veber, en plus du nombre de liaisons rotatives, l'une de ces deux valeurs doit être respectée. (**) Dans l'utilisation de la « règle de 3 » ces critères sont optionnels, permettant éventuellement d'affiner les résultats.*

Selon des méthodes similaires, une « règle de 3 » a été définie pour la recherche de fragments « lead-like » et la construction de chimiothèques de fragments.[61] Ces critères ont été établis à partir des propriétés physico-chimiques de « hits » validés par cristallographie à rayons X à haut débit, à partir de fragments de 100 à 250 Da et sur une variété de cibles. Cette approche expérimentale est intéressante, la cristallographie ayant l'avantage d'identifier des « hits » de fragments qui n'auraient pas été mesurables par les autres méthodes disponibles.[61] Cependant,

les auteurs ne donnent que très peu de détails sur les méthodes ayant abouti à l'établissement de cette « règle de 3 ».

2.2.2.1.4. Toxicologie

Complémentairement aux caractéristiques ADME, il est également possible de filtrer les chimiothèques suivant la toxicité potentielle des composés. Les « hits » identifiés lors des criblages constitueront ainsi une base de meilleure confiance avant d'engager le processus d'optimisation et les essais pré-cliniques puis cliniques suivants. Parmi les caractéristiques de toxicité les plus importantes, la carcinogénicité, mutagénicité, cardiotoxicité et hépatotoxicité doivent être étudiées.[82] L'émergence des méthodes de toxicologie computationnelle a permis la prédiction de telles caractéristiques, en s'appuyant principalement sur deux techniques : l'établissement de modèles QSAR et les méthodes de docking.[82,83]

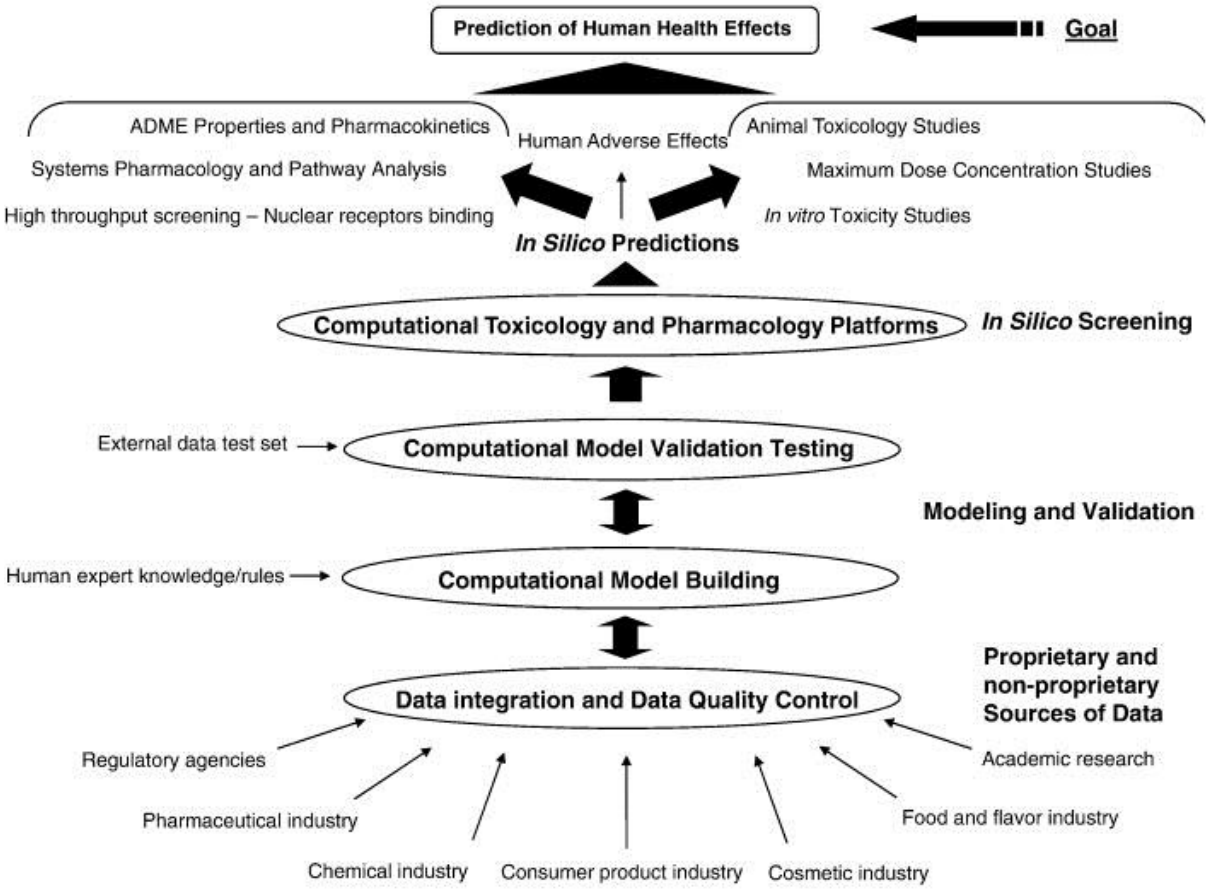


Figure 9. *Illustration du processus de prédiction de toxicité d'une molécule à partir des différents types de données disponibles. Des modèles de toxicité sont établis puis mis en œuvre pour la filtration des chimiothèques et la caractérisation des « hits » et « leads » sélectionnés. L'expérimentation in vitro et in vivo est requise durant les phases pré-cliniques et cliniques.*[83]

La mise en œuvre de ces techniques est possible grâce à la disponibilité de données toxicologiques de sources variées : industrie pharmaceutique et recherche académique, mais aussi industrie cosmétique, alimentaire et des produits de consommation courants (***Figure 9***). Un recensement exhaustif de ces bases de données est disponible,[82] comportant notamment la base données de pharmacovigilance de la FDA, très détaillée (openFDA, système FAERS ou FDA Adverse Event Reporting System).[84]

L'utilisation de modèles QSAR (voir points 2.3.4 et 2.4.4) est la technique la plus répandue pour la prédiction de toxicité d'un composé. Ils reposent sur la construction de modèles statistiques afin d'estimer la toxicité d'un composé, grâce à l'utilisation de données expérimentales obtenues précédemment sur des composés similaires.[36,82,85] La recherche pharmaceutique fut la première à tirer profit de ces modèles. Cependant, l'application récente des législations européennes REACH (Registration, Evaluation, Authorization and Restriction of Chemicals)[86] et CLP (Classification, Labelling and Packaging)[87] oblige actuellement chaque industrie à identifier et caractériser la toxicologie de l'ensemble des composés et substances qu'elle requière.[86,87] Des modèles QSAR préliminaires peuvent être établis dans ce cadre, avec l'objectif d'estimer les études expérimentales nécessaires à l'enregistrement des composés et les financements associés. Les données recueillies dans le cadre de l'application de la législation REACH, massives et rigoureusement formatées, constituent en retour une nouvelle manne pour l'établissement de modèles QSAR. La recherche pharmaceutique en bénéficiera particulièrement, ces données étant d'ores et déjà largement étudiées.[85,88–90]

Une seconde approche à la prédiction de toxicité, complémentaire aux modèles QSAR, repose sur l'estimation du potentiel de liaison d'un composé sur des cibles connues comme étant impliquées dans des mécanismes de toxicité. Les méthodes de docking (voir point 2.4.6) peuvent être appliquées à ces fins. L'exemple du canal potassique hERG, dont l'inhibition provoque une cardiotoxicité forte,[91] illustre le succès des méthodes de docking dans la prédiction de toxicité. Du et al. proposent en 2007 une approche efficace pour l'identification de composés cardiotoxiques, alliant méthodes de docking et modélisation par homologie (***Figure 10***).[92] Après avoir modélisé la structure du canal hERG à partir d'une structure cristallographique du canal potassique KcsA, ceux-ci utilisent le logiciel GOLD[93] pour prédire l'affinité de composés au site de liaison du canal hERG. Ils obtiennent une bonne concordance entre les valeurs d'affinité prédites et déterminées expérimentalement, validant ainsi la capacité prédictive du modèle pour l'identification de composés cardiotoxiques.[92]

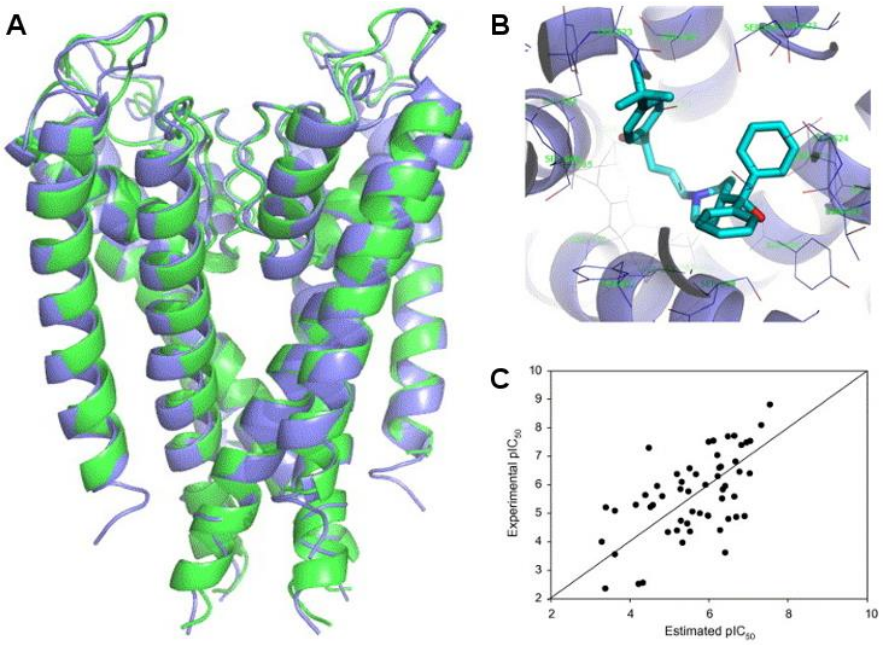


Figure 10. *Illustrations issues de la méthode de Du et al. pour la prédiction de cardiotoxicité d'un composé.*[92] *(A) Superposition de la structure cristallographique de référence du canal potassique KcsA (violet) et de la structure du canal hERG modélisée par homologie (vert), représentations cartoon. (B) Conformation amarrée de la terfenadine dans le site de liaison du canal hERG modélisé. (C) Les valeurs d'affinité prédites par le logiciel GOLD*[93] *sont en bonne concordance avec les valeurs d'affinité expérimentales (pIC*$_{50}$*), disponibles pour 56 composés lors de cette étude.*

2.2.2.3. Préparation des conformations 3D

Les molécules sont stockées dans les chimiothèques principalement suivant les formats sdf et mol2, souvent en 2D, décrivant notamment les types d'atomes, leurs charges, leur connectivité, les types de liaisons et la stéréochimie. Après avoir sélectionné et filtré une chimiothèque d'intérêt, il convient de préparer les conformations 3D des molécules (ou conformères) à partir des données chimiques brutes informatisées. Les méthodes de docking, recherche de similarité, RD-QSAR et les approches pharmacophoriques requièrent toutes la génération de conformères, réalisée soit à la volée par les logiciels, soit en amont de leur exécution. L'objectif des méthodes de génération de conformères est d'obtenir une ou plusieurs conformations de basse énergie d'une molécule, qui représenteront potentiellement ses conformations bioactives.[94]

La génération des conformères peut être réalisée principalement par des algorithmes systématiques ou stochastiques. Pour une approche systématique, l'espace global des conformères est exploré par l'augmentation incrémentielle des angles de torsions de toutes les liaisons rotatives par une valeur prédéfinie (***Figure 11***).[95] Tous les conformères disponibles suivant ce paramètre sont énumérés. Les conformères de plus basse énergie seront ensuite

sélectionnés grâce à l'usage de champs de force et selon des critères de diversité, en fonction du nombre de conformères demandés par l'utilisateur. Suivant cette approche, de nombreux outils rendent de bonnes performances, parmi lesquels : CORINA[96], OMEGA,[97] ROTATE,[98] ConfGen,[94] et Confab.[99]

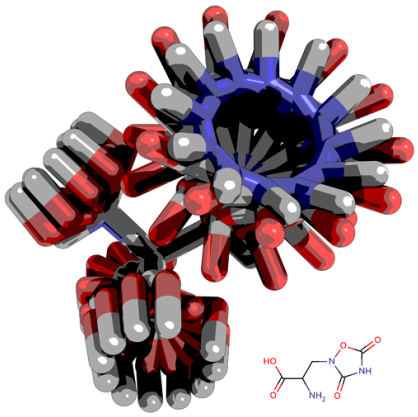


Figure 11. *Représentation artistique de conformères de l'acide quisqualique générés par incrémentation des angles de torsions par 30 degrés.*[95] *L'acide quisqualique est un agoniste des récepteurs AMPA*[100,101] *et des récepteurs métabotropes du glutamate, intervenant dans les synapses glutamatergiques.*[102] *En bas à droite, l'acide quisqualique décrit suivant les fichiers de stockage.*

Cependant, l'approche systématique peut demander des temps de calcul importants du fait de l'augmentation exponentielle du nombre d'états conformationnels pour des molécules possédant un grand nombre de liaisons rotatives. Ainsi, l'utilisation de cette méthode n'est pas automatique, notamment lorsque l'on s'intéresse à de larges chimiothèques. Les approches stochastiques pallient ce problème en explorant l'espace des conformations par des algorithmes génétiques,[103–105] de distance géométrique[106,107] ou de Monte Carlo[108–110] (voir point 2.4.6.2.2). Les outils suivant obtiennent de bonnes performances dans la mise en œuvre de ces méthodes, respectivement : Balloon,[104] RDKIT[111] et Frog2.[112]

2.3. Criblage virtuel « ligand-based »

Les méthodes de criblage « ligand-based » reposent sur la connaissance préalable de ligands ayant une activité sur la cible thérapeutique. Il sera ainsi possible d'utiliser ces ligands comme une première base de « hits » afin d'identifier d'autres composés similaires, présentant des caractéristiques d'activité communs aux ligands connus de la cible.[113,114] Différents types de descripteurs moléculaires pourront être calculés pour quantifier la similitude entre composés. Suivant le nombre de ligands connus de la cible thérapeutique, plusieurs méthodes peuvent être

employées : la recherche de similarité, le criblage de pharmacophores, ou les approches QSAR.[114]

2.3.1. Descripteurs moléculaires

Les descripteurs moléculaires peuvent être obtenus mathématiquement ou expérimentalement et sont classés suivant leur dimension : 1D (général), 2D (topologie) ou 3D (conformation).[115] Les descripteurs 1D représentent des propriétés globales de la molécule et peuvent être obtenus à partir de sa formule chimique (poids moléculaire, pK_a, logP, nombres absolus et relatifs d'atomes, d'hétéroatomes, de liaisons, de cycles et de groupes fonctionnels d'un type donné, etc.). Les descripteurs 2D décrivent et sont obtenus à partir de la topologie de la molécule (indices de connectivité, de complexité, de ramification, quantiques, etc.), tandis que les descripteurs 3D décrivent directement les caractéristiques des conformations stériques d'une molécule (indices de forme ou de volume, moments d'inertie, valeurs d'angles dièdres, etc.) (***Figure 12***).[115]

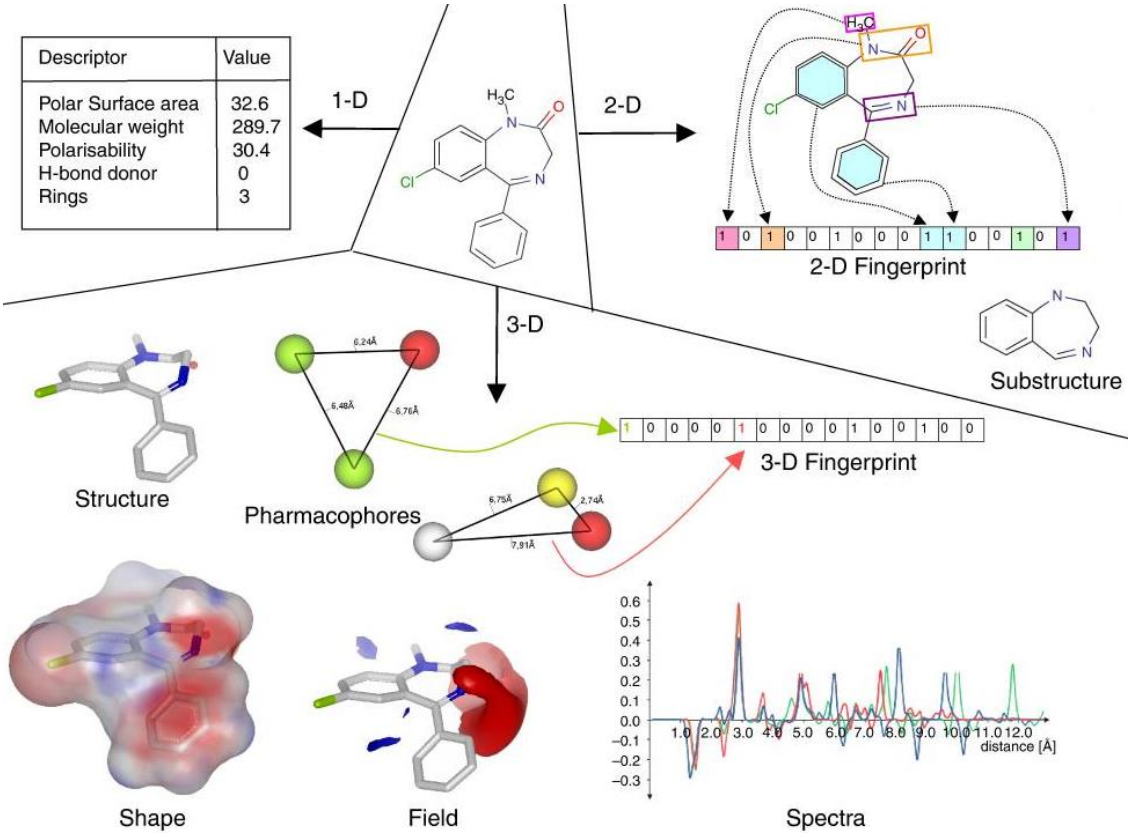


Figure 12. *Exemples de descripteurs inscrits dans les différentes dimensions d'une molécule.*[115]

2.3.2. Recherche de similarité

La recherche de similarité peut être employée dès lors qu'un ligand connu de la cible thérapeutique est identifié. Des métriques de similarité (voir point 2.3.2.5) permettront ensuite de comparer les ligands de référence aux molécules criblées, à partir de descripteurs moléculaires adaptés, afin de prédire leurs profils d'activité.

2.3.2.1. Structures communes maximales

Une des premières méthodes utilisées pour la recherche de similarité, appliquée dès 1977, consiste en l'identification des sous-structures communes maximales (Maximum Common Substructure ou MCS) des molécules.[116] Cette technique repose sur l'observation que des structures moléculaires similaires partagent des propriétés physico-chimiques.[113,116] Algorithmiquement, il est possible de représenter les molécules sous forme de graphes, permettant la comparaison des structures moléculaires pour l'identification des MCS et de nombreux autres descripteurs de similarités locales (***Figure 13***). De la même manière, l'identification de sous-structures communes maximales discontinues (Topologically constrained Disconnected MCS ou TD-MCS) peut s'avérer informative (***Figure 14***).[117–119] Plusieurs logiciels permettent la mise en œuvre de ces méthodes, dont OpenEye GraphSim TK,[120] MultiMCS,[117] RDKIT,[111] ChemAxon JKlustor.[121]

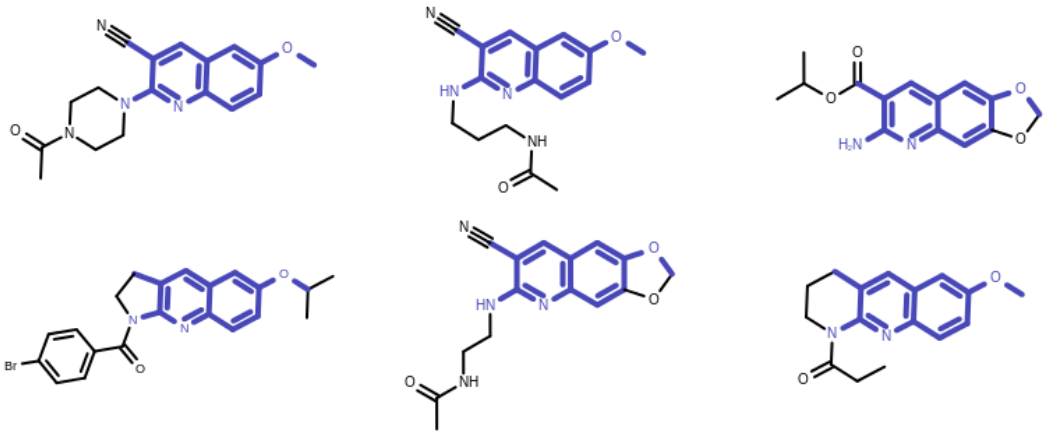


Figure 13. *Représentation 2D de la sous-structure commune maximale (MCS, bleu) entre six molécules différentes.*[120]

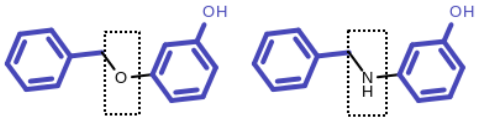


Figure 14. *Représentation 2D de la sous-structure commune maximale discontinue (TD-MCS, bleu) entre deux molécules différentes. Des caractéristiques physico-chimiques proches peuvent être détectées avec une meilleure sensibilité par rapport à l'usage des MCS.*[120]

2.3.2.2. Empreintes 2D

Les empreintes 2D (ou « 2D fingerprints ») reposent également sur la topologie et permettent une comparaison très rapide et efficace des molécules.[122] Dans cette approche, chaque molécule est représentée par un vecteur binaire. Chaque bit de ce vecteur représente alors la présence (1) ou l'absence (0) d'un fragment moléculaire. La similarité entre deux molécules peut ensuite être évaluée par la comparaison des deux vecteurs définis pour chaque molécule (***Figure 15***).[123] Deux types d'empreintes 2D peuvent être discernés en fonction de la procédure utilisée dans la définition des fragments.

Les empreintes 2D dites « keyed fingerprints » utilisent un dictionnaire de fragments prédéfini.[122] La plus courte empreinte, qui est également la plus utilisée, ne comporte que 166 bits qui représentent la majorité des caractéristiques physico-chimiques utiles à l'identification de molécules d'intérêt thérapeutique (empreintes MACCS[124]). Les empreintes utilisées par PubChem[125] pour la recherche de similarité comportent 881 bits (implémentées dans ChemFP[126] et CDK[127,128]). Les empreintes BCI[129] permettent à l'utilisateur de sélectionner les fragments utilisés et donc le nombre de bits total de l'empreinte, le dictionnaire inclus par défaut comportant 1052 entrées.[130]

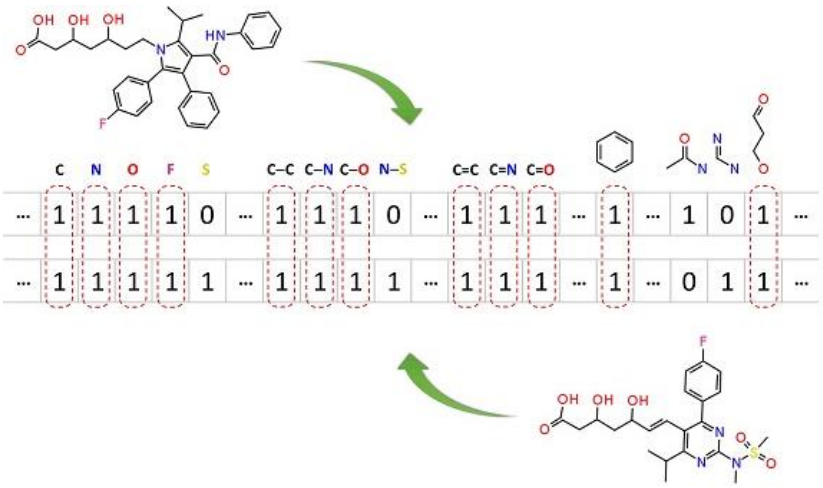


Figure 15. *Comparaison de deux molécules par la méthode des empreintes 2D. Chaque bit représente la présence (1) ou l'absence (0) d'un fragment moléculaire prédéfini. Les deux vecteurs binaires peuvent ensuite être comparés pour obtenir un indice de similarité des molécules.*

Les empreintes 2D dites « hashed fingerprints » utilisent un dictionnaire de fragments dynamique.[122] Les molécules de référence seront analysées algorithmiquement afin de générer l'ensemble des fragments existants jusqu'à un nombre défini de liaisons interatomiques (par défaut, n+7 pour les empreintes Daylight[131]), linéaires ou circulaires. La taille du vecteur binaire

associé à chaque molécule dépendra donc du nombre de fragments identifiés dynamiquement, permettant ainsi une meilleure exploration des structures de référence et une meilleure discrimination des caractéristiques physico-chimiques aptes à conférer une activité aux molécules. Il existe également des méthodes reposant sur un dictionnaire de fragments à la fois prédéfinis et dynamiques, dont SYBYL-X (Unity 2D)[132], OpenEye GraphSim TK[120] et MP-MFP.[133]

2.3.2.3. Empreintes 3D

Bien que les empreintes 2D soient particulièrement efficaces pour la recherche de composés dont la structure est proche des molécules de référence, elles ne sont cependant pas conçues pour retrouver des composés structurellement différents et qui présenteraient un mode de liaison similaire sur la cible d'intérêt.[122,134] Hors, cette dernière approche présente de nombreux avantages et s'est dernièrement développée sous le nom de « scaffold hopping ».[135] Il est ainsi possible de focaliser la recherche sur les groupements fonctionnels, conférant leur activité aux molécules de référence, tout en acceptant des modifications des structures internes (ou squelettes). Le « scaffold hopping » permet notamment de rechercher de nouvelles séries de composés hors du domaine couvert par d'éventuels brevets. De la même manière, il sera possible de substituer un squelette rigide à un squelette trop flexible afin d'augmenter l'affinité d'un composé à sa cible ou ses propriétés ADME-Tox.[135] Cette approche peut également être source d'alternatives lorsque la synthèse de certains composés se révèle difficile ou impossible.

Ainsi, les empreintes 3D permettent de décrire une molécule suivant la présence ou l'absence de caractéristiques géométriques. Suivant cette méthode, chaque bit de l'empreinte moléculaire représente la présence ou l'absence de paires, triplets ou quadruplets d'atomes spécifiques espacés suivants des plages de distances prédéfinies et s'inscrivant ou non dans des plages de valeurs prédéfinies d'angles de valence et de torsion.[122] La construction d'une empreinte 3D pourra être réalisée à partir d'un pharmacophore (voir point 2.3.3), pour accélérer la recherche de similarité, ou sans modèle pharmacophorique, par exemple en représentant les distances géométriques par un nombre de liaisons interatomiques.[122] Plusieurs implémentations sont disponibles pour la mise en œuvre de recherches de similarité à partir d'empreintes 2D ou 3D, dont jCompoundMapper,[136] ChemAxon JChem,[137] MOE[138] et Schrödinger Canvas.[139]

2.3.2.4. Comparaison de formes

La recherche par comparaison de formes représente une autre approche pour l'identification de molécules similaires, reposant sur le principe que des molécules adoptant des formes 3D proches possèderaient des propriétés physico-chimiques communes.[140,141] Les méthodes disponibles pour la comparaison de formes sont regroupées en deux catégories suivant qu'elles nécessitent, ou non, une superposition préalable à la comparaison de deux molécules. Dans les deux cas, la flexibilité d'une molécule sera décrite par la génération de ses diverses conformations (voir point 2.2.2.3), exécutée en amont de la recherche de similarité.

Les méthodes dites « de superposition » requièrent donc une superposition préalable des conformations moléculaires afin de procéder à la comparaison de leurs formes (***Figure 16***). En 1991, Meyer et al. furent les premiers à proposer des approches rapides, permettant ce type de comparaisons à grande échelle avec les moyens de calcul de l'époque.[140,142] Leur méthode consiste à approximer la surface moléculaire de van der Waals de manière discrète dans une grille afin d'accélérer la comparaison des conformations moléculaires.

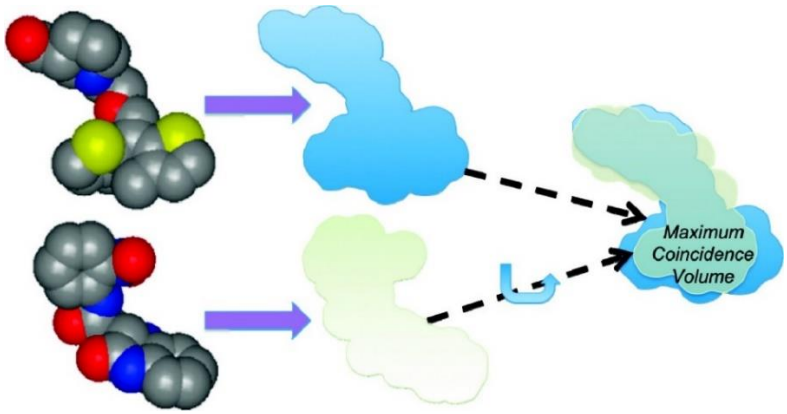


Figure 16. *Illustration du concept de similarité de forme pour la recherche de molécules similaires. Après avoir superposé les deux molécules, un indice de similarité peut être calculé en fonction du chevauchement de leurs volumes.*[141]

L'adaptation de fonctions gaussiennes pour la représentation de l'espace de densité électronique des atomes permis par la suite de réaliser des comparaisons plus flexible de différentes conformations.[143,144] Parmi les implémentations disponibles pour la comparaison de formes, le logiciel ROCS (Rapid Overlay of Chemical Structures) est considéré comme une référence.[141] Celui-ci utilise des fonctions gaussiennes centrées sur chaque atome afin de décrire la surface d'une molécule de manière flexible et permet d'inclure des caractéristiques physico-chimiques dans la recherche de similarité (hydrophobicité et donneurs et accepteurs de liaisons hydrogènes) (***Figure 17***).[141,145]

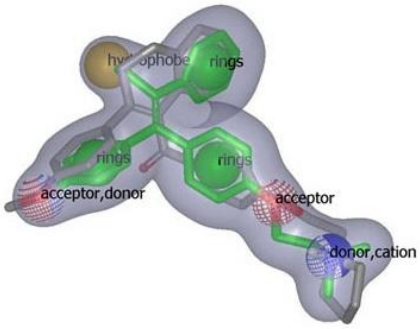


Figure 17. *Illustration de la comparaison d'une molécule (gris) avec une forme de référence (vert) incluant des propriétés physico-chimiques, réalisée avec le logiciel ROCS.*[145]

D'autres méthodes dites « de non-superposition » furent développées en parallèle avec l'objectif principal de réaliser des recherches de similarité sans dépendance à l'alignement préalable des conformations moléculaires. En effet, la pertinence de l'alignement impacte directement la qualité des résultats et celui-ci représente une part importante des temps de calculs nécessaires à l'exécution des méthodes de superposition.[146]

Zauhar et al. proposent en 2003 une approche intéressante permettant de décrire les caractéristiques 3D d'une surface moléculaire en une « signature de forme » en 1D.[147] Dans un premier temps, une représentation triangulée de la surface moléculaire est obtenue grâce à l'algorithme SMART (SMooth moleculAR surface Triangulator).[148] Les dimensions de cette surface sont ensuite explorées suivant une variété d'angles grâce à une méthode de « ray-tracing » et suivant les principes de réflexion optique, permettant d'obtenir une distribution de probabilité de la longueur des rayons (***Figure 18***). Les signatures de forme 1D ainsi obtenues sont suffisamment discriminantes pour permettre une comparaison efficace et rapide des différentes formes moléculaires.[147]

Cependant, si cette méthode permet de s'affranchir d'un alignement des molécules à comparer, celle-ci reste exigeante en calculs du fait de la construction d'une surface moléculaire et du procédé de « ray-tracing » utilisés préalablement à la comparaison très rapide des signatures de forme. Le besoin croissant de méthodes rapides pour la comparaison de conformations moléculaires, lié à la croissance des bases de données de molécules, a donc entrainé la recherche de nouvelles heuristiques.[149,150]

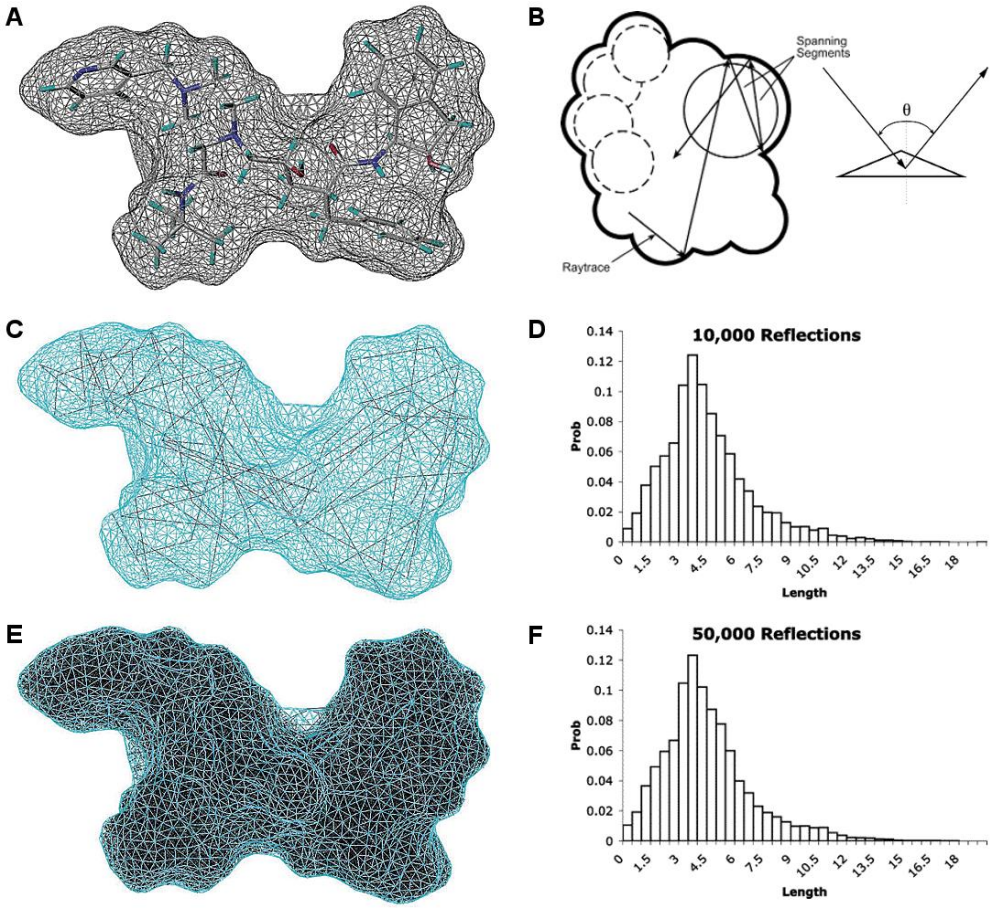


Figure 18. *Illustration du principe des « signatures de forme » de Zauhar et al. (A) Surface accessible au solvant de l'indinavir générée par l'algorithme SMART. (B) Illustration du principe de « ray-tracing » appliqué pour décrire la géométrie de la surface moléculaire. (C) Traces obtenues pour l'indinavir à basse densité (100 traces). (D) Signature de forme de l'indinavir à 10.000 réflexions. (E) Traces obtenues pour l'indinavir à 50.000 réflexions. (F) Signature de forme de l'indinavir à 50.000 rélexions.*[147]

L'algorithme USR (Ultrafast Shape Recognition), proposé dans sa première version en 2006, reste le plus rapide à ce jour.[151,152] Cette approche met en application un théorème statistique qui démontre qu'une distribution est complètement déterminée par ses moments[153] et repose uniquement sur les distances interatomiques (hydrogènes non-compris). Chaque conformation moléculaire est représentée par un unique vecteur de 12 descripteurs représentant les 3 moments statistiques (moyenne, variance, asymétrie) des distances interatomiques par rapport à 4 positions de référence : centre de masse (P_1), atome le plus proche de P_1 (P_2), atome le plus éloigné de P_1 (P_3), atome le plus éloigné de P_3 (P_4).[151] Une fonction de score normalisée permet ensuite de quantifier la similarité entre deux conformations moléculaires (***Figure 19***).

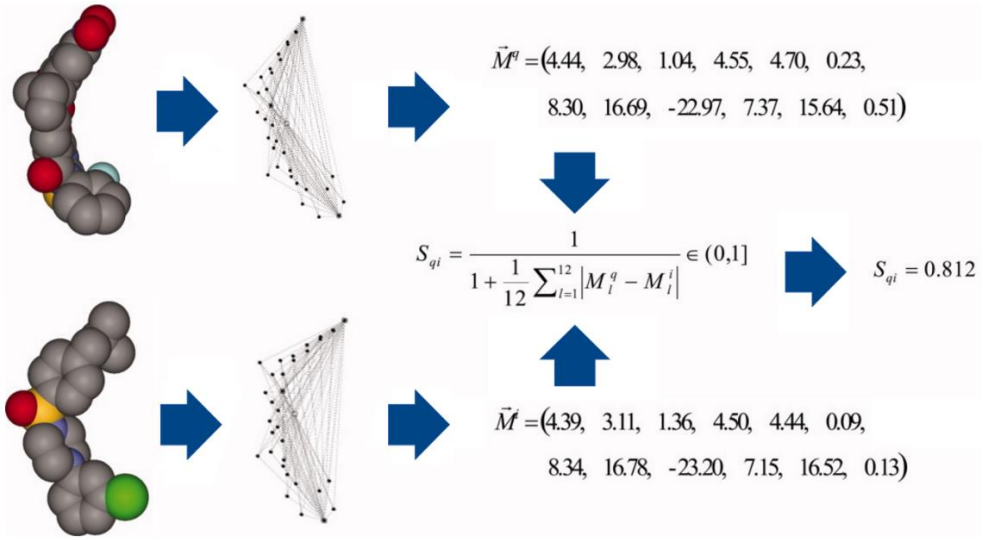


Figure 19. *Exemple d'une comparaison de formes réalisée par la méthode USR entre deux molécules. Les temps de calcul nécessaires à l'exécution d'une telle requête sont de l'ordre de quelques millisecondes pour ces molécules comportant 33 et 26 atomes lourds.*[151,152]

Les heuristiques utilisées permettent donc de réduire drastiquement les temps de calcul nécessaires par rapport aux autres méthodes de comparaison de formes (plusieurs ordres de magnitude).[151,152] D'autres versions de l'algorithme USR ont récemment été développées afin de prendre en compte des informations pharmacophoriques (USRCAT[154], UFSRAT[155]), d'inclure directement des critères électrostatiques (ElectroShape[156]), ou de mieux gérer la chiralité des molécules (CSR[157]) tout en conservant une grande optimisation des temps de calculs. Bien que l'engouement pour ces méthodes très rapides soit assez fort, elles présentent la caractéristique commune d'être sensibles aux variations de taille entre conformations moléculaires, ce qui n'est pas le cas des empreintes topologiques 2D et 3D.[154] Cette spécificité est illustrée par l'évaluation de l'algorithme USRCAT réalisé par Schreyer et al. sur la banque de données DUD-E[158] (Directory of Useful Decoys Enhanced), qui comporte des ligands de taille très variée (***Tableau 3***).[154] L'évaluation de la méthode ElectroShape sur la banque DUD (Directory of Useful Decoys)[159] révèle de meilleures performances (***Figure 20***),[156] plus proches de celles des méthodes d'empreinte 2D.[160] Ainsi, l'algorithme USR et ses améliorations représentent des méthodes particulièrement efficaces pour les recherches de « hits » de type « scaffold hopping ».[152,161,162] Les banques d'évaluation sont détaillées en point 3.2.1.

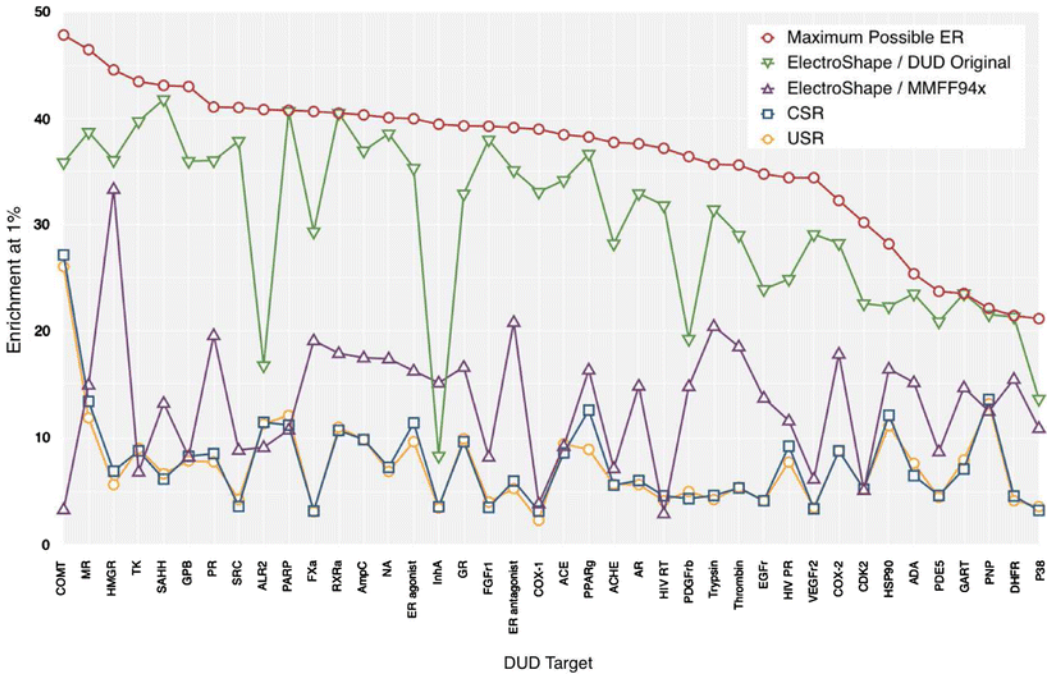


Figure 20. *Facteurs d'enrichissement obtenus à 1% des différents jeux de données lors de criblages rétrospectifs sur la banque d'évaluation DUD*[159] *avec les méthodes USR,*[151] *CSR*[157] *et ElectroShape*[156] *réalisés par Armstrong et al. pour l'évaluation des performances de l'algorithme ElectroShape.*[156] *Deux types de charges atomiques ont été utilisées avec ElectroShape : celles assignées initialement dans les données DUD et des charges calculées par le champ de force MMFF94x*[163] *implémenté dans MOE.*[138] *Les auteurs indiquent que les résultats obtenus avec les charges atomiques assignées dans les données DUD (vert) sont anormaux et ne doivent pas être considérés comme un indice de performance.*[156]

Type de méthode	**Méthode**	**Facteurs d'enrichissement moyens**		
		1.0%	**0.5%**	**0.25%**
USR et améliorations	USR[151]	5.00	6.71	8.84
	ElectroShape[156]	8.40	11.27	14.48
	USRCAT[154]	8.62	11.99	15.64
Empreintes topologiques 2D	Circular FP[120]	32.14	42.54	49.72
	Path FP[120]	24.50	35.14	44.27
	Tree FP[120]	19.28	28.98	38.59
	MACCS166[120,124]	20.90	28.85	36.60

Tableau 3. *Facteurs d'enrichissement moyens obtenus lors d'un criblage rétrospectif de la banque d'évaluation DUD-E*[158] *réalisé par Schreyer et al. pour l'évaluation des performances de l'algorithme USRCAT.*[154] *Les méthodes USR et assimilées sont peu performantes, du fait de la grande variété de taille des ligands de la banque DUD-E.*[154]

2.3.2.5. Métriques de similarité

Ainsi, les méthodes de recherche de similarité représentent fréquemment les molécules sous forme de vecteurs qui regroupent les différents descripteurs moléculaires. Les méthodes d'empreintes utilisent des vecteurs de variables dichotomiques (0 ou 1), tandis que les différentes méthodes USR et la méthode des signatures de formes, par exemple, utilisent des vecteurs de variables continues.[147,151,152] Dans les deux cas, ces représentations vectorielles permettent la comparaison rapide de molécules ou de conformations moléculaires, deux à deux, grâce à différentes métriques de similarité. Ces métriques peuvent être classées en trois catégories : les mesures de distance, les coefficients de similarité et les coefficients de corrélation. Les métriques de distance peuvent être converties en métriques de similarité et inversement (***Équation 1***).[123] Les métriques de distance sont définies sur l'intervalle [0,+∞], les coefficients de similarité sur l'intervalle [0,1] (1 : similarité maximale), tandis que les coefficients de corrélation sont définis sur l'intervalle [-1,1] (-1 : corrélation négative, 0 : absence de corrélation, 1 : corrélation positive).

$$Similarité = \frac{1}{1 + Distance}$$

Équation 1. *Relations entre les métriques de similarité et les métriques de distance.*[123]

Les formules des principales métriques utilisées par les méthodes de recherche de similarité sont données en ***Tableau 4***, dans le contexte de variables continues ou dichotomiques. Pour la comparaison de leurs signatures de formes, Zauhar et al. choisissent d'utiliser des distances de Manhattan, éventuellement pondérées (***Figure 18***).[147] La méthode USR et ses améliorations utilisent également une distance de Manhattan normalisée pour la comparaison des moments définissant la forme d'une conformation moléculaire (***Figure 19***).[151,152] Les distances de Manhattan et Euclidienne sont dites monotoniques l'une de l'autre, c'est-à-dire qu'elles résultent en un classement identique de la distance des composés par rapport à un ligand de référence, bien que les valeurs de distances diffèrent. Parmi les mesures de similarité appliquées par les méthodes d'empreintes (voir points 2.3.2.2 et 2.3.2.3), le coefficient de Tanimoto est le plus fréquemment utilisé.[123] Dans le contexte de variables dichotomiques utilisées par ces méthodes, les coefficients de similarité permettent donc de quantifier la présence de caractéristiques communes, inversement aux métriques de distance qui considèrent également l'absence commune d'une caractéristique comme une preuve de similarité.[164] Ainsi, la seule faiblesse du coefficient de Tanimoto (T_c) apparaît lorsque les empreintes comparées comportent peu de bits positifs, conduisant à un biais vers de faibles valeurs du T_c. La vaste majorité des

méthodes d'empreintes utilisent toutefois plusieurs centaines de bits, dont suffisamment sont positifs pour que l'utilisation du T_c soit pertinente et efficace.[123] Le coefficient de Dice est monotonique du coefficient de Tanimoto, tandis que le coefficient de Cosine est fortement corrélé au T_c sans lui être strictement monotonique.[123] Initialement, le coefficient de Cosine fut proposé pour la comparaison de vecteurs (moments) et permet d'estimer la similarité de leurs orientations, sans prendre en compte leurs magnitudes.[123,164]

Métrique	Formule pour des variables continues	Formule pour des variables dichotomiques
Distance de Manhattan	$D_{A,B} = \sum_{i=1}^{n} \lvert x_{iA} - x_{iB} \rvert$	$D_{A,B} = a + b - 2c$
Distance Euclidienne	$D_{A,B} = \sqrt{\sum_{i=1}^{n} (x_{iA} - x_{iB})^2}$	$D_{A,B} = \sqrt{a + b - 2c}$
Coefficient de Cosine	$S_{A,B} = \frac{\sum_{i=1}^{n} x_{iA} x_{iB}}{\sqrt{\sum_{i=1}^{n} {x_{iA}}^2 \sum_{i=1}^{n} {x_{iB}}^2}}$	$S_{A,B} = \frac{c}{\sqrt{ab}}$
Coefficient de Dice	$S_{A,B} = \frac{2 \sum_{i=1}^{n} x_{iA} x_{iB}}{\sum_{i=1}^{n} {x_{iA}}^2 + \sum_{i=1}^{n} {x_{iB}}^2}$	$S_{A,B} = \frac{2c}{a + b}$
Coefficient de Tanimoto	$S_{A,B} = \frac{\sum_{i=1}^{n} x_{iA} x_{iB}}{\sum_{i=1}^{n} {x_{iA}}^2 + \sum_{i=1}^{n} {x_{iB}}^2 - \sum_{i=1}^{n} x_{iA} x_{iB}}$	$S_{A,B} = \frac{c}{a + b - c}$

Tableau 4. *Principales métriques utilisées par les méthodes de recherche de similarité dans le contexte de variables continues ou dichotomiques entre deux composés A et B (D : distance, S : similarité, n : taille du vecteur, x_{iA} : variable du vecteur décrivant le composé A à l'indice i, a : nombre de bits positifs pour le composé A, b : nombre de bits positifs pour le composé B, c : nombre de bits positifs communs aux composés A et B).*[123]

Enfin, les coefficients de corrélation permettent de quantifier l'association linéaire entre deux variables.[165] La corrélation de Pearson exprime le ratio de covariance des variables étudiées par rapport à leurs déviations standards et s'applique également au contexte de variables continues (***Équation 2***) ou dichotomiques (***Équation 3***).

$$Pearson_{A,B} = \frac{\sum_{i=1}^{n} x_{iA} x_{iB} - \frac{1}{n} \sum_{i=1}^{n} x_{iA} \sum_{i=1}^{n} x_{iB}}{\sqrt{\left(\sum_{i=1}^{n} {x_{iA}}^2 - \frac{1}{n} \sum_{i=1}^{n} {x_{iA}}^2\right)\left(\sum_{i=1}^{n} {x_{iB}}^2 - \frac{1}{n} \sum_{i=1}^{n} {x_{iB}}^2\right)}}$$

Équation 2. *Formule du coefficient de corrélation de Pearson appliqué à des variables continues (n : taille du vecteur, x_{iA} : variable du vecteur décrivant le composé A à l'indice i).*

$$Pearson_{A,B} = \frac{nc - ab}{\sqrt{nab(n-b)(n-a)}}$$

Équation 3. *Formule du coefficient de corrélation de Pearson appliqué à des variables dichotomiques (a : nombre de bits positifs pour le composé A, b : nombre de bits positifs pour le composé B, c : nombre de bits positifs communs aux composés A et B).*

2.3.3. Modèles pharmacophoriques « ligand-based »

L'idée que les groupes chimiques d'une molécule soient responsables de son action biologique fut évoquée pour la première fois par Ehrlich à la fin du XIX^{e} siècle.[166] Le premier modèle de pharmacophore, publié par Beckett et al. en 1963, défini des distances entre les groupements fonctionnels du site de liaison des récepteurs muscariniques à partir de molécules dont l'activité est confirmée sur ces récepteurs (***Figure 21***).[167] Kier affine ensuite ce modèle fonctionnel et utilise pour la première fois le terme de « pharmacophore » en 1971.[168]

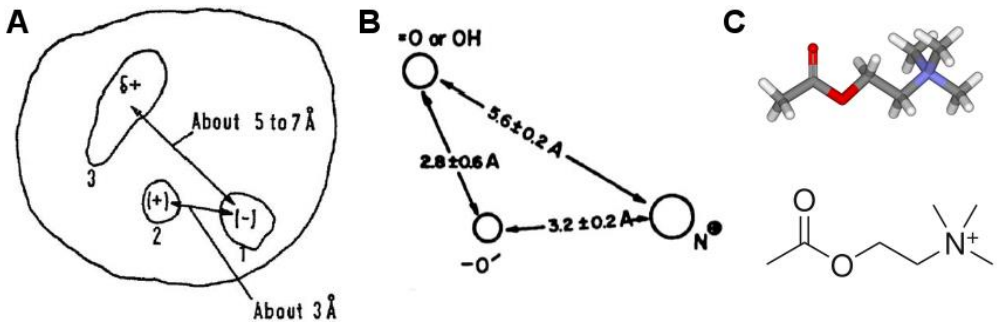


Figure 21. *Illustrations originales des premiers pharmacophores publiés par Kier. (A) Modèle des récepteurs muscariniques comportant un point chargé négativement permettant d'accueillir l'amine quaternaire de l'acétylcholine et de ses analogues (1) et deux points chargés positivement (2 et 3). (B) Pharmacophore basé sur les ligands des récepteurs muscariniques proposé en 1971. (C) Structure de l'acétylcholine.*[167,168]

Un pharmacophore est désormais défini comme l'ensemble des propriétés moléculaires stériques et électrostatiques nécessaires à établir des interactions supramoléculaires optimales avec une cible biologique spécifique de manière à déclencher ou bloquer sa réponse biologique.[169] Suivant cette définition, des molécules partageant des pharmacophores similaires devraient se lier de manière similaire sur un récepteur donné. Un pharmacophore peut donc être utilisé comme référence pour procéder au criblage de chimiothèques, en recherchant les molécules qui s'inscrivent dans celui-ci. Ces approches reposent donc sur la définition de points pharmacophoriques complémentaires les uns des autres, considérés comme des groupes fonctionnels et non plus des groupes d'atomes. Les points pharmacophoriques recherchés sont généralement les donneurs et accepteurs de liaisons hydrogènes, les points d'interaction

électrostatique et les groupes aromatiques et hydrophobes.[170] Un pharmacophore est dit « ligand-based » lorsqu'il est déterminé à partir de composés actifs utilisés comme référence sans prendre en compte la structure de leur récepteur. Inversement, un pharmacophore est dit « structure-based » lorsqu'il est construit à partir de la structure du site de liaison de la cible étudiée (voir point 2.4.3).

2.3.3.1. Pharmacophores 2D

Les approches pharmacophoriques 2D reposent uniquement sur la topologie des molécules pour définir et comparer les points pharmacophoriques en présence. De nombreuses approches ont été développées suivant ce principe, parmi lesquelles les « Similog keys »,[171] les descripteurs CATS 2D (Chemical Advanced Template Search),[172] la méthode ErG (Extended reduced Graph)[173] et les « feature trees ».[174]

La méthode des « Similog keys » est à l'interface de la recherche de similarité par les méthodes d'empreintes 2D et pharmacophoriques 2D. Une clé similog est un vecteur encodant les positions relatives et les caractéristiques de trois groupements d'atomes fonctionnels présents dans une molécule. Chaque groupe d'atomes est encodé suivant une combinaison de 4 bits : donneur de liaisons hydrogènes (code 1000), accepteur de liaisons hydrogènes (code 0100), encombrement (code 0010) et électropositivité (code 0001) suivant des seuils définis (***Figure 22***).[171] Par exemple, un groupement alcool est encodé 1100 du fait de son encombrement réduit et de son caractère à la fois donneur et accepteur de liaisons hydrogènes.

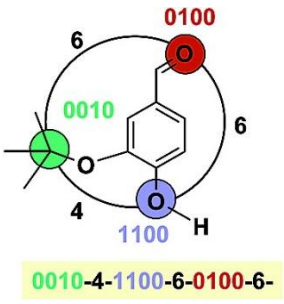


Figure 22. *Exemple de clé similog. Le groupement alcool est encodé 1100 du fait de son encombrement réduit et de son caractère à la fois donneur et accepteur de liaisons hydrogènes.*[171]

Les positions relatives des groupes d'atomes fonctionnels sont décrites par le nombre de liaisons les séparant. L'énumération des clés similog uniques de la chimiothèque étudiée permettra ensuite d'obtenir une empreinte vectorielle binaire de ces clés pour chaque molécule. Ces empreintes similog seront ensuite utilisées pour le criblage fonctionnel de chimiothèques.

Schuffenhauer et al. indiquent que les empreintes similog ainsi construites sur la chimiothèque de Novartis de 2003 comportaient 5989 clés similog uniques.[171]

La méthode ErG[173] est une variante de la méthode des graphes réduits,[175] qui définit les groupes fonctionnels de manière plus abstraite. Chaque groupement fonctionnel est ici représenté par un unique nœud d'un graphe moléculaire et une attention particulière est portée aux systèmes cycliques. La première étape de la construction d'un graphe ErG consiste en l'ionisation de la molécule selon les conditions physiologiques et l'identification des groupements donneurs et accepteurs de liaisons hydrogènes. Les groupements hydrophobes terminaux constitués de trois atomes sont ensuite isolés, puis les groupements aromatiques et cycliques sont représentés de manière réduite (***Figure 23***). Le graphe obtenu est ensuite converti en descripteurs radiaux, de manière comparable aux clés similog, avec l'utilisation de distances topologiques calculées sur le graphe réduit.[173] Chaque molécule est finalement décrite par un vecteur binaire capturant la présence ou l'absence de ces descripteurs radiaux et permettant le criblage de chimiothèques.

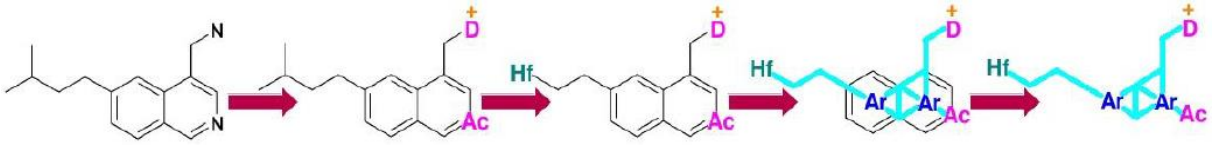


Figure 23. *Conversion d'une structure moléculaire en graphe réduit suivant la méthode ErG (D : donneur de liaisons hydrogènes, Ac : accepteur de liaisons hydrogènes, Hf : groupement hydrophobe, Ar : groupement aromatique).*[173]

La méthode ErG obtient de meilleures performances que les empreintes topologiques Daylight[131] et se prête efficacement aux criblages de type « scaffold hopping », principalement grâce à sa gestion particulière des structures cycliques (***Figure 24***) alliée à une approche pharmacophorique.[173]

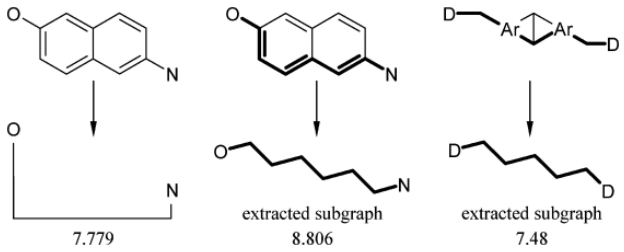


Figure 24. *Illustration de la gestion des structures cycliques par la méthode ErG. Les distances interatomiques réelles (A) et attribuées à une représentation topologique « classique » (B) ou « abstraite » (C) sont indiquées en Å. Les distances interatomiques encodées suivant la méthode ErG sont plus proches de la vérité terrain en comparaison aux distances topologiques précédemment décrites.*[173]

Les descripteurs CATS 2D[172] représentent une autre approche à l'interface des méthodes d'empreintes 2D et pharmacophoriques 2D. Dans sa seconde version, CATS 2D utilise six caractéristiques pour décrire les types pharmacophoriques de chaque atome, selon leur nature : lipophile (L), aromatique (R), donneur (D) ou accepteur (A) de liaisons hydrogènes, chargé positivement (P) ou négativement (N).[176] La distance topologique la plus courte est ensuite calculée entre chaque atome de la molécule afin de réaliser l'énumération des paires de points pharmacophoriques de chaque type (21 combinaisons) en fonction d'une distance topologique définie (***Figure 25***). Le nombre d'occurrences de chaque paire de points pharmacophoriques en fonction des distances topologiques constitue le vecteur utilisé pour la comparaison des molécules, à l'aide de métriques telles que la distance euclidienne.[176]

Bien que les descripteurs CATS 2D soient désormais dépassés par les méthodes d'empreintes pharmacophoriques radiales 2D comme les « Similog keys » et la méthode ErG dans la cadre de recherches de similarité, ils présentent de bonnes performances pour les criblages de type « scaffold hopping ».[176] L'algorithme CATS 2D, proposé en 1999, est considéré comme un des pionniers de son domaine.

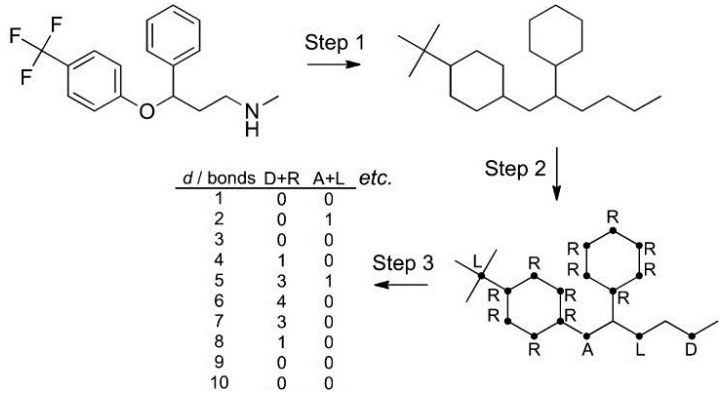


Figure 25. *Principe de calcul des descripteurs CATS 2D. Les points pharmacophoriques sont considérés par paires et énumérés en fonction de leurs distances topologiques.*[176]

Les « feature trees »,[174] proposés en 1998, font également partie des méthodes pionnières de la représentation pharmacophorique 2D. Cet algorithme fut le premier à permettre une alternative à la représentation linéaire des molécules, utilisée par les empreintes pharmacophoriques type CATS, afin de mieux prendre en compte les liaisons interatomiques aux structures cycliques. Les graphes moléculaires construits suivant cette méthode reposent sur des critères à la fois stériques et physico-chimiques (***Figure 26***). La superposition maximale de deux graphes moléculaires permet ensuite d'évaluer la similarité de deux molécules à faible coût computationnel.[174] Cette approche est particulièrement efficace pour l'identification de cycle et

d'hétérocycles alternatifs au sein des structures moléculaires et reste performante dans les criblages de type « scaffold hopping ».[135,172]

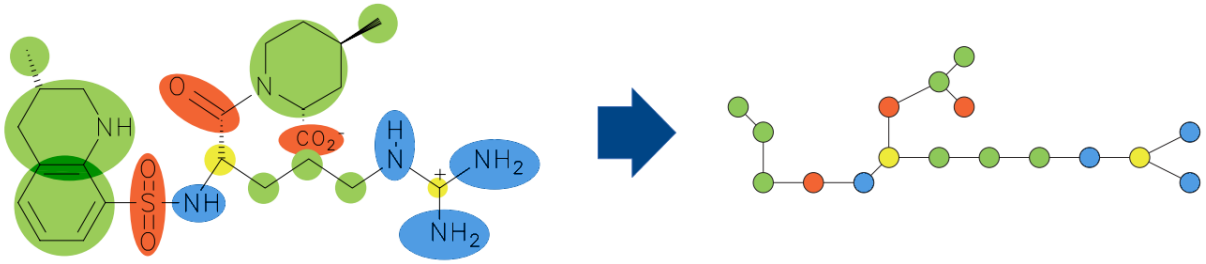


Figure 26. *Illustration de la construction d'un « feature tree ». Chaque nœud du graphe moléculaire encode un point pharmacophorique : donneur (bleu) ou accepteur (orange) de liaisons hydrogènes, point hydrophobe (vert) ou absence d'interaction directe (jaune).*[174]

2.3.3.2. Pharmacophores 3D

Contrairement aux approches 2D, les approches pharmacophoriques « ligand-based » 3D décrivent l'arrangement spatial des propriétés physico-chimiques conférant leur activité à une ou plusieurs molécules de manière géométrique.[177] Des pharmacophores 3D peuvent également être construits à partir de la structure 3D d'un site de liaison à la surface d'une protéine (voir point 2.4.3). Un pharmacophore « ligand-based » est construit et appliqué selon les étapes suivantes.[177]

2.3.3.2.1. Sélection des ligands de référence

Le choix des ligands utilisés pour la construction d'un pharmacophore a un impact direct sur le modèle résultant. Lorsqu'un grand nombre de ligands est disponible, les molécules utilisées pour la construction du pharmacophore sont sélectionnées selon des critères de diversité afin de favoriser l'identification des caractéristiques clés de leurs interactions avec la cible d'intérêt.[177] Cependant, il est important de quantifier les effets des « outliers » dans la construction d'un modèle pharmacophorique, puisque des ligands très différents pourraient être liés à différents sites actifs à la surface de la protéine cible et fausser la construction du modèle.[178]

La majorité des méthodes repose sur l'utilisation de ligands connus de la protéine étudiée et ne considère pas les données d'activité quantifiées des ligands (K_i, IC_{50}, pIC_{50}, EC_{50}, ou autres).[177] Quelques autres approches ont également été proposées. L'algorithme CLEW,[179] mis au point en 1998, permet la construction de pharmacophores à partir de composés actifs et inactifs, affinant les modèles grâce à la l'identification de caractéristiques physico-chimiques négligeables ou réduisant l'activité des composés.[177] Le logiciel HypoGen,[180] proposé en 2000,

met à profit les données d'activité quantifiées des ligands pour réaliser des prédictions d'activité sur la chimiothèque étudiée (méthode 3D-QSAR, voir point 2.3.4.2).

