



HAL
open science

Méthodes de criblage virtuel in silico : importance de l'évaluation et application à la recherche de nouveaux inhibiteurs de l'interleukine 6.

Nathalie Lagarde

► **To cite this version:**

Nathalie Lagarde. Méthodes de criblage virtuel in silico : importance de l'évaluation et application à la recherche de nouveaux inhibiteurs de l'interleukine 6.. Bio-Informatique, Biologie Systémique [q-bio.QM]. Conservatoire national des arts et metiers - CNAM, 2014. Français. NNT : 2014CNAM0943 . tel-01132490

HAL Id: tel-01132490

<https://theses.hal.science/tel-01132490>

Submitted on 17 Mar 2015

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

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

THÈSE présentée par :

Nathalie LAGARDE

soutenue le : **29 octobre 2014**

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

Discipline/ Spécialité : Bioinformatique

**Méthodes de criblage virtuel *in silico* :
importance de l'évaluation et application à
la recherche de nouveaux inhibiteurs de
l'interleukine 6**

THÈSE dirigée par :

M. ZAGURY Jean-François
M. MONTES Matthieu

Professeur, Directeur de thèse, Cnam
Docteur, Co-directeur de thèse, Cnam

RAPPORTEURS :

M. LANGER Thierry
M. SAETTEL Nicolas

Professeur, Université de Vienne
Docteur, Université de Caen-Basse Normandie, Institut Curie

JURY :

M. DALLEMAGNE Patrick
Mme. DUMAS Françoise
M. RODRIGO DE LOSADA Jordi

Professeur, Université de Caen-Basse Normandie
Directrice de recherche CNRS, Université Paris-Sud
Docteur, Université Paris-Sud

« Aucun de nous, en agissant seul, ne peut atteindre le succès »

Nelson Mandela, 10 mai 1994

Remerciements

Je tiens à exprimer tous mes remerciements au professeur Jean-François Zagury de m'avoir accueilli dans son laboratoire, de sa confiance et son soutien tout au long de cette thèse. Qu'il soit assuré de ma gratitude.

J'adresse également tous mes remerciements au docteur Matthieu Montes pour m'avoir encadré durant toutes mes années de thèse, pour tout le temps qu'il a passé à me former, pour avoir relu et corrigé ce manuscrit et pour avoir fait de moi la chercheuse que je suis !

Je suis très reconnaissante au Professeur Thierry Langer et au Docteur Nicolas Saettel d'avoir accepté de juger mon travail de thèse en tant que rapporteur.

Je remercie également le Professeur Patrick Dallemagne, le Docteur Françoise Dumas et le Docteur Jordi Rodrigo De Losada de m'avoir fait l'honneur de participer au jury de cette thèse.

Je remercie toutes les personnes qui m'ont aidé tout au long de mon cursus et qui ont contribué, d'une manière ou d'une autre, à la réalisation de cette thèse, en particulier les équipes du CERMN et du laboratoire BioCIS, avec un grand merci à Jordi Rodrigo de Losada pour avoir initié ma formation de chercheuse et m'avoir poussé à continuer dans cette voie et à Françoise Dumas qui m'a présenté au Professeur Jean-François Zagury et m'a permis d'obtenir ce sujet de thèse.

Il n'est pas facile de trouver les mots pour remercier toute l'équipe du laboratoire GBA, qui a toujours été là pour moi durant cette thèse, et qui a supporté mes envolées lyriques sans (trop) se plaindre. Il me faut bien évidemment commencer par la partie drug design de l'équipe : Nesrine, avec qui j'ai tout partagé : le bureau-bunker, le chocolat, les bons moments et les moins bons, chokran beaucoup ; Hélène, ma fournisseuse officielle de Games of Thrones et toujours là pour rigoler et enfin Charly et son « bon esprit, gros(se) ». Même si je ne comprends pas tout lorsqu'ils parlent « travail », je remercie également énormément les génomiciens (par ordre alphabétique car je vous apprécie tous autant malgré nos différents politiques) : Cédric Federer pour les nombreux débats du midi, pari ligue1, et autres activités cohésives ; Damien pour avoir partagé mon bureau, n'avoir pendu mon lapin duracell que 2 ou 3 fois et pour tous les bons moments de rigolade ; Jean-Louis pour sa disponibilité et sa gentillesse ; Josselin pour sa maîtrise de l'anglais et sa sympathie ; Lieng pour avoir adoré me détester (ou l'inverse ?), les blind tests mythiques et les origamis ; Pierre pour sa gentillesse, sa souris d'ordinateur, sa chaise de bureau, son écran d'ordinateur ; Sigrid, ah Sigfried euh Sigrid, pour à peu près tout de la gym suédoise aux pauses, en passant par les séances shopping et les discussions boucles d'oreilles, mais surtout pour sa franchise et son énorme gentillesse ; Taoufik le sage, pour être toujours de bonne humeur, disponible et prêt à remonter le moral ; et enfin Vincent, pour ses punch lines qui sont comme des coups francs de Zlatan, pour tous les bons moments de musique, de rire et de trolls,

pour ses imitations inimitables mais aussi pour son aide du point de vue scripto-informatique. Je n'oublie pas non plus Sophie et Olivier qui sont partis quand je suis arrivée, coïncidence ? je ne crois pas... Je termine donc par l'équipe basée à Cochin : Gaby alias « Magic Mike », Hadley grâce à qui tant de verres du vendredi soir ont été organisés et qui a largement contribué à les rendre incontournables ; Julie pour les bons moments partagés à Ermenonville et Chissay, pour avoir réalisé les criblages expérimentaux, me les avoir patiemment et clairement expliqué et pour ses corrections de cette partie de ce manuscrit et enfin Lucille, adversaire coriace à pari ligue 1, pour tous les bons moments passés à parler de tout (vernissages, acteurs qui ne dormiraient pas dans la baignoire...) et de rien (publication, thèse...). Un grand merci aussi à Hervé, non seulement pour mon salaire mensuel, mais aussi pour les éclats de rire et l'animation musicale du bureau (ceci n'inclut pas la chronique de Zemmour). Tous mes remerciements vont aussi à Christiane, Janine et Juanjo pour leur gentillesse, leur disponibilité et leur aptitude à résoudre tous nos petits et gros problèmes. Je remercie aussi Aurore, Stéphanie, Marc et Lourdes les stagiaires avec qui j'ai pu travailler.

Je tiens à remercier aussi tous mes amis qui me supportent depuis très longtemps pour certains maintenant (n'est-ce pas Marie ?), et qui m'ont tous apporté leur soutien durant cette thèse, en particulier Anne-Sophie et Marlène, qui un jour posséderont probablement la pharmacie Lanciot-Mouchel, Céline qui en a eu marre que je l'assomme avec ma thèse et qui a préféré fuir en Australie, Anne et Greg évidemment que je ne vois pas aussi souvent que je le voudrais. Je tiens à remercier particulièrement Alexis pour la centaine de bagels et autant de cheesecakes partagés, et pour son amitié infailible.

Enfin, je remercie toute ma famille, et en particulier mes parents et mon frère qui m'ont appris à ne jamais baisser les bras, m'ont toujours poussé à me dépasser et m'ont toujours soutenue. J'espère les rendre fiers de moi comme moi je suis fière d'eux.

Résumé

Le criblage virtuel est largement employé pour la recherche de nouveaux médicaments.

La sélection de structures pour les méthodes de criblage virtuel basées sur la structure reste problématique. Nous avons montré que les propriétés physico-chimiques du site de liaison, critères simples et peu coûteux en temps de calcul, pouvaient être utilisés pour guider celle-ci.

L'évaluation des méthodes de criblage virtuel, critique pour vérifier leur fiabilité, repose sur la qualité de banques d'évaluation. Nous avons construit la NRLiSt BDB, n'incluant que des données vérifiées manuellement et prenant en compte le profil pharmacologique des ligands. Une étude à l'aide du logiciel Surflex-Dock montre qu'elle devrait devenir la base de données de référence, pour l'évaluation des méthodes de criblage virtuel et pour rechercher de nouveaux ligands des récepteurs nucléaires.

L'application d'un protocole hiérarchique de criblage *in silico/in vitro*, a permis d'identifier de nouveaux composés inhibiteurs de l'IL-6, potentiellement utilisables dans le traitement de la polyarthrite rhumatoïde. Les résultats *in vitro* devront être confirmés par des tests *in vivo*.

Mots clés : criblage virtuel, docking, flexibilité, banque d'évaluation, récepteurs nucléaires, interleukine 6, polyarthrite rhumatoïde

Résumé en anglais

Virtual screening is widely used in drug discovery processes.

Structure selection in structure-based virtual screening methods is still problematic. We showed that simple and “low cost” binding site physico-chemical properties could be used to guide structure selection.

The evaluation of virtual screening methods, necessary to ensure their reliability, relies on benchmarking databases quality. We created the NRLiSt BDB, gathering only manually curated data and taking into account ligands pharmacological profiles. A study using Surflex-Dock showed that the NRLiSt BDB should become the reference, both for the evaluation of virtual screening methods and for the identification of new ligands of the nuclear receptors.

The use of a *in silico/in vitro* hierarchical approach screening allowed to identify new IL-6 inhibitors, that could be used in rheumatoid arthritis treatment. *In vitro* results should be confirmed *in vivo*.

Key words: Virtual screening, docking, flexibility, benchmarking database, nuclear receptors, interleukin 6, rheumatoid arthritis

Table des matières

Remerciements	1
Résumé	3
Résumé en anglais	4
Table des matières	5
Liste des tableaux	10
Liste des figures	12
Liste des équations	22
Liste des abréviations.....	27
Première partie Introduction	32
1 Découverte de nouveaux médicaments.....	33
1.1 Histoire de la découverte des médicaments	33
1.2 Schéma général de R&D.....	34
1.2.1 Choix d'une cible thérapeutique	35
1.2.2 Identification de hits	37
1.2.3 Génération et optimisation des leads.....	38
1.2.4 Tests pré-cliniques.....	38
1.2.5 Tests cliniques	39
2 Méthodes de criblages <i>in silico</i>	41
2.1 Généralités.....	41
2.2 Les chimiothèques	43
2.2.1 Différents types de chimiothèques	44
2.2.1.1 Chimiothèques dans l'espace réel et l'espace global	44
2.2.1.2 Chimiothèques dans l'espace tangible et virtuel	49
2.2.2 Formats de chimiothèques virtuelles.....	50
2.2.2.1 Les formats de fichiers 2D	50
2.2.2.2 Les formats de fichiers 3D	54
2.2.3 Préparation d'une chimiothèque.....	59
2.2.3.1 États d'ionisation, mésomérie et tautomérie	60
2.2.3.2 Génération des conformations 3D.....	60
2.2.3.3 Filtres ADME-Tox	61
3 Criblage virtuel « ligand-based ».....	71
3.1 Recherche de similarité.....	71

3.1.1	Descripteurs de similarité.....	71
3.1.1.1	Descripteurs 2D.....	72
3.1.1.2	Descripteurs 3D.....	75
3.1.2	Métriques de similarité.....	78
3.2	Modèles pharmacophoriques « ligand-based ».....	82
3.2.1	Approches pharmacophoriques 2D.....	84
3.2.2	Approches pharmacophoriques 3D.....	87
3.2.2.1	Elucidation du pharmacophore.....	87
3.2.2.2	Criblage de chimiothèques.....	93
3.3	Modèles de relations quantitatives structure-activité (QSAR).....	94
3.3.1	QSAR-2D.....	95
3.3.2	QSAR 3D.....	97
3.3.2.1	Analyse comparative des champs moléculaires : CoMFA.....	97
3.3.2.2	Analyse comparative d'indices de similarité moléculaire : CoMSIA.....	99
3.3.2.3	GRID/GOLPE.....	101
3.3.2.4	Phase.....	102
3.4	Succès du criblage virtuel « ligand-based ».....	104
4	Criblage virtuel « structure-based ».....	106
4.1	Identification du site de liaison.....	106
4.1.1	Structure co-cristallisée avec un ligand.....	107
4.1.2	Outils de prédiction de site de liaison.....	107
4.1.2.1	Outils de prédiction basés sur la géométrie.....	107
4.1.2.2	Outils de prédiction basés sur les énergies.....	108
4.1.2.3	Outils de prédiction basés sur la connaissance.....	109
4.2	Modèles pharmacophoriques basés sur la structure du récepteur.....	110
4.2.1	Approche basée sur le récepteur.....	111
4.2.2	Approche basée sur le complexe récepteur-ligand.....	112
4.3	RD-QSAR (Receptor Dependent-Quantitative Structure-Activity Relationship).....	113
4.4	Conception de novo.....	115
4.4.1	Identification des sites d'interactions dans le site de liaison.....	118
4.4.2	Assemblage des blocs de construction.....	118
4.4.3	Recherche combinatoire.....	120
4.4.4	Attribution de scores.....	122

4.5	<i>Méthodes de docking</i>	123
4.5.1	Docking avec ligand rigide	123
4.5.2	Docking avec ligand flexible.....	124
4.5.2.1	Algorithmes de recherche.....	125
4.5.2.2	Scoring	130
4.5.3	Principaux logiciels de docking	139
4.5.4	Problématique lié aux méthodes de docking.....	140
4.5.4.1	Disponibilité des structures des protéines	141
4.5.4.2	Importance du solvant	142
4.5.4.3	Gestion de la flexibilité de la protéine	144
4.6	<i>Succès du criblage virtuel basé sur la structure</i>	146
5	Evaluation des méthodes de criblage virtuel	148
5.1	<i>Précision du positionnement</i>	148
5.1.1	Ecart quadratique moyen ou RMSD	149
5.1.2	L'erreur relative de déplacement RDE.....	150
5.1.3	L'espace réel du facteur R RSR	150
5.1.4	Classification de précision basée sur les interactions IBAC	152
5.2	<i>Enrichissement d'une chimiothèque</i>	152
5.2.1	Les banques d'évaluation	153
5.2.2	Les métriques de performance	155
5.2.2.1	Facteurs d'enrichissement	156
5.2.2.2	Courbes de ROC (Receiver Operating Characteristic)	157
5.2.2.3	Robust Initial Enhancement (RIE)	159
5.2.2.4	Boltzmann-enhanced discrimination of ROC (BEDROC)	160
6	Objectifs de thèse	162
	Deuxième partie Résultats	163
1	Evaluation des méthodes de criblage virtuel	164
1.1	<i>SBVLS : Définition de critères basés sur les propriétés du site de liaison pour optimiser la sélection de la (ou des) structure(s) de référence</i>	164
1.1.1	Introduction	164
1.1.2	Publication.....	166
1.1.3	Discussion	186
1.1.4	Analyse critique de l'étude.....	188

1.1.4.1	La banque d'évaluation DUD	188
1.1.4.2	Déroulement de l'étude	189
1.1.5	Conclusion.....	191
1.2	<i>La NRLiSt BDB : une banque d'évaluation validée manuellement dédiée aux ligands et aux structures des récepteurs nucléaires.....</i>	193
1.2.1	Introduction	193
1.2.1.1	Les récepteurs nucléaires	193
1.2.1.2	Récepteurs nucléaires et évaluation des méthodes de criblage virtuel.....	196
1.2.2	Publication.....	198
1.2.3	Discussion	208
1.2.3.1	Séparation des jeux de données « agoniste » et « antagoniste »	208
1.2.3.2	Sélection des RNs, des structures et des ligands à inclure dans la NRLiSt BDB	208
1.2.3.3	Tentative de profilage des ligands agonistes et antagonistes à l'aide de descripteurs structuraux.....	210
1.2.3.4	Présentation du site web de la NRLiSt BDB.....	212
1.2.4	Analyse critique de l'étude.....	213
1.2.4.1	Critique de la base de données ChEMBL	213
1.2.4.2	Déséquilibre inter- et intra- jeux de données	215
1.2.4.3	Diversité structurale de la NRLiST BDB.....	215
1.2.4.4	Améliorations	215
1.2.5	Conclusion.....	216
1.3	<i>Importance du profil pharmacologique du ligand co-cristallisé et utilisation de « decoys ligands ».....</i>	217
1.3.1	Introduction	217
1.3.2	Publication.....	218
1.3.3	Discussion	249
1.3.3.1	Influence de l'utilisation de jeux de données séparés sur les performances du docking	249
1.3.3.2	Importance du ligand co-cristallisé dans la structure utilisée pour le criblage	250
1.3.3.3	Recherche de nouveaux decoys.....	251
1.3.4	Analyse critique de l'étude.....	252