2.3.3.2.2. Recherche conformationnelle

Les conformations spécifiques adoptées par les ligands dans le site de liaison de la cible sont généralement inconnues et diffèrent de leurs conformations non-liées de plus basse énergie dans la vaste majorité des cas.[181] L'espace conformationnel de chaque ligand doit donc être exploré et représenté de manière exhaustive (voir point 2.2.2.3) afin de procéder à la recherche de pharmacophores communs.[182]

La plupart des méthodes réalisent cette étape de recherche conformationnelle dans un stage initial précédant l'identification des points pharmacophoriques, parmi lesquelles : RAPID,[183] MPHIL,[184] DISCO,[185] PHASE,[186] LigandScout,[187] MOE,[138] HipHop[188] et HypoGen.[180] Ces approches déterministes requièrent généralement un grand nombre de conformations par ligand (jusqu'à 100 conformations) afin de garantir une bonne précision des résultats.[181,189] Une autre stratégie consiste à réaliser la recherche des conformations moléculaires conjointement à l'identification des points pharmacophoriques, principalement grâce à des algorithmes de recherche aléatoire permettant d'explorer l'espace conformationnel de façon continue (SCAMPI,[190] GAMMA,[191] GASP,[192] GALAHAD,[193] algorithme de Cottrell et al.[194]).[182]

2.3.3.2.3. Identification des points pharmacophoriques des ligands

Les propriétés physico-chimiques utiles à la définition de pharmacophores sont principalement définies selon trois niveaux d'abstraction : atomique (atomes lourds), topologique (cycle phényl, groupe alcool, etc.) et fonctionnel (généralement : donneur ou accepteur de liaisons hydrogènes, caractère hydrophobe, aromatique et charges).[177]

Les méthodes RAPID,[183] MPHIL,[184] DISCO,[185] GAMMA,[191] GASP,[192] GALAHAD[193] et l'algorithme de Cottrell et al.[194] utilisent directement les types atomiques et leurs coordonnées spatiales afin d'assigner des caractéristiques physico-chimiques fonctionnelles. Inversement, les algorithmes SCAMPI,[190] LigandScout,[187] MOE,[138] PHASE,[186] HipHop[188] et HypoGen[180] reposent sur l'identification de groupes d'atomes topologiquement proches pour l'assignement de leurs caractéristiques pharmacophoriques, notamment grâce à des dictionnaires de fragments parfois couplés à des critères géométriques (***Tableau 5***).[182,195]

Ainsi, les logiciels représentent les caractéristiques pharmacophoriques selon des critères géométriques (sphères, plans et vecteurs) en admettant une certaine tolérance de leurs positionnements (***Figure 27A*** et ***Tableau 5***).[177,195] La position d'un point pharmacophorique donneur ou accepteur de liaisons hydrogènes est généralement centrée sur l'atome impliqué. Dans le cas de régions hydrophobes, le ou les points pharmacophoriques pourront être représentés par des sphères ou des plans et positionnés sur le centre de masse des atomes impliqués.[177] Les points donneurs ou accepteurs de liaisons hydrogènes peuvent être représentés par des vecteurs, permettant d'inclure la direction de la liaison hydrogène. Lorsque la direction d'une liaison hydrogène ne peut pas être définie de manière simple, plusieurs vecteurs pourront être définis afin de mieux capturer les propriétés géométriques du point pharmacophorique (***Figure 27B***).[177,195]

Propriétés	**Méthode**				
	SCAMPI	**LigandScout**	**MOE**	**PHASE**	**Catalyst**
Liaisons hydrogènes	Positions des atomes lourds	Sphères et vecteurs centrés sur les atomes lourds	Sphères et/ou vecteurs *	Vecteur de l'hydrogène donneur et sphère sur l'atome lourd accepteur	Sphères et vecteurs sur l'hydrogène donneur et l'atome lourd récepteur **
Groupes hydrophobes	Positions des atomes lourds	Sphères	Sphères et/ou plans *	Sphères, exclusif des cycles aromatiques	Sphères
Interactions aromatiques	Position du centre de masse	Sphères et plans	Sphères et/ou plans *	Sphères et plans	Sphères et plans
Interactions électrostatiques	Positions des atomes lourds	Sphères, charges optionnelles	Sphères, charges nécessaires	Sphères, charges optionnelles	Sphères, charges optionnelles
Définition des groupes fonctionnels	Coordonnées atomiques	Dictionnaire SMARTS avec géométrie	Coordonnées atomiques et/ou dictionnaire SMARTS avec géométrie *	Dictionnaire SMARTS avec géométrie	Coordonnées atomiques

Tableau 5. *Résumé des méthodes d'abstraction et de représentation des caractéristiques physico-chimiques des ligands implémentées dans SCAMPI,*[190] *LigandScout,*[187] *MOE,*[138] *PHASE,*[186] *Catalyst HipHop*[188] *et Catalyst HypoGen.*[180] *Ces cinq logiciels permettent l'édition manuelle des définitions des groupes fonctionnels par une interface graphique, des fichiers de configuration ou des scripts. (*) Plusieurs configurations de contraintes géométriques sont disponibles. (**) Un atome peut être exclusivement donneur ou accepteur de liaisons hydrogènes.*[182,195]

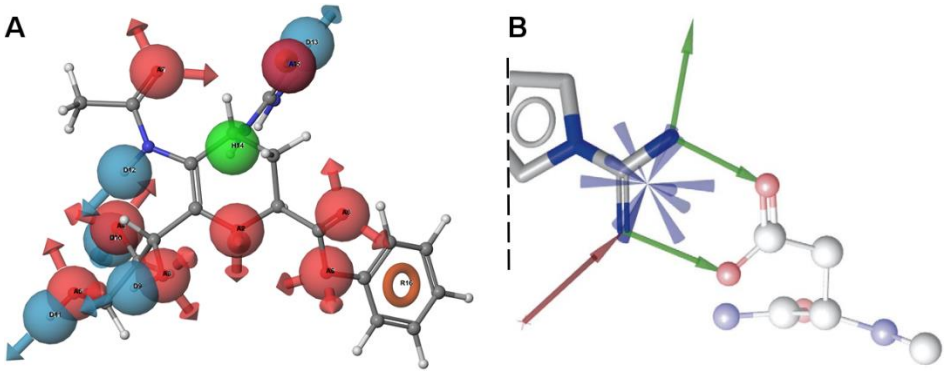


Figure 27. *(A) Exemple de pharmacophore construit et visualisé avec MOE*[138] *(anneau orange : cycle aromatique, vert : point hydrophobe, bleu : donneur de liaison hydrogène, rouge : accepteur de liaison hydrogène). (B) Illustration de la gestion des liaisons hydrogènes par LigandScout*[187] *lorsqu'une délocalisation électronique est observée (PDB 2gde,*[196] *Thrombine en complexe avec l'inhibiteur SN3401). LigandScout assigne trois vecteurs donneurs de liaisons hydrogènes (vert), un accepteur (rouge) et détecte un transfert de charge (étoile bleue).*[195]

2.3.3.2.4. Construction des modèles de pharmacophores

L'étape suivante a pour objectif de construire un ou plusieurs pharmacophores regroupant des points pharmacophoriques communs entre les ligands. L'approche la plus populaire repose sur l'identification de sous-structures communes maximales (MCS, voir point 2.3.2.1) afin de réaliser un alignement de ces points pharmacophoriques.[177,182] Les algorithmes de recherche de MCS se déclinent de nombreuses manières,[197] dont les algorithmes de « clique-detection » implémentés dans les méthodes DISCO,[185] MPHIL[184] et MOE,[138] les algorithmes de recherche exhaustive (RAPID,[183] SCAMPI,[190] GALAHAD,[193] HipHop,[188] HypoGen[180] et PHASE[186]) ou les algorithmes génétiques (GAMMA[191] et GASP[192]). Parmi les premières approches proposées, MPHIL[184] défini un unique modèle de pharmacophore comme le plus petit nombre de points pharmacophoriques avec lequel chaque ligand a *p* points en commun, *p* étant défini par l'utilisateur. Des approches alternatives aux algorithmes de recherche de MCS ont également été implémentées, notamment dans LigandScout[187] qui utilise sa propre méthode de recherche et d'alignement de motifs pharmacophoriques.[198] L'importance de l'alignement des ligands dans la construction d'un pharmacophore est illustrée en ***Figure 28***.[198]

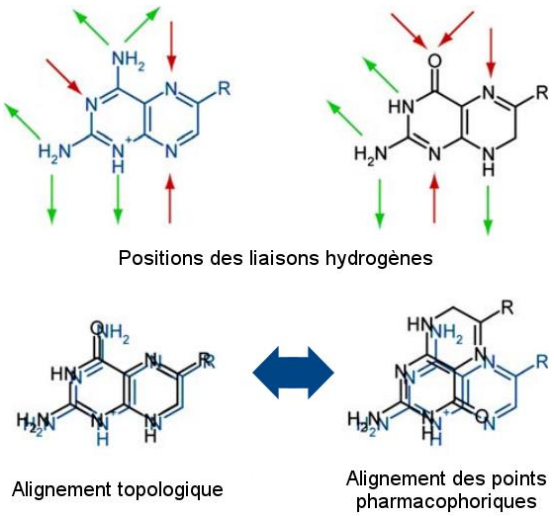


Figure 28. *Illustration de l'alignement du methotrexate et du dihydrofolate, deux inhibiteurs de la dihydrofolate reductase, de manière topologique et pharmacophorique. Dans cet exemple, seuls les points donneurs (vert) et accepteurs (rouge) de liaisons hydrogènes sont utilisés. Un alignement exclusivement topologique ne permet pas de percevoir l'alignement correct des propriétés pharmacophoriques.*[198]

2.3.3.2.5. Sélection des modèles de pharmacophores

A cette étape, plusieurs pharmacophores sont donc proposés par les méthodes. Ceux-ci regroupent différents nombres de points pharmacophoriques supposés sélectifs des molécules utilisées pour leurs constructions. Les pharmacophores pré-sélectionnés sont ensuite classés par pertinence grâce à des fonctions de score reposant sur le nombre et la qualité de la superposition des points pharmacophoriques, l'énergie des conformations moléculaires, leurs volumes de recouvrement ou la fréquence des points pharmacophoriques dans les différents pharmacophores proposés.[177,182,199]

En effet, parmi les points pharmacophoriques identifiés, certains ne seront pas suffisamment spécifiques et discriminants d'une interaction entre les ligands et la cible biologique puisqu'ils pourront être retrouvés dans un large nombre de molécules présentant des profils d'activité divers.[200] Les modèles comprenant de tels points pharmacophoriques devront donc être qualifiés comme peu pertinents par les fonctions de score et ignorés dans la suite des études.[182] Par exemple, les points d'interaction électrostatique sont généralement plus rares que les points hydrophobes et permettent la construction de pharmacophores plus sélectifs et pertinents.[200] Un candidat pharmacophore composé de trois points d'interaction électrostatique partagés par *n* ligands sera généralement plus pertinent qu'un pharmacophore composé de quatre points hydrophobes partagés par *n-1* ligands.[200]

Ainsi, les approches dites « weighted pharmacophores » utilisent une pondération des différents points pharmacophoriques en fonction du nombre de ligands qui les présentent.[182,199] Les candidats pharmacophores pourront ensuite être scorés selon la somme des poids établis (HipHop,[188] HypoGen[180] et PHASE[186]). Avec HipHop,[188] l'utilisateur défini également le nombre de ligands devant s'inscrire partiellement ou complètement dans chaque candidat pharmacophore et réalise un classement en conséquence. Les méthodes reposant sur des algorithmes génétiques calculent le score des candidats pharmacophores à chaque étape de sélection (« fitness evaluation ») durant la génération des pharmacophores (GAMMA[191] et GASP[192]).

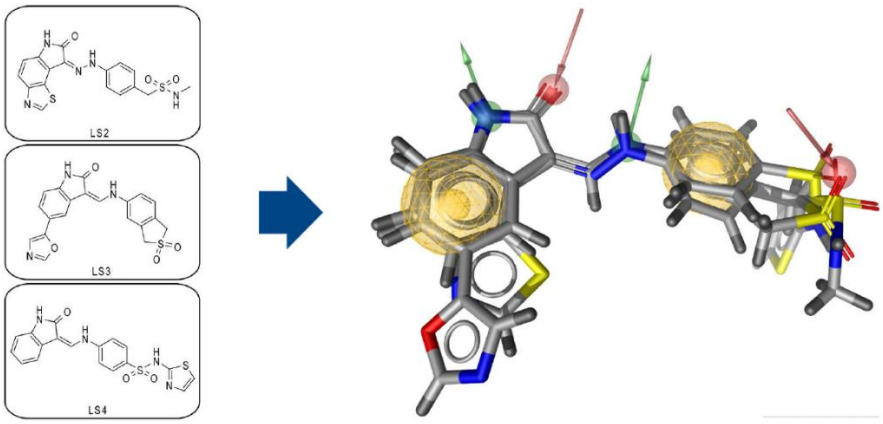


Figure 29. *Exemple de modèle pharmacophorique construit et représenté par LigandScout*[187] *à partir de trois inhibiteurs de la kinase dépendante des cyclines 2 (CDK2) dans leurs conformations liées au récepteur (vert : points pharmacophoriques donneurs de liaisons hydrogènes, rouge : accepteurs, sphères jaunes : points hydrophobes).*[189]

2.3.3.2.6. Criblage de chimiothèques

Après avoir obtenu un pharmacophore suffisamment discriminant et sélectif des ligands de la cible étudiée, celui-ci pourra être utilisé pour la recherche de nouveaux « hits » par le criblage de chimiothèques.[177,189] De la même manière que lors des étapes de construction des pharmacophores, la bonne prise en compte de la flexibilité des molécules est capitale dans ce type de criblages puisqu'ils reposent sur la définition de points pharmacophoriques contraints géométriquement.

La plupart des méthodes utilisent des bases de données de conformations moléculaires et réalisent un alignement des conformations sur le pharmacophore choisi afin d'évaluer la superposition de leurs points pharmacophoriques, dont HipHop,[188] HypoGen,[180] LigandScout,[187] PHASE[186] et MOE.[138] Des métriques telles que le RMSD (Root Mean Square Deviation, voir point 3.1) permettent d'évaluer l'écart entre les points pharmacophoriques des

molécules et du pharmacophore de référence à chaque modulation de l'alignement des conformations moléculaires.[189] L'approche alternative principale consiste à ajuster la conformation des molécules de la chimiothèques à la volée, ce qui permet potentiellement une meilleure couverture de l'espace de recherche conformationnel mais demande des temps de calcul plus importants.[177,189] Cette méthode est préférée par défaut dans PHASE.[186] Les approches d'alignement rigide et flexible peuvent également être combinées. Dans ce cas, après avoir réalisé une première étape d'alignement des conformations moléculaires rigides avec le pharmacophore, les meilleurs alignements seront utilisés pour une étape d'ajustement incluant la flexibilité des molécules (disponible avec HipHop,[188] HypoGen[180] et PHASE[186]).

Puisque ces étapes d'alignement représentent un coût calculatoire important, les logiciels utilisent généralement une étape de pré-filtrage de la chimiothèque qui a pour objectif d'identifier rapidement les composés qui ne pourront pas être alignés sur le pharmacophore.[189] La simple comparaison du nombre et du type des points pharmacophoriques entre chaque molécule et le pharmacophore permet de filtrer une chimiothèque efficacement et très rapidement.[201] Dans ce cas, pour chaque type de point pharmacophorique, seules les molécules qui présentent au minimum le même nombre de points pharmacophoriques que le pharmacophore de référence seront considérées. D'autres descripteurs 1D peuvent être utilisés de manière similaire.[202] Les méthodes d'empreinte 3D (voir point 2.3.2.3) telles que les « pharmacophore keys » sont également efficaces (***Figure 30***).[189]

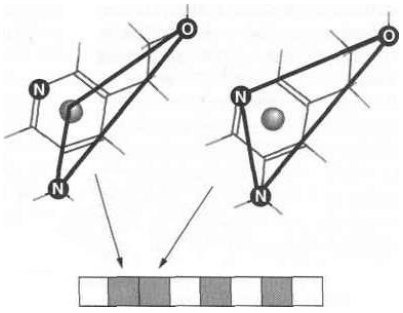


Figure 30. *Illustration de l'approche des empreintes 3D avec l'utilisation de triplets de points pharmacophoriques. Chaque bit de l'empreinte représente la présence ou l'absence de trois points pharmacophoriques avec des distances interatomiques prédéfinies par plages.*[189]

La plupart des programmes utilisent donc un pré-filtrage qui permet une grande rapidité d'exécution et entraîne parfois l'élimination d'une petite partie des molécules qui auraient pu s'inscrire dans le pharmacophore étudié, ce qui est considéré comme acceptable.[189] Inversement, des méthodes comme LigandScout[187] appliquent des filtres plus stricts qui garantissent que l'ensemble des composés éliminés ne pourront pas être alignés sur le

pharmacophore.[189] Les étapes d'alignement suivantes, très précises, permettent l'identification de nouveaux « hits » qui peuvent être très similaires comme très différents des ligands de référence. Les « hits » obtenus peuvent ainsi initier la découverte de nouvelles classes de composés thérapeutiques. Toutefois, si les méthodes pharmacophoriques permettent d'identifier les propriétés physico-chimiques importantes à l'activité d'une molécule sur sa cible, celles-ci ne permettent pas de quantifier leurs importances relatives.[33,189] Les méthodes QSAR peuvent être appliquées dans ce but.

2.3.4. Quantitative Structure-Activity Relationship (QSAR) « ligand-based »

Les méthodes QSAR « ligand-based » ont pour objectif d'établir des relations entre les propriétés physico-chimiques de ligands de référence avec leurs propriétés biologiques (activité expérimentale, toxicité, solubilité, etc.) grâce à l'usage de modèles statistiques. Les relations structure-activité établies par ces modèles permettront notamment de prédire les propriétés biologiques d'autres molécules « hits » ou « leads » obtenus lors de phases de criblage ou d'optimisation des candidats médicaments.[36,203]

Dimension	Structure du récepteur	Méthode de prédiction
1D-QSAR	Ignorée	Corrèle l'activité à des descripteurs moléculaires 1D constitutionnels (poids moléculaire, pKa, logP, nombres absolus et relatifs d'atomes, d'hétéroatomes, de liaisons, de cycles et de groupes fonctionnels d'un type donné, etc.)
2D-QSAR	Ignorée	Corrèle l'activité à des descripteurs topologiques sans prendre en compte la géométrie (indices de connectivité, de complexité, de ramification, quantiques, etc.)
3D-QSAR	Optionnelle	Corrèle l'activité à l'arrangement spatial des champs d'interactions non-liées
4D-QSAR	Optionnelle	QSAR-3D avec inclusion de plusieurs conformations des ligands
5D-QSAR	Ignorée	QSAR-4D avec représentation de différents modèles « induced-fit »
6D-QSAR	Ignorée	QSAR-5D avec inclusion de différents modèles de solvatation
7D-QSAR	Nécessaire	QSAR-6D avec prise en compte explicite de la structure du récepteur

Tableau 6. *Classification des méthodes QSAR selon la dimension des descripteurs utilisés. Les modèles 4D-QSAR et de dimension moindre sont les plus couramment appliqués.*[36,204]

De la même manière que dans les approches de recherche de similarité, les descripteurs utilisés pour la conception de modèles QSAR peuvent être généraux (1D-QSAR), topologiques (2D-QSAR) ou géométriques (3D-QSAR) (***Tableau 6***). D'autres méthodes dites 4D à 7D-QSAR,

plus récentes, se concentrent spécifiquement sur une meilleure prise en compte des différentes conformations et formes moléculaires (4D-QSAR[205]), de l'adaptation structurale de la cible induite par l'accommodation d'un ligand (« induced-fit », voir point 2.4.6.2) (5D-QSAR[206]), des interactions faibles impliquant le solvant (6D-QSAR)[207] et sur une utilisation optimale des données structurales de la cible (7D-QSAR).[204] Les modèles utilisant la structure du récepteur sont dits RD QSAR (Receptor Dependent) (voir point 2.4.4), en opposition aux modèles RI QSAR (Receptor Independent) qui utilisent uniquement les informations issues des ligands.[204]

2.3.4.1. RI 1D et 2D-QSAR

La première publication d'une équation qui peut être considérée comme un modèle QSAR est attribuable à Crum-Brown & Fraser en 1868, qui déclarent qu'il « ne peut y avoir de doute raisonnable quant à l'existence d'une relation entre l'action physiologique d'une substance et sa composition chimique ».[208] Le premier modèle 2D-QSAR appliqué à la biologie fut ensuite proposé en 1962 par Hansch & Muir, qui relient l'activité de régulateurs de la croissance de végétaux à des descripteurs d'hydrophobicité en utilisant l'équation de Hammett[209] et initient le développement de ces méthodes.[210] Les méthodes 1D et 2D-QSAR modernes représentent les propriétés physico-chimiques des composés au moyen de descripteurs constitutionnels, topologiques et/ou quantiques sans prendre directement en compte leur géométrie.[115] Le processus de conception de modèles QSAR est illustré en ***Figure 31***.

Wiener introduit dès 1947 le premier descripteur topologique lié à la connectivité moléculaire, qu'il applique avec succès à la prédiction des points d'ébullition de paraffines.[211] L'indice de Wiener d'une molécule est la somme de toutes ses liaisons covalentes entre toutes ses paires d'atomes, hormis les hydrogènes. De nombreux autres descripteurs topologiques tels que les indices de Randic,[212] Galvez,[213] Balaban,[214] Schultz,[215] Zagreb,[216] Szeged,[217] Hosoya[218] et Kier-Hall[219] reposent également sur la connectivité des molécules. L'indice de Randic peut être interprété comme une mesure de la surface accessible au solvant de la molécule,[212] tandis que l'indice de Galvez décrit les potentiels transferts de charge au sein de la molécule.[213] Les descripteurs de fragments moléculaires sont également largement utilisés par les méthodes QSAR. L'approche de Hansch permet d'estimer précisément le logP d'une molécule à partir de ses différents fragments,[220] tandis que les équations de Hammett[209] et Taft[221] décrivent les effets de polarisation et stériques des fragments sur la réactivité d'une molécule. Enfin, les descripteurs quantiques reposent sur les informations énergétiques, vibrationnelles et orbitalaires issues de la chimie quantique. Parmi les plus utilisés, les descripteurs de charges

atomiques partielles de Mulliken[222] et Gasteiger[223,224] décrivent les champs de forces électrostatiques moléculaires au moyen, respectivement, de calculs heuristiques du recouvrement des orbitales moléculaires et de calculs itératifs des contributions électrostatiques relevant d'une proximité topologique.

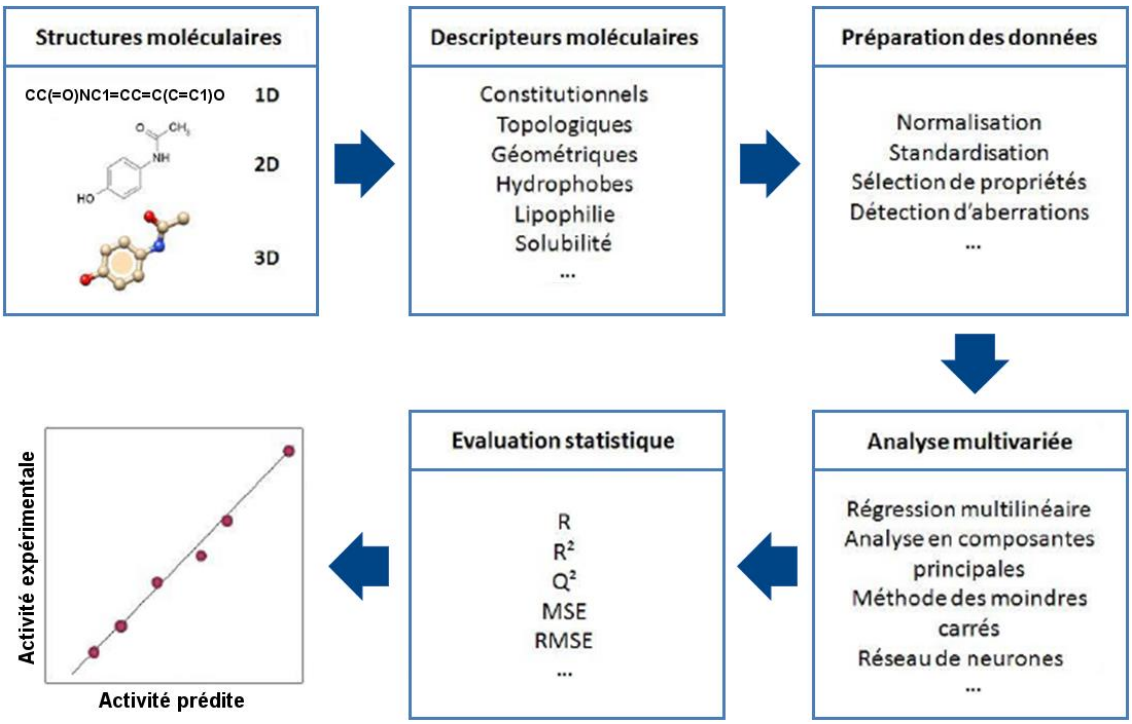


Figure 31. *Illustration du processus de conception de modèles QSAR aboutissant à un modèle statistique expliquant ici l'activité des ligands de référence à partir de leurs descripteurs moléculaires. Les modèles obtenus permettront ensuite de prédire l'activité d'autres composés.*[203]

Des logiciels comme MOLCONN,[225] Dragon,[226] MOE,[138] CODESSA,[227,228] SYBYL-X[132] et Pipeline Pilot[229] permettent de calculer efficacement jusqu'à 5270 descripteurs moléculaires sur de larges chimiothèques (Dragon version 7). Il est possible de construire des modèles QSAR efficaces en utilisant des centaines de descripteurs.[230] Cependant, plusieurs raisons motivent leur utilisation en nombre très réduit : (i) la précision et la valeur prédictive d'un modèle peuvent être accrues après exclusion de descripteurs redondants ou non-pertinents, (ii) des modèles plus simples permettent généralement une meilleure interprétation des résultats et (iii) certaines méthodes restreignent le nombre de descripteurs utilisables du fait des temps calculatoires associés.[36,203,230] De nombreuses méthodes statistiques peuvent être mises en œuvre pour la construction de modèles QSAR, parmi les plus accessibles : les régressions linéaires multiples (Multiple Linear Regression ou MLR) ou de moindres carrés (Partial Least Squares ou PLS), les arbres de décision, l'analyse en composantes principales (Principal

Component Analysis ou PCA) et la méthode des plus proches voisins.[231] Les méthodes des forêts aléatoires,[232] des machines à vecteurs de support[77,78] ou des réseaux de neurones[77,79,80] peuvent également être appliquées à ces fins.[231] Plus récemment, des techniques de « deep learning » ont également été appliquées à la prédiction de la solubilité de molécules « drug-like ».[233]

2.3.4.2. RI 3D-QSAR

Les approches 3D-QSAR regroupent toutes les approches QSAR qui corrèlent les propriétés biologiques de ligands de référence à des descripteurs dérivés directement de leurs représentations spatiales.[36] En 1979, Cramer & Milne proposent une première approche 3D-QSAR utilisant une grille tridimensionnelle pour la représentation des champs d'interactions de ligands de référence de manière discrète.[234] Les développements successifs de cette approche, couplée à l'utilisation de modèles PCA et PLS pour l'analyse des champs d'interaction et la prédiction de propriétés biologiques, aboutissent aux prédécesseurs des logiciels 3D-QSAR actuels : DYLOMMS en 1983 (DYnamic Lattice-Oriented Molecular Modeling System),[235] puis CoMFA en 1988 (Comparative Molecular Field Analysis).[236] Plusieurs méthodes ont désormais prouvé leur efficacité pour la conduite d'études 3D-QSAR, parmi lesquelles CoMFA,[236] CoMSIA,[237] COMPASS,[238] GERM,[239] CoMMA[240] et SoMFA.[241] Des logiciels commerciaux comme LigandScout,[187] MOE,[138] HypoGen[180] et PHASE[186] permettent également de concevoir des modèles 3D-QSAR à partir de pharmacophores. De la même manière que pour la construction de modèles pharmacophoriques, l'alignement pertinent des conformations moléculaires est crucial au succès d'une étude 3D-QSAR (voir point 2.3.3.2.2).[36]

Lors d'une étude CoMFA,[236] les ligands de référence sont alignés et placés dans une grille régulière d'une résolution par défaut de 2 Å. Des sondes sont positionnées à chaque point de cette grille et représentent la présence hypothétique de carbones sp^3 de charge +1.0 (d'autres types de sondes peuvent également être utilisés). Les potentiels d'interactions stériques (potentiel de Lennard-Jones 12-6)[242] et électrostatiques (Coulomb) sont ensuite calculés entre chaque sonde et chacun des ligands de référence.[236] Les données d'interaction ainsi obtenues permettent la construction d'un modèle de régression PLS comprenant un grand nombre de coefficients (***Figure 32***). Afin de faciliter la compréhension de ce modèle, les résultats sont généralement représentés sous la forme d'aires de contour indiquant les régions favorables et défavorables aux interactions stériques et électrostatiques (***Figure 33***).[236]

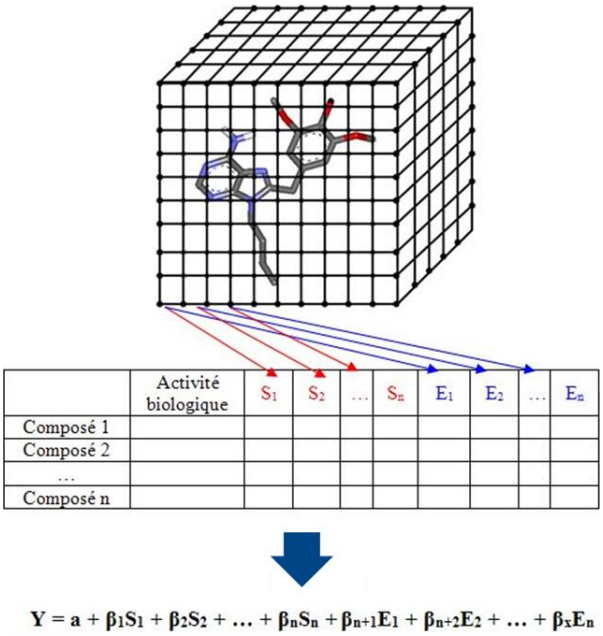


Figure 32. *Illustration du processus d'une étude CoMFA.*[236] *Après alignement des ligands, les énergies d'interaction stériques (S) et électrostatiques (E) sont calculées à chaque point de la grille et pour chaque ligand de référence. Un modèle de régression PLS est ensuite mis en œuvre pour corréler ces valeurs d'énergies d'interaction aux valeurs d'activité des ligands.*

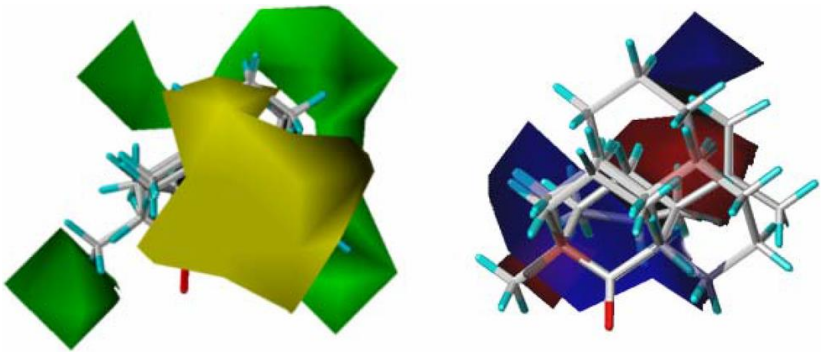


Figure 33. *Représentation d'alignements de ligands réalisés lors d'une étude CoMFA*[236] *avec visualisation des zones de contour des contributions stériques favorables (vert) et défavorables (jaune) à l'activité des ligands. Les zones de contour des contributions électrostatiques favorables (rouge) ou défavorables (bleu) à l'activité des ligands sont représentées à droite.*[243]

Cependant, puisque la méthode CoMFA[236] repose sur l'utilisation d'un espace discret, il est fréquent que les potentiels de Lennard-Jones et de Coulomb prennent des valeurs nulles (sonde isolée) ou maximales (sonde à l'intérieur du volume d'un ligand de référence, des valeurs maximales des potentiels sont alors utilisées), ce qui nuit à la précision des modèles.[244] Pour pallier ce problème, Klebe et al. proposent la méthode CoMSIA (Comparative Molecular Similarity Indices Analysis),[237] qui diffère de CoMFA[236] principalement par son estimation des potentiels d'interaction au moyen d'indices de similarité adaptés de l'algorithme SEAL (Steric and Electrostatic ALignment).[245] Les ligands de référence sont alignés puis placés dans une

grille régulière d'une résolution par défaut de 1 Å. Les indices de similarité sont ensuite calculés, toujours au moyen de sondes à chaque point de la grille, avec une fonction gaussienne approximant les potentiels de Lennard-Jones et Coulomb (***Figure 34***). Cette définition permet de prendre en compte les propriétés stériques, électrostatiques et hydrophobes des molécules de manière plus progressive et résout le problème des valeurs de potentiels limites trop fréquentes.[237] Comme pour la méthode CoMFA,[236] un modèle de régression PLS est ensuite mis en œuvre, cette fois à partir des indices de similarité.[237]

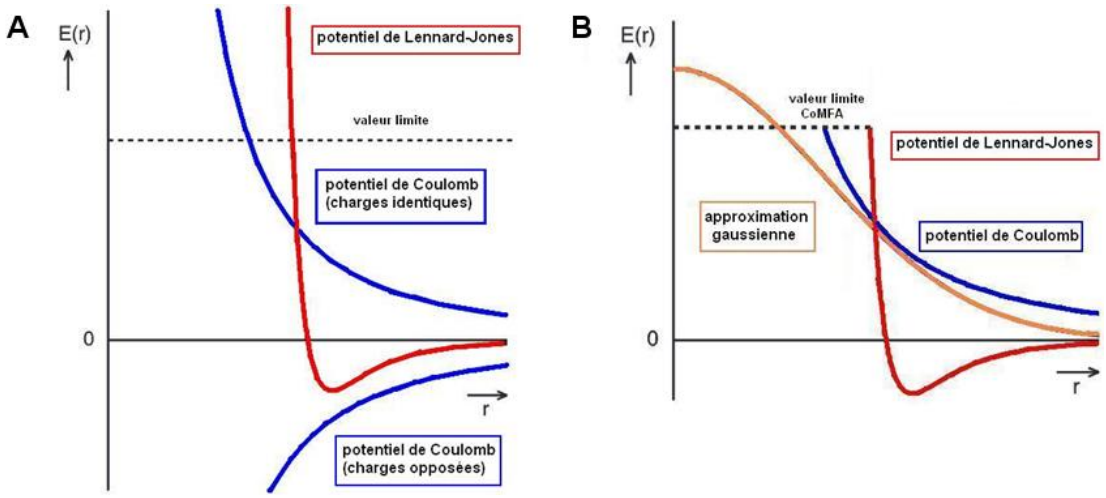


Figure 34. *(A) Courbes des potentiels de Lennard-Jones (rouge) et Coulomb (bleu) utilisées dans les études CoMFA.*[236] *(B) La fonction gaussienne utilisée pour le calcul des indices de similarité (orange) approxime les potentiels de Lennard-Jones et Coulomb de manière plus progressive.*[246]

Parmi les solutions commerciales plus récentes, de nombreux logiciels initialement conçus pour la construction de modèles de pharmacophores ont ensuite été adaptés pour la réalisation d'études 3D-QSAR, dont LigandScout,[187] MOE[138] et PHASE.[186] Ce dernier propose deux approches pour la construction de modèles 3D-QSAR, soit à partir de l'ensemble des atomes des ligands de référence, soit à partir de leurs points pharmacophoriques.

L'approche atomique représente les caractéristiques physico-chimiques localement en utilisant des sphères de van der Waals avec recouvrement, suivant six catégories : donneur ou accepteur de liaisons hydrogènes, charge électrostatique positive ou négative, atome hydrophobe ou autre.[186] Les ligands alignés sont placés dans une grille d'une résolution de 1 Å, dont le centre de chaque cube peut être inclus, ou non, dans les sphères de van der Waals définies par les atomes de chaque ligand (***Figure 35***).[186] Pour chaque ligand, un cube peut ainsi être occupé par plus d'un atome et détecter plus d'une caractéristique physico-chimique. Des empreintes binaires sont ensuite construites pour chaque ligand, dont la taille unique est définie de manière

dynamique, le nombre de bits étant donné par l'occupation de la grille pour l'ensemble des ligands (un seul bit à 0 pour un cube qui ne recouvre aucun atome des ligands et jusqu'à six bits pour un cube recouvrant les six types physico-chimiques parmi les différents atomes des ligands). Cette représentation permet de traiter chaque bit comme une variable indépendante pour la construction de modèles 3D-QSAR. L'application de régressions PLS permet ainsi d'obtenir des coefficients pour chaque bit, ce qui facilite l'interprétation des résultats et l'identification de caractéristiques physico-chimiques liées, ou non, à l'activité des ligands de référence. Le nombre de coefficients de la régression PLS est limité, au maximum, à 1/5^{e} du nombre de ligands de référence.[186]

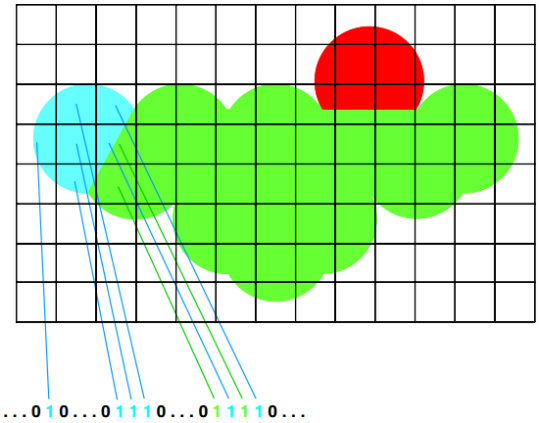


Figure 35. *Illustration en 2D de la méthode utilisée par PHASE*[186] *pour la construction d'empreintes binaires à partir de sphères de van der Waals représentant les atomes de chaque ligand. Chaque cube peut donner lieu à la définition de un à six bits. La taille de chaque empreinte et la signification de chaque bit reste constante pour l'ensemble des molécules.*[186]

L'approche pharmacophorique proposée par PHASE[186] pour la construction de modèles 3D-QSAR est similaire à son approche atomique concernant la recherche de l'empreinte de chaque ligand et l'élaboration statistique du modèle. Les différences majeures sont l'utilisation des points pharmacophoriques des ligands à la place de sphères de van der Waals atomiques et la différenciation des atomes aromatiques ou hydrophobes. Cette approche s'appuie donc sur la construction préalable d'un pharmacophore (voir point 2.3.3.2.3).

2.3.4.3. RI 4D-QSAR

Les approches 4D-QSAR consistent en une extension des méthodes 3D-QSAR afin de traiter chaque molécule comme un ensemble de différentes conformations, orientations, formes tautomères, stéréoisomères ou états de protonation.[247] La première méthode RI 4D-QSAR fut proposée en 1997 par Hopfinger et al.[205] Celle-ci repose sur l'utilisation de dynamiques moléculaires pour la génération des différentes conformations des ligands, dont leurs

probabilités d'observation sont ensuite calculées avec l'équation de Boltzmann. Une grille d'occupation est utilisée pour détecter et quantifier la présence des caractéristiques physico-chimiques, similairement à ce qui est réalisé par le logiciel PHASE,[186] cette fois avec des ensembles de conformations moléculaires alignées et sélectionnées de manière iterative.[205] La présence d'une caractéristique physico-chimique en un cube de la grille est alors pondérée par la probabilité d'observation d'un état conformationnel, permettant une meilleure prise en compte de la flexibilité moléculaire dans le développement du modèle QSAR.[205]

2.3.4.4. RI 5D et 6D-QSAR

Les approches dites 5D et 6D-QSAR, implémentées dans le logiciel Quasar,[248] ont été proposées par Vedani & Dobler en 2001 et 2005.[206,207] L'approche 5D-QSAR est caractérisée par la simulation de différents modèles d'adaptation du récepteur lors de l'accommodation d'un ligand (« induced-fit »).[206,249] Pour ce faire, Quasar génère plusieurs modèles hypothétiques du récepteur à partir des surfaces de van der Waals des conformations des ligands de références, typées localement selon leurs propriétés physico-chimiques. Une « surface moyenne » du récepteur est ensuite calculée puis adaptée progressivement à chacun des ligands grâce à un algorithme génétique, dans la limite d'une déviation atomique de 4 Å (***Figure 36***).[206] Les biais des méthodes 4D-QSAR sont ainsi réduits grâce à la détermination du scénario « induced-fit » le plus pertinent pour chacun des ligands, engageant une modélisation hypothétique du récepteur et reposant sur un alignement dynamique des ligands.[206] L'approche 6D-QSAR prend également en compte différents modèles de solvatation dans la simulation des complexes récepteur-ligand.[207]

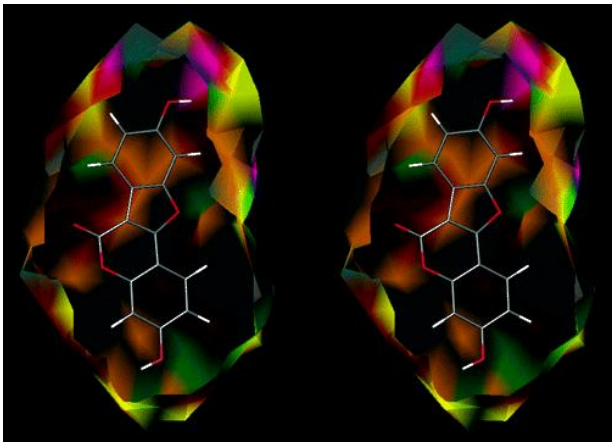


Figure 36. *Vue stéréo de la surface hypothétique d'un récepteur aux œstrogènes construite à partir du coumestrol (vert : donneur de liaisons hydrogènes, jaune : accepteur, violet : accepteur ou donneur (« flip-flop »),*[250] *marron clair : hydrophobe chargé positivement, marron foncé : hydrophobe chargé négativement, gris : hydrophobe non-chargé).*[207]

Grâce au développement de programmes accessibles et intuitifs, notamment commerciaux, l'intérêt pour les approches QSAR multidimensionnelles a largement augmenté au cours des années 2000 (***Figure 37***). Cependant, ces méthodes ne sont pas applicables à de larges chimiothèques. Leur principal intérêt réside en la prédiction des propriétés biologiques de « hits » ou « leads » obtenus lors de phases de criblage ou d'optimisation des candidats médicaments.[36,203]

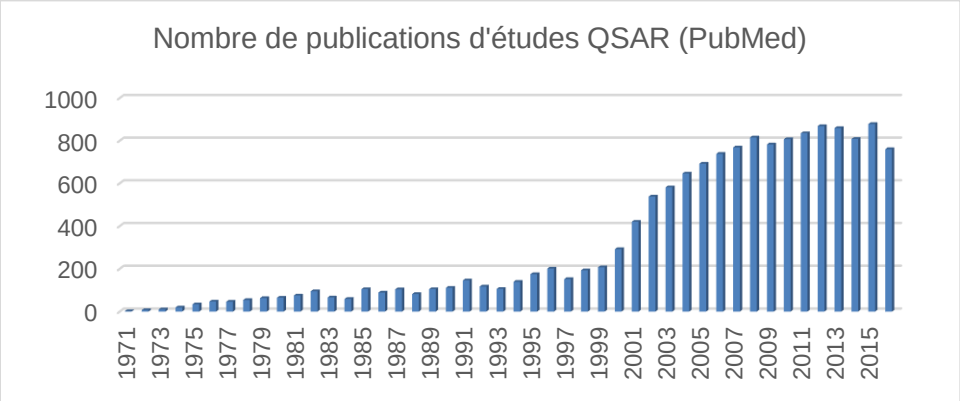


Figure 37. *Nombre de publications retrouvées dans la base de données PubMed*[251] *avec le mot-clé unique « QSAR » entre 1971 et 2016. Les approches CoMFA*[236] *et CoMSIA*[237] *furent publiées en 1988 et 1999, respectivement.*

2.3.5. Succès du criblage virtuel « ligand-based »

Ainsi, les méthodes de criblage virtuel « ligand-based » permettent de rationaliser le processus de recherche de nouveaux « hits » ou « leads » grâce à la connaissance préalable de ligands ayant une activité sur la cible thérapeutique. La mise en œuvre de ces méthodes a permis la découverte de nouveaux composés actifs lors de différentes études, qu'elles reposent sur des approches QSAR, pharmacophoriques ou de recherche de similarité (***Tableau 7***).

Les méthodes de criblage virtuel « structure-based » et « ligand-based » ont notamment contribué à l'autorisation et la mise sur le marché de la première thérapie contre les cancers métastatiques du poumon à grandes cellules (environ 80% des cancers du poumon) pour lesquels le récepteur du facteur de croissance épidermique (EGFR) est surexprimé (environ 10% des cas).[252] L'utilisation du gefitinib, commercialisé par AstraZeneca & Teva sous le nom Iressa®, fut approuvée dans ce contexte en 2015 suite à l'identification d'un premier « hit » inhibiteur d'EGFR en 1994 (***Figure 38***).[252,253] L'application du gefitinib pour le traitement de patients atteints d'autres types de tumeur solides a également été explorée. Cet inhibiteur a largement contribué à une meilleure compréhension des mécanismes de signalisation cellulaire impliquant EGFR.[252]

Approche	Méthode	Composés	Référence
Empreintes 2D	Empreintes CATS	Antagonistes du canal calcique de type T	Schneider et al.[172]
Empreintes 2D	Empreintes CATS	Inhibiteurs de la glycogène synthase kinase 3 (GSK3)	Naerum et al.[254]
Pharmacophores 3D	Catalyst	Antagonistes de l'antigène α4β1	Singh et al.[255]
Pharmacophores 3D	Catalyst	Antagonistes du récepteur de l'urotensine II	Flohr et al.[256]
Pharmacophores 3D	GALAHAD	Inhibiteurs de la formation de l'hétérodimère Myc-Max	Mustata et al.[257]
Similarité de forme	ROCS	Inhibiteurs de ZipA	Rush et al.[258]
Similarité de forme	ROCS	Antagonistes du récepteur cannabinoïde 1 (CB1)	Boström et al.[259]
Empreintes 2D, similarité de forme et pharmacophores 3D	Empreintes DayLight, ROCS, ALMOND	Agonistes du récepteur couplé aux G-protéines 30 (GRP30)	Bologa et al.[260]
2D-QSAR, 3D-QSAR et similarité de forme	Surflex-Sim, ROCS, CoMFA	Inhibiteurs de la cruzaïne et de la cathepsine L	Freitas et al.[261]
2D-QSAR et pharmacophores 3D	Catalyst	Inhibiteurs de HSP90α	Al-Sha'er et al.[262]

Tableau 7. *Quelques exemples de succès dans la découverte de nouveaux composés actifs à objectif thérapeutique pour lesquels la mise en œuvre de criblages virtuels « ligand-based » a joué un rôle important.*

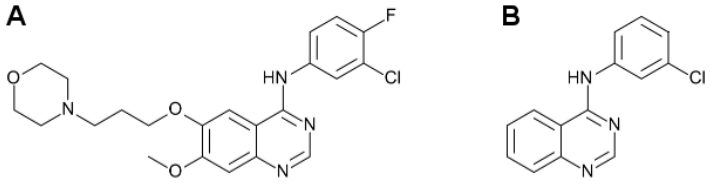


Figure 38. *Structure du gefitinib (A) et du premier « hit » identifié en 1994 qui servit de base à son développement (B).*[252,253]

2.4. Criblage virtuel « structure-based »

Le criblage virtuel « structure-based » repose sur la connaissance de la structure de la cible thérapeutique pour l'identification de composés actifs. La structure 3D d'une protéine peut être déterminée principalement par les méthodes de cristallographie à rayons X ou de résonance magnétique nucléaire (RMN).[196] Les structures obtenues par cristallographie à rayons X ont l'avantage d'être d'une bonne résolution (78% de structures ont une résolution inférieure à 2.5 Å).[196] Cependant, cette technique décrit un état cristallin de la protéine, qui peut parfois différer largement de son état *in vivo* en phase aqueuse. Inversement, la RMN permet d'obtenir un

ensemble de conformations illustrant la dynamique et la flexibilité d'une protéine en phase aqueuse. La Protein Data Bank (PDB)[196] regroupe et mets à disposition ces structures, dont le nombre augmente de manière stable et exponentielle depuis 1990 (***Figure 39***). Cette base de données comporte aujourd'hui plus de 110.000 structures cristallographiques et croît de manière stable, doublant de volume tous les 4 à 6 ans.[263] La plus grande disponibilité des structures a ainsi permis l'essor des méthodes de criblage « structure-based ».

Lorsque la structure de la cible thérapeutique étudiée n'est pas disponible, il est également possible de recourir à la modélisation par homologie.[33,44,264] Une ou plusieurs protéines dont la structure est résolue et qui présentent une bonne homologie de séquence avec la cible serviront de support pour en construire un modèle. Cette méthode permet donc d'obtenir rapidement et à moindre coût l'élément clé nécessaire à la mise en œuvre des méthodes de criblage « structure-based » de tous types. Ainsi, près de 25% des études de docking publiées entre 2000 et 2009 auraient utilisé une structure construite par homologie de séquence.[44]

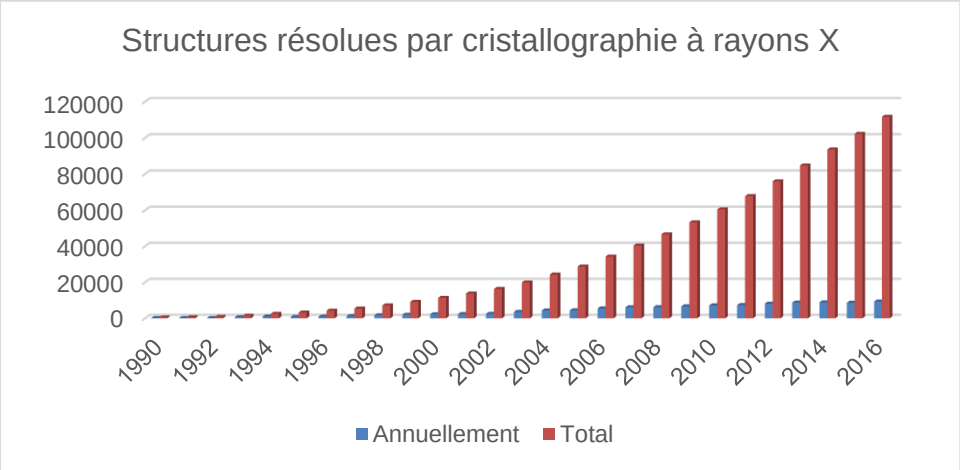


Figure 39. *Nombre de structures résolues par cristallographie à rayons X et déposées dans la PDB entre 1990 et 2016. La taille de cette base de données augmente de manière stable et exponentielle.*[196]

Après avoir obtenu une structure de la cible thérapeutique, plusieurs approches « structure-based » peuvent être employées pour l'identification de composés actifs : les méthodes de docking, la construction de modèles pharmacophoriques, la conception de ligands *de novo*, ou les méthodes RD-QSAR. Chacune de ces méthodes requiert la définition préalable d'un site de liaison sur la cible thérapeutique.

2.4.1. Sélection d'un site de liaison

Lorsque la structure de la cible a été déterminée expérimentalement, celle-ci peut être co-cristallisée avec un ligand. Auquel cas, le site de liaison du ligand est connu et l'on pourra définir un espace de recherche centré sur celui-ci afin de procéder à la mise en œuvre des méthodes. Dans le cas contraire, des outils de prédiction de sites de liaison peuvent être mis en œuvre.

2.4.2. Prédiction du site de liaison

Les outils de prédiction de poches « druggables » permettent d'identifier des sites de liaisons potentiels d'un ligand sur la cible. Ces outils peuvent être classés suivant trois catégories : les méthodes basées sur la connaissance, les méthodes basées sur la géométrie et les méthodes basées sur l'énergie d'interaction.[265]

2.4.2.1. Prédiction basée sur la connaissance

Les outils de prédiction basés sur la connaissance reposent sur les informations extraites de protéines qui possèdent une séquence ou une structure similaire à la cible. En 1965, Zuckerkandl & Pauling formulent l'hypothèse que les sites de liaison d'une protéine soient conservés au cours de l'évolution de sorte à maintenir l'interaction avec leurs ligands naturels et les fonctions biologiques qui en dépendent.[266] La recherche de similarité de séquence entre la cible et d'autres protéines dont les sites d'interactions sont connus permettra donc de prédire des poches « druggables » potentielles. Cette méthode est utilisée par différents outils d'estimation de poches (Consurf,[267] FREPS,[268] etc.) qui considèrent les séquences homologues à la cible et identifient des positions très conservées.

L'accroissement du nombre de complexes protéine-ligand co-cristallisés disponibles dans la PDB et le développement d'outils de modélisation de structure ont permis de mettre en place des outils de recherche de poches basés sur la similarité structurale. Il existe trois méthodes d'alignement structural visant à détecter des sites de liaison sur des protéines similaires à la cible, reposant sur : un alignement global (FINDSITE[269]), un alignement local (ProBiS[270]) et celles combinant les deux approches (COFACTOR[271]). Différentes bases de données regroupent les sites de liaisons identifiés expérimentalement et peuvent être utilisées comme bases de référence (CavBase,[272] PINTS,[273] SiteEngines,[274] eF-site,[275] ProFunc,[276] SitesBase,[277] Catalytic Site Atlas[278]). A titre d'exemple, ProBiS utilise un algorithme d'alignement structural

local basé sur la théorie des graphes qui permet d'identifier des sites similaires entre la cible et une protéine de référence (***Figure 40***).

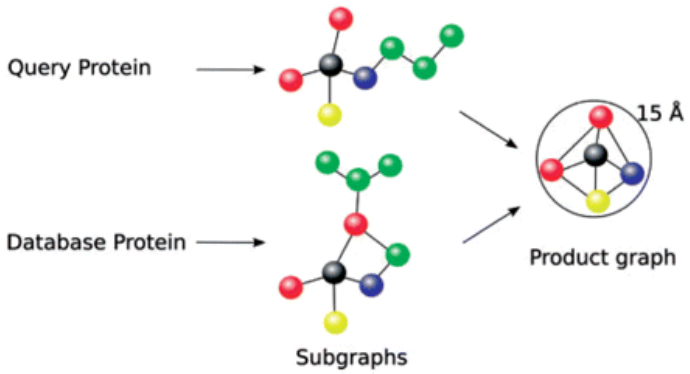


Figure 40. *Illustration du fonctionnement du logiciel ProBis. Des labels sont assignés aux différents groupes fonctionnels de chaque protéine. Des graphes de la cible et de chaque protéine de la banque de données utilisée sont générés, puis comparés pour en déduire un graphe commun. Ce graphe représente la superposition des groupes fonctionnels des deux protéines comparées.*[270]

2.4.2.2. Prédiction basée sur la géométrie

Les sites de liaisons de petites molécules sont généralement associés à des cavités présentes à la surface des protéines. Différentes méthode d'estimation de poches cherchent donc à localiser ces cavités en analysant la géométrie de la surface protéique. Pour cela, la plupart des outils utilisent une grille tridimensionnelle visant à définir la surface moléculaire (MOLCAD,[279] LIGSITE,[280] VolSite[281]). D'autres outils font appel à des sondes sphériques (FPocket,[282] Surface Racer[283], SurfNet[284]). Le programme FPocket utilise l'approche des sphères alpha, qui consiste à utiliser une sonde sphérique en contact avec 4 atomes de la protéine pour refléter la courbure locale à la surface d'une protéine (***Figure 41***).[282,285]

Une tesselation de Voronoï[286] est réalisée pour obtenir la surface moléculaire de la protéine, puis les sphères alpha sont positionnées à chaque sommet du maillage. Le rayon de chaque sphère alpha est ensuite adapté en fonction de la distance aux atomes de la protéine, permettant de définir un tétraèdre grâce à 4 points de contact avec celle-ci. Un filtre est ensuite appliqué pour omettre les sphères de petite taille, qui correspondent aux zones inaccessibles au solvant, et celles de grandes tailles, qui correspondent aux zones trop exposées. Les sphères sont ensuite regroupées de sorte chaque à ce que chaque groupe corresponde à un site de liaison potentiel (***Figure 41F***).[285]

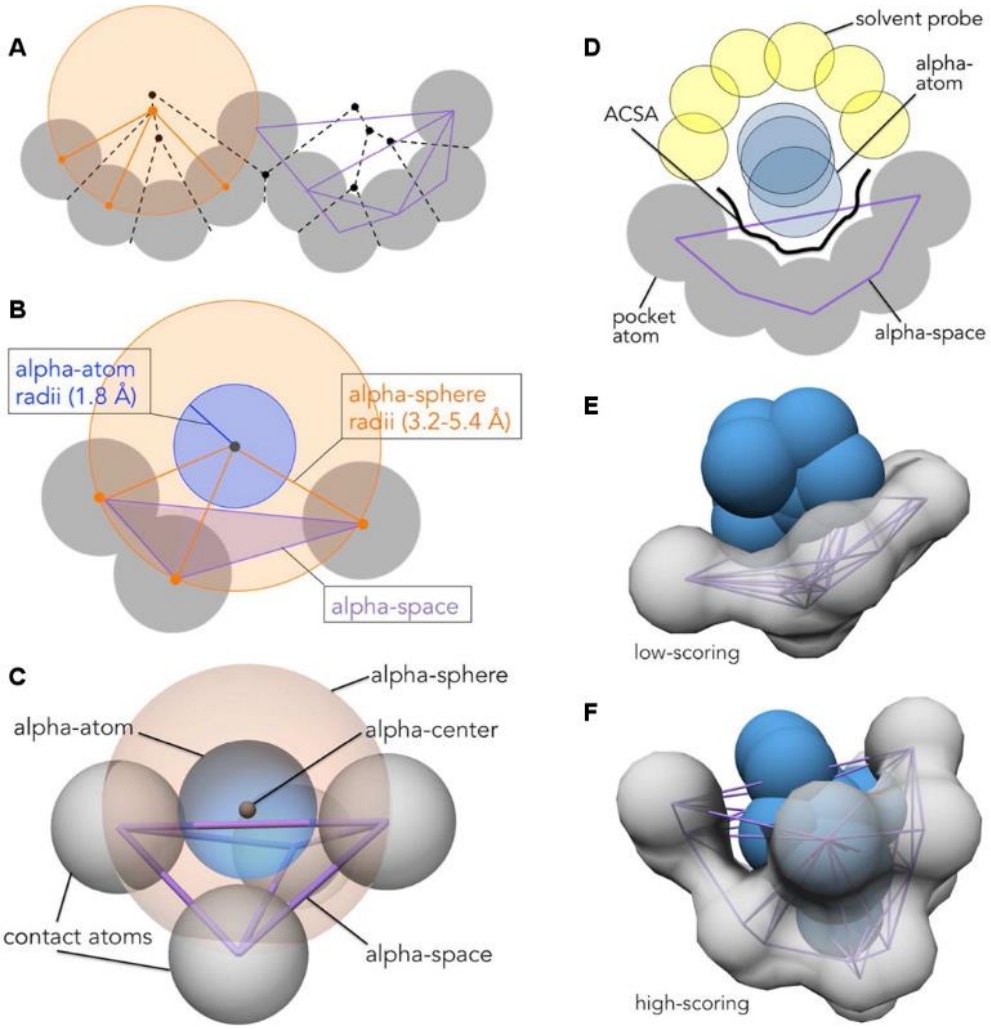


Figure 41. *Illustration du fonctionnement de FPocket. (A) Représentation schématique 2D : sphère alpha (orange), maillage de Voronoï (noir), atomes de la protéine (gris) et triangulation d'une poche (violet). (B) Système alpha vu en 2D avec son atome alpha (bleu). (C) Système alpha vu en 3D. (D) Définition du score des poches grâce à des sondes type solvant (jaune) et des atomes alpha pour le calcul d'un « Alpha-cluster Contact Surface Area » (ACSA, noir). (E) Poche exposée, score bas. (F) Poche profonde, score haut.*[285]

2.4.2.3. Prédiction basée sur l'énergie d'interaction

Les méthodes de prédiction basées sur des critères énergétiques utilisent des sondes (groupement méthyle ou hydroxyle par exemple) pour définir, à la surface d'une protéine, les points d'interaction pouvant être associés à une poche « druggable ». Parmi les différents programmes employant cette approche (AutoSite,[287] FTMap,[45] FTSite[288]), FTSite utilise 16 sondes distinctes et identifie les zones d'interaction favorables en utilisant une fonction d'évaluation empirique (***Figure 42***). Les sondes de même type sont regroupées selon des critères de distance, puis les zones de chevauchement entre les différents groupes sont identifiées et considérées comme sites de liaison potentiels (***Figure 43***). Il existe également des

approches consensus qui combinent plusieurs des méthodes précédemment citées. Par exemple, MetaPocket[289] combine trois outils basés sur la géométrie (LIGSITE,[280] PASS,[290] Surfnet[284]) à un outil basé sur l'énergie d'interaction (Q-SiteFinder[291]) dans sa recherche des poches « druggables ».

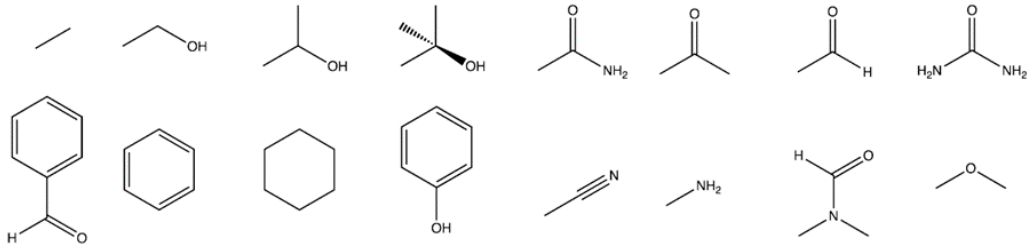


Figure 42. *Les 16 petites molécules utilisées comme sondes par FTSite présentent différents caractères hydrophobes, polaires et aromatiques. Cette sélection permet une bonne caractérisation des sites de liaison, qui présentent dans la majorité des cas de forts potentiels d'énergie.*[288]

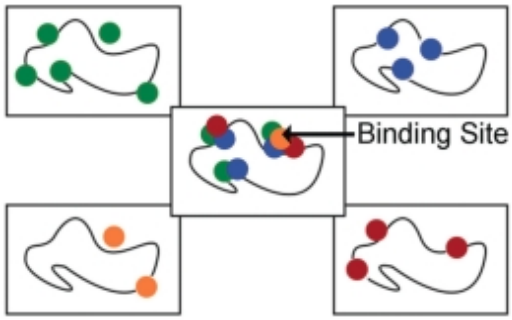


Figure 43. *FTSite défini des régions impliquées dans différents types d'interactions grâce à l'usage de sondes variées. En combinant ces différentes cartes d'interaction de la surface de la protéine, il est possible de scorer le potentiel des différentes régions à constituer des poches « druggables ». Les couleurs représentent les différents types d'interaction à la surface de la protéine.*[288]

2.4.3. Modèles pharmacophoriques « structure-based »

Comme décrit précédemment, les modèles pharmacophoriques « ligand-based » permettent l'identification de nouveaux composés actifs à partir de la connaissance de ligands de la cible étudiée (voir point 2.3.3). Cependant, lorsque la structure de la cible a été résolue, l'utilisation des informations de forme, de volume ou physico-chimiques du récepteur peut rendre les méthodes pharmacophoriques plus puissantes.[177,292] Les modèles pharmacophoriques « structure-based » peuvent être construits soit à partir de la structure du récepteur, soit à partir d'un complexe récepteur-ligand.[177,292]

2.4.3.1. Méthodes basées sur le récepteur

Les approches pharmacophoriques basées sur le récepteur requièrent, en premier lieu, la définition d'un site de liaison à la surface de la cible (***Figure 44A***). Celui-ci doit être identifié préalablement par l'utilisateur, par exemple grâce à la mise en œuvre de méthodes de prédiction (voir point 2.4.2). Expérimentalement, un site de liaison peut être validé par mutagénèse dirigée : si la mutation d'un résidu affecte l'affinité d'un composé avec la cible étudiée, alors ce résidu doit être impliqué dans l'interaction et faire partie du site de liaison.[293,294]

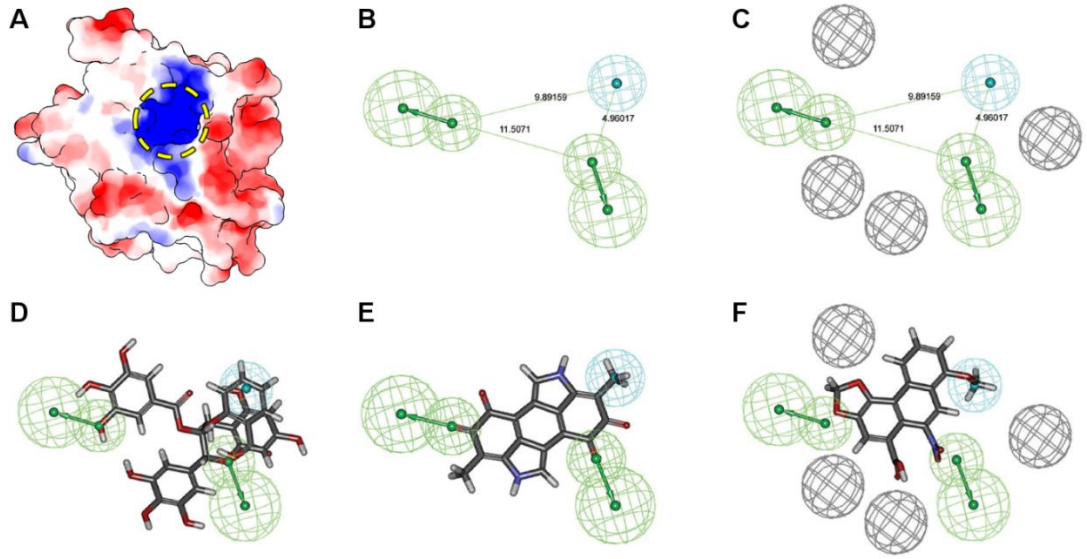


Figure 44. *Illustration du processus de construction et d'utilisation d'un modèle pharmacophorique « structure-based ». (A) Représentation du potentiel électrostatique en surface de la cystéine protéase RTX de* Vibrio cholerae *(PDB 3eeb*[196]*). Une cavité à fort potentiel électropositif (bleu) est identifiée (rouge : potentiel électronégatif, blanc : potentiel neutre). (B) Exemple de pharmacophore sélectionné à partir de la carte d'interaction négative du récepteur (vert : liaisons hydrogènes, bleu : point hydrophobe). (C) Ajout de volumes d'exclusion afin de restreindre le volume accessible aux molécules qui seront alignées sur la pharmacophore. (D,E) Exemples d'alignements de composés sur le pharmacophore proposé. (F) Ajout de volumes d'exclusion au pharmacophore : l'alignement de la molécule est contraint et affiné selon, par exemple, l'encombrement du récepteur.*[295]

Une fois le site de liaison identifié, celui-ci peut être analysé de différentes manières afin de construire une image négative du récepteur représentant les caractéristiques pharmacophoriques de ses ligands potentiels. Une propriété de type donneur de liaisons hydrogènes identifiée dans le site de liaison sera donc complétée par une propriété de type accepteur sur la carte d'interaction (***Figure 44B***). Parmi les premières méthodes proposées dans les années 1990, le logiciel LUDI[296,297] identifie les groupes hydrophobes, aromatiques, accepteurs ou donneurs de

liaisons hydrogènes grâce à des sondes représentant les groupes fonctionnels, afin de proposer différents pharmacophores complémentaires au site de liaison. Le logiciel BOSS[298] propose un couplage de l'algorithme de LUDI[296,297] à l'utilisation de simulations de dynamique moléculaire pour la prise en compte de la flexibilité du récepteur.[298–300] Les logiciels commerciaux procèdent selon des méthodes similaires (LigandScout,[187] MOE,[138] PHASE,[186] HipHop,[188] etc.).

Quelle que soit la méthode utilisée, la carte d'interaction obtenue pour le récepteur comporte généralement un grand nombre de points pharmacophoriques, particulièrement lorsque le site de liaison défini est vaste ou lorsqu'il peut accommoder une variété de ligands selon différents modes de liaison.[177] Pour pallier ce problème, les points pharmacophoriques de même type et spatialement proches peuvent être regroupés et représentés par un point unique. Par exemple, avec le logiciel HipHop,[188] le barycentre du groupe de points pharmacophoriques est calculé, puis le point plus proche du barycentre est sélectionné. Une autre approche consiste à sélectionner manuellement les points pharmacophoriques d'intérêt à partir de la carte d'interaction proposée.[293,294] Les logiciels commerciaux, notamment, disposent d'une interface qui facilite cette sélection. Lorsque des ligands de la cible sont connus, sans que des données de co-cristallisation soient disponibles, ceux-ci peuvent également être utilisés pour guider la sélection des points pharmacophoriques.[293,294] Différents pharmacophores sont alors générés, comportant chacun un plus petit nombre de points pharmacophoriques, afin de représenter l'ensemble de leurs combinaisons potentielles. La capacité de chacun de ces pharmacophores à accommoder les ligands connus sera ensuite évaluée afin d'identifier les modèles les plus pertinents.[293,294]

Enfin, les pharmacophores peuvent être complétés et affinés par l'ajout de volumes d'exclusion, destinés à restreindre l'espace accessible aux molécules qui seront alignées sur le pharmacophore (***Figure 44C***). Les volumes d'exclusion peuvent être définis de manière explicite à partir des données structurales du récepteur ou positionnés manuellement par l'utilisateur. Les pharmacophores « structure-based » ainsi construits peuvent ensuite être utilisés pour le criblage de molécules, de la même manière qu'avec des pharmacophores « ligand-based » (voir point 2.3.3.2.6) (***Figure 44D-F***).