1.3.5	Conclusion.....	253
2	Réalisation d'un criblage virtuel à la recherche de composés inhibant l'interleukine IL-6	256
2.1	<i>La polyarthrite rhumatoïde</i>	256
2.2	<i>Interleukine IL-6</i>	260
2.3	<i>Protocole du criblage</i>	263
2.3.1	Chimiothèque de criblage.....	264
2.3.2	Sélection de la structure et identification du site actif	264
2.3.3	Réalisation du criblage virtuel.....	265
2.3.4	Tests biologiques.....	266
2.3.4.1	Criblage expérimental par test cellulaire HEK-BLUE™ IL-6.....	266
2.3.4.2	Essai de spécificité des produits confirmés.....	268
2.3.4.3	Test de liaison IL-6/IL-R.....	269
2.3.5	Résultats préliminaires	271
	Troisième partie Conclusion	282
	Bibliographie	285
	Liste des publications	312
	Liste des communications orales	313
	Posters	313
	Résumé	314
	Résumé en anglais	314

Liste des tableaux

Tableau 1. Liste de quelques compagnies proposant des chimiothèques commerciales focalisées pour différentes familles de cibles biologiques (d'après ³³) (27/01/2014).....	45
Tableau 2. Classification des principales chimiothèques généralistes en fonction du nombre de composés proposés, et de leur caractère commercial ou publique (d'après ³³) (27/01/2014).....	46
Tableau 3. Classification de quelques chimiothèques de fragments en fonction du nombre de composés mis à disposition (d'après ³³) (30/01/2014).....	48
Tableau 4. Classification de quelques chimiothèques de produits naturels en fonction du nombre de composés mis à disposition (31/01/2014).....	49
Tableau 5. Valeurs seuils des différentes propriétés physico-chimiques définissant le caractère « drug-like » ou « lead-like » d'un composé. Pour la règle de Veber et al., le symbole « * » signifie que l'un ou l'autre de ces critères doit être respecté en plus du nombre de liaisons rotatives.....	63
Tableau 6. Présentation de quelques bases de données toxicologiques publiques en fonction de leur contenu (d'après ⁸²).....	67
Tableau 7. Liste de quelques programmes publics et commerciaux permettant d'obtenir des modèles QSAR de toxicologie prédictifs (d'après ⁸²).....	69
Tableau 8. Mode de représentation des points pharmacophoriques et exemple de modèles de pharmacophore obtenu avec les logiciels CATALYST ¹⁵⁷ , MOE ¹⁵⁸ , PHASE ¹⁵⁹ et SCAMPI ¹⁶⁰ (d'après ¹⁶¹).....	90
Tableau 9. Résumé des caractéristiques de quelques approches de pharmacophores 3D.....	92
Tableau 10. Quelques succès obtenus dans la découverte de nouveaux composés actifs en utilisant des méthodes LBVLS.....	105
Tableau 11. Caractéristiques des principaux logiciels, classés par ordre chronologique et utilisés pour la conception de novo(DFS :Depth-First-Search, BFS : Breadth-First-Search, Rnd : Random, MC : Monte-Carlo, AE : Algorithme Evolutionnaire ; Gr : Growth, Lk : Link, Lat : Lattice, DM : dynamique moléculaire, Sto : Stochastique ^{297, 298}).....	117
Tableau 12. Quelques exemples de logiciels de docking classés selon leur gestion de la flexibilité du ligand, leur approche de la recherche conformationnelle des ligands, leurs algorithmes de recherche et leurs fonctions de score.....	139

Tableau 13. Principales sources d'erreur lors d'un docking, classé par complexité décroissante de la solution à apporter pour éviter cette erreur ³⁵⁹	140
Tableau 14. Quelques exemples de médicaments pour lesquels la structure 3D a été rationnellement et significativement mise à profit (RSA : Relations Structure-Activité, HIV : Human Immunodeficiency Virus, LMC : Leucémie Myéloïde Chronique) (d'après ³⁸³).....	147
Tableau 15. Classification des 39 cibles de la DUD en fonction du volume moyen et de l'hydrophobie moyenne des sites de liaison de leurs différentes structures	186
Tableau 16. Les récepteurs nucléaires sont des cibles potentielles pour un grand nombre de maladies (indiquées en gras) et de dysfonctionnements physiologiques (indiqués en italique) (d'après ⁴⁷³).....	195
Tableau 17. Classification du nombre de publications et de composés associés à tort à un récepteur nucléaire en fonction de la cause de l'erreur, pour 19 des 27 RNs de la NRLiSt BDB.....	214
Tableau 18. Présentation des différents AMRR actuellement disponible (Ac: Anticorps, AICAR: 5-aminoimidazole-4-carboxamide ribonucleotide, ATIC : 5-aminoimidazole-4-carboxamide ribonucléotide transformylase, DHFR : Dihydrofolate reductase, DHPDH : Dehydroorotate dehydrogenase, TLR : Toll-like receptor, TS : Thymidilate synthase) (d'après ⁵⁰²)	259
Tableau 19. Résultats préliminaires obtenus pour les 353 composés ayant montré un pourcentage de neutralisation de l'activité de l'IL-6 sur la lignée cellulaire HEK-Blue™ (h-IL6 neutra.) supérieur à 20 pourcent, lors de la reconfirmation différentes doses (h-IL neutra. RECONF), lors de l'étude de spécificité des produits pour l'IL-6 à l'encontre du TNF α (h-TNF α neutra.), de l'IL-1 (h-IL1 β neutra.) et de l'IL-4 (h-IL4 neutra) et du test de liaison IL-6/IL-6R (liaison IL-6) (avec Neutra : pourcentage de neutralisation ; Neutra. COR. : pourcentage de neutralisation corrigé).....	281

Liste des figures

Figure 1. Quelques grandes étapes de la découverte de médicaments au cours du XIXe et XXe siècle (d'après ⁷).....	34
Figure 2. Modèle de R&D pour développer avec succès un nouveau médicament présentant les différentes étapes ainsi que le nombre de molécules et les coûts en terme monétaire et temporel associés à chaque étape (d'après ⁹).....	35
Figure 3. Taux de succès des candidats médicaments à chaque étape des phases cliniques et lors des étapes d'enregistrement (Reg) et de demande d'autorisation de mise sur le marché (App.) ³¹	40
Figure 4. Classification des méthodes de criblage virtuel « ligand-based » et « structure-based » (avec QSAR : Quantitative Structure-Activity Relationship ou modèle de relations quantitatives structure-activité et RD-QSAR : Receptor Dependent-Quantitative Structure-Activity Relationship ou modèle de relations quantitatives structure-activité dépendant du récepteur).....	42
Figure 5. Evolution du nombre de publications de 1997 à 2013 dans la base de données PUBMED en utilisant les mots clés « virtual screening » (VS), « ligand-based virtual screening » (LBVS) et « structure-based virtual screening » (SBVS) (08/01/2014).....	43
Figure 6. Potentiel en candidats médicaments de collections de criblage commerciales (1 à 7 et 9 à 13) et publique (8 : Chimiothèque Nationale) ³⁸	44
Figure 7. Exemple de construction d'un code SMILES d'un cycle aromatique avec le benzène.....	51
Figure 8. Représentation par les codes SMILES de la configuration E (a) et Z (b) du but-2-ène.	51
Figure 9. Exemple de deux codes InChI pour une molécule, l'alanine, avec (a) ou sans (b) détail de stéréochimie.....	53
Figure 10. Code InChIKey standard de la L-alanine.	54
Figure 11. Fichier SDF de la L-alanine téléchargé dans la base de données ZINC ⁴³ , illustrant la disposition en bloc des informations structurales.....	55
Figure 12. Fichier PDB de la L-alanine généré avec la version en ligne de Corina ⁷⁰ . Dans la section HEADER, l'identifiant PDB (non présent ici puisqu'il s'agit d'un ligand et pas d'une protéine), la date de publication dans la banque (ici, date de génération par CORINA) et la classification de la molécule sont indiqués (ici, unknown). Dans la section REMARK, les informations usuellement proposées sont le nom de la molécule,	

l'espèce dont la molécule est extraite, les auteurs, des références bibliographiques, et d'autres informations générales sur la protéine.	57
Figure 13. Format MOL2 de la L-alanine téléchargé de la base de données ZINC ⁴³	59
Figure 14. Exemple de deux formes tautomères (a), mésomère (b) ou états d'ionisation différents (c).	60
Figure 15. Mécanisme de liaison d'un ligand à son récepteur par sélection de la meilleure conformation du récepteur par le ligand (passage de l'état 1 à 2) suivi d'un ajustement de la protéine et du ligand l'un à l'autre selon un mécanisme induced fit (passage de l'état 3 à l'état 4) ⁷³	61
Figure 16. Evolution des causes d'attrition des candidats médicaments lors des phases cliniques entre 1991 et 2000 ³¹	61
Figure 17. Comparaison de la distribution de huit propriétés entre les leads (en noir) et les médicaments (en gris) : le poids moléculaire (a), le logP calculé (b), la solubilité logS (c), la fraction d'aire de surface polaire (d), et non polaire (e), la somme du nombre de donneurs et d'accepteurs de liaison (f), le nombre de cycles aromatiques (g) et le nombre de liaison rotatives (h). L'axe vertical représente le nombre de composés. Les valeurs ml et mm indiquent la valeur médiane de chaque propriété pour les leads et les médicaments respectivement (d'après ⁷⁹).	65
Figure 18. Classification des descripteurs selon leurs dimensions ⁹⁹	72
Figure 19. Exemple d'empreinte 2D possible pour le paracétamol	73
Figure 20. Exemples de quelques fragments (rang de liaison de 4) créé par un algorithme permettant d'obtenir des empreintes de type « hashed fingerprints »	74
Figure 21. Exemple d'empreinte 3D pour le paracétamol codant une paire de pharmacophores (jaune), un triplet de pharmacophores (grenat) et une paire d'atomes (violet)	75
Figure 22. (a) Principe du suivi d'un rayon lumineux et de sa réflexion pour décrire la forme d'une molécule. (b) Traces des rayons pour l'indinavir à basse (100 réflexions) résolution ¹¹⁴	76
Figure 23. Représentation de (a) la forme à l'aide du modèle CPK (Corey-Pauling-Koltun) (b) la position de tous les atomes. (c) toutes les distances atomiques par rapport aux quatre points de références (d) la définition des douze descripteurs géométriques ¹¹⁰	77
Figure 24. Illustration de la comparaison des formes de deux molécules en fonction du chevauchement de leur volume ¹⁰⁹	78

Figure 25. Présentation des premiers pharmacophores à avoir été publié, définissant un pharmacophore pour les agonistes muscariniques. Le modèle de Beckett (a) datant de 1963 définit des ordres de grandeur de distances entre la zone 1, une cavité anionique chargée négativement pour accueillir une amine quaternaire, la zone 2, un point chargé positivement permettant la liaison ou de l'acétylcholine et ses analogues et la zone 3, chargée pour interagir avec le OH de la muscarine, le C-O de l'acétylcholine et ses analogues ou la double liaison des analogues furaniques de la muscarine ¹³² . Le modèle de Kier (b) propose quant à lui des distances calculées entre 3 atomes clés communs à l'acétylcholine, la muscarine et la muscarone ¹³³	83
Figure 26. Pharmacophore généré à l'aide du logiciel CATALYST à partir de 5 antagonistes du récepteur de la sérotonine 5-HT5A dont l'EMDT (2-Ethyl-5-methoxy-N,N-dimethyltryptamine) ici représenté. Les différents types de points pharmacophoriques sont illustrés par des sphères de couleur bleue pour les groupements hydrophobes, rouge pour les groupements chargés positivement, orange pour les aromatiques et verte pour les accepteurs de liaison hydrogène.....	84
Figure 27. Exemple de « Similog keys » (d'après ¹³⁶).....	85
Figure 28. Conversion de la structure chimique en graphe réduit en 4 étapes (D : donneur de liaisons hydrogènes, Ac : accepteur de liaison hydrogène, Hf : groupement hydrophobe, Ar : groupement aromatique, + : charge positive) (d'après ¹³⁷).....	85
Figure 29. Principe de l'utilisation des descripteurs CATS 2D (R : Aromatique, L : Lipophile, A : Accepteur de liaisons hydrogènes, D : Donneurs de liaisons hydrogènes) (d'après ¹⁴⁰).....	86
Figure 30. Exemple de construction d'un « feature tree (en vert les nœuds principalement hydrophobes, en orange les accepteurs de liaisons hydrogènes, en bleu les donneurs de liaisons hydrogène et en jaune les atomes ne formant pas d'interactions directes) (d'après ¹⁴²).....	87
Figure 31. Principales étapes de l'élucidation d'un pharmacophore	88
Figure 32. Schéma général des différentes étapes d'une étude QSAR (d'après ¹⁷³)	95
Figure 33. Déroulement d'une étude CoMFA	98
Figure 34. (a) Alignement des 14 molécules sur la structure cristallisée du PU-H71 (code PDB : 2FWZ) et (b) Représentation du modèle CoMFA permettant de visualiser la contribution stérique (région où des substituants encombrants sont favorables (vert) ou défavorable (jaune) à l'activité, et la contribution électrostatique (région où un potentiel	

négatif est favorable (bleu) ou défavorable (rouge) à l'activité) (source : rapport de M2 de Nathalie Lagarde, laboratoire BioCIS, 2010).....	99
Figure 35. (a) Courbes des potentiels de Lennard-Jones (rouge) et Coulomb (bleu) utilisées dans les études CoMFA et définition d'une valeur limite supérieure. (b) La fonction gaussienne en forme de cloche des champs SEAL (orange) utilisée lors des études CoMSIA est une bonne approximation des potentiels de Lennard-Jones et Coulomb tout en présentant l'avantage d'être plus lissée (d'après ¹⁸⁷).....	100
Figure 36. Aires de contour d'un modèle CoMSIA permettant de visualiser la contribution hydrophobique favorable (jaune) ou défavorables (gris), et la contribution des accepteurs de liaisons hydrogène favorable (violet) ou défavorable (vert) à l'activité ¹⁹⁰	101
Figure 37. Aires de contour obtenues par une approche GRID/GOLPE pour des inhibiteurs des histones déacetylases (en blanc la trichostatine A et en vert le composé TAA_5A). Les aires de contours colorées en cyan illustrent les zones de contribution négative à l'activité alors que celles colorées en jaune indiquent les zones favorables à l'activité ¹⁹³	102
Figure 38. Principales étapes du déroulement d'une étude 3D-QSAR à l'aide du logiciel Phase	103
Figure 39. (a) Hypothèse de pharmacophore AADHRR pour des inhibiteurs de la HSP90 (A : accepteur de liaison hydrogène, D : donneur de liaisons hydrogènes, H : hydrophobe, R : aromatique) et (b) Représentation du modèle 3D-QSAR obtenu (cubes bleus : coefficient positif, rouges : coefficient négatif) et contribution de chaque classe d'atomes : (c) donneur de liaisons hydrogène, (d) hydrophobe, (e) ionisation positive et (f) électroattracteur (source : rapport de M2 de Nathalie Lagarde, laboratoire BioCIS, 2010)	104
Figure 40. Méthodes pour obtenir la structure 3D d'une cible biologique d'intérêt: expérimentale (cristallographie aux rayons X et RMN) ou de novo (prédiction par homologie)	106
Figure 41. Schématisation de la triangulation de Delaunay pour un modèle simplifié d'atomes possédant tous le même rayon (a). Lorsque l'on relie le centre de tous les atomes, un polygone est formé (b) qui peut être triangulé de manière à ce que tout le polygone soit couvert sans superposition de triangles (c). Selon la méthode « discrete flow », un triangle agit comme un « puit » pour les triangles voisins et la poche est définie (d). Dans	

certains cas, ce « puit » ne peut pas être créé et CASTp ne considère donc pas cette partie comme une poche. (d'après ²¹⁸)	108
Figure 42. Site de liaison (code PDB: 1BBP) pour la protéine liant la biline (Biling Binding Protein BBP) prédit avec Q-SiteFinder. Les sondes utilisées pour prédire ce site correspondent aux nœuds de la grille. (²¹²)	109
Figure 43. Exemples de modèles de sites proposés par WebFEATURE représentant l'environnement 3D en utilisant différentes propriétés physico-chimiques : (a) site de liaison pour le calcium, (b) site de pont disulfure et (c) site actif de sérine protéase. (²³¹) (en bleu : carbones, en bleu foncé : azotes, en rouge : oxygènes, en vert : calcium, jaune : soufre).....	110
Figure 44. Résultats de la recherche d'un site de liaison sur le domaine C-terminal de la Hsp90 d'un modèle construit par homologie à l'aide des logiciels Pocket-Finder (a), CASTp (b) Q-SiteFinder(c) et WebFEATURE (d).....	110
Figure 45. Schématisation des différentes étapes jalonnant la construction d'un modèle de pharmacophore basé sur la structure du récepteur (d'après ¹⁴³).....	111
Figure 46. Modèle de pharmacophore généré pour le complexe Angiotensin Converting Enzyme (ACE) lisinopril, code PDB 1O86 (flèche verte : donneur de liaison hydrogène, flèche rouge : accepteur de liaison hydrogène, sphère jaune : hydrophobe, sphère noire : volume exclus, sphère rouge : groupe ionisable négativement, sphère bleue : groupe ionisable positivement) ²³⁸	113
Figure 47. Exemple d'application de la méthode VALIDATE pour la recherche d'inhibiteurs de la HIV-1 protease (les contours oranges et bleus représentent des zones où des groupes chargés respectivement positivement et négativement diminuent l'activité, les contours jaune et vert symbolisent des zones dans lesquelles la présence de groupements encombrants est respectivement défavorable et favorable) ²⁴¹	114
Figure 48. Coloration de la surface du récepteur BACE-1 (code PDB 1W51) avec 46 inhibiteurs superposés dans le site de liaison, à l'aide des coefficients PLS de van der Waals (A) et électrostatiques (B) ²⁴³	115
Figure 49. Illustration de la conception de novo de ligands par construction incrémentale par liaison après identification du site actif (a) et des sites d'interaction (b). Les différents fragments sont placés dans le site actif (c) puis liés les uns aux autres pour obtenir la molécule finale (d) ²⁹⁷	119
Figure 50. Conception de novo de ligands par croissance (d'après ²⁹⁶).....	119