2.4.3.2. Méthodes basées sur le complexe récepteur-ligand

Lorsque des données de co-cristallisation protéine-ligand sont disponibles pour la cible étudiée, l'analyse des complexes récepteur-ligand permet d'obtenir des informations précises sur le

mode d'interaction des ligands connus.[295,301,302] La carte d'interaction peut être extraite directement du complexe récepteur-ligand et permet ensuite la sélection de différents pharmacophores, comportant l'ensemble ou une grande partie des interactions détectées (***Figure 45)***.[302]

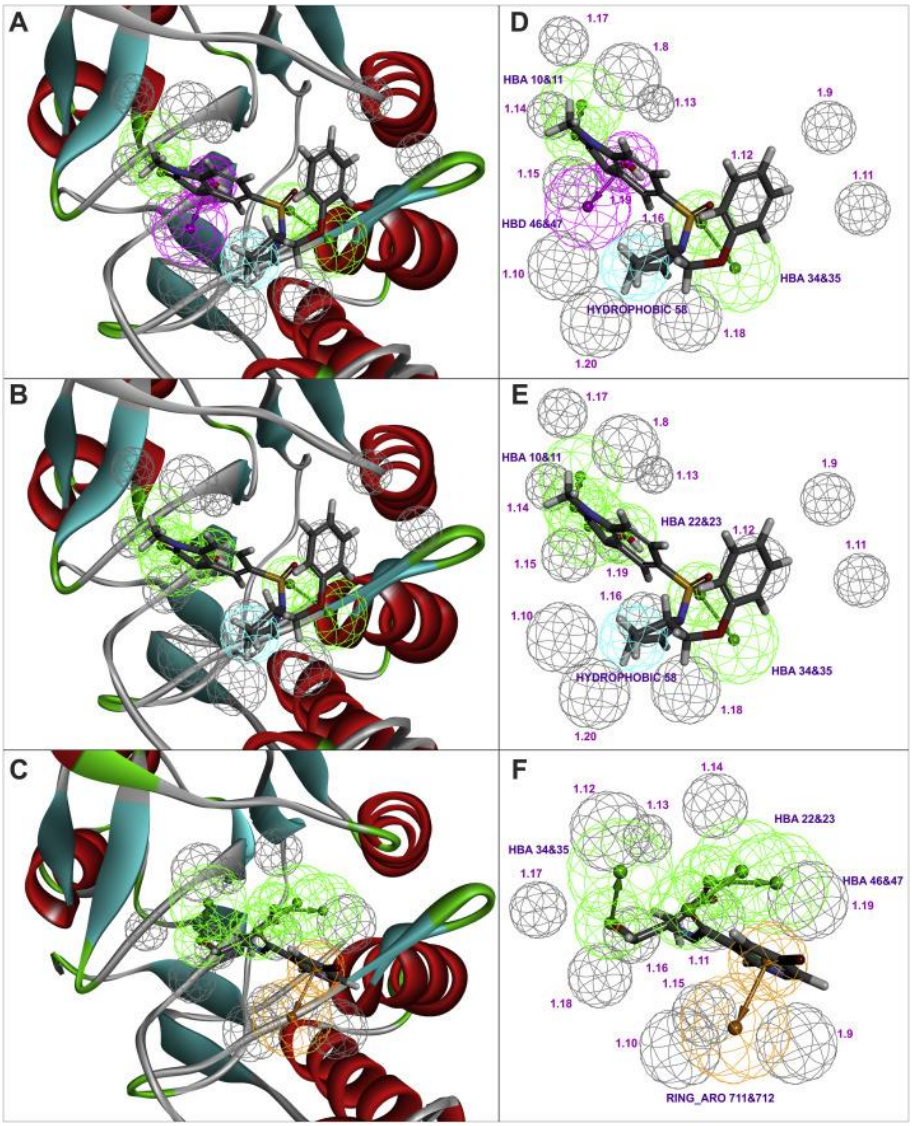


Figure 45. *Exemples de différentes hypothèses de pharmacophores obtenues à partir de données de co-cristallisation protéine-ligand de la caspase 3 (CASP3) avec Discovery Studio.*[303] *(A,B) Deux modèles pharmacophoriques différents proposés à partir de la même structure co-cristallisée (PDB 1gfw*[196]*). (C) Modèle proposé à partir d'une autre co-cristallisation de la caspase 3 (PDB 1re1*[196]*). (D,E) Modèles pharmacophoriques A et B isolés. Les volumes d'exclusion sont les mêmes, sélectionnés à partir de l'encombrement du récepteur. Deux modèles sont construits pour prendre en compte le caractère accepteur ou donneur de liaisons hydrogènes d'un groupement carboxylique (HDB46&47 – HBA22&23). (F) Modèle pharmacophorique C isolé. Ce modèle diffère largement de ceux obtenus à partir de la co-cristallisation 1gfw.*[302]

Les logiciels disponibles pour ce type d'analyse, dont LigandScout,[187] MOE,[138] PHASE,[186] HipHop[188] et de nombreux autres, différencient généralement six types d'interaction : donneur ou accepteur de liaisons hydrogènes, charge électrostatique positive ou négative, points hydrophobes et aromatiques. Dans le cas où plusieurs co-cristallisations de la cible sont disponibles, particulièrement lorsqu'elles accommodent différents ligands, il est important de comparer les pharmacophores obtenus afin d'estimer leur robustesse et leur fiabilité (***Figure 45)***.[302] L'alignement préalable des complexes permet généralement un bon alignement de leurs groupes fonctionnels et facilite la comparaison des différents pharmacophores.[295,301,302] Grâce à l'augmentation du nombre de structures co-cristallisées disponibles, les méthodes pharmacophoriques reposant sur le complexe récepteur-ligand sont de plus en plus utilisées puisqu'elles reposent sur l'utilisation de données expérimentales d'une grande fiabilité et permettent la construction de modèles très pertinents.[295,301,302]

2.4.4. Quantitative Structure-Activity Relationship (QSAR) « structure-based »

Les méthodes QSAR « ligand-based » ont initialement été proposées et développées à une période pendant laquelle un faible nombre de structure de protéines étaient disponibles. Le logiciel CoMSIA[237] fut proposé en 1999 alors que la PDB[196] comportait environ 9000 structures cristallographiques. Par la suite, les méthodes QSAR ont été améliorées afin de permettent l'utilisation des données structurales de la cible pour déduire les potentiels modes de liaison des ligands et affiner leurs modèles, permettant ainsi de meilleurs prédiction des propriétés biologiques des composés. Les données structurales utilisées peuvent être des cristallographies ou co-cristallographies à rayons X, en fonction des méthodes utilisées et du nombre de ligands connus. Les méthodes QSAR « structure-based », également dites RD QSAR, sont généralement classées selon trois catégories : 3D, 4D ou 7D-QSAR.[36,204] Le terme 7D-QSAR dénomme l'utilisation de méthodes RD 4D-QSAR prenant spécifiquement en compte l'adaptation du récepteur lors de l'accommodation d'un ligand (« induced-fit ») et différents modèles de solvatation des complexes récepteur-ligand simulés.[204]

2.4.4.1. RD 3D-QSAR

Parmi les premières méthodes RD 3D-QSAR proposées, le logiciel VALIDATE[304] permet la prédiction de l'activité biologique de nouveaux composés grâce à un modèle de régression PLS prenant en compte douze mesures des propriétés physico-chimiques et stériques d'un complexe récepteur-ligand : énergie d'interaction électrostatique et stérique, logP, nombre de liaisons rotatives, enthalpie de la formation du complexe récepteur-ligand et six mesures de surfaces

hydrophiles et hydrophobes du récepteur lié ou non-lié.[304] Les méthodes RD 3D-QSAR telles que VALIDATE[304] peuvent être mises en œuvre de manière directe lorsque plusieurs co-cristallisations récepteur-ligand sont disponibles. Il est également possible de réaliser un alignement préalable des ligands connus de la cible, par exemple avec des méthodes RI 3D-QSAR ou pharmacophoriques 3D, puis de positionner cet alignement dans le récepteur soit manuellement, soit sur la base d'une co-cristallisation existante pour l'un des ligands. Une étape de minimisation permettra ensuite de relaxer les conflits stériques, aboutissant généralement à une modélisation des interactions récepteur-ligand suffisamment fiable pour construire des modèles RD 3D-QSAR pertinents.[304]

Une autre approche proposée par le logiciel COMBINE[305] utilise un modèle de régression PLS construit à partir des énergies d'interaction calculées localement. Dans un premier temps, les énergies libres et liées du récepteur et des ligands sont calculées grâce à un champ de force (voir point 2.4.6.3.1). Les données structurales des complexes récepteur-ligand, obtenues par co-cristallographie ou par modélisation, sont décomposées en fragments des ligands et en régions du récepteur.[305] Les énergies d'interaction stériques et électrostatiques sont ensuite évaluées entre chacun de ces fragments et chacune de ces régions. Les changements énergétiques internes à chaque fragment et région, induits par la formation du complexe récepteur-ligand, sont également quantifiés par rapport à leurs énergies libres.[305] Un modèle de régression PLS est ensuite construit pour corréler l'activité biologique des ligands aux variations énergétiques locales des fragments et régions.

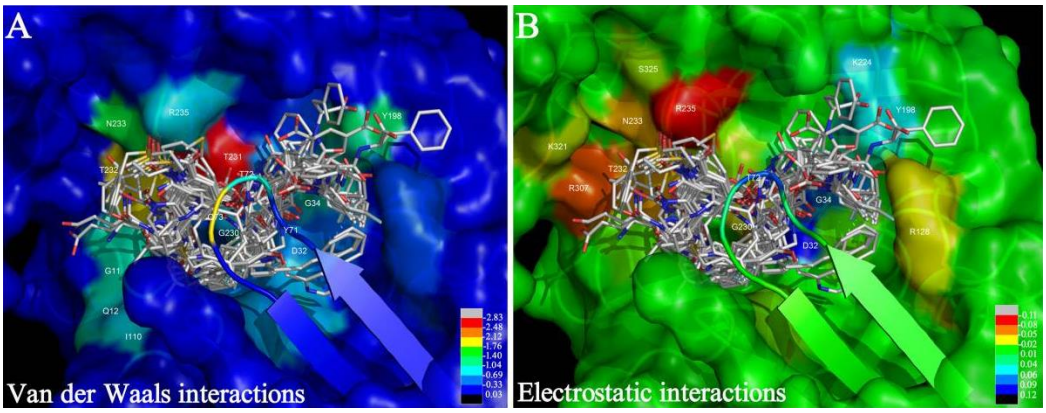


Figure 46. *Structures de 46 inhibiteurs liés au récepteur de la beta-secretase 1 (BACE-1) (PDB 1w51*[196]*). Visualisation PyMol*[306] *de la surface du récepteur, représentée de manière semi-transparente et colorée selon un spectre correspondant à l'échelle des coefficients d'une régression PLS construite avec gCOMBINE*[307] *pour expliquer l'activité des ligands. (A) Coefficients obtenus pour les interactions stériques entre fragments des ligands et régions du récepteur. (B) Coefficients obtenus pour les interactions électrostatiques.*[308]

Le logiciel gCOMBINE[307] implémente cette méthode avec une interface qui permet la génération de graphiques pour l'analyse des résultats. Cette approche reposant sur l'évaluation locale des changements énergétiques a l'avantage de permettre une visualisation directe des interactions favorisant l'activité des ligands dans le modèle statistique établi (***Figure 46***).

2.4.4.2. RD 4D-QSAR

Le formalisme des méthodes RD 4D-QSAR a été proposé en 2003 par Pan et al.[309] à l'initiative de Hopfinger, qui introduit précédemment les méthodes RI 4D-QSAR. L'approche proposée repose sur l'utilisation de simulations de dynamique moléculaire pour la modélisation des complexes récepteur-ligand.[309] Comme pour les méthodes RI 4D-QSAR (voir point 2.3.4.3), une grille d'occupation permet de détecter les caractéristiques physico-chimiques favorables ou défavorables à l'interaction des ligands (***Figure 47A***). L'utilisation de simulations de dynamique moléculaire requiert la définition d'un référentiel stable pour le positionnement de la grille d'occupation. L'utilisateur doit donc définir plusieurs points d'interaction fixes entre le récepteur et chacun des ligands, sur la base de données expérimentales obtenues préalablement. Le cas d'application présenté par Pan et al.[309] utilise 47 inhibiteurs de la glycogène phosphorylase b (GPb), tous analogues du glucose, permettant de figer spatialement trois points d'interaction communs au glucose au cours des simulations. Les atomes du récepteur situés entre 10 et 12 Å de ces points d'interaction sont également figés afin de maintenir la structure du récepteur au cours des simulations. Après alignement des ligands sur le récepteur, leurs positions sont minimisées avant de procéder à de courtes dynamiques moléculaires. La grille d'occupation est ensuite analysée à chaque étape des simulations grâce à des régressions PLS et MLR, afin de détecter les points d'interactions favorables ou défavorables à l'activité des ligands ainsi que leurs types physico-chimiques (***Figure 47B***).[309]

Dans la même étude, Pan et al.[309] comparent leur approche RD 4D-QSAR à une étude RI 4D-QSAR réalisée à partir d'un alignement initial identique de ces inhibiteurs de la GPb. La qualité prédictive des modèles obtenus est similaire, bien que différents points d'interaction soient identifiés (***Figure 47C***). Les différences obtenues peuvent être expliquées, notamment, par un alignement plus permissif des composés rigides avec l'approche RI 4D-QSAR, pour lesquels des effets « induced-fit » seront observés avec l'approche RD 4D-QSAR.[309]

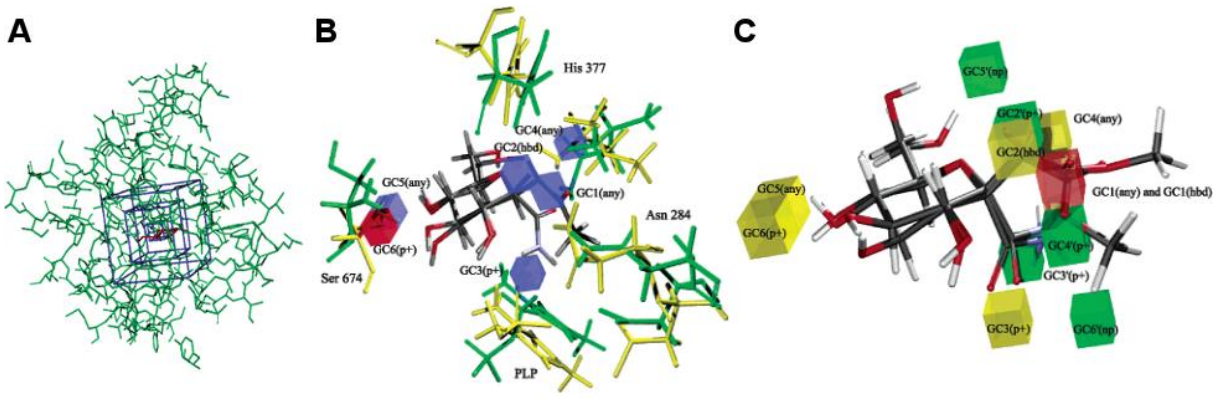


Figure 47. *Approches 4D-QSAR appliquées à 47 inhibiteurs de la GPb.*[309] *(A) Représentation du positionnement de la grille d'occupation à la surface du récepteur. Les atomes du récepteur présents au bord de la grille sont figés au cours des simulations. (B) Représentation du meilleur modèle RD 4D-QSAR obtenu (bleu : interactions favorables à l'activité des ligands, rouge : défavorables). Plusieurs résidus du récepteur peuvent être réorientés (vert, jaune). (C) Comparaison des meilleurs modèles RD et RI 4D-QSAR obtenus pour ces 47 inhibiteurs à partir d'alignements initiaux identiques (jaune : RD 4D-QSAR, vert : RI 4D-QSAR, rouge : point d'interaction commun aux modèles).*[309]

2.4.5. Approches *de novo*

Les méthodes dites de construction *de novo* reposent sur la connaissance du site actif de la cible étudiée pour construire, de manière incrémentielle ou combinatoire, des composés qui lui seront spécifiquement adaptés.[310] Ces méthodes ont initialement été développées afin de réaliser des assemblages d'atomes uniques ou de fragments moléculaires (approches « fragment-based »). Si les approches utilisant un assemblage d'atomes permettent d'explorer une plus grande diversité de molécules, elles présentent également le désavantage de générer un très grand nombre de composés dont il peut être difficile d'extraire les « hits » les plus pertinents.[310] Le développement des méthodes *de novo* a donc été focalisé sur l'assemblage de fragments, permettant d'explorer l'espace chimique théorique de manière rationnelle sans recourir à une recherche exhaustive trop exigeante en temps de calcul. La majorité des méthodes reposent sur l'utilisation de chimiothèques de fragments et d'algorithmes stochastiques permettant une optimisation locale à chaque étape de la construction des molécules dans le site de liaison de la cible.[310] Une solution de conception *de novo* doit être efficace en trois points principaux : (i) l'assemblage des différentes fragments moléculaires, (ii) l'échantillonnage de l'espace chimique accessible et (iii) l'évaluation de la qualité des assemblages.[310]

2.4.5.1. Identification des sites d'interaction

La mise en œuvre des approches *de novo* nécessite, dans un premier temps, l'identification des points d'interaction potentiels du site de liaison de la cible. Les interactions recherchées sont généralement de type liaisons hydrogènes, interactions électrostatiques et hydrophobes.[310] Après avoir identifié le site de liaison (voir point 2.4.1), celui-ci peut être analysé de différentes manières afin de construire une image négative du récepteur représentant ses points d'interaction potentiels avec un ligand, similairement à ce qui est réalisé pour la construction de pharmacophores « structure-based ». Des méthodes basées sur les règles, dont HSITE,[311] LUDI[296,297] ou HIPPO,[312] utilisent des critères géométriques dérivés de données expérimentales. D'autres types de méthodes, comme GRID,[313] GREEN[314] ou LigBuilder,[315] reposent sur l'utilisation de grilles permettant de discrétiser l'espace à la surface du site de liaison. Des sondes représentant la présence hypothétique d'un atome ou d'un groupe fonctionnel à chaque point de la grille permettent ensuite de calculer des potentiels d'interaction à la surface du site de liaison, puis d'identifier les sites d'interaction.

D'une manière similaire, la méthode MCSS (Multiple Copy Simultaneous Search)[316] permet également de déterminer les positions et orientations énergétiquement favorables de fragments moléculaires dans le site de liaison. De nombreuses copies de fragments moléculaires fonctionnels sont d'abord positionnées à la surface du récepteur, utilisées comme sondes. Un champ de force permet ensuite de minimiser leurs positions, simultanément et indépendamment (voir point 2.4.6.2.3). Chacun de ces fragments minimisés sera ensuite soit éliminé, soit conservé comme hypothèse de départ pour l'étape d'assemblage, en fonction de son énergie d'interaction avec le récepteur.[310,316]

2.4.5.2. Assemblage des fragments moléculaires

L'assemblage des fragments peut être réalisé selon différentes stratégies, éventuellement combinées : par croissance (« fragment-growth »), par liaison (« fragment-linking »), à l'aide d'une grille, par l'usage de dynamiques moléculaires ou de méthodes stochastiques.[310]

L'assemblage par croissance s'initie par le placement d'un premier fragment au niveau d'un site d'interaction du récepteur. Le site d'interaction et le fragment initial peuvent être choisis par l'utilisateur, par exemple lorsque des données expérimentales valident l'importance d'un site d'interaction pour l'inhibition de l'activité de la cible. Le fragment initial sera étendu progressivement par l'ajout d'autres fragments, en maximisant leurs interactions avec le

récepteur (***Figure 48A-C***).[310] Cette approche a été implémentée dernièrement dans les logiciels Contour,[317] AutoGrow, BOMB[318] et FlexNoVo.[319]

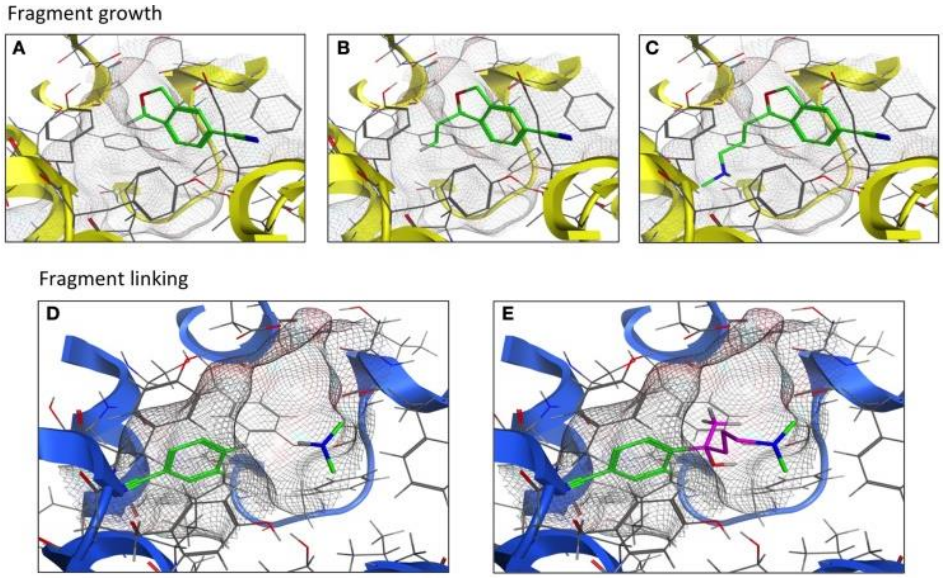


Figure 48. *Illustration du fonctionnement des méthodes « fragment-based » reposant sur la croissance de fragments et liaison de fragments. (A-C) Croissance des fragments : le fragment (vert) est étendu progressivement en maximisant ses interactions dans le site de liaison. (D, E) Liaison des fragments : les fragments (vert) sont joints par l'ajout d'un connecteur (violet). Les interactions définies par le positionnement sont conservées dans la conception du nouveau ligand.*[43]

Pour un assemblage par liaison, la chimiothèque de fragments est explorée afin de sélectionner et positionner plusieurs fragments complémentaires à chaque site d'interaction du récepteur. Les hypothèses de liaison de ces fragments sont ensuite formulées de manière exhaustive, toujours en satisfaisant aux contraintes stériques du récepteur (***Figure 48D-E***).[310] Cette approche est notamment utilisée par les logiciels MED-Hybridise,[320] LigBuilder,[315] PRO-LIGAND[321] et GrowMol.[322]

Il est également possible de réaliser un assemblage de fragments ou d'atomes grâce à une grille de points positionnée entre les sites d'interaction du récepteur.[310,323] Après avoir positionné les différents fragments ou atomes adaptés aux sites d'interaction, ceux-ci sont reliés selon les plus courts chemins passant par les points de la grille. Ces données permettent ensuite d'envisager plusieurs hypothèses de liaison covalente des fragments ou atomes porteurs des caractéristiques physico-chimiques nécessaires à l'interaction avec le récepteur. Il est également possible de commencer par la définition d'un squelette moléculaire que l'on rendra fonctionnel dans un

second temps (***Figure 49***). Ces approches sont notamment implémentées dans les logiciels BUILDER[324,325] et Diamond Lattice.[323]

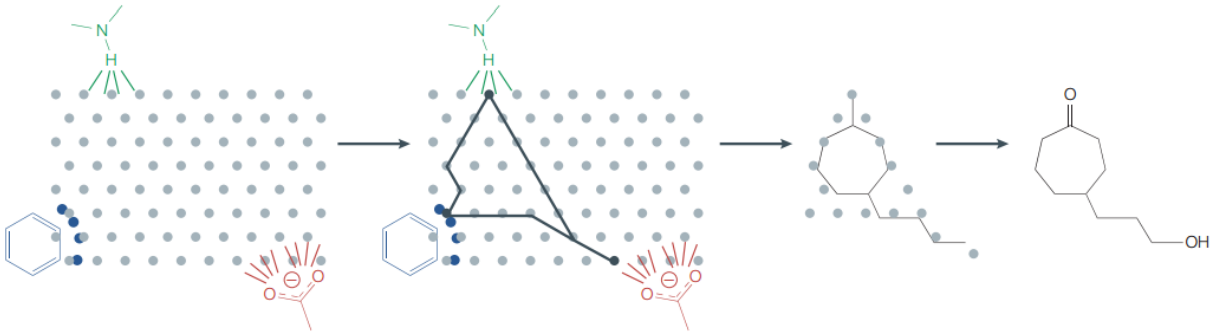


Figure 49. *Illustration d'une conception moléculaire* de novo *à l'échelle atomique à l'aide d'une grille de points. Les sites d'interaction sont reliés selon les plus courts chemins passant par les points de la grille. Le chemin obtenu permet la construction de plusieurs squelettes moléculaires, qui sont ensuite rendus fonctionnels par l'ajout d'atomes porteurs des caractéristiques physico-chimiques adaptées au récepteur.*[310]

Enfin, quelques méthodes reposent sur l'utilisation de dynamiques moléculaires pour assurer un positionnement optimal des fragments dans le récepteur avant de procéder à l'assemblage.[310] Après avoir placé les fragments aléatoirement dans le site actif, ceux-ci sont liés de manière covalente selon une procédure stochastique, puis une dynamique moléculaire est lancée. A chaque intervalle de temps défini, un ou plusieurs fragments sont sélectionnés, toutes les liaisons précédemment établies sont clivées puis reconstruites aléatoirement et chaque assemblage est enregistré pour évaluation. Cette approche fut proposée pour la première fois en 1996 avec le logiciel CONCERTS[326] et considérait à la fois les interactions fragments-récepteur et fragments-fragments. D'autres approches furent ensuite développées pour prendre exclusivement en compte les interactions fragments-récepteur et proposer différents modèles de clivage et reconstruction des liaisons covalentes entre fragments.[327–330]

2.4.5.3. Echantillonnage de l'espace chimique

Quelle que soit la stratégie d'assemblage moléculaire choisie, les méthodes de conception *de novo* doivent échantillonner l'espace chimique accessible de manière efficace, malgré le très grand nombre de molécules pouvant théoriquement s'inscrire dans le récepteur étudié.[310] Deux approches principales sont distinguées, dites « depth-first search » (ou « recherche en profondeur) et « breadth-first search » (ou « recherche en largeur »).[310] La recherche en profondeur ne retient que quelques solutions partielles à chaque étape du processus de construction des molécules, en ne considérant qu'un nombre limité d'hypothèses de départ (i.e. fragments initiaux) (***Figure 50***).

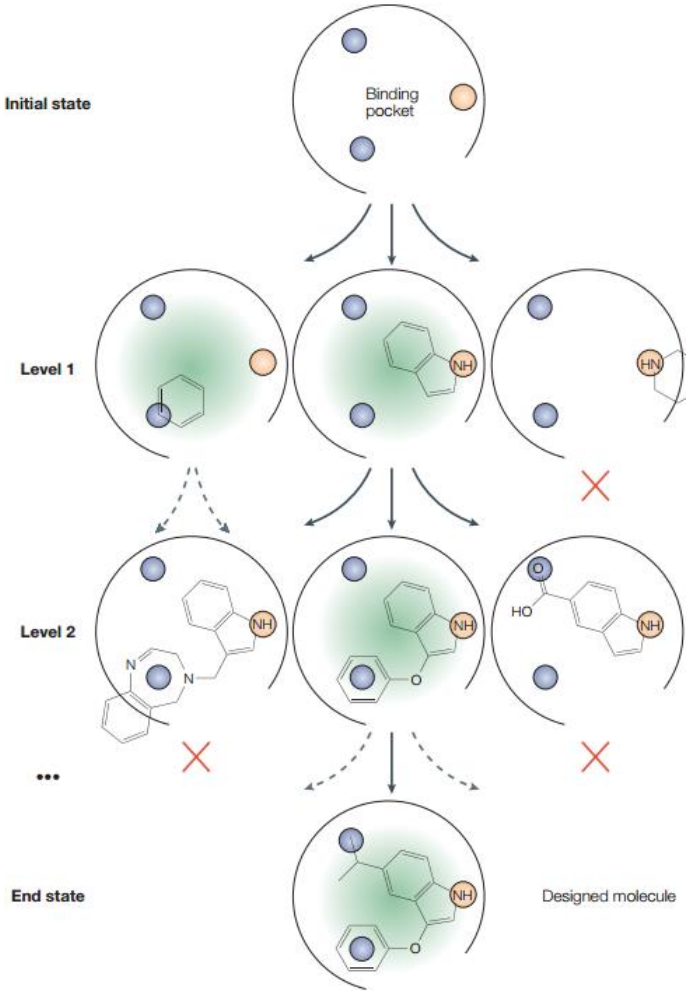


Figure 50. *Illustration du processus d'exploration de l'espace chimique réalisé par les méthodes « fragment-based » utilisant une recherche en profondeur. Seules quelques solutions partielles sont conservées à chaque étape de construction des molécules (représenté au milieu).*[310]

Inversement, une recherche en largeur conserve et évalue l'ensemble des solutions partielles générées à chaque étape, permettant une couverture exhaustive de l'espace chimique étudié.[310] Les logiciels implémentant une recherche en largeur sont généralement ceux qui utilisent un assemblage de fragments par liaison. L'espace de recherche étant initialement restreint par un pré-positionnement optimal des différents fragments, cette procédure intensive en calculs peut alors être mise en oeuvre.[310]

Il existe également quelques méthodes qui procèdent à un échantillonnage aléatoire de l'espace chimique, dont les logiciels LEGEND[331] et CONCEPTS[328] qui emploient une approche de Monte Carlo[110] combinée au critère de Metropolis. D'autres logiciels utilisent des algorithmes génétiques ou évolutionnaires qui miment une sélection biologique des composés les plus pertinents. Parmi ceux-ci, LEA[332] et sa nouvelle version LEA3D[333] représentent les molécules

par leurs codes SMILES sur lesquels des opérateurs génétiques sont appliqués (mutation et recombinaison) (***Figure 51***).

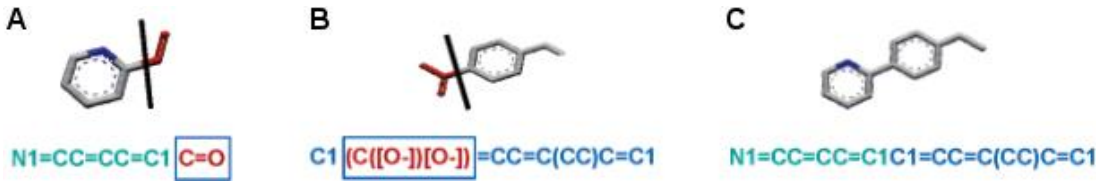


Figure 51. *Représentation d'une étape de recombinaison réalisée par l'algorithme génétique du logiciel LEA3D.*[333] *Les codes SMILES des molécules (A) et (B) sont combinés pour produire la molécule (C).*[333]

2.4.5.4. Evaluation de la qualité des assemblages

Lors de l'assemblages des molécules, celles-ci sont évaluées à chaque étape par une fonction de score afin de guider la construction de ligands optimaux.[310] Les premières approches *de novo* utilisaient un score calculé uniquement à partir des contraintes stériques du récepteur (BUILDER[325] et NEWLEAD[334]). De nombreuses fonctions de score ont ensuite été développées, principalement réparties selon trois catégories : les fonctions de score basées sur les champs de force, sur les connaissances et les fonctions de score empiriques.[27,33,335] Les fonctions de score utilisées par les méthodes de conception *de novo*, très similaires à celles utilisées par les méthodes de docking, sont détaillées en point 2.4.6.3.

2.4.6. Méthodes de docking

Les méthodes de docking ont pour objectif de prédire la capacité d'une molécule à se lier au site d'actif d'une protéine, grâce à une simulation de la conformation et de l'orientation qu'elle adopte lors de sa liaison au récepteur.[28,44] Elles sont généralement considérées comme complémentaires aux criblages HTS *in vivo*, puisqu'elles permettent un premier filtrage *in silico* des chimiothèques pour la détection des molécules les plus aptes à constituer des « hits » réels. Pour ce faire, les logiciels combinent l'utilisation d'un algorithme de recherche conformationnelle à une fonction de score, permettant respectivement de générer différents modes de liaison potentiels d'une molécule dans le récepteur (ou « poses ») et de calculer un indice d'affinité molécule-récepteur.[28,44] Les méthodes de docking ont deux objectifs principaux : (i) l'identification de ligands de la cible étudiée à partir de larges chimiothèques, grâce à la prédiction de scores d'affinité ou de valeurs d'affinité réelles (K_i, IC_{50}, etc.) et (ii) déterminer les modes de liaison adoptées par les ligands dans le récepteur, permettant notamment de procéder à des optimisations rationnelles des composés.[28,44]

2.4.6.1. Docking de molécules rigides

Jusqu'au milieu du XX^{e} siècle, le mécanisme de liaison d'un ligand à une protéine était compris comme un processus statique impliquant deux formes stériques complémentaires (concept « clé-serrure »).[336] Les premières approches de docking furent développées selon ce modèle, en considérant les molécules et le site de liaison de la protéine cible comme des entités rigides. Désormais populaire, le logiciel DOCK[337] fut proposé dans sa première version en 1982 pour le docking rigide de conformations moléculaires. Parmi les logiciels de ce type, DOCK (version 1),[337] FLOG[338] et FRED[339] procèdent selon des méthodes assez similaires.[28,44] L'approche proposée par FRED[339] est la suivante. Après avoir généré les différentes conformations moléculaires, leurs poses potentielles sont obtenues par énumération de l'ensemble des translations et rotations dans un espace de recherche englobant le site de liaison de la protéine cible (***Figure 52***). Une image négative du site de liaison permet ensuite d'éliminer les poses stériquement incompatibles avec le récepteur, puis les meilleures poses sont optimisées. L'intérêt des méthodes de docking rigide réside dans leur rapidité d'exécution qui permet de réaliser un premier filtrage de très larges chimiothèques avant de procéder, par exemple, à la mise en œuvre de méthodes de docking flexible.[28,44]

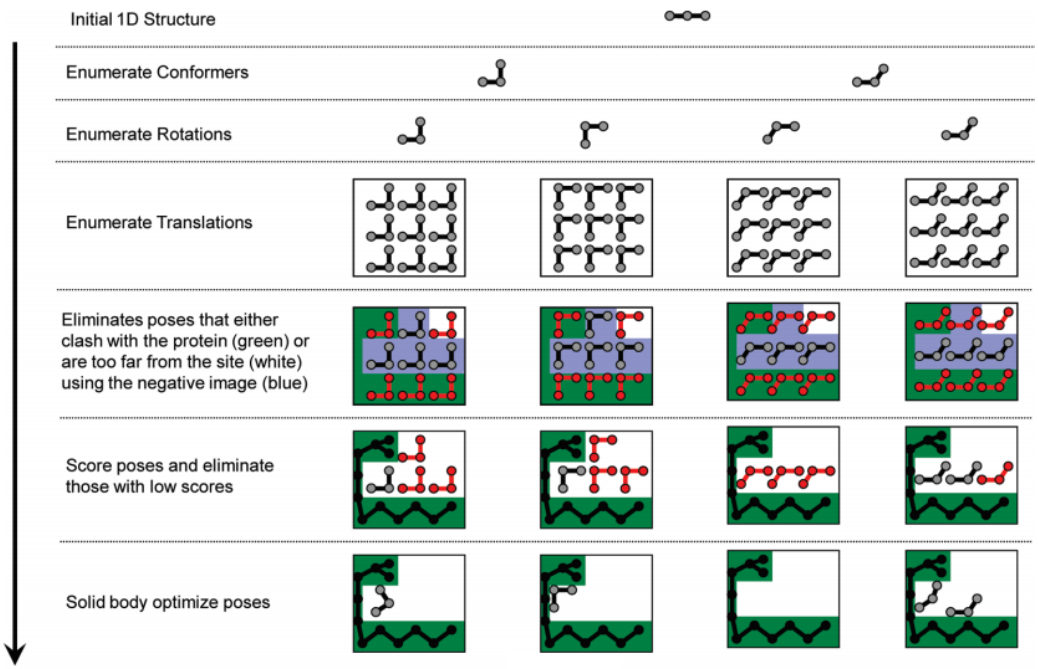


Figure 52. *Illustration du protocole de docking rigide implémenté dans le logiciel FRED.*[339] *Les poses de basse énergie d'interaction ou stériquement incompatibles avec le site de liaison (rouge) sont éliminées par étapes successives. Les meilleures poses sont ensuite optimisées, toujours de manière rigide, par quelques dernières translations et rotations utilisant des paramètres plus précis.*[339]

2.4.6.2. Docking de molécules flexibles

Bien que les méthodes de docking rigide permettent l'obtention rapide de résultats préliminaires, elles reposent toutefois sur des heuristiques non-représentatives du mécanisme biologique d'interaction entre un ligand et son récepteur.[28,44] En réalité, une protéine peut être très flexible et exister dans des états transitoires variés, dont certains seront plus aptes à amorcer la liaison d'un ligand donné (***Figure 53***). Ce mécanisme est aujourd'hui connu et accepté sous le nom de sélection de conformation (« conformational selection »).[340–342] Inversement, un ligand peut induire une modification du récepteur comme, par exemple, un élargissement du site de liaison ou un repositionnement des chaînes latérales des acides aminés impliqués dans l'interaction. Ce mécanisme est connu sous le nom d'effets « induced-fit » (***Figure 53***).[249,340]

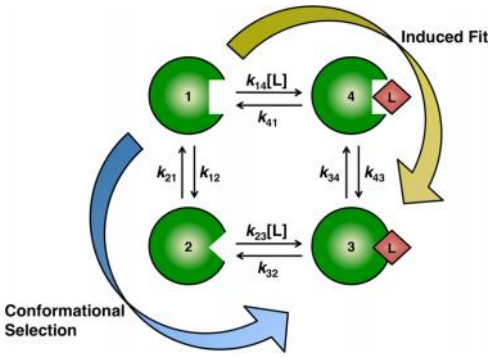


Figure 53. *Illustration du mécanisme général de reconnaissance et de liaison d'une molécule à une protéine. Le processus de sélection de conformation implique principalement la flexibilité de la protéine cible, permettant d'atteindre une conformation apte à l'accommodation du ligand (1-2-3), tandis que l'effet « induced-fit » est une modification structurale du récepteur initiée par la fixation du ligand (4-3).*[340]

De nouvelles méthodes de docking ont ainsi été développées afin de prendre en compte ces mécanismes de flexibilité, à la fois des molécules et de leurs récepteurs. La simulation de la flexibilité d'un récepteur est désormais proposée par de nombreux logiciels de docking, bien qu'elle ne soit pas systématiquement envisagée du fait des coûts calculatoires associés.[8,33,44] Les algorithmes de docking permettant de simuler une flexibilité moléculaire peuvent être classés en trois catégories : les algorithmes de recherche systématique, stochastique ou déterministe.[9,343,344] Une liste des principaux logiciels de docking rigide ou flexible est disponible en point 2.4.6.4.

2.4.6.2.1. Recherche systématique

Les logiciels de docking utilisant une recherche systématique cherchent à approximer une exploration de tous les degrés de liberté des liaisons covalentes rotatives d'une molécule, de 0

à 360° selon un pas prédéfini, tout en évaluant le positionnement de ces conformations dans le site de liaison.[343] Cette approche peut générer un très grand nombre de conformations moléculaires, conduisant à un problème connu comme l'explosion combinatoire du nombre de possibilités (***Équation 4***).

$$N_{conformations} = \left(\frac{360}{\theta}\right)^{n}$$

Équation 4. *Nombre de conformations moléculaires obtenues par une recherche systématique utilisant un angle incrémentiel de* θ° *sur un composé comportant* n *liaisons rotatives.*

Par exemple, la représentation d'une molécule comportant 7 liaisons rotatives avec un angle de recherche incrémentiel de 60° peut produire jusqu'à 279.936 conformations différentes. Une énumération de ce type, sans critères énergétiques, peut toutefois produire des conformations aberrantes. Deux types de recherches systématiques permettent de pallier ce problème pour procéder à un docking flexible : les méthodes de recherche exhaustive et les méthodes de reconstruction incrémentielle.[343]

Parmi les méthodes reposant sur une recherche exhaustive, le logiciel GLIDE[345] utilise une première étape d'échantillonnage des conformations moléculaires tout en limitant l'explosion combinatoire grâce à une évaluation optimisée de leurs énergies libres : le processus d'énumération est rapidement arrêté lorsque certaines valeurs d'angles produisent des conformations de trop haute énergie. Ensuite, pour faciliter la recherche de poses, GLIDE[345] considère chaque molécule comme la somme de son squelette rigide (« core ») et de ses régions flexibles (***Figure 54***).

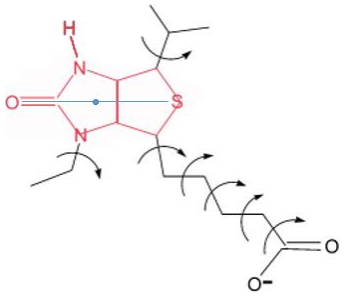


Figure 54. *Représentation d'une molécule par GLIDE, comme la somme de son squelette rigide (rouge) et de ses régions flexibles. Le barycentre (point bleu) et l'axe du squelette rigide, donnée par l'axe de ses atomes les plus distants (bleu), permettent de placer et d'orienter les conformations moléculaires dans le site de liaison.*[345]

Le barycentre de chaque squelette rigide est ensuite positionné en chaque point d'une grille englobant le site de liaison, puis leurs différentes orientations sont énumérées. Lorsque la position et l'orientation d'un squelette rigide ne provoque pas de conflits stériques et que celui-

ci est suffisamment proche de la surface du site de liaison, les différentes conformations de la molécule associée lui sont superposées, puis scorées.[345]

Les méthodes de reconstruction incrémentielle reposent sur une fragmentation des molécules pour procéder ensuite à leur reconstruction en maximisant leurs interactions avec le site de liaison.[346] Un des avantages principaux de cette approche, utilisée notamment dans DOCK (version 6),[347] FlexX[348] et Surflex-dock,[349,350] est de permettre des temps de calcul linéaires à l'augmentation du nombre de liaisons rotatives des molécules. Ce dernier utilise une variante de l'algorithme Hammerhead[346] qui présente certaines similitudes avec les approches de conception *de novo*. La première étape du protocole de Surflex-dock[349,350] consiste à cartographier le site de liaison grâce à trois types de fragments sondes (CH_4, C=O et N-H) placés et optimisés de manière à maximiser leurs interactions avec le récepteur, tout en couvrant l'intégralité de son volume. Cette procédure permet l'obtention d'un prototype exhaustif de ligand idéal dit « protomol » (***Figure 55***).[349,350]

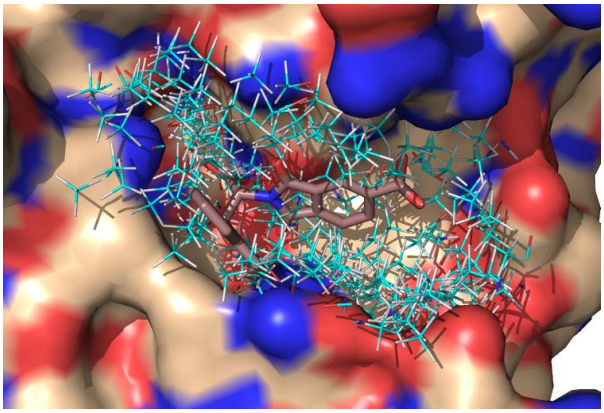


Figure 55. *Protomol généré pour la β-lactamase Adénosine MonoPhosphate cyclique (AMPc) par extension de l'espace de recherche à 4 Å autour d'un de ses inhibiteurs co-cristallisé (PDB 2r9w*[196]*).*

La procédure de docking utilise une fragmentation des molécules par rupture des liaisons covalentes non-aromatiques, puis une recherche des conformations des fragments. Chacune des conformations de fragments obtenues est d'abord alignée sur le protomol grâce à une fonction de similarité morphologique.[351] La position des fragments qui présentent les meilleurs scores d'affinité avec le récepteur est ensuite optimisée localement, puis sert de point de départ pour positionner les fragments suivants et reconstruire des poses moléculaires par superposition maximale avec le protomol.[346,349–351] Les poses finales sont évaluées de manière similaire, par une fonction de superposition au protomol donnée par l'algorithme Hammerhead.[346]

2.4.6.2.2. Recherche stochastique

Les logiciels de docking utilisant une recherche stochastique procèdent à des changements aléatoires des variables de translations, rotations et torsions conditionnant la position et la conformation des molécules dans le site de liaison de la cible. Les algorithmes génétiques[343,344] et de Monte Carlo[110] comptent parmi les approches les plus utilisées.[9,343,344]

Les méthodes de Monte Carlo[110] évaluent donc l'impact de chaque modification aléatoire itérative de la position et/ou de la conformation d'une molécule sur l'énergie de sa pose dans le site de liaison. Lorsque la modification opérée produit une pose de plus haute énergie qu'à l'étape précédente, cette modification peut être acceptée ou rejetée en fonction du critère de Metropolis (***Figure 56***).[110] Ces étapes sont généralement répétées entre 100 et 10.000 fois pour chaque molécule, en fonction du nombre de pas initialement choisi ou jusqu'à atteindre une énergie minimale.[352]

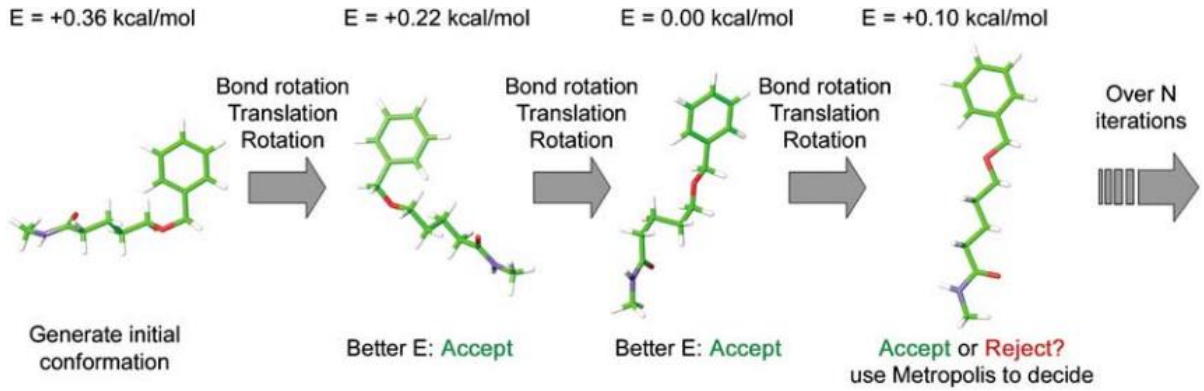


Figure 56. *Illustration du fonctionnement d'un algorithme de Monte Carlo[110] appliqué au docking flexible de molécules. Une ou plusieurs modifications aléatoires peuvent être réalisées à chaque étape.*[27]

Les logiciels MCDOCK,[353] ProDOCK,[354] AADS,[355] QXP,[356] ParDock,[357] DockVision[358] et ICM[359] reposent principalement sur une procédure de Monte Carlo.[110] Le logiciel ICM[359] utilise une première étape de placement aléatoire de conformations moléculaires dans une boite centrée sur le site de liaison, qui permet également de limiter l'espace de recherche pour la suite de la procédure. Pour optimiser l'évaluation de l'énergie des poses, le récepteur est représenté grâce à une grille de potentiels hydrophobes, stériques, électrostatiques et accepteurs ou donneurs de liaisons hydrogènes.[359]

Les algorithmes génétiques miment l'évolution des systèmes biologiques telle que décrite par la théorie de l'évolution de Darwin.[360] Les opérateurs génétiques (réplication, recombinaison, mutation, sélection) sont ici adaptés afin de manipuler les variables (chromosomes)

conditionnant la position et la conformation des molécules dans le site de liaison.[343,344] Après avoir positionné un petit nombre de conformations initiales dans une boite englobant le récepteur, celles-ci sont modifiées itérativement par recombinaison de deux conformations (échange d'une partie de leurs variables d'état) ou par mutation d'une conformation (modification aléatoire d'une ou plusieurs variables d'état) (***Figure 57***). Après chaque modification, une fonction de score permet de sélectionner les conformations qui seront à nouveau modifiées pour produire la génération suivante. La population de conformations moléculaires initiales produit donc, successivement, de nouvelles générations de conformations qui tendent vers une solution optimale. Ce type d'algorithme est implémenté, notamment, dans les logiciels AutoDock[361] et GOLD.[93]

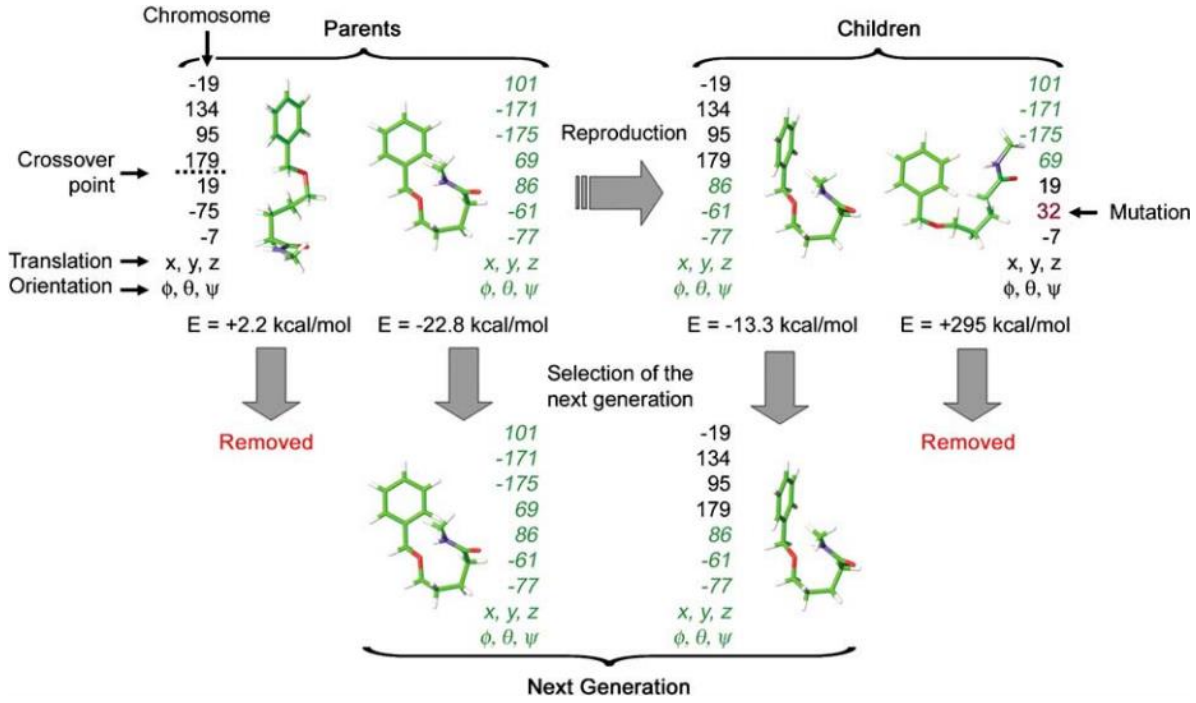


Figure 57. *Illustration du fonctionnement d'un algorithme génétique appliqué au docking flexible de molécules. Le chromosome représenté contient les valeurs des angles de torsions d'une conformation moléculaire.*[27]

2.4.6.2.3. Recherche déterministe

Les logiciels de docking peuvent également utiliser des méthodes de recherche déterministe : les simulations de dynamique moléculaire et les procédures de minimisation d'énergie.[343,362] Les méthodes de dynamique moléculaire reposent sur l'équation du mouvement décrite par Newton pour simuler l'évolution d'un système récepteur-ligand au cours du temps à l'échelle atomique.[362,363] La simulation est initialisée par l'attribution de vitesses à chaque atome du système selon une distribution caractérisant une température donnée. La position de chaque

atome est ensuite recalculée à très courts intervalles de temps en fonction des forces qui lui sont appliquées, dérivées des énergies potentielles et cinétiques des atomes environnants (***Équation 5***). Cette technique permet donc d'observer finement la trajectoire d'une molécule dans le site de liaison de la protéine cible et la stabilité des interactions établies. Cependant, avec des temps de calcul raisonnables, les méthodes de dynamique moléculaire tendent à positionner les molécules selon des minimums locaux d'énergie et ne permettent pas systématiquement d'identifier une pose d'énergie minimale globale dans le site de liaison.[362,363] Une stratégie envisageable pour pallier ce problème consiste à utiliser différentes poses initiales pour lancer plusieurs dynamiques moléculaires.[344] Si les temps de simulation restent trop élevés pour utiliser cette approche à large échelle, elle peut toutefois être envisagée pour optimiser et valider un petit nombre de poses moléculaires.[362,363]

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

Équation 5. *Equation de Newton utilisée pour le calcul de la nouvelle position d'un atome après un intervalle de temps (d^2r_i/dt^2) en fonction de sa masse (m_i) et de la somme des forces qui lui sont appliquées (F_i).*

Les méthodes de minimisation d'énergie reposent sur l'usage de champs de force (voir point 2.4.6.3.1) pour optimiser, dans le cas des méthodes de docking, les positions atomiques d'un complexe récepteur-ligand selon leurs énergies à la fois intra et intermoléculaires. Les temps d'exécution associés sont très courts et permettent d'obtenir des poses plus proches de la réalité biologique, notamment en affinant les positions des chaînes latérales des acides aminés impliqués dans l'interaction. Cette approche identifie donc des minimums énergétiques locaux et est majoritairement utilisée pour affiner les poses moléculaires.

2.4.6.3. Calcul des scores d'affinité

Après avoir obtenu des poses moléculaires, les méthodes de docking utilisent des fonctions de score permettant de calculer l'affinité de liaison entre une molécule et son récepteur.[27,33,335] Cette dernière étape est tout aussi importante que l'étape d'échantillonnage précédente, puisque les nombreuses poses générées pour chaque molécule doivent désormais être triées par les fonctions de score afin d'identifier la conformation bioactive d'un ligand. Inversement, lorsque la molécule dockée est inactive sur la cible, les fonctions de score doivent attribuer des scores d'interaction défavorables à l'ensemble des poses. L'échantillonnage des poses permet généralement d'obtenir la conformation bioactive d'un ligand parmi les solutions proposées. Cependant, les fonctions de score actuelles n'attribuent pas systématiquement les meilleurs

scores d'affinité à ces conformations bioactives. Leur développement et leur amélioration fait l'objet d'un effort de recherche important.[27,33,335]

Les fonctions de score calculent une énergie d'interaction en réalisant la somme de termes indépendants qui permettent d'approximer l'énergie des phénomènes physiques impliqués : complémentarité de forme, interactions électrostatiques et de van der Waals, interactions hydrophobes et énergies de désolvatation. Ces énergies peuvent être calculées directement ou indirectement par les fonctions de score grâce à des mesures de complémentarité géométrique, de chevauchement intra et intermoléculaire, d'aires de contact, de proximité et de charge atomique, des contacts électrostatiques et des énergies de solvatation.[27,33,335] Chacun de ces termes apporte une contribution favorable ou défavorable au score d'interaction. Les fonctions de score peuvent être classées selon trois catégories : les fonctions de score basées sur les champs de force, sur les connaissances et les fonctions de score empiriques.

2.4.6.3.1. Fonctions basées sur les champs de force

Les fonctions de score basées sur les champs de force utilisent un calcul indépendant de l'énergie d'interaction récepteur-ligand et des énergies internes de la molécule et de la protéine cible. Cette procédure est notamment intéressante pour l'implémentation de logiciels de docking puisque la majorité d'entre eux utilise une unique conformation de la protéine cible, ce qui accélère les calculs de scores.[27] L'expression d'un champ de force prend généralement la forme donnée en ***Équation 6***.

$$E = \sum_{Liaisons} K_L(r - r_0)^2 + \sum_{Valences} K_V(\theta - \theta_0)^2 + \sum_{Torsions} K_T[1 + \cos(n\varphi - \varphi_0)] + \sum_{VDW} 4\varepsilon_{ij}\left(\frac{\sigma_{ij}^{12}}{r_{ij}^{12}} - \frac{\sigma_{ij}^{6}}{r_{ij}^{6}}\right) + \sum_{Coulomb} \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}}$$

Équation 6. *Forme générale de l'expression d'un champ de force. Les trois premiers termes décrivent les contributions intramoléculaires à l'énergie totale du système. Les deux derniers termes correspondent aux interactions intermoléculaires de van der Waals (ici représentées par un potentiel de Lennard-Jones 12-6) et électrostatiques (K_L, K_V et K_T : facteurs de pénalité pour les écarts de distance des liaisons covalentes, des angles de valence et de torsion, r et r_0 : longueurs des liaisons covalentes mesurées et de référence, θ et θ_0 : valeurs des angles de valence mesurées et de référence, φ et φ_0 : valeurs des angles de torsion mesurées et de référence, $\mathcal{E}$: profondeur du puits de potentiel, σ_{ij} : rayons de van der Waals, r_{ij} : distance interatomique, $\mathcal{E}_0$: constante diélectrique, q_i et q_j : charges des atomes i et j).*[364]

Les fonctions de score peuvent reposer sur différents champs de force utilisant des constantes spécifiques, des termes supplémentaires ou se limitant au calcul des interactions stériques et électrostatiques.[27] Parmi les logiciels de docking, AutoDock[361] et DOCK (version 6)[347] reposent sur le champ de force AMBER,[365,366] GOLD[93] repose sur le champ de force Tripos[367] et ICM[359] repose sur le champ de force ECEPP/3.[368] La fonction de score hybride d'ICM[359] ajoute notamment des termes empiriques permettant de prendre en compte l'entropie des angles de torsion et de considérer les liaisons hydrogènes de manière géométrique. GOLD[93] et AutoDock[361] utilisent également des termes spécifiques aux liaisons hydrogènes.

2.4.6.3.2. Fonctions empiriques

Les fonctions de score empiriques sont conçues pour reproduire des valeurs d'affinité expérimentales en utilisant des termes pondérés et calibrés par des méthodes de régression statistique ou d'apprentissage sur des données de complexes protéine-ligand co-cristallisés. Si ces fonctions présentent l'avantage d'être plus facilement interprétables que celles basées sur les champs de force, elles peuvent toutefois être très dépendantes des jeux de données utilisés pour leur calibration.[369] La nature, la forme et les coefficients des termes implémentés varient selon les fonctions de score. Par exemple, le logiciel LUDI[296,297] utilisait initialement deux termes pour la considération isolée des liaisons hydrogènes neutres et des liaisons hydrogènes ioniques, tandis que la fonction ChemScore[370] utilise un terme unique. LUDI[296,297] prend en compte les interactions hydrophobes grâce à un calcul des surfaces accessibles au solvant, tandis que ChemScore[370] évalue la proximité de paires d'atomes hydrophobes. La fonction F-Score[348] utilise un terme spécifique aux interactions aromatiques. Différents modèles ont également été proposés pour le calcul des énergies de solvatation et de désolvatation.[371–373] Parmi les développements récents, la fonction ID-Score[374] a été calibrée sur 2278 complexes co-cristallisés et utilise 50 descripteurs réparties en 9 catégories : interactions de van der Waals, électrostatiques, aromatiques, hydrogènes et métal-ligand, effets de désolvatation, entropie, complémentarité et propriétés des surfaces moléculaires.

2.4.6.3.3. Fonctions basées sur les connaissances

Les fonctions de score basées sur les connaissances (« knowledge-based ») sont également construites grâce à des données de co-cristallisation et d'affinité expérimentales, selon une méthodologie différente. Dans un premier temps, une analyse statistique permet de déterminer les fréquences et les distributions de distances de chaque type de paires d'atomes intervenant dans une interaction récepteur-ligand. Ces données permettent d'obtenir des potentiels

d'énergie distance-dépendants pour chaque type de chaque paire atomique. Un score d'interaction peut ensuite être donné par la somme des potentiels énergétiques attribués à chaque contact interatomique entre une molécule et son récepteur, en fonction des distances de contact et par rapport à une distribution de référence calculée pour des contacts atomiques internes à une protéine (fonction de distribution radiale) (***Équation 7***). Ce type de fonctions de score permet notamment de modéliser de manière implicite des phénomènes difficiles à quantifier explicitement, dont les effets entropiques et les énergies liées au solvant.[375–378] Un autre avantage des fonctions de score basées sur les connaissances est leur vitesse de calcul qui les rend particulièrement adaptées au criblage de larges chimiothèques. DrugScore,[379] SMoG,[380] FlexX[348] et PMF[375–378] font partie des plus utilisées. PMF[375–378] a été déclinée dans de nombreuses formes et calibrée en utilisant jusqu'à 44.585 complexes protéine-ligand co-cristallisés.[377]

$$Score\ PMF = \sum_{contacts} A_{ij}(r) = \sum_{contacts} -k_B T \ln\left(f_{vol}(r)\frac{\rho^{ij}(r)}{\rho^{ij}_{bulk}}\right)$$

Équation 7. *Equation de la fonction de score de PMF dans sa version initiale.*[376] *Le quotient $\rho^{ij}(r)/\rho^{ij}_{bulk}$ correspond à une fonction de distribution radiale décrivant la probabilité d'observer une paire d'atomes impliquée dans une interaction récepteur-ligand en fonction de leur distance et par rapport à une distribution de distances de référence ($A_{ij}(r)$: énergie attribuée à un contact interatomique ij en fonction de sa distance (r), k_B : constante de Boltzmann, T : température absolue du système, $\rho^{ij}(r)$: densité d'une paire d'atomes ij en fonction de leur distance, calibrée sur des complexes récepteur-ligand, ρ^{ij}_{bulk} : densité d'une paire d'atomes ij en fonction de leur distance, calibrée sur des protéines de référence non-complexées, $f_{vol}(r)$: facteur de correction du volume du ligand).*[376]

2.4.6.3.4. Scores consensus

Bien que l'utilisation des fonctions de score présente un intérêt certain, chaque type de fonction de score possède ses imperfections et aucune d'entre elles ne permet actuellement d'attribuer à coup sûr le meilleur score d'affinité à la conformation bioactive d'un ligand parmi les différentes poses proposées par les logiciels de docking.[27,33,335] Ainsi, l'utilisation de méthodes consensus a été proposée pour combiner et tirer le meilleur profit de plusieurs logiciels de docking et/ou fonctions de score appliquées dans le cadre d'une même étude. L'hypothèse sous-jacente est que la probabilité qu'une molécule soit active doit augmenter si cette molécule est associée à de bons scores d'affinité selon plusieurs fonctions de score. De plus, puisque les fonctions de score prennent en compte différents aspects des interactions récepteur-ligand, celles-ci peuvent se compléter pour mieux décrire une interaction. Deux types d'approches

peuvent être différenciés : (i) le consensus de plusieurs fonctions de score appliquées aux poses obtenues lors d'une expérience de docking unique et (ii) le consensus de plusieurs expériences de docking utilisant chacune leur fonction de score.[381–383]

Model	Score A		Score B		Score C		Score D		Consensus
	Score	Rank score	Score	Rank score	Score	Rank score	Score	Rank score	
1	12	**0**	55	**0**	43	**0**	241	**0**	0
2	22	**0**	46	**0**	113	**0**	283	**1**	1
3	112	**1**	92	**1**	221	**1**	299	**1**	**4**
4	78	**0**	82	**1**	182	**0**	251	**0**	1
5	98	**1**	77	**0**	192	**1**	263	**0**	2

Tableau 8. *Illustration du fonctionnement d'une méthode consensus d'intersection avec 5 composés, 4 méthodes de docking et un score binaire positif pour les composés présents dans les 40% de tête d'un classement.*[382]

La première application d'un consensus aux méthodes de docking a été décrite par Charifson et al. selon une approche d'intersection.[382] Pour chaque méthode de docking utilisée, les composés sont d'abord classés selon leurs scores d'affinité en considérant uniquement la meilleure des poses obtenues. Un score binaire est ensuite attribué à chacun en fonction de sa présence ou absence dans un pourcentage des composés les mieux classés, indépendamment pour chaque fonction de score. Le score consensus obtenu correspond à la somme de ces scores binaires (***Tableau 8***). Dans le cadre de cette étude, cette approche permit de sélectionner des fractions de composés pour lesquelles le taux d'enrichissement en actifs est toujours supérieur aux taux obtenus sur des fractions de mêmes tailles avec l'usage de fonctions de score indépendantes.[382] Cette procédure peut toutefois produire des intersections consensus de petites tailles, particulièrement lorsque : (i) le pourcentage de composés sélectionnés dans chaque classement est faible, (ii) ces classements se recouvrent peu ou (iii) un grand nombre de méthodes de docking ou de fonctions de score est utilisé.[382] En général, les taux d'enrichissement ainsi obtenus sont satisfaisants bien que le taux d'actifs total retrouvés reste faible.[381–383]

D'autres approches permettent de calculer un consensus pour réordonner une liste complète de molécules. Par exemple, il est possible de réaliser la somme des rangs de chaque composé dans les classements obtenus par différentes méthodes de docking (« sum-ranks »), le minimum des rangs (« min-rank », seul le meilleur rang est conservé) ou la somme des meilleurs rangs (« deprecated sum-ranks », le plus haut rang d'un composé est éliminé et la moyenne des rangs est calculée).[381] Pour l'approche « deprecated sum-ranks », lorsque le consensus utilise de nombreuses méthodes de docking, il est possible d'éliminer les deux ou trois plus hauts rangs.

Ces consensus peuvent également être appliqués directement sur les scores d'affinité attribués aux composés (après normalisation) ou leurs z-scores (nombre de déviations standard d'un score par rapport à leur moyenne). L'usage de régressions statistiques a également été décrit, avec des performances comparables aux différents types de consensus précédemment cités.[384,385] Les méthodes consensus sont également discutées dans nos résultats en point 1.2.

2.4.6.4. Principaux logiciels de docking

Parmi les nombreux logiciels disponibles, AutoDock,[361] GOLD[93] ou GLIDE[345] ont été utilisés dans plus de 50% des études de docking publiées entre 1990 et 2013.[386] Une liste exhaustive des logiciels de docking a été compilée en 2015 par Chen (extrait en ***Tableau 9***).[386]

Logiciel	Récepteur flexible	Algorithme de recherche conformationnelle	Algorithme de recherche de pose	Fonction de score
DOCK v6[347]	Oui	Systématique	Fragmentation reconstruction	Basée sur un champ de force
FlexX[348]	Non	Systématique	Fragmentation reconstruction	Basée sur les connaissances
Surflex-dock[350]	Non	Systématique	Fragmentation reconstruction	Empirique
PSI-DOCK[387]	Non	Stochastique	Génétique Tabou[388]	Empirique
AutoDock[361]	Oui	Stochastique	Génétique	Basée sur un champ de force
GOLD[93]	Oui	Stochastique	Génétique	Basée sur un champ de force
GLIDE[345]	Non	Stochastique	Monte Carlo	Empirique
FRED[339] *	Non	Systématique	Exhaustif	Empirique
ICM[359]	Oui	Stochastique	Monte Carlo	Basée sur un champ de force / Empirique
MCDOCK[353]	Non	Stochastique	Monte Carlo	Basée sur un champ de force
ProDOCK[354]	Oui	Stochastique	Monte Carlo	Basée sur un champ de force
AADS[355]	Non	Stochastique	Monte Carlo	Basée sur les connaissances
QXP[356]	Non	Stochastique	Monte Carlo	Basée sur un champ de force
ParDock[357]	Non	Stochastique	Monte Carlo	Basée sur un champ de force
DockVision[358]	Non	Stochastique	Monte Carlo	Basée sur un champ de force

Tableau 9. *Principaux logiciels permettant la mise en œuvre de docking rigide ou flexible de molécules. (*) FRED*[339] *est conçu uniquement pour un docking de conformations rigides.*[27,386]

2.4.6.5. Limites des méthodes de docking

Dans la vaste majorité des cas, les études de criblage virtuel sont conduites sur de larges chimiothèques de composés en maintenant un récepteur rigide et en simulant la flexibilité des molécules criblées.[44] Bien que cette approche présente un grand intérêt pour procéder au premier filtrage d'une chimiothèque, les modèles utilisés ne réalisent qu'une approximation de des phénomènes biologiques de sélection de conformation, des effets « induced-fit » et des interactions qui régissent l'affinité d'un ligand pour le site de liaison d'une protéine.[249,340–342] Le choix et la paramétrisation d'un logiciel de docking doivent donc être adaptés à la cible et à la chimiothèque étudiée.