Figure 51. Conception de ligands à l'aide d'un treillis de points qui remplit le site de liaison (a). Les sites d'interaction sont ensuite reliés en suivant le plus court chemin passant par des points de la grille (b). Le chemin ainsi obtenu est ensuite transformé en un squelette moléculaire adéquat (c) puis en une molécule finale (d). ²⁹⁷	120
Figure 52. Modèle d'exploration de l'espace de recherche dans les méthodes de conception de novo. L'approche « depth-first search » ne conserve qu'une seule solution à chaque niveau (ici, le chemin du milieu) alors que l'approche « breadth-first search » conserve toutes les solutions plausibles à chaque niveau (ici, les chemins de gauche et du milieu pour le niveau 1, le chemin du milieu pour le niveau 2 etc...) (d'après ²⁹⁷).....	121
Figure 53. Le logiciel LEA3D utilise la représentation 3D (a), SMILES (b) et basée sur les fragments (c). Les molécules Mol_a et Mol_b sont combinées par une opération de recombinaison ou (crossover) et la molécule Mol_c est modifiée par cyclisation à l'aide de l'opérateur de mutation ²⁸³	122
Figure 54. Schématisation du protocole de docking ligand rigide à l'aide du logiciel FRED (d'après ³⁰³)	124
Figure 55. Protomol généré pour la beta lactamase AMPc par extension de 4 Å autour du ligand (code PDB: 2R9W)	126
Figure 56. Schématisation du processus de docking à l'aide d'un algorithme génétique ²³ ..	128
Figure 57. Influence de l'utilisation de fonctions de score consensus sur l'enrichissement en vrais actifs (en noir : FlexX ou Dock seuls ; en gris foncé : combinaison FlexX/Dock ou FlexX/Gold, en gris clair : combinaison FlexX/Dock/Fresno ou FlexX/Gold/Fresno) ³⁵⁹	138
Figure 58. Utilisation des différents logiciels de docking en 2010-2011 (fréquence de citation des noms des logiciels dans les titres et les résumés dans PubMed) (d'après ³⁶¹).....	140
Figure 59. Evolution du nombre de structures disponibles dans la PDB depuis sa création à 2014 (en bleu: nombre de nouvelles structures par an, en rose: nombre total de structures cumulées) (18/03/2014) ³⁶²	141
Figure 60. Liaison du bosutinib (en beige) dans le site actif de la Src kinase (en jaune) mettant en jeu des liaisons hydrogènes médiées par des molécules d'eau (W1 et W2) ³⁷²	143
Figure 61. Différence de positionnement de l'hélice H12 du récepteur PPAR_alpha lors de la liaison (a) d'un agoniste (code 1I7G) ou (b) d'un antagoniste (code 1KKQ).....	144
Figure 62. Illustration de la création de la protéine "unifiée" de FlexE ³⁸⁰	145

Figure 63. Calcul des valeurs de RSR pour le ligand co-cristallisé (RSRc) et la pose prédite (RSRd) à partir des cartes de densité électronique théoriques du ligand de référence (jaune) ou de la pose prédite (bleu) et de la densité électronique mesurée expérimentalement (vert) ⁴¹²	151
Figure 64. Protocole de sélection des decoys de la DUD dans la ZINC (avec PM: Poids Moléculaire, HBD: donneur de liaison hydrogène, HBA: accepteur de liaison hydrogène, logP: coefficient de partage eau-octanol, rotB, nombre de liaisons rotatives) d'après ⁴²³	154
Figure 65. Les méthodes 1 et 2 présentent le même taux d'actifs retrouvés pour la fraction considérée de la chimiothèque et donc des facteurs d'enrichissement similaires tout en affichant des distributions d'actifs et de decoys tout à fait dissemblables. Les valeurs d'EF indiquent que les deux méthodes considérées possèdent des performances identiques, alors qu'en réalité la méthode 1 semble plus performante.	157
Figure 66. Les courbes de ROC représentent l'évolution, pour chaque fraction de la chimiothèque, de la sensibilité (Se) et de la spécificité (Sp), avec Se fonction de 1 – Sp. Une classification aléatoire des composés de la chimiothèque est représentée par une diagonale allant du point (0,0) au point (1,1) du graphique. Si une méthode permet de discriminer les actifs et les decoys de manière plus efficace que le hasard, la majorité des points de la courbe de ROC est située au-dessus de la diagonale. Pour une distribution idéale, c'est-à-dire que les actifs sont tous classés en premier, la courbe monte en ligne droite jusqu'au point (0,1) (Se = Sp = 1 pour tous les actifs) puis viennent ensuite les inactifs (Se = 1, Sp = 0 pour tous les inactifs) représentés par une ligne horizontale entre les points (0,1) et (1,1) du graphique (d'après ⁴²⁹)	158
Figure 67. Les AUC des courbes 1 et 2 présentent des valeurs similaires, indiquant que les deux méthodes de criblages virtuels présentent des performances globales identiques. Cependant, la méthode associée à la courbe rouge est plus efficace pour discriminer de façon précoce les actifs des decoys ⁴⁰⁹	Erreur ! Signet non défini.
Figure 68. Protocole de sélection des structures de l'EGFR pour l'étude, en fonction du volume et de l'ouverture du site de liaison et utilisation de ces structures pour le docking des ligands et des decoys du jeu de données correspondant de la DUD	165
Figure 69. Ligands de la GPB inclus dans la DUD mais ne se fixant pas au site catalytique étudié mais à un site allostérique.....	189
Figure 70. Classification des cibles de la DUD selon les grandes familles de protéines.....	191

Figure 71. Classification des récepteurs nucléaires en « endocrins » ou « orphelins » selon l'identification préalable d'un ligand endogène. Les RNs « endocrins » peuvent former des homodimères ou des hétérodimères pour se lier à des parties spécifiques de l'ADN	470	193
Figure 72. Illustration du contrôle transcriptionnel ligand dépendant exercé par les récepteurs nucléaires. La liaison du ligand provoque la dimérisation du récepteur et sa translocation dans le noyau cellulaire. Le dimère ainsi formé reconnaît et se lie à des éléments spécifiques de la séquence de l'ADN situés dans les régions régulatrices des gènes cibles pour pouvoir contrôler leur transcription grâce au recrutement de co-régulateurs	470	194
Figure 73. Représentation schématique des différents domaines constituant un récepteur nucléaire. Les RNs partagent une très grande similarité structurale et de fonctions. Cependant, le domaine A/B est la région la moins conservée au sein des RNs et contient une fonction d'activation transcriptionnelle (AF1) indépendante de la liaison d'un ligand. Le domaine C est hautement conservé et constitue le domaine de liaison à l'ADN (DNA Binding Domain DBD) sous la forme de deux doigts de zinc. Le domaine de liaison des ligands est quant à lui inclus dans la partie C-terminale E/F et est composé de 12 hélices formant en son centre une poche hydrophobe pouvant accueillir un ligand. Ce LBD contient aussi une fonction d'activation transcriptionnelle (AF2), qui est cette fois, ligand dépendante.	470	196
Figure 74. Contenu des jeux de données agoniste et antagoniste de la NRLiSt BDB		198
Figure 75. Représentation schématique des données contenues dans la NRLiSt BDB		210
Figure 76. Superposition des sites de liaison de RAR α co-cristallisé avec un ligand antagoniste (bleu) et du RAR γ co-cristallisé avec un agoniste (rose) illustrant le clash stérique (flèche rouge) se produisant entre l'extension du ligand antagoniste et l'hélice 12 en position agoniste	480	211
Figure 77. Stratégie prévisionnelle d'intégration de nouveaux decoys expérimentaux dans la version mise à jour de la NRLiSt BDB		253
Figure 78. Représentation schématique d'une articulation normale (a) et d'une articulation atteinte dans un cas de polyarthrite rhumatoïde (b). Dans une articulation, deux extrémités osseuses se font face, couvertes par une couche de cartilage, séparées par un espace et entourées de la membrane synoviale et de la capsule articulaire. L'atteinte articulaire observée en cas de polyarthrite rhumatoïde est initialement caractérisée par une inflammation de la membrane synoviale résultant de l'afflux et l'activation locale de		

cellules variées (lymphocytes B et T, cellules plasmatiques et dendritiques, macrophages, mastocytes). Les ostéoclastes sont responsables des destructions survenant au niveau articulaire observées dans la polyarthrite rhumatoïde (c et d : radiographies respectives d'une main saine et atteinte de polyarthrite rhumatoïde) et qui débutent généralement à la jonction membrane synoviale-cartilage-os (la portion détruite de la membrane synoviale est appelé « pannus »). ⁵⁰¹	257
Figure 79. L'IL-6 régule l'action de nombreux partenaires cellulaires, soit par induction de leur production, inhibition de leur synthèse ou stimulation de leur différenciation. ⁵¹¹ .	261
Figure 80. Formation du complexe hexamérique activé IL-6/IL-6R/gp130 par assemblage séquentiel et coopératif des 3 partenaires. Après liaison de l'IL-6 à l'IL-6R, un complexe trimérique est formé par addition de la gp130. Deux complexes ainsi formés peuvent ensuite interagir pour former le complexe hexamérique fonctionnel (avec D1 : domaine d'activation « immunoglobulin-like » de la gp130 et D2D3 : région homologue de liaison des cytokines CHR (Cytokine binding Homology Region) de la gp130 et de l'IL-6R) ⁵²²	262
Figure 81. Le tocilizumab, commercialisé sous le nom Actemra® est un anticorps anti-IL-6R (avec sIL-6 Receptor : la forme soluble de l'IL-6R et mIL-6 Receptor : la forme transmembranaire de l'IL-6R) ⁵²⁷	263
Figure 82. Représentation schématique du protocole mis en place au laboratoire GBA pour la découverte d'inhibiteurs de l'IL-6.....	263
Figure 83. Structure du complexe IL-6 (beige) / IL-6R (rose) / gp130 (jaune) décrits dans la PDB sous le code 1P9M par Boulanger et ses collègues ⁵²²	264
Figure 84. Structure du complexe IL-6/gp130 utilisée pour le docking avec le protocole généré avec le logiciel Surflex-Dock et édité manuellement pour couvrir le site de liaison choisi	265
Figure 85. La formation du complexe actif IL-6/IL-6R/gp130 permet l'activation des protéines de la famille Janus : Janus Kinase 1 (JAK1) et 2 (JAK2) et Tyrosine kinase 2 (Tyk2). Ces kinases ainsi activées induisent la phosphorylation, la dimérisation et la translocation dans le noyau du facteur de transcription STAT3 (Signal transducer and activator of transcription 3) qui se lie à des éléments activateurs des gènes inductibles par l'IL-6. Dans les cellules HEK-Blue TM IL-6, l'activation de la voie JAK/STAT3 provoque la sécrétion de SEAP (Secreted Embryonic Alkaline Phosphatase) ⁵²⁹	266

Figure 86. Clivage du XTT tetrazolium en XTT formazan par l'activité succinate deshydrogenase mitochondriale en présence de phenazine methosulfate (PMS).	267
Figure 87. Les tests HEK-Blue TM TNF α /IL-1 β (a) et HEK-Blue TM IL-4 (b) reposent sur l'activation de la production de la protéine SEAP ensuite détectée par le test colorimétrique QUANTI-Blue TM . (a) Les cellules HEK-Blue TM TNF α /IL-1 β permettent de mesurer le potentiel du TNF α et de l'IL-1 β à se fixer sur leurs récepteurs et plus particulièrement à induire l'activation de la voie NF-kB ⁵³² . (b) Les cellules HEK-Blue TM IL-4 mettent à profit la voie STAT6 pour évaluer le potentiel de l'IL-4 à se fixer sur son récepteur ⁵³³	269
Figure 88. Le test de liaison IL-6/IL-6R se déroule en plusieurs étapes successives séparées par des phases de lavage des puits. (a) Fixation de la gp130 au fond des puits, et saturation des puits avec de l'albumine de sérum bovin (BSA). (b) Incubation de l'IL-6 avec son récepteur l'IL-6R et avec le ligand à tester. Le ligand ne se fixe pas à l'IL-6, celle-ci est donc libre d'interagir avec son récepteur IL-6R pour ensuite former un trimère avec la gp130 fixée dans le puits. Après lavage, au cours duquel le ligand qui ne s'est pas fixé est éliminé, un anticorps anti-IL6 associé à la biotine est ajouté. Celui-ci peut se lier à l'IL-6 du complexe fixé dans le puits et n'est pas éliminé par lavage. La révélation de la présence des anticorps anti-IL-6 (et donc de la présence de l'IL-6) se fait avec de l'avidine-HRP (Horseradish Peroxidase) qui va se coller à la biotine sur l'anticorps et va générer un produit coloré, en réaction avec son substrat le TMB (3,3',5,5'-tétraméthylbenzidine). La réaction est stoppée avec du H ₂ SO ₄ et l'absorbance est mesurée par spectrophotométrie à 450 nM. (c) Incubation de l'IL-6 avec son récepteur l'IL-6R et avec le ligand à tester. Le ligand se fixe à l'IL-6 qui ne peut donc pas interagir avec son récepteur IL-6R. L'IL-6 et l'IL-6R ne peuvent donc pas se fixer dans les puits et sont éliminés par lavage des puits. Lorsque l'anticorps anti-IL6 est ajouté, il ne trouve pas d'IL-6 auquel se lier et est donc éliminé par lavage des puits. Aucune réaction colorée n'est alors observée.	270

Liste des équations

Équation 1. Description d'une molécule A par un vecteur de n attributs x	78
Équation 2. Mesures de distance de Hamming (t=1) et Euclidienne (t=2) entre 2 composés A et B ¹²³	79
Équation 3. Mesures de distance de Hamming (t=1) et Euclidienne (t=2) entre 2 composés A et B lorsque des descripteurs dichotomiques sont employés, avec a et b les nombres de bits des vecteurs des descripteurs égaux à 1 pour respectivement la molécule A et la molécule B, et c le nombre de bits des descripteurs égaux à 1 communs aux deux molécules A et B ¹²³	79
Équation 4. Formule du coefficient de Tanimoto pour deux molécules A et B en utilisant des empreintes 2D ou 3D comme descripteurs, avec a et b les nombres de bits des vecteurs des descripteurs égaux à 1 pour respectivement la molécule A et la molécule B, et c le nombre de bits des descripteurs égaux à 1 communs aux deux molécules A et B ¹²³	80
Équation 5. Formule du coefficient de Tanimoto pour deux molécules A et B en utilisant des variables continues	80
Équation 6. Formule du coefficient de Dice pour deux molécules A et B en utilisant des empreintes 2D ou 3D comme descripteurs, avec a et b les nombres de bits des vecteurs des descripteurs égaux à 1 pour respectivement la molécule A et la molécule B, et c le nombre de bits des descripteurs égaux à 1 communs aux deux molécules A et B ¹²³	80
Équation 7. Formule du coefficient de Dice pour deux molécules A et B en utilisant des variables continues ¹²³	81
Équation 8. Formule du coefficient de Cosine pour deux molécules A et B en utilisant des empreintes 2D ou 3D comme descripteurs, avec a et b les nombres de bits des vecteurs des descripteurs égaux à 1 pour respectivement la molécule A et la molécule B, et c le nombre de bits des descripteurs égaux à 1 communs aux deux molécules A et B ¹²³	81
Équation 9. Formule du coefficient de Cosine pour deux molécules A et B en utilisant des variables continues ¹²³	81
Équation 10. Formule du coefficient de Pearson pour deux molécules A et B en utilisant des empreintes 2D ou 3D comme descripteurs, avec a et b les nombres de bits des vecteurs des descripteurs égaux à 1 pour respectivement la molécule A et la molécule B, c le nombre de bits des descripteurs égaux à 1 communs aux deux molécules A et B et n la longueur de la chaîne de bits ¹²⁸	82

Équation 11. Formule du coefficient de Pearson pour deux molécules A et B lorsque les descripteurs constituent des variables continues	82
Équation 12. Calcul de la similarité des parties communes entre deux molécules (sim(m)) avec s un facteur de pondération approprié (d'après ¹⁴²)	87
Équation 13. Calcul du nombre de conformations possibles ($N_{\text{conformations}}$) pour un ligand avec N le nombre de liaisons rotatives, n_{inc} le nombre d'incrémentations et $\theta_{i,j}$ la valeur de l'angle incrémental rotationnel j pour la liaison i ³⁰⁰	125
Équation 14. Probabilité P d'acceptation d'une nouvelle conformation selon le critère de Metropolis (E_{new} : énergie de la nouvelle conformation, E_{old} : énergie de l'ancienne conformation, k : constante de Boltzmann et T : température de simulation)	127
Équation 15. Equation du mouvement de Newton pour un système atomique (F: force, m: masse, a : accélération).....	129
Équation 16. Calcul de la force (F) de chaque atome par changement dans l'énergie potentielle (dE) entre deux positions de distance r_i	130
Équation 17. Calcul des positions atomiques à chaque intervalle de temps (d^2r_i/dt^2) en fonction de la force F et de la masse atomique m	130
Équation 18. L'enthalpie libre de liaison (ΔG_{bind}) peut être exprimée en fonction de la constante des gaz parfaits (R), de la température (T) et de la constante d'équilibre du complexe ligand-récepteur, égale au ratio constante d'association (k_a) constante de dissociation (k_d).....	131
Équation 19. A température (T) et pression constante, l'enthalpie libre de liaison (ΔG_{bind}) peut être reliée aux variations d'enthalpie (ΔH) et d'entropie (ΔS).....	131
Équation 20. Calcul de l'énergie potentielle d'un système (E_{MM}) par les champs de force de la mécanique moléculaire (K_r , K_θ et K_ϕ : facteurs de pénalité pour les liaisons, les angles et les angles dièdres ; r et r_{ref} : longueurs des liaisons mesurées et de référence ; θ et θ_{ref} : valeurs des angles mesurées et de référence ; ϕ et ϕ_{ref} : valeurs des angles dièdres mesurées et de référence ; A_{ij} et B_{ij} : constantes attractives et répulsives mesurées expérimentalement ; R_{ij} : distance entre les atomes i et j ; q_i et q_j : charges des atomes i et j ; ϵ : constante diélectrique).....	133
Équation 21. Evaluation de l'enthalpie libre de liaison ΔG_{bind} par les potentiels de Lennard-Jones et Coulomb	133
Équation 22. Equation de la fonction de score d'ICM sommant des termes de van der Waals (E_{vdw}) de liaisons hydrogène (E_{HB}), de torsion (E_{torsion}), électrostatique (E_{el}) de	

solvation (E_{solv}), de déformation de liaisons (E_{bond}), de flexion d'angle (E_{angles}), de déformation de l'angle de phase (E_{phases}) de restrictions de distances ($E_{\text{dist.restr.}}$), d'attaches (E_{tethers}) et de pénalité de restrictions de variables ($E_{\text{var.restr.}}$) (avec $F=0,5$ pour les atomes séparés par 3 liaisons covalentes et $F=1$ dans tous les autres cas, A et C : constantes de van der Waals, d : distance entre les atomes, A' et B : constantes de liaisons hydrogènes, K, K_b , K_o , K_Φ : constantes de force pour la torsion, les déformations de liaisons, les flexions d'angles et les déformations de l'angle de phase dièdre ; φ : angle de torsion, q_α et q_β : charges des atomes ; ϵ : constante diélectrique ; K_s : paramètres de solvation, AAS : surface atomique accessible au solvant ; b_0 , ω_0 et Φ_0 : valeurs de référence de distances de liaisons, d'angles et d'angles de phase dièdres ; W_p et W_τ : coefficients de restrictions de distance et d'attaches ; D_U et D_L : valeurs hautes et basses pour les restrictions de liaisons ; U : constante de restriction de variable ; δ : distance normalisée ; $F' : F'$: fraction de la dimension occupée par le fond plat de la fonction)³¹⁷ 134

Équation 23. Equation de la fonction de score Surfex-dock tentant de prédire l'affinité de liaison $-\log(K_d)$ d'un complexe à l'aide de termes stérique, polaire et d'entropie (l_1 :facteur stérique gaussien d'attraction, l_2 :facteur stérique sigmoïde de répulsion, l_3 : facteur stérique de pénétration, l_4 :facteur polaire gaussien d'attraction, l_5 : facteur polaire sigmoïde de répulsion, l_6 : facteur polaire d'inadéquation, l_7 : facteur du nombre de liaisons rotatives, l_8 : facteur du poids moléculaire, n_1 : position gaussienne stérique, n_2 : propagation gaussienne stérique ; n_3 :raideur de la pente sigmoïde, n_4 : point d'inflexion sigmoïde stérique, n_5 : tolérance de van der Waals aux clashes stériques, n_6 :position polaire gaussienne, n_7 : propagation gaussienne stérique, n_8 : point d'inflexion sigmoïde polaire, n_9 : tolérance de van der Waals pour les clashes polaires, n_{10} :point d'inflexion de la sigmoïde de direction polaire, n_{11} : facteur de charge, n_{12} : position gaussienne de la répulsion polaire, n_{13} : propagation gaussienne de la répulsion polaire)³⁴⁷ 136

Équation 24. La fonction de score basée sur les connaissances PMF peut être reliée à l'enthalpie de liaison libre ΔG_{bind} à l'aide d'un facteur d'échelle (ϵ) représentant tous les termes traités implicitement dans le modèle. Elle est définie comme étant la somme de toutes les interactions de paires d'atomes du complexe protéine-ligand (avec $r_{\text{cut_off}}^{ij}$ la valeur seuil de distance pour les paires kl d'atomes de type ij ; r : distance entre paires d'atomes ; k_B : constante de Boltzmann, T : température absolue ; $f_{\text{Vol_corr}}^j(r)$: facteur de correction du volume du ligand ; $\rho_{\text{seg}}^{ij}(r)$: densité de paires d'atomes ij pour une certaine

distance « seg » ; ρ_{bulk}^{ij} : distribution de i et j lorsqu'il n'y a pas d'interaction entre i et j) 351	137
Équation 25. Calcul du RMSD (avec v_i et w_i les atomes identiques de la structure expérimentale et de la structure prédite respectivement ; x, y et z les coordonnées cartésiennes ; n le nombre total d'atomes)	149
Équation 26. Calcul de l'erreur relative de déplacement (RDE) pour un ligand de N atomes i par rapport au ligand co-cristallisé de N atomes i' (avec L: le paramètre d'échelle définissant l'échelle de précision et dont la valeur est couramment située entre 1,5 et 3 Å ; et $D_{ii'}$ la déviation de l'atome i prédit par rapport à l'atome i' de référence) ⁴¹¹	150
Équation 27. Calcul du RSR à l'aide de la densité électronique observée expérimentalement (ρ_{obs}) et de la densité électronique calculée à partir des coordonnées atomiques du modèle : le ligand co-cristallisé ou la pose prédite (ρ_{calc}) ⁴¹²	150
Équation 28. Calcul du ratio RSR_n à partir des valeurs de RSR des poses prédites (RSR_d) et du ligand co-cristallisé (RSR_c) ⁴¹²	151
Équation 29. La sélectivité (Se) est définie comme le ratio du nombre d'actifs retrouvés par la méthode de docking dans une fraction donnée de la chimiothèque classée ($N_{\text{actifs}[fraction]}$ ou VP : vrais positifs) sur le nombre d'actifs total de la chimiothèque ($N_{\text{actifs}[fraction]}$) (avec FN les faux négatifs, c'est-à-dire les actifs non reconnus en tant que tels par la méthode de criblage virtuel) ⁴⁰⁹	155
Équation 30. La spécificité (Sp) représente le ratio du nombre de decoys non classés dans la fraction de la chimiothèque ($N_{\text{inactifs non présents}[fraction]}$ ou VN : vrais négatifs) sur le nombre de decoys total ($N_{\text{inactifs}[total]}$) (avec FP : faux positifs) ⁴⁰⁹	155
Équation 31. Calcul du facteur d'enrichissement (EF) pour la fraction des $100 \cdot n/N$ premiers pourcents de la chimiothèque (avec n: nombre de composés dans la fraction de la chimiothèque étudiée, N : nombre total de composés dans la chimiothèque, VP : vrais positifs, et FN : faux négatifs)	156
Équation 32. L'AUC est calculée en sommant tous les rectangles formés par les valeurs de Se et $1-\text{Sp}$ de chaque point de la courbe de ROC correspondant à la molécule classée à la position i ⁴⁰⁹	158
Équation 33. Le RIE est calculé comme le ratio de la somme des poids de tous les actifs (S) sur la moyenne de la somme ($\langle S \rangle$) obtenue par 1000 tests aléatoires de classification des actifs (avec i le $i^{\text{ème}}$ actif de la chimiothèque et a : le nombre de composés sélectionnés) 430	160

Équation 34. Calcul de la métrique de performance BEDROC à partir du descripteur RIE (avec R_a : le ratio d'actifs et α : le paramètre de reconnaissance précoce calculé à partir du pourcentage θ du score total à z pourcent du rang par l'équation $0=\theta(1-e^{-\alpha})+1-e^{-\alpha z}-1$. La valeur de α usuellement utilisée est de 20 pour que les 8 premiers pourcents (z) des rangs relatifs contribuent pour 80 pourcents (θ) de la valeur de BEDROC) ⁴²⁸	161
Équation 35. Approximation de la valeur de BEDROC lorsque la valeur de αR_a est très inférieure à 1.....	161
Équation 36. Calcul du pourcentage de neutralisation à l'aide des densités optiques du test QUANTI-Blue TM des puits contenant les produits à tester et des différents puits contrôles ($DO_{produit}$: Densité Optique du puit contenant le produit à tester, des cellules et de l'IL-6; $DO_{cellule}$: Densité Optique du puit contenant des cellules uniquement; DO_{IL6} : Densité Optique du puit contenant des cellules et de l'IL-6).....	267
Équation 37. Calcul du pourcentage de neutralisation à l'aide des densités optiques du test XTT des puits contenant les produits à tester et des différents puits contrôles ($DO_{produit}$: Densité Optique du puit contenant le produit à tester, des cellules et de l'IL-6; DO_{blanc} : Densité Optique du puit contenant du milieu de culture uniquement, sans produit ni cellule ni IL-6; DO_{IL6} : Densité Optique du puit contenant des cellules et de l'IL-6)...	268
Équation 38. Calcul du pourcentage de neutralisation à l'aide des densités optiques du test QUANTI-BLUE TM des puits contenant les produits à tester et des différents puits contrôles et du pourcentage de survie ($DO_{produit}$: Densité Optique du puit contenant le produit à tester, des cellules et de l'IL-6; $DO_{cellule}$: Densité Optique du puit contenant des cellules uniquement; DO_{IL6} : Densité Optique du puit contenant des cellules et de l'IL-6).....	268

Liste des abréviations

ACD : Advanced Chemical Directory

ADME-Tox : Absorption Distribution Métabolisme Elimination – Toxicité

ADN : Acide DésoxyriboNucléique

ALH : Accepteur de Liaison Hydrogène

AMM : Autorisation de Mise sur le Marché

ANSM : Agence Nationale de Sécurité du Médicament

APROPOS : Automatic PROtein POcket Search

AR : Androgen Receptor: Récepteur des androgènes

ARMM : Anti-Rhumatisme Modificateur de la Maladie

ARN : Acide RiboNucléique

ASCII : American Standard Code for Information Interchange

ASP : Aire de la Surface Polaire

AUC : Area Under the Curve: Aire sous la courbe

BCUT : Burden-CAS-University of Texas

BEDROC : Boltzmann-Enhanced Discrimination of ROC: discrimination de Boltzmann améliorée des courbes de ROC

BSA : Bovin Serum Albumin : Albumine de sérum bovin

CAR : Constitutive Androstane Receptor: Récepteur constitutive des androstanes

CASTp : Computed Atlas of Surface Topography of proteins

CATS : Chemical Advanced Template Search

CHR : Cytokine binding Homology Region : région d'homologie de liaison des cytokines

CNAM : Conservatoire National des Arts et Métiers

COMBINE : COMparative BINding Energy

CoMFA : Comparative Molecular Field Analysis

CoMSIA : Comparative Molecular Similarity Indices Analysis

CPK : Corey-Pauling-Koltun

CPP : Comité de Protection des Personnes

CRP : C Reactive Protein : Protéine C Réactive

DBD : DNA Binding Domain : Domaine de liaison de l'ADN

DISCO : DIStance COmparison

DLH : Donneur de Liaison Hydrogène
DMARD : Disease Modifying AntiRheumatic Drug
DO : Densité Optique
DUD : Directory of Useful Decoys
DYLOMMS : DYnamic Latice-Oriented Molecular Modeling System
EF : Enrichment Factor: Facteur d'enrichissement
ER : Estrogen Receptor: Récepteur des œstrogènes
ErG : Extended reduced Graph
ERR : Estrogen Related Receptor: Récepteur associé à l'œstrogène
FDA : Food and Drug Administration
FN : Faux négatif
FP : Faux positif
FXR : Farnesoid X Receptor: Récepteur de l'acide biliaire
GAMMA : Genetic Algorithm for Multiple Molecular Alignment
GASP : Genetic Algorithm Superposition Program
GBA : Génomique, Bioinformatique et Applications
GB/SA : Generalized Born / Surface Area
GCNF: Germ Cell Nuclear Factor: Facteur nucléaire des cellules germinales
GDD : GPCR Decoys Database
GLL : GPCR Ligands Library
GOLPE : Generating Optimal Linear PLS Estimations
GPB : Glycogen Phosphorylase Beta
GR : Glucocorticoid Receptor: Récepteur des glucocorticoïdes
GWAS : Genome Wide Association Study: Etude d'association pangénomique
HBA : Hydrogen Bond Acceptor: Accepteur de liaison hydrogène
HBD : Hydrogen Bond Donor: Donneur de liaison hydrogène
hERG : human Ether-a-go-go Related Gene
HIV : Human Immunodeficiency Virus: virus de l'immunodéficience humaine
HNF4 : Hepatocyte Nuclear Factor 4: Facteur nucléaire hépatocytaire 4
HRP : HorseRadish Peroxide
HTS : High Throughput Screening : Criblage à haut débit
HypoGen : Hypothesis Generator

IBAC : Interactions-Based Accuracy Classification: Classification de précision basée sur les interactions

ICM : Internal Coordinate Mechanic

IFREDA : ICM Flexible REceptor Docking Algorithm

IL-1 : Interleukine 1

IL-4 : Interleukine 4

IL-6 : Interleukine 6

IL-6R : Récepteur de l'interleukine 6

IUPAC : International Union of Pure and Applied Chemistry

IUPHAR : International Union of basic and clinical PHARmacology database

JAK : JA Kinase

LBD : Ligand Binding Domaine: Domaine de liaison du ligand

LBVLS : Ligand Based Virtual Ligand Screening: Criblage virtuel basé sur les ligands

LBVS : Ligand Based Virtual Screening: Criblage virtuel basé sur les ligands

LMC : Leucémie Myéloïde Chronique

logP : coefficient de partition eau / octanol

LXR : Liver X Receptor : Récepteur des oxystérols

MCS : Maximun Common Substructure: Sous-structure commune maximale

MCSS : Multiple Copy Silmutaneous Search

MPHIL : Mapping PHarmacophore In Ligands

MR : Mineralocorticoid Receptor: Récepteur des minéralocorticoïdes

NGFIB : Nerve Growth Factor IB: Facteur de croissance des nerfs IB

NRLiSt BDB : Nuclear Receptors Ligands and Structures Benchmarking DataBase

Nurr1 : Nuclear receptor related 1: Récepteur nucléaire associé

NURSA : NUclear Receptor Signaling Atlas

PARP :Poly ADP-Ribose Polymerase

PCA : Principal Component Analysis: Analyse en composante principale

PDB : Protein Data Bank

PDGFR : Platelet Derived Growth Factor Receptor

PHASE : PHarmacophore Alignment and Scoring Engine

PINTS : Patterns In Non-homologous Tertiary Structures

PLS : Partial Least Square: Méthode des moindres carrés partiels

PM : Poids Moléculaire

PMF : Potential of Mean Force

PPAR : Peroxisome Proliferator Activated Receptor

PR : Progesterone Receptor: Récepteur de la progestérone

PXR : Pregnane X Receptor : Récepteur des xénobiotiques

QSAR : Quantitative Structure-Activity Relationship: Relation quantitative structure-activité