2.4.6.5.1. Flexibilité du site de liaison

La structure d'un site de liaison est directement affectée par la flexibilité d'une protéine à petite comme à large échelle. Les changements conformationnels observés sont généralement limités aux mouvements des chaînes latérales des acides aminés du site de liaison. Cependant, ils peuvent également consister en une restructuration ou déstructuration d'une partie de la protéine, ce qui peut moduler le volume du récepteur (***Figure 58***). A l'extrême, les protéines très flexibles constituent des cibles difficiles à étudier avec les méthodes de docking du fait de la complexité de la paramétrisation et des temps de calcul nécessaires à leur succès.

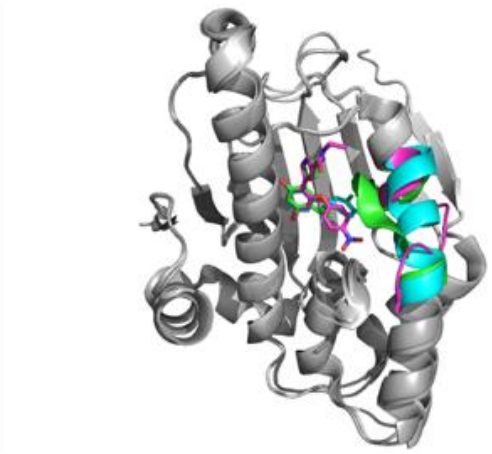


Figure 58. *Superposition de 3 structures cristallographiques de la protéine de choc thermique 90 (HSP90) illustrant différents degrés de déstructuration d'une hélice α impliquée dans le site de liaison (PDB*[196] *4lwe (vert), 4bqg (cyan) et 4bqj (rose)). La structure de cette hélice a un impact direct sur le mode de liaison des ligands puisque celle-ci peut être orientée soit vers l'intérieur, soit vers l'extérieur du site de liaison.*

Différentes approches permettent de prendre en compte la flexibilité d'une protéine ou de son site de liaison. Les méthodes de dynamique moléculaire considèrent explicitement l'ensemble

du système protéine-ligand comme flexible au prix de calculs importants (***Figure 59A***). Celles-ci ne peuvent donc pas être appliquées à large échelle et sont généralement utilisées pour une dernière étape de validation des poses ou pour le docking de fragments lors de courtes simulations. Parmi les approches alternatives, plus économes en calculs, il est possible de : (i) traiter seulement quelques résidus du site de liaison comme flexibles, (ii) utiliser plusieurs conformations de la protéine cible (« Ensemble docking ») ou (iii) utiliser une fonction de score permissive concernant les conflits stériques (« Soft docking »).[389]

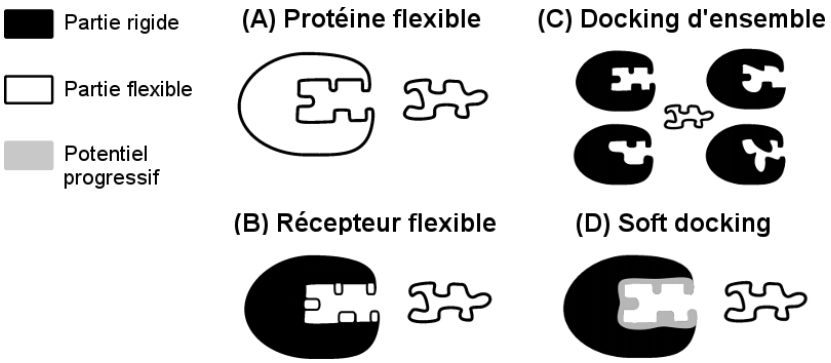


Figure 59. *Schématisation des différentes approches permettant de prendre en compte la flexibilité d'une protéine ou de son récepteur.*[389]

Lorsque l'on souhaite simuler uniquement la flexibilité des chaînes latérales, quelques résidus du site de liaison peuvent être définis comme explicitement flexibles au cours de la simulation (***Figure 59B***). Différents angles de valence et de torsions de ces chaînes latérales pourront alors être explorés par les logiciels, similairement à ce qui est réalisé pour explorer les conformations moléculaires. Il est également possible d'utiliser des banques de rotamères pour limiter le nombre des conformations étudiées à celles qui sont le plus fréquemment observées expérimentalement.[390–394] Une autre approche, proposée notamment par le logiciel FlexE,[395] permet d'aligner plusieurs structures d'une protéine pour regrouper différents types de mouvements des chaînes latérales et considérer uniquement leurs instances représentatives lors du docking. Cette procédure nécessite toutefois d'utiliser des structures pour lesquelles le squelette carboné est très similaire.[395]

L'approche de docking d'ensemble permet d'utiliser un ensemble de conformations qui présentent de plus larges variations pour procéder à un docking selon chaque état du récepteur (***Figure 59C***). Ces conformations peuvent être choisies par l'étude de différentes structures cristallographiques, de données de RMN ou générées par dynamique moléculaire.[344] La première approche de docking d'ensemble a été proposée en 2007 et repose sur une procédure de recherche de pose optimisée pour choisir à moindre coût computationnel la conformation du

récepteur la plus apte à accommoder une molécule donnée.[396] Un docking « classique » est ensuite réalisé en utilisant cette conformation.[396] Une autre approche plus coûteuse mais toutefois efficace consiste à réaliser le docking de chaque molécule sur chaque conformation de la protéine cible, puis sélectionner la pose ayant obtenu le meilleur score d'activité pour chaque molécule.[397]

Enfin, le principe du « soft docking » prend en compte la flexibilité du récepteur lors de la phase de calcul des scores d'affinité (***Figure 59D***).[398,399] Le potentiel de Lennard-Jones des fonctions de score utilisées, habituellement 12-6, est ici remplacé par un potentiel 9-6 ou par d'autres termes plus permissifs autorisant des conflits stériques mineurs (voir point 2.3.4.2 et ***Figure 34***). Cette approche permet ainsi de simuler un mouvement de retrait des chaînes latérales vers l'intérieur du site de liaison lors de la fixation d'un ligand.[398,399]

2.4.6.5.2. Rôle du solvant

D'autre part, la majorité des logiciels de docking ignorent les effets liés au solvant dans leur génération des poses moléculaires et leur calcul des scores d'affinité. Cependant, le solvant peut interagir naturellement avec une molécule en formant des liaisons hydrogènes avec celle-ci. Lors de la fixation d'un ligand dans un site de liaison, de nombreuses molécules d'eau qui sont liées à ces deux entités doivent donc rompre leurs interactions et laisser place à la formation de nouvelles interactions récepteur-ligand (***Figure 60***).

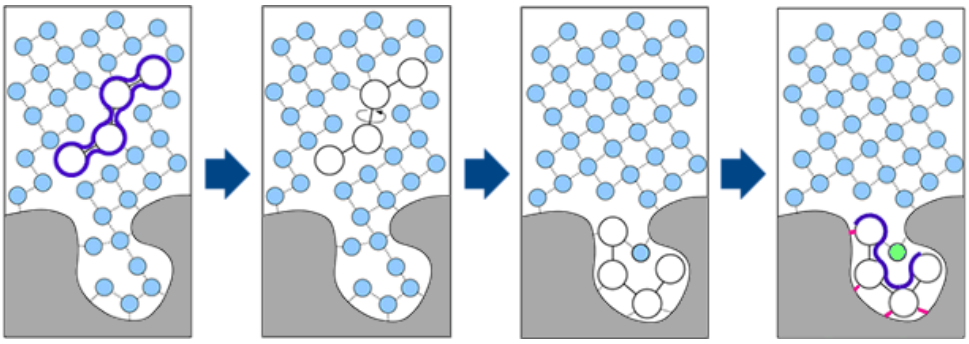


Figure 60. *Illustration de la fixation d'un ligand dans un site de liaison en présence de solvant. Une molécule d'eau structurale (vert) intervient dans le mode de liaison (violet : surface accessible au solvant du ligand, bleu : molécules d'eau, rouge : interactions récepteur-ligand).*[400]

Les méthodes de dynamique moléculaire permettent de prendre en compte ces phénomènes de désolvatation de manière explicite mais ne sont pas adaptées au criblage à large échelle. Ainsi, des modèles de mécanique moléculaire tels que PB/SA (Poisson-Boltzmann Solvant Area)[401] et GB/SA (Generalized Born Solvant Area)[402] ont été développés afin de considérer le solvant

de manière implicite. Ceux-ci permettent d'ajouter au calcul des scores d'interaction des termes pénalisants en fonction de la taille des surfaces moléculaires désolvatées ou du nombre et des énergies des molécules d'eau dont l'interaction est rompue.[403] Ces modèles font régulièrement l'objet d'optimisations afin de réduire les coûts calculatoires nécessaires à leur application dans le cadre de criblages haut débit, mais peuvent déjà être utilisés en routine pour affiner ou valider les prédictions d'énergies d'interaction de « hits ».[404–407]

Certaines molécules d'eau, que l'on nomme molécules d'eau structurales, peuvent toutefois rester impliquées dans une interaction récepteur-ligand et jouent un rôle de « relai » des liaisons hydrogènes tout en participant à l'organisation tridimensionnelle du site de liaison (***Figure 60***). Une étude menée par Klebe en 2015 démontre que des ligands très similaires, adoptant des poses très proches dans un même site de liaison, peuvent accepter ou déplacer des molécules d'eau structurales en fonction de l'énergie libre des poses obtenues (***Figure 61***).[408]

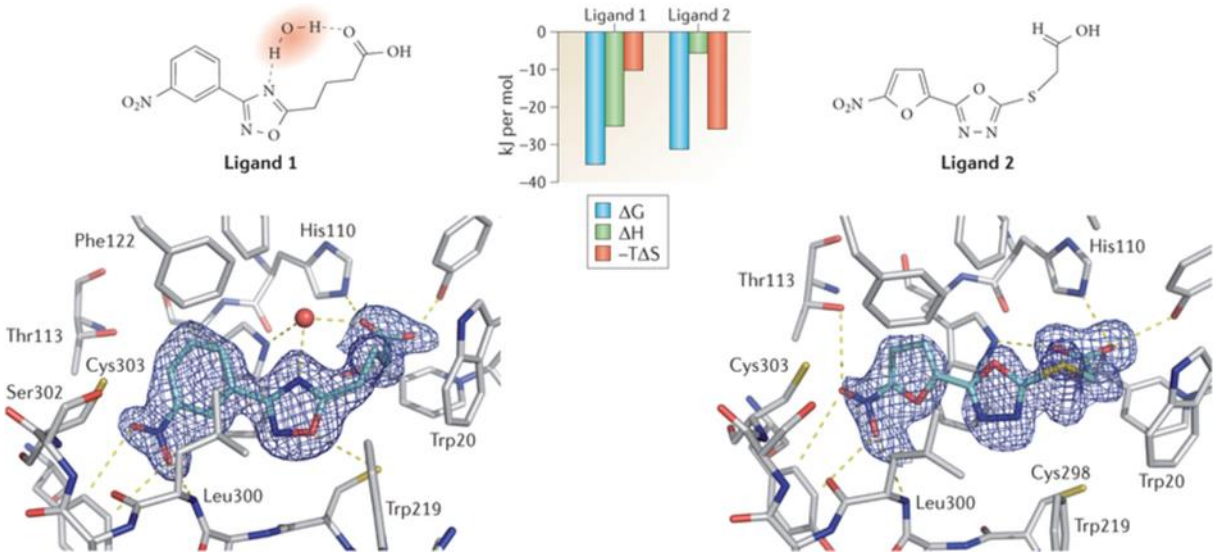


Figure 61. *Poses cristallographiques de deux ligands dans le site de liaison de l'aldose réductase (ALR2). La pose du ligand 1 comprend une interaction avec une molécule d'eau structurale qui diminue l'enthalpie du système. Le ligand 2 déplace cette molécule d'eau et permet de diminuer l'entropie de cette pose. Les énergies libres des poses des ligands 1 et 2 sont proches.*[408]

Lorsque des données cristallographiques, par exemple, indiquent qu'une molécule d'eau est présente de manière stable dans un site de liaison, celle-ci peut être explicitement considérée comme faisant partie du récepteur lors du docking. La majorité des méthodes de docking permettent d'ajouter des molécules d'eau structurales à un récepteur en les considérant comme figées dans un récepteur rigide, ce qui limite les performances de ces approches.[409,410] Le logiciel rDock[411] pallie ce problème en laissant à l'utilisateur le choix des positions initiales et

des paramètres de mobilité des molécules d'eau structurales, tandis que SLIDE[412] utilise une approche basée sur les connaissances pour identifier les molécules d'eau susceptibles d'être conservées dans un complexe récepteur-ligand et pénalise les poses qui déplacent des molécules d'eau considérées comme structurales. Le logiciel FlexX[348] propose une option qui autorise l'intégration de molécules d'eau dans le processus de construction incrémentale des poses. Les positions des molécules d'eau structurales peuvent être précisées par l'utilisateur ou prédites par FlexX.[348] La construction incrémentale des poses est ensuite réalisée en fonction des interactions récepteur-ligand établies et des entropies des deux entités selon différentes combinaisons de présence des molécules d'eau structurales, permettant ainsi de déterminer les meilleures poses en présence de solvant.[348]

2.4.7. Succès du criblage virtuel « structure-based »

Les différentes approches « structure-based » précédemment citées font désormais partie intégrante du processus de développement de nouveaux médicaments.[413] Les criblages virtuels, qu'ils soient réalisés par des méthodes de docking, des approches *de novo* ou à moindre échelle grâce à des modèles pharmacophoriques, permettent d'augmenter le nombre de « hits » obtenus expérimentalement pour amorcer ce processus. Les modèles pharmacophoriques et QSAR constituent ensuite un outil précieux pour l'optimisation des « leads ». Le ***Tableau 10*** recense 29 exemples de médicaments commercialisés entre 1995 et 2009 pour lesquels l'étude de la structure de la cible et des relations structure-activité des composés a joué un rôle crucial.[413]

Par ailleurs, le succès de ces approches rationnelles a conjointement donné lieu à une modernisation des moyens de protection de la propriété intellectuelle.[414] Jusque récemment les brevets utilisaient majoritairement des structures de Markush et des descriptions de classes de composés pour protéger l'utilisation commerciale d'une molécule spécifique dans un ou plusieurs cadres thérapeutiques définis. Désormais, il est également possible d'utiliser une description pharmacophorique géométrique pour revendiquer un ensemble de composés à ces fins.[414] Le premier brevet de ce genre a été déposé en 1997 et utilise plusieurs descriptions pharmacophoriques couplées à des structures de Markush (WO 98/004913).[415] Quelques dizaines de brevets de ce type ont été déposés depuis (WO 98/046630,[416] US 7897162,[417] etc.).

Nom du composé	Année	Cible	Indication	Référence
Dorzolamide	1995	Anyhdrase carbonique	Glaucome	418
Saquinavir	1995	Protéase du VIH	SIDA	419
Ritonavir	1996	Protéase du VIH	SIDA	420
Indinavir	1996	Protéase du VIH	SIDA	421
Brinzolamide	1999	Anyhdrase carbonique	Glaucome	422
Nelfinavir	1999	Protéase du VIH	SIDA	423
Amprenavir	1999	Protéase du VIH	SIDA	424
Lopinavir	1999	Protéase du VIH	SIDA	425
Zanamivir	1999	Neuraminidases variées	Grippe	426
Oseltamivir	1999	Neuraminidases variées	Grippe	427
Imatinib	2001	Kinase BCR-ABL	Leucémie myéloïde chronique	428,429
Gefitinib	2003	EGFR	Cancer du poumon à grandes cellules	430
Atazanavir	2003	Protéase du VIH	SIDA	431
Fosamprenavir	2003	Protéase du VIH	SIDA	432
Erlotinib	2004	EGFR	Cancer du poumon à grandes cellules	433
Sorafenib	2005	VEGFR	Cancer du rein	434
Tipranavir	2005	Protéase du VIH	SIDA	435
Udenafil	2005	Phosphodiestérase 5 (PDE5)	Dysfonctionnements érectiles	436,437
Sunitinib	2006	Kinases variées	Cancer du rein	438
Darunavir	2006	Protéase du VIH	SIDA	439
Vorinostat	2006	Histone désacétylase	Lymphome T cutané	440
Dasatinib	2006	Tyrosine kinases variées	Leucémie myéloïde chronique	441
Nilotinib	2006	Kinase BCR-ABL	Leucémie myéloïde chronique	442
Aliskiren	2007	Rénine	Hypertension	443
Lapatinib	2007	Kinases variées dont HER2	Cancer du sein	444
Rivaroxaban	2008	Facteur X	Thromboses veineuses	445
Dabigatran	2008	Thrombine	Thromboses veineuses	446
Etravirine	2008	Transcriptase inverse	SIDA	447
Pazopanib	2009	Kinases variées dont VEGFR	Cancer des ovaires	448

Tableau 10. *Exemple de 29 médicaments commercialisés entre 1995 et 2009. Les références associées décrivent les méthodes « structure-based » engagées dans leurs développements.*[413]

3. Evaluation des méthodes de criblage virtuel

Comme nous l'avons vu précédemment, de nombreux paramètres conditionnent le succès d'une expérience de criblage virtuel. Leur mise en œuvre requiert donc une bonne connaissance du système étudié, dont principalement la flexibilité de la protéine cible, les caractéristiques de la chimiothèque utilisée, les algorithmes de génération des poses et les fonctions de score appliquées. Puisque les résultats obtenus par criblage virtuel restent des prédictions et que leur fiabilité peut parfois être mise en doute, il est crucial de s'assurer des performances des différentes méthodes afin d'obtenir des résultats optimaux lors des études expérimentales réalisées en aval.

L'évaluation des méthodes (« benchmarking ») permet de guider les bioinformaticiens dans le choix des protocoles performants sur un système donné. Cette évaluation est réalisée de manière rétrospective principalement grâce à des banques d'évaluations construites à ces fins.[29,158,159] Il est également possible d'utiliser des résultats expérimentaux précédents (HTS ou autres données d'activité quantifiées) pour évaluer la capacité d'un logiciel à prédire des scores d'affinité proches des données recueillies expérimentalement. Cependant, les données de HTS sont rarement mises à disposition librement puisqu'elles représentent des coûts importants et un atout conséquent dans le processus de conception de médicaments.[335,449] Les données de co-cristallographie protéine-ligand permettent également d'évaluer la capacité d'une méthode à repositionner un ligand selon son mode de liaison biologique. Ainsi, il est possible de comprendre pourquoi certaines prédictions des criblages virtuels peuvent être biaisées, soit par un mauvais positionnement des molécules dans le récepteur ciblé, soit par un calcul des scores d'affinité qui restituerait mal les énergies d'interaction réelles.[29,158,159]

3.1. Précision du positionnement des composés

La précision d'une pose prédite peut être évaluée par différentes métriques en comparaison à des données de co-cristallographie protéine-ligand.[450] La première et la plus utilisée de ces métriques est l'écart quadratique moyen (Root Mean Square Deviation ou RMSD) entre les coordonnées atomiques de deux molécules dans un même référentiel.[451,452] D'autres métriques peuvent également être utilisées, comme l'erreur relative de déplacement (Relative Displacement Error ou RDE)[453] ou l'espace réel du facteur R (Real Space R-factor ou RSR).[454]

3.1.1. Root Mean Square Deviation (RMSD)

Cette mesure géométrique est basée sur la distance entre les positions atomiques d'une pose prédite et d'une pose co-cristallisée d'un ligand (***Équation 8***).[450] Il est ainsi possible d'évaluer la superposition des deux poses, qui doit idéalement être parfaite. Plus la valeur de RMSD est faible, plus la pose prédite est proche de celle du ligand co-cristallisé.

$$RMSD(A,B) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}((A_{ix} - B_{ix})^2 + \left(A_{iy} - B_{iy}\right)^2 + (A_{iz} - B_{iz})^2)}$$

Équation 8. *Formule du RMSD entre deux composés A et B (x, y et z : coordonnées cartésiennes, n : nombre d'atomes du composé étudié).*[450]

En général, pour évaluer la précision du positionnement, seuls les atomes lourds sont considérés.[455] Cette métrique a pour principaux avantages d'être sensible, simple d'utilisation et automatisable. Sa simplicité amène toutefois quelques désavantages : (i) le RMSD ne fournit aucune information sur les interactions établies avec le récepteur et n'indique donc pas si la pose évaluée établie les mêmes interactions que le ligand co-cristallisé et (ii) il est également très sensible aux orientations symétriques d'un composé, puisqu'il utilise une numérotation prédéfinie des atomes du composé étudié.[456] Ainsi, pour un ligand symétrique, la comparaison de deux poses symétriquement inversées dans un site de liaison peut produire des valeurs de RMSD importantes (***Figure 62***). L'usage de ce seul RMSD lors d'une évaluation automatisée indiquerait alors une mauvaise prédiction du positionnement du ligand alors que les interactions clés de son mode de liaison peuvent être conservées.

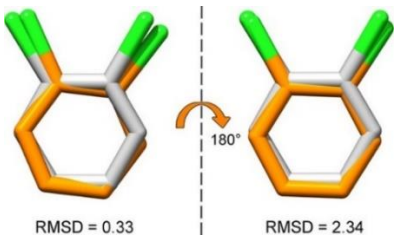


Figure 62. *Illustration du problème de symétrie dans le calcul de RMSD pour le composé 1,2-dichlorobenzene. La pose de référence co-cristallisée (gris) et la pose amarrée (orange) sont inversées par rapport à un axe de 180°.*[456]

D'autres effets peuvent biaiser l'interprétation de la valeur du RMSD, qui est notamment dépendant du poids moléculaire des composés. Ainsi, les petits composés sont fréquemment associés à de faibles valeurs de RMSD tandis que des composés plus larges, bien que positionnés correctement dans une poche, obtiendront des valeurs de RMSD plus hautes.[455]

Cette métrique très utile doit donc être utilisée avec recul et idéalement couplée à une observation humaine du positionnement des composés.

3.1.2. RMSD corrigé pour la symétrie

Dans le cas de molécules symétriques, une correction du RMSD est nécessaire concernant la définition des couples d'atomes à comparer. Deux types de corrections existent : le RMSD de distance minimale et le RMSD de correspondance optimale.

3.1.2.1. RMSD de distance minimale

Trott & Olson introduisent une première correction du RMSD en définissant les couples d'atomes à comparer sur le critère de la distance minimale (***Équation 9***).[457] Dans cette méthode, les atomes a_i de la pose de référence A sont comparés itérativement à tous les atomes b_j de la pose B qui sont du même élément chimique. Les couples d'atomes (a_i, b_j) à sont définis comme ayant une distance minimale. Cette procédure est ensuite répétée en utilisant la pose B comme référence. Le RMSD de distance minimale entre les deux poses A et B est le maximum des deux RMSD calculés.

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

Avec

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

Équation 9. *Formule de calcul du RMSD de distance minimale (a_i et b_j : positions des composés A et B, n : nombre d'atomes du composé). Les atomes sélectionnés pour établir une distance minimale doivent être du même élément chimique.*[457]

Cette méthode permet de corriger le problème de la symétrie dans le calcul du RMSD. Cependant, elle comporte des inconvénients majeurs : un atome peut être comparé à plusieurs autres atomes lorsqu'il est impliqué dans plusieurs couples de correspondance et inversement, certains atomes peuvent être ignorés s'ils ne sont présents dans aucun couple de correspondance.

3.1.2.2. RMSD de correspondance optimale

Le RMSD de correspondance optimale, plus abouti, repose sur l'utilisation de l'algorithme hongrois pour déterminer les couples optimaux de correspondance atomique (***Équation***

10).[456,458] Celui-ci a pour avantages de ne négliger aucun atome et de corriger efficacement les problèmes associés à l'évaluation du positionnement des composés symétriques (***Figure 63***).[456]

$$RMSD_{corr.opt}(A,B) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(cor\|a_i - b_j\|)^2}$$

Équation 10. *Formule de calcul du RMSD de correspondance optimale (a_i et b_j : positions atomiques des composés A et B, n : nombre d'atomes du composé étudié). Les correspondances atomiques optimales sont déterminées par la fonction cor dont l'algorithme est le suivant : pour chaque atome de la pose A, son unique correspondant de la pose B est celui qui minimise la somme des distances entre toutes les paires d'atomes possibles.*[456]

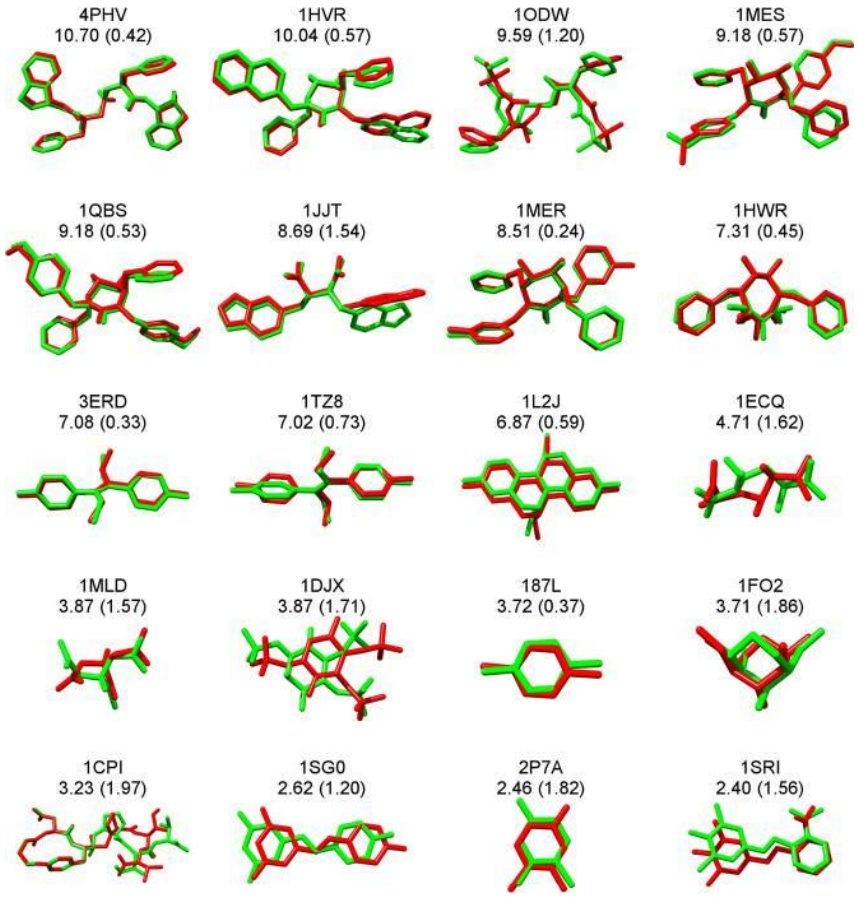


Figure 63. *Illustration des différences entre RMSD et RMSD de correspondance optimale par la comparaison de 20 composés amarrés (vert) à leurs poses cristallographiques de référence (rouge). Le RMSD de correspondance optimale corrige les problèmes associés à la symétrie. Le code PDB est indiqué en haut, suivi du RMSD non corrigé (Å). Le RMSD de correspondance optimale est indiqué entre parenthèses (Å).*[456]

3.2. Enrichissement d'une chimiothèque

Les méthodes de criblage virtuel doivent permettre de différencier les composés capables de se lier à la cible thérapeutique, dits « actifs », du reste des composés de la chimiothèque utilisée, dits « inactifs ». Dans un scénario idéal, après avoir prédit le mode d'interaction correct d'un ligand dans le site de liaison, le calcul du score d'affinité entre le ligand et son récepteur indique une interaction favorable. Inversement, les composés inactifs n'arriveront pas à être positionnés sur le récepteur ou seront placés selon un mode de liaison improbable, aboutissant au calcul de scores d'interaction défavorables. Les scores obtenus sur l'ensemble de la chimiothèque étudiée permettront d'établir une liste classée des composés en fonction de leur potentiel d'interaction avec la cible. Lorsque l'on retrouve un fort taux d'actifs dans la première fraction de cette liste, idéalement très supérieur au taux d'actifs dans l'ensemble de la chimiothèque, la méthode de criblage est considérée comme efficace. Les différents logiciels de criblage sont évalués principalement rétrospectivement sur les données de banques d'évaluation grâce à différentes métriques de performance.

3.2.1. Banques d'évaluation

Les banques d'évaluation rassemblent plusieurs protéines dont la structure cristallographique est disponible et pour lesquelles de nombreux composés actifs et inactifs ont été répertoriés. Les composés actifs sont sélectionnés à partir de données expérimentales validant leur activité sur une cible. Concernant les composés inactifs, puisque les données d'inactivité ne sont généralement pas mises à disposition, l'usage est d'utiliser des molécules « leurres » (« decoys ») présumées inactives.[455]

3.2.1.1. Composition des banques d'évaluation

La première banque d'évaluation a été proposée par Bissantz et al. en 2000 et comporte deux cibles : le récepteur aux œstrogènes α (ERα) et la thymidine kinase de type 1 (TK).[459] Pour chaque cible, le jeu de données d'évaluation est constitué d'une structure cristallographique extraite de la PDB,[196] de 10 ligands et de 990 decoys sélectionnés au hasard dans la base de données Advanced Chemical Directory (Accelerys) filtrée de manière à éliminer les composés réactifs, inorganiques ou de poids moléculaire inadéquat (inférieur à 250 Da ou supérieur à 500 Da). Bissantz et al. utilisent cette banque pour évaluer et comparer les performances de 3 logiciels de docking et de 7 fonctions de score.[459] La publication de cette étude a amorcé le développement d'autres banques d'évaluation avec l'objectif d'identifier les méthodes de

criblage virtuel les plus performantes, ainsi que les faiblesses des logiciels afin de guider leur développement. Les principales banques d'évaluation construites entre 2000 et 2014 sont recensées en ***Tableau 11***. Les banques récentes regroupent plusieurs dizaines de milliers de ligands et decoys et jusqu'à 102 cibles thérapeutiques populaires dans le cas de la DUD-E.[158]

Nom	Année	Origine des ligands (L)	Origine des decoys (D)	Ratio D/L	Accès
« Rognan's decoy set »	2000	Littérature	Advanced Chemical Directory	99	http://bioinfo-pharma.u-strasbg.fr/labwebsite/download.html
« Shoichet's decoy set »	2002	Littérature	MDDR	-	-
« Diller's decoy set »	2003	Littérature	MDDR	-	-
« Jain's decoy set »	2005	PDBbind	ZINC « drug-like »	-	http://www.jainlab.org/downloads.html
Directory of Useful Decoys (DUD)	2006	Littérature et PDBbind	ZINC « drug-like »	36	http://dud.docking.org
DUD Clusters	2007	Ligands de la banque DUD	Decoys de la banque DUD	-	http://dud.docking.org
WOMBAT Datasets	2007	WOMBAT	Decoys de la banque DUD	-	http://dud.docking.org
Charge Matched DUD	2010	Ligands de la banque DUD	ZINC	36	http://dud.docking.org
Virtual Decoys Sets (VDS)	2011	Ligands de la banque DUD	Molécules virtuelles	36	http://compbio.cs.toronto.edu/VDS
DEKOIS	2011	BindingDB	ZINC	30	http://www.dekois.com/dekois_orig.html
GPCR Ligand Library (GLL) / Decoy Database (GDD)	2011	GLIDA	ZINC	39	http://cavasotto-lab.net/Databases/GDD/Download/
DecoyFinder	2012	Fournis par l'utilisateur	ZINC	Défini par l'utilisateur	http://urvnutrigenomica-ctns.github.io/DecoyFinder
DUD Enhanced (DUD-E)	2012	ChEMBL	ZINC	50	http://dude.docking.org
DEKOIS 2.0	2013	BindingDB	ZINC	30	http://www.dekois.com
NRLiSt BDB	2014	ChEMBL et PubChem	ZINC et générateur de decoys DUD-E	50	http://nrlist.drugdesign.fr

Tableau 11. *Principales banques développées entre 2000 et 2014 pour l'évaluation des méthodes de criblage virtuel « structure-based ».*[460]

Puisque les banques d'évaluation représentent une référence pour le développement des logiciels, leur construction requiert une attention particulière. La sélection de structures cristallographiques adéquates constitue un premier point clé de la construction de banques d'évaluation de qualité, mais une grande attention doit également être portée au choix des

ligands et des decoys. Si la sélection des ligands est relativement aisée à partir des données de la littérature, les bases de données d'activité peuvent toutefois contenir des erreurs et une revue manuelle des données est préférable.[461,462] La sélection des decoys peut également biaiser l'évaluation des méthodes, notamment lorsque ceux-ci appartiennent à un espace chimique très différent de celui des ligands du jeu de données. Par exemple, une différence de poids moléculaire significative entre les ligands et les decoys peut biaiser une évaluation en augmentant ou en réduisant les taux d'enrichissement obtenus de manière artificielle.[463] La banque de données DUD[159] a été construite de sorte à limiter ces biais, en choisissant des decoys dont les propriétés physico-chimiques (poids moléculaire, logP, nombre de donneurs et accepteurs de liaisons hydrogène) sont proches de celles des ligands, mais dont les structures et topologies diffèrent. La banque de données VDS[464] propose l'utilisation de decoys virtuels, dont la synthèse chimique n'est pas systématiquement maîtrisée, afin de minimiser les différences des propriétés physico-chimiques entre les decoys et les ligands.

3.2.1.2. Limitations spécifiques aux banques d'évaluation

Malgré ces améliorations, les banques d'évaluation comportent toujours une faible diversité de cibles et les kinases sont généralement surreprésentées au détriment des autres familles de protéines (plus de 25% de kinases dans la banque DUD).[159] L'espace chimique couvert par les ligands de chaque banque est également limité par les procédés de validation des actifs, qui consistent généralement à tester expérimentalement un nombre restreint de classes de composés. Par ailleurs, les bases de données précédemment citées ont été construites pour évaluer la performance des logiciels à discriminer les composés actifs des inactifs et il n'existe à ce jour aucune banque publique permettant d'évaluer la capacité d'une méthode à ordonner les ligands en fonction de leurs affinités relatives pour une cible.[465] Une telle banque d'évaluation devrait renseigner : (i) les poses cristallographiques de nombreux ligands sur des cibles diverses et (ii) leurs affinités expérimentales déterminées dans des conditions similaires, afin de pouvoir évaluer les fonctions de score indépendamment des méthodes d'échantillonnage des poses. Bien que certains groupes pharmaceutiques commencent à publier des données de ce type, notamment au travers des concours CSAR (Community Structure-Activity Resource)[466] et D3R (Drug Design Data Resource),[467] il reste difficile de regrouper des données fiables et comparables pour construire une banque d'évaluation qui satisferait à ces critères.[461]

3.2.2. Métriques de performance

Plusieurs métriques permettent d'évaluer rétrospectivement la capacité d'une méthode de criblage virtuel à ordonner une chimiothèque en fonction des prédictions d'activité des composés. Les métriques peuvent être groupées en deux catégories : (i) les métriques globales, qui quantifient la qualité des prédictions sur l'ensemble d'un jeu de données et (ii) les métriques partielles, utilisées pour évaluer plus finement la meilleure fraction des prédictions.[455]

3.2.2.1. Courbes de ROC (Receiver Operating Characteristics)

Les courbes de ROC permettent de visualiser les fraction de vrais positifs et de faux positifs retrouvées à chaque rang d'une chimiothèque classée (***Figure 64***, ***Équation 11*** et ***Équation 12***).[468] Cette représentation permet de comparer la capacité de discrimination des méthodes sur une échelle commune, qu'elles soient appliquées à un unique ou à différents systèmes (cible et chimiothèque étudiée, etc.) (***Figure 65***). Lorsqu'une méthode n'est pas capable de distinguer les vrais positifs des faux positifs, les composés sont classés de manière aléatoire et la courbe de ROC associée correspond à une diagonale tracée entre les points (0,0) et (1,1) du graphique. Inversement, lorsqu'une méthode permet de discriminer parfaitement les ligands des decoys, la courbe de ROC associée atteint directement le point (0,1) avant de rejoindre le point (1,1).[468]

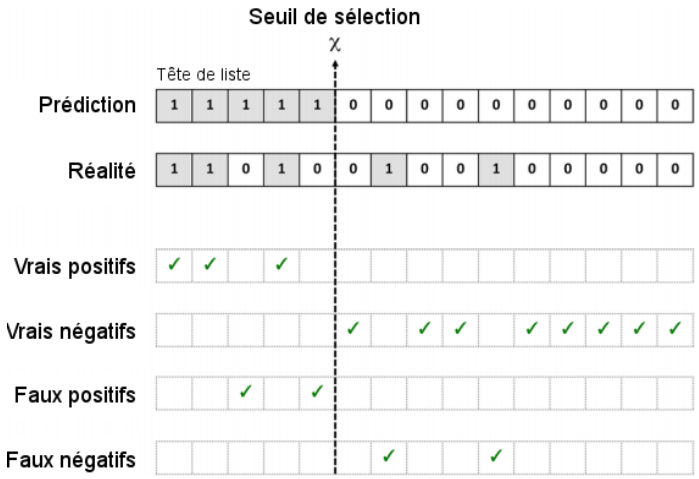


Figure 64. *Illustration du compte des vrais positifs (VP), vrais négatifs (VN), faux positifs (FP) et faux négatifs (FN) pour une fraction du jeu de données. Les ligands et decoys sont représentés de manière binaire, respectivement 1 et 0.*[455]

$$Se = FVP = \frac{VP_{fraction}}{VP_{total} + FN_{total}}$$

Équation 11. *Formule de calcul de la sélectivité (Se) pour une fraction d'un jeu de données, aussi appelée sensibilité ou fraction des vrais positifs (FVP).*[455]

$$Sp = FVN = \frac{VN_{hors\ fraction}}{VN_{total} + FP_{total}} = 1 - FFP$$

Équation 12. *Formule de calcul de la spécificité (Sp) pour une fraction d'un jeu de données, aussi appelée fraction des vrais négatifs (FVN) (FFP : fraction des faux positifs).*[455]

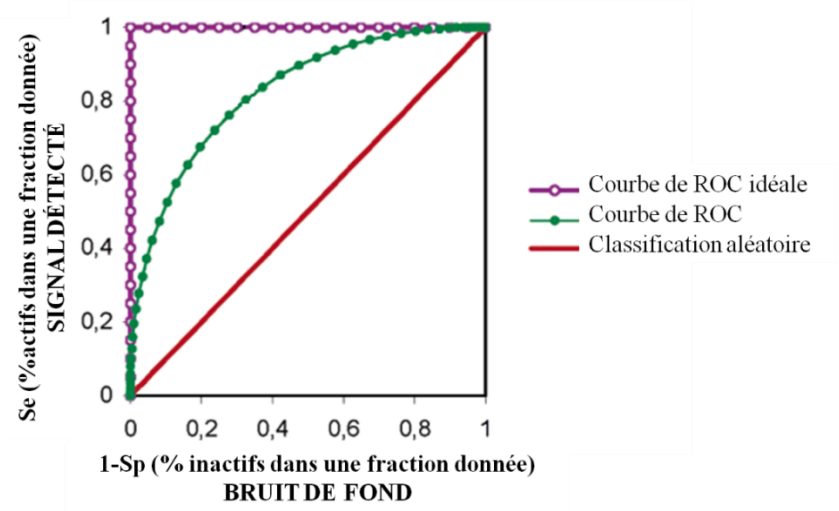


Figure 65. *Illustration de courbes de ROC dans le cas d'une classification aléatoire (rouge), supérieure à l'aléatoire (vert) ou parfaite (violet).*[468]

L'aire sous la courbe de ROC (Area Under the Curve ou AUC) résume la capacité de discrimination globale d'une méthode en un unique indicateur chiffré qui peut être utilisé pour comparer rapidement les performances de plusieurs méthodes de criblage virtuel (***Équation 13***). Une classification aléatoire produit une AUC égale à 0,5 tandis qu'une discrimination parfaite est caractérisée par une AUC égale à 1. Plus une AUC est proche de 1, plus la classification réalisée est discriminante. Cependant, les valeurs d'AUC ne reflètent pas les différences de classement des ligands (***Figure 66***). Pour évaluer correctement les méthodes, il est important d'inspecter visuellement les courbes de ROC et d'utiliser des métriques partielles permettant d'évaluer l'enrichissement en ligands de la première fraction d'une chimiothèque classée, également appelé reconnaissance précoce (« early recognition »).[455] Une AUC partielle (partial AUC ou pAUC) peut être calculée à partir d'une courbe de ROC pour quantifier la capacité de discrimination d'une méthode sur une fraction de la classification étudiée (***Équation 14***). Les AUC partielles permettent de répondre correctement au problème posé en ***Figure 66***.

$$AUC = \sum_{i} FVP_i(FFP_i - FFP_{i-1})$$

Équation 13. *Formule de calcul de l'aire sous la courbe de ROC (Area Under the Curve ou AUC) à partir des fractions des vrais positifs (FVP) et des faux positifs (FFP) retrouvées à chaque rang i de la chimiothèque classée.*[455]

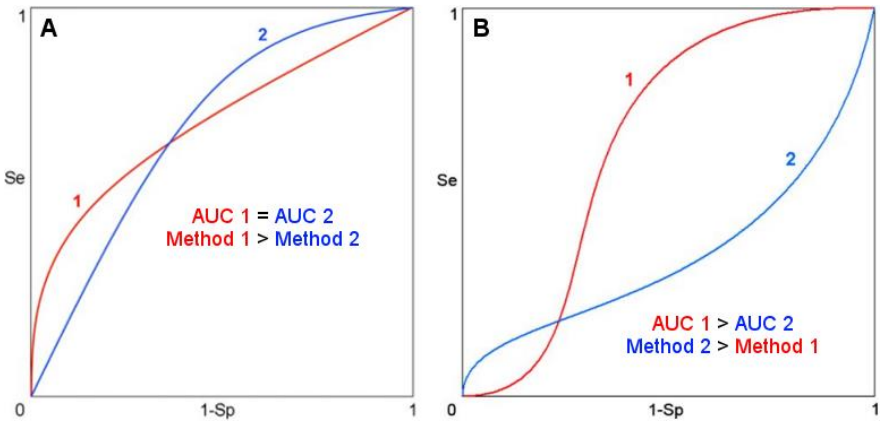


Figure 66. *Illustration du problème d'évaluation de la reconnaissance précoce en utilisant les seules valeurs d'AUC. (A) Les classifications 1 et 2 présentent des valeurs d'AUC identiques bien que la méthode 1 attribue de meilleurs rangs aux ligands. (B) La classification 1 présente une valeur d'AUC supérieure bien que la classification 2 comprenne plus de ligands dans sa première fraction.*[455]

$$pAUC(n) = \frac{\sum_{i=1}^{n} FVP_i(FFP_i - FFP_{i-1})}{FFP_n}$$

Équation 14. *Formule de calcul d'une aire partielle sous la courbe de ROC (partial Area Under the Curve ou pAUC) à partir des valeurs FVP et FFP de chaque fraction des composés à un rang i de la chimiothèque classée, jusqu'à un rang n. Les pAUC sont normalisées entre 0 et 1.*[455]

3.2.2.2. Facteurs et courbes d'enrichissement

Les facteurs d'enrichissement sont couramment utilisés pour évaluer la reconnaissance précoce des ligands et sont généralement couplés à l'usage des courbes de ROC pour l'évaluation des méthodes de criblage virtuel. Ils quantifient le gain obtenu dans une fraction de la chimiothèque classée, par rapport à une sélection aléatoire des composés, et sont habituellement calculés à 1, 2, 5 ou 10% du classement selon la formule suivante (***Équation 15***).

$$EF_{100(n/N)\%} = \frac{VP_{fraction}/n}{(VP_{total} + FN_{total})/N}$$

Équation 15. *Formule de calcul du facteur d'enrichissement (EF) pour la fraction des 100*n/N premiers pourcents d'une chimiothèque classée (n : nombre de composés dans la fraction étudiée, N : nombre de composés de la chimiothèque, VP : vrais positifs, FN : faux négatifs).*

Cette métrique présente néanmoins le désavantage d'être dépendante du taux de ligands de la chimiothèque étudiée, qui détermine sa valeur maximale.[469] Elle est donc adaptée à la comparaison des performances de différentes méthodes sur un même jeu de données, mais ne peut pas être utilisée pour comparer des performances obtenues sur chimiothèques présentant

des taux de ligands différents. La seconde faiblesse des facteurs d'enrichissement est liée au fait qu'ils utilisent le nombre de ligands retrouvés dans une fraction de la classification, sans tenir compte des rangs des ligands au sein de cette fraction. Ainsi, deux méthodes qui retrouvent le même nombre de ligands à un pourcentage fixe du jeu de données produisent des valeurs d'EF identiques, même si, au sein de la fraction considérée, une méthode classe les ligands en premier et l'autre en dernier (***Figure 67***).[469]

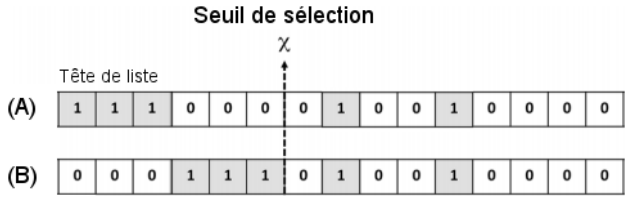


Figure 67. *Les méthodes A et B comprennent le même nombre de ligands dans la fraction sélectionnée. Les valeurs d'EF associées sont identiques, bien que la méthode A soit plus performante.*

Les courbes d'enrichissement peuvent également être utilisées pour comparer les performances des méthodes et estimer la reconnaissance précoce des ligands, grâce à la visualisation des fractions des vrais positifs retrouvées à chaque rang d'une chimiothèque classée sur une échelle logarithmique (***Figure 68***).

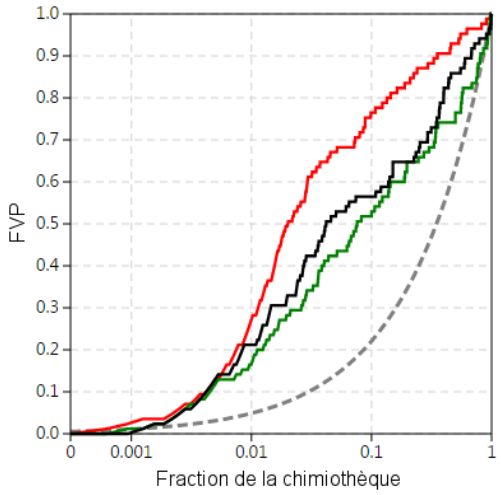


Figure 68. *Illustration de la comparaison de trois méthodes de criblage (couleurs) grâce à leurs courbes d'enrichissement (pointillés gris : classification aléatoire) (FVP : fraction des vrais positifs).*

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

Le RIE est une autre métrique permettant de quantifier la reconnaissance précoce des ligands, qui utilise une exponentielle décroissante pour pondérer les rangs d'une chimiothèque classée :

le poids des ligands classés en tête est proche de 1, puis diminue avec l'augmentation de leurs rangs.[470] Le RIE correspond au ratio de la somme des poids des ligands sur la moyenne de cette somme obtenue avec 1000 tests de classification aléatoire. Il quantifie donc le gain obtenu dans le classement de l'ensemble des ligands par rapport à une distribution aléatoire et prend directement en compte la reconnaissance précoce. Plus une valeur de RIE est supérieure à 1, plus une méthode de criblage virtuel est performante dans sa discrimination des ligands et des decoys. Une formule simplifiée du RIE a ensuite été proposée par Truchon et al. afin de s'affranchir des tests aléatoires (***Équation 16***).[469] Ceux-ci introduisent également la métrique BEDROC, qui correspond à une normalisation du RIE entre 0 et 1 pour faciliter la comparaison des méthodes (***Équation 19***).[469]

Les métriques RIE et BEDROC présentent toutefois deux faiblesses : (i) tout comme les facteurs d'enrichissement, celles-ci sont dépendantes du taux de ligands de la chimiothèque étudiée et ne doivent donc pas être utilisées pour comparer des classifications obtenues sur des jeux de données différents, et (ii) leur utilisation nécessite le choix d'un paramètre α qui détermine le poids attribué à la fraction précoce des composés.

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

Équation 16. *Formule de calcul du RIE (n : nombre de ligands, N : nombre de composés de la chimiothèque étudiée, x_i : rang normalisé, α : paramètre contrôlant le poids attribué à la fraction précoce des composés).*[469]

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

Équation 17. *Formule de calcul de la valeur minimale du RIE (R_a : taux de ligands dans la chimiothèque étudiée, α : paramètre contrôlant le poids attribué à la fraction précoce des composés).*

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

Équation 18. *Formule de calcul de la valeur maximale du RIE (R_a : taux de ligands dans la chimiothèque étudiée, α : paramètre contrôlant le poids attribué à la fraction précoce des composés).*

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

Équation 19. *Formule de calcul de la métrique BEDROC à partir de la valeur minimale (RIE_{min}), maximale (RIE_{max}) et mesurée du RIE.*

4. Objectifs de thèse

Dans le cadre de cette thèse, nous nous sommes particulièrement intéressés aux méthodes d'évaluation des expériences de criblage virtuel. Dans un premier temps, nous avons proposé l'utilisation d'une nouvelle métrique d'évaluation qui peut être utilisée conjointement aux métriques actuelles afin d'évaluer plus finement les performances des logiciels : la Courbe de Prédictivité (Predictiveness Curve ou PC). Appliquée à l'évaluation des logiciels de docking, cette métrique permet de prendre directement en compte les variations de score d'affinité pour évaluer leur sensibilité dans la détection des composés actifs. Les courbes de prédictivité complètent efficacement les courbes de ROC pour évaluer visuellement les performances des méthodes et sont particulièrement adaptées à l'évaluation de la reconnaissance précoce des ligands.

Après avoir décrit l'application de cette métrique dans le cadre de criblages virtuels, nous avons développé un outil web permettant de réaliser une évaluation complète, rapide et intuitive des performances des logiciels : Screening Explorer. Cet outil regroupe l'ensemble des métriques utilisées pour l'évaluation des méthodes, qu'elles soient graphiques ou chiffrées, globales ou partielles, et inclus plusieurs approches consensus simples pour combiner les résultats de plusieurs méthodes de criblage virtuel.

D'autre part, nous avons également amorcé un projet de recherche et développement de potentiels nouveaux inhibiteurs du VIH. Cette démarche fait suite à des études génomiques de type GWAS[18] et eQTL[19] réalisées récemment dans notre laboratoire qui ont permis d'identifier des gènes dont l'expression favorise la non-progression[15,17] ou la progression rapide[16] vers le SIDA. Nous avons donc exploré ces résultats à la recherche de cibles « druggables » et de « hits » potentiels.

Deuxième partie : Résultats

1. Evaluation des méthodes de criblage virtuel

1.1. Les courbes de prédictivité pour l'analyse des criblages virtuels

1.1.1. Introduction & Publication

Dans le cas des criblages virtuels, différentes métriques sont utilisées rétrospectivement pour comparer l'efficacité des méthodes afin d'obtenir de meilleurs résultats dans les recherches prospectives de composés. Comme nous l'avons vu précédemment, les métriques disponibles reposent exclusivement sur le classement résultant des composés et les rangs des ligands au sein de ce classement (voir partie 3.2.2). Ainsi, ces métriques ne permettent pas de quantifier l'amplitude des variations des scores attribués par les méthodes pour expliquer leur pertinence dans la détection de composés actifs. Or, les scores produits par une méthode de criblage virtuel donnée sont directement conditionnés par les caractéristiques de la chimiothèque utilisée, sa taille, sa composition, sa diversité, les caractéristiques des molécules qu'elle regroupe, mais aussi par les caractéristiques du récepteur ciblé, son volume, sa flexibilité et de nombreux autres facteurs.[8,28,30–32]

Quantifier les écarts de score entre composés actifs et inactifs nous semble primordial, puisque ceux-ci représentent un indicateur direct de la puissance de discrimination des méthodes, ainsi que de leur capacité à conserver cette puissance de discrimination lorsqu'elles sont appliquées dans différentes conditions. L'évaluation de la reconnaissance précoce reste également problématique et de nouvelles métriques sont attendues à ces fins.[455,469] Disposer de nouvelles méthodes d'évaluation, efficaces pour mesurer la reconnaissance précoce et s'appuyant directement sur les scores des composés, permettrait de mieux évaluer les performances des méthodes afin de mieux guider, notamment, le développement des fonctions de score appliquées en criblage virtuel « structure-based ».[455,469]

Pour pallier les faiblesses des métriques actuelles, nous avons emprunté au domaine de l'épidémiologie clinique l'utilisation des courbes de prédictivité, que nous transposons et adaptons au cas des criblages virtuels. En épidémiologie, il est courant de construire des modèles de risque pour prédire l'occurrence d'une pathologie au sein d'une population de patients, grâce à l'usage de biomarqueurs et de modèles statistiques. Par exemple, Pepe et al. modélisent le risque d'occurrence d'un cancer de la prostate sur un panel de 5300 patients en utilisant 4 biomarqueurs et un modèle de régression logistique (***Équation 20***).[471] L'utilisation

des courbes de prédictivité permet ensuite d'interpréter intuitivement les résultats du modèle, puis d'estimer visuellement les fractions de patients à risque faible, intermédiaire ou élevé (***Figure 69***).[471] Par la suite, lorsque l'on réalise l'estimation du risque de cancer de la prostate d'un nouveau patient, la courbe de prédictivité permet d'informer les patients de leur situation par rapport au reste d'une population. Cet outil assiste donc la prise de décision, par exemple lorsqu'il s'agit d'entreprendre des examens invasifs ou des traitements médicaux.[471,472]

$$Risque = \exp(Y)\,/(1 + \exp(Y)$$

Avec

$$Y = -5.94 + 1.30\ \log(PSA) + 0.03\ \hat{a}ge + 0.99\ DRE - 0.37\ biopsie$$

Équation 20. *Modèle du risque de cancer de la prostate proposé par Pepe et al. à partir de 4 descripteurs et biomarqueurs : la concentration de l'antigène prostatique (PSA) retrouvée dans le sérum du patient, son âge et les résultats d'une biopsie et d'un examen digital rectal (DRE) précédents.*[471]

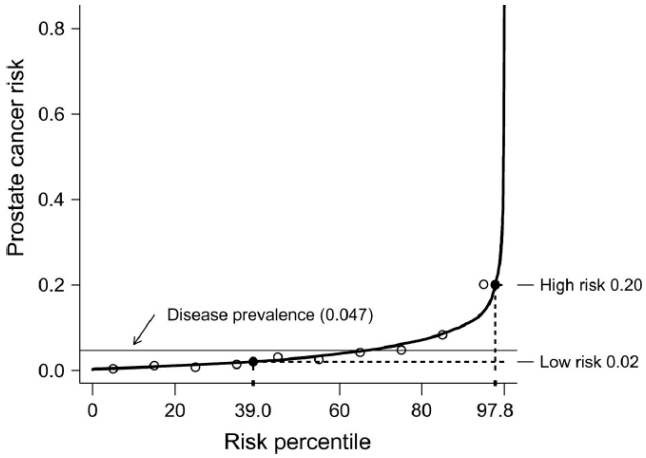


Figure 69. *Courbe de prédictivité du modèle de risque d'occurrence du cancer de la prostate proposé par Pepe et al. Cette visualisation permet d'estimer intuitivement les fractions de patients à risque faible, intermédiaire ou élevé.*[471]

De façon similaire, nous proposons d'utiliser des modèles de régression logistique couplés aux courbes de prédictivité pour expliquer et mesurer la capacité d'une méthode de criblage virtuel à discriminer les composés actifs des inactifs. Un modèle linéaire généralisé (Generalized Linear Model ou GLM) utilisant une fonction de distribution binomiale et un lien logit permet de calculer une probabilité d'activité pour chaque composé, en fonction de son score et par rapport à l'ensemble des autres scores attribués aux composés actifs ou inactifs d'une chimiothèque. Concernant le calcul des métriques partielles, destinées à quantifier la reconnaissance précoce des ligands, les métriques issues des courbes de prédictivité sont bien restreintes à une fraction du jeu de données classé mais restent ici dépendantes des scores de chacun des composés, qu'ils soient compris ou non dans la fraction sélectionnée. La

visualisation des courbes de prédictivité permet ensuite d'estimer intuitivement : (i) la fraction des composés actifs identifiés par une méthode de criblage virtuel et (ii) la capacité de cette méthode à produire des écarts de score pertinents entre composés actifs et inactifs, qui se traduit par des écarts entre les probabilités d'activité calculées.[30]

Empereur-mot *et al. J Cheminform (2015) 7:52*
DOI 10.1186/s13321-015-0100-8

Journal of Cheminformatics

RESEARCH ARTICLE Open Access

Predictiveness curves in virtual screening

Charly Empereur-mot[1], Hélène Guillemain[1], Aurélien Latouche[2], Jean-François Zagury[1], Vivian Viallon[3,4,5] and Matthieu Montes[1*]

Abstract

Background: In the present work, we aim to transfer to the field of virtual screening the predictiveness curve, a metric that has been advocated in clinical epidemiology. The literature describes the use of predictiveness curves to evaluate the performances of biological markers to formulate diagnoses, prognoses and assess disease risks, assess the fit of risk models, and estimate the clinical utility of a model when applied to a population. Similarly, we use logistic regression models to calculate activity probabilities related to the scores that the compounds obtained in virtual screening experiments. The predictiveness curve can provide an intuitive and graphical tool to compare the predictive power of virtual screening methods.

Results: Similarly to ROC curves, predictiveness curves are functions of the distribution of the scores and provide a common scale for the evaluation of virtual screening methods. Contrarily to ROC curves, the dispersion of the scores is well described by predictiveness curves. This property allows the quantification of the predictive performance of virtual screening methods on a fraction of a given molecular dataset and makes the predictiveness curve an efficient tool to address the early recognition problem. To this last end, we introduce the use of the total gain and partial total gain to quantify recognition and early recognition of active compounds attributed to the variations of the scores obtained with virtual screening methods. Additionally to its usefulness in the evaluation of virtual screening methods, predictiveness curves can be used to define optimal score thresholds for the selection of compounds to be tested experimentally in a drug discovery program. We illustrate the use of predictiveness curves as a complement to ROC on the results of a virtual screening of the Directory of Useful Decoys datasets using three different methods (Surflex-dock, ICM, Autodock Vina).

Conclusion: The predictiveness curves cover different aspects of the predictive power of the scores, allowing a detailed evaluation of the performance of virtual screening methods. We believe predictiveness curves efficiently complete the set of tools available for the analysis of virtual screening results.

Background

Structure-based and ligand-based virtual screening of compound collections has become extensively used in drug discovery programs to reduce the number of compounds going into high throughput screening procedures [1]. The aim of virtual screening methods is to enrich a subset of molecules in potentially active compounds while discarding the compounds supposed to be inactive according to a scoring function [2]. One of the issues with their use in prospective screening is to choose an optimal score selection threshold for experimental testing. It is usually estimated empirically through the analysis of retrospective virtual screening outputs on benchmarking datasets, which include known active compounds and putative inactive compounds (also known as decoys).

In this context, different metrics have emerged to evaluate the performance of virtual screening methods: enrichment factors (EFs), receiver operating characteristics (ROC) curves [2], the area under the ROC curve

*Correspondence: matthieu.montes@cnam.fr
[1] Laboratoire Génomique Bioinformatique et Applications, EA 4627, Conservatoire National des Arts et Métiers, 292 rue Saint Martin, 75003 Paris, France
Full list of author information is available at the end of the article

(ROC AUC) [2], the partial area under the ROC curve (pAUC) [3], the Boltzmann-enhanced discrimination of ROC (BEDROC) [4], the robust initial enhancement (RIE) [5]; ROC and EF being the most widely used. The ROC curves and their AUC provide a common scale to compare the performances of virtual screening methods. However, the ROC curves and their AUC suffer from two limitations. First, virtual screening methods are used to prioritize a subset of the screened compound collection for experimental testing, whereas ROC curves and ROC AUC summarize the ability of a method to rank a database over its entirety [4, 6]. Second, these two metrics are exclusively based on the ranks obtained by the compounds according to the score they obtained with the virtual screening method and do not take into account the difference in score between successively ranked compounds. Additionally, ROC curves are not suited to estimate the size of the molecular fraction selected at a given threshold. The true positive fraction (TPF) and false positive fraction (FPF) of the ROC plot can reflect a very different number of compounds on an identical scale, which can be misleading for analyzing the early recognition of active compounds.

EFs are more reliable towards the early recognition problem, since they are focused on the true positive fraction [2]. However, with EFs, the "ranking goodness" before the fractional threshold is not taken into account and their maximum value is strongly dependent on the ratio of active compounds over decoys in the benchmarking dataset (i.e. prevalence of activity) [2, 4, 7]. Another problem reported in previous studies is that metrics that seem to be statistically different such as ROC AUC, BEDROC, the area under the accumulation curve (AUAC) and the average rank of actives are in fact intimately related [4, 7, 8].

Different metrics have been proposed to overcome the limitations of the widely used EF and ROC curves, such as pAUC [3], BEDROC [4] and RIE [5], which better address early recognition. However, some limitations still persist: (1) the rank-based problems of ROC AUC are inherited by pAUC; (2) the maximum RIE value is dependent on the ratio of active compounds over decoys (similarly to EFs) [4]; and 3. BEDROC is dependent on a single parameter that embodies its overall sensitivity and that has to be selected according to the importance given to the early ranks. Unbiased comparisons between different evaluations are then rendered difficult by such a sensitive parameter [4, 6].

In the present work, we aimed to transfer to the field of virtual screening the Predictiveness Curve (PC) [9], a metric that has already been advocated in clinical epidemiology [10–14], where the values of biomarkers are used to formulate diagnoses, prognoses and assess disease risks. The use of PCs is described in the literature to evaluate the performance of given biological markers, to assess the fit of risk models and to estimate the clinical utility of a model when applied to a population. The dispersion of the scores attributed to the compounds by a given method is emphasized with the predictiveness curve, providing complementary information to classical metrics such as ROC and EF. Predictiveness curves can be used to (1) quantify and compare the predictive power of scoring functions above a given score quantile; and (2) define a score threshold for prospective virtual screening, in order to select an optimal number of compounds to be tested experimentally in a drug discovery program. In this study, we show how PCs can be used to graphically assess the predictive capacities of virtual screening methods, especially useful when considering the early recognition problem. Next, we applied the PC to the analysis of retrospective virtual screening results on the DUD database [15] using three different methods: Surflex-dock [16], ICM [17], and Autodock Vina [18]. We introduced the use of the total gain (TG) [19] to quantify the contribution of virtual screening scores to the explanation of compound activity. Standardized TG (noted as TG) ranges from 0 (no explanatory power) to 1 ("perfect" explanatory power) and can be visualized directly from the predictiveness curve [19]. Similarly, the partial total gain (pTG) [20] allows the explanatory power of virtual screening scores in the early part of the benchmarking dataset to be quantified as a partial summary measure of the PC. By monitoring the performances of three virtual screening methods using the predictiveness curve, TG and pTG on the DUD dataset, we have proposed a new approach to define optimal score thresholds adjusted to each target. Finally, we have discussed the interests of using predictiveness curves, total gain and partial total gain in addition to the ROC curves to better assess the performances of virtual screening methods and optimize the selection of compounds to be tested experimentally in prospective studies.

Methods

The directory of useful decoys (DUD) dataset

The DUD is a public benchmarking dataset designed for the evaluation of docking methods containing known active compounds for 40 targets, including 36 decoys for each active compound [15]. We selected for each target its corresponding DUD-own dataset that comprises only its associated active compounds and decoys. In our study, we used DUD release 2 dataset available at http://dud.docking.org.

Selection and preparation of the protein structures

We selected for this study the 39 targets issued from the DUD for which at least one experimental structure was

available. Target PDGFR-β was thus excluded since it was obtained through homology modeling. Hydrogen atoms were added using Chimera [21].

Computational methods

Surflex-dock

Surflex-dock is based on a modified Hammerhead fragmentation-reconstruction algorithm to dock compound flexibly into the binding site [16]. The query molecule is decomposed into rigid fragments that are superimposed to the Surflex protomol (i.e. molecular fragments covering the entire binding site). The docking poses were evaluated by an empirical scoring function. For each structure, the binding site was defined at 4Å around the co-crystallized ligand for the protomol generation step. In this study, Surflex-dock version 2.5 was used for all calculations.

ICM

ICM is based on Monte Carlo simulations in internal coordinates to optimize the position of molecules using a stochastic global optimization procedure combined with pseudo-Brownian positional/torsional steps and fast local gradient minimization [17]. The docking poses were evaluated using the ICM-VLS empirical scoring function [22]. The binding sites defined for docking were adjusted to be similar to the Surflex protomol. ICM version 3.6 was used for all calculations.

AutoDock Vina

Autodock Vina generates docking poses using an iterated local search global optimizer [23] which consists in a succession of steps of stochastic mutations and local optimizations [18]. At each step, the Broyden-Fletcher-Goldfarb-Shanno algorithm (BFGS) is used for local optimization [24]. Autodock Vina evaluated docking poses using its own empirical scoring function. The binding sites have been defined identically to the ones used for Surflex-dock and ICM calculations to obtain similar spatial search areas in all of the docking experiments. We used Autodock Vina version 1.1.2 for all calculations.

ROC curves analysis

The ROC curve applied to the retrospective analysis of a virtual screening experiment is a plot of the true positive fractions (TPF, y-axis) versus false positive fractions (FPF, x-axis) for all compounds in a ranked dataset [2, 6]. Each point of the ROC curve then represents a unique TPF/FPF pair corresponding to a particular fraction of the molecular dataset. A scoring function that would be able to perform perfect discrimination (i.e. no overlap between the two distributions of active and inactive compounds according to their calculated scores of binding affinity) has a ROC curve that passes through the upper left corner of the plot, where the TPF is 1 (perfect sensitivity) and the FPF is 0 (perfect specificity). The theoretical ROC curve resulting from an experiment in which the scoring function would have no discrimination is a 45° diagonal line from the lower left corner to the upper right corner. Qualitatively, the closer the curve is to the upper left corner, the higher the overall accuracy of the test. The area under the ROC curve (ROC AUC) summarizes the overall performance of a virtual screening experiment [2], whereas the partial area under the ROC curve (pAUC) allows to focus on a specific region of the curve and is usually calculated at a given early FPF value [3].