R&D : Recherche et Développement

RANKL : Receptor Activator of NF- κ B Ligand: récepteur activateur du ligand NF- κ B

RAPID : RAndomized Pharmacophore Identification for Drug design

RAR : Retinoid Acid Receptor : Récepteur de l'acide rétinoïque

RD-QSAR : Receptor Dependent Quantitative Structure-Activity Relationship : Relation quantitative structure-activité dépendante du récepteur

RDE : Relative Displacement Error: Erreur relative de déplacement

REACH : Registration, Evaluation, Autorisation and Restriction of Chemicals

RIE : Robust Initial Enhancement: Amélioration robuste initiale

RMN : Résonance Magnétique Nucléaire

RMS : Root Mean Square: Moyenne quadratique

RMSD : Root Mean Square Deviation: Ecart quadratique moyen

RNs : Récepteurs Nucléaires

ROC : Receiver Operating Characteristic

ROCS : Rapid Overlay of Chemical Structures

ROR : Retinoid-related Orphan Receptor : Récepteur orphelin lié aux rétinoïdes

RSA : Relations Structure-Activité

RSR : Real Space R-factor : Espace réel du facteur R

RTI : Record Type Indicator

RXR : Retinoid X Receptor : Récepteur des rétinoïdes X

SBVLS : Structure Based Virtual Ligand Screening: Criblage virtuel basé sur les structures

SBVS : Structure Based Virtual Screening: Criblage virtuel basé sur les structures

SCAMPI : Statistical Classification of Activities of Molecules for Pharmacophore Identification

SDF : Structure Data File

Se : Sensibilité

SEAP : Secreted Embryonic Alkaline Phosphatase

SIDA : Syndrôme d'ImmunoDéficiency Acquis

SMART : SMOoth molecularAR surface Triangulation

SMILES : Simplified Molecular-Input Line-Entry System

Sp : Spécificité

STAT3 : Signal Transducer and Activator of Transcription 3

Tc : coefficient de Tanimoto

TMB : 3,3',5,5'-tetraméthylbenzidine

TNF : Tumor Necrosis Factor : Facteur de nécrose tumorale

TPSA : Topologic Polar Surface Area : Aire de surface polaire topologique

TR : Thyroid Receptor: Récepteur des hormones thyroïde

Tyk: Tyrosine kinase

USR : Ultrafast method for Shape Recognition

VDR : Vitamin D Receptor : Récepteur de la vitamine D

VEGF : Vascular Endothelium Growth Factor : Facteur de croissance de l'endothélium vasculaire

VP: Vrai positif

VN: Vrai négatif

VS : Virtual Screening : Criblage virtuel

WDI : World Drug Index

XTT : 2,3-Bis(2-Methoxy-4-Nitro-5-Sulfohenyl)-2H-Tetrazolium-5-carboxanilide

Première partie

Introduction

1 Découverte de nouveaux médicaments

1.1 Histoire de la découverte des médicaments

La transmission des savoirs ancestraux médicaux, reposant sur la capacité de partager les connaissances acquises, est depuis toujours un facteur déterminant de la découverte et de l'utilisation des médicaments. Longtemps limitée à la transmission orale, que ce soit lors de simples conversations ou par l'entremise d'histoires et de poèmes récités par les bardes, la publication rapide et mondiale des résultats scientifiques a largement contribué aux nombreux succès de la recherche moderne de médicaments ¹. Pendant longtemps, des substances naturelles, principalement d'origine végétale mais aussi minérale et animale, étaient utilisées. Ces substances étaient sélectionnées par l'observation empirique de leurs effets sur le cours des maladies. Cependant, le début du XIXe siècle marque un tournant historique dans la recherche de nouveaux médicaments grâce à l'isolement de principes actifs de substances naturelles précédemment utilisées dans la médecine traditionnelle. Ainsi, la morphine est isolée du pavot en 1803 et ses effets sont décrits par Friedrich Wilhelm Adam Sertürner quelques temps plus tard ² (Figure 1). L'essor des médicaments de synthèse débute dans le milieu du XIXe siècle avec la synthèse en 1853 de l'acide acétylsalicylique par Charles Gerhardt ^{3,4} puis sa commercialisation vingt ans plus tard sous le nom de marque « Aspirin » par les Laboratoires Bayer ⁵. L'exploration de ces savoirs ancestraux est toujours d'actualité et constitue une discipline à part entière, l'ethnopharmacologie. De même, certaines substances naturelles traditionnelles sont toujours utilisées pour leurs vertus thérapeutiques⁶. Cependant, la nécessité de développer d'autres sources de découverte de médicaments est rapidement apparue et a permis de nombreuses avancées tout au long du XXe siècle et jusqu'à aujourd'hui (Figure 1).

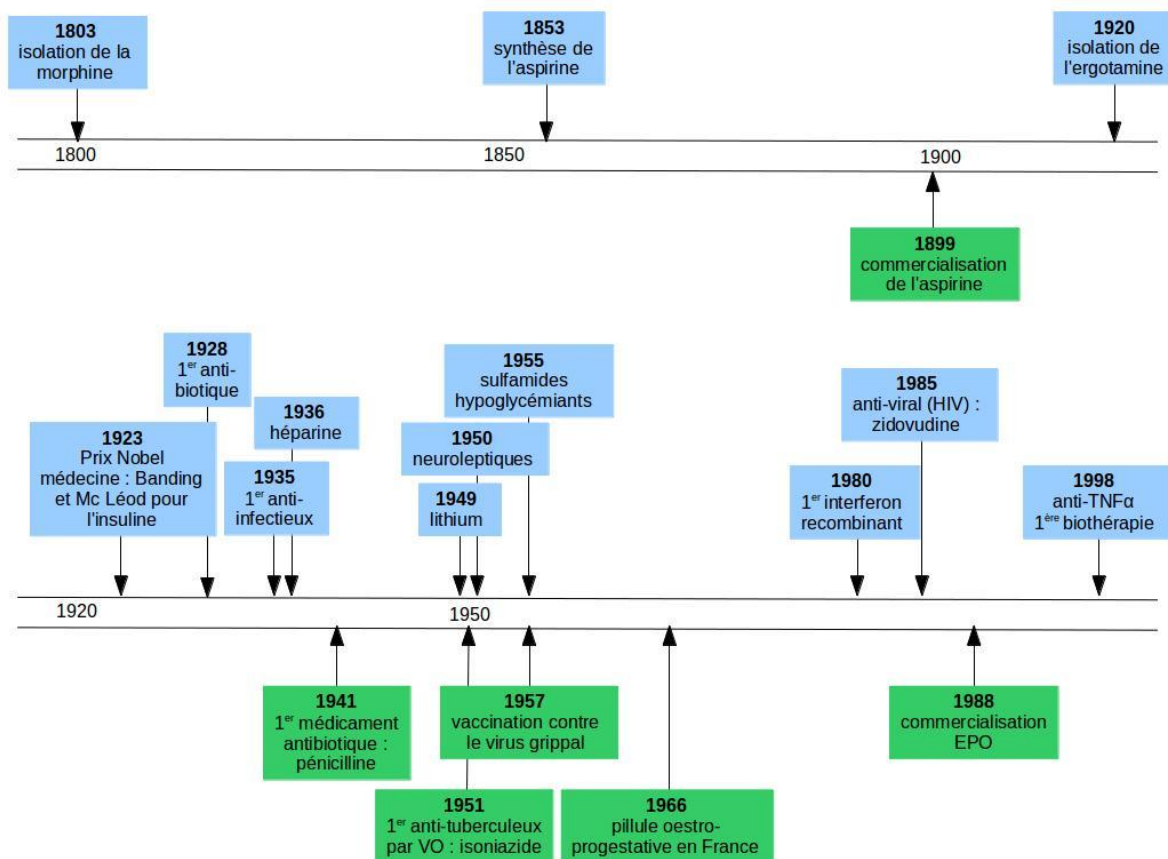


Figure 1. Quelques grandes étapes de la découverte de médicaments au cours du XIXe et XXe siècle (d'après ⁷)

Quatre grandes disciplines ont totalement révolutionné la recherche de nouveaux médicaments et abouti aux processus actuellement utilisés dans les phases précoces de recherche et développement (R&D). Il s'agit de la génétique qui a permis la rationalisation du choix et de l'utilisation de cibles biologiques définies au niveau moléculaire ⁸, la pharmacologie avec le développement de techniques de criblage à haut débit (High Throughput Screening ou HTS), la bioinformatique avec la mise en pratique de criblage virtuel, et la chimie, avec le développement de la chimie combinatoire.

1.2 Schéma général de R&D

Le processus de R&D de découverte de nouveaux médicaments, ou « drug discovery », est extrêmement long et coûteux et regroupe l'ensemble des étapes aboutissant à la mise sur le marché du nouveau médicament (Figure 2).



Figure 2. Modèle de R&D pour développer avec succès un nouveau médicament présentant les différentes étapes ainsi que le nombre de molécules et les coûts en terme monétaire et temporel associés à chaque étape (d'après ⁹)

Ainsi, le temps s'écoulant entre la première étape, correspondant au choix d'une cible thérapeutique adéquate et pertinente, et la dernière étape qu'est la mise sur le marché d'un médicament, est estimé entre 12 et 14 ans avec un coût total entre 800 millions et 1 milliard de dollars. ^{9,10}

A chaque étape, les procédés utilisés doivent donc être optimisés pour tenter de réduire le coût et la longueur des étapes, mais aussi en vue d'obtenir les molécules les plus actives possibles et les plus susceptibles d'aboutir à un médicament.

1.2.1 Choix d'une cible thérapeutique

Dans l'article L. 5111-1 du Code la Santé Publique ¹¹, un médicament est défini par ses « propriétés curatives ou préventives à l'égard des maladies ». Le processus de découverte d'un nouveau médicament doit donc débiter par la définition d'une maladie pour laquelle le défaut de traitement adapté et efficace engendre un réel besoin médical. Il s'agit le plus

souvent de pathologies touchant une grande partie de la population, largement étudiées et donc avec un fort potentiel commercial, comme par exemple les maladies neurodégénératives, les cancers, le Syndrome d'ImmunoDéficiency Acquis (SIDA), etc.... Cependant, des maladies dites « orphelines » qui sont des maladies rares, c'est à dire touchant une faible portion de la population et donc moins attractives au niveau financier, présentent un réel besoin de nouvelles thérapies adéquates puisqu'aucun traitement n'est actuellement disponible et peuvent donc constituer des thèmes de recherche très stimulants ¹².

Une fois la maladie à cibler définie, l'étape suivante correspond à l'identification d'une cible biologique potentielle et la validation de sa pertinence. Cette cible peut être diverse selon la maladie étudiée, dans notre cas il s'agira de protéines (voir paragraphe Deuxième Partie Résultats), mais les gènes ou encore les ARNs peuvent aussi constituer des cibles biologiques intéressantes. La dernière estimation portait à 500 le nombre de cibles qui seraient modulées par des médicaments commercialisés, mais ce nombre est en constante augmentation. ¹³ Trois critères sont à prendre en compte pour définir le potentiel d'une entité biologique à devenir une bonne cible pour un médicament : son efficacité, sa sécurité, mais aussi et surtout son caractère « druggable ». La « druggabilité » d'une cible est définie par sa capacité à être modulée par la liaison de petites molécules chimiques ou de produits biologiques plus larges.¹⁴ Elle est couramment estimée par évaluation de l'appartenance de la cible à des familles de gènes déjà ciblées avec succès par des médicaments (récepteurs couplés aux protéines G, kinases...) ¹⁵, mais cette approche est trop réductrice et de nouvelles méthodes permettant d'affiner son estimation sont recherchées ¹⁶.

Parmi les approches utilisées pour identifier une cible biologique, il est possible de citer la recherche bibliographique, la recherche de l'expression d'ARN messagers (ARNm) ou de protéines et la recherche d'associations génétiques. ¹⁷

La recherche bibliographique dans la littérature scientifique de données en rapport avec la maladie est la première méthode à utiliser. L'exploration de données ou « data mining » qui utilise des approches bioinformatiques pour trier des données issues d'une grande variété de sources (publications, brevets, données d'expression des gènes, protéomique, ...) a permis d'améliorer grandement le nombre de cibles identifiées par cette approche ¹⁸.

La recherche de l'expression d'ARNm ou de protéines dans la pathologie étudiée permet aussi d'identifier des cibles biologiques. La corrélation entre l'expression des entités et la progression ou non de la maladie définit de quelle manière la cible doit être modulée. ¹⁷

La recherche d'associations génétiques entre des polymorphismes génétiques et le risque de développer une maladie ou la rapidité de sa progression est une méthode émergente et très prometteuse d'identification de cibles.¹⁷ Ainsi, des études d'associations pangénomiques (ou Genome Wide Association Studies GWAS) recherchent des cibles biologiques impliquées dans la non progression¹⁹ ou au contraire dans la progression rapide²⁰ vers le SIDA.

Avant de rechercher des molécules capables d'agir sur la cible biologique identifiée, il est nécessaire de procéder à la validation de cette cible. Cette validation consiste d'une part à s'assurer de l'effet bénéfique de la modulation de la cible sur la pathologie étudiée tout en vérifiant d'autre part que les conséquences de ces altérations ne seront pas néfastes. De nombreux outils *in vitro* et *in vivo* sont utilisés lors de cette étape de validation, parmi lesquels les animaux transgéniques, les petits ARN interférents (ou small interfering RNA siRNA), les anticorps monoclonaux ou encore la chémogénomique (dont le but est de fournir une petite molécule pour chaque protéine codée par le génome pour explorer les fonctions cellulaires et guider la découverte de nouveaux médicaments²¹).¹⁷

1.2.2 Identification de hits

Une fois la cible identifiée et validée, l'étape suivante est logiquement de tenter de moduler ses effets pour pouvoir agir favorablement sur le processus pathologique, et ceci par interaction avec des molécules. Des composés capables de tels effets sont appelés des touches ou « hits ». La recherche de hits se fait par criblage (ou « screening ») de larges banques de composés aussi appelées « chimiothèques », soit *in vitro*, le plus souvent grâce aux techniques d'HTS, soit *in silico* par criblage virtuel sur ordinateur. Les processus actuels de R&D utilisent très fréquemment une combinaison de ces deux techniques. En effet, selon les cibles biologiques choisies, la mise en place de tests expérimentaux peut s'avérer très compliquée et leur utilisation lors d'un criblage de milliers de composés est parfois tout simplement irréalisable aux vues des coûts aussi bien financiers que temporels. Ainsi, les criblages biologiques et les tests pré-cliniques représenteraient environ 14 % du budget total de R&D et le coût d'un simple programme de HTS est estimé à 75000\$²². Pour pallier ces difficultés, les techniques de criblage *in silico* ont été développées. A l'inverse des techniques *in vitro*, ces méthodes présentent le triple avantage d'être plus aisées à mettre en œuvre, peu coûteuses et relativement rapides selon les capacités de calcul disponibles²³. Cependant, l'inconvénient majeur du criblage virtuel est que ses résultats ne sont que prédictifs et ils doivent donc ensuite être impérativement validés par des tests expérimentaux. Les techniques *in silico*

peuvent donc être utilisées en préambule des techniques *in vitro* en tant que premier filtre des chimiothèques dont les molécules les plus prometteuses seront ensuite criblées expérimentalement²⁴. Aux vues de leur activité, mais aussi de critères annexes tels que leur originalité ou leur stabilité, les composés les plus à même de devenir des médicaments sont sélectionnés comme hits et sont ensuite optimisés.

1.2.3 Génération et optimisation des leads

Cette étape nécessite l'intervention des chimistes médicaux et de chémoinformaticiens qui vont tenter, à partir des hits précédemment identifiés, d'obtenir de nouvelles molécules que l'on qualifiera de têtes de séries ou « leads ». Les leads idéaux sont des composés qui sont plus actifs et plus sélectifs que les hits dont ils sont issus, tout en présentant des propriétés pharmacocinétiques optimales¹⁷. Pour obtenir ces leads, de vastes études de relations structure-activité (RSA) sont menées en faisant varier les structures chimiques des hits par modification des groupements fonctionnels tout en gardant leurs squelettes de bases. Lors de ces études, les informations (activité, sélectivité, propriétés physico-chimiques,...) des composés nouvellement synthétisés sont comparées à celles des hits et permettent de guider la sélection de nouveaux leads²⁵. L'analyse des résultats de RSA pouvant être fastidieuse et complexe, des approches bioinformatiques ont été développées pour assister ces études^(26,27). Les meilleurs leads seront ensuite optimisés en prenant garde de conserver les propriétés favorables d'activité et « drug-like » (voir paragraphe 2.2.3.3) tout en tentant d'améliorer l'affinité, la sélectivité (en vue de minimiser d'éventuels effets secondaires délétères) ainsi que la perméabilité (pour s'assurer que le médicament pourra atteindre sa cible). De plus, lors de la phase d'optimisation des leads, il est important de rechercher les preuves que l'effet biologique observé est bien induit par interaction du lead avec la cible²⁸.