Predictiveness curves calculation

The approach we used in this study relies on the use of logistic regression to model how the scores issued by virtual screening methods explain the activity of the compounds in a virtual screening experiment. We used generalized linear models with a binomial distribution function and the canonical log link to calculate each compound probability of activity from the scores obtained by the compounds in a virtual screening experiment. Parameters were fit using the iteratively reweighted least squares algorithm. The predictiveness curve was then built as a cumulative distribution function (CDF) of activity probabilities. Let A denote a binary outcome termed compound activity where $A = 1$ for active and $A = 0$ for inactive. The probability of a compound to be active given its VS score $Y = y$ is $P_{\text{act}}(\text{y}) = P[A = 1 \mid Y = y]$. We proposed the use of the predictiveness plots, $R(\nu)$ versus ν, to describe the predictive capacity of a VS method, where $R(\nu)$ is the activity probability associated with the νth quantile of the VS scores: $R(\nu) = P[A = 1 \mid Y = F^{-1}(\nu)]$, and F is the CDF of VS scores. Hence, predictiveness plots provide a common scale for making comparisons between VS methods that may not be comparable on their original scales [12]. Suppose p_L and p_H are two thresholds that define "low probability of activity" and "high probability of activity". Then the proportions of the compounds with low, high, and equivocal probabilities of activity are $R^{-1}(p_L)$, $1 - R^{-1}(p_H)$ and $R^{-1}(p_H) - R^{-1}(p_L)$, respectively, using the inverse function of $R(\nu)$. Virtual screening scores that are uninformative about compound activity assign equal activity probabilities to all compounds, $P_{\text{act}}(\text{Y}) = \text{P}[\text{A} = 1 \mid Y] = \text{P}[\text{A} = 1] = p$, where p is the prevalence of activity in the molecular dataset. On the other hand, perfect VS scores assign $P_{\text{act}}(\text{Y}) = 1$ for the proportion p of compounds with $\text{A} = 1$ and $P_{\text{act}}(\text{Y}) = 0$ for the proportion $1 - p$ with $\text{A} = 0$. Correspondingly, its PC is the step function $R(\nu) = I[(1 - p) < \nu]$, where I is the indicator function. Most scoring functions are imperfect, yielding activity probabilities between these extremes. Good

predictions issued from virtual screening methods yield steeper predictiveness curves corresponding to wider variations of activity probabilities.

Predictiveness plots analysis

The ability of the models to highlight score gaps between compounds and relate those differences to activity probabilities allowed us to quantify the predictive power of virtual screening methods in terms of both scoring and ranking. Displaying the PC then allows for an intuitive analysis of the performances of virtual screening methods. The visualization of the total gain, partial total gain and the size of the molecular subset enables a straightforward interpretation of the results (Fig. 1a). For a completely uninformative model the PC would correspond to a horizontal line at the level of activity prevalence (Fig. 1). Inversely, steep predictiveness curves enable the observation of an inflexion point from which the curve rises. Hence, additionally to its benchmarking interests, PC provides a guidance to choose an optimal score threshold from VS results, allowing one to assess decision criteria from multiple points of view. Visualizing the curve allows to determine if activity probability variations are important enough to induce the selection of a threshold for prospective virtual screenings. Usual metrics can also be interpreted from the predictiveness curve: the true positive fraction (TPF), false positive fraction (FPF), positive predictive value (PPV) and negative predictive value (NPV) (Fig. 1b).

Performance metrics

Statistical analysis was conduced using the R software [25]. The package ROCR [26] was used to plot ROC curves and perform ROC and partial ROC AUC calculations.

Enrichment factors were computed as follows:

$$EF_{x\%} = \frac{Hits_{x\%}/N_{x\%}}{Hits_{t\%}/N_{x\%}}$$

where $Hits_{x\%}$ is the number of active compounds in the top $x\%$ of the ranked dataset, $Hits_t$ is the total number of active compounds in the dataset, $N_{x\%}$ is the number of compounds in the $x\%$ of the dataset and N_t is the total number of compounds in the dataset.

The contribution of virtual screening scores to the explanation of compounds activity can be quantified over a dataset using the standardized total gain (TG) [19], introduced by Bura et al. as a summary measure of the predictiveness curve:

$$\overline{TG}(v) = \frac{\int_0^1 |R(v) - p|\, dv}{2p\,(1-p)}$$

where p is the prevalence of activity in the molecular dataset and $R(v)$ is the value of the activity probability at the vth quantile. The total gain is normalized by its maximum value, so that TG values are in the range [0,1] (null to perfect explanatory power). TG summarizes the proportion of variance in a binomial outcome explained by

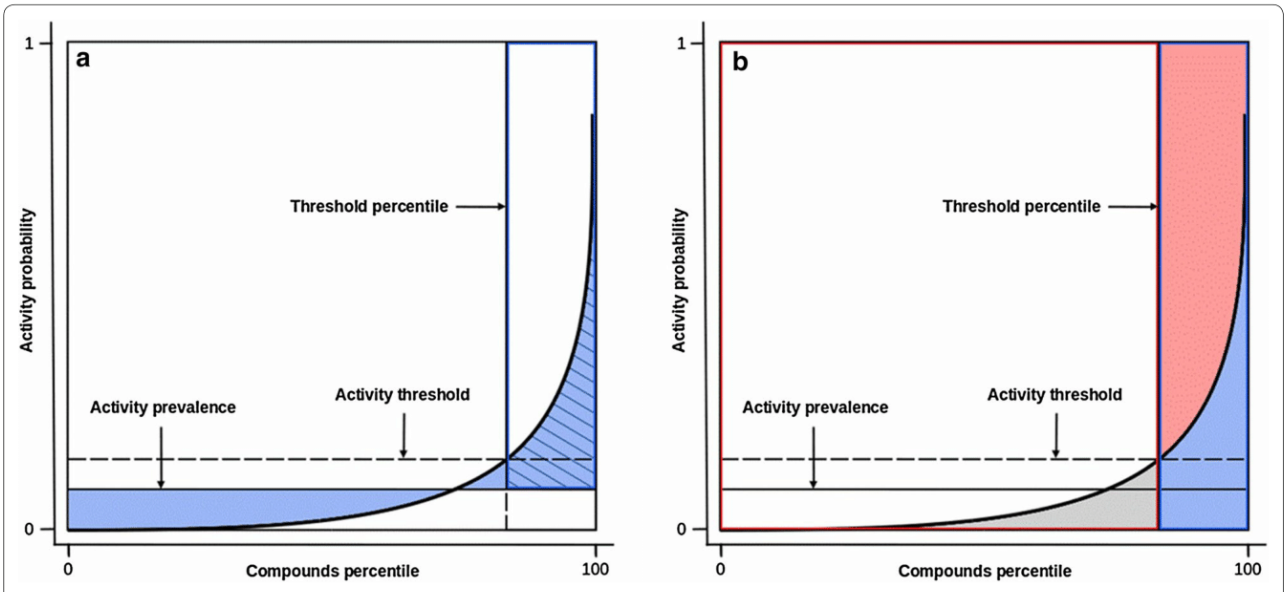


Fig. 1 Schematic diagram presenting how performance metrics relate to the predictiveness curve. Displaying the PC allows for an intuitive selection of thresholds. Performance metrics related to a chosen threshold are easily interpreted from the curve. **a** Partial total gain (pTG): hatched area/*blue* frame; total gain (TG): *blue* area. **b** True positive fraction (TPF): *blue* area/area under activity prevalence; false positive fraction (FPF): *red* area/(1-area under activity prevalence); positive predictive value (PPV): *blue* area/*blue* frame; negative predictive value (NPV): *white area/red* frame

the model. In our application, TG quantifies the success of a VS method to rank and score compounds depending on activity, over the complete molecular dataset.

The predictive performance of VS scores can be quantified above the νth quantile of the molecular dataset using the partial total gain (pTG) [20], recently introduced by Sachs et al. as a partial summary measure of the PC, defined as:

$$pTG(\nu) = \frac{\int_{\nu}^{1} |R(\nu) - p| \, d\nu}{(1-\nu)(1-p)}$$

where p is the prevalence of activity in the molecular dataset and $R(\nu)$ is the value of the activity probability at the νth quantile of the dataset. The denominator term is a standardization factor leading to pTG values in the range of 0 to 1 and makes pTG prevalence independent. pTG summarizes the proportion of variance in a binomial outcome explained by the model above the νth quantile. In our application, pTG quantifies the contribution of virtual screening scores to the explanation of compounds activity above the νth quantile of the molecular dataset.

Results

Assessment of the predictive power of a scoring function

We first illustrated the use of the predictiveness curve as a complement to the ROC curve with the results obtained from Surflex-dock, ICM, and Autodock Vina on target retinoic X receptor (RXR) of the DUD dataset (Fig. 2). For these methods, the ROC AUCs indicated that the discrimination of active compounds over inactive compounds within the complete dataset was successful (Surflex-dock: 0.907, ICM: 0.812, Autodock Vina: 0.944). The ROC curve profiles suggested that acceptable early recognition has been achieved by the three methods (Surflex-dock pAUC2 %: 0.167, ICM pAUC2 %: 0.342, Autodock Vina pAUC2 %: 0.330), which was confirmed in terms of enrichment (Surflex-dock EF2 %: 16.84, ICM EF2 %: 24.06, Autodock Vina EF2 %: 26.47). Under these conditions, following the first described use of the ROC curves for the analysis of virtual screening results [2], score selection thresholds could be extracted from the curve points prior to FPF = 0.2 by maximizing the sensitivity or the specificity of the method.

In the present case, the analysis of the predictiveness curves brought complementary insights. Total gain values indicated that the detection of the activity of the compounds is related to more important score variations with Autodock Vina, compared to ICM and Surflex-dock (Surflex-dock TG = 0.675, ICM TG = 0.124, Autodock Vina TG = 0.740). The contributions of each scoring function to the early detection of active compounds can be quantified using the partial total gain (Surflex-dock pTG2 %: 0.308, ICM pTG2 %: 0.026, Autodock Vina pTG2 %: 0.653), which enables a straightforward comparison of the performances of the methods in a limited range of the dataset. In the case of ICM, even if the ROC curve profile supported that global and early enrichments are achieved, the associated PC corresponded to a *quasi* null-model, associated to a low TG value. Even if ICM was able to rank the active compounds satisfactorily, the analysis of the PC informed us that the score variations

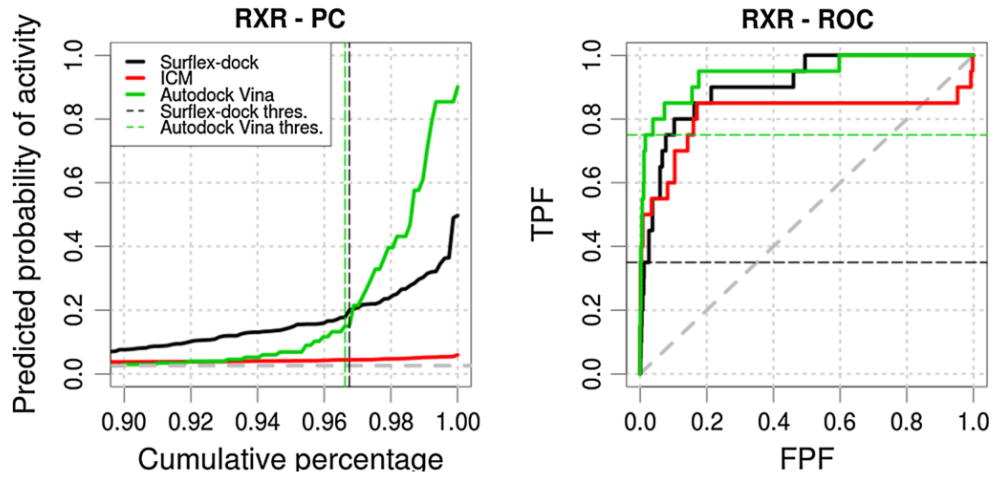


Fig. 2 Predictiveness and ROC curves for the virtual screening of target retinoic X receptor (RXR) from the DUD dataset using Surflex-dock, ICM and Autodock Vina (black, red and green curves, respectively). Dashed gray lines indicate the prevalence of activity and random picking of compounds. Dashed black and green lines show the two thresholds we set manually from the reading of the PC. Those two thresholds have been reported on the ROC curves.

between the active compounds and the decoys were not representative of the activity of the compounds. Then, deriving score thresholds from the analysis of retrospective virtual screening experiments with ICM would not be relevant for the prospective detection of active compounds on RXR.

The PCs could graphically emphasize the performance of each method on early enrichment, highlighting that the most predictive method towards the activity of the compounds on RXR was Autodock Vina, over Surflex-dock and ICM.

Selection of optimal score thresholds

A visual analysis of the PCs for RXR clearly displayed that Autodock Vina outperformed Surflex-dock and ICM in terms of early enrichment and that its scoring function would be more predictive of activity within its high scores. In particular, for Autodock Vina on this target, an inflexion point was observable where the PC rose steeply (3.38 % of the ranked dataset), which allowed the retrieval of a score selection threshold from which the scores are highly associated with the activity of the compounds in the corresponding subset (Autodock Vina pTG3.38 %: 0.488, Autodock Vina EF3.38 %: 21.39) (Fig. 2, vertical dashed green line). The pTG of 0.488 in the selected subset signified that each compound in this subset has an average probability gain of 0.488 of being active over the random picking of compounds. For Surflex-dock the PC showed a different profile, gradually increasing to reach activity probabilities over 0.5. In this particular case, the threshold selection is graphically estimated depending on the size of the selected subset. We have estimated the optimal selection threshold for Surflex-dock at 3.25 % of the ranked dataset (Surflex-dock pTG3.25 %: 0.265, Surflex-dock EF3.25 %: 10.37) (Fig. 2, vertical dashed black line), which was close to the optimal threshold retrieved with Autodock Vina. We then projected these two thresholds on the ROC curves (Fig. 2, horizontal colored dashed lines). Interestingly, the visualization of these two thresholds on the PC and ROC curves emphasized the bias induced by the ROC towards the estimation of the size of the selected subset. For the two close selected thresholds the corresponding points on the ROC curves largely differ emphasizing that the ROC curves are not adapted to visualize the size of the selected datasets (Surflex-dock TPF3.25 %: 0.350, Surflex-dock FPF3.25 %: 0.025, Autodock Vina TPF3.38 %: 0.750, Autodock Vina FPF3.38 %: 0.016).

Emphasize on the different early recognition profiles

We performed virtual screening experiments on 39 targets from the DUD dataset using Surflex-dock, ICM and Autodock Vina. For 9 out of the 39 targets (ACHE, AMPC, FGFR1, GR, HIVRT, HSP90, PR, TK and VEGFR2), none of the three virtual screening methods yielded differences in score that were predictive of the activity of the compounds, resulting in PCs *quasi* null-model profile and very low TG values.

Surflex-dock, ICM and Autodock Vina screenings of the remaining datasets resulted in PCs with a profile that allowed an estimation of an optimal score selection threshold at the steepest inflexion point of the PC for respectively 22, 19 and 17 datasets. ROC AUC and TG are presented in Table 1. PCs and ROC plots are presented in Figs. 3, 4, 5 and 6 and include the display of the score selection thresholds (dashed colored lines). Score selection thresholds, pTGs, pAUCs and EFs for each virtual screening method in the resulting subsets are presented in Tables 2, 3 and 4.

The score selection thresholds for each method varied with the datasets (Surflex-dock: 6.73–12.83, ICM: −52.17 to −22.69, Autodock Vina: −12.10 to −9.00). Mean EF and median EF in the resulting subsets for each virtual screening method were superior to 13.00. The analysis thus allowed to identify target specific optimal score selection thresholds that yielded satisfying EFs, up to two digits, for 57 out of the 117 possible method/dataset associations (Figs. 3, 4, 5, 6). For 1 out of the 117 possible method/dataset associations, the defined threshold resulted in no enrichment (Surflex-dock on SAHH). For the remaining 59 method/dataset associations, the predictiveness curves suggested a defect of association between the scores obtained by the compounds and their activity.

The score selection thresholds for each method varied with the datasets (Surflex-dock: 6.73–12.83, ICM: −52.17 to −22.69, Autodock Vina: −12.10 to −9.00). Mean EF and median EF in the resulting subsets for each virtual screening method were superior to 13.00. The analysis thus allowed to identify target specific optimal score selection thresholds that yielded satisfying EFs, up to two digits, for 57 out of the 117 possible method/dataset associations (Fig. 3, 4, 5, 6). For 1 out of the 117 possible method/dataset associations, the defined threshold resulted in no enrichment (Surflex-dock on SAHH). For the remaining 59 method/dataset associations, the predictiveness curves suggested a defect of association between the scores obtained by the compounds and their activity.

We finally highlighted systems that illustrated the interest of using the PCs as a complement to the ROC curves: (1) Surflex-dock and ICM applied to the HMGR dataset represented one of the best-achieved early recognition cases, both PCs displaying a steep inflexion point. In this case, the analysis of the PC validated the profile of the ROC

Table 1 Description of the benchmarking dataset from the DUD, including global metrics of the virtual screens performed using Surflex-dock, ICM and Autodock Vina

Target	Nb of actives	Nb of compounds	Prevalence	Surflex-dock		ICM		Autodock vina	
				TG	ROC AUC	TG	ROC AUC	TG	ROC AUC
ACE	49	1846	0.0265	0.035	0.464	0.299	0.655	0.189	0.408
ACHE	107	3999	0.0268	0.012	0.512	0.115	0.614	0.107	0.662
ADA	39	966	0.0404	0.310	0.699	0.250	0.320	0.110	0.438
ALR2	26	1021	0.0255	0.250	0.536	0.228	0.647	0.295	0.677
AMPC	21	807	0.0260	0.227	0.687	0.134	0.534	0.214	0.325
AR	79	2933	0.0269	0.067	0.684	0.151	0.691	0.396	0.745
CDK2	72	2146	0.0336	0.186	0.608	0.364	0.734	0.212	0.620
COMT	11	479	0.0230	0.392	0.733	0.256	0.698	0.001	0.440
COX-1	25	936	0.0267	0.078	0.587	0.372	0.727	0.348	0.726
COX-2	426	13715	0.0311	0.396	0.784	0.056	0.555	0.461	0.736
DHFR	410	8777	0.0467	0.387	0.715	0.198	0.618	0.337	0.737
EGFR	475	16471	0.0288	0.018	0.461	0.352	0.697	0.159	0.605
ER ago	67	2637	0.0254	0.301	0.708	0.462	0.772	0.533	0.833
ER antago	39	1487	0.0262	0.412	0.758	0.263	0.631	0.176	0.562
FGFR1	120	4670	0.0257	0.134	0.569	0.097	0.403	0.083	0.441
FXA	146	5891	0.0248	0.521	0.860	0.326	0.702	0.132	0.616
GART	40	919	0.0435	0.555	0.881	0.492	0.783	0.287	0.710
GPB	52	2192	0.0237	0.218	0.675	0.434	0.835	0.361	0.757
GR	78	3025	0.0258	0.010	0.564	0.050	0.450	0.126	0.560
HIVPR	62	2100	0.0295	0.517	0.808	0.175	0.649	0.317	0.743
HIVRT	43	1562	0.0275	0.185	0.621	0.191	0.622	0.234	0.633
HMGR	35	1515	0.0231	0.642	0.878	0.438	0.723	0.080	0.545
HSP90	37	1016	0.0364	0.098	0.598	0.224	0.340	0.136	0.612
INHA	86	3352	0.0257	0.112	0.551	0.032	0.524	0.203	0.544
MR	15	651	0.0230	0.492	0.796	0.401	0.732	0.614	0.844
NA	49	1923	0.0255	0.633	0.870	0.764	0.923	0.198	0.350
P38	454	9595	0.0473	0.231	0.651	0.127	0.367	0.087	0.572
PARP	35	1386	0.0253	0.435	0.738	0.440	0.755	0.324	0.728
PDE5	88	2066	0.0426	0.062	0.524	0.465	0.775	0.121	0.582
PNP	50	1086	0.0460	0.404	0.755	0.072	0.635	0.034	0.536
PPAR	85	3212	0.0265	0.676	0.901	0.415	0.748	0.499	0.801
PR	27	1068	0.0253	0.109	0.527	0.130	0.686	0.080	0.525
RXR	20	770	0.0260	0.675	0.907	0.124	0.812	0.740	0.944
SAHH	33	1379	0.0239	0.391	0.811	0.330	0.751	0.338	0.717
SRC	159	6478	0.0245	0.162	0.569	0.420	0.748	0.288	0.694
THR	72	2528	0.0285	0.447	0.787	0.420	0.798	0.331	0.706
TK	22	913	0.0241	0.139	0.668	0.015	0.453	0.110	0.583
TRP	49	1713	0.0286	0.767	0.953	0.155	0.637	0.140	0.619
VEGFR2	88	2994	0.0294	0.092	0.558	0.201	0.625	0.034	0.504
Minimum	11	479	0.0230	0.010	0.461	0.015	0.320	0.001	0.325
Maximum	475	16471	0.0473	0.767	0.953	0.764	0.923	0.740	0.944
Mean	97	3134	0.0294	0.302	0.691	0.268	0.650	0.242	0.625
Median	50	1923	0.0265	0.250	0.687	0.250	0.686	0.203	0.619

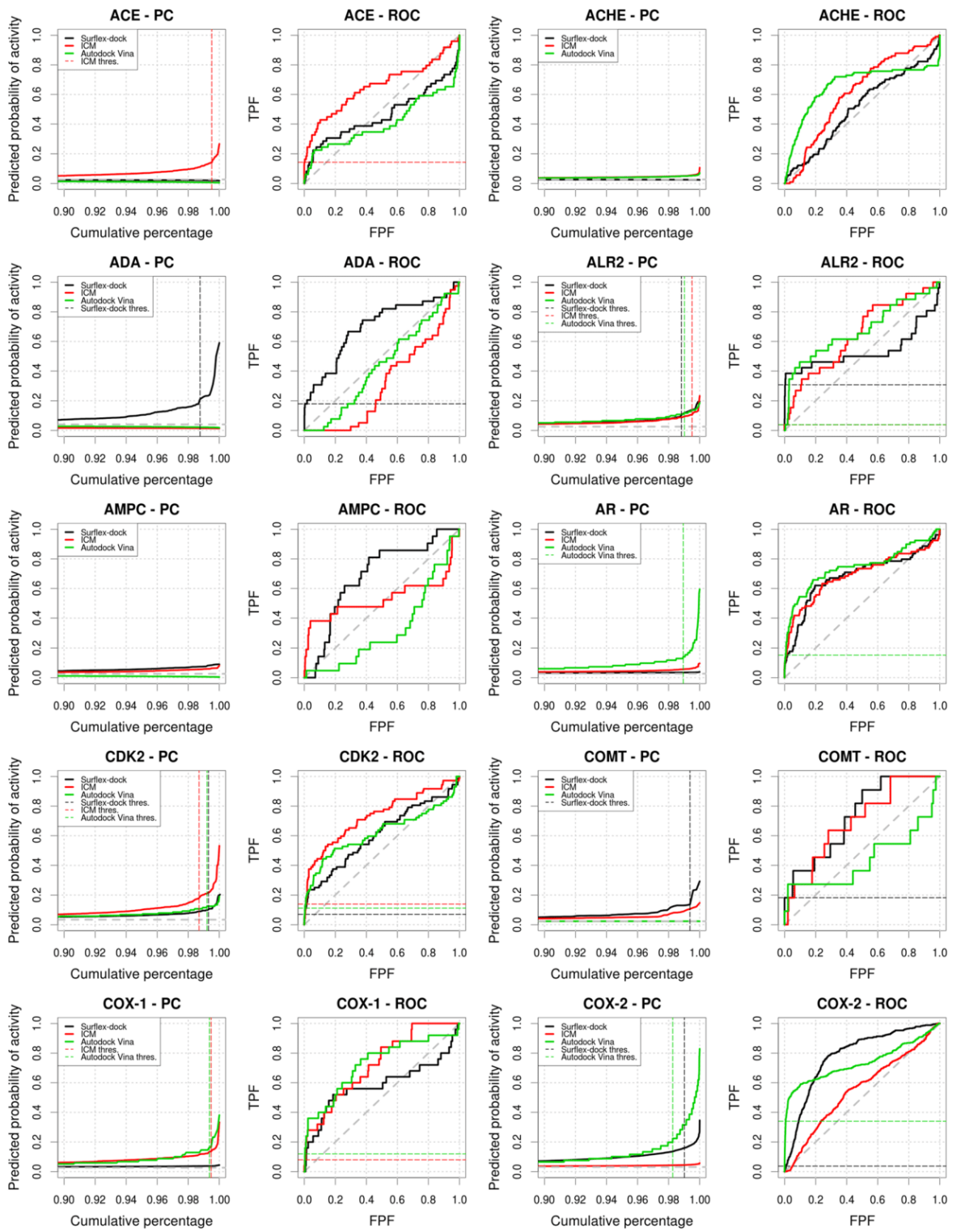


Fig. 3 Predictiveness and ROC curves for the virtual screenings of ACE, ACHE, ADA, ALR2, AMPC, AR, CDK2, COMT, COX1 and COX2 selected from the DUD datasets using Surflex-dock, ICM and Autodock Vina (black, red and green curves, respectively). Dashed gray lines indicate the prevalence of activity and random picking of compounds. Vertical dashed lines represent the thresholds we manually selected from the analysis of the curves. Metrics associated to the selected thresholds are available in Tables 2, 3, 4. Partial metrics at 2 % and 5 % of the ranked dataset are available in Additional file 1: Table S1; Additional file 2: Table S2 and Additional file 3: Table S3.

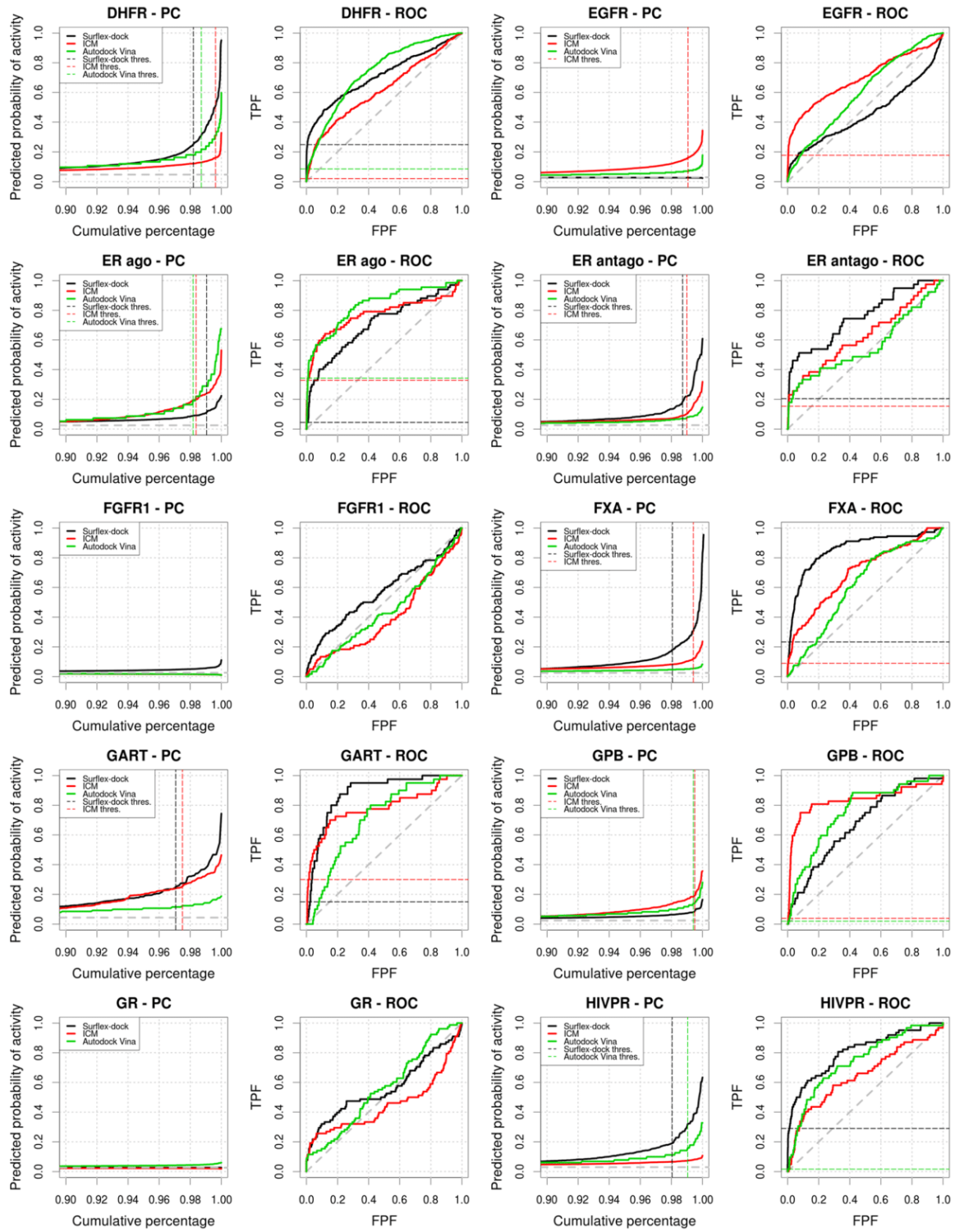


Fig. 4 Predictiveness and ROC curves for the virtual screenings of DHFR, EGFR, ER ago, ER antago, FGFR1, FXA, GART, GPB, GR and HIVPR selected from the DUD datasets using Surflex-dock, ICM and Autodock Vina (black, red and green curves, respectively). Dashed gray lines indicate the prevalence of activity and random picking of compounds. Vertical dashed lines represent the thresholds we manually selected from the analysis of the curves. Metrics associated to the selected thresholds are available in Tables 2, 3, 4. Partial metrics at 2 % and 5 % of the ranked dataset are available in Additional file 1: Table S1; Additional file 2: Table S2 and Additional file 3: Table S3.

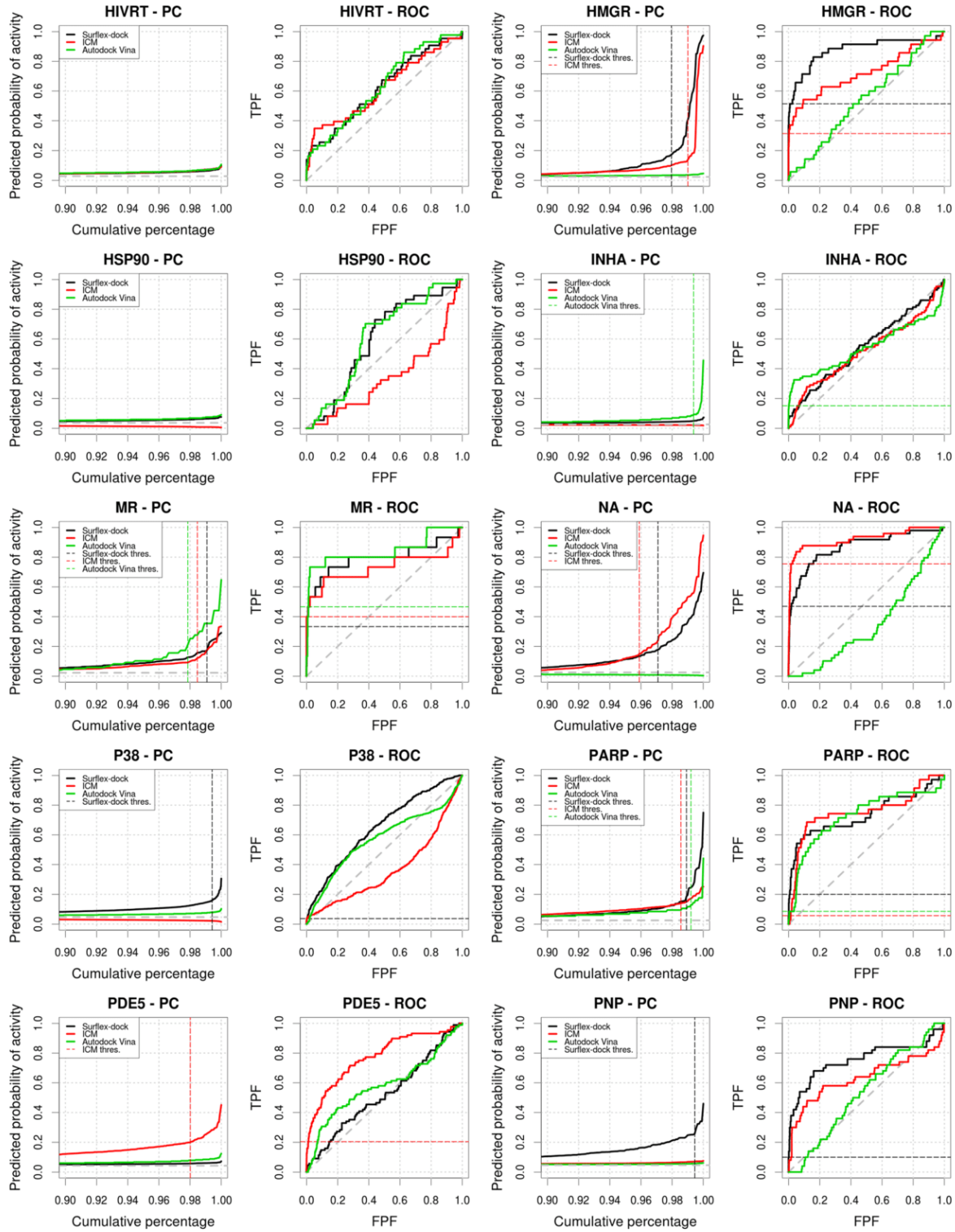


Fig. 5 Predictiveness and ROC curves for the virtual screenings of HIVRT, HMGR, HSP90, INHA, MR, NA, P38, PARP, PDE5 and PNP selected from the DUD datasets using Surflex-dock, ICM and Autodock Vina (black, red and green curves, respectively). Dashed gray lines indicate the prevalence of activity and random picking of compounds. Vertical dashed lines represent the thresholds we manually selected from the analysis of the curves. Metrics associated to the selected thresholds are available in Tables 2, 3, 4. Partial metrics at 2 % and 5 % of the ranked dataset are available in Additional file 1: Table S1; Additional file 2: Table S2 and Additional file 3: Table S3.

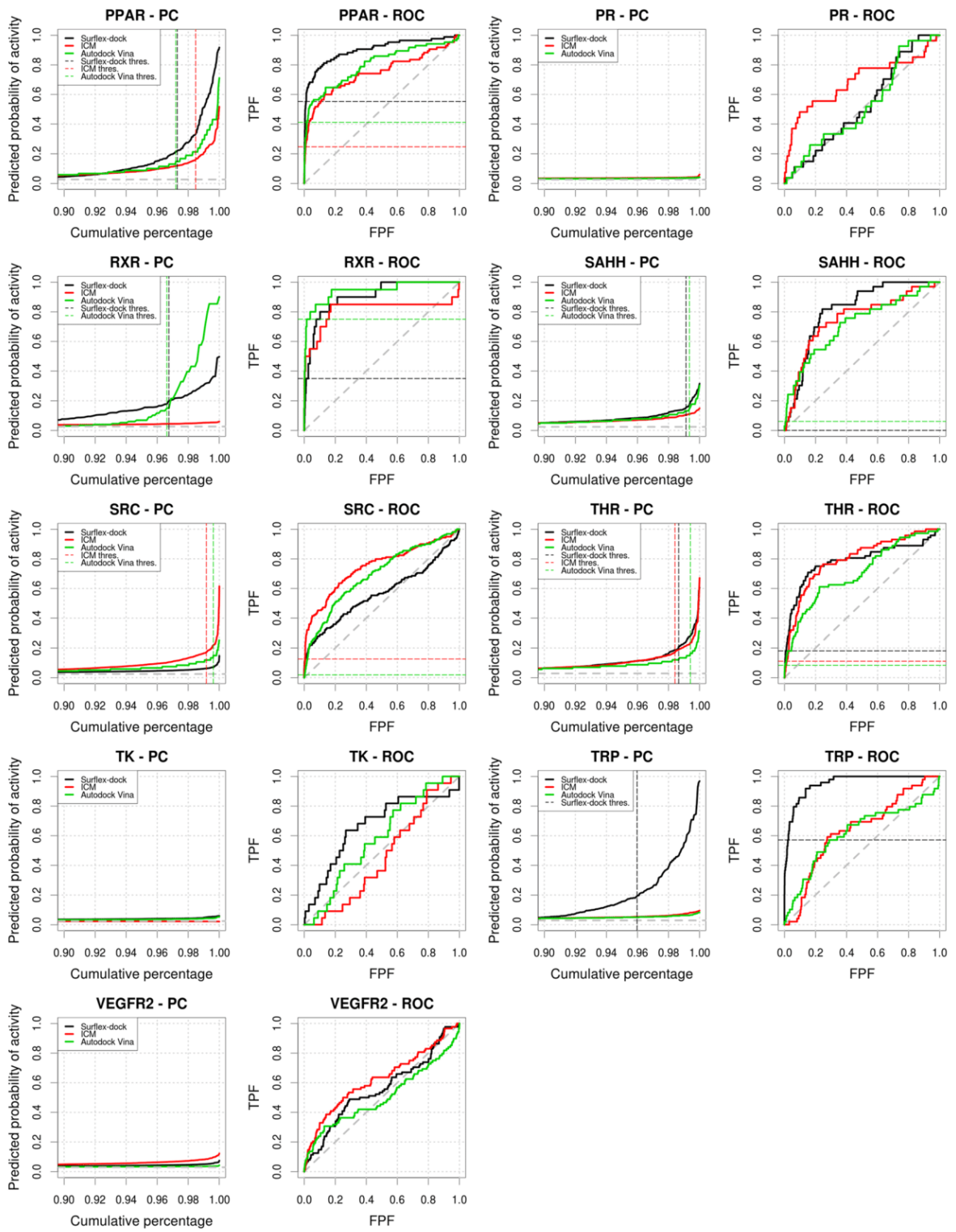


Fig. 6 Predictiveness and ROC curves for the virtual screenings of PPAR, PR, RXR, SAHH, SRC, THR and VEGFR2 selected from the DUD datasets using Surflex-dock, ICM and Autodock Vina (black, red and green curves, respectively). Dashed gray lines indicate the prevalence of activity and random picking of compounds. Vertical dashed lines represent the thresholds we manually selected from the analysis of the curves. Metrics associated to the selected thresholds are available in Tables 2, 3, 4. Partial metrics at 2 % and 5 % of the ranked dataset are available in Additional file 1: Table S1; Additional file 2: Table S2 and Additional file 3: Table S3.

Table 2 Summary of the partial metrics associated to the thresholds we selected manually from the virtual screens performed using Surflex-dock

Target	Surflex-dock—manual thresholds							
	Rank threshold	Activity threshold	pTG	pAUC	EF	Score	Actives	Cpds
ACE	–	–	–	–	–	–	–	–
ACHE	–	–	–	–	–	–	–	–
ADA	1.24	0.211	0.293	0.128	13.34	9.78	7	13
ALR2	1.18	0.103	0.119	0.231	24.17	6.73	8	13
AMPC	–	–	–	–	–	–	–	–
AR	–	–	–	–	–	–	–	–
CDK2	0.70	0.103	0.115	0.052	8.77	8.82	5	17
COMT	0.63	0.136	0.206	0.182	21.77	7.29	2	4
COX-1	–	–	–	–	–	–	–	–
COX-2	0.98	0.160	0.165	0.017	3.79	8.24	16	136
DHFR	1.80	0.251	0.389	0.207	13.73	9.36	102	159
EGFR	–	–	–	–	–	–	–	–
ER ago	0.95	0.114	0.131	0.032	4.54	8.07	3	26
ER antago	1.28	0.185	0.315	0.071	15.25	10.07	8	20
FGFR1	–	–	–	–	–	–	–	–
FXA	1.94	0.181	0.314	0.126	11.53	10.04	34	119
GART	2.94	0.249	0.345	0.062	4.92	12.47	6	28
GPB	–	–	–	–	–	–	–	–
GR	–	–	–	–	–	–	–	–
HIVPR	1.95	0.190	0.313	0.227	14.52	9.80	18	42
HIVRT	–	–	–	–	–	–	–	–
HMGR	2.05	0.173	0.471	0.449	24.35	8.91	18	32
HSP90	–	–	–	–	–	–	–	–
INHA	–	–	–	–	–	–	–	–
MR	0.92	0.172	0.227	0.200	31.00	7.32	5	7
NA	2.91	0.180	0.320	0.339	15.84	11.37	23	57
P38	0.58	0.161	0.155	0.029	6.30	8.75	17	57
PARP	1.08	0.187	0.332	0.143	17.32	7.12	7	16
PDE5	–	–	–	–	–	–	–	–
PNP	0.55	0.258	0.303	0.100	15.51	7.65	5	7
PPAR	2.71	0.221	0.441	0.395	20.18	12.83	47	88
PR	–	–	–	–	–	–	–	–
RXR	3.25	0.200	0.265	0.263	10.37	10.84	7	26
SAHH	0.87	0.154	0.200	0.000	0.00	10.08	0	13
SRC	–	–	–	–	–	–	–	–
THR	1.34	0.209	0.283	0.097	13.04	9.54	13	35
TK	–	–	–	–	–	–	–	–
TRP	4.03	0.193	0.414	0.442	13.98	8.80	28	70
VEGFR2	–	–	–	–	–	–	–	–
Minimum	0.55	0.103	0.115	0.000	0.00	6.73	0	4
Maximum	4.03	0.258	0.471	0.449	31.00	12.83	102	159
Mean	1.63	0.181	0.278	0.172	13.83	9.27	17	45
Median	1.26	0.183	0.298	0.136	13.86	9.13	8	27

Table 3 Summary of the partial metrics associated to the thresholds we selected manually from the virtual screens performed using ICM

Target	ICM—manual thresholds							
	Rank threshold	Activity threshold	pTG	pAUC	EF	Score	Actives	Cpds
ACE	0.49	0.14	0.166	0.136	26.37	−31.64	7	10
ACHE	–	–	–	–	–	–	–	–
ADA	–	–	–	–	–	–	–	–
ALR2	0.49	0.11	0.118	0.038	6.54	−31.46	1	6
AMPC	–	–	–	–	–	–	–	–
AR	–	–	–	–	–	–	–	–
CDK2	1.30	0.18	0.231	0.059	10.28	−29.72	10	29
COMT	–	–	–	–	–	–	–	–
COX-1	0.53	0.14	0.165	0.050	12.48	–29.72	2	6
COX-2	–	–	–	–	–	–	–	–
DHFR	0.38	0.16	0.148	0.018	5.04	−30.36	8	34
EGFR	0.93	0.16	0.179	0.101	18.79	−33.60	84	155
ER ago	1.63	0.20	0.252	0.197	19.68	−31.57	22	44
ER antago	1.01	0.10	0.157	0.062	14.30	−34.06	6	16
FGFR1	–	–	–	–	–	–	–	–
FXA	0.59	0.12	0.140	0.067	14.57	−33.80	13	36
GART	2.50	0.25	0.293	0.142	11.49	−52.17	12	24
GPB	0.50	0.19	0.240	0.010	7.03	−35.50	2	12
GR	–	–	–	–	–	–	–	–
HIVPR	–	–	–	–	–	–	–	–
HIVRT	–	–	–	–	–	–	–	–
HMGR	0.99	0.14	0.454	0.257	29.76	−26.21	11	16
HSP90	–	–	–	–	–	–	–	–
INHA	–	–	–	–	–	–	–	–
MR	1.54	0.12	0.185	0.400	23.67	−29.12	6	11
NA	4.11	0.15	0.385	0.580	18.15	−22.69	37	80
P38	–	–	–	–	–	–	–	–
PARP	1.44	0.14	0.162	0.015	3.77	−36.20	2	21
PDE5	1.98	0.20	0.232	0.151	10.06	−28.00	18	42
PNP	–	–	–	–	–	–	–	–
PPAR	1.53	0.16	0.237	0.146	15.87	−42.49	21	50
PR	–	–	–	–	–	–	–	–
RXR	–	–	–	–	–	–	–	–
SAHH	–	–	–	–	–	–	–	–
SRC	0.83	0.17	0.220	0.082	14.82	−35.14	20	55
THR	1.58	0.17	0.236	0.050	6.85	−26.85	8	41
TK	–	–	–	–	–	–	–	–
TRP	–	–	–	–	–	–	–	–
VEGFR2	–	–	–	–	–	–	–	–
Minimum	0.38	0.10	0.118	0.010	3.77	−52.17	1	6
Maximum	4.11	0.25	0.454	0.580	29.76	−22.69	84	155
Mean	1.28	0.16	0.221	0.135	14.19	−32.65	15	36
Median	1.01	0.16	0.220	0.082	14.30	−31.57	10	29

Table 4 Summary of the partial metrics associated to the thresholds we selected manually from the virtual screens performed using Autodock Vina

Target	Autodock Vina—manual thresholds							
	Rank threshold	Activity threshold	pTG	pAUC	EF	Score	Actives	Cpds
ACE	–	–	–	–	–	–	–	–
ACHE	–	–	–	–	–	–	–	–
ADA	–	–	–	–	–	–	–	–
ALR2	0.98	0.117	0.114	0.019	3.57	−10.10	1	11
AMPC	–	–	–	–	–	–	–	–
AR	1.06	0.141	0.214	0.113	13.92	−9.90	12	32
CDK2	0.79	0.124	0.103	0.054	13.25	−10.30	8	18
COMT	–	–	–	–	–	–	–	–
COX-1	0.64	0.165	0.225	0.000	16.05	−9.00	3	7
COX-2	1.74	0.224	0.345	0.201	19.53	−10.20	145	239
DHFR	1.29	0.218	0.250	0.051	6.57	−9.90	35	114
EGFR	–	–	–	–	–	–	–	–
ER ago	1.82	0.191	0.328	0.175	18.47	−9.50	23	49
ER antago	–	–	–	–	–	–	–	–
FGFR1	–	–	–	–	–	–	–	–
FXA	–	–	–	–	–	–	–	–
GART	–	–	–	–	–	–	–	–
GPB	0.59	0.139	0.164	0.007	3.01	−9.20	1	14
GR	–	–	–	–	–	–	–	–
HIVPR	0.95	0.159	0.192	0.012	1.61	−10.70	1	21
HIVRT	–	–	–	–	–	–	–	–
HMGR	–	–	–	–	–	–	–	–
HSP90	–	–	–	–	–	–	–	–
INHA	0.63	0.093	0.136	0.138	23.03	−11.10	13	22
MR	2.15	0.210	0.336	0.250	20.25	−10.30	7	15
NA	–	–	–	–	–	–	–	–
P38	–	–	–	–	–	–	–	–
PARP	0.79	0.130	0.160	0.048	9.90	−10.30	3	12
PDE5	–	–	–	–	–	–	–	–
PNP	–	–	–	–	–	–	–	–
PPAR	2.80	0.148	0.258	0.267	14.53	−12.10	35	91
PR	–	–	–	–	–	–	–	–
RXR	3.38	0.150	0.488	0.500	21.39	−10.60	15	27
SAHH	0.65	0.146	0.181	0.030	8.36	−9.00	2	10
SRC	0.39	0.149	0.154	0.005	4.70	−9.60	3	26
THR	0.59	0.164	0.187	0.056	13.17	−10.40	6	16
TK	–	–	–	–	–	–	–	–
TRP	–	–	–	–	–	–	–	–
VEGFR2	–	–	–	–	–	–	–	–
Minimum	0.39	0.093	0.103	0.000	1.61	−12.10	1	7
Maximum	3.38	0.224	0.488	0.500	23.03	−9.00	145	239
Mean	1.25	0.157	0.226	0.113	12.43	−10.13	18	43
Median	0.95	0.149	0.192	0.054	13.25	−10.20	7	21

curve and informed us that the scores obtained by both methods were highly associated to the detection of active compounds; (2) For the PARP dataset, the analysis of the PCs allowed to easily estimate an optimal score selection threshold for Surflex-dock whereas ROC AUCs and ROC curve profiles were very close for all methods; (3) For the GART dataset, the PCs emphasized a better predictive performance of Surflex-dock scores over ICM's in the early part of the dataset, whereas the ROC curves profiles could lead to an opposite interpretation of the results.

Discussion

The goal of virtual screening methods in drug discovery programs is to predict the potential activity of the compounds of a compound collection on a specific target. The result is a list of compounds ranked by a scoring function that estimates the activity on the target (binding affinity, equilibrium constant, binding energy), which will be confirmed experimentally. Since scoring functions are still the most limiting factor in virtual screening in particular to predict activity, it is usual to select empirically the top scoring compounds for experimental tests [27–29]. Several performance metrics were developed over the years to evaluate the performance of virtual screening methods and guide the definition of the best protocols. The most used metrics suffer from three main limitations; (1) they focus on the predicted ranks of the compounds according to the scoring function instead of taking into account the value of the score; (2) they do not focus particularly on the top scoring compounds; (3) they do not allow an intuitive estimation of the score threshold that would give the best confidence into finding active compounds. In the present work, we suggested the use of a metric that tackles these limitations, the Predictiveness Curve.

As expected, the score values issued from scoring functions differ from one system to another rendering direct score comparisons between different systems difficult. That is why benchmarking metrics use specificity and selectivity to focus on the ranks of the compounds according to the scoring functions instead of the score values. In prospective virtual screening experiments, since score values and resulting ranks are available to the expert, both should be used to perform the compounds selection for experimental tests. As pointed out by Triballeau et al., a ROC AUC of 0.9 means that a randomly selected active molecule has a higher score than a randomly selected inactive 9 times out of 10 [2]. However, it does not mean that a hit would be confirmed experimentally with a probability of 0.9. ROC curves characterize the overall inherent quality of a virtual screening experiment and by no means are indicative of the quality of a particular compound or of a given subset of the initial compound collection. Finally, ROC plots do not allow a direct estimation of the size of an optimal subset in terms of activity potential, which is a critical task of virtual screening. We suggested in the present work the use of logistic regression and PC analysis to provide activity probabilities related to the scores obtained by the compounds after virtual screening.

Considering early recognition, it seems surprising that in other fields where this problem occurs, such as information retrieval, the metrics that are commonly used are not particularly efficient [30]. Likewise, there is still no consensus on the optimal metric to use to analyze the performance of virtual screening methods. ROC and EF are not able to discriminate the "ranking goodness" before the fractional threshold [4]. Furthermore, if two ranked lists display similar initial enhancements, but differ significantly just after the selection threshold, they would not be differentiated using EF or partial ROC metrics [2, 4, 31]. Since the overall distribution of the scores after virtual screening is taken into account by predictiveness models, the PC is able to perform efficient differentiation in this case. Hence, by summarizing the PC over a restricted range of compounds, pTG quantifies the enhancement of activity in the early part of the ranked molecular dataset and is a function of the overall success of the virtual screening experiment [20].

Now considering the choice of score selection thresholds towards prospective virtual screening experiments, Neyman and Pearson, who pioneered hypothesis testing, asserted that there is no general rule for balancing errors [32]. In any given case, the determination of "how the balance [between wrong and correct classifications] should be struck, must be left to the investigator" [32]. In summary, balancing false-positive and false-negative rates has "nothing to do with statistical theory but is based instead on context-dependent pragmatic considerations where informed personal judgment plays a vital role" [33]. Triballeau et al. transferred the ROC curve to the field of virtual screening and described how to retrieve score thresholds by maximizing either specificity or sensitivity from the ROC analysis [2]. The PC has the advantage to provide a probability-related interpretation of the scores by taking into account their variations, which efficiently complements the ROC curve for benchmarking purposes. Predictiveness curves allow for the detection of optimal score selection thresholds in an intuitive and straightforward way; a task for which the ROC curves are not adapted. Through the analysis of PCs, we were able to estimate optimal score selection thresholds for each virtual screening method used in the study, which were associated to satisfying EFs in each resulting subset. We were also able to detect an absence of association between the scores obtained by the compounds after virtual screening and the activity of the compounds, in

particular for experiments that yielded high ROC AUC values. We demonstrated these usages on the DUD dataset for three virtual screening methods, providing all PC and ROC curves with scores and metrics associated to each resulting subset (Figs. 3, 4, 5, 6; Tables 2, 3, 4).

The first objective of this paper is to introduce to the field of virtual screening the predictiveness curves for the purpose of benchmarking retrospective virtual screening experiments. We believe that benchmarking metrics have to take into account the values of the scores calculated in a virtual screening experiment for a better understanding of its results; which may also support the enhancement of the performances of scoring functions. The second objective of this paper is to provide a method to define score selection thresholds to be used for prospective virtual screenings, in order to select an optimal number of compounds to be tested experimentally in drug discovery programs. The predictiveness curves graphically emphasize the differences in scores that are relevant for the detection of active compounds in a virtual screening experiment and ease the process of defining optimal thresholds. When retrospective studies on a specific target allowed to detect optimal score selection thresholds, considering that a prospective virtual screening experiment could be performed under similar conditions, we can expect score variations to be reproducible and the corresponding score thresholds to be transferable. Therefore, the resulting subset of compounds selected when applying the estimated score threshold would be expected to be highly enriched in active compounds. However, score selection thresholds defined in retrospective studies must be considered carefully when applied for the selection of molecular subsets in prospective studies. It is important to keep in mind that all performance measures should be interpreted in the context of the composition of the benchmarking datasets [34, 35] and that the score selection thresholds that would be estimated during the benchmark should be adapted to the composition of the dataset that will be used for prospective screening.

Conclusion

The value of a continuous test in predicting a binary outcome can be assessed by considering two aspects: discrimination and outcome prediction. In the present study, we proposed predictiveness curves as a complement to the existing methods to analyze the results of virtual screening methods. Logistic regression models can be used to evaluate the probability of each compound to be active given the score it obtained through the virtual screening method. The PC then provides an intuitive way to visualize the data and allows for an efficient comparison of the performance of virtual screening methods, especially considering the early recognition problem. Performance metrics are easily estimated from the predictiveness plots: TG, pTG, PPV, NPV, TPF and NPF. PC also ease the process of extracting optimal score selection thresholds from virtual screening results, which is a valuable step to proceed to prospective virtual screening. The enhancement of activity attributed to the variations of virtual screening scores can then be quantified in the resulting subsets of compounds using the pTG.

Visualizing both the predictiveness curve and the ROC curve empowers the analysis of virtual screening results. The two measures, however, summarize different aspects of the predictive performance of scores and thus answer different questions [14, 20]. On the one hand, we are interested in the ROC curve because it summarizes the inherent capacity of a virtual screening method to distinguish between active and inactive compounds. This information would aid in the decision to whether or not apply a virtual screening method in the first place. On the other hand, the predictiveness curve informs us on the association between virtual screening scores and the activity of the compounds. This information would aid in decision making when performing prospective virtual screening experiments. By simultaneously displaying PC and ROC, we believe researchers will be better equipped to analyze and understand the results of virtual screening experiments.

Additional files

Additional file 1: Table S1. Summary of the partial metrics at 2 % and 5 % of the ordered dataset for virtual screens performed using Surflex-dock.

Additional file 2: Table S2. Summary of the partial metrics at 2 % and 5 % of the ordered dataset for virtual screens performed using ICM.

Additional file 3: Table S3. Summary of the partial metrics at 2 % and 5 % of the ordered dataset for virtual screens performed using Autodock Vina.

Abbreviations

VS: virtual screening; PC: predictiveness curve; EF: enrichment factor; ROC: receiver operating characteristic; AUC: area under the curve; pAUC: partial AUC; BEDROC: Boltzmann-enhanced discrimination of ROC; RIE: robust initial enhancement; AUAC: area under the accumulation curve; TPF: true positive fraction; FPF: false positive fraction; TG: total gain; pTG: partial total gain; DUD: directory of useful decoys; CDF: cumulative distribution function; ACE: angiotensin-converting enzyme; ACHE: acetylcholin esterase; ADA: adenosine deaminase; ALR2: aldose reductase; AMPC: AmpC beta lactamase; AR: androgen receptor; CDK2: cyclin dependent kinase 2; COMT: catechol O-methyltransferase; COX-1: cyclooxygenase-1; COX-2: cyclooxygenase-2; DHFR: dihydrofolate reductase; EGFR: epidermal growth factor receptor kinase; ER: agoestrogen receptor agonist; ER: antagoestrogen receptor antagonist; FGFR1: fibroblast growth factor receptor kinase; FXA: factor Xa; GART: glycinamide ribonucleotide transformylase; GPB: glycogen phosphorylase beta; GR: glucocorticoid receptor; HIVPR: HIV protease; HIVRTHIV: reverse transcriptase; HMGR: hydroxymethylglutaryl-CoA reductase; HSP90: human heat shock protein 90 kinase; INHA: enoyl ACP reductase; MR: mineralocorticoid receptor; NA: neuraminidase; P38: P38 mitogen activated protein kinase; PARP: poly(ADP-ribose)

polymerase; PDE5: phosphodiesterase V; PDGFR-β: platelet derived growth factor receptor kinase beta; PNP: purine nucleoside phosphorylase; PPAR: peroxisome proliferator activated receptor gamma; PR: progesterone receptor; RXR: retinoic X receptor alpha; SAHH: S-adenosyl-homocystein hydrolase; SRC: tyrosine kinase SRC; THR: thrombin; TK: thymidine kinase; TRP: trypsin; VEGFR2: vascular endothelial growth factor receptor kinase; NR: nuclear receptors.

Authors' contributions
Conceived and designed the experiments: AL, JFZ, VV and MM. Performed the experiments: CE and HG. Analyzed the data: CE, HG and MM. Wrote the paper: CE, HG, AL, VV, MM. All authors discussed the results and commented on the manuscript. All authors read and approved the final manuscript.

Author details
[1] Laboratoire Génomique Bioinformatique et Applications, EA 4627, Conservatoire National des Arts et Métiers, 292 rue Saint Martin, 75003 Paris, France. [2] Equipe MSDMA, Laboratoire CEDRIC, EA 4629, Conservatoire National des Arts et Métiers, 292 rue Saint Martin, 75003 Paris, France. [3] Université de Lyon, 69622 Lyon, France. [4] UMRESTTE, Université Lyon 1, 69373 Lyon, France. [5] UMRESTTE, IFSTTAR, 69675 Bron, France.

Acknowledgements
We thank Dr N. Lagarde for fruitful discussions. We thank Pr. Jain for generously providing the Surflex-dock software and Molsoft LLC for providing academic licenses for the ICM suite. HG is recipient of an ANSM fellowship. CE is recipient of a MNESR fellowship.

Competing interests
The authors declare that they have no competing interests.

Received: 23 July 2015 Accepted: 20 October 2015
Published online: 04 November 2015

References

1. Alvarez JC (2004) High-throughput docking as a source of novel drug leads. Curr Opin Chem Biol 8:365–370
2. Triballeau N, Acher F, Brabet I, Pin J-P, Bertrand H (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
3. McClish DK (1989) Analyzing a portion of the ROC Curve. Med Decis Mak 9:190–195
4. Truchon J, Bayly C (2007) Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem. J Chem Inf Model 47:488–508
5. Sheridan RP, Singh SB, Fluder EM, Kearsley SK (2001) Protocols for bridging the peptide to nonpeptide gap in topological similarity searches. J Chem Inf Comput Sci 41:1395–1406
6. Zhao W, Hevener K, White S, Lee R, Boyett J (2009) A statistical framework to evaluate virtual screening. BMC Bioinformatics 10:225
7. Kairys V, Fernandes MX, Gilson MK (2006) Screening drug-like compounds by docking to homology models: a systematic study. J Chem Inf Model 46:365–379
8. Nicholls A (2008) What do we know and when do we know it? J Comput Aided Mol Des 22:239–255
9. Copas J (1999) The effectiveness of risk scores: the logit rank plot. J R Stat Soc Ser C Appl Stat 48:165–183
10. Huang Y, Sullivan Pepe M, Feng Z (2007) Evaluating the predictiveness of a continuous marker. Biometrics 63:1181–1188
11. Pepe MS, Feng Z, Huang Y, Longton G, Prentice R, Thompson IM, Zheng Y (2008) Integrating the predictiveness of a marker with its performance as a classifier. Am J Epidemiol 167:362–368
12. Huang Y, Pepe MS (2009) A parametric ROC model-based approach for evaluating the predictiveness of continuous markers in case-control studies. Biometrics 65:1133–1144
13. Huang Y, Pepe MS (2010) Semiparametric methods for evaluating the covariate-specific predictiveness of continuous markers in matched case-control studies. J R Stat Soc Ser C Appl Stat 59:437–456
14. Viallon V, Latouche A (2011) Discrimination measures for survival outcomes: connection between the AUC and the predictiveness curve. Biom J 53:217–236
15. Huang N, Shoichet B, Irwin J (2006) Benchmarking sets for molecular docking. J Med Chem 49:6789–6801
16. Jain AN (2003) Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine. J Med Chem 46:499–511
17. Abagyan R, Totrov M, Kuznetsov D (1994) ICM—a new method for protein modeling and design: applications to docking and structure prediction from the distorted native conformation. J Comput Chem 15:488–506
18. 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
19. Bura E, Gastwirth JL (2001) The binary regression quantile plot: assessing the importance of predictors in binary regression visually. Biometrical J 43:5–21
20. Sachs MC, Zhou XH (2013) Partial summary measures of the predictiveness curve. Biom J 55:589–602
21. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE (2004) UCSF Chimera—a visualization system for exploratory research and analysis. J Comput Chem 25:1605–1612
22. Schapira M, Abagyan R, Totrov M (2003) Nuclear hormone receptor targeted virtual screening. J Med Chem 46:3045–3059
23. Baxter J (1981) Local optima avoidance in depot location. J Oper Res Soc 32:815–819
24. Nocedal J, Wright SJ (199) Numerical optimization. Springer, New York **(Springer Series in Operations Research and Financial Engineering)**
25. R Core Team (2015) R: a language and environment for statistical computing. R Foundation for Statistical Computing
26. Sing T, Sander O, Beerenwinkel N, Lengauer T (2005) ROCR: visualizing classifier performance in R. Bioinformatics 21:3940–3941
27. Pan Y, Huang N, Cho S, MacKerell AD (2003) Consideration of molecular weight during compound selection in virtual target-based database screening. J Chem Inf Comput Sci 43:267–272
28. Cross JB, Thompson DC, Rai BK, Baber JC, Fan KY, Hu Y, Humblet C (2009) Comparison of several molecular docking programs: pose prediction and virtual screening accuracy. J Chem Inf Model 49:1455–1474
29. Verdonk M, Berdini V, Hartshorn M, Mooij W, Murray C, Taylor R, Watson P (2004) Virtual screening using protein-ligand docking: avoiding artificial enrichment. J Chem Inf Comput Sci 44:793–806
30. Edgar S, Holliday J, Willett P (2000) Effectiveness of retrieval in similarity searches of chemical databases: a review of performance measures. J Mol Graph Model 18:343–357
31. Jain AN (2008) Bias, reporting, and sharing: computational evaluations of docking methods. J Comput Aided Mol Des 22:201–212
32. Neyman J, Pearson E (1992) On the problem of the most efficient tests of statistical hypotheses. Philos Trans R Soc 231:289–337
33. Berk KN, Carlton MA, Statistician TA (2003) Confusion over measures of evidence (p's) versus errors (alpha's) in classical statistical testing. Am Stat 57:171–182
34. Muegge I, Enyedy IJ (2004) Virtual screening for kinase targets. Curr Med Chem 11:693–707
35. Rohrer SG, Baumann K (2008) Impact of benchmark data set topology on the validation of virtual screening methods: exploration and quantification by spatial statistics. J Chem Inf Model 48:704–718

Target	Surflex-dock – Top 2% dataset					Surflex-dock – Top 5% dataset				
	pTG 2%	pAUC 2%	EF 2%	Actives 2%	Cpds 2%	pTG 5%	pAUC 5%	EF 5%	Actives 5%	Cpds 5%
ACE	0.006	0.040	4.07	4	37	0.005	0.091	2.84	7	93
ACHE	0.002	0.022	2.34	5	80	0.002	0.048	1.68	9	200
ADA	0.237	0.154	8.67	7	20	0.153	0.186	5.05	10	49
ALR2	0.097	0.300	18.70	10	21	0.064	0.364	7.55	10	52
AMPC	0.054	0.000	0.00	0	17	0.045	0.000	0.00	0	41
AR	0.008	0.106	6.92	11	59	0.008	0.147	3.54	14	147
CDK2	0.078	0.091	6.77	10	44	0.053	0.167	4.65	17	109
COMT	0.146	0.182	8.71	2	10	0.095	0.182	3.63	2	24
COX-1	0.012	0.040	7.88	4	19	0.010	0.137	3.98	5	47
COX-2	0.141	0.033	3.04	26	275	0.106	0.075	2.86	61	686
DHFR	0.371	0.214	13.38	110	176	0.225	0.294	7.07	145	439
EGFR	0.003	0.076	5.04	48	330	0.003	0.109	2.86	68	824
ER ago	0.101	0.057	8.17	11	53	0.068	0.184	5.96	20	132
ER antago	0.253	0.119	16.52	13	30	0.144	0.325	9.15	18	75
FGFR1	0.035	0.058	4.55	11	94	0.026	0.098	2.99	18	234
FXA	0.310	0.129	11.24	34	122	0.181	0.269	8.77	65	299
GART	0.392	0.029	4.84	4	19	0.280	0.123	6.49	13	46
GPB	0.060	0.042	3.83	4	44	0.043	0.092	3.07	8	110
GR	0.001	0.094	7.63	12	61	0.001	0.149	4.34	17	152
HIVPR	0.309	0.227	14.18	18	43	0.198	0.328	8.95	28	106
HIVRT	0.040	0.143	7.95	7	32	0.032	0.183	4.60	10	79
HMGR	0.481	0.444	25.13	18	31	0.253	0.510	11.96	21	76
HSP90	0.029	0.000	0.00	0	21	0.024	0.005	0.54	1	51
INHA	0.024	0.076	3.95	7	69	0.019	0.098	2.54	11	169
MR	0.179	0.322	21.70	7	14	0.124	0.469	10.52	8	33
NA	0.380	0.259	20.13	20	39	0.237	0.423	10.52	26	97
P38	0.114	0.052	4.40	40	192	0.086	0.095	2.91	66	480
PARP	0.242	0.183	14.14	10	28	0.142	0.310	8.49	15	70
PDE5	0.019	0.039	2.79	5	42	0.016	0.058	1.81	8	104
PNP	0.231	0.152	13.82	14	22	0.173	0.309	8.29	21	55
PPAR	0.517	0.278	25.00	43	65	0.302	0.544	13.38	57	161
PR	0.011	0.026	1.80	1	22	0.010	0.033	0.73	1	54
RXR	0.308	0.167	16.84	7	16	0.223	0.320	10.86	11	39
SAHH	0.153	0.023	2.98	2	28	0.104	0.067	3.03	5	69
SRC	0.042	0.123	7.52	24	130	0.031	0.169	4.40	35	324
THR	0.243	0.131	10.33	15	51	0.156	0.237	8.29	30	127
TK	0.025	0.062	4.37	2	19	0.021	0.096	2.71	3	46
TRP	0.577	0.348	19.98	20	35	0.364	0.479	13.41	33	86
VEGFR2	0.024	0.036	3.40	6	60	0.019	0.065	2.27	10	150
Minimum	0.001	0.000	0.00	0	10	0.001	0.000	0.00	0	24
Maximum	0.577	0.444	25.13	110	330	0.364	0.544	13.41	145	824
Mean	0.160	0.125	9.30	15	63	0.104	0.201	5.56	23	157
Median	0.101	0.094	7.63	10	39	0.068	0.167	4.40	14	97

Table S1. Summary of the partial metrics at 2% and 5% of the ordered dataset for virtual screens performed using Surflex-dock.

Target	ICM – Top 2% dataset					ICM – Top 5% dataset				
	pTG 2%	pAUC 2%	EF 2%	Actives 2%	Cpds 2%	pTG 5%	pAUC 5%	EF 5%	Actives 5%	Cpds 5%
ACE	0.111	0.161	9.16	9	37	0.077	0.217	5.67	14	93
ACHE	0.027	0.000	0.00	0	80	0.022	0.003	0.19	1	200
ADA	0.027	0.000	0.00	0	20	0.027	0.000	0.00	0	49
ALR2	0.080	0.038	1.87	1	21	0.055	0.081	3.02	4	52
AMPC	0.035	0.111	9.04	4	17	0.028	0.211	7.50	8	41
AR	0.034	0.132	9.44	15	59	0.026	0.216	6.31	25	147
CDK2	0.199	0.089	9.01	13	43	0.132	0.228	7.45	27	108
COMT	0.079	0.000	0.00	0	10	0.051	0.087	3.63	2	24
COX-1	0.114	0.114	9.85	5	19	0.086	0.216	5.58	7	47
COX-2	0.015	0.002	0.23	2	275	0.012	0.009	0.75	16	686
DHFR	0.103	0.042	3.89	32	176	0.078	0.097	3.71	76	439
EGFR	0.143	0.190	13.77	131	330	0.101	0.280	6.99	166	824
ER ago	0.238	0.232	17.08	23	53	0.160	0.342	9.24	31	132
ER antago	0.111	0.136	11.44	9	30	0.068	0.207	5.08	10	75
FGFR1	0.012	0.023	2.07	5	94	0.011	0.040	1.33	8	234
FXA	0.093	0.111	7.18	21	118	0.065	0.171	5.61	41	295
GART	0.308	0.081	12.09	10	19	0.244	0.268	8.49	17	46
GPB	0.167	0.087	13.41	14	44	0.118	0.335	11.88	31	110
GR	0.005	0.103	5.72	9	61	0.005	0.154	4.34	17	152
HIVPR	0.048	0.019	2.36	3	43	0.038	0.069	3.83	12	106
HIVRT	0.046	0.076	4.54	4	32	0.037	0.155	5.52	12	79
HMGR	0.281	0.324	18.15	13	31	0.148	0.374	8.54	15	76
HSP90	0.029	0.000	0.00	0	21	0.027	0.004	0.54	1	51
INHA	0.005	0.013	1.15	2	68	0.004	0.041	2.32	10	168
MR	0.163	0.400	21.70	7	14	0.105	0.475	10.52	8	33
NA	0.557	0.311	25.16	25	39	0.336	0.629	15.37	38	97
P38	0.026	0.045	2.75	25	192	0.024	0.055	1.28	29	480
PARP	0.150	0.030	4.24	3	28	0.112	0.120	5.66	10	70
PDE5	0.232	0.149	10.06	18	42	0.178	0.240	6.55	29	104
PNP	0.024	0.065	4.94	5	22	0.019	0.176	5.92	15	55
PPAR	0.211	0.179	13.95	24	65	0.134	0.304	8.45	36	161
PR	0.017	0.092	7.19	4	22	0.014	0.167	5.13	7	54
RXR	0.026	0.342	24.06	10	16	0.021	0.467	10.86	11	39
SAHH	0.084	0.021	2.98	2	28	0.064	0.071	3.03	5	69
SRC	0.168	0.169	14.42	46	130	0.116	0.281	7.29	58	324
THR	0.216	0.062	6.20	9	51	0.145	0.160	6.36	23	127
TK	0.002	0.000	0.00	0	19	0.002	0.000	0.00	0	46
TRP	0.047	0.000	0.00	0	35	0.035	0.008	0.41	1	86
VEGFR2	0.059	0.039	3.97	7	60	0.046	0.106	3.86	17	150
Minimum	0.002	0.000	0.00	0	10	0.002	0.000	0.00	0	24
Maximum	0.557	0.400	25.16	131	330	0.336	0.629	15.37	166	824
Mean	0.110	0.102	7.77	13	63	0.076	0.181	5.34	21	157
Median	0.080	0.081	6.20	7	39	0.055	0.167	5.58	14	97

Table S2. Summary of the partial metrics at 2% and 5% of the ordered dataset for virtual screens performed using ICM.

Target	Autodock Vina – Top 2% dataset					Autodock Vina – Top 5% dataset				
	pTG 2%	pAUC 2%	EF 2%	Actives 2%	Cpds 2%	pTG 5%	pAUC 5%	EF 5%	Actives 5%	Cpds 5%
ACE	0.020	0.048	3.05	3	37	0.019	0.075	2.84	7	93
ACHE	0.024	0.038	3.74	8	80	0.019	0.107	4.11	22	200
ADA	0.020	0.000	0.00	0	20	0.018	0.000	0.00	0	49
ALR2	0.098	0.028	3.74	2	21	0.071	0.154	6.80	9	52
AMPC	0.021	0.013	2.26	1	17	0.019	0.034	0.94	1	41
AR	0.161	0.157	11.96	19	59	0.108	0.268	7.83	31	147
CDK2	0.087	0.117	9.70	14	43	0.063	0.190	5.24	19	108
COMT	0.000	0.091	4.35	1	10	0.000	0.182	5.44	3	24
COX-1	0.154	0.113	11.82	6	19	0.102	0.250	7.17	9	47
COX-2	0.322	0.234	18.03	154	275	0.193	0.397	10.14	216	686
DHFR	0.215	0.070	5.47	45	176	0.150	0.118	3.56	73	439
EGFR	0.048	0.038	3.26	31	330	0.036	0.071	2.19	52	824
ER ago	0.314	0.192	17.08	23	53	0.186	0.383	9.84	33	132
ER antago	0.059	0.110	8.90	7	30	0.040	0.173	5.08	10	75
FGFR1	0.012	0.003	0.83	2	94	0.010	0.016	0.67	4	234
FXA	0.029	0.011	1.37	4	118	0.023	0.036	1.50	11	295
GART	0.108	0.000	0.00	0	19	0.087	0.005	1.00	2	46
GPB	0.113	0.026	2.87	3	44	0.081	0.101	4.22	11	110
GR	0.023	0.099	5.72	9	61	0.019	0.111	2.55	10	152
HIVPR	0.147	0.038	4.73	6	43	0.099	0.091	3.51	11	106
HIVRT	0.047	0.121	7.95	7	32	0.038	0.161	4.14	9	79
HMGR	0.015	0.035	2.79	2	31	0.012	0.049	1.14	2	76
HSP90	0.039	0.000	0.00	0	21	0.032	0.004	0.54	1	51
INHA	0.079	0.191	12.04	21	68	0.051	0.257	6.50	28	168
MR	0.346	0.229	18.60	6	14	0.215	0.517	14.47	11	33
NA	0.019	0.000	0.00	0	39	0.018	0.000	0.00	0	97
P38	0.031	0.012	1.54	14	192	0.026	0.049	2.29	52	480
PARP	0.114	0.071	4.24	3	28	0.080	0.091	3.39	6	70
PDE5	0.047	0.009	1.68	3	42	0.037	0.043	1.81	8	104
PNP	0.011	0.000	0.00	0	22	0.009	0.000	0.00	0	55
PPAR	0.304	0.219	16.28	28	65	0.183	0.372	10.33	44	161
PR	0.012	0.009	1.80	1	22	0.010	0.027	1.47	2	54
RXR	0.653	0.330	26.47	11	16	0.362	0.620	14.81	15	39
SAHH	0.126	0.069	8.95	6	28	0.086	0.174	4.84	8	69
SRC	0.099	0.053	5.64	18	130	0.070	0.135	4.78	38	324
THR	0.129	0.097	7.57	11	51	0.091	0.149	3.87	14	127
TK	0.019	0.000	0.00	0	19	0.015	0.000	0.00	0	46
TRP	0.037	0.037	3.00	3	35	0.029	0.069	2.03	5	86
VEGFR2	0.007	0.062	4.54	8	60	0.006	0.101	2.72	12	150
Minimum	0.000	0.000	0.00	0	10	0.000	0.000	0.00	0	24
Maximum	0.653	0.330	26.47	154	330	0.362	0.620	14.81	216	824
Mean	0.105	0.076	6.20	12	63	0.070	0.143	4.20	20	157
Median	0.048	0.048	4.24	6	39	0.038	0.101	3.51	10	97

Table S3. Summary of the partial metrics at 2% and 5% of the ordered dataset for virtual screens performed using Autodock Vina.

1.1.2. Discussion

L'objectif de cette étude était de proposer une nouvelle métrique adaptée à l'évaluation de la reconnaissance précoce des ligands et permettant de prendre directement en compte les scores calculés par les logiciels de criblage virtuel. Pour ce faire, nous avons décrit l'usage des courbes de prédictivité dans le cadre d'une application « structure-based » en utilisant la base de données DUD.[159] Le modèle logistique utilisé permet de lier les variations de score aux rangs des ligands pour répondre, notamment, aux questions suivantes : (i) à partir de quelle valeur de score obtient-on une probabilité donnée de détecter un ligand ? et (ii) lorsque l'on augmente un score d'affinité par une valeur donnée, comment évolue la probabilité de détecter un ligand ? Les courbes de prédictivité permettent ensuite de visualiser les résultats des modèles de manière intuitive, pour quantifier la performance globale d'une méthode ou sélectionner des scores d'affinité pertinents pour la recherche de ligands lors de criblages virtuels prospectifs.

1.1.2.1. Sélection de seuils de score

Au sujet de la sélection de seuils de score, nous pouvons observer les métriques obtenues lors des criblages réalisés avec Surflex-dock[349,350] sur les jeux de données NA (NeurAminidase) et HMGR (HydroxyMethylGlutaryl-CoA Reductase) de la banque DUD.[159] Ces deux jeux de données sont de taille similaire et présentent des taux de ligands proches (respectivement, 35 ligands pour 1480 decoys, soit un taux de ligands de 2.31%, et 49 ligands pour 1874 decoys, soit un taux de ligands de 2.55%). Les AUC ROC retrouvées sont très proches et les courbes de ROC présentent des profils très similaires (***Figure 70A*** et ***Figure 71A***). Cependant, les distributions de score obtenues lors de ces deux expériences diffèrent (***Figure 70B*** et ***Figure 71B***). Cette information n'étant pas disponible en utilisant les seules courbes de ROC, les courbes de prédictivité permettent de compléter l'analyse en quantifiant les écarts de score sous la forme de probabilités de détection des ligands (***Figure 70C*** et ***Figure 71C***). Tout comme les métriques ROC, les métriques issues des courbes de prédictivité sont indépendantes des taux d'actifs dans les chimiothèques étudiées. Ici, la visualisation des courbes de prédictivité et la comparaison des valeurs de pTG (partial Total Gain) permettent d'affirmer que Surflex-dock[349,350] a produit de plus larges variations entre les scores attribués aux ligands et aux decoys sur le jeu de données HMGR, par rapport au jeu de données NA (***Tableau 12***).

À la manière de ce qui est réalisé en épidémiologie, il est également possible d'identifier des fractions de composés pour lesquelles les scores présentent une probabilité faible, intermédiaire ou forte d'être attribués à un composé actif. Dans ces deux expériences, nous avons défini des

fractions de probabilité intermédiaire entre 1.6% et 7% du jeu de données NA classé (scores respectifs : 12.13 et 10.26, ***Figure 70C***) et entre 1.7% et 6% du jeu de données HMGR classé (scores respectifs : 9.06 et 8.11, ***Figure 71C***). Les scores obtenus à chaque seuil sont directement dépendants de la cible choisie, de la chimiothèque étudiée et des conditions de ces deux expériences. Cette analyse permet donc de détecter et de sélectionner des seuils de score d'exigence variable pour estimer, ultérieurement, la pertinence des scores d'affinité obtenus lors d'expériences de criblage prospectives (sous réserve que celles-ci soient conduites selon des conditions similaires).

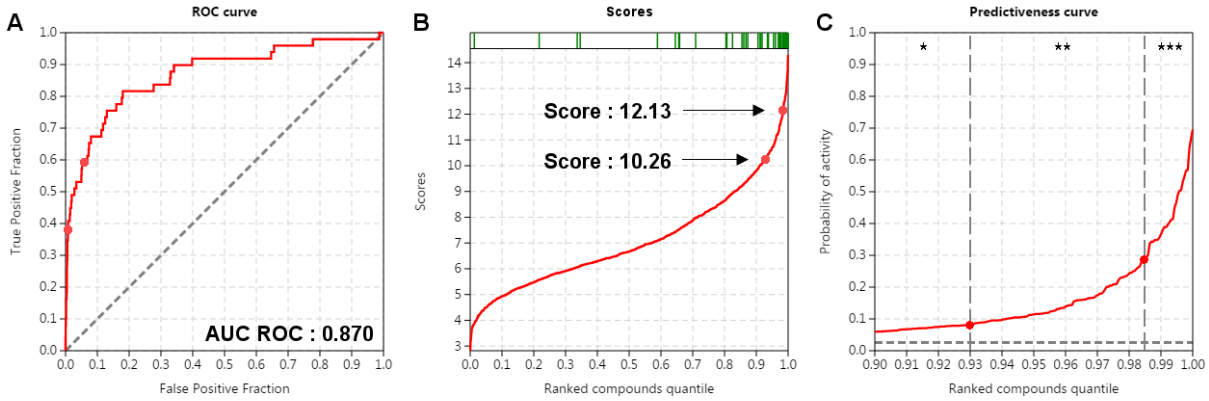


Figure 70. *(A) Courbe de ROC, (B) distribution des scores et (C) courbe de prédictivité des résultats de criblage virtuel obtenus avec Surflex-dock*[349,350] *sur le jeu de données NA de la banque DUD.*[159] *Les rangs des ligands sont représentés en vert avec les distributions de scores. Les étoiles représentent les probabilités faible (*), intermédiaire (**) ou forte (***) de détecter des composés actifs dans les fractions respectives. Les pointillés gris foncé correspondent à un tirage aléatoire.*[30]

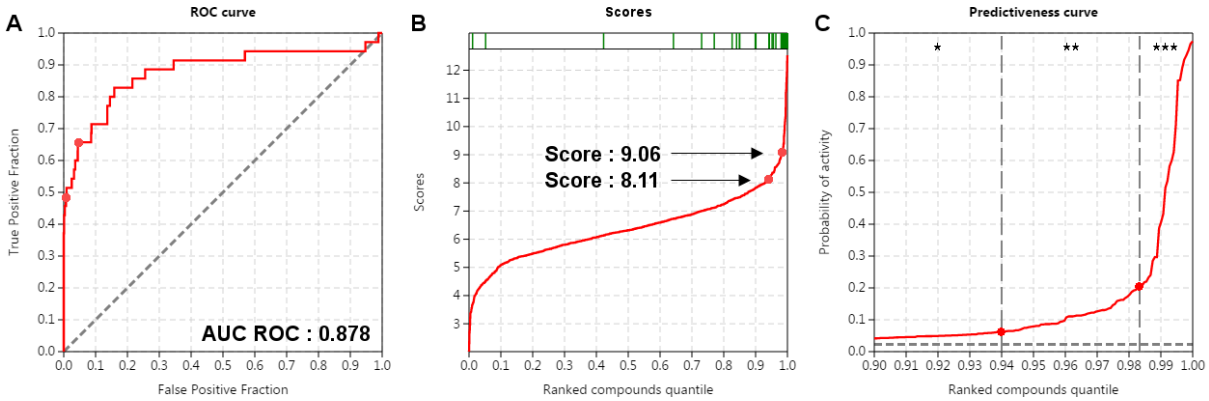


Figure 71. *(A) Courbe de ROC, (B) distribution des scores et (C) courbe de prédictivité des résultats de criblage virtuel obtenus avec Surflex-dock*[349,350] *sur le jeu de données HMGR de la banque DUD.*[159] *Les rangs des ligands sont représentés en vert avec les distributions de scores. Les étoiles représentent les probabilités faible (*), intermédiaire (**) ou forte (***) de détecter des composés actifs dans les fractions respectives. Les pointillés gris foncé correspondent à un tirage aléatoire.*[30]

Jeu de données	Total Gain (TG)	partial Total Gain (pTG)			
		1%	3%	5%	10%
NA	0.633	0.476	0.317	0.237	0.148
HMGR	0.641	0.749	0.365	0.253	0.142

Tableau 12. *Métriques obtenues lors des criblages des jeux de données HMGR et NA de la banque DUD*[159] *avec Surflex-dock.*[349,350]

1.1.2.2. Développement des fonctions de score

Dans une démarche de développement d'une fonction de score, l'utilisateur pourrait également étudier les différentes fractions de composés définies afin de chercher à comprendre les raisons pour lesquelles la fonction de score étudiée a produit, ou non, de larges variations de score. En étendant ce type d'analyse à des évaluations large échelle, les courbes de prédictivité permettent notamment : (i) d'identifier des classes de composés pour lesquelles une fonction de score serait particulièrement performante et (ii) d'optimiser une fonction de score de manière à la rendre plus « générale », ou plus spécifiquement performante pour certaines classes de composés. Dans le cadre de ces deux exemples d'application, les courbes de prédictivité constituent un outil puissant qui complète efficacement l'analyse réalisée grâce aux courbes de ROC et aux autres métriques utilisées en criblage virtuel.

1.1.2.3. Aspects théoriques et signification des métriques

Les métriques de prédictivité permettent de quantifier l'association entre les variations des scores et le classement des composés actifs de manière globale et partielle, respectivement avec le Total Gain (TG) et le partial Total Gain (pTG). Le TG quantifie la capacité d'une méthode à maximiser les écarts de score entre composés actifs et inactifs sur l'ensemble d'un jeu de données. Pour quantifier la reconnaissance précoce, le pTG évalue le gain de probabilité de détecter un composé actif dans une fraction du jeu de données, au-delà de la probabilité du tirage aléatoire.

Pour mieux comprendre le comportement des métriques de prédictivité, nous nous proposons ici d'étudier 7 distributions de score différentes (***Figure 72A***) rapportées à une unique distribution des composés actifs (***Figure 72B-C***). Le jeu de données théorique utilisé comporte 10.000 composés dont 400 sont actifs (taux d'actifs : 4%) et distribués de la manière suivante : aucun n'est retrouvé parmi les 20% des rangs les plus hauts, puis ceux-ci sont attribués aléatoirement aux 80% des rangs les plus bas selon une loi de densité de distribution

exponentielle. Les courbes de prédictivité sont ensuite calculées pour chaque distribution de score (***Figure 72D-F***), avec leurs métriques associées.

Les métriques issues de chaque distribution de score permettent de les séparer en deux groupes (***Tableau 13***). Les fourchettes de TG obtenues pour le groupe 1 (distributions 1, 2, 4 et 5) et le groupe 2 (distributions 3, 6 et 7) sont respectivement de 0.489-0.516 et 0.297-0.403, ce qui permet d'affirmer que les distributions de score du groupe 1 sont celles qui maximisent les écarts de score entre les composés actifs et inactifs. Plus spécifiquement, si l'on évalue les fractions précoces des jeux de données grâce aux pTG, les distributions de score du groupe 2 sont celles qui permettent de sélectionner les seuils de score plus pertinents, à partir desquels la probabilité d'identifier un composé actif deviendra plus importante. Ici, le jeu de données 7 est celui qui comporte les plus larges variations de scores au sein des composés actifs, ce qui produit les plus larges variations de probabilité parmi les différentes courbes de prédictivité.

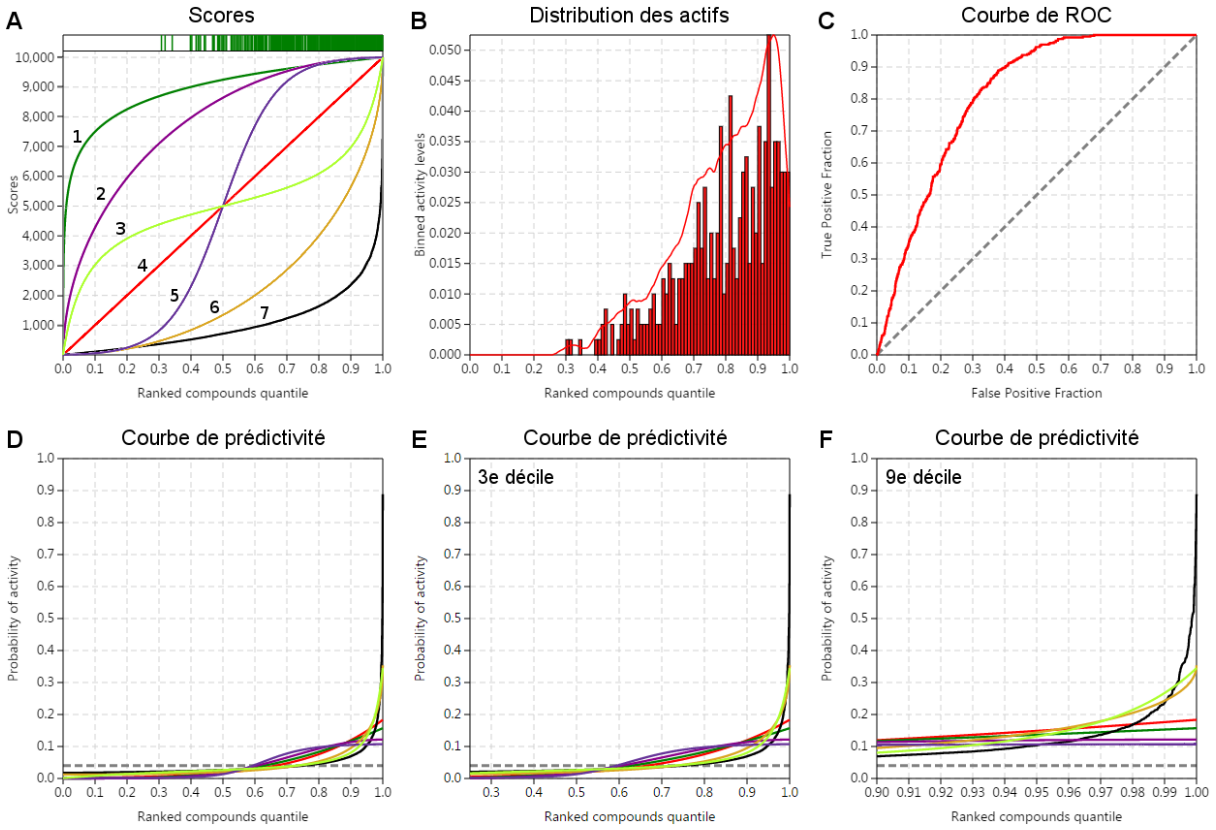


Figure 72. *Distribution des scores (A), des actifs (B) et courbe de ROC (C) obtenues pour les jeux de données simulés. Les courbes de prédictivité sont représentées sur l'ensemble des jeux de données classés (D) ou au-delà des 3^{e} (E) et 9^{e} (F) déciles. Les rangs des ligands sont représentés en vert avec les distributions de scores. La courbe rouge représentée sur la distribution des actifs correspond à une estimation de densité réalisée avec une fenêtre de 5% et un noyau Epanechnikov.*

Jeu de données	Groupe	Total Gain (TG)	partial Total Gain (pTG)			
			1%	3%	5%	10%
Jeu 1	1	0.497	0.120	0.115	0.111	0.101
Jeu 2	1	0.506	0.085	0.085	0.084	0.082
Jeu 3	2	0.364	0.272	0.215	0.175	0.120
Jeu 4	1	0.489	0.145	0.138	0.131	0.114
Jeu 5	1	0.516	0.070	0.070	0.070	0.069
Jeu 6	2	0.403	0.245	0.199	0.169	0.125
Jeu 7	2	0.297	0.332	0.208	0.157	0.101

Tableau 13. *Métriques obtenues pour les jeux de données simulés présentés en* ***Figure 72****.*

Par ailleurs, si l'on conserve ces 7 distributions de score en modifiant progressivement la distribution des composés actifs pour la faire approcher une classification idéale (tous les meilleurs scores sont associés aux actifs), les écarts de TG diminuent et les valeurs de TG tendent vers 1 pour chacun des jeux de données théoriques. Selon le même modèle, les valeurs de pTG dans les fractions précoces augmentent progressivement et leurs meilleures valeurs sont associées aux distributions de score du groupe 2 (distributions 3, 6 et 7).

1.1.2.4. Autres applications

Dans cette publication, nous avons présenté l'usage des courbes et métriques de prédictivité dans le cas de criblages virtuels « structure-based ». Cependant, tout comme les métriques ROC et les autres métriques utilisées en criblage virtuel, les métriques de prédictivité peuvent également être utilisées pour l'analyse de criblages virtuels « ligand-based ».

1.2. Screening Explorer : Un outil interactif pour l'évaluation des méthodes de criblage

1.2.1. Introduction & Publication

Après avoir proposé l'utilisation des courbes de prédictivité pour l'évaluation des méthodes de criblage virtuel et suite à de bons retours du public lors de séminaires, nous avons souhaité rendre l'utilisation de cette métrique accessible à une large audience. Afin d'éviter l'installation de logiciels et pour permettre l'accès le plus rapide aux méthodes d'analyse, nous avons choisi le format web pour créer un outil regroupant les principales métriques utilisées en criblage virtuel, ainsi que les métriques de prédictivité. Screening Explorer est composé de deux modules. Le premier permet d'analyser et de comparer les performances de plusieurs méthodes de criblage virtuel, grâce à l'ensemble des métriques de performance citées en point 3.2.2 (***Figure 74***), tandis que le second permet de réaliser différents types de reclassements consensus (***Figure 75***). La capture d'écran donnée en ***Figure 73*** présente la page d'accueil de l'outil, qui est disponible gratuitement à l'adresse : http://stats.drugdesign.fr.

Screening Explorer permet de générer facilement des graphiques interactifs des : (i) courbes de prédictivité, (ii) courbes de ROC, (iii) courbes d'enrichissement, (iv) distributions des scores et (v) distributions des composés actifs, pour comparer jusqu'à 10 méthodes de criblage virtuel exécutées sur un même jeu de données. Les différentes métriques globales sont calculées lors de la génération des graphiques, tandis que les métriques partielles sont calculées à la volée lorsque l'utilisateur interagit avec un graphique (***Figure 74***). Cette fonction est particulièrement utile puisqu'elle autorise une grande accessibilité aux métriques et rend l'analyse intuitive. Afin de faciliter l'interprétation des métriques, lorsque l'utilisateur sélectionne une fraction de composés sur l'un des graphiques, cette même fraction est reportée et représentée sur tous les autres graphiques. Après génération des graphiques, ceux-ci peuvent être facilement exportés au format png. Les métriques globales et partielles peuvent également être enregistrées au format texte. Ainsi, les figures utilisées dans la partie résultats précédente (1.1.2) ont été générées à l'aide de Screening Explorer (***Figure 70*** à ***Figure 72***).

Pour permettre une prise en main rapide de cet outil, nous avons réalisé une interface minimale et permis l'accès à deux jeux de données de démonstration (***Figure 73***). Nous avons également inséré de nombreuses info-bulles (« tooltips ») pour aider l'utilisateur à comprendre les paramètres et les fonctions de cet outil, mais aussi pour faciliter la compréhension de la

signification des différentes métriques. Screening Explorer peut donc être utilisé à des fins éducatives, particulièrement dans le cadre de travaux pratiques, tout en restant un outil complet qui permet une analyse précise des résultats de criblage virtuel.

Le second module de Screening Explorer permet d'utiliser des méthodes consensus (voir point 2.4.6.3.4) pour évaluer leurs performances dans le reclassement des composés (***Figure 75***). Bien que l'usage de méthodes consensus soit désormais fréquent dans le processus de développement de nouveaux médicaments, leur application ne conduit pas automatiquement à une amélioration des résultats VHTS. Nous avons donc implémenté quelques méthodes consensus simples dans Screening Explorer, pour permettre l'évaluation des performances des méthodes consensus sur les données fournies par l'utilisateur.

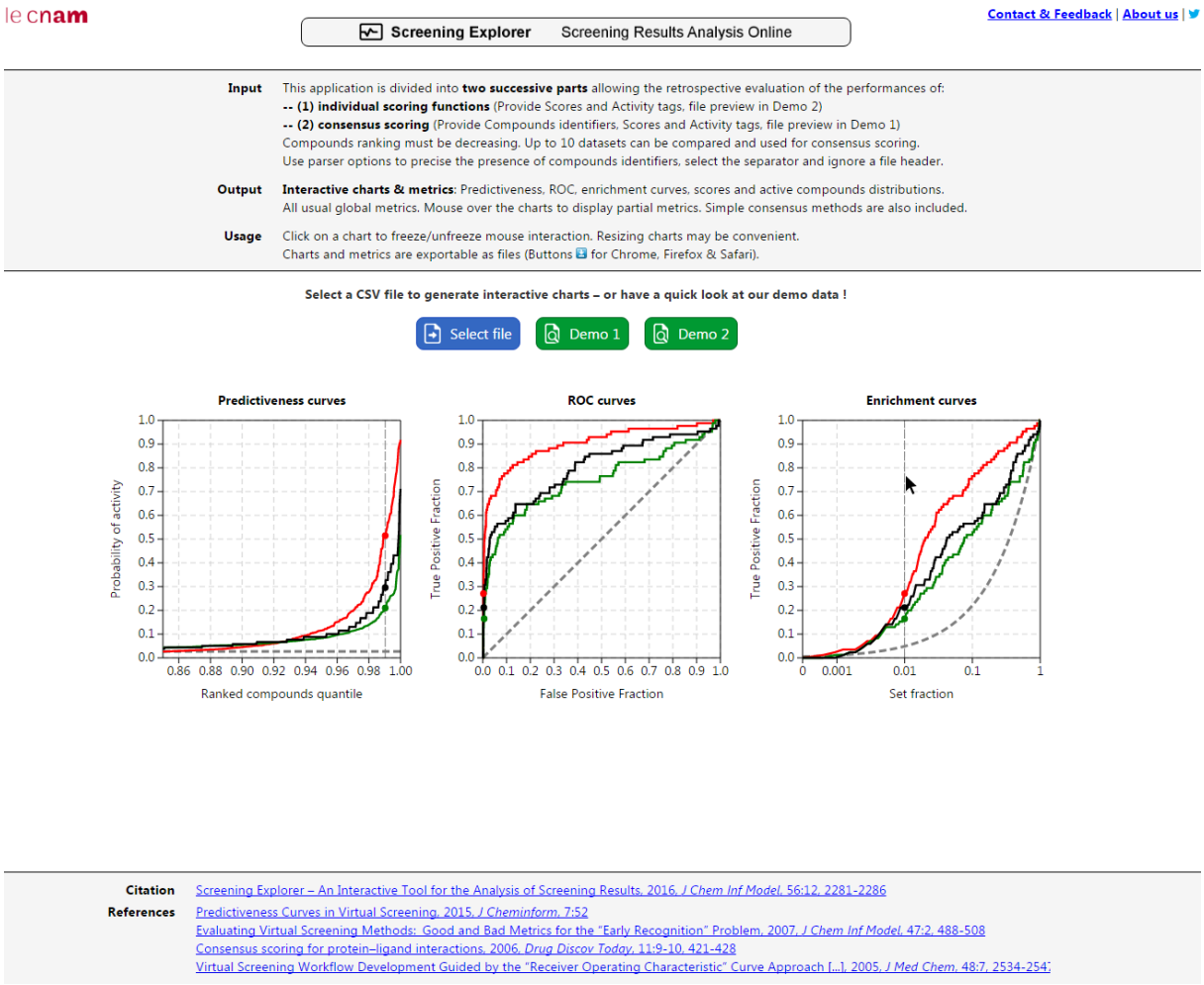


Figure 73. *Capture d'écran de la page d'accueil de Screening Explorer. L'utilisateur peut entrer ses données (bouton bleu) ou essayer l'outil grâce à deux jeux de données de démonstration (boutons verts). Ici, les graphiques sont figés et présentés à titre illustratif.*

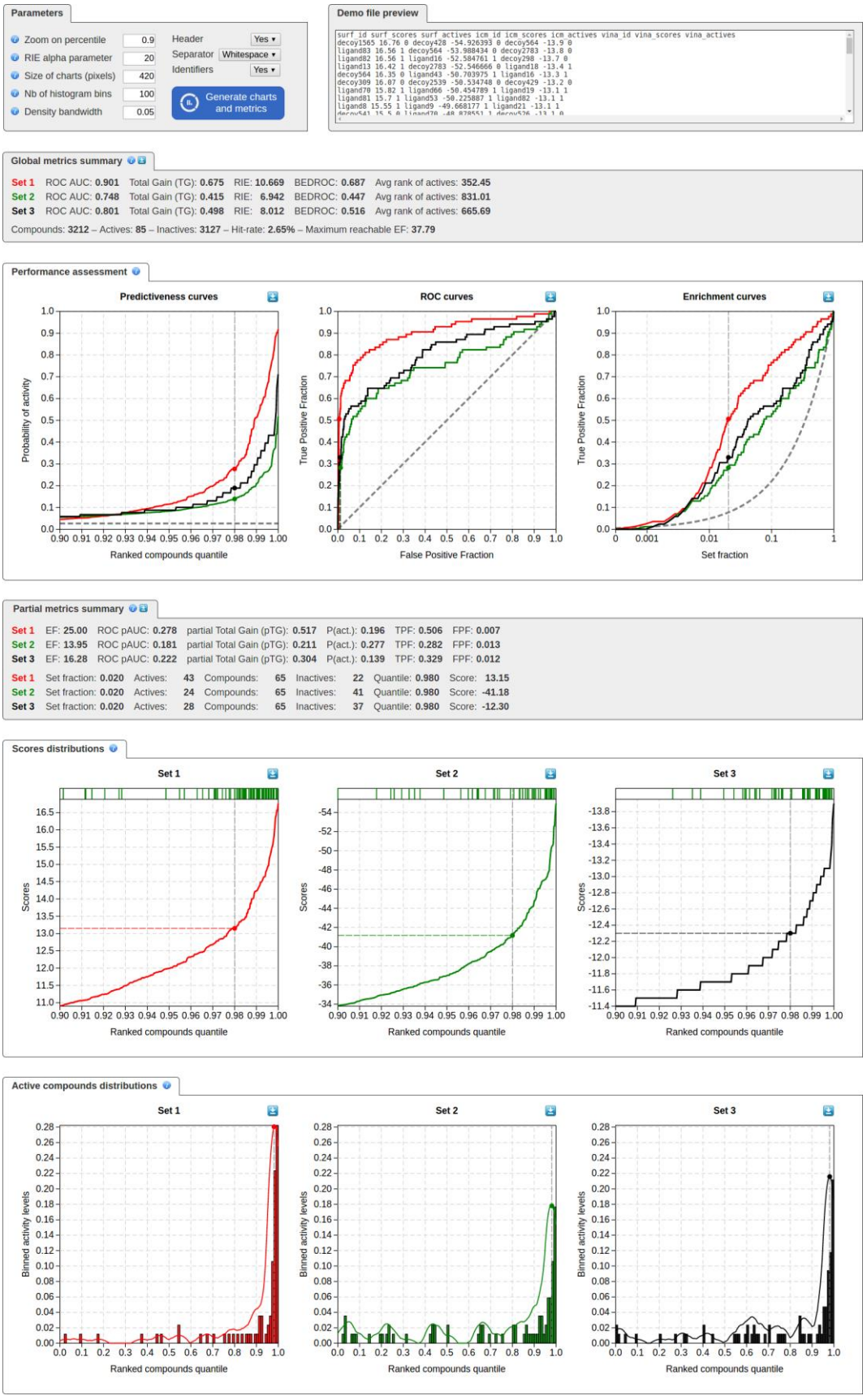


Figure 74. *Capture d'écran des outils d'analyse des résultats de criblage virtuel : courbes de prédictivité, de ROC, d'enrichissement, distributions des scores et des ligands. Les résultats de trois méthodes sont représentés en sélectionnant la fraction à 2% des jeux de données classés.*

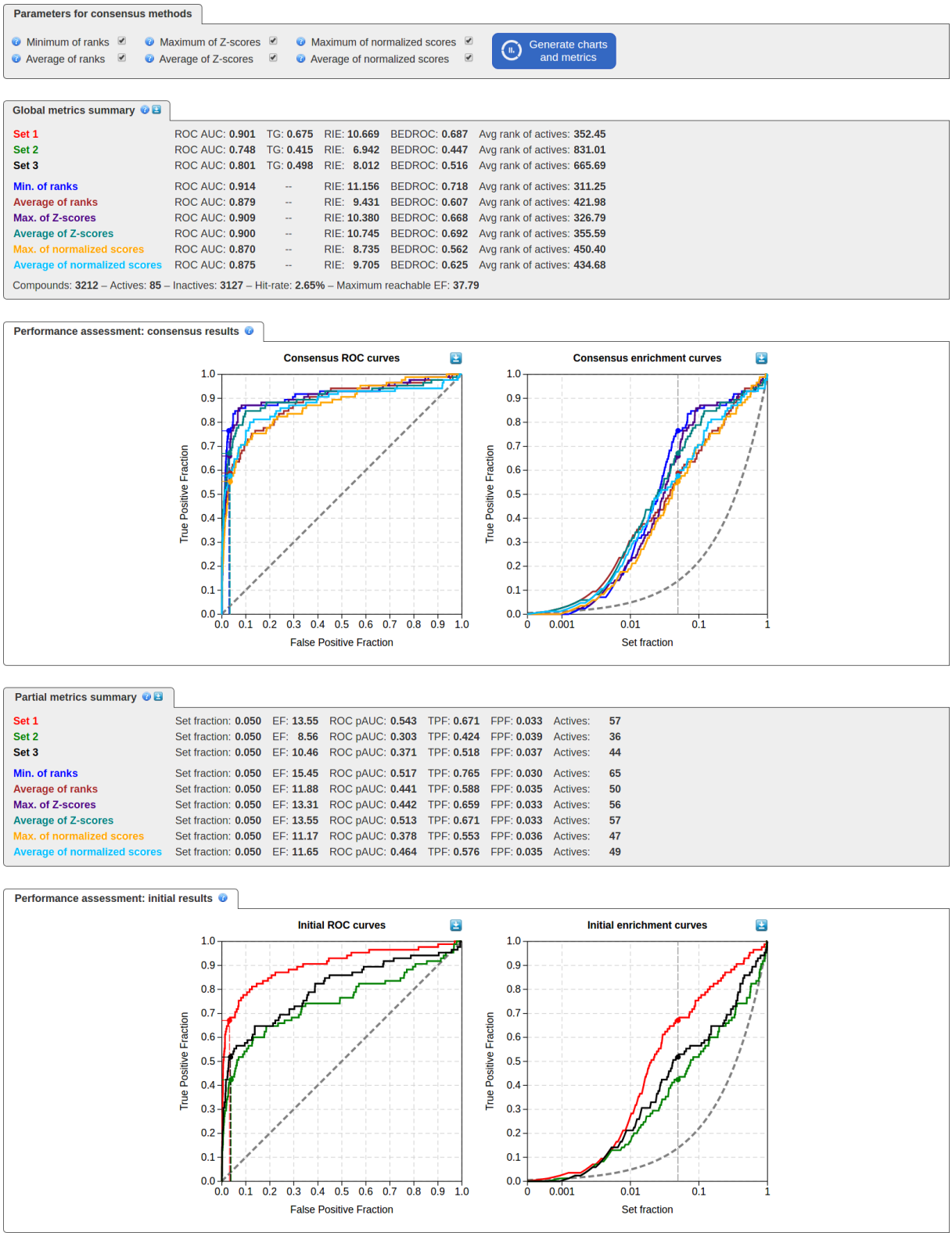


Figure 75. *Capture d'écran des outils d'analyse des résultats de méthodes consensus simples : courbes de ROC et d'enrichissement. Les résultats de six méthodes consensus appliquées aux classements obtenus par trois méthodes de criblage virtuel sont représentés et comparés en sélectionnant la fraction à 5% des jeux de données classés.*

JOURNAL OF CHEMICAL INFORMATION AND MODELING

Application Note

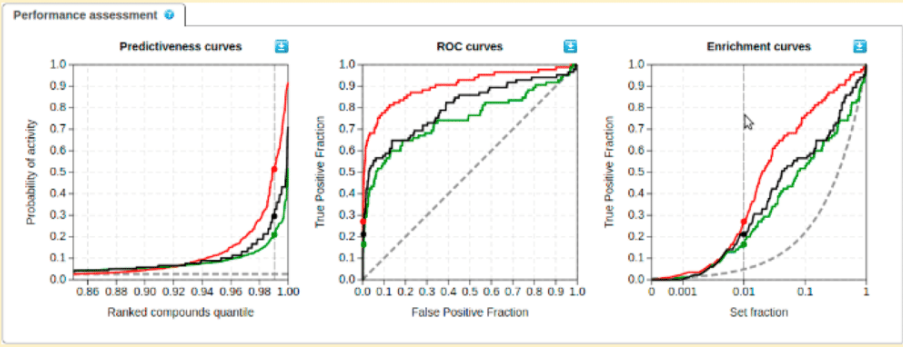
pubs.acs.org/jcim

Screening Explorer−An Interactive Tool for the Analysis of Screening Results

Charly Empereur-Mot, Jean-François Zagury, and Matthieu Montes*

Laboratoire Génomique Bioinformatique et Applications, EA 4627, Conservatoire National des Arts et Métiers, 292 rue Saint Martin, 75003 Paris, France

Supporting Information



ABSTRACT: Screening Explorer is a web-based application that allows for an intuitive evaluation of the results of screening experiments using complementary metrics in the field. The usual evaluation of screening results implies the separate generation and apprehension of the ROC, predictiveness, and enrichment curves and their global metrics. Similarly, partial metrics need to be calculated repeatedly for different fractions of a data set and there exists no handy tool that allows reading partial metrics simultaneously on different charts. For a deeper understanding of the results of screening experiments, we rendered their analysis straightforward by linking these metrics interactively in an interactive usable web-based application. We also implemented simple consensus scoring methods based on scores normalization, standardization (z-scores), and compounds ranking to evaluate the enrichments that can be expected through methods combination. Two demonstration data sets allow the users to easily apprehend the functions of this tool that can be applied to the analysis of virtual and experimental screening results. Screening Explorer is freely accessible at http://stats.drugdesign.fr.

INTRODUCTION

In the last two decades, virtual screening of compound collections has become extensively used in drug discovery programs to reduce the number of compounds going into high throughput screening procedures.[1] The aim of virtual screening methods is to enrich a subset of molecules in potentially active compounds while discarding the compounds supposed to be inactive according to a scoring function.[1,2] However, scoring functions are still imperfect and their performances can vary largely depending on the conditions of the experiment (i.e., for structure-based experiments, the target of interest, the preparation of the binding site, the addition of water molecules to the pocket, and the parameters of the conformational search algorithm or, more generally, the characteristics of the molecular data set and the parameters of the scoring function).[3,4] Therefore, it is important to perform benchmarking studies to (1) calibrate the parameters of virtual screening experiments for specific systems and (2) enable an efficient selection of the compounds to be tested experimentally in drug discovery programs, in the light of previously understood results.[5,6]

Different metrics have emerged to evaluate the performances of virtual screening methods retrospectively. Among them, the receiver operating characteristics (ROC) curves,[2] enrichment curves, and the recently introduced Predictiveness Curves (PC)[7] provide intuitive visualizations and cover different aspects of the results. The ROC metric is a chart of the true positive fraction (TPF, or active compounds fraction, y-axis) versus the false positive fraction (FPF, or inactive compounds fraction, x-axis) for each compound in an ordered data set. Each point of the ROC curve then represents a unique FPF/TPF pair corresponding to a particular fraction of the molecular data set, which allows to estimate the overall success of a screening method in the discrimination of active compounds over inactive compounds.[2] Enrichment curves allow to quantify the early recognition of active compounds by visualizing the TPF (y-axis) for each fraction of the ordered data set on a logarithmic scale (x-axis).[6] The PC quantifies the discrimination of active compounds attributable to the variations of the scores issued by

Received: May 20, 2016
Published: November 3, 2016

DOI: 10.1021/acs.jcim.6b00283
J. Chem. Inf. Model. 2016, 56, 2281−2286

a scoring function. A generalized linear model is built to issue probabilities of activity (y-axis) for each compound of the ordered data set (x-axis), as a function of the scores and the rank of the active compounds.[7] Consequently, the PC allows to detect potential score gaps and variations issued by a scoring function in the detection of the actives, which correspond to gaps in probabilities of activity.[7]

Global measures summarize each of these aspects of the results. The widely used area under the ROC curve (ROC AUC) accounts for the overall discrimination of the actives while the Boltzmann enhanced discrimination of ROC (BEDROC)[6] and the robust initial enhancement (RIE)[8] aim at quantifying their early recognition. RIE and BEDROC metrics emphasize early recognition using a continuously decreasing exponential weight as a function of the active compounds ranks. The standardized total gain ($\overline{TG}$),[9] calculated from the PC, summarizes the discrimination of active compounds imputable to the variation of the scores over a complete molecular data set.[7] The integral of the probabilities of activity is calculated over the random picking of compounds and normalized, bounding its value between 0 and 1. $\overline{TG}$ values over 0.25 combined with a ROC AUC over 0.5 generally signify that score variations are relevant in the discrimination of the actives.[7] High $\overline{TG}$ values (over 0.4) combined with a ROC AUC over 0.5 therefore indicate that the screening method performed well and that its performance would be reproducible in similar experimental conditions. Additionally, partial metrics aim at quantifying the enhancement of activity over a limited, early part of the ordered molecular data set. Enrichment factors (EF) and the partial area under the ROC curve (pAUC)[10] allow for the evaluation of the performances of a scoring function within a given fractional threshold. The partial total gain (pTG)[11] quantifies the enhancement of activity related to score variations within a given fractional threshold.[7]

Several programs and R packages are available for the analysis of virtual screening results, principally: ROCR,[12] Bioconductor ROC,[13] pROC,[14] and enrichvs[15] They all allow generation of ROC curves and calculation of AUCs. Bioconductor ROC and pROC also allow calculation of pAUCs, whereas enrichvs brings the calculation of the RIE, BEDROC, and EFs. PredictABEL[16] allows generation of PCs, and pseval[17] can be used to calculate $\overline{TG}$s and pTGs, recently introduced for the analysis of virtual screening results.[7] Thus, all metrics are not easily accessible to the users, who have to generate metrics separately and calculate partial metrics repeatedly for each fractional threshold of interest. There exists no handy tool that regroup these metrics for the task of evaluating virtual screening performances retrospectively.

Therefore, we developed a usable web-based interactive application that covers all the aspects of the analysis of the results and allows to calculate partial metrics on the fly for specific fractions of the data sets: Screening Explorer (http://stats.drugdesign.fr). Our primary interest in developing this program is to provide an intuitive tool to facilitate the apprehension of the results of virtual and experimental screening experiments retrospectively, which should support the development of scoring functions and help in the process of selecting active compounds in drug discovery programs. We also focused on making the application user-friendly and suitable for educational purposes, notably by providing tooltip explanations on the content of each chart and the signification of the different metrics.

Additionally to the first part of the program oriented toward the analysis of the output of individual scoring functions, we implemented several consensus scoring methods based on scores normalization, standardization (or z-scores), and compounds ranking.[18] Consensus scoring methods aim at combining the results of different scoring functions to improve the final enrichment and diminish the importance of the choice of a particular scoring function.[19] Their application has now become common in drug discovery and the majority of published work in the field of consensus scoring reports better enrichments compared to the use of single scoring functions. However, it appears that each of the combined scoring function must yield reasonable results to enhance the retrieval of active compounds through consensus scoring. A blind combination of some arbitrarily chosen scoring functions will not necessarily lead to better results.[18,20,21] Therefore, we included simple consensus methods to Screening Explorer to facilitate the apprehension of consensus scoring results as well as the output of individual scoring functions.

USAGE

Input. This application is divided into two successive parts: the evaluation of the performances of (1) individual scoring functions and (2) consensus scoring of these functions. Therefore, two types of input are accepted as column text files to proceed to 1 only and both 1 and 2.

For the evaluation of the performances of individual scoring functions (1) the required input takes the following form, using only screening scores and activity tags (AT) (1 or 0: active or inactive):

set1_score1	set1_AT1	set2_score1	set2_AT1	[...]
set1_score2	set1_AT2	set2_score2	set2_AT2	[...]
set1_score3	set1_AT3	set2_score3	set2_AT3	[...]

Up to 10 data sets can be compared and thereafter used for consensus scoring. To proceed to the evaluation of the performances of individual scoring functions (1) and consensus scoring (2), compounds identifiers (CI), screening scores, and activity tags are required under the following format:

set1_CI1	set1_score1	set1_AT1	set2_CI1	set2_score1	set2_AT1	[...]
set1_CI2	set1_score2	set1_AT2	set2_CI2	set2_score2	set2_AT2	[...]
set1_CI3	set1_score3	set1_AT3	set2_CI3	set2_score3	set2_AT3	[...]

We included two demonstration cases to the application to ease the apprehension of its features and enable the preview of its two possible input types.

Parameters. Five parameters modulate the display of the charts and metrics: 1. "Zoom on percentile" allows to focus on an early part of the data sets in the predictiveness and scores distributions charts. 2. "Alpha parameter" allows to choose the alpha parameter used for the RIE and BEDROC calculations. 3. "Size of charts" allows to scale all the charts at the convenience of the user for better readability. 4. "Nb of histogram bins" allows to choose the number of bins to display in the histograms of the active compounds distributions. 5. "Density bandwidth" allows to choose the bandwidth of the kernel density estimation superposed to the histograms of the active compounds distributions (kernel: Epanechnikov). Tooltips define the parameter ranges for each of them. Parser options allow the user to define the presence or absence of compounds

DOI: 10.1021/acs.jcim.6b00283
J. Chem. Inf. Model. 2016, 56, 2281−2286

identifiers in the input file ("Identifiers"), select the column separator ("Separator"), and ignore a file header ("Header").

MATERIALS AND METHODS

Metrics Calculation. Though the RIE and BEDROC were originally calculated through a Monte Carlo simulation, we adopted the simplified formulation proposed by Truchon et al. to reduce computational costs.[6] The RIE uses a decreasing exponential weight as a function of the actives ranks. Its denominator corresponds to the average sum of the exponential when n actives are uniformly distributed in the ordered data set containing N compounds, x_i is the relative scaled rank, and α controls the extent of the weight importance as a parameter.

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

The RIE minimum and maximum values can be calculated as follows, R_a being the ratio of actives in the data set:

$$\mathrm{RIE}_{\min} = \frac{1-e^{\alpha R_a}}{R_a(1-e^{\alpha})}$$

$$\mathrm{RIE}_{\max} = \frac{1-e^{-\alpha R_a}}{R_a(1-e^{-\alpha})}$$

The BEDROC metric is then a normalization of the RIE, bounded by 0 and 1:

$$\mathrm{BEDROC} = \frac{\mathrm{RIE}-\mathrm{RIE}_{\min}}{\mathrm{RIE}_{\max}-\mathrm{RIE}_{\min}}$$

The PC is calculated using a generalized linear model (binomial distribution function and canonical log link) to issue probabilities of activity from the scores and actives ranks.[7] We implemented the PC using the glm.js Javascript library,[22] which yields the same results as the R glm function.[23] The standardized total gain is calculated from the probabilities of activity $R(v)$ at each vth quantile over the prevalence of activity $p = R_a$ and normalized by its maximum value. The $\overline{\mathrm{TG}}$ quantifies the discrimination of actives over inactives attributable to score variations, bounded by 0 and 1.

$$\overline{\mathrm{TG}}(v) = \frac{\int_0^1 |R(v)-p|\,\mathrm{d}v}{2p(1-p)}$$

Consensus Calculation. For the methods based on the maximum and average normalized score, and the minimum and average rank, potential equalities are split using the maximum and average z-scores from the individual scoring functions, respectively. Similarly, for the methods based on z-scores, equalities are split using the ranks. For scoring functions that rank compounds from positive to negative scores, z-scores are negated before proceeding to consensus calculations. Z-scores are calculated as follows, where x is the raw score from the given scoring function, μ is the average, and σ is the standard deviation of the raw scores.

$$z = \frac{x-\mu}{\sigma}$$

Virtual Screening Methods. Surflex-dock is based on a modified Hammerhead fragmentation−reconstruction algorithm to dock compound flexibly into the binding site.[24] The query molecule is decomposed into rigid fragments that are superimposed to the Surflex protomol (i.e., molecular fragments covering the entire binding site). The binding sites were extended to 4 Å around the cocrystallized ligand from the Directory of Useful Decoys (DUD)[25] for the protomol generation step. ICM is based on Monte Carlo simulations to optimize the position of molecules using a stochastic global optimization procedure.[26] Autodock Vina (referred afterward as Vina) generates docking poses using an iterated local search global optimizer, which consists in a succession of steps of stochastic mutations and local optimizations.[27] The binding sites used with ICM and Vina were adjusted to be similar to the Surflex protomol. Surflex-dock version 2.5, ICM version 3.6, and Vina version 1.1.2 were used for this study.

Others. The d3.js Javascript library was used for charts implementation.[28] Since the calculations are performed in the browser client-side, input files will not be uploaded to any server.

RESULTS AND DISCUSSION

Interactive Charts and Performance Metrics. The performances of individual scoring functions can be visually compared and summarized in three charts: the PC, ROC curves, and enrichment curves. They respectively allow the evaluation of the discrimination power of the scores, the overall success of a method in the ranking of active compounds, and the enrichment obtained in the early part of a data set.[2,6,7] Additionally, scores distributions charts that include a rugplot to locate the ranks of the actives allow to preview score variations related to the detection of the actives, although this information is already compiled and summarized by the predictiveness charts.[7] Histograms of the active compounds distributions allow to overview their ranking and include a kernel density estimation (kernel: Epanechnikov).

Global metrics are calculated along with charts. Partial metrics will be displayed on mouse-over on charts, which can be frozen by clicking on a chart to set a threshold on a fraction of interest of a data set. When browsing a chart, the current threshold will be displayed on all of the other charts, allowing the visualization where a fraction of a data set lies from one metric to another. The predictiveness, enrichment, scores distributions, and active compound distribution charts can be browsed by fraction of the data set. The ROC curves can be browsed either by TPF or FPF. Charts and metrics can be exported as image and text files.

Consensus Scoring Methods. Since individual scoring functions usually describe binding affinity as a sum of independent terms, they all suffer from the same disadvantages: dependence to molecular weight (the larger the molecule, the higher the probability of a favorable score), they may ignore entropic effects (rigid receptor and single ligand binding modes), and some do not take into account solvation and desolvation effects.[21] In an ideal consensus scoring experiment, each scoring function should both yield reasonable results and contribute something unique to the consensus to obtain an improvement in the detection of active compounds. Therefore, it is important to assess the complementarity of individual scoring functions to take the advantages and balance the deficiencies of different scoring functions when using consensus scoring.

Because different scoring functions usually have different scales, their combination is not trivial. To evaluate the results of consensus scoring, we implemented methods based on (1)

DOI: 10.1021/acs.jcim.6b00283
J. Chem. Inf. Model. 2016, 56, 2281−2286

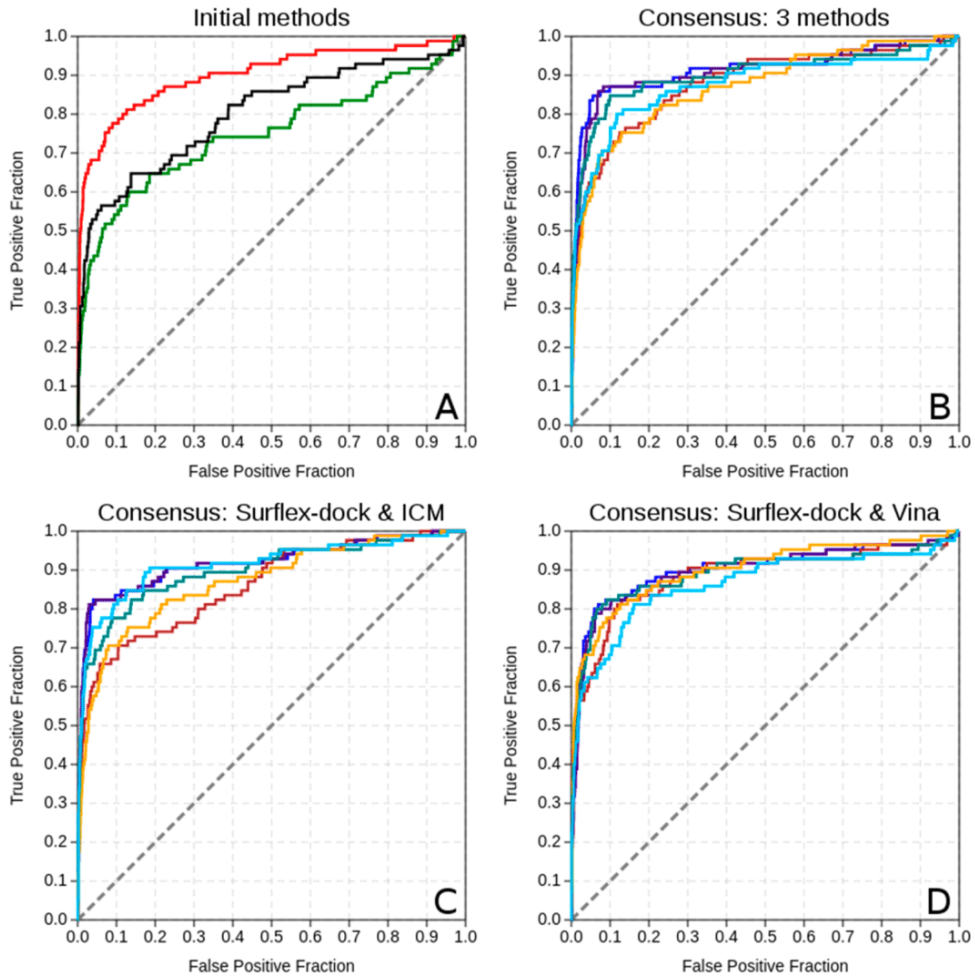


Figure 1. ROC curves of the following experiments on the PPAR gamma: (A) Surflex-dock (red), ICM (green), and Vina (black). (B) Consensus scoring with the three methods. (C) Consensus scoring with Surflex-dock and ICM. (D) Consensus scoring with Surflex-dock and Vina. Consensus methods: minimum of ranks (blue), average of ranks (dark red), maximum of z-scores (purple), average of z-scores (dark green), maximum of normalized scores (orange), average of normalized scores (cyan).

Table 1. ROC AUC and BEDROC Metrics of the Score Combinations of Surflex-dock, ICM, and Vina on the PPAR Gamma of the DUD Data Set Using Consensus Methods Based on Ranks, Score Standardization, and Normalization

	three methods		Surflex-dock, ICM		ICM, Vina		Surflex-dock, Vina	
	AUC	BEDROC	AUC	BEDROC	AUC	BEDROC	AUC	BEDROC
minimum of ranks	0.914	0.718	0.918	**0.722**	0.859	0.592	0.902	0.687
average of ranks	0.879	0.607	0.859	0.594	0.804	0.499	0.887	0.654
maximum of z-scores	0.909	0.668	0.919	**0.743**	0.857	0.559	0.896	0.646
average of z-scores	0.900	0.692	0.899	0.673	0.826	0.557	0.893	0.691
maximum of normalized scores	0.870	0.562	0.870	0.564	0.748	0.446	0.900	0.679
average of normalized scores	0.875	0.625	0.914	0.706	0.819	0.535	0.864	0.614

score normalization, that allows to rescale the scores though it is sensible to the presence of outliers; (2) score standardization, that allows to detect extremely high values which should be associated with active compounds; and (3) rank-based methods, which have a strong inherent averaging effect that reduce the effect of outliers. For each method, maximum and average of normalized and standardized scores, and minimum and average of the compounds ranks have been implemented. The minimum of ranks, maximum of z-scores, and maximum of normalized scores are useful for the selection of the best scored actives from each individual scoring function, whereas averages of ranks, z-scores or normalized scores can balance the advantages and deficiencies of different scoring functions. ROC and enrichment charts allow to compare the refinement obtained with each selected method and calculate partial metrics interactively, additionally to the usual global metrics.

Case Study: Structure-Based Virtual Screening on the Peroxisome Proliferator Activated Receptor (PPAR) Gamma. We analyzed the results of retrospective structure-based virtual screenings on target PPAR gamma from the DUD data set[25] using Surflex-dock,[24] ICM,[26] and Vina[27] (data available as Demo1). Individual scoring functions were

DOI: 10.1021/acs.jcim.6b00283
J. Chem. Inf. Model. 2016, 56, 2281−2286

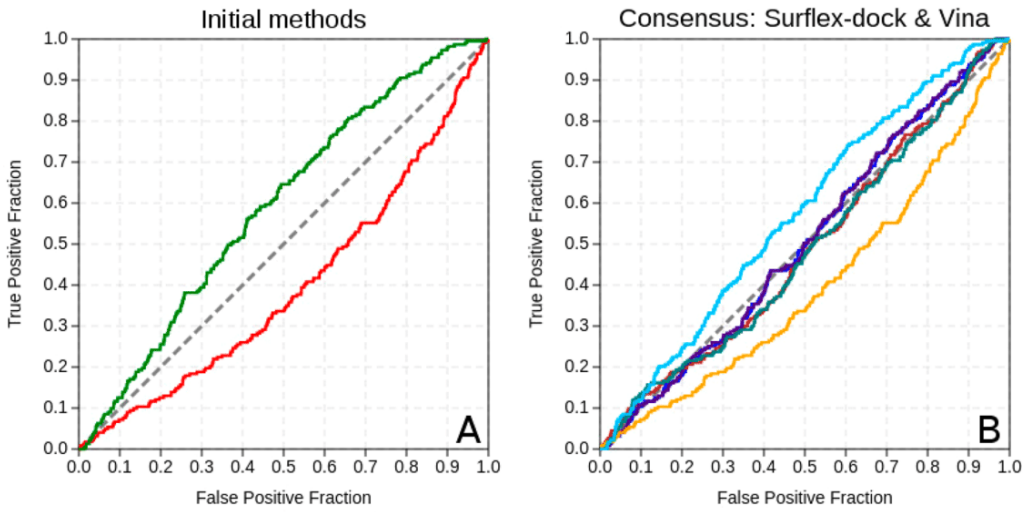


Figure 2. ROC curves of the following experiments on the thrombin alpha: (A) Surflex-dock (red) and Vina (green). (B) Consensus scoring with the two methods. Consensus methods: minimum of ranks (blue), average of ranks (dark red), maximum of z-scores (purple), average of z-scores (dark green), maximum of normalized scores (orange), and average of normalized scores (cyan).

successful in the discrimination of active compounds over inactives (Surflex-dock AUC 0.901, ICM AUC 0.748, Vina AUC 0.801) and the early recognition of active compounds was very satisfying, especially for Surflex-dock (Surflex-dock BEDROC 0.687, ICM BEDROC 0.447, Vina BEDROC 0.516). The three methods produced meaningful score variations in the detection of the actives (Surflex-dock $\overline{TG}$ 0.675, ICM $\overline{TG}$ 0.415, Vina $\overline{TG}$ 0.498).

The consensus study showed that the combination of the 3 methods using the minimum of ranks or the maximum of z-scores could increase the early recognition compared to the best performing individual scoring function: Surflex-dock, while averaging the 3 methods by the ranks, the z-scores or the normalized scores did not result in better enrichments (Figure 1). Interestingly, the combination of the methods by pairs (Table 1) indicated that Surflex-dock and ICM were the most complementary methods (Figure 1C). Even if Vina slightly outperformed ICM in the conditions of this experiment, the consensus combination of Surflex-dock and ICM by the minimum of ranks or maximum of z-scores resulted in an improvement of the early recognition of the active compounds (minimum of ranks TPF 5% 0.741, maximum of z-scores TPF 5%: 0.812, Surflex-dock TPF 5% 0.671).

This result reflects that Surflex-dock and Vina were successful in the early recognition of a similar part of the active compounds, whereas ICM could score different actives correctly. The minimum of ranks and maximum of z-scores allowed to pick the best scored actives from each individual method whereas the maximum of normalized scores did not perform as well, affected by outliers. Since the $\overline{TG}$ values of Surflex-dock and ICM were high, they are expected to yield good performances in further prospective screenings on target PPAR gamma under similar experimental conditions. In this case, the consensus combination of Surflex-dock and ICM by the minimum of ranks or maximum of z-scores should yield increased enrichments through the correct early recognition of different active compounds.

Case Study: Structure-Based Virtual Screening on the Androgen Receptor (AR). We analyzed the results of retrospective structure-based virtual screenings on target AR from the DUD data set[25] using Surflex-dock,[24] ICM,[26] and Vina[27] to illustrate how consensus selection can improve the final enrichments (Supporting Information).

Case Study: Structure-Based Virtual Screening on the Thrombin Alpha. In this case study, we used the data sets from two experimental assays to evaluate retrospectively the performances of structure-based virtual screening experiments on the Thrombin alpha. High-throughput screening assay (Pubchem[29] AID1046) included 217 250 compounds tested for inhibition of the thrombin alpha activity and yielded 557 compounds qualified as actives. A dose response confirmation assay (Pubchem[29] AID1215) followed using 529 of the previous 557 actives, validating 223 compounds as inhibitors. We performed virtual screening experiments using Surflex-dock[24] and Vina[27] with this molecular data set of 217 250 compounds including 223 actives (actives rate 0.10%) on the thrombin alpha structure from the DUD (Figure 2).[25]

For this large data set the generation of the metrics for individual methods evaluation took approximately 30 s for Screening Explorer in the web-browser Chrome. Consensus scoring and associated metrics were generated in 5 min. On this HTS assay that provided few active compounds (0.1% of the data set), Vina and Surflex-dock displayed under-average performances (Surflex-dock AUC 0.397, Vina AUC 0.592). Both methods failed to achieve early recognition (Surflex-dock BEDROC 0.036, Vina BEDROC 0.058). Consensus scoring could not, in this context, increase the performance of the results.

■ CONCLUSION

We described the functions of an interactive and usable web-based application that covers different aspects of the analysis of the results of screening experiments: Screening Explorer (http://stats.drugdesign.fr). We believe that the interactive evaluation of screening results using complementary metrics in the field allows for a more complete interpretation of their performances, which should enhance our understanding of the remaining weaknesses of scoring functions and help in their future developments. Furthermore, the study of consensus scoring coupled to the assessment of the performance of each

DOI: 10.1021/acs.jcim.6b00283
J. Chem. Inf. Model. 2016, 56, 2281−2286

scoring function allows to evaluate the reproducibility of the results of screening experiments from benchmarking to prospective assays, as illustrated in the two case studies. We also tried to make the comprehension of the metrics and their results more accessible to nonexpert users and to students. Although this program was initially designed for the interactive analysis of the results of virtual screening methods retrospectively, it can be used equally for the calibration of scoring functions with respect to experimental assays.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.6b00283.

Retrospective structure-based virtual screenings on target AR from the DUD data set[25] using Surflex-dock,[24] ICM,[26] and Vina[27] (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: matthieu.montes@cnam.fr (M.M.).

ORCID

Matthieu Montes: 0000-0001-5921-460X

Author Contributions

C.E-M. implemented the program. C.E-M. and M.M. analyzed the data. The manuscript was written by C.E-M. with contributions from all authors. All authors read and approved the final manuscript.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

C.E-M. is recipient of a MNESR fellowship. We thank Prof. Jain for providing Surflex and Molsoft for providing an academic license for ICM.

■ REFERENCES

(1) Alvarez, J. C. High-throughput Docking as a Source of Novel Drug Leads. *Curr. Opin. Chem. Biol.* **2004**, *8*, 365−370.

(2) 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*, 2534−2547.

(3) Ben Nasr, N.; Guillemain, H.; Lagarde, N.; Zagury, J. F.; Montes, M. Multiple Structures for Virtual Ligand Screening: Defining Binding Site Properties-based Criteria to Optimize the Selection of the Query. *J. Chem. Inf. Model.* **2013**, *53*, 293−311.

(4) Jain, A. N. Bias, Reporting, and Sharing: Computational Evaluations of Docking Methods. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 201−212.

(5) Nicholls, A. What do we Know and When do we Know it? *J. Comput.-Aided Mol. Des.* **2008**, *22*, 239−255.