1.2.4 Tests pré-cliniques

Les leads optimisés subissent ensuite une batterie de tests (sur cellules et sur animaux) qui ont pour but de préparer l'utilisation de ces composés à l'échelle humaine²⁸. Cette étape est très importante puisque la poursuite ou non du développement des composés et leur entrée en phases cliniques sont décidées à cette étape. Cette décision représente pour l'équipe de R&D une double prise de risque aux vues du grand coût financier de telles études mais aussi et surtout de la grande responsabilité incombée puisque les candidats médicaments vont être testés à l'échelle humaine. Pour faire ce choix, la faisabilité synthétique à grande échelle de composés, les problèmes de formulation, la pharmacologie et la toxicologie doivent donc être

minutieusement étudiés. Lorsque l'équipe de R&D décide d'engager des études cliniques, toutes les informations obtenues lors des tests pré-cliniques sont rassemblées dans un dossier qui est étudié par les autorités compétentes (en France, il s'agit du comité de protection des personnes, le CPP, et de l'agence nationale de sécurité du médicament et des produits de santé, l'ANSM). Après analyse de ces données, le CPP et l'ANSM vont autoriser ou non le(s) candidat(s) médicament(s) à être testé(s) chez l'homme ¹⁷.

1.2.5 Tests cliniques

Les essais cliniques constituent l'étape critique de tout processus de R&D mais aussi l'étape la plus longue et la plus coûteuse. L'efficacité du candidat médicament chez l'homme, ainsi que sa pharmacocinétique et sa sécurité d'emploi sont évaluées et consignées en vue de la demande d'autorisation sur le marché (AMM) ²⁹. L'administration se faisant chez l'homme, l'encadrement de ces essais au niveau législatif et réglementaire est très strict ³⁰. Ces essais cliniques sont divisés en quatre phases, les trois premières permettent de constituer le dossier d'AMM alors que la quatrième commence dès l'obtention de l'AMM et dure toute la durée de commercialisation du médicament. Au cours de chacune de ses phases, si des effets indésirables inacceptables sont découverts, ces études cliniques peuvent prendre fin et le candidat ou médicament abandonné.

La phase I des essais cliniques se déroule sur un faible nombre de volontaires sains pour s'assurer de la sécurité du candidat médicament en recherchant les effets secondaires qui apparaissent lors de l'administration croissante du candidat médicament. Généralement, 70 % des composés testés passent avec succès cette première phase. Lors de la phase II, un essai randomisé, généralement contre placebo, et incluant une centaine de patients malades est mené pour évaluer l'efficacité du candidat médicament. Seul un tiers des essais cliniques débutés accèdent à la phase III. Celle-ci consiste en une confirmation du potentiel thérapeutique du candidat médicament sur une très large population de malades puisque plusieurs milliers de patients peuvent être inclus. Lors de cette phase, une confirmation de l'efficacité est recherchée et peut être comparée avec celle du traitement de référence et le profil de tolérance (interactions, effets indésirables...) est établi. ²⁸ A la fin de ces 3 phases, toutes les informations obtenues sont rassemblées et le dossier d'AMM est constitué. Seul un candidat médicament sur dix entrant en phase clinique obtient une AMM et est commercialisé ¹⁷ (Figure 3). C'est alors le début de la phase IV des essais cliniques ou pharmacovigilance. Lors de cette phase, le médicament est administré à une immense population hétérogène et de

nouveaux effets secondaires (indésirables ou non) et des interactions inexplorées peuvent apparaître. La surveillance continue permet donc de garantir aux patients une plus grande sécurité d'utilisation.

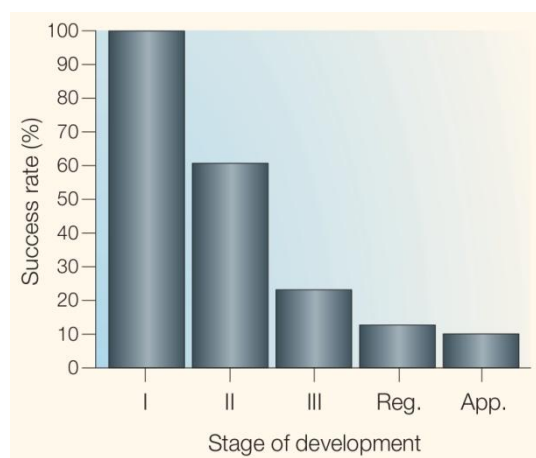


Figure 3. Taux de succès des candidats médicaments à chaque étape des phases cliniques et lors des étapes d'enregistrement (Reg) et de demande d'autorisation de mise sur le marché (App.)³¹

Tout au long du processus de développement d'un médicament, des approches bioinformatiques peuvent être utilisées pour guider la sélection des composés et particulièrement lors des phases précoces par réalisation d'un criblage virtuel.

2 Méthodes de criblages *in silico*

2.1 Généralités

Le criblage virtuel, analogue *in silico* de l'HTS, peut schématiquement être assimilé à un entonnoir dans lequel on verse un grand nombre (généralement entre cent mille et dix millions) de composés, constituant la chimiothèque à cribler, pour obtenir, à l'aide d'un algorithme de criblage, un plus faible nombre (quelques dizaines à quelques milliers) de composés qui seront ensuite testés expérimentalement²⁴.

Le rôle des méthodes de criblage est donc d'éliminer les composés supposés inactifs ou les molécules indésirables tout en priorisant les composés les plus susceptibles d'être actifs. Cependant, très souvent, la liste des composés proposés n'est pas utilisée en l'état et les composés à tester expérimentalement sont sélectionnés manuellement par des experts, c'est ce qu'on appelle le « cherry picking »²⁴.

Deux grandes familles de méthodes de criblages sont distinguées et le choix de leur utilisation est basé sur la disponibilité des données au démarrage du projet³². Ainsi, lorsque la structure tridimensionnelle de la cible biologique a été résolue (cristallographie aux rayons X, résonance magnétique nucléaire ou modèle de structure), les méthodes basées sur la structure (ou « structure-based »), évaluent la capacité des ligands à établir des interactions avec le site de liaison étudié pour sélectionner les molécules capables de se lier à la cible. Les méthodes basées sur les ligands (ou « ligand-based ») peuvent pour leur part être mises en œuvre lorsque les valeurs d'activité pour la cible étudiée d'un ensemble de ligands sont disponibles. Les relations structure-activité de ces molécules sont alors analysées pour découvrir de nouveaux composés susceptibles d'être actifs³³. Lorsque ces deux types de données sont disponibles simultanément, les méthodes « ligand-based » et « structure-based » peuvent toutes deux être utilisées l'une à la suite de l'autre³⁴. Au sein de ces deux grandes familles, différentes méthodes ont été développées (Figure 4).

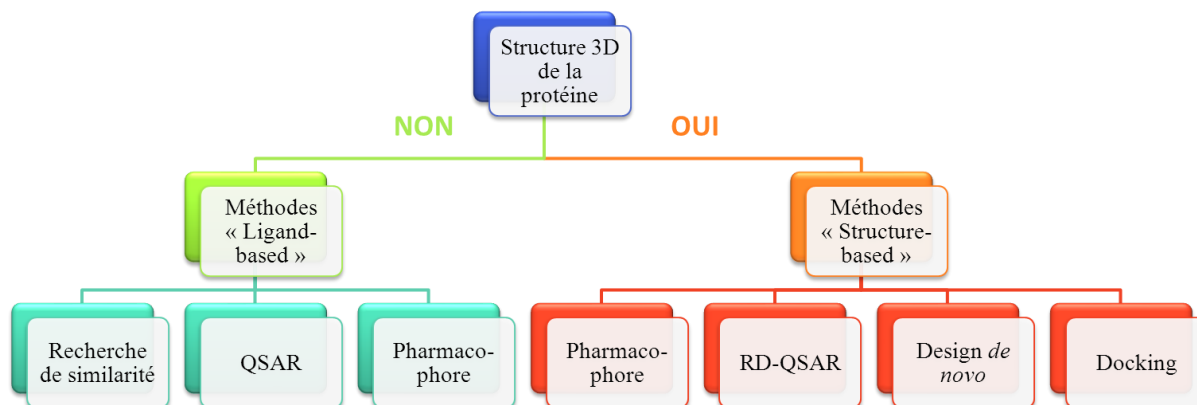


Figure 4. Classification des méthodes de criblage virtuel « ligand-based » et « structure-based » (avec QSAR : Quantitative Structure-Activity Relationship ou modèle de relations quantitatives structure-activité et RD-QSAR : Receptor Dependent-Quantitative Structure-Activity Relationship ou modèle de relations quantitatives structure-activité dépendant du récepteur)

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

L'intérêt porté aux méthodes de criblage virtuel a littéralement explosé au cours des vingt dernières années, comme en témoigne le nombre de publications recensées dans la base de données PUBMED depuis 1997, date de la première publication évoquant le criblage virtuel³⁵ (Figure 5). Il est à noter que les progrès informatiques ont rendu possible les améliorations théoriques des méthodes de screening *in silico* et ont contribué à leur démocratisation.

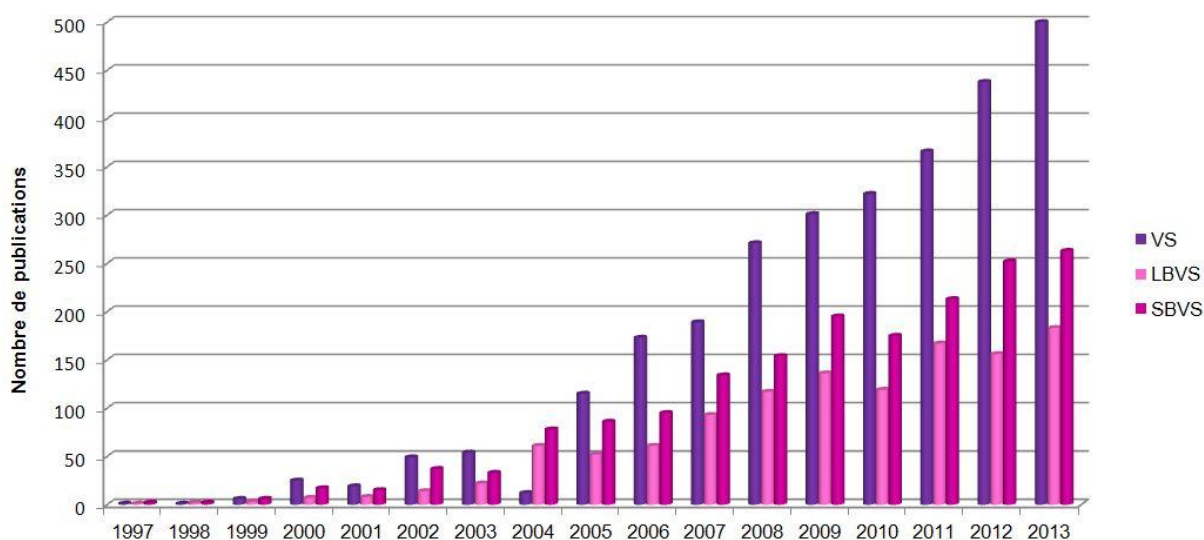


Figure 5. Evolution du nombre de publications de 1997 à 2013 dans la base de données PUBMED en utilisant les mots clés « virtual screening » (VS), « ligand-based virtual screening » (LBVS) et « structure-based virtual screening » (SBVS) (08/01/2014)

Ces publications s'intéressent non seulement à l'utilisation de ces méthodes dans le processus de R&D de nouveaux médicaments mais aussi aux techniques permettant d'évaluer leur fiabilité.

2.2 Les chimiothèques

La probabilité d'identifier un composé capable de se lier spécifiquement à une cible biologique et de moduler son action est relativement faible mais augmente avec le nombre de composés testés et leur diversité. Cependant, avec les techniques actuelles, il serait impossible d'explorer par criblage l'espace chimique dans sa totalité (en général, la taille des chimiothèques criblées ne dépasse pas deux millions de composés) et peu efficace. En effet, le nombre de composés différents théoriquement synthétisable, définissant l'espace chimique virtuel, est estimé ³⁶ à 10^{60} , cependant, l'espace chimique médicinal, c'est à dire les molécules potentiellement utilisables comme médicaments, ne représente qu'une faible portion de l'espace chimique total ³⁷. La proportion en candidats médicaments potentiels (composés « drug-like », voir paragraphe 2.2.3.3.1) des chimiothèques actuelles n'excèdent que rarement le taux de cinquante pour cent ³⁸ (Figure 6) et la qualité des chimiothèques doit donc être améliorée, notamment en terme de diversité, par des méthodes de synthèse divergentes (Diversity-Oriented Synthesis DOS) ou orientées (biology-oriented synthesis), ou par homologie avec les produits naturels qui présentent des squelettes très différents des molécules de synthèse ³³.

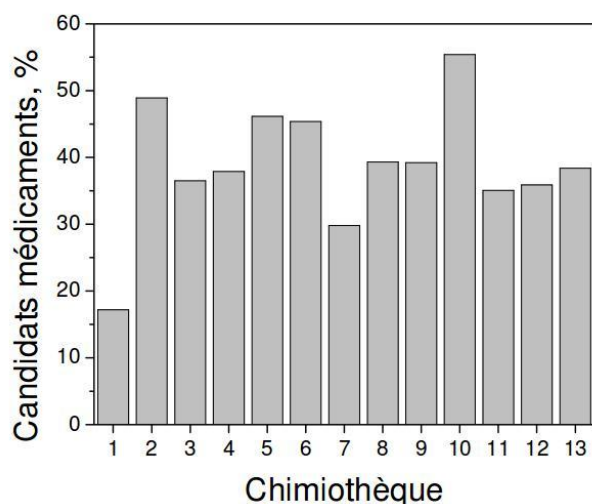


Figure 6. Potentiel en candidats médicaments de collections de criblage commerciales (1 à 7 et 9 à 13) et publique (8 : Chimiothèque Nationale) ³⁸

2.2.1 Différents types de chimiothèques

2.2.1.1 Chimiothèques dans l'espace réel et l'espace global

Les chimiothèques dans l'espace réel (ensemble des composés constituant les chimiothèques privées des laboratoires pharmaceutiques) et l'espace global (ensemble des composés déjà synthétisés) sont constituées de composés potentiellement immédiatement disponibles pour criblage expérimental. ³⁹ Il s'agit le plus souvent de collections virtuelles mise à disposition gratuitement par des fournisseurs ou des institutions et des produits correspondant, conditionnés en vrac ou en microplaques, afin de permettre leur évaluation biologique et disponibles à l'achat auprès de différents fournisseurs.

2.2.1.1.1 Chimiothèques focalisées et généralistes

Les chimiothèques focalisées sont dédiées à une famille de cibles précises. C'est notamment le cas de la NRLiSt BDB ⁴⁰, construite dans notre laboratoire et qui regroupe 9908 ligands pour les récepteurs nucléaires (voir Deuxième Partie Résultats paragraphe 1.2). De nombreuses autres chimiothèques focalisées sont disponibles et quelques-unes sont résumées dans le Tableau 1.

Compagnie	Nom de la chimiothèque	Nombre de composés	Site web
Asinex	Gram negative antibacterial	6045	www.asinex.com
	Lipid GPCRs	2099	
	Ion Channel	11643	
SPECS	Kinase-targeted Library	2720	www.specs.net
Timtec	ActiTarg-G GPCR Ligands	2300	www.timtec.net
	ActiTarg-K Kinase Modulators	6600	
	ActiTarg-P Protease Inhibitors	2000	
	ActiTarg-S Serine Proteinase Inhibitors	900	
	ActiTarg-I Potassium Channels Modulators	1460	
	ActiTarg-N Nuclear Receptors Ligands	1040	
	ActiTarg-H Histone deacetylase inhibitors	1700	
ChemBridge	KINASet	11000	http://www.chembridge.com
	KINACore	6000	
	GPCR Library	13000	
	ION Channel Set	5000	
	IONCore Library	4000	
	NHRCore Library	1200	
Life Chemicals	Kinase Targeted Screening Library	2200	www.lifechemicals.com
	Ion Channel Targeted Libraries	5000	
	GPCR Targeted Libraries	16000	
	Nuclear Receptor Targeted Libraries	9900	
	Protease Targeted Libraries	11500	
	Phosphatase	32000	
	HIV Protease and HIVRT	6700	
	Glutamate Receptors	25000	
	Helicase	6000	
	Polymerase	17700	
	ATPase	3500	
	CYP Inhibitors	13300	
Hepatitis C Virus	16400		

Tableau 1. Liste de quelques compagnies proposant des chimiothèques commerciales focalisées pour différentes familles de cibles biologiques (d'après ³³) (27/01/2014).