(6) Truchon, J.; Bayly, C. Evaluating Virtual Screening Methods: Good and Bad Metrics for the "Early Recognition" Problem. *J. Chem. Inf. Model.* **2007**, *47*, 488−508.

(7) Empereur-Mot, C.; Guillemain, H.; Latouche, A.; Zagury, J. F.; Viallon, V.; Montes, M. Predictiveness Curves in Virtual Screening. *J. Cheminf.* **2015**, *7*, 1−17.

(8) 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. Model.* **2001**, *41*, 1395−1406.

(9) Bura, E.; Gastwirth, J. L. The Binary Regression Quantile Plot: Assessing the Importance of Predictors in Binary Regression Visually. *Biom. J.* **2001**, *43*, 5−21.

(10) McClish, D. K. Analyzing a Portion of the ROC Curve. *Med. Decis. Making.* **1989**, *9*, 190−195.

(11) Sachs, M. C.; Zhou, X. H. Partial Summary Measures of the Predictiveness Curve. *Biom. J.* **2013**, *55*, 589−602.

(12) Sing, T.; Sander, O.; Beerenwinkel, N.; Lengauer, T. ROCR: visualizing classifier performance in R. *Bioinformatics* **2005**, *21*, 3940−3941.

(13) Gentleman, R. C.; Carey, V. J.; Bates, D. M.; Bolstad, B.; Dettling, M.; Dudoit, S.; Ellis, B.; Gautier, L.; Ge, Y.; Gentry, J.; Hornik, K.; Hothorn, T.; Huber, W.; Iacus, S.; Irizarry, R.; Leisch, F.; Li, C.; Maechler, M.; Rossini, A. J.; Sawitzki, G.; Smith, C.; Smyth, G.; Tierney, L.; Yang, J. Y.; Zhang, J. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **2004**, *5*, R80.

(14) Robin, X.; Turck, N.; Hainard, A.; Tiberti, N.; Lisacek, F.; Sanchez, J. C.; Müller, M. pROC: an open-source package for R and S + to analyze and compare ROC curves. *BMC Bioinf.* **2011**, *12*, 77−84.

(15) Hiroaki, Y. Package "enrichvs". 2015.

(16) Kundu, S.; Aulchenko, Y. S.; van Duijn, C. M.; Janssens, A. C. PredictABEL: an R package for the assessment of risk prediction models. *Eur. J. Epidemiol.* **2011**, *26*, 261−264.

(17) Sachs, M. C.; Gabriel, E. E. pseval: Methods for Evaluating Principal Surrogates of Treatment Response. https://cran.r-project.org/web/packages/pseval/index.html (accessed Nov 2016).

(18) Feher, M. Consensus Scoring for Protein-ligand Interactions. *Drug Discovery Today* **2006**, *11*, 421−428.

(19) Jorgensen, W. L. The Many Roles of Computation in Drug Discovery. *Science* **2004**, *303*, 1813−1818.

(20) Du, J.; Bleylevens, I. W.; Bitorina, A. V.; Wichapong, K.; Nicolaes, G. A. Optimization of Compound Ranking for Structure-based Virtual Ligand Screening Using an Established FRED-surflex Consensus Approach. *Chem. Biol. Drug Des.* **2014**, *83*, 37−51.

(21) Schulz-Gasch, T.; Stahl, M. Scoring Functions for Protein-ligand Interactions: a Critical Perspective. *Drug Discovery Today: Technol.* **2004**, *1*, 231−239.

(22) Generalized Linear Models with Javascript. http://github.com/rallysf/glm (accessed Aug 10, 2016).

(23) R Core Team. *RA Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, 2016.

(24) Jain, A. N. Surflex: Fully Automatic Flexible Molecular Docking Using a Molecular Similarity-based Search Engine. *J. Med. Chem.* **2003**, *46*, 499−511.

(25) Huang, N.; Shoichet, B.; Irwin, J. Benchmarking Sets for Molecular Docking. *J. Med. Chem.* **2006**, *49*, 6789−6801.

(26) Abagyan, R.; Totrov, M.; Kuznetsov, D. ICM − A New Method for Protein Modeling and Design: Applications to Docking and Structure Prediction from the Distorted Native Conformation. *J. Comput. Chem.* **1994**, *15*, 488−506.

(27) Trott, O.; Olson, A. J. AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading. *J. Comput. Chem.* **2009**, *31*, 455−461.

(28) Data-Driven Documents. https://d3js.org (accessed Aug 10, 2016).

(29) Wang, Y.; Suzek, T.; Zhang, J.; Wang, J.; He, S.; Cheng, T.; Shoemaker, B. A.; Gindulyte, A.; Bryant, S. H. PubChem BioAssay: 2014 update. *Nucleic Acids Res.* **2014**, *42*, 1075−1082.

Supporting Information

Screening Explorer – An Interactive Tool for the Analysis of Screening Results

Charly Empereur-Mot[1], Jean-François Zagury[1], Matthieu Montes[1§]

[1]Laboratoire Génomique Bioinformatique et Applications, EA 4627, Conservatoire National des Arts et Métiers, 292 rue Saint Martin, 75003 Paris, France

[§]Corresponding author

Case study: Structure-based virtual screening on the Androgen Receptor (AR)

We analyzed the results of retrospective structure-based virtual screenings on target AR from the DUD dataset[1] using Surflex-dock,[2] ICM[3] and Vina.[4] The 3 methods performed above random in the discrimination of active compounds over inactives (Surflex-dock AUC: 0.684, ICM AUC: 0.691, Vina AUC: 0.745) and the early recognition of the actives was best achieved by Vina[4] (Surflex-dock BEDROC: 0.257, ICM BEDROC: 0.350, Vina BEDROC: 0.423) (Figure S1). The $\overline{TG}$s indicate that score variations issued by Vina[4] were the most relevant in the detection of active compounds, whereas Surflex-dock[2] was the least sensitive method (Surflex-dock $\overline{TG}$: 0.067, ICM $\overline{TG}$: 0.151, Vina $\overline{TG}$: 0.395).

The consensus study revealed that the combination of the 3 methods by averaging the ranks, the z-scores or the normalized scores increased both the discrimination and the early recognition of active compounds (Average of ranks BEDROC: 0.536, Average of z-scores BEDROC: 0.534, Average of normalized scores BEDROC: 0.536), while the use of the minimum of ranks, the maximum z-scores and normalized scores did not improve the enrichment compared to the best performing individual scoring function: Autodock Vina[4] (Minimum of ranks BEDROC: 0.355, Maximum of z-scores BEDROC: 0.402, Maximum of normalized scores BEDROC: 0.266) (Figure S1B). The combination of scoring functions by pairs did not yield considerably better enrichments than the combination of the 3 methods (Table S1).

In the conditions of this experiment, averaging the scores issued by Surflex-dock,[2] ICM[3] and Vina[4] could improve the final enrichments and eliminate the choice of a particular scoring function. However, since the $\overline{TG}$ of Surflex-dock[2] was very low it is not assured that its scoring function would yield satisfying enrichments in prospective screenings of similar conditions. ICM[3] and Vina[4] achieved early recognition but the minimum of ranks, maximum of normalized or z-scores could not improve the final enrichments compared to the best performing scoring

function, highlighting that ICM[3] and Vina[4] scored similar compounds in their early fraction. Averaging the results of ICM[3] and Vina[4] still improved the enrichments, which means these scoring functions performed well in the calculation of different components of the binding affinity, thus proving complementarity. For further prospective screenings on target AR, the combination of ICM[3] and Vina[4] by averaging their ranks should yield the most robust results and the best final enrichments. Averaging standardized scores, normalized scores or using other consensus methods could also be considered to improve the results.

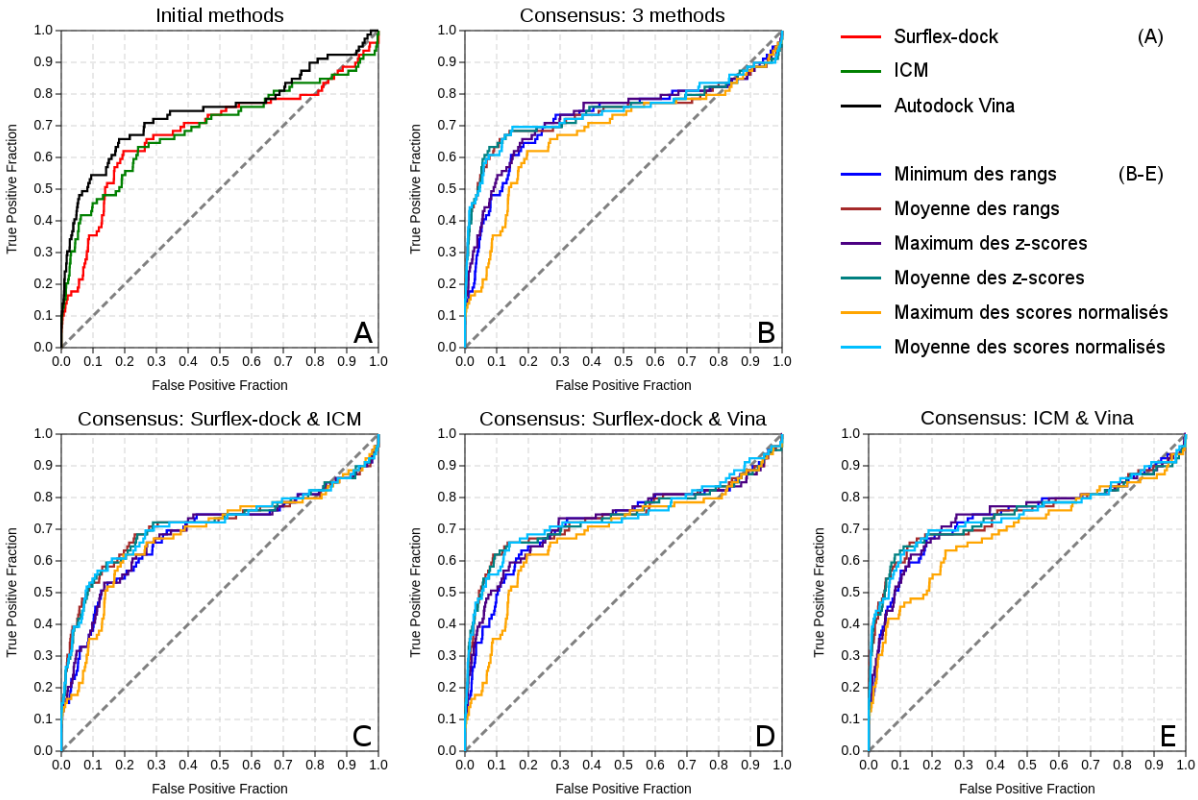


Figure S1. ROC curves of the following experiments on the AR: (A) Surflex-dock (red), ICM (green) and Vina (black); (B) Consensus scoring with the 3 methods; (C) Consensus scoring with Surflex-dock and ICM; (D) Consensus scoring with Surflex-dock and Vina; (E) Consensus scoring with ICM and Vina. Consensus methods: Minimum of ranks (blue), Average of ranks (dark red), Maximum of z-scores (purple), Average of z-scores (dark green), Maximum of normalized scores (orange), Average of normalized scores (cyan).

	3 methods		Surflex-dock, ICM		ICM, Vina		Surflex-dock, Vina	
	AUC	BEDROC	AUC	BEDROC	AUC	BEDROC	AUC	BEDROC
Minimum of ranks	0.725	0.355	0.684	0.303	0.739	0.396	0.722	0.363
Average of ranks	0.744	**0.536**	0.713	0.419	0.747	**0.519**	0.740	0.480
Maximum of z-scores	0.733	0.402	0.688	0.314	0.742	0.409	0.728	0.412
Average of z-scores	0.746	**0.534**	0.713	0.408	0.747	**0.518**	0.738	0.491
Maximum of normalized scores	0.681	0.266	0.681	0.266	0.691	0.350	0.684	0.258
Average of normalized scores	0.747	**0.536**	0.713	0.412	0.747	**0.516**	0.740	0.483

Table S1. ROC AUC and BEDROC metrics of the scores combinations of Surflex-dock, ICM and Vina on the AR of the DUD dataset using consensus methods based on ranks, scores standardization and normalization.

References

(1) Huang N. ; Shoichet B. ; Irwin J. Benchmarking Sets for Molecular Docking. *J. Med. Chem.* **2006**, *49*, 6789-6801.

(2) Jain A.N. Surflex : Fully Automatic Flexible Molecular Docking Using a Molecular Similarity-based Search Engine. *J. Med. Chem.* **2003**, *46*, 499-511.

(3) Abagyan R. ; Totrov M. ; Kuznetsov D. ICM – A New Method for Protein Modeling and Design : Applications to Docking and Structure Prediction from the Distorted Native Conformation. *J. Comput. Chem.* **1994**, *15*, 488-506.

(4) Trott O. ; Olson A.J. AutoDock Vina : Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading. *J. Comput. Chem.* **2010**, *31*, 455-461.

1.2.2. Discussion

Depuis sa publication sous forme d'article ASAP dans *Journal of Chemical Information and Modeling* le 3 novembre 2016, Screening Explorer a reçu 562 visiteurs uniques réels, soit 3.23 visites par jour (au 25 avril 2017). En moyenne, les visiteurs restent 118 secondes sur cet outil, ce qui est cohérent avec le temps nécessaire à l'analyse d'un résultat de criblage virtuel grâce aux outils statistiques mis à disposition. A ce jour, deux sessions de travaux pratiques ont été réalisées avec Screening Explorer pendant lesquelles les étudiants utilisaient l'outil pendant environ 15 minutes. La part des visiteurs récurrents ayant tendance à augmenter, il semble que le public ciblé soit satisfait de cet outil et nous espérons que sa fréquentation augmentera dans le futur.

1.2.2.1. Cas d'application et reclassements consensus

Dans cet article, nous avons décrit les fonctions de Screening Explorer à travers l'analyse de trois expériences de criblage virtuel « structure-based ». Les deux premiers cas d'application ont été réalisés sur des données de la banque DUD[159] (PPAR gamma et AR), tandis que le troisième cas évalue la capacité des logiciels à reproduire des résultats de HTS expérimentaux (Thrombine alpha, Pubchem AID 1046).[473] Pour les cibles PPAR gamma et AR, trois méthodes de criblage ont été appliquées : Surflex-dock,[349,350] ICM[359] et Autodock Vina.[457] Le jeu de données de la Thrombine alpha comportant un plus grand nombre de composés, nous avons utilisé uniquement Surflex-dock[349,350] et Autodock Vina[457] sur cette cible afin de réduire les temps de calcul associés. Cette étude est toutefois restée peu concluante, puisque les logiciels n'arrivaient pas à classer correctement les composés actifs dans les conditions de cette expérience. Comme cas complémentaire, nous ajoutons ici l'analyse des résultats des trois logiciels précités sur le jeu de données ER (Estrogen Receptor) agoniste de la banque DUD.[159]

Lors de l'étude réalisée sur le jeu de données PPAR gamma, les trois logiciels ont permis de bonnes reconnaissances précoces et des classements satisfaisants des ligands, associés à des variations de score pertinentes (AUCs > 0.748, BEDROCs > 0.447, TGs > 0.415). Les résultats étaient comparables pour le jeu de données ER agoniste, avec une reconnaissance précoce moins marquée (AUCs > 0.708, BEDROC > 0.299, TGs > 0.301) (***Tableau 14*** et ***Figure 76A***). Pour l'expérience réalisée sur le jeu de données AR les performances obtenues étaient plus mitigées, mais restent supérieures à un tirage aléatoire des composés (AUCs > 0.684, BEDROC > 0.257, TGs > 0.067).

Dans le cas de PPAR gamma, nous avons pu observer que les combinaisons consensus des résultats des trois logiciels ne produisaient pas automatiquement de meilleures performances que leurs combinaisons deux à deux. Les consensus obtenus en utilisant trois logiciels améliorent la reconnaissance précoce des ligands ainsi que leur classement global, en utilisant le minimum des rangs ou le maximum des z-scores et par rapport au logiciel le plus performant (Surflex-dock). Cependant, avec ces mêmes méthodes consensus, la combinaison des deux logiciels les plus et moins efficaces (respectivement, Surflex-dock et ICM) permet d'obtenir une reconnaissance précoce encore meilleure. Ce résultat s'explique par le fait que les fractions précoces identifiées par Surflex-dock et ICM comportent une grande part de ligands différents, tandis que Surflex-dock et Autodock Vina tendent à attribuer de bons scores aux mêmes ligands.

Inversement, dans le cas de l'expérience réalisée sur le récepteur AR, les stratégies visant à reclasser les composés selon leur meilleur rang, z-score ou score normalisé sont inefficaces. Par contre, l'utilisation de consensus moyens (moyenne des rangs, z-scores ou scores normalisés) améliore les résultats de manière marquée, particulièrement concernant la reconnaissance précoce et lorsque les résultats des trois logiciels sont combinés. Le succès des consensus moyens dans cette expérience indique une complémentarité des méthodes de criblage virtuel dans la prise en compte des phénomènes régissant les interactions récepteur-ligand, aboutissant à une meilleure approximation des phénomènes d'affinité.

Dans le cas du jeu de données ER agoniste, les consensus utilisant les rangs et les z-scores permettent tous d'améliorer les résultats par rapport à la méthode criblage la plus performante, Autodock Vina, particulièrement concernant la reconnaissance précoce (***Tableau 15*** et ***Figure 76***). La moyenne des scores normalisés est également efficace, contrairement au maximum des scores normalisés. Ce résultat s'explique par la grande sensibilité aux valeurs extrêmes de ces deux dernières stratégies.

	AUC	RIE	BEDROC	TG
Surflex-dock	0.708	4.682	0.299	0.301
ICM	0.772	7.776	0.496	0.462
Autodock Vina	0.833	8.066	0.514	0.533

Tableau 14. *Métriques AUC, RIE et BEDROC obtenues sur le jeu de données ER agoniste de la banque DUD*[159] *avec Surflex-dock,*[349,350] *ICM*[359] *et Autodock Vina.*[457]

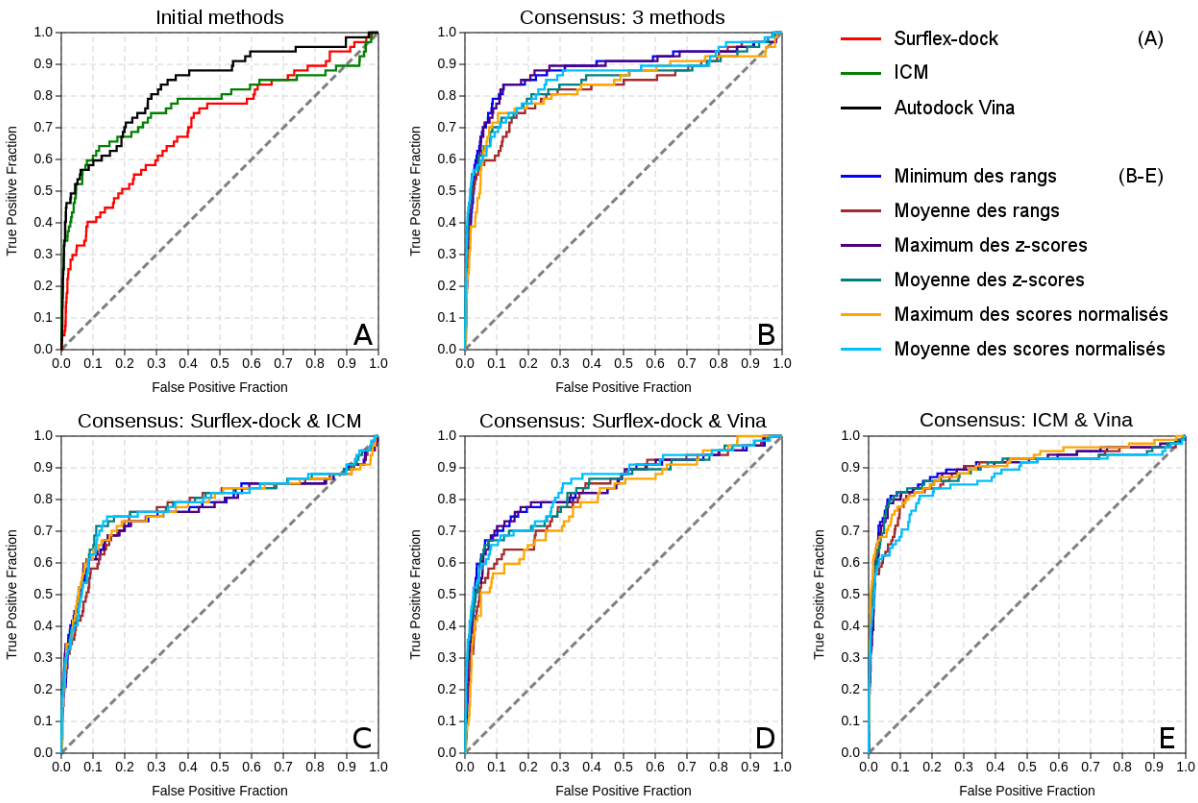


Figure 76. *Courbes de ROC des résultats obtenus sur le jeu de données ER agoniste de la banque DUD.*[159] *(A) Criblages virtuel réalisés avec Surflex-dock,*[349,350] *ICM*[359] *et Autodock Vina.*[457] *(B) Reclassements consensus obtenus en combinant les résultats des trois logiciels. (C-E) Reclassements consensus obtenus en combinant les résultats des logiciels deux à deux.*

	3 méthodes		Surflex-dock, ICM		ICM, Vina		Surflex-dock, Vina	
	AUC	BEDROC	AUC	BEDROC	AUC	BEDROC	AUC	BEDROC
Minimum des rangs	0.879	**0.577**	0.771	0.450	0.863	**0.622**	0.839	0.513
Moyenne des rangs	0.823	**0.549**	0.774	0.436	0.837	**0.575**	0.817	0.468
Maximum des z-scores	0.881	**0.600**	0.770	0.470	0.864	**0.619**	0.841	0.526
Moyenne des z-scores	0.840	**0.600**	0.788	0.476	0.848	**0.619**	0.834	0.543
Maximum des scores normalisés	0.828	0.505	0.775	0.459	0.822	0.543	0.796	0.405
Moyenne des scores normalisés	0.853	**0.596**	0.785	0.460	0.860	**0.604**	0.845	0.551

Tableau 15. *Métriques AUC et BEDROC obtenues sur le jeu de données ER agoniste de la banque DUD*[159] *en combinant les résultats de Surflex-dock,*[349,350] *ICM*[359] *et Autodock Vina*[457] *grâce aux méthodes consensus de Screening Explorer.*

Les méthodes consensus constituent donc un outil intéressant, notamment du fait de leur efficacité et de leur simplicité. Cependant, leur application ne peut pas être envisagée de manière systématique dans le cadre d'études prospectives. Pour assurer leur succès, chaque méthode de criblage virtuel utilisée dans le consensus doit : (i) obtenir des résultats satisfaisants

de manière indépendante (supérieurs à un tirage aléatoire des composés et avec une reconnaissance précoce des ligands) et (ii) apporter une information unique au consensus. Hors de ces deux conditions, les approches consensus ne peuvent pas améliorer les résultats des criblages virtuels au-delà des performances de la méthode la plus efficace individuellement. Screening Explorer permet d'évaluer rapidement les résultats de méthodes consensus simples de manière rétrospective, afin d'estimer la complémentarité des logiciels de criblage virtuel et la pertinence de l'utilisation de consensus dans le cadre d'applications prospectives.

1.2.2.2. Autres applications

Bien que nous ayons présenté Screening Explorer en analysant exclusivement des expériences de criblage virtuel « structure-based », cet outil peut tout aussi bien être appliqué à l'analyse de résultats « ligand-based ». Rétrospectivement, les performances de chaque approche de criblage peuvent être évaluées et optimisées à partir des données issues des banques d'évaluation ou, souvent de manière plus exigeante, à partir de données issues de HTS expérimentaux.

2. Recherche de nouveaux inhibiteurs du VIH

En parallèle aux deux projets décrits précédemment, supportant l'évaluation et le développement des méthodes de criblage virtuel, nous avons également amorcé un projet de recherche et développement de nouveaux inhibiteurs du VIH. Suite à une étude récente de l'équipe génomique de notre laboratoire, de nouveaux gènes ont été identifiés comme favorisant la non-progression[15,17] ou la progression rapide[16] vers le SIDA, grâce à des méthodes GWAS[18] et eQTL.[19] Nous avons donc exploré ces résultats à la recherche de cibles « druggables » et de « hits » potentiels.

Parmi les 36 gènes identifiés par notre équipe génomique, 3 sont impliqués dans la formation du complexe majeur d'histocompatibilité (CMH) et ont donc été ignorés dans la suite de notre étude. Parmi les 33 gènes restant, l'expression de 18 d'entre eux favoriserait la non-progression vers le SIDA, tandis que l'expression des 15 autres favoriserait une progression rapide de la maladie. Dans une démarche thérapeutique, il convient ici de chercher à inhiber l'interaction des protéines issues des gènes favorisant la progression rapide vers le SIDA avec leurs partenaires biologiques. Pour ce faire, afin d'estimer la possibilité d'utiliser des criblages virtuels, nous avons évalué la disponibilité des structures de ces protéines cibles grâce à des alignements de séquence réalisés avec l'outil BLASTp.[474] Pour 8 de ces cibles, la structure est directement disponible dans la PDB[196] et résolue par cristallographie à rayons X ou microscopie électronique, avec une forte couverture et une forte identité des alignements paire à paire. Pour 4 de ces cibles, aucune structure n'est ressortie avec un alignement suffisement significatif pour pouvoir réaliser, à minima, une modélisation par homologie. Enfin, pour les 3 dernières cibles dont l'expression favorise la progression rapide vers le SIDA, les taux de couverture et d'identité des séquences alignées paire à paire sont suffisants pour réaliser des modèles par homologies et amorcer des campagnes de criblage virtuel.

Dans un premier temps, nous avons commencé par rechercher les inhibiteurs connus de ces cibles, recensés dans la littérature sans pour autant avoir été testés dans le cadre d'études portant sur la progression du SIDA. Cette recherche a permis d'identifier 63 molécules, dont 47 ont pu être commandées auprès des fournisseurs. Ces composés sont actuellement à l'étude par la société Peptinov, partenaire de notre laboratoire, afin d'estimer le potentiel thérapeutique de ces composés dans des essais *in vitro* (i) d'infection par le VIH, (ii) de prolifération virale et (iii) de réactivation virale.

Troisième partie : Conclusion

Les méthodes de criblage virtuel sont communément employées dans le processus de recherche et développement de médicaments, en amont des criblages expérimentaux à haut débit, afin d'optimiser leurs rendements. Cependant, malgré l'intérêt indéniable de ces méthodes, il reste difficile d'établir des lignes directrices pour assurer la performance des criblages virtuels et l'obtention de « hits » réels lors des campagnes HTS. Particulièrement, il est difficile de choisir des seuils de score d'affinité qui serviront, à l'issu d'un criblage virtuel, à sélectionner la fraction des composés qui sera utilisée pour procéder aux essais HTS.

Nous avons donc proposé l'usage des Courbes de Prédictivité pour compléter les outils d'évaluation actuels et améliorer la compréhension des résultats de criblage virtuel. Cette métrique permet de prendre directement en compte les scores d'affinité calculés par les méthodes afin d'estimer la probabilité d'identifier des composés actifs, rétrospectivement, afin d'évaluer ensuite la qualité des résultats de criblages virtuels prospectifs. Les Courbes de Prédictivité, couplées aux courbes de ROC et aux courbes d'enrichissement, permettent une analyse graphique, intuitive et rapide des résultats.

Nous avons ensuite choisi de développer Screening Explorer en nous focalisant sur deux objectifs : (i) promouvoir et faciliter l'utilisation des Courbes de Prédictivité et (ii) rendre l'analyse des résultats de criblage virtuel accessible et intuitive au plus grand nombre, particulièrement aux étudiants et aux utilisateurs non-experts, tout en fournissant un outil complet et efficace. Nous avons également souhaité inclure quelques approches consensus simples dans Screening Explorer, puisque celles-ci peuvent permettre une augmentation importante de la reconnaissance précoce des composés actifs, à condition d'utiliser une combinaison de méthodes de criblage un minimum performantes individuellement.

Enfin, en parallèle, nous avons réalisé une première exploration des données fournies par notre équipe génomique, portant sur la recherche de gènes dont l'expression favoriserait la progression vers le SIDA. La société Peptinov, associée à notre laboratoire, a commencé récemment des essais *in vitro* afin d'estimer le potentiel thérapeutique des 43 inhibiteurs identifiés lors de notre recherche de « hits », qui portent sur 3 gènes distincts. Ces essais in vitro ont pour objectif d'évaluer le potentiel de ces composés à inhiber les processus : (i) d'infection par le VIH, (ii) de prolifération virale et (iii) de réactivation virale.

Bibliographie

1. Stone, E. An Account of the Success of the Bark of the Willow in the Cure of Agues. In a Letter to the Right Honourable George Earl of Macclesfield, President of R. S. from the Rev. Mr. Edmund Stone, of Chipping-Norton in Oxfordshire. *Philos. Trans.* **53,** 195–200 (1763).
2. Warner, T. D. & Mitchell, J. A. Cyclooxygenase-3 (COX-3): Filling in the gaps toward a COX continuum? *Proc. Natl. Acad. Sci.* **99,** 13371–13373 (2002).
3. International Narcotics Control Board 2016 report. Available at: https://www.incb.org/incb/en/narcotic-drugs/estimates/nacotic-drugs-estimates.html. (Accessed: 16th December 2016)
4. Kowalik, M. *et al.* Parallel Optimization of Synthetic Pathways within the Network of Organic Chemistry. *Angew. Chemie Int. Ed.* **51,** 7928–7932 (2012).
5. Hughes, J. P., Rees, S., Kalindjian, S. B. & Philpott, K. L. Principles of early drug discovery. *Br. J. Pharmacol.* **162,** 1239–1249 (2011).
6. Morgan, S., Grootendorst, P., Lexchin, J., Cunningham, C. & Greyson, D. The cost of drug development: A systematic review. *Health Policy (New. York).* **100,** 4–17 (2011).
7. Paul, S. M. *et al.* How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat Rev Drug Discov* **9,** 203–214 (2010).
8. Tanrikulu, Y., Krüger, B. & Proschak, E. The holistic integration of virtual screening in drug discovery. *Drug Discov. Today* **18,** 358–364 (2013).
9. Duffy, B. C., Zhu, L., Decornez, H. & Kitchen, D. B. Early phase drug discovery: Cheminformatics and computational techniques in identifying lead series. *Bioorganic Med. Chem.* **20,** 5324–5342 (2012).
10. Overington, J. P., Al-lazikani, B. & Hopkins, A. L. How many drug targets are there ? *Nat. Rev. Drug Discov.* **5,** 993–996 (2006).
11. Owens, J. Determining druggability. *Nat. Rev. Drug Discov.* **6,** 2275 (2007).
12. Yang, Y., Adelstein, S. J. & Kassis, A. I. Target discovery from data mining approaches. *Drug Discov. Today* **17S,** S16–S23 (2012).
13. Chan, J. N. Y., Nislow, C. & Emili, A. Recent advances and method development for drug target identification. *Trends in Pharmacological Sciences* **31,** 82–88 (2010).
14. Wang, S., Sim, T. B., Kim, Y. S. & Chang, Y. T. Tools for target identification and validation. *Curr. Opin. Chem. Biol.* **8,** 371–377 (2004).
15. Limou, S. *et al.* Genomewide Association Study of an AIDS - Nonprogression Cohort Emphasizes the Role Played by HLA Genes (ANRS Genomewide Association Study 02). *J. Infect. Dis.* **199,** 419–426 (2009).
16. Le Clerc, S. *et al.* Genomewide Association Study of a Rapid Progression Cohort Identifies New Susceptibility Alleles for AIDS (ANRS Genomewide Association Study 03). *J. Infect. Dis.* **200,** 1194–1201 (2009).
17. Spadoni, J., Rucart, P., Le Clerc, S. & Manen, D. Van. Identification of Genes Whose Expression Profile Is Associated with Non-Progression towards AIDS Using eQTLs. *PLoS One* 1–13 (2015).
18. Bush, W. S. & Moore, J. H. Chapter 11: Genome-Wide Association Studies. *PLoS Comput. Biol.* **8,** e1002822 (2012).
19. Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat Genet* **48,** 481–487 (2016).

20. Pauling, L., Itano, H. A., Singer, S. & Wells, I. C. Sickle Cell Anemia, a Molecular Disease. *Science (80-.).* **110,** 543–548 (1949).
21. Ingram, V. M. A specific chemical difference between the globins of normal human and sickle-cell anæmia hæmoglobin. *Nature* **178,** 792–794 (1956).
22. Levy, E. D. & Teichmann, S. Structural, evolutionary, and assembly principles of protein oligomerization. *Prog Mol Biol Transl Sci* **117,** 25–51 (2013).
23. Breindl, M., Doehmer, J., Willecke, K., Dausman, J. & Jaenisch, R. Germ line integration of Moloney leukemia virus: identification of the chromosomal integration site. *Proc. Natl. Acad. Sci. U. S. A.* **76,** 1938–42 (1979).
24. Palmiter, R. D. & Brinster, R. L. Transgenic mice. *Cell* **41,** 343–345 (1985).
25. Forss-Petter, S. *et al.* Transgenic mice expressing B-galactosidase in mature neurons under neuron-specific enolase promoter control. *Neuron* **5,** 187–197 (1990).
26. Bajorath, J. Integration of virtual and high-throughput screening. *Nat. Rev. Drug Discov.* **1,** 882–894 (2002).
27. Moitessier, N., Englebienne, P., Lee, D., Lawandi, J. & Corbeil, C. R. Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go. *Br. J. Pharmacol.* **153,** S7–S26 (2008).
28. Lyne, P. D. Structure-based virtual screening: an overview. *Drug Discov. Today* **7,** 1047–1055 (2002).
29. McGaughey, G. B. *et al.* Comparison of topological, shape, and docking methods in virtual screening. *J. Chem. Inf. Model.* **47,** 1504–19 (2007).
30. Empereur-Mot, C. *et al.* Predictiveness curves in virtual screening. *J. Cheminform.* **7,** 1–17 (2015).
31. Cross, J. B. *et al.* Comparison of several molecular docking programs: Pose prediction and virtual screening accuracy. *J. Chem. Inf. Model.* **49,** 1455–1474 (2009).
32. Ben Nasr, N., Guillemain, H., Lagarde, N., Zagury, J. F. & Montes, M. Multiple structures for virtual ligand screening: Defining binding site properties-based criteria to optimize the selection of the query. *J. Chem. Inf. Model.* **53,** 293–311 (2013).
33. Lavecchia, A. & Giovanni, C. D. Virtual screening strategies in drug discovery: A critical review. *Curr. Med. Chem.* **20,** 2839–2860 (2013).
34. Rohrer, S. G. & Baumann, K. Impact of benchmark data set topology on the validation of virtual screening methods: exploration and quantification by spatial statistics. *J. Chem. Inf. Model.* **48,** 704–18 (2008).
35. Copeland, R. A., Pompliano, D. L. & Meek, T. D. Drug-target residence time and its implications for lead optimization. *Nat Rev Drug Discov* **5,** 730–739 (2006).
36. Verma, J., Khedkar, V. M. & Coutinho, E. C. 3D-QSAR in Drug Design - A Review. *Curr. Top. Med. Chem.* **10,** 95–115 (2010).
37. Guha, R. & Van Drie, J. H. Structure−Activity Landscape Index: Identifying and Quantifying Activity Cliffs. *J. Chem. Inf. Model.* **48,** 646–658 (2008).
38. Hubbard, R. E. 3D structure and the drug-discovery process. *Mol. Biosyst.* **1,** 391–406 (2005).
39. Bender, A. *et al.* Analysis of pharmacology data and the prediction of adverse drug reactions and off-target effects from chemical structure. *ChemMedChem* **2,** 861–73 (2007).
40. Wang, D. & Bakhai, A. *Clinical Trials: A Practical Guide to Design, Analysis, and Reporting*. (Remedica Medical Education and Publishing, 2006).
41. Kola, I. & Landis, J. Can the pharmaceutical industry reduce attrition rates? *Nat Rev*

Drug Discov **3,** 711–716 (2004).
42. Booth, B., Glassman, R. & Ma, P. Oncology's trials. *Nat Rev Drug Discov* **2,** 609–610 (2003).
43. Wasko, M. J., Pellegrene, K. A., Madura, J. D. & Surratt, C. K. A Role for Fragment-Based Drug Design in Developing Novel Lead Compounds for Central Nervous System Targets. *Frontiers in Neurology* **6,** 197 (2015).
44. Tuccinardi, T. Docking-based virtual screening: Recent developments. *Comb. Chem. High Throughput Screen.* **12,** 303–314 (2009).
45. Kozakov, D. *et al.* The FTMap family of web servers for determining and characterizing ligand-binding hot spots of proteins. *Nat. Protoc.* **10,** 733–755 (2015).
46. Butkiewicz, M. *et al.* Benchmarking ligand-based virtual high-throughput screening with the pubchem database. *Molecules* **18,** 735–756 (2013).
47. CAS Chemical Substances. Available at: http://www.cas.org/. (Accessed: 3rd December 2016)
48. Ruddigkeit, L., Blum, L. C. & Reymond, J.-L. Visualization and Virtual Screening of the Chemical Universe Database GDB-17. *J. Chem. Inf. Model.* **53,** 56–65 (2013).
49. Blum, L. C. & Reymond, J.-L. 970 Million Druglike Small Molecules for Virtual Screening in the Chemical Universe Database GDB-13. *J. Am. Chem. Soc.* **131,** 8732–8733 (2009).
50. Fink, T. & Reymond, J.-L. Virtual Exploration of the Chemical Universe up to 11 Atoms of C, N, O, F: Assembly of 26.4 Million Structures (110.9 Million Stereoisomers) and Analysis for New Ring Systems, Stereochemistry, Physicochemical Properties, Compound Classes, and Drug Discov. *J. Chem. Inf. Model.* **47,** 342–353 (2007).
51. 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.* **52,** 2864–2875 (2012).
52. Klebe, G. Virtual ligand screening: strategies, perspectives and limitations. *Drug Discov. Today* **11,** 580–594 (2006).
53. Kolb, P., Ferreira, R. S., Irwin, J. J. & Shoichet, B. K. Docking & Chemoinformatic Screens for New Ligands and Targets. *Curr. Opin. Biotechnol.* **20,** 429–436 (2009).
54. Geppert, H., Vogt, M. & Bajorath, J. Current Trends in Ligand-Based Virtual Screening: Molecular Representations, Data Mining Methods, New Application Areas, and Performance Evaluation. *J. Chem. Inf. Model.* **50,** 205–216 (2010).
55. ChemBridge screening libraries. Available at: http://www.chembridge.com/screening_libraries/. (Accessed: 3rd December 2016)
56. ASINEX screening libraries. Available at: http://www.asinex.com/libraries.html. (Accessed: 3rd December 2016)
57. Life Chemicals screening libraries. Available at: http://www.lifechemicals.com/services. (Accessed: 3rd December 2016)
58. TimTec screening libraries. Available at: http://www.timtec.net/screening-compound-libraries.html. (Accessed: 3rd December 2016)
59. Maybridge screening libraries. Available at: http://www.maybridge.com/. (Accessed: 3rd December 2016)
60. Bienstock, R. J. in *Library Design, Search Methods, and Applications of Fragment-Based Drug Design* **1076,** 1 (American Chemical Society, 2011).
61. Congreve, M., Carr, R., Murray, C. & Jhoti, H. A 'Rule of Three' for fragment-based lead discovery? *Drug Discov. Today* **8,** 876–877 (2003).

62. Erlanson, D. A., McDowell, R. S. & O'Brien, T. Fragment-Based Drug Discovery. *J. Med. Chem.* **47,** 3463–3482 (2004).
63. Dömling, A., Wang, W. & Wang, K. Chemistry and Biology Of Multicomponent Reactions. *Chem. Rev.* **112,** 3083–3135 (2012).
64. Lagunin, A. & Poroikov, D. F. and V. Multi-Targeted Natural Products Evaluation Based on Biological Activity Prediction with PASS. *Current Pharmaceutical Design* **16,** 1703–1717 (2010).
65. Cragg, G. M. Paclitaxel (Taxol®): A success story with valuable lessons for natural product drug discovery and development. *Med. Res. Rev.* **18,** 315–331 (1998).
66. Milletti, F., Storchi, L., Sforna, G., Cross, S. & Cruciani, G. Tautomer Enumeration and Stability Prediction for Virtual Screening on Large Chemical Databases. *J. Chem. Inf. Model.* **49,** 68–75 (2009).
67. Kochev, N. T., Paskaleva, V. H. & Jeliazkova, N. Ambit-Tautomer: An Open Source Tool for Tautomer Generation. *Mol. Inform.* **32,** 481–504 (2013).
68. ChemAxon Marvin. Available at: https://www.chemaxon.com/products/marvin/. (Accessed: 9th December 2016)
69. ARChem SPARC. Available at: http://www.archemcalc.com/sparc.html. (Accessed: 9th December 2016)
70. Greenwood, J. R., Calkins, D., Sullivan, A. P. & Shelley, J. C. Towards the comprehensive, rapid, and accurate prediction of the favorable tautomeric states of drug-like molecules in aqueous solution. *J. Comput. Aided. Mol. Des.* **24,** 591–604 (2010).
71. Shelley, J. C. *et al.* Epik: a software program for pKa prediction and protonation state generation for drug-like molecules. *J. Comput. Aided. Mol. Des.* **21,** 681–691 (2007).
72. Knox, A. J. S., Meegan, M. J., Carta, G. & Lloyd, D. G. Considerations in Compound Database Preparation'Hidden' Impact on Virtual Screening Results. *J. Chem. Inf. Model.* **45,** 1908–1919 (2005).
73. Lipinski, C. A., Lombardo, F., Dominy, B. W. & Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* **23,** 3–26 (1997).
74. Veber, D. F. *et al.* Molecular Properties That Influence the Oral Bioavailability of Drug Candidates. *J. Med. Chem.* **45,** 2615–2623 (2002).
75. Zheng, S. *et al.* A New Rapid and Effective Chemistry Space Filter in Recognizing a Druglike Database. *J. Chem. Inf. Model.* **45,** 856–862 (2005).
76. Muegge, I., Heald, S. L. & Brittelli, D. Simple Selection Criteria for Drug-like Chemical Matter. *J. Med. Chem.* **44,** 1841–1846 (2001).
77. Zernov, V. V, Balakin, K. V, Ivaschenko, A. A., Savchuk, N. P. & Pletnev, I. V. Drug Discovery Using Support Vector Machines. The Case Studies of Drug-likeness, Agrochemical-likeness, and Enzyme Inhibition Predictions. *J. Chem. Inf. Comput. Sci.* **43,** 2048–2056 (2003).
78. Cortes, C. & Vapnik, V. Support-vector networks. *Mach. Learn.* **20,** 273–297 (1995).
79. Ajay, Walters, W. P. & Murcko, M. A. Can We Learn To Distinguish between 'Drug-like' and 'Nondrug-like' Molecules? *J. Med. Chem.* **41,** 3314–3324 (1998).
80. Sadowski, J. & Kubinyi, H. A Scoring Scheme for Discriminating between Drugs and Nondrugs. *J. Med. Chem.* **41,** 3325–3329 (1998).
81. Hann, M. M. & Oprea, T. I. Pursuing the leadlikeness concept in pharmaceutical research. *Curr. Opin. Chem. Biol.* **8,** 255–263 (2004).
82. Toropov, A. A., Toropova, A. P., Raska Jr, I., Leszczynska, D. & Leszczynski, J.

Comprehension of drug toxicity: Software and databases. *Comput. Biol. Med.* **45,** 20–25 (2014).
83. Valerio Jr., L. G. In silico toxicology for the pharmaceutical sciences. *Toxicol. Appl. Pharmacol.* **241,** 356–370 (2009).
84. FDA Adverse Event Reporting System. Available at: https://open.fda.gov/drug/event/reference/. (Accessed: 3rd December 2016)
85. Lahl, U. & Gundert-Remy, U. The Use of (Q)SAR Methods in the Context of REACH. *Toxicol. Mech. Methods* **18,** 149–158 (2008).
86. EUR-Lex - Access to European Union Law - REACH regulation. Available at: http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32006R1907. (Accessed: 3rd December 2016)
87. EUR-Lex - Access to European Union Law - CLP classification. Available at: http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32008R1272. (Accessed: 3rd December 2016)
88. Verheyen, G. R., Braeken, E., Van Deun, K. & Van Miert, S. Evaluation of existing (Q)SAR models for skin and eye irritation and corrosion to use for REACH registration. *Toxicol. Lett.* **265,** 47–52 (2017).
89. Levet, A. *et al.* Acute aquatic toxicity of organic solvents modeled by QSARs. *J. Mol. Model.* **22,** 288 (2016).
90. Nendza, M. *et al.* Screening for potential endocrine disruptors in fish: evidence from structural alerts and in vitro and in vivo toxicological assays. *Environ. Sci. Eur.* **28,** 26 (2016).
91. Sanguinetti, M. C. & Tristani-Firouzi, M. hERG potassium channels and cardiac arrhythmia. *Nature* **440,** 463–469 (2006).
92. Du, L., Li, M., You, Q. & Xia, L. A novel structure-based virtual screening model for the hERG channel blockers. *Biochem. Biophys. Res. Commun.* **355,** 889–894 (2007).
93. 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.* **267,** 727–748 (1997).
94. Watts, K. S. *et al.* ConfGen: A Conformational Search Method for Efficient Generation of Bioactive Conformers. *J. Chem. Inf. Model.* **50,** 534–546 (2010).
95. Ebejer, J.-P., Morris, G. M. & Deane, C. M. Freely Available Conformer Generation Methods: How Good Are They? *J. Chem. Inf. Model.* **52,** 1146–1158 (2012).
96. Sadowski, J., Gasteiger, J. & Klebe, G. Comparison of Automatic Three-Dimensional Model Builders Using 639 X-ray Structures. *J. Chem. Inf. Comput. Sci.* **34,** 1000–1008 (1994).
97. 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.* **50,** 572–584 (2010).
98. Molecular Networks - ROTATE Classic. Available at: https://www.mn-am.com/products/rotate. (Accessed: 16th December 2016)
99. O'Boyle, N. M., Vandermeersch, T., Flynn, C. J., Maguire, A. R. & Hutchison, G. R. Confab - Systematic generation of diverse low-energy conformers. *J. Cheminform.* **3,** 8 (2011).
100. Zhang, W., Robert, A., Vogensen, S. B. & Howe, J. R. The Relationship between Agonist Potency and AMPA Receptor Kinetics. *Biophys. J.* **91,** 1336–1346 (2006).
101. Jin, R., Horning, M., Mayer, M. L. & Gouaux, E. Mechanism of Activation and

Selectivity in a Ligand-Gated Ion Channel: Structural and Functional Studies of GluR2 and Quisqualate,. *Biochemistry* **41,** 15635–15643 (2002).

102. Kuang, D. & Hampson, D. R. Ion dependence of ligand binding to metabotropic glutamate receptors. *Biochem. Biophys. Res. Commun.* **345,** 1–6 (2006).
103. Mekenyan, O., Dimitrov, D., Nikolova, N. & Karabunarliev, S. Conformational Coverage by a Genetic Algorithm. *J. Chem. Inf. Comput. Sci.* **39,** 997–1016 (1999).
104. Vainio, M. J. & Johnson, M. S. Generating Conformer Ensembles Using a Multiobjective Genetic Algorithm. *J. Chem. Inf. Model.* **47,** 2462–2474 (2007).
105. Liu, X. *et al.* Cyndi: a multi-objective evolution algorithm based method for bioactive molecular conformational generation. *BMC Bioinformatics* **10,** 101 (2009).
106. Spellmeyer, D. C., Wong, A. K., Bower, M. J. & Blaney, J. M. Conformational analysis using distance geometry methods. *J. Mol. Graph. Model.* **15,** 18–36 (1997).
107. Havel, T. F., Kuntz, I. D. & Crippen, G. M. The theory and practice of distance geometry. *Bull. Math. Biol.* **45,** 665–720 (1983).
108. Chang, G., Guida, W. C. & Still, W. C. An internal-coordinate Monte Carlo method for searching conformational space. *J. Am. Chem. Soc.* **111,** 4379–4386 (1989).
109. Wilson, S. R., Cui, W., Moskowitz, J. W. & Schmidt, K. E. Applications of simulated annealing to the conformational analysis of flexible molecules. *J. Comput. Chem.* **12,** 342–349 (1991).
110. Metropolis, N. & Ulam, S. The Monte Carlo method. *J. Am. Stat. Assoc.* **44,** 335–341 (1949).
111. RDKIT: Open-source cheminformatics. Available at: http://www.rdkit.org. (Accessed: 2nd December 2016)
112. Miteva, M. A., Guyon, F. & Tufféry, P. Frog2: Efficient 3D conformation ensemble generator for small compounds. *Nucleic Acids Res.* **38,** 622–627 (2010).
113. Schulz, K.-P. Concepts and Applications of Molecular Similarity. *Berichte der Bunsengesellschaft für Phys. Chemie* **96,** 1087 (1992).
114. Hawkins, P. C. D., Skillman, A. G. & Nicholls, A. Comparison of Shape-Matching and Docking as Virtual Screening Tools. *J Med Chem* **50,** (2007).
115. Rognan, D. Chemogenomic approaches to rational drug design. *Br. J. Pharmacol.* **152,** 38–52 (2007).
116. Cone, M. M., Venkataraghavan, R. & McLafferty, F. W. Molecular structure comparison program for the identification of maximal common substructures. *J. Am. Chem. Soc.* **99,** 7668–7671 (1977).
117. Hariharan, R. *et al.* MultiMCS: A Fast Algorithm for the Maximum Common Substructure Problem on Multiple Molecules. *J. Chem. Inf. Model.* **51,** 788–806 (2011).
118. Kawabata, T. Build-Up Algorithm for Atomic Correspondence between Chemical Structures. *J. Chem. Inf. Model.* **51,** 1775–1787 (2011).
119. Kawabata, T. & Nakamura, H. 3D Flexible Alignment Using 2D Maximum Common Substructure: Dependence of Prediction Accuracy on Target-Reference Chemical Similarity. *J. Chem. Inf. Model.* **54,** 1850–1863 (2014).
120. OpenEye toolkit. Available at: http://www.eyesopen.com/oechem-tk. (Accessed: 9th December 2016)
121. ChemAxon JKlustor. Available at: https://www.chemaxon.com/products/jklustor/. (Accessed: 9th December 2016)
122. Cereto-Massagué, A. *et al.* Molecular fingerprint similarity search in virtual screening. *Methods* **71,** 58–63 (2015).

123. Bajusz, D., Rácz, A. & Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J. Cheminform.* **7,** 20 (2015).
124. Durant, J. L., Leland, B. A., Henry, D. R. & Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.* **42,** 1273–1280 (2002).
125. Bolton, E. E., Wang, Y., Thiessen, P. A. & Bryant, S. H. in *Annual Reports in Computational Chemistry* **4,** 217–241 (2008).
126. ChemFP. Available at: http://chemfp.com/. (Accessed: 9th December 2016)
127. Steinbeck, C. *et al.* The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo- and Bioinformatics. *J. Chem. Inf. Comput. Sci.* **43,** 493–500 (2003).
128. Steinbeck, C., Hoppe, C., Kuhn, S., Floris, M. & Willighagen, R. G. and E. L. Recent Developments of the Chemistry Development Kit (CDK) - An Open-Source Java Library for Chemo- and Bioinformatics. *Current Pharmaceutical Design* **12,** 2111–2120 (2006).
129. Barnard, J. M. & Downs, G. M. Chemical Fragment Generation and Clustering Software. *J. Chem. Inf. Comput. Sci.* **37,** 141–142 (1997).
130. Tovar, A., Eckert, H. & Bajorath, J. Comparison of 2D Fingerprint Methods for Multiple-Template Similarity Searching on Compound Activity Classes of Increasing Structural Diversity. *ChemMedChem* **2,** 208–217 (2007).
131. Empreintes Daylight. Available at: http://www.daylight.com/. (Accessed: 9th December 2016)
132. Certara SYBYL-X. Available at: https://www.certara.com/software/molecular-modeling-and-simulation/sybyl-x-suite/cheminformatics/. (Accessed: 10th December 2016)
133. Xue, L., Godden, J. W., Stahura, F. L. & Bajorath, J. Design and Evaluation of a Molecular Fingerprint Involving the Transformation of Property Descriptor Values into a Binary Classification Scheme. *J. Chem. Inf. Comput. Sci.* **43,** 1151–1157 (2003).
134. Lengauer, T., Lemmen, C., Rarey, M. & Zimmermann, M. Novel technologies for virtual screening. *Drug Discov. Today* **9,** 27–34 (2004).
135. Böhm, H. J., Flohr, A. & Stahl, M. Scaffold hopping. *Drug Discov. Today Technol.* **1,** 217–224 (2004).
136. Hinselmann, G., Rosenbaum, L., Jahn, A., Fechner, N. & Zell, A. jCompoundMapper: An open source Java library and command-line tool for chemical fingerprints. *J. Cheminform.* **3,** 3 (2011).
137. ChemAxon JChem. Available at: https://www.chemaxon.com/download/jchem-suite/. (Accessed: 10th December 2016)
138. Molecular Operating Environment. Available at: https://www.chemcomp.com/MOE-Molecular_Operating_Environment.htm. (Accessed: 10th December 2016)
139. Schrödinger Canvas. Available at: https://www.schrodinger.com/Canvas/. (Accessed: 11th December 2016)
140. Meyer, A. Y. & Richards, W. G. Similarity of molecular shape. *J. Comput. Aided. Mol. Des.* **5,** 427–439 (1991).
141. Nicholls, A. *et al.* Molecular Shape and Medicinal Chemistry: A Perspective. *J. Med. Chem.* **53,** 3862–3886 (2010).
142. Meyer, A. Y. The size of molecules. *Chem. Soc. Rev.* **15,** 449–474 (1986).
143. Good, A. C. & Richards, W. G. Rapid evaluation of shape similarity using Gaussian functions. *J. Chem. Inf. Comput. Sci.* **33,** 112–116 (1993).
144. Grant, J. A. & Pickup, B. T. A Gaussian Description of Molecular Shape. *J. Phys. Chem.* **99,** 3503–3510 (1995).

145. OpenEye ROCS. Available at: http://www.eyesopen.com/rocs. (Accessed: 14th December 2016)
146. Giganti, D. *et al.* Comparative Evaluation of 3D Virtual Ligand Screening Methods: Impact of the Molecular Alignment on Enrichment. *J. Chem. Inf. Model.* **50,** 992–1004 (2010).
147. Zauhar, R. J., Moyna, G., Tian, L., Li, Z. & Welsh, W. J. Shape Signatures: A New Approach to Computer-Aided Ligand- and Receptor-Based Drug Design. *J. Med. Chem.* **46,** 5674–5690 (2003).
148. Zauhar, R. J. SMART: A solvent-accessible triangulated surface generator for molecular graphics and boundary element applications. *J. Comput. Aided. Mol. Des.* **9,** 149–159 (1995).
149. Claus, B. L. & Underwood, D. J. Discovery informatics: its evolving role in drug discovery. *Drug Discov. Today* **7,** 957–966 (2002).
150. Dobson, C. M. Chemical space and biology. *Nature* **432,** 824–828 (2004).
151. Ballester, P. J. & Richards, W. G. Ultrafast shape recognition to search compound databases for similar molecular shapes. *J. Comput. Chem.* **28,** 1711–1723 (2007).
152. Li, H., Leung, K.-S., Wong, M.-H. & Ballester, P. J. USR-VS: a web server for large-scale prospective virtual screening using ultrafast shape recognition techniques. *Nucleic Acids Res.* (2016).
153. Hall, P. A distribution is completely determined by its translated moments. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **62,** 355–359 (1983).
154. Schreyer, A. M. & Blundell, T. USRCAT: real-time ultrafast shape recognition with pharmacophoric constraints. *J. Cheminform.* **4,** 27 (2012).
155. Shave, S. *et al.* UFSRAT: Ultra-Fast Shape Recognition with Atom Types –The Discovery of Novel Bioactive Small Molecular Scaffolds for FKBP12 and 11βHSD1. *PLoS One* **10,** e0116570 (2015).
156. Armstrong, M. S. *et al.* ElectroShape: fast molecular similarity calculations incorporating shape, chirality and electrostatics. *J Comput Aided Mol Des* **24,** (2010).
157. Armstrong, M. S., Morris, G. M., Finn, P. W., Sharma, R. & Richards, W. G. Molecular similarity including chirality. *J Mol Graph Model* **28,** (2009).
158. Mysinger, M. M., Carchia, M., Irwin, J. J. & Shoichet, B. K. Directory of useful decoys, enhanced (DUD-E): Better ligands and decoys for better benchmarking. *J. Med. Chem.* **55,** 6582–6594 (2012).
159. Huang, N., Shoichet, B. K. & Irwin, J. J. Benchmarking sets for molecular docking. *J. Med. Chem.* **49,** 6789–6801 (2006).
160. Riniker, S. & Landrum, G. A. Open-source platform to benchmark fingerprints for ligand-based virtual screening. *J. Cheminform.* **5,** 26 (2013).
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 Interfac / R Soc* **7,** (2010).
162. Ballester, P. J. Ultrafast shape recognition: method and applications. *Future Med. Chem.* **3,** 65–78 (2010).
163. Halgren, T. A. Merck molecular force field. *J. Comput. Chem.* **17,** 490–641 (1996).
164. Willett, P., Barnard, J. M. & Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **38,** 983–996 (1998).
165. Mukaka, M. M. A guide to appropriate use of Correlation coefficient in medical research. *Malawi Med. J.* **24,** 69–71 (2012).

166. Ehrlich, P. Über die constitution des diphtheriegiftes. *DMW-Deutsche Medizinische Wochenschrift* **24,** 597–600 (1898).
167. Beckett, A. H., Harper, N. J. & Clitherow, J. W. The Importance of Stereoisomerism in Muscarinic Activity. *J. Pharm. Pharmacol.* **15,** 362–371 (1963).
168. Kier, L. B. *Molecular orbital theory in drug research*. (Academic Press, 1971).
169. Glossary of terms used in medicinal chemistry (IUPAC Recommendations 1998) . *Pure and Applied Chemistry* **70,** 1129 (1998).
170. Varnek, A. & Tropsha, A. *Chemoinformatics approaches to virtual screening*. (Royal Society of Chemistry, 2008).
171. Schuffenhauer, A., Floersheim, P., Acklin, P. & Jacoby, E. Similarity Metrics for Ligands Reflecting the Similarity of the Target Proteins. *J. Chem. Inf. Comput. Sci.* **43,** 391–405 (2003).
172. Schneider, G., Neidhart, W., Giller, T. & Schmid, G. 'Scaffold-Hopping' by Topological Pharmacophore Search: A Contribution to Virtual Screening. *Angew. Chemie Int. Ed.* **38,** 2894–2896 (1999).
173. Stiefl, N., Watson, I. A., Baumann, K. & Zaliani, A. ErG: 2D Pharmacophore Descriptions for Scaffold Hopping. *J. Chem. Inf. Model.* **46,** 208–220 (2006).
174. Rarey, M. & Dixon, J. S. Feature trees: A new molecular similarity measure based on tree matching. *J. Comput. Aided. Mol. Des.* **12,** 471–490 (1998).
175. Gillet, V. J., Willett, P. & Bradshaw, J. Similarity Searching Using Reduced Graphs. *J. Chem. Inf. Comput. Sci.* **43,** 338–345 (2003).
176. Reutlinger, M. *et al.* Chemically Advanced Template Search (CATS) for Scaffold-Hopping and Prospective Target Prediction for 'Orphan' Molecules. *Mol. Inform.* **32,** 133–138 (2013).
177. Dror, O., Shulman-Peleg, A. & Wolfson, R. N. and H. J. Predicting Molecular Interactions in silico: I. A Guide to Pharmacophore Identification and its Applications to Drug Design. *Current Medicinal Chemistry* **11,** 71–90 (2004).
178. Ghose, A. K., Viswanadhan, V. N. & Wendoloski, J. J. THE FUNDAMENTALS OF PHARMACOPHORE MODELING IN COMBINATORIAL CHEMISTRY. *J. Recept. Signal Transduct.* **21,** 357–375 (2001).
179. Dolata, D. P., Parrill, A. L. & Walters, W. P. CLEW: The Generation of Pharmacophore Hypotheses Through Machine Learning. *SAR QSAR Environ. Res.* **9,** 53–81 (1998).
180. Li, H., Sutter, J. & Hoffman, R. in *Pharmacophore Perception, Development and Use in Drug Design SE - IUL Biotechnology Series* (ed. Guner, O.) 171–189 (International University Line, 2000).
181. Günther, S., Senger, C., Michalsky, E., Goede, A. & Preissner, R. Representation of target-bound drugs by computed conformers: implications for conformational libraries. *BMC Bioinformatics* **7,** 293 (2006).
182. Dror, O., Schneidman-Duhovny, D., Inbar, Y., Nussinov, R. & Wolfson, H. J. Novel Approach for Efficient Pharmacophore-Based Virtual Screening: Method and Applications. *J. Chem. Inf. Model.* **49,** 2333–2343 (2009).
183. Finn, P. W. *et al.* RAPID: Randomized pharmacophore identification for drug design. *Comput. Geom.* **10,** 263–272 (1998).
184. Holliday, J. D. & Willett, P. Using a Genetic Algorithm to Identify Common Structural Features in Sets of Ligands. *J. Mol. Graph. Model.* **15,** 221–232 (1997).
185. Martin, Y. C. *et al.* A fast new approach to pharmacophore mapping and its application to dopaminergic and benzodiazepine agonists. *J. Comput. Aided. Mol. Des.* **7,** 83–102

(1993).
186. Dixon, S. L. *et al.* PHASE: a new engine for pharmacophore perception, 3D QSAR model development, and 3D database screening: 1. Methodology and preliminary results. *J. Comput. Aided. Mol. Des.* **20,** 647–671 (2006).
187. Wolber, G. & Langer, T. LigandScout: 3-D Pharmacophores Derived from Protein-Bound Ligands and Their Use as Virtual Screening Filters. *J. Chem. Inf. Model.* **45,** 160–169 (2005).
188. Clement, O. A. & Mehl, A. T. in *Pharmacophore Perception, Development and Use in Drug Design SE - IUL Biotechnology Series* 69–84 (2000).
189. Seidel, T., Ibis, G., Bendix, F. & Wolber, G. Strategies for 3D pharmacophore-based virtual screening. *Drug Discov. Today Technol.* **7,** e221–e228 (2010).
190. Chen, X., Rusinko, Tropsha, A. & Young, S. S. Automated Pharmacophore Identification for Large Chemical Data Sets. *J. Chem. Inf. Comput. Sci.* **39,** 887–896 (1999).
191. Handschuh, S. & Gasteiger, J. The Search for the Spatial and Electronic Requirements of a Drug. *Mol. Model. Annu.* **6,** 358–378 (2000).
192. Jones, G., Willett, P. & Glen, R. in *Pharmacophore Perception, Development and Use in Drug Design SE - IUL Biotechnology Series* 85–106 (2000).
193. Richmond, N. J. *et al.* GALAHAD: Pharmacophore Identification by Hypermolecular Alignment of Ligands in 3D. *J. Comput. Aided. Mol. Des.* **20,** 567–587 (2006).
194. Cottrell, S. J., Gillet, V. J., Taylor, R. & Wilton, D. J. Generation of multiple pharmacophore hypotheses using multiobjective optimisation techniques. *J. Comput. Aided. Mol. Des.* **18,** 665–682 (2004).
195. Wolber, G., Seidel, T., Bendix, F. & Langer, T. Molecule-pharmacophore superpositioning and pattern matching in computational drug design. *Drug Discov. Today* **13,** 23–29 (2008).
196. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28,** 235–242 (2000).
197. Raymond, J. W. & Willett, P. Maximum common subgraph isomorphism algorithms for the matching of chemical structures. *J. Comput. Aided. Mol. Des.* **16,** 521–533 (2002).
198. Wolber, G., Dornhofer, A. A. & Langer, T. Efficient overlay of small organic molecules using 3D pharmacophores. *J. Comput. Aided. Mol. Des.* **20,** 773–788 (2006).
199. Gillet, V. J. in *Physico-Chemical and Computational Approaches to Drug Discovery* 151–170 (The Royal Society of Chemistry, 2012).
200. Barnum, D., Greene, J., Smellie, A. & Sprague, P. Identification of common functional configurations among molecules. *J. Chem. Inf. Comput. Sci.* **36,** 563–571 (1996).
201. Leach, A. R., Gillet, V. J., Lewis, R. A. & Taylor, R. Three-Dimensional Pharmacophore Methods in Drug Discovery. *J. Med. Chem.* **53,** 539–558 (2010).
202. Sheridan, R. P. & Kearsley, S. K. Why do we need so many chemical similarity search methods? *Drug Discov. Today* **7,** 903–911 (2002).
203. Damale, M. G., Harke, S. N., Khan, F. A. K. & Sangshetti, D. B. S. and J. N. Recent Advances in Multidimensional QSAR (4D-6D): A Critical Review. *Mini-Reviews in Medicinal Chemistry* **14,** 35–55 (2014).
204. Polanski, J. Receptor Dependent Multidimensional QSAR for Modeling Drug - Receptor Interactions. *Current Medicinal Chemistry* **16,** 3243–3257 (2009).
205. Hopfinger, A. J. *et al.* Construction of 3D-QSAR Models Using the 4D-QSAR Analysis Formalism. *J. Am. Chem. Soc.* **119,** 10509–10524 (1997).
206. Vedani, A. & Dobler, M. 5D-QSAR: The Key for Simulating Induced Fit? *J. Med.*

Chem. **45,** 2139–2149 (2002).
207. 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.* **48,** 3700–3703 (2005).
208. Brown, A. C. & Fraser, T. R. On the Connection between Chemical Constitution and Physiological Action. *Trans. R. Soc. Edinburgh* **25,** 693–739 (1869).
209. Hammett, L. P. The Effect of Structure upon the Reactions of Organic Compounds. Benzene Derivatives. *J. Am. Chem. Soc.* **59,** 96–103 (1937).
210. Hansch, C., Maloney, P. P., Fujita, T. & Muir, R. M. Correlation of Biological Activity of Phenoxyacetic Acids with Hammett Substituent Constants and Partition Coefficients. *Nature* **194,** 178–180 (1962).
211. Wiener, H. Structural Determination of Paraffin Boiling Points. *J. Am. Chem. Soc.* **69,** 17–20 (1947).
212. Randic, M. Characterization of molecular branching. *J. Am. Chem. Soc.* **97,** 6609–6615 (1975).
213. Galvez, J., Garcia, R., Salabert, M. T. & Soler, R. Charge Indexes. New Topological Descriptors. *J. Chem. Inf. Comput. Sci.* **34,** 520–525 (1994).
214. Balaban, A. T. Highly discriminating distance-based topological index. *Chem. Phys. Lett.* **89,** 399–404 (1982).
215. Schultz, H. P. Topological organic chemistry. 1. Graph theory and topological indices of alkanes. *J. Chem. Inf. Comput. Sci.* **29,** 227–228 (1989).
216. Graovac, A., Gutman, I., Trinajstić, N. & Živković, T. Graph theory and molecular orbitals. *Theor. Chim. Acta* **26,** 67–78 (1972).
217. Gutman, I. Selected properties of the Schultz molecular topological index. *J. Chem. Inf. Comput. Sci.* **34,** 1087–1089 (1994).
218. Hosoya, H. Topological Index. A Newly Proposed Quantity Characterizing the Topological Nature of Structural Isomers of Saturated Hydrocarbons. *Bull. Chem. Soc. Jpn.* **44,** 2332–2339 (1971).
219. Kier, L. B. & Hall, L. H. Derivation and significance of valence molecular connectivity. *J. Pharm. Sci.* **70,** 583–589 (1981).
220. Leo, A., Jow, P. Y. C., Silipo, C. & Hansch, C. Calculation of hydrophobic constant (log P) from pi and f constants. *J. Med. Chem.* **18,** 865–868 (1975).
221. Taft, R. W. Linear Steric Energy Relationships. *J. Am. Chem. Soc.* **75,** 4538–4539 (1953).
222. Mulliken, R. S. Electronic Population Analysis on LCAO–MO Molecular Wave Functions. I. *J. Chem. Phys.* **23,** 1833–1840 (1955).
223. Gasteiger, J. & Marsili, M. A new model for calculating atomic charges in molecules. *Tetrahedron Lett.* **19,** 3181–3184 (1978).
224. Gasteiger, J. & Marsili, M. Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges. *Tetrahedron* **36,** 3219–3228 (1980).
225. Hall Associates Consulting MOLCONN. Available at: http://www.molconn.com/. (Accessed: 8th January 2017)
226. Kode Dragon. Available at: https://chm.kode-solutions.net/products_dragon.php. (Accessed: 8th January 2017)
227. Semichem CODESSA. Available at: http://www.semichem.com/codessa/default.php. (Accessed: 8th January 2017)
228. Ivanciuc, O. CODESSA Version 2.13 for Windows. *J. Chem. Inf. Comput. Sci.* **37,** 405–

406 (1997).
229. BIOVIA Pipeline Pilot. Available at: http://accelrys.com/products/collaborative-science/biovia-pipeline-pilot/. (Accessed: 8th January 2017)
230. Shahlaei, M. Descriptor Selection Methods in Quantitative Structure–Activity Relationship Studies: A Review Study. *Chem. Rev.* **113,** 8093–8103 (2013).
231. Lewis, R. A. & Wood, D. Modern 2D QSAR for drug discovery. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **4,** 505–522 (2014).
232. Breiman, L. Random Forests. *Mach. Learn.* **45,** 5–32 (2001).
233. Lusci, A., Pollastri, G. & Baldi, P. Deep Architectures and Deep Learning in Chemoinformatics: The Prediction of Aqueous Solubility for Drug-Like Molecules. *J. Chem. Inf. Model.* **53,** 1563–1575 (2013).
234. Cramer, R. D. & Milne, M. The lattice model: a general paradigm for shape-related structure/activity correlation. *Abstr. Pap. Am. Chem. Soc. Comput. Chem. Sect.* (1979).
235. Wise, M., Cramer, R. D., Smith, D. & Exman, I. Progress in three-dimensional drug design: the use of real time color graphics and computer postulation of bioactive molecules in DYLOMMS. *Quant. Approaches to Drug Des.* 145–146 (1983).
236. 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.* **110,** 5959–5967 (1988).
237. 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.* **37,** 4130–4146 (1994).
238. Jain, A. N., Koile, K. & Chapman, D. Compass: predicting biological activities from molecular surface properties. Performance comparisons on a steroid benchmark. *J. Med. Chem.* **37,** 2315–2327 (1994).
239. Walters, D. E. in (eds. Kubinyi, H., Folkers, G. & Martin, Y. C.) 159–166 (Springer Netherlands, 1998). doi:10.1007/0-306-46858-1_10
240. Silverman, B. D. & Platt, D. E. Comparative Molecular Moment Analysis (CoMMA): 3D-QSAR without Molecular Superposition. *J. Med. Chem.* **39,** 2129–2140 (1996).
241. Robinson, D. D., Winn, P. J., Lyne, P. D. & Richards, W. G. Self-Organizing Molecular Field Analysis: A Tool for Structure−Activity Studies. *J. Med. Chem.* **42,** 573–583 (1999).
242. Jones, J. E. On the Determination of Molecular Fields. II. From the Equation of State of a Gas. *Proc. R. Soc. London. Ser. A* **106,** 463 LP-477 (1924).
243. Carballeira, D. J., Quezada, A. M., Álvarez, E. & Sinisterra, V. J. High Throughput Screening and QSAR-3D/CoMFA: Useful Tools to Design Predictive Models of Substrate Specificity for Biocatalysts†. *Molecules* **9,** (2004).
244. Kubinyi, H. *3D QSAR in Drug Design: Volume 30: CoMFA: Scope and Limitations*. (Springer, 1993).
245. Kearsley, S. K. & Smith, G. M. An alternative method for the alignment of molecular structures: Maximizing electrostatic and steric overlap. *Tetrahedron Comput. Methodol.* **3,** 615–633 (1990).
246. Böhm, H. J., Klebe, G. & Kubinyi, H. *Wirkstoffdesign: der Weg zum Arzneimittel*. (Spektrum Akademischer Verlag, 2002).
247. Sliwoski, G., Kothiwale, S., Meiler, J. & Lowe, E. W. Computational Methods in Drug Discovery. *Pharmacol. Rev.* **66,** 334–395 (2014).
248. Biograf Quasar. Available at:

http://www.biograf.ch/index.php?id=software&subid=quasar. (Accessed: 29th January 2017)

249. Koshland, D. E. Application of a Theory of Enzyme Specificity to Protein Synthesis. *Proc. Natl. Acad. Sci. U. S. A.* **44,** 98–104 (1958).
250. Koehler, J. E. H., Saenger, W. & van Gunsteren, W. F. The flip-flop hydrogen bonding phenomenon. *Eur. Biophys. J.* **16,** 153–168 (1988).
251. NCBI PubMed. Available at: https://www.ncbi.nlm.nih.gov/. (Accessed: 18th January 2017)
252. Herbst, R. S., Fukuoka, M. & Baselga, J. Gefitinib - a novel targeted approach to treating cancer. *Nat Rev Cancer* **4,** 979–987 (2004).
253. Ward, W. H. *et al.* Epidermal growth factor receptor tyrosine kinase. Investigation of catalytic mechanism, structure-based searching and discovery of a potent inhibitor. *Biochem. Pharmacol.* **48,** 659–666 (1994).
254. Naerum, L., Norskov-Lauritsen, L. & Olesen, P. H. Scaffold hopping and optimization towards libraries of glycogen synthase kinase-3 inhibitors. *Bioorg. Med. Chem. Lett.* **12,** 1525–1528 (2002).
255. Singh, J. *et al.* Identification of Potent and Novel α4β1 Antagonists Using in Silico Screening. *J. Med. Chem.* **45,** 2988–2993 (2002).
256. Flohr, S. *et al.* Identification of Nonpeptidic Urotensin II Receptor Antagonists by Virtual Screening Based on a Pharmacophore Model Derived from Structure−Activity Relationships and Nuclear Magnetic Resonance Studies on Urotensin II. *J. Med. Chem.* **45,** 1799–1805 (2002).
257. Mustata, G. *et al.* Discovery of Novel Myc−Max Heterodimer Disruptors with a Three-Dimensional Pharmacophore Model. *J. Med. Chem.* **52,** 1247–1250 (2009).
258. Rush, T. S., Grant, J. A., Mosyak, L. & Nicholls, A. A Shape-Based 3-D Scaffold Hopping Method and Its Application to a Bacterial Protein−Protein Interaction. *J. Med. Chem.* **48,** 1489–1495 (2005).
259. Boström, J., Berggren, K., Elebring, T., Greasley, P. J. & Wilstermann, M. Scaffold hopping, synthesis and structure–activity relationships of 5,6-diaryl-pyrazine-2-amide derivatives: A novel series of CB1 receptor antagonists. *Bioorg. Med. Chem.* **15,** 4077–4084 (2007).
260. Bologa, C. G. *et al.* Virtual and biomolecular screening converge on a selective agonist for GPR30. *Nat. Chem. Biol.* **2,** 207–212 (2006).
261. Freitas, R. F., Oprea, T. I. & Montanari, C. A. 2D QSAR and similarity studies on cruzain inhibitors aimed at improving selectivity over cathepsin L. *Bioorg. Med. Chem.* **16,** 838–853 (2008).
262. Al-Sha'er, M. A. & Taha, M. O. Elaborate ligand-based modeling reveals new nanomolar heat shock protein 90alpha inhibitors. *J. Chem. Inf. Model.* **50,** 1706–1723 (2010).
263. Abad-Zapatero, C. Notes of a protein crystallographer: on the high-resolution structure of the PDB growth rate. *Acta Crystallogr. Sect. D* **68,** 613–617 (2012).
264. Vyas, V. K., Ukawala, R. D., Ghate, M. & Chintha, C. Homology Modeling a Fast Tool for Drug Discovery: Current Perspectives. *Indian J. Pharm. Sci.* **74,** 1–17 (2012).
265. Methods for the Prediction of Protein-Ligand Binding Sites for Structure-Based Drug Design and Virtual Ligand Screening. *Current Protein & Peptide Science* **7,** 395–406 (2006).
266. Zuckerkandl, E. & Pauling, L. Evolutionary divergence and convergence in proteins. *Evol. Genes Proteins* 97–166 (1965).