A l'opposé, les chimiothèques généralistes (Tableau 2) peuvent être utilisées lors de tous les criblages, leur construction n'ayant pris en compte aucun a priori biologique. Les chimiothèques généralistes présentent globalement un nombre beaucoup plus élevé de composés, leur taille pouvant atteindre plus de 48 millions de composés comme dans la base de données Pubchem substances ⁴¹.

Chimiothèque	Type	Nombre de composés	Données biologiques	Site web
DrugBank	Publique	6825	oui	http://www.drugbank.ca/
Maybridge	Commerciale	53000	non	http://www.maybridge.com/
Chimiothèque nationale	Commerciale	53430	non	http://chimiotheque-nationale.enscm.fr/
MDDR	Commerciale	150000	oui	http://accelrys.com/products/databases/bioactivity/mddr.html
NCI Open database	Publique	250250	non	http://cactus.nci.nih.gov/ncidb2.2/
WOMBAT	Commerciale	270918	oui	http://www.sunsetmolecular.com
Drug Discovery Center Collection	Commerciale	340000	non	http://drugdiscovery.uc.edu/
ChemBridge	Commerciale	950000	non	http://www.chembridge.com
ChemBank	Publique	1200000	oui	http://chembank.broadinstitute.org/
ChEMBL	Publique	1324941	oui	https://www.ebi.ac.uk/chembl/
ChemDiv	Commerciale	1500000	non	http://eu.chemdiv.com/
Enamine	Commerciale	1800000	non	http://www.enamine.net/
ChemDB	Publique	5000000	non	http://cdb.ics.uci.edu/
Cococo	Publique	6981556	non	http://cococo.unibo.it/cococo
eMolecules	Commerciale	6000000	non	http://www.emolecules.com/
ACD	Commerciale	7000000	non	http://accelrys.com/products/databases/sourcing/available-chemicals-directory.html
ZINC	Publique	21000000	oui	http://zinc.docking.org/
ChemSpider	Publique	29000000	oui	http://www.chemspider.com/
PubChem	Publique	30000000	oui	http://pubchem.ncbi.nlm.nih.gov/

Tableau 2. Classification des principales chimiothèques généralistes en fonction du nombre de composés proposés, et de leur caractère commercial ou publique (d'après ³³) (27/01/2014).

Les concepteurs de chimiothèques, qu'elles soient focalisées ou généralistes, peuvent renseigner les données d'activité biologiques correspondant aux composés de la chimiothèque. Ces données peuvent être diverses selon la chimiothèque étudiée. Ainsi, la

ChEMBL ⁴², une chimiothèque publique, se base sur plus de 48000 publications issues de 47 journaux scientifiques pour fournir les données d'activités biologiques lorsque celles-ci sont disponibles. Ces données peuvent être des valeurs d'affinité (constante d'inhibition K_i , concentration inhibitrice médiane IC50), d'activité (concentration efficace médiane EC50 ou inhibitrice médiane IC50, Activity, Inhibition) ou autres (Other). La ZINC ⁴³, une autre chimiothèque publique qui regroupe plus de 21 millions de composés, ne propose pas d'annotation directe de ses composés avec des valeurs d'activités biologique mais plutôt des liens vers d'autres bases de données, et notamment la ChEMBL, qui contiennent ces informations.

2.2.1.1.2 Chimiothèques de fragments

Les chimiothèques de fragments ont été développées pour augmenter la diversité de l'espace chimique exploré ⁴⁴. En effet, à partir d'une petite base de données de 100 fragments, l'ensemble des combinaisons obtenues par assemblage de trois fragments différents permet d'obtenir un million de composés ⁴⁵. L'utilisation de fragments permet donc aussi une réelle économie de temps et d'argent pour les criblages virtuels ou expérimentaux tout en présentant un taux de succès d'identification de hits de trois à cinq pourcent, supérieur à celui obtenu lors des criblages expérimentaux à haut débit qui atteint à peine un pourcent ⁴⁴. Par analyse des hits identifiés lors d'approches basées sur les fragments, une « règle de 3 », à l'instar de la « règle de 5 » de Lipinski ⁴⁶ (voir paragraphe 2.2.3.3), a été proposée pour la construction de chimiothèques de fragments optimisées ⁴⁷. Cette règle pose des valeurs seuils pour 6 propriétés physico-chimiques : le poids moléculaire (< 300 Da) le nombre de donneurs et d'accepteurs de liaisons hydrogène (≤ 3), le coefficient de partage clogP (≤ 3), le nombre de liaisons rotatives (≤ 3), et l'aire de surface polaire ($< 60 \text{ \AA}^2$). Le Tableau 3 liste des chimiothèques de fragments respectant cette « règle de 3 ».

Chimiothèque	Nombre de fragments	Site web
Prestwick Fragment Library	2230	http://www.prestwickchemical.com/
Maybridge Ro3 Library	2500	http://www.maybridge.com/
ASINEX's BioFragments	3500	http://www.asinex.com/
ChemDiv Fragment Based Library	4947	http://us.chemdiv.com/
ChemBridge Fragment Library	7000	http://www.chembridge.com/
OTAVA Fragment Library	8445	http://www.otavachemicals.com/
Keyorganics BIONET Fragment Library	13631	http://www.keyorganics.co.uk/
Enamine Fragment Library	28043	http://www.enamine.net/
Lifechemicals General Fragments Library	47500	http://www.lifechemicals.com/

Tableau 3. Classification de quelques chimiothèques de fragments en fonction du nombre de composés mis à disposition (d'après ³³) (30/01/2014).

2.2.1.1.3 Chimiothèques de produits naturels

Avec l'émergence des nouvelles techniques de découverte de nouveaux médicaments (HTS, criblage virtuel ...), l'intérêt porté aux produits naturels a quelque peu décliné. Cependant, les produits naturels présentent de nombreux avantages essentiels pour leur potentiel de médicaments. Ils possèdent notamment une sélectivité très importante pour leurs cibles cellulaires ⁴⁸. De plus, leurs structures sont très différentes des produits classiques de synthèse, ce qui leur permet d'explorer des zones de l'espace chimique auparavant inaccessibles. Ainsi, sur les 13 médicaments issus de la recherche sur des produits naturels ayant reçu l'autorisation de mise sur le marché par la FDA (Food and Drug Administration) entre 2005 et 2007 ⁴⁹, cinq sont les tout premiers représentants de nouvelles classes de médicaments (l'exenatide BYETTA®, la ziconotide PRIALT®, l'ixabepilone IXEMPRA®, la rétapamuline ALTARGO® et la trabectédine YONDELIS®). Pour finir, les produits naturels présentent généralement une meilleure absorption que les produits de synthèse ⁵⁰. Aux vues de ces avantages, un regain d'intérêt a lieu depuis quelques années pour ces produits. Ainsi, en

2008, 126 produits naturels étaient en phases cliniques, et 99 en phases pré-cliniques, pour la majorité issus de plantes mais aussi de bactéries, de champignons, ou encore d'animaux avec de nombreuses indications thérapeutiques parmi lesquelles le traitement du cancer et des infections ⁵¹. Les chimiothèques de produits naturels se sont donc développées pour permettre d'accélérer et de guider la découverte de nouveaux médicaments issus de produits naturels. Ces chimiothèques peuvent être focalisées sur un ensemble de produits naturels traditionnellement utilisés par une population, comme par exemple la TCM Database@Taiwan ⁵² ou l'AfroDb ⁵³, ou être la plus généraliste possible comme l'Universal Natural Products Database ⁵⁴ ou le Dictionary of Natural Products (Tableau 4).

Chimiothèque	Nombre de composés	Site web
TimTec NPL	720	http://www.timtec.net/
AfroDb	954	http://zinc.docking.org/catalogs/afronp
TCM Database@Taiwan	32364	http://tcm.cmu.edu.tw/
Universal Natural Products Database	197201	http://pkuxxj.pku.edu.cn/UNPD
Dictionary of Natural Products	226000	http://dnp.chemnetbase.com/
AMRI Natural Product Libraries	300000	http://www.amriglobal.com/
Super natural II	355076	http://bioinf-applied.charite.de/supernatural_new/

Tableau 4. Classification de quelques chimiothèques de produits naturels en fonction du nombre de composés mis à disposition (31/01/2014).

2.2.1.2 Chimiothèques dans l'espace tangible et virtuel

Les molécules de l'espace tangible (c'est-à-dire qui peuvent être synthétisées facilement par des voies chimiques classiques) et de l'espace virtuel n'ont pas encore toutes été synthétisées et/ou isolées. Leurs structures doivent donc être générées *in silico* puis assemblées dans des formats adéquats pour constituer des chimiothèques virtuelles. Comme nous l'avons vu précédemment, il est actuellement impossible de générer puis d'évaluer l'ensemble des molécules de ces deux espaces. La solution consiste donc à réaliser des chimiothèques virtuelles focalisées par sélection d'un sous-ensemble au sein de la totalité des produits potentiellement synthétisables avec les fragments de départ ou à concevoir *de novo* un ligand présentant la structure la mieux adaptée au site actif et donc théoriquement une activité biologique optimale ⁵⁵.

2.2.2 Formats de chimiothèques virtuelles

Les chimiothèques virtuelles stockent les molécules la constituant dans divers formats chemoinformatiques. On distingue deux grands types de formats : les formats 2D et les formats 3D qui diffèrent des premiers par l'apport d'information sur la conformation spatiale des molécules. Le choix du format adéquat est réalisé en fonction des capacités de stockage disponibles, et de l'utilisation attendue de la chimiothèque. Cependant, ce choix n'est pas forcément limitant puisqu'il existe des logiciels permettant de convertir les chimiothèques d'un format à un autre (par exemple Openbabel ⁵⁶, Omega ⁵⁷, Corina ⁵⁸, LigPrep ⁵⁹ ...)

2.2.2.1 Les formats de fichiers 2D

Les formats 2D décrivent de manière relativement simple la structure des molécules, mais sans aucune indication sur les coordonnées spatiales des atomes. Parmi les différents formats 2D disponibles, les codes SMILES, InChI et InChIKey sont les plus populaires.

2.2.2.1.1 Les formats de fichiers SMILES

Le format SMILES (Simplified Molecular-Input Line-Entry System) introduit en 1988 ⁶⁰ permet la représentation d'une molécule comme une succession d'atomes et de liaisons. Dans ce système, les atomes sont représentés par leurs symboles atomiques entre crochets, sauf pour les éléments classiques de la chimie organique (B, C, N, O, P, S, F, Cl, Br et I) pour lesquels l'écriture entre crochets est réservée uniquement aux cas où la charge, la masse, un isotope ou la stéréochimie sont précisés. Le symbole atomique est en majuscule (par exemple C pour le carbone) lorsque l'atome appartient à un groupement aliphatique ou en minuscule lorsqu'il fait partie d'un groupement aromatique (par exemple un phényle est représenté par le code SMILES : c1ccccc1). Les liaisons simples, doubles, triples et aromatiques sont codées respectivement par les symboles suivants « - », « = », « # » et « : ». Les liaisons simples et aromatiques peuvent être omises pour simplifier le code, des atomes adjacents sont donc liés par une liaison simple ou aromatique, la distinction étant instinctive selon la casse des symboles atomiques (par exemple, CCCC et cccc représentent respectivement le butane et le 1,3-butadiène). Les ramifications, impossible à représenter telles quelles puisque le code SMILES est un enchaînement linéaire d'atomes, sont spécifiées entre parenthèses (par exemple, CC(C)C pour l'isobutane). De même, les cycles aromatiques sont construits en cassant une liaison du cycle et l'enchaînement des atomes du cycle est indiqué classiquement

mais avec un chiffre suivant le symbole atomique de chaque atome impliqué dans la liaison rompue (Figure 7).

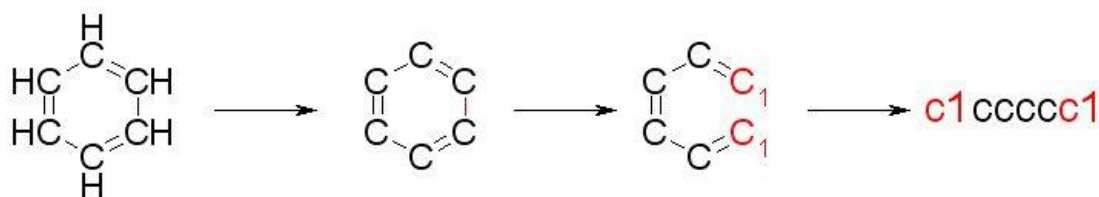


Figure 7. Exemple de construction d'un code SMILES d'un cycle aromatique avec le benzène.

Lorsque des informations sur un isotope, la configuration des doubles liaisons et la chiralité sont décrites, le code SMILES utilisé est alors dit « isomérique ». Les informations isotopiques sont renseignées en inscrivant entre crochets le symbole atomique précédé du nombre représentant sa masse atomique (par exemple [14C] pour le carbone 14). La configuration des double liaisons est précisée à l'aide des symboles « / » et « \ » qui sont considérés comme des liaisons directionnelles. Ainsi, la configuration E est codée par une combinaison de 2 symboles parallèles « /C=C/ » ou « \C=C\ » et une configuration Z par une combinaison de 2 symboles anti-parallèles « /C=C\ » ou « \C=C/ » (Figure 8).

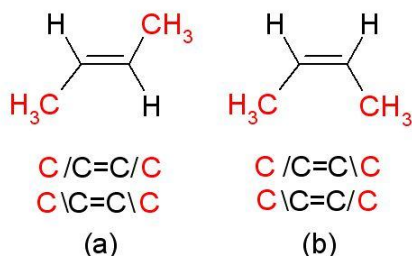


Figure 8. Représentation par les codes SMILES de la configuration E (a) et Z (b) du but-2-ène.

Un atome présentant une chiralité est représenté entre crochet par son symbole atomique suivi du symbole « @ », d'un code de deux lettres indiquant la classe chirale (TH pour tétraédrique, AL pour les allènes, SP pour square-planar ou plan carré, TB pour trigonale-bipyramidale et OH pour octaédrale) et d'un désignateur numérique de permutation chirale. Pour un carbone tétraédrique, la chiralité est donc notée [C@TH1] simplifié en [C@] et [C@TH2] simplifié en [C@@]. Les désignateurs de permutation chirale 1 et 2, ou « @ » et « @@ », sont utilisés lorsque les substituants sont disposés dans un sens anti-horaire ou horaire autour du centre tétraédrique respectivement.

Le principal problème des codes SMILES isomériques est que chaque code SMILES représente bien une seule et unique molécule mais qu'une même molécule peut être codée par différentes formules SMILES, en fonction du choix (arbitraire) du premier atome à coder et du sens de lecture. Ceci pose un problème majeur lors d'une recherche spécifique d'une molécule par comparaison de codes SMILES et un risque de redondances dans les chimiothèques. David Weininger et ses collaborateurs ne tardent pas à proposer une solution, puisque un an à peine après avoir présenté le code SMILES, la canonisation des codes SMILES est publiée ⁶¹. La méthode proposée pour obtenir un code SMILES unique pour chaque molécule, ou SMILES canonique, est appelée CANGENE et consiste en l'utilisation successive de deux algorithmes CANON et GENES. L'algorithme CANON se base sur six critères (le nombre de connections, le nombre de liaisons non hydrogène, le nombre atomique, le signe de la charge, la charge absolue et le nombre d'hydrogènes attachés) pour attribuer à chaque atome un nombre canonique. L'algorithme GENES permet ensuite la génération de SMILES uniques, en débutant par l'atome avec le nombre canonique le plus petit, et lorsqu'il y a plusieurs choix possibles, en progressant toujours vers le nombre canonique le plus faible. Cependant, malgré cela, il n'y a jamais eu de standardisation formelle du format SMILES et différentes implémentations ont été réalisées pouvant conduire à des codes SMILES qui ne sont plus uniques selon le logiciel utilisé ⁶². Un autre système de notation a donc été proposé, les codes InChI (IUPAC International Chemical Identifier) ⁶³

2.2.2.1.2 Les formats de fichiers InChI et InChIKey

Les composés chimiques ont des identifiants chimiques standardisés et internationaux, définis par la nomenclature IUPAC (International Union of Pure and Applied Chemistry). Le format de fichier InChI a été pensé et développé comme l'équivalent informatique de ce système de notation. Les codes InChI utilisent une succession de champs d'information pour décrire la structure chimique, chaque nouveau champ permettant d'ajouter de nouveaux détails ⁶⁴. Chaque code InChI correspond à une seule et même molécule et une molécule aura un seul code InChI. En effet, même si les champs générés dépendent du niveau des détails structuraux disponibles, l'un des avantages de l'InChI est que, pour deux structures avec des niveaux de détails différents (par exemple, une structure dessinée avec des informations stéréochimiques ou de chiralité et pas l'autre), le code InChI de la structure avec le moins de détails sera un sous-ensemble du code de la seconde ⁶⁴ (Figure 9).