267. Ashkenazy, H., Erez, E., Martz, E., Pupko, T. & Ben-Tal, N. ConSurf 2010: Calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res.* **38,** 529–533 (2010).
268. Nemoto, W. & Toh, H. Functional region prediction with a set of appropriate homologous sequences-an index for sequence selection by integrating structure and sequence information with spatial statistics. *BMC Struct. Biol.* **12,** 11 (2012).
269. Brylinski, M. & Skolnick, J. A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 129–134 (2008).
270. Konc, J. & Janežič, D. ProBiS algorithm for detection of structurally similar protein binding sites by local structural alignment. *Bioinformatics* **26,** 1160–1168 (2010).
271. Roy, A., Yang, J. & Zhang, Y. COFACTOR: an accurate comparative algorithm for structure-based protein function annotation. *Nucleic Acids Res.* 1–7 (2012).
272. Schmitt, S., Kuhn, D. & Klebe, G. A New Method to Detect Related Function Among Proteins Independent of Sequence and Fold Homology. *J. Mol. Biol.* **323,** 387–406 (2002).
273. Stark, A. & Russell, R. B. Annotation in three dimensions. PINTS: Patterns in Non-homologous Tertiary Structures. *Nucleic Acids Res.* **31,** 3341–3344 (2003).
274. Shulman-Peleg, A., Nussinov, R. & Wolfson, H. J. SiteEngines: recognition and comparison of binding sites and protein–protein interfaces. *Nucleic Acids Res.* **33,** W337–W341 (2005).
275. Kinoshita, K. & Nakamura, H. Identification of protein biochemical functions by similarity search using the molecular surface database eF-site. *Protein Sci.* **12,** 1589–1595 (2003).
276. Laskowski, R. A., Watson, J. D. & Thornton, J. M. ProFunc: a server for predicting protein function from 3D structure. *Nucleic Acids Res.* **33,** W89–W93 (2005).
277. Gold, N. D. & Jackson, R. M. SitesBase: a database for structure-based protein–ligand binding site comparisons. *Nucleic Acids Res.* **34,** D231–D234 (2006).
278. Porter, C. T., Bartlett, G. J. & Thornton, J. M. The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res.* **32,** D129–D133 (2004).
279. Waldherr-Teschner, M. *et al.* in *Advances in Scientific Visualization* (eds. Post, F. H. & Hin, A. J. S.) 58–67 (Springer Berlin Heidelberg, 1992).
280. Hendlich, M., Rippmann, F. & Barnickel, G. LIGSITE: Automatic and efficient detection of potential small molecule-binding sites in proteins. *J. Mol. Graph. Model.* **15,** 359–363 (1997).
281. 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.* **52,** 2287–2299 (2012).
282. Le Guilloux, V., Schmidtke, P. & Tuffery, P. Fpocket: An open source platform for ligand pocket detection. *BMC Bioinformatics* **10,** 168 (2009).
283. Tsodikov, O. V, Record, M. T. & Sergeev, Y. V. Novel computer program for fast exact calculation of accessible and molecular surface areas and average surface curvature. *J. Comput. Chem.* **23,** 600–9 (2002).
284. Laskowski, R. A. SURFNET: A program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J. Mol. Graph.* **13,** 323–330 (1995).
285. Rooklin, D., Wang, C., Katigbak, J., Arora, P. S. & Zhang, Y. AlphaSpace: Fragment-Centric Topographical Mapping To Target Protein–Protein Interaction Interfaces. *J.*

Chem. Inf. Model. **55,** 1585–1599 (2015).

286. Richards, F. M. The interpretation of protein structures: Total volume, group volume distributions and packing density. *J. Mol. Biol.* **82,** 1–14 (1974).
287. Ravindranath, P. A. & Sanner, M. F. AutoSite: an automated approach for pseudoligands prediction – From ligand binding sites identification to predicting key ligand atoms. *Bioinforma.* (2016).
288. Ngan, C.-H. *et al.* FTSite: high accuracy detection of ligand binding sites on unbound protein structures. *Bioinforma.* **28,** 286–287 (2012).
289. Huang, B. MetaPocket: A Meta Approach to Improve Protein Ligand Binding Site Prediction. *Omi. A J. Integr. Biol.* **13,** 325–330 (2009).
290. Brady, G. P. J. & Stouten, P. F. Fast prediction and visualization of protein binding pockets with PASS. *J. Comput. Aided. Mol. Des.* **14,** 383–401 (2000).
291. Laurie, A. T. R. & Jackson, R. M. Q-SiteFinder: an energy-based method for the prediction of protein–ligand binding sites. *Bioinforma.* **21,** 1908–1916 (2005).
292. Pirhadi, S., Shiri, F. & Ghasemi, J. B. Methods and applications of structure based pharmacophores in drug discovery. *Curr. Top. Med. Chem.* **13,** 1036–1047 (2013).
293. Thangavelu, K., Chong, Q. Y., Low, B. C. & Sivaraman, J. Structural Basis for the Active Site Inhibition Mechanism of Human Kidney-Type Glutaminase (KGA). *Sci. Rep.* **4,** 3827 (2014).
294. Pascarella, S. *et al.* The structure of serine hydroxymethyltransferase as modeled by homology and validated by site-directed mutagenesis. *Protein Sci.* **7,** 1976–1982 (1998).
295. Kumar, R., Suresh, M. & Priya, B. Pharmacophore modeling, *in silico* screening, molecular docking and molecular dynamics approaches for potential alpha-delta bungarotoxin-4 inhibitors discovery. *Pharmacogn. Mag.* **11,** 19–28 (2015).
296. Böhm, H.-J. The computer program LUDI: A new method for the de novo design of enzyme inhibitors. *J. Comput. Aided. Mol. Des.* **6,** 61–78 (1992).
297. Böhm, H.-J. LUDI: rule-based automatic design of new substituents for enzyme inhibitor leads. *J. Comput. Aided. Mol. Des.* **6,** 593–606 (1992).
298. Jorgensen, W. L. & Tirado–Rives, J. Molecular modeling of organic and biomolecular systems using BOSS and MCPRO. *J. Comput. Chem.* **26,** 1689–1700 (2005).
299. Güner, O. F. in *Pharmacophore Perception, Development, and Use in Drug Design* (International University Line, 2000).
300. Carlson, H. A. *et al.* Developing a Dynamic Pharmacophore Model for HIV-1 Integrase. *J. Med. Chem.* **43,** 2100–2114 (2000).
301. Khedkar, S. A., Malde, A. K., Coutinho, E. C. & Srivastava, S. Pharmacophore modeling in drug discovery and development: an overview. *Med. Chem. (Los. Angeles).* **3,** 187–197 (2007).
302. Kumar, S. P. & Jha, P. C. Multi-level structure-based pharmacophore modelling of caspase-3-non-peptide complexes: Extracting essential pharmacophore features and its application to virtual screening. *Chem. Biol. Interact.* **254,** 207–220 (2016).
303. BIOVIA Discovery Studio. Available at: http://accelrys.com/products/collaborative-science/biovia-discovery-studio/. (Accessed: 23rd January 2017)
304. Head, R. D. *et al.* VALIDATE: A New Method for the Receptor-Based Prediction of Binding Affinities of Novel Ligands. *J. Am. Chem. Soc.* **118,** 3959–3969 (1996).
305. Ortiz, A. R., Pisabarro, M. T., Gago, F. & Wade, R. C. Prediction of Drug Binding Affinities by Comparative Binding Energy Analysis. *J. Med. Chem.* **38,** 2681–2691 (1995).

306. Delano, W. L. The PyMOL Molecular Graphics System. (2002).
307. Gil-Redondo, R., Klett, J., Gago, F. & Morreale, A. gCOMBINE: A graphical user interface to perform structure-based comparative binding energy (COMBINE) analysis on a set of ligand-receptor complexes. *Proteins Struct. Funct. Bioinforma.* **78,** 162–172 (2010).
308. Liu, S., Fu, R., Cheng, X., Chen, S.-P. & Zhou, L.-H. Exploring the binding of BACE-1 inhibitors using comparative binding energy analysis (COMBINE). *BMC Struct. Biol.* **12,** 21 (2012).
309. 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.* **43,** 1591–1607 (2003).
310. Schneider, G. & Fechner, U. Computer-based de novo design of drug-like molecules. *Nat Rev Drug Discov* **4,** 649–663 (2005).
311. Danziger, D. J. & Dean, P. M. Automated site-directed drug design: a general algorithm for knowledge acquisition about hydrogen-bonding regions at protein surfaces. *Proc. R. Soc. London. Ser. B, Biol. Sci.* **236,** 101–113 (1989).
312. Gillet, V. J., Myatt, G., Zsoldos, Z. & Johnson, A. P. SPROUT, HIPPO and CAESA: Tools for de novo structure generation and estimation of synthetic accessibility. *Perspect. Drug Discov. Des.* **3,** 34–50 (1995).
313. Goodford, P. J. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.* **28,** 849–857 (1985).
314. Tomioka, N. & Itai, A. GREEN: a program package for docking studies in rational drug design. *J. Comput. Aided. Mol. Des.* **8,** 347–366 (1994).
315. Wang, R., Gao, Y. & Lai, L. LigBuilder: A Multi-Purpose Program for Structure-Based Drug Design. *Mol. Model. Annu.* **6,** 498–516 (2000).
316. Miranker, A. & Karplus, M. Functionality maps of binding sites: A multiple copy simultaneous search method. *Proteins Struct. Funct. Bioinforma.* **11,** 29–34 (1991).
317. Ishchenko, A. *et al.* Structure-Based Design Technology Contour and Its Application to the Design of Renin Inhibitors. *J. Chem. Inf. Model.* **52,** 2089–2097 (2012).
318. Jorgensen, W. L. *et al.* Computer-aided design of non-nucleoside inhibitors of HIV-1 reverse transcriptase. *Bioorg. Med. Chem. Lett.* **16,** 663–667 (2006).
319. Degen, J. & Rarey, M. FlexNovo: structure-based searching in large fragment spaces. *ChemMedChem* **1,** 854–868 (2006).
320. Moriaud, F. *et al.* Computational Fragment-Based Approach at PDB Scale by Protein Local Similarity. *J. Chem. Inf. Model.* **49,** 280–294 (2009).
321. Clark, D. E. *et al.* PRO-LIGAND: an approach to de novo molecular design. 1. Application to the design of organic molecules. *J. Comput. Aided. Mol. Des.* **9,** 13–32 (1995).
322. Bohacek, R. S. & McMartin, C. Multiple Highly Diverse Structures Complementary to Enzyme Binding Sites: Results of Extensive Application of a de Novo Design Method Incorporating Combinatorial Growth. *J. Am. Chem. Soc.* **116,** 5560–5571 (1994).
323. Lewis, R. A. Automated site-directed drug design: approaches to the formation of 3D molecular graphs. *J. Comput. Aided. Mol. Des.* **4,** 205–210 (1990).
324. Roe, D. C. & Kuntz, I. D. BUILDER v.2: Improving the chemistry of a de novo design strategy. *J. Comput. Aided. Mol. Des.* **9,** 269–282 (1995).

325. Lewis, R. A. *et al.* Automated site-directed drug design using molecular lattices. *J. Mol. Graph.* **10,** 66–78,106 (1992).
326. Pearlman, D. A. & Murcko, M. A. CONCERTS: Dynamic Connection of Fragments as an Approach to de Novo Ligand Design. *J. Med. Chem.* **39,** 1651–1663 (1996).
327. Zhu, J., Fan, H., Liu, H. & Shi, Y. Structure-based ligand design for flexible proteins: Application of new F-DycoBlock. *J. Comput. Aided. Mol. Des.* **15,** 979–996 (2001).
328. Pearlman, D. A. & Murcko, M. A. CONCEPTS: New dynamic algorithm for de novo drug suggestion. *J. Comput. Chem.* **14,** 1184–1193 (1993).
329. Zhu, J., Yu, H., Fan, H., Liu, H. & Shi, Y. Design of new selective inhibitors of cyclooxygenase-2 by dynamic assembly of molecular building blocks. *J. Comput. Aided. Mol. Des.* **15,** 447–463 (2001).
330. Liu, H., Duan, Z., Luo, Q. & Shi, Y. Structure-based ligand design by dynamically assembling molecular building blocks at binding site. *Proteins* **36,** 462–470 (1999).
331. Nishibata, Y. & Itai, A. Automatic creation of drug candidate structures based on receptor structure. Starting point for artificial lead generation. *Tetrahedron* **47,** 8985–8990 (1991).
332. Douguet, D., Thoreau, E. & Grassy, G. A genetic algorithm for the automated generation of small organic molecules: Drug design using an evolutionary algorithm. *J. Comput. Aided. Mol. Des.* **14,** 449–466 (2000).
333. Douguet, D., Munier-Lehmann, H., Labesse, G. & Pochet, S. LEA3D: A Computer-Aided Ligand Design for Structure-Based Drug Design. *J. Med. Chem.* **48,** 2457–2468 (2005).
334. Tschinke, V. & Cohen, N. C. The NEWLEAD program: a new method for the design of candidate structures from pharmacophoric hypotheses. *J. Med. Chem.* **36,** 3863–3870 (1993).
335. Cheng, T., Li, Q., Zhou, Z., Wang, Y. & Bryant, S. H. Structure-based virtual screening for drug discovery: a problem-centric review. *AAPS J.* **14,** 133–41 (2012).
336. Fischer, E. Einfluss der Configuration auf die Wirkung der Enzyme. *Berichte der Dtsch. Chem. Gesellschaft* **27,** 2985–2993 (1894).
337. Kuntz, I. D., Blaney, J. M., Oatley, S. J., Langridge, R. & Ferrin, T. E. A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.* **161,** 269–288 (1982).
338. 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.* **8,** 153–174 (1994).
339. McGann, M. FRED Pose Prediction and Virtual Screening Accuracy. *J. Chem. Inf. Model.* **51,** 578–596 (2011).
340. Vogt, A. D. & Di Cera, E. Conformational Selection Is a Dominant Mechanism of Ligand Binding. *Biochemistry* **52,** 5723–5729 (2013).
341. Boehr, D. D., Nussinov, R. & Wright, P. E. The role of dynamic conformational ensembles in biomolecular recognition. *Nat. Chem. Biol.* **5,** 789–796 (2009).
342. James, L. C. & Tawfik, D. S. Conformational diversity and protein evolution--a 60-year-old hypothesis revisited. *Trends Biochem. Sci.* **28,** 361–368 (2003).
343. Brooijmans, N. & Kuntz, I. D. Molecular recognition and docking algorithms. *Annu. Rev. Biophys. Biomol. Struct.* **32,** 335–373 (2003).
344. Kitchen, D. B., Decornez, H., Furr, J. R. & Bajorath, J. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat Rev Drug Discov* **3,** 935–949 (2004).

345. Friesner, R. A. *et al.* Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* **47,** 1739–1749 (2004).
346. Welch, W., Ruppert, J. & Jain, A. N. Hammerhead: fast, fully automated docking of flexible ligands to protein binding sites. *Chem. Biol.* **3,** 449–462 (1996).
347. Lang, P. T. *et al.* DOCK 6: combining techniques to model RNA-small molecule complexes. *RNA* **15,** 1219–1230 (2009).
348. Rarey, M., Kramer, B., Lengauer, T. & Klebe, G. A Fast Flexible Docking Method using an Incremental Construction Algorithm. *J. Mol. Biol.* **261,** 470–489 (1996).
349. Jain, A. N. Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine. *J. Med. Chem.* **46,** 499–511 (2003).
350. Jain, A. N. Surflex-Dock 2.1: robust performance from ligand energetic modeling, ring flexibility, and knowledge-based search. *J. Comput. Aided. Mol. Des.* **21,** 281–306 (2007).
351. Jain, A. N. Morphological similarity: a 3D molecular similarity method correlated with protein-ligand recognition. *J. Comput. Aided. Mol. Des.* **14,** 199–213 (2000).
352. Trosset, J.-Y. & Scheraga, H. A. Reaching the global minimum in docking simulations: A Monte Carlo energy minimization approach using Bezier splines. *Proc. Natl. Acad. Sci. U. S. A.* **95,** 8011–8015 (1998).
353. Liu, M. & Wang, S. MCDOCK: A Monte Carlo simulation approach to the molecular docking problem. *J. Comput. Aided. Mol. Des.* **13,** 435–451 (1999).
354. Trosset, J.-Y. & Scheraga, H. A. Prodock: Software package for protein modeling and docking. *J. Comput. Chem.* **20,** 412–427 (1999).
355. Singh, T., Biswas, D. & Jayaram, B. AADS--an automated active site identification, docking, and scoring protocol for protein targets based on physicochemical descriptors. *J. Chem. Inf. Model.* **51,** 2515–2527 (2011).
356. Mcmartin, C. & Bohacek, R. S. QXP: Powerful, rapid computer algorithms for structure-based drug design. *J. Comput. Aided. Mol. Des.* **11,** 333–344 (1997).
357. Gupta, A., Gandhimathi, A. & Jayaram, P. S. and B. ParDOCK: An All Atom Energy Based Monte Carlo Docking Protocol for Protein-Ligand Complexes. *Protein & Peptide Letters* **14,** 632–646 (2007).
358. Hart, T. N. & Read, R. J. A multiple-start Monte Carlo docking method. *Proteins Struct. Funct. Bioinforma.* **13,** 206–222 (1992).
359. Abagyan, R., Totrov, M. & Kuznetsov, D. ICM - A new method for protein modeling and design: Applications to docking and structure prediction from the distorted native conformation. *J. Comput. Chem.* **15,** 488–506 (1994).
360. Darwin, C. *On the origin of species*. (1859).
361. Morris, G. M. *et al.* AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility. *J. Comput. Chem.* **30,** 2785–2791 (2009).
362. Paquet, E. & Viktor, H. L. Molecular dynamics, monte carlo simulations, and langevin dynamics: a computational review. *Biomed Res. Int.* **2015,** 183918 (2015).
363. Nichols, S. E., Baron, R., Ivetac, A. & McCammon, J. A. Predictive Power of Molecular Dynamics Receptor Structures in Virtual Screening. *J. Chem. Inf. Model.* **51,** 1439–1446 (2011).
364. González, M. A. Force fields and molecular dynamics simulations. *JDN* **12,** 169–200 (2011).
365. Weiner, S. J. *et al.* A new force field for molecular mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc.* **106,** 765–784 (1984).

366. Weiner, P. K. & Kollman, P. A. AMBER: Assisted model building with energy refinement. A general program for modeling molecules and their interactions. *J. Comput. Chem.* **2,** 287–303 (1981).
367. Clark, M., Cramer, R. D. & Van Opdenbosch, N. Validation of the general purpose tripos 5.2 force field. *J. Comput. Chem.* **10,** 982–1012 (1989).
368. Nemethy, G. *et al.* Energy parameters in polypeptides. 10. Improved geometrical parameters and nonbonded interactions for use in the ECEPP/3 algorithm, with application to proline-containing peptides. *J. Phys. Chem.* **96,** 6472–6484 (1992).
369. Sousa, S. F., Fernandes, P. A. & Ramos, M. J. Protein–ligand docking: Current status and future challenges. *Proteins Struct. Funct. Bioinforma.* **65,** 15–26 (2006).
370. Eldridge, M. D., Murray, C. W., Auton, T. R., Paolini, G. V & Mee, R. P. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput. Aided. Mol. Des.* **11,** 425–445 (1997).
371. Huo, S., Wang, J., Cieplak, P., Kollman, P. A. & Kuntz, I. D. Molecular dynamics and free energy analyses of cathepsin D-inhibitor interactions: insight into structure-based ligand design. *J. Med. Chem.* **45,** 1412–1419 (2002).
372. Rognan, D., Lauemøller, S. L., Holm, A., Buus, S. & Tschinke, V. 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 (1999).
373. Kollman, P. A. *et al.* Calculating Structures and Free Energies of Complex Molecules: Combining Molecular Mechanics and Continuum Models. *Acc. Chem. Res.* **33,** 889–897 (2000).
374. Li, G.-B., Yang, L.-L., Wang, W.-J., Li, L.-L. & Yang, S.-Y. ID-Score: A New Empirical Scoring Function Based on a Comprehensive Set of Descriptors Related to Protein–Ligand Interactions. *J. Chem. Inf. Model.* **53,** 592–600 (2013).
375. Muegge, I. Effect of ligand volume correction on PMF scoring. *J. Comput. Chem.* **22,** 418–425 (2001).
376. Muegge, I. & Martin, Y. C. A General and Fast Scoring Function for Protein−Ligand Interactions: A Simplified Potential Approach. *J. Med. Chem.* **42,** 791–804 (1999).
377. Zheng, M. *et al.* Knowledge-Based Scoring Functions in Drug Design: 3. A Two-Dimensional Knowledge-Based Hydrogen-Bonding Potential for the Prediction of Protein–Ligand Interactions. *J. Chem. Inf. Model.* **51,** 2994–3004 (2011).
378. Muegge, I. PMF Scoring Revisited. *J. Med. Chem.* **49,** 5895–5902 (2006).
379. Gohlke, H., Hendlich, M. & Klebe, G. Knowledge-based scoring function to predict protein-ligand interactions. *J. Mol. Biol.* **295,** 337–356 (2000).
380. DeWitte, R. S. & Shakhnovich, E. I. SMoG: de Novo Design Method Based on Simple, Fast, and Accurate Free Energy Estimates. 1. Methodology and Supporting Evidence. *J. Am. Chem. Soc.* **118,** 11733–11744 (1996).
381. Feher, M. Consensus scoring for protein-ligand interactions. *Drug Discov. Today* **11,** 421–428 (2006).
382. 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.* **42,** 5100–5109 (1999).
383. Wang, R. & Wang, S. How Does Consensus Scoring Work for Virtual Library Screening? An Idealized Computer Experiment. *J. Chem. Inf. Comput. Sci.* **41,** 1422–1426 (2001).

384. Terp, G. E., Johansen, B. N., Christensen, I. T. & Jørgensen, F. S. A New Concept for Multidimensional Selection of Ligand Conformations (MultiSelect) and Multidimensional Scoring (MultiScore) of Protein−Ligand Binding Affinities. *J. Med. Chem.* **44,** 2333–2343 (2001).
385. Jacobsson, M., Lidén, P., Stjernschantz, E., Boström, H. & Norinder, U. Improving Structure-Based Virtual Screening by Multivariate Analysis of Scoring Data. *J. Med. Chem.* **46,** 5781–5789 (2003).
386. Chen, Y.-C. Beware of docking! *Trends Pharmacol. Sci.* **36,** 78–95 (2015).
387. Pei, J. *et al.* PSI-DOCK: Towards highly efficient and accurate flexible ligand docking. *Proteins Struct. Funct. Bioinforma.* **62,** 934–946 (2006).
388. Baxter, C. A., Murray, C. W., Clark, D. E., Westhead, D. R. & Eldridge, M. D. Flexible docking using Tabu search and an empirical estimate of binding affinity. *Proteins* **33,** 367–382 (1998).
389. Antunes, D. A., Devaurs, D. & Kavraki, L. E. Understanding the challenges of protein flexibility in drug design. *Expert Opin. Drug Discov.* **10,** 1301–1313 (2015).
390. Leach, A. R. Ligand docking to proteins with discrete side-chain flexibility. *J. Mol. Biol.* **235,** 345–356 (1994).
391. Tuffery, P., Etchebest, C., Hazout, S. & Lavery, R. A new approach to the rapid determination of protein side chain conformations. *J. Biomol. Struct. Dyn.* **8,** 1267–1289 (1991).
392. Lovell, S. C., Word, J. M., Richardson, J. S. & Richardson, D. C. The penultimate rotamer library. *Proteins* **40,** 389–408 (2000).
393. De Maeyer, M., Desmet, J. & Lasters, I. All in one: a highly detailed rotamer library improves both accuracy and speed in the modelling of sidechains by dead-end elimination. *Fold. Des.* **2,** 53–66 (1997).
394. Ponder, J. W. & Richards, F. M. Tertiary templates for proteins. *J. Mol. Biol.* **193,** 775–791 (1987).
395. Claußen, H., Buning, C., Rarey, M. & Lengauer, T. FlexE: efficient molecular docking considering protein structure variations1 . *J. Mol. Biol.* **308,** 377–395 (2001).
396. Huang, S.-Y. & Zou, X. Ensemble docking of multiple protein structures: Considering protein structural variations in molecular docking. *Proteins Struct. Funct. Bioinforma.* **66,** 399–421 (2007).
397. Ellingson, S. R., Miao, Y., Baudry, J. & Smith, J. C. Multi-Conformer Ensemble Docking to Difficult Protein Targets. *J. Phys. Chem. B* **119,** 1026–1034 (2015).
398. Jiang, F. & Kim, S.-H. 'Soft docking': Matching of molecular surface cubes. *J. Mol. Biol.* **219,** 79–102 (1991).
399. Ferrari, A. M., Wei, B. Q., Costantino, L. & Shoichet, B. K. Soft Docking and Multiple Receptor Conformations in Virtual Screening. *J. Med. Chem.* **47,** 5076–5084 (2004).
400. Reulecke, I., Lange, G., Albrecht, J., Klein, R. & Rarey, M. Towards an Integrated Description of Hydrogen Bonding and Dehydration: Decreasing False Positives in Virtual Screening with the HYDE Scoring Function. *ChemMedChem* **3,** 885–897 (2008).
401. Srinivasan, J., Cheatham, T. E., Cieplak, P., Kollman, P. A. & Case, D. A. Continuum Solvent Studies of the Stability of DNA, RNA, and Phosphoramidate−DNA Helices. *J. Am. Chem. Soc.* **120,** 9401–9409 (1998).
402. Still, W. C., Tempczyk, A., Hawley, R. C. & Hendrickson, T. Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.* **112,** 6127–6129 (1990).

403. Wang, W., Donini, O., Reyes, C. M. & Kollman, P. A. Biomolecular simulations: recent developments in force fields, simulations of enzyme catalysis, protein-ligand, protein-protein, and protein-nucleic acid noncovalent interactions. *Annu. Rev. Biophys. Biomol. Struct.* **30,** 211–243 (2001).
404. Genheden, S. & Ryde, U. The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opin. Drug Discov.* **10,** 449–461 (2015).
405. Michel, J., Verdonk, M. L. & Essex, J. W. Protein-Ligand Binding Affinity Predictions by Implicit Solvent Simulations: A Tool for Lead Optimization? *J. Med. Chem.* **49,** 7427–7439 (2006).
406. Liu, H.-Y., Kuntz, I. D. & Zou, X. Pairwise GB/SA Scoring Function for Structure-based Drug Design. *J. Phys. Chem. B* **108,** 5453–5462 (2004).
407. Homeyer, N. & Gohlke, H. Free Energy Calculations by the Molecular Mechanics Poisson−Boltzmann Surface Area Method. *Mol. Inform.* **31,** 114–122 (2012).
408. Klebe, G. Applying thermodynamic profiling in lead finding and optimization. *Nat Rev Drug Discov* **14,** 95–110 (2015).
409. Liebeschuetz, J. W. Evaluating docking programs: keeping the playing field level. *J. Comput. Aided. Mol. Des.* **22,** 229–238 (2008).
410. Ladbury, J. E. Just add water! The effect of water on the specificity of protein-ligand binding sites and its potential application to drug design. *Chem. Biol.* **3,** 973–980 (1996).
411. Ruiz-Carmona, S. *et al.* rDock: A Fast, Versatile and Open Source Program for Docking Ligands to Proteins and Nucleic Acids. *PLOS Comput. Biol.* **10,** e1003571 (2014).
412. Schnecke, V. & Kuhn, L. A. Virtual screening with solvation and ligand-induced complementarity. *Perspect. Drug Discov. Des.* **20,** 171–190 (2000).
413. Alex, A. A. & Millan, D. S. in *Drug Design Strategies: Quantitative Approaches* 108–163 (The Royal Society of Chemistry, 2012). doi:10.1039/9781849733410-00108
414. Shimbo, I., Nakajima, R., Yokoyama, S. & Sumikura, K. Patent protection for protein structure analysis. *Nat Biotech* **22,** 109–112 (2004).
415. Singh, J. *et al.* Molecular model for VLA-4 inhibitors. (1998).
416. Hart, T. & Quibell, M. Hepatitis C NS3 protease inhibitors. (1998).
417. Gupta, R. K., Bhattacharjee, A. K. & Lee, D. M. Arthropod repellent pharmacophore models, compounds identified as fitting the pharmacophore models, and methods of making and using thereof. (2011).
418. Baldwin, J. J. *et al.* Thienothiopyran-2-sulfonamides: novel topically active carbonic anhydrase inhibitors for the treatment of glaucoma. *J. Med. Chem.* **32,** 2510–2513 (1989).
419. Roberts, N. A. *et al.* Rational design of peptide-based HIV proteinase inhibitors. *Science (80-.).* **248,** 358 LP-361 (1990).
420. Kempf, D. J. *et al.* Discovery of Ritonavir, a Potent Inhibitor of HIV Protease with High Oral Bioavailability and Clinical Efficacy. *J. Med. Chem.* **41,** 602–617 (1998).
421. Dorsey, B. D. *et al.* L-735,524: The Design of a Potent and Orally Bioavailable HIV Protease Inhibitor. *J. Med. Chem.* **37,** 3443–3451 (1994).
422. Silver, L. H. Dose-response evaluation of the ocular hypotensive effect of brinzolamide ophthalmic suspension (Azopt). Brinzolamide Dose-Response Study Group. *Surv. Ophthalmol.* **44 Suppl 2,** S147-53 (2000).
423. Kaldor, S. W. *et al.* Viracept (Nelfinavir Mesylate, AG1343): A Potent, Orally Bioavailable Inhibitor of HIV-1 Protease. *J. Med. Chem.* **40,** 3979–3985 (1997).
424. Kim, E. E. *et al.* Crystal structure of HIV-1 protease in complex with VX-478, a potent

and orally bioavailable inhibitor of the enzyme. *J. Am. Chem. Soc.* **117,** 1181–1182 (1995).

425. Sham, H. L. *et al.* ABT-378, a highly potent inhibitor of the human immunodeficiency virus protease. *Antimicrob. Agents Chemother.* **42,** 3218–3224 (1998).
426. von Itzstein, M. *et al.* Rational design of potent sialidase-based inhibitors of influenza virus replication. *Nature* **363,** 418–423 (1993).
427. Kim, C. U. *et al.* Structure-activity relationship studies of novel carbocyclic influenza neuraminidase inhibitors. *J. Med. Chem.* **41,** 2451–2460 (1998).
428. Zimmermann, J. *et al.* Phenylamino-pyrimidine (PAP) — derivatives: a new class of potent and highly selective PDGF-receptor autophosphorylation inhibitors. *Bioorg. Med. Chem. Lett.* **6,** 1221–1226 (1996).
429. Zimmermann, J., Buchdunger, E., Mett, H., Meyer, T. & Lydon, N. B. Potent and selective inhibitors of the Abl-kinase: phenylamino-pyrimidine (PAP) derivatives. *Bioorg. Med. Chem. Lett.* **7,** 187–192 (1997).
430. Hennequin, L. F. *et al.* Design and Structure−Activity Relationship of a New Class of Potent VEGF Receptor Tyrosine Kinase Inhibitors. *J. Med. Chem.* **42,** 5369–5389 (1999).
431. Robinson, B. S. *et al.* BMS-232632, a highly potent human immunodeficiency virus protease inhibitor that can be used in combination with other available antiretroviral agents. *Antimicrob. Agents Chemother.* **44,** 2093–2099 (2000).
432. Becker, S. & Thornton, L. Fosamprenavir: advancing HIV protease inhibitor treatment options. *Expert Opin. Pharmacother.* **5,** 1995–2005 (2004).
433. Pollack, V. A. *et al.* Inhibition of epidermal growth factor receptor-associated tyrosine phosphorylation in human carcinomas with CP-358,774: dynamics of receptor inhibition in situ and antitumor effects in athymic mice. *J. Pharmacol. Exp. Ther.* **291,** 739–748 (1999).
434. Wilhelm, S. M. *et al.* BAY 43-9006 exhibits broad spectrum oral antitumor activity and targets the RAF/MEK/ERK pathway and receptor tyrosine kinases involved in tumor progression and angiogenesis. *Cancer Res.* **64,** 7099–7109 (2004).
435. Turner, S. R. *et al.* Tipranavir (PNU-140690): A Potent, Orally Bioavailable Nonpeptidic HIV Protease Inhibitor of the 5,6-Dihydro-4-hydroxy-2-pyrone Sulfonamide Class. *J. Med. Chem.* **41,** 3467–3476 (1998).
436. Oh, T. Y., Kang, K. K., Ahn, B. O., Yoo, M. & Kim, W. B. Erectogenic effect of the selective phosphodiesterase type 5 inhibitor, DA-8159. *Arch. Pharm. Res.* **23,** 471–476 (2000).
437. Ahn, B. O. *et al.* Efficacy of DA-8159, a new PDE5 inhibitor, for inducing penile erection in rabbits with acute spinal cord injury. *Int. J. Impot. Res.* **15,** 405–411 (2003).
438. Mendel, D. B. *et al.* In vivo antitumor activity of SU11248, a novel tyrosine kinase inhibitor targeting vascular endothelial growth factor and platelet-derived growth factor receptors: determination of a pharmacokinetic/pharmacodynamic relationship. *Clin. Cancer Res.* **9,** 327–337 (2003).
439. Ghosh, A. K., Dawson, Z. L. & Mitsuya, H. Darunavir, a conceptually new HIV-1 protease inhibitor for the treatment of drug-resistant HIV. *Bioorg. Med. Chem.* **15,** 7576–7580 (2007).
440. Kelly, W. K. *et al.* Phase I study of an oral histone deacetylase inhibitor, suberoylanilide hydroxamic acid, in patients with advanced cancer. *J. Clin. Oncol.* **23,** 3923–3931 (2005).
441. Lombardo, L. J. *et al.* Discovery of N-(2-Chloro-6-methyl-phenyl)-2-(6-(4-(2-

hydroxyethyl)-piperazin-1-yl)-2-methylpyrimidin-4-ylamino)thiazole-5-carboxamide (BMS-354825), a Dual Src/Abl Kinase Inhibitor with Potent Antitumor Activity in Preclinical Assays. *J. Med. Chem.* **47,** 6658–6661 (2004).

442. Weisberg, E. *et al.* Characterization of AMN107, a selective inhibitor of native and mutant Bcr-Abl. *Cancer Cell* **7,** 129–141 (2005).
443. Rahuel, J. *et al.* Structure-based drug design: the discovery of novel nonpeptide orally active inhibitors of human renin. *Chem. Biol.* **7,** 493–504 (2000).
444. Rusnak, D. W. *et al.* The effects of the novel, reversible epidermal growth factor receptor/ErbB-2 tyrosine kinase inhibitor, GW2016, on the growth of human normal and tumor-derived cell lines in vitro and in vivo. *Mol. Cancer Ther.* **1,** 85–94 (2001).
445. Roehrig, S. *et al.* Discovery of the novel antithrombotic agent 5-chloro-N-({(5S)-2-oxo-3-[4-(3-oxomorpholin-4-yl)phenyl]-1,3-oxazolidin-5-yl}methyl)thiophene-2-carboxamide (BAY 59-7939): an oral, direct factor Xa inhibitor. *J. Med. Chem.* **48,** 5900–5908 (2005).
446. Sorbera, L. A., Bozzo, J. & Castaner, J. Dabigatran/Dabigatran etexilate: Pevention of DVT prevention of ischemic stroke thrombin inhibitor. *Drugs Future* **30,** 877–885 (2005).
447. Das, K. *et al.* Roles of conformational and positional adaptability in structure-based design of TMC125-R165335 (etravirine) and related non-nucleoside reverse transcriptase inhibitors that are highly potent and effective against wild-type and drug-resistant HIV-1 varia. *J. Med. Chem.* **47,** 2550–2560 (2004).
448. Sleijfer, S. *et al.* Pazopanib, a multikinase angiogenesis inhibitor, in patients with relapsed or refractory advanced soft tissue sarcoma: a phase II study from the European organisation for research and treatment of cancer-soft tissue and bone sarcoma group (EORTC study 620. *J. Clin. Oncol.* **27,** 3126–3132 (2009).
449. Inglese, J., Shamu, C. E. & Guy, R. K. Reporting data from high-throughput screening of small-molecule libraries. *Nat Chem Biol* **3,** 438–441 (2007).
450. Hawkins, P. C. D., Warren, G. L., Skillman, A. G. & Nicholls, A. How to do an evaluation: pitfalls and traps. *J. Comput. Aided. Mol. Des.* **22,** 179–190 (2008).
451. Levitt, M. A simplified representation of protein conformations for rapid simulation of protein folding. *J. Mol. Biol.* **104,** 59–107 (1976).
452. Nishikawa, K. & Ooi, T. Tertiary Structure of Proteins. II. Freedom of Dihedral Angles and Energy Calculation. *J. Phys. Soc. Japan* **32,** 1338–1347 (1972).
453. Abagyan, R. A. & Totrov, M. M. Contact area difference (CAD): a robust measure to evaluate accuracy of protein models. *J. Mol. Biol.* **268,** 678–685 (1997).
454. Yusuf, D., Davis, A. M., Kleywegt, G. J. & Schmitt, S. An Alternative Method for the Evaluation of Docking Performance: RSR vs RMSD. *J. Chem. Inf. Model.* **48,** 1411–1422 (2008).
455. 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.* **22,** 213–228 (2008).
456. Allen, W. J. & Rizzo, R. C. Implementation of the Hungarian Algorithm to Account for Ligand Symmetry and Similarity in Structure-Based Design. *J. Chem. Inf. Model.* **54,** 518–529 (2014).
457. Trott, O. & Olson, A. J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **31,** 455–461 (2010).

458. Kuhn, H. W. The Hungarian method for the assignment problem. *Nav. Res. Logist. Q.* **2,** 83–97 (1955).
459. Bissantz, C., Folkers, G. & Rognan, D. Protein-Based Virtual Screening of Chemical Databases. 1. Evaluation of Different Docking/Scoring Combinations. *J. Med. Chem.* **43,** 4759–4767 (2000).
460. 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* **0,** 146–157 (2015).
461. Tiikkainen, P., Bellis, L., Light, Y. & Franke, L. Estimating Error Rates in Bioactivity Databases. *J. Chem. Inf. Model.* **53,** 2499–2505 (2013).
462. 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.* **55,** 1297–1307 (2015).
463. Verdonk, M. L. *et al.* Virtual Screening Using Protein - Ligand Docking : Avoiding Artificial Enrichment. *J. Chem. Inf. Comput. Sci* **44,** 793–806 (2004).
464. Wallach, I. & Lilien, R. Virtual Decoy Sets for Molecular Docking Benchmarks. *J. Chem. Inf. Model.* **51,** 196–202 (2011).
465. Novikov, F. N. *et al.* CSAR Scoring Challenge Reveals the Need for New Concepts in Estimating Protein–Ligand Binding Affinity. *J. Chem. Inf. Model.* **51,** 2090–2096 (2011).
466. Community Structure-Activity Resource (CSAR). Available at: http://www.csardock.org. (Accessed: 8th March 2017)
467. Drug Design Data Resource (D3R). Available at: http://drugdesigndata.org. (Accessed: 8th March 2017)
468. Triballeau, N., Acher, F., Brabet, I., Pin, J.-P. & Bertrand, H. 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 (2005).
469. Truchon, J. & Bayly, C. Evaluating virtual screening methods: good and bad metrics for the 'early recognition' problem. *J. Chem. Inf. Model.* **47,** 488–508 (2007).
470. 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. Model.* **41,** 1395–1406 (2001).
471. Pepe, M. S. *et al.* Integrating the predictiveness of a marker with its performance as a classifier. *Am. J. Epidemiol.* **167,** 362–368 (2008).
472. Huang, Y., Sullivan Pepe, M. & Feng, Z. Evaluating the predictiveness of a continuous marker. *Biometrics* **63,** 1181–1188 (2007).
473. Wang, Y. *et al.* PubChem BioAssay: 2014 update. *Nucleic Acids Res.* **42,** D1075–D1082 (2014).
474. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410 (1990).
475. Koishi, M. *et al.* Quercetin, an Inhibitor of Heat Shock Protein Synthesis, Inhibits the Acquisition of Thermotolerance in a Human Colon Carcinoma Cell Line. *Japanese J. Cancer Res.* **83,** 1216–1222 (1992).
476. Hosokawa, N. *et al.* Flavonoids inhibit the expression of heat shock proteins. *Cell Struct. Funct.* **15,** 393–401 (1990).
477. Cassel, J. A., Ilyin, S., McDonnell, M. E. & Reitz, A. B. Novel inhibitors of heat shock

protein Hsp70-mediated luciferase refolding that bind to DnaJ. *Bioorg. Med. Chem.* **20,** 3609–3614 (2012).

478. Richardson, P. G. *et al.* Inhibition of heat shock protein 90 (HSP90) as a therapeutic strategy for the treatment of myeloma and other cancers. *Br. J. Haematol.* **152,** 367–379 (2011).
479. Sharma, S. V, Agatsuma, T. & Nakano, H. Targeting of the protein chaperone, HSP90, by the transformation suppressing agent, radicicol. *Oncogene* **16,** 2639–2645 (1998).
480. Whitesell, L. & Lindquist, S. L. HSP90 and the chaperoning of cancer. *Nat. Rev. Cancer* **5,** 761–772 (2005).
481. Tian, Z.-Q. *et al.* Synthesis and biological activities of novel 17-aminogeldanamycin derivatives. *Bioorg. Med. Chem.* **12,** 5317–5329 (2004).
482. Sharp, S. Y. *et al.* Inhibition of the heat shock protein 90 molecular chaperone in vitro and in vivo by novel, synthetic, potent resorcinylic pyrazole/isoxazole amide analogues. *Mol. Cancer Ther.* **6,** 1198–1211 (2007).
483. Sharp, S. Y. *et al.* In vitro biological characterization of a novel, synthetic diaryl pyrazole resorcinol class of heat shock protein 90 inhibitors. *Cancer Res.* **67,** 2206–2216 (2007).
484. Bussenius, J. *et al.* Discovery of XL888: a novel tropane-derived small molecule inhibitor of HSP90. *Bioorg. Med. Chem. Lett.* **22,** 5396–5404 (2012).
485. Gopalsamy, A. *et al.* Discovery of Benzisoxazoles as Potent Inhibitors of Chaperone Heat Shock Protein 90. *J. Med. Chem.* **51,** 373–375 (2008).
486. Donnelly, A. & Blagg, B. S. J. Novobiocin and additional inhibitors of the Hsp90 C-terminal nucleotide-binding pocket. *Curr. Med. Chem.* **15,** 2702–2717 (2008).
487. Williamson, D. S. *et al.* Novel Adenosine-Derived Inhibitors of 70 kDa Heat Shock Protein, Discovered Through Structure-Based Design. *J. Med. Chem.* **52,** 1510–1513 (2009).
488. Ko, S.-K. *et al.* A small molecule inhibitor of ATPase activity of HSP70 induces apoptosis and has antitumor activities. *Chem. Biol.* **22,** 391–403 (2015).
489. Hanson, B. E. & Vesole, D. H. Retaspimycin hydrochloride (IPI-504): a novel heat shock protein inhibitor as an anticancer agent. *Expert Opin. Investig. Drugs* **18,** 1375–1383 (2009).
490. Braunstein, M. J. *et al.* Antimyeloma Effects of the Heat Shock Protein 70 Molecular Chaperone Inhibitor MAL3-101. *J. Oncol.* **2011,** 232037 (2011).
491. Yi, F. & Regan, L. A Novel Class of Small Molecule Inhibitors of Hsp90. *ACS Chem. Biol.* **3,** 645–654 (2008).
492. Leu, J. I.-J., Pimkina, J., Pandey, P., Murphy, M. E. & George, D. L. HSP70 inhibition by the small-molecule 2-phenylethynesulfonamide impairs protein clearance pathways in tumor cells. *Mol. Cancer Res.* **9,** 936–947 (2011).
493. Blagosklonny, M. V. Hsp-90-associated oncoproteins: multiple targets of geldanamycin and its analogs. *Leukemia* **16,** 455–462 (2002).
494. Brough, P. A. *et al.* 4,5-Diarylisoxazole Hsp90 Chaperone Inhibitors: Potential Therapeutic Agents for the Treatment of Cancer. *J. Med. Chem.* **51,** 196–218 (2008).
495. Brandt, G. E. L., Schmidt, M. D., Prisinzano, T. E. & Blagg, B. S. J. Gedunin, a novel hsp90 inhibitor: semisynthesis of derivatives and preliminary structure-activity relationships. *J. Med. Chem.* **51,** 6495–6502 (2008).
496. Menezes, D. L. *et al.* The novel oral Hsp90 inhibitor NVP-HSP990 exhibits potent and broad-spectrum antitumor activities in vitro and in vivo. *Mol. Cancer Ther.* **11,** 730–739 (2012).

497. Martin, C. J. *et al.* Molecular characterization of macbecin as an Hsp90 inhibitor. *J. Med. Chem.* **51,** 2853–2857 (2008).
498. McCleese, J. K. *et al.* The novel HSP90 inhibitor STA-1474 exhibits biologic activity against osteosarcoma cell lines. *Int. J. cancer* **125,** 2792–2801 (2009).
499. Wang, Y., Trepel, J. B., Neckers, L. M. & Giaccone, G. STA-9090, a small-molecule Hsp90 inhibitor for the potential treatment of cancer. *Curr. Opin. Investig. Drugs* **11,** 1466–1476 (2010).
500. Ying, W. *et al.* Ganetespib, a unique triazolone-containing Hsp90 inhibitor, exhibits potent antitumor activity and a superior safety profile for cancer therapy. *Mol. Cancer Ther.* **11,** 475–484 (2012).
501. Lin, T.-Y. *et al.* The novel HSP90 inhibitor STA-9090 exhibits activity against Kit-dependent and -independent malignant mast cell tumors. *Exp. Hematol.* **36,** 1266–1277 (2008).
502. Nakashima, T. *et al.* New molecular and biological mechanism of antitumor activities of KW-2478, a novel nonansamycin heat shock protein 90 inhibitor, in multiple myeloma cells. *Clin. Cancer Res.* **16,** 2792–2802 (2010).
503. Massey, A. J. *et al.* Preclinical antitumor activity of the orally available heat shock protein 90 inhibitor NVP-BEP800. *Mol. Cancer Ther.* **9,** 906–919 (2010).
504. Lundgren, K. *et al.* BIIB021, an orally available, fully synthetic small-molecule inhibitor of the heat shock protein Hsp90. *Mol. Cancer Ther.* **8,** 921–929 (2009).
505. Boll, B. *et al.* Heat shock protein 90 inhibitor BIIB021 (CNF2024) depletes NF-kappaB and sensitizes Hodgkin's lymphoma cells for natural killer cell-mediated cytotoxicity. *Clin. Cancer Res.* **15,** 5108–5116 (2009).
506. Yin, X. *et al.* BIIB021, a novel Hsp90 inhibitor, sensitizes head and neck squamous cell carcinoma to radiotherapy. *Int. J. cancer* **126,** 1216–1225 (2010).
507. Zhang, H. *et al.* BIIB021, a synthetic Hsp90 inhibitor, has broad application against tumors with acquired multidrug resistance. *Int. J. cancer* **126,** 1226–1234 (2010).
508. Graham, B. *et al.* The heat shock protein 90 inhibitor, AT13387, displays a long duration of action in vitro and in vivo in non-small cell lung cancer. *Cancer Sci.* **103,** 522–527 (2012).
509. Mandler, R., Kobayashi, H., Hinson, E. R., Brechbiel, M. W. & Waldmann, T. A. Herceptin-geldanamycin immunoconjugates: pharmacokinetics, biodistribution, and enhanced antitumor activity. *Cancer Res.* **64,** 1460–1467 (2004).
510. Mandler, R., Kobayashi, H., Davis, M. Y., Waldmann, T. A. & Brechbiel, M. W. Modifications in synthesis strategy improve the yield and efficacy of geldanamycin-herceptin immunoconjugates. *Bioconjug. Chem.* **13,** 786–791 (2002).
511. Mandler, R. *et al.* Immunoconjugates of Geldanamycin and Anti-HER2 Monoclonal Antibodies: Antiproliferative Activity on Human Breast Carcinoma Cell Lines. *JNCI J. Natl. Cancer Inst.* **92,** 1573–1581 (2000).
512. Wisén, S. *et al.* Binding of a Small Molecule at a Protein–Protein Interface Regulates the Chaperone Activity of Hsp70–Hsp40. *ACS Chem. Biol.* **5,** 611–622 (2010).
513. Ishida, R. *et al.* Cisplatin differently affects amino terminal and carboxyl terminal domains of HSP90. *FEBS Lett.* **582,** 3879–3883 (2008).
514. Palermo, C. M., Westlake, C. A. & Gasiewicz, T. A. Epigallocatechin gallate inhibits aryl hydrocarbon receptor gene transcription through an indirect mechanism involving binding to a 90 kDa heat shock protein. *Biochemistry* **44,** 5041–5052 (2005).
515. Okawa, Y. *et al.* SNX-2112, a selective Hsp90 inhibitor, potently inhibits tumor cell growth, angiogenesis, and osteoclastogenesis in multiple myeloma and other

hematologic tumors by abrogating signaling via Akt and ERK. *Blood* **113,** 846–855 (2009).

516. Huang, K. H. *et al.* Discovery of novel 2-aminobenzamide inhibitors of heat shock protein 90 as potent, selective and orally active antitumor agents. *J. Med. Chem.* **52,** 4288–4305 (2009).
517. Lamoureux, F. *et al.* A novel HSP90 inhibitor delays castrate-resistant prostate cancer without altering serum PSA levels and inhibits osteoclastogenesis. *Clin. Cancer Res.* **17,** 2301–2313 (2011).
518. Jain, L. *et al.* Determination of PF-04928473 in human plasma using liquid chromatography with tandem mass spectrometry. *J. Chromatogr. B. Analyt. Technol. Biomed. Life Sci.* **878,** 3187–3192 (2010).
519. Westerheide, S. D., Kawahara, T. L. A., Orton, K. & Morimoto, R. I. Triptolide, an inhibitor of the human heat shock response that enhances stress-induced cell death. *J. Biol. Chem.* **281,** 9616–9622 (2006).
520. Liu, L. L. *et al.* Resveratrol induces antioxidant and heat shock protein mRNA expression in response to heat stress in black-boned chickens. *Poult. Sci.* **93,** 54–62 (2014).
521. Toullec, D. *et al.* The bisindolylmaleimide GF 109203X is a potent and selective inhibitor of protein kinase C. *J. Biol. Chem.* **266,** 15771–15781 (1991).
522. Furuhashi, M. *et al.* Treatment of diabetes and atherosclerosis by inhibiting fatty-acid-binding protein aP2. *Nature* **447,** 959–965 (2007).
523. Macia, E. *et al.* Dynasore, a cell-permeable inhibitor of dynamin. *Dev. Cell* **10,** 839–850 (2006).
524. Baqi, Y., Weyler, S., Iqbal, J., Zimmermann, H. & Muller, C. E. Structure-activity relationships of anthraquinone derivatives derived from bromaminic acid as inhibitors of ectonucleoside triphosphate diphosphohydrolases (E-NTPDases). *Purinergic Signal.* **5,** 91–106 (2009).
525. Chen, B. C., Lee, C. M. & Lin, W. W. Inhibition of ecto-ATPase by PPADS, suramin and reactive blue in endothelial cells, C6 glioma cells and RAW 264.7 macrophages. *Br. J. Pharmacol.* **119,** 1628–1634 (1996).
526. Müller, C. E. *et al.* Polyoxometalates—a new class of potent ecto-nucleoside triphosphate diphosphohydrolase (NTPDase) inhibitors. *Bioorg. Med. Chem. Lett.* **16,** 5943–5947 (2006).
527. Kukulski, F. *et al.* NTPDase1 controls IL-8 production by human neutrophils. *J. Immunol.* **187,** 644–653 (2011).
528. Lecka, J., Rana, M. S. & Sevigny, J. Inhibition of vascular ectonucleotidase activities by the pro-drugs ticlopidine and clopidogrel favours platelet aggregation. *Br. J. Pharmacol.* **161,** 1150–1160 (2010).
529. Thupari, J. N., Landree, L. E., Ronnett, G. V & Kuhajda, F. P. C75 increases peripheral energy utilization and fatty acid oxidation in diet-induced obesity. *Proc. Natl. Acad. Sci. U. S. A.* **99,** 9498–9502 (2002).
530. Eskens, F. A. L. M. *et al.* Phase I pharmacokinetic and pharmacodynamic study of the first-in-class spliceosome inhibitor E7107 in patients with advanced solid tumors. *Clin. Cancer Res.* **19,** 6296–6304 (2013).
531. Christian, F. *et al.* Small molecule AKAP-protein kinase A (PKA) interaction disruptors that activate PKA interfere with compartmentalized cAMP signaling in cardiac myocytes. *J. Biol. Chem.* **286,** 9079–9096 (2011).

Liste des publications

1. Lv Z., Tek A., Da Silva F., Empereur-mot C., Chavent M., Baaden M. **Game on, Science – How Video Game Technology May Help Biologists Tackle Visualization Challenges**, *PLoS One*, 2013, 8(3).

2. Empereur-mot C., Guillemain H., Latouche A., Zagury J.F., Viallon V., Montes M. **Predictiveness Curves in Virtual Screening**, *Journal of Cheminformatics*, 2015, 7(1), 1-17.

3. Empereur-mot C., Zagury J.F., Montes M. **Screening Explorer – An Interactive Tool for the Analysis of Screening Results**, *Journal of Chemical Information and Modeling*, 2016, 56, 2281-2286.

4. Garcia-seisdedos H., Empereur-mot C., Nadav E., Levy E. **Proteins Evolve on the Edge of Supramolecular Self-Assembly**, *Nature*, Accepté le 11/05/2017.

5. Lagarde N., Delahaye S., Jérémie A., Ben Nasr N., Guillemain H., Empereur-mot C., Laville V., Labib T., Réau M., Langenfeld F., Zagury J.F., Montes M. **Discriminating agonist from antagonist ligands of the nuclear receptors using different chemoinformatics approaches**, *Molecular Informatics*, En révision.

Liste des communications orales

1. Empereur-mot C., Guillemain H., Latouche A., Zagury J.F., Viallon V., Montes M. **Courbes de Prédictivité Appliquées au Criblage Virtuel**, *47e Journées de Statistique de la Société Française de Statistiques*, Université de Lille, France, Juin 2015.

2. Empereur-mot C., Guillemain H., Latouche A., Zagury J.F., Viallon V., Montes M. **Predictiveness Curves in Virtual Screening**, *7e Journées de la Société Française de Chemoinformatique*, Université de Nice, France, Octobre 2015.

Liste des posters

1. Empereur-mot C., Guillemain H., Latouche A., Zagury J.F., Viallon V., Montes M. **Courbes de Prédictivité Appliquées au Criblage Virtuel**, *Journées des doctorants de 2e année de l'école doctorale SMI*, ENSAM, Paris, Juin 2015.

2. Empereur-mot C., Guillemain H., Latouche A., Zagury J.F., Viallon V., Montes M. **Predictiveness Curves in Virtual Screening**, *7e Journées de la Société Française de Chemoinformatique*, Université de Nice, France, Octobre 2015.

3. Empereur-mot C., Guillemain H., Latouche A., Zagury J.F., Viallon V., Montes M. **Predictiveness Curves in Virtual Screening**, *23e Rencontre des jeunes chercheurs*, Université de Lille, France, Février 2016.

4. Empereur-mot C., Zagury J.F., Montes M. **Screening Explorer – Un outil interactif pour l'analyse des résultats de criblage**, *5e Ecole d'été de chémoinformatique*, Université de Strasbourg, France, Juin 2016.

Charly EMPEREUR-MOT

Développement d'outils statistiques d'évaluation de méthodes de criblage virtuel : courbes de prédictivité & Screening Explorer

Résumé

Les méthodes de criblage virtuel sont largement utilisées dans le processus de conception de médicaments afin de réduire le nombre de composés à tester expérimentalement. Cependant, les résultats obtenus par criblage virtuel ne sont que des prédictions et leur fiabilité n'est pas garantie. L'évaluation de ces méthodes est donc essentielle pour guider le bioinformaticien dans le choix de l'outil et du protocol adaptés dans les conditions de son expérience. Dans une première étude, nous avons développé une nouvelle métrique pour l'analyse des résultats de criblage : la Courbe de Prédictivité. Cette métrique permet une analyse fine de la pertinence des scores d'affinité pour la détection de composés actifs et complète les métriques existantes, permettant une meilleure compréhension des résultats de criblage. Lors de notre projet suivant, nous avons souhaité faciliter ce processus d'analyse en intégrant l'ensemble des métriques de criblage virtuel dans un outil web interactif : Screening Explorer. Une seconde partie de ma thèse a consisté en la recherche de nouveaux inhibiteurs du VIH. L'équipe génomique de notre laboratoire a identifié plusieurs gènes dont l'expression influence le développement du SIDA, révèlant ainsi de potentielles cibles thérapeutiques. Une étude bibliographique a permis d'identifier plusieurs composés inhibiteurs de ces cibles. La société Peptinov, associée à notre laboratoire, va prochainement estimer le potentiel thérapeutique de ces composés dans des essais *in vitro* (i) d'infection par le VIH, (ii) de prolifération virale et (iii) de réactivation virale.

Mots-clés : Criblage virtuel, Evaluation, Métriques, Courbe de prédictivité, Conception de médicaments, Petites molécules, VIH, SIDA.

Résumé en anglais

Virtual screening methods are widely used in drug discovery processes in order to reduce the number of compounds to test experimentally. However, virtual screening results are only predictions and their reliability is not guaranteed. Evaluating these methods is crucial to guide the bioinformatician in the choice of the right tool and protocol according to the conditions of his experiment. In a first study, we developed a new metric to analyze the results of virtual screening: the Predictiveness Curve. This metric allows to finely analyze the relevance of binding scores for the detection of active compounds and complete existing metrics, allowing a better comprehension of screening results. In a following project, we facilitated the analysis process by integrating all of the virtuel screening metrics in an interactive tool: Screening Explorer. The second part of my thesis consisted in the research of novel HIV inhibitors. The genomic team of our laboratory identified several genes whose expression influence the development of AIDS, therefore revealing potential therapeutic targets. A bibliographic study allowed to identify compounds that can inhibit those targets. The company Peptinov, associated to our laboratory, is currently evaluating the therapeutic potential of the compounds *in vitro* in assays of (i) HIV infection, (ii) viral proliferation and (iii) viral reactivation.

Keywords : Virtual screening, Evaluation, Metrics, Predictiveness curve, Drug design, Small molecules, HIV, AIDS.