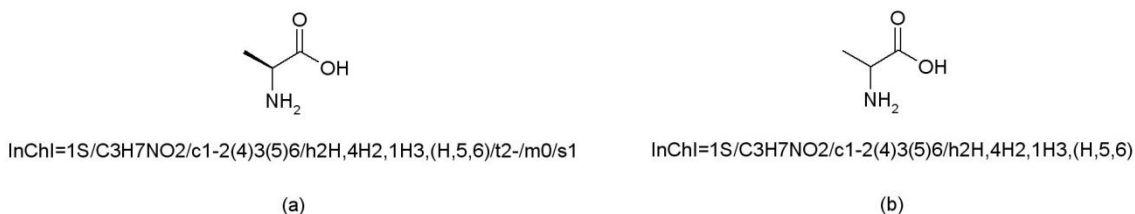
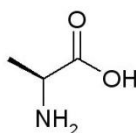


Figure 9. Exemple de deux codes InChI pour une molécule, l'alanine, avec (a) ou sans (b) détail de stéréochimie.

Les différents champs d'un code InChI concernent la formule brute, la connectivité, les isotopes, la stéréochimie et les tautomères. Ces champs sont séparés par le symbole « / », suivi d'une lettre en minuscule (excepté pour le champ de la formule brute). Ainsi, dans la Figure 9, les différents champs inclus dans les deux codes InChI (a) et (b) sont, dans l'ordre, la formule brute et la connectivité (c pour la connectivité 1-1 en excluant les hydrogènes terminaux et h pour la connectivité 1-2 incluant les hydrogènes terminaux). Le code (a) possède aussi des champs d'information sur la chiralité (t pour la parité tétraédrique) et la stéréochimie (m pour la parité inversée [m1] ou non [m0] déterminant la stéréochimie relative et s pour la stéréochimie : absolue [s1], relative [s2] ou racémique [s3]). Avant le premier champ pour les deux codes InChI, la mention 1S précise qu'il s'agit d'un code standard réalisé avec la première version de l'algorithme.

Les codes sont générés à l'aide d'un algorithme qui procède en trois étapes : la normalisation (permettant d'éviter les informations redondantes), la canonisation (pour s'assurer d'obtenir un code unique) et la sérialisation (permettant le codage de l'information sous la forme d'une suite d'informations plus petites) ⁶³.

Le code InChI étant constitué d'une succession de champs, il peut être très long voire trop pour certains outils de recherche qui vont lui faire subir des cassures imprévisibles et indésirables, rendant la recherche impossible ⁶⁴. Pour pallier à ce problème, une version hash code de 27 caractères, appelée InChIKey, a été développée. Le premier bloc de 14 lettres code pour le squelette moléculaire (l'équivalent de la formule brute et de la connectivité précédente). Le second bloc est formé de 8 lettres représentant la stéréochimie et la composition isotopique, suivies de deux lettres, S indiquant que l'InChIKey a été obtenu à partir d'un InChI standard, et A indiquant que la version 1 a été utilisée. Le dernier bloc est constitué d'une seule lettre traduisant le nombre de proton (la lettre N correspondant au terme neutre). ⁶⁴ Chaque bloc est séparé du précédent par un tiret « - ». (Figure 10)



InChIKey=QNAYBMKLOCPYGJ-REOHCLBHSA-N

Figure 10. Code InChIKey standard de la L-alanine.

2.2.2.2 Les formats de fichiers 3D

Les formats de fichiers 3D décrivent la structure de la molécule en présentant l'avantage de préciser les coordonnées cartésiennes spatiales. Cependant, en contrepartie, la taille des fichiers est beaucoup plus importante et leur stockage et leur manipulation nécessite de plus grandes capacités informatiques. Les formats 3D les plus couramment utilisés sont les formats SDF⁶⁵, PDB⁶⁶ et MOL2⁶⁷.

2.2.2.2.1 Le format SDF

Le format SDF (ou SDfile, Structure-Data file) fait partie d'une grande famille de formats de fichiers appelée CTable ou Chemical Table file. Ce format utilise une table de connexion pour décrire les relations et les propriétés structurales d'un groupe d'atomes, partiellement ou complètement connectés les uns aux autres. Un fichier SD est ainsi divisé en blocs placés les uns en dessous des autres (Figure 11), débutant par une ligne de chiffres, appelée « Counts line », qui apporte des informations notamment sur le nombre d'atomes et de liaisons. A la suite de cette première ligne, deux blocs consacrés respectivement aux atomes (précisant les coordonnées cartésiennes, le symbole atomique, la charge, la stéréochimie et les hydrogènes associés) puis aux liaisons (indiquant les atomes mis en jeu, le type, la stéréochimie et le topologie de la liaison) sont disposés⁶⁵. A ces trois blocs peut venir s'ajouter un dernier bloc consacré aux propriétés générales de la molécule (masse molaire, activité biologique...). Les fichiers aux formats SDF possèdent une extension « .sdf ».

ZINC04658553
-OEChem-02031410013D

13 12 0 1 0 0 0 0 0999 V2000	Count line
-0.0187 1.5258 0.0104 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0.0021 -0.0041 0.0020 C 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0.5123 -0.3556 -0.8948 H 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0.7309 -0.5033 1.2229 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0.1165 -0.7205 2.2534 O 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1.9350 -0.6899 1.1802 O 0 5 0 0 0 0 0 0 0 0 0 0 0 0 0 0 -1.3761 -0.5125 0.0124 N 0 3 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1.0039 1.9031 0.0027 H 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 -0.5459 1.8868 -0.8726 H 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 -0.5289 1.8773 0.9072 H 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 -1.8641 -0.1783 -0.8050 H 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 -1.3624 -1.5214 0.0069 H 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 -1.8484 -0.1871 0.8426 H 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	atomes
1 2 1 0 0 0 0 1 8 1 0 0 0 0 1 9 1 0 0 0 0 1 10 1 0 0 0 0 2 3 1 0 0 0 0 2 4 1 0 0 0 0 2 7 1 0 0 0 0 4 5 2 0 0 0 0 4 6 1 0 0 0 0 7 11 1 0 0 0 0 7 12 1 0 0 0 0 7 13 1 0 0 0 0	liaisons
M CHG 2 6 -1 7 1 M END	propriétés
\$\$\$\$	

Figure 11. Fichier SDF de la L-alanine téléchargé dans la base de données ZINC ⁴³, illustrant la disposition en bloc des informations structurales.

2.2.2.2.2 Le format PDB

La PDB (Protein Data Bank) ⁶⁸ est une base de données créée en 1971 pour archiver les structures cristallographiques de macromolécules. Alors qu'au commencement cette base de données ne contenait que sept structures, aujourd'hui 97362 structures sont proposées sur le site internet ⁶⁹. Pour faire face à cet afflux immense de données, un nouveau format de fichier, capable de stocker de manière standardisée et optimisée l'ensemble des informations structurales à disposition, a été développé. Il s'agit du format PDB ⁶⁶, nommé ainsi en référence à la banque dont il est issu et dont l'extension est « .pdb ». Dans ce format de fichiers, les coordonnées atomiques sont stockées sous forme de ligne de 80 caractères dans une section appelée ATOM (pour les 20 résidus classiques des protéines) ou HETATM (pour tous les autres atomes : ligands, molécules d'eau...). Chaque ligne est consacrée à un atome,

les six premières colonnes étant réservées aux identifiants (type atomique, résidu auquel appartient l'atome et identifiant de séquence), les colonnes 7 à 70 aux données (coordonnées spatiales exprimées en Å, et les facteurs d'occupation et de température) tandis que les colonnes 71 à 80 peuvent contenir des informations sur la séquence (symbole et charge de l'élément). Les atomes sont présentés groupés par résidus, en commençant par les atomes du squelette (N-C α -C-O) puis les atomes de la chaîne latérale à partir du C α . Les informations de connectivité (atomes mis en jeu dans la liaison et type de liaisons) des atomes de la catégorie HETATM sont quant à elles renseignées dans une section CONECT. Il est à noter que les informations de connectivité des atomes de la catégorie ATOM sont implicites. Mises à part les coordonnées atomiques, d'autres informations (par exemple, le nom de la protéine, des références bibliographiques, les conditions de cristallisation, la séquence en acides aminés dans la section SEQRES, les informations sur la structure secondaire dans les sections HELIX et TURN etc...) sont aussi contenues dans un fichier PDB, triées dans différentes sections (Figure 12).

HEADER	UNK	14-02-04	1UNK						
REMARK	1 corina 3.48 0000 08.02.2010								
HETATM	1	C1 <0>	1	-0.019	1.526	0.010	0.00	0.00	C
HETATM	2	N1 <0>	1	-1.376	-0.512	0.012	0.00	0.00	N
HETATM	3	O1 <0>	1	0.116	-0.720	2.253	0.00	0.00	O
HETATM	4	C2 <0>	1	0.002	-0.004	0.002	0.00	0.00	C
HETATM	5	O2 <0>	1	1.935	-0.690	1.180	0.00	0.00	O
HETATM	6	C3 <0>	1	0.731	-0.503	1.223	0.00	0.00	C
HETATM	7	H1 <0>	1	0.512	-0.356	-0.895	0.00	0.00	H
HETATM	8	H2 <0>	1	1.004	1.903	0.003	0.00	0.00	H
HETATM	9	H3 <0>	1	-0.546	1.887	-0.873	0.00	0.00	H
HETATM	10	H4 <0>	1	-0.529	1.877	0.907	0.00	0.00	H
HETATM	11	H5 <0>	1	-1.864	-0.178	-0.805	0.00	0.00	H
HETATM	12	H6 <0>	1	-1.362	-1.521	0.007	0.00	0.00	H
HETATM	13	H7 <0>	1	-1.848	-0.187	0.843	0.00	0.00	H
CONECT	1	4	8	9	10				
CONECT	2	4	11	12	13				
CONECT	3	6							
CONECT	3	6							
CONECT	4	1	2	6	7				
CONECT	5	6							
CONECT	6	3							
CONECT	6	3							
CONECT	6	4	5						
CONECT	7	4							
CONECT	8	1							
CONECT	9	1							
CONECT	10	1							
CONECT	11	2							
CONECT	12	2							
CONECT	13	2							

atomes

liaisons

END

Figure 12. Fichier PDB de la L-alanine généré avec la version en ligne de Corina⁷⁰. Dans la section *HEADER*, l'identifiant PDB (non présent ici puisqu'il s'agit d'un ligand et pas d'une protéine), la date de publication dans la banque (ici, date de génération par CORINA) et la classification de la molécule sont indiqués (ici, unknown). Dans la section *REMARK*, les informations usuellement proposées sont le nom de la molécule, l'espèce dont la molécule est extraite, les auteurs, des références bibliographiques, et d'autres informations générales sur la protéine.

2.2.2.2.3 Le format MOL2

Le format MOL2⁶⁷ a initialement été développé pour le logiciel SYBYL par son éditeur Tripos. Il s'agit d'un fichier ASCII (American Standard Code for Information Interchange), d'extension « .mol2 », et réunissant toutes les informations nécessaires pour reconstruire une molécule. Les informations sont présentées dans différentes sections, chacune débutant par un RTI (Record Type Indicator), qui est une ligne de caractère ASCII commençant toujours par le symbole « @ » et permettant d'explicitier le type de données de la section. Pour une molécule, la première ligne sera toujours le RTI « @<TRIPOS>MOLECULE »,

éventuellement suivie de lignes de commentaires (nom et informations sur la molécule). Parmi les sections les plus courantes figurent les sections @<TRIPOS>ATOM, @TRIPOS<BOND> et @TRIPOS<SUBSTRUCTURE>. Dans la section @<TRIPOS>ATOM, chaque ligne est consacrée à un atome. Les différentes informations fournies pour chaque atome sont son nom, ses coordonnées cartésiennes, son type atomique, l'identifiant de la sous-structure à laquelle il appartient, sa charge et éventuellement son statut interne défini par SYBYL (DSPMOD, TYPECOL, CAP, BACKBONE, DICT, ESSENTIAL, WATER, DIRECT). De même, dans la section @<TRIPOS>BOND, chaque ligne définit une seule liaison. Chaque liaison est définie par un identifiant numérique, les identifiants numériques des atomes formant la liaison, le type de liaison (simple [1], double [2], triple [3], amide [am], aromatique [ar], pseudo [du], inconnue [un] ou non connecté [nc]) et un statut interne défini par SYBYL (TYPECOL, GROUP, CAP, BACKBONE, DICT, INTERRES). Chaque ligne de la section @<TRIPOS>SUBSTRUCTURE donne des informations sur une sous-structure: son identifiant, l'atome racine de la sous-structure, le type de sous-structure (temporaire [temp] ou permanente [perm], résidu [residue], groupe [group] ou domaine [domain]), le dictionnaire associé à la sous-structure, la chaîne correspondante, le nombre de liaisons inter sous-structures et le statut interne (LEAF, ROOT, TYPECOL, DICT, BACKWARD et BLOCK) (Figure 13).

2.2.2.2.4 Comparaison des formats

Le format MOL2 permet d'assigner à chaque atome une charge partielle. Le format PDB, pour sa part, possède un champ propre pour les facteurs de température (ou facteur B) dérivés de la précision de la densité électronique. Enfin, le format SDF contient des lignes de 0 permettant d'intégrer des propriétés ou des caractéristiques du produit tels que le numéro de référence dans un catalogue, la quantité disponible etc...

<pre>@<TRIPOS>MOLECULE ZINC04658553 13 12 0 0 0 SMALL USER_CHARGES</pre>							informations générales sur la molécule
<pre>@<TRIPOS>ATOM 1 C1 -0.0187 1.5258 0.0104 C.3 1 <0> -0.1611 2 C2 0.0021 -0.0041 0.0020 C.3 1 <0> -0.0551 3 H1 0.5123 -0.3556 -0.8948 H 1 <0> 0.1319 4 C3 0.7309 -0.5033 1.2229 C.2 1 <0> 0.4852 5 O1 0.1165 -0.7205 2.2534 O.co2 1 <0> -0.6701 6 O2 1.9350 -0.6899 1.1802 O.co2 1 <0> -0.6379 7 N1 -1.3761 -0.5125 0.0124 N.4 1 <0> -0.6135 8 H2 1.0039 1.9031 0.0027 H 1 <0> 0.0992 9 H3 -0.5459 1.8868 -0.8726 H 1 <0> 0.0680 10 H4 -0.5289 1.8773 0.9072 H 1 <0> 0.0758 11 H5 -1.8641 -0.1783 -0.8050 H 1 <0> 0.4150 12 H6 -1.3624 -1.5214 0.0069 H 1 <0> 0.4300 13 H7 -1.8484 -0.1871 0.8426 H 1 <0> 0.4328</pre>							atomes
<pre>@<TRIPOS>BOND 1 1 2 1 2 1 8 1 3 1 9 1 4 1 10 1 5 2 3 1 6 2 4 1 7 2 7 1 8 4 5 2 9 4 6 1 10 7 11 1 11 7 12 1 12 7 13 1</pre>							liaisons

Figure 13. Format MOL2 de la L-alanine téléchargé de la base de données ZINC⁴³

L'obtention des molécules dans le format adéquat fait partie intégrante et constitue une étape critique du processus de préparation d'une chimiothèque. Cependant, des étapes supplémentaires sont nécessaires avant d'obtenir une chimiothèque correctement préparée.

2.2.3 Préparation d'une chimiothèque

Pour garantir la réussite d'un criblage virtuel, une attention très précise doit être apportée aux paramètres et à la qualité des outils utilisés à chaque étape. L'utilisation d'une chimiothèque optimisée et préparée avec soin est déterminante pour cribler uniquement des composés présentant une structure adéquate (notamment par prise en compte des différents états d'ionisation, de mésomérie et de tautomérie et génération de plusieurs conformères) et un bon potentiel de candidat-médicament (déterminé par des filtres ADME-Tox), tout en éliminant avant criblage les composés dont certaines propriétés (physico-chimiques, groupements toxiques...), au contraire, sont rédhibitoires et incompatibles avec un candidat-médicament. Toutes ces étapes de préparation doivent pouvoir être automatisées pour assurer la viabilité du projet puisque une chimiothèque peut comporter des millions de molécules.

2.2.3.1 États d'ionisation, mésomérie et tautomérie

Pour construire une chimiothèque, il est important de prendre en compte toutes les formes ionisées, mésomériques et tautomériques pertinentes d'un composé. En effet, en fonction des conditions physiologiques, certains composés peuvent exister sous différentes formes à cause de phénomènes de tautomérie (transfert intramoléculaire de protons), de mésomérie (délocalisation électronique dans une molécule conjuguée) ou d'ionisation (Figure 14). En fonction de la forme envisagée, les groupements fonctionnels seront placés de manière permettant ou non l'établissement d'interactions avec le site de liaison.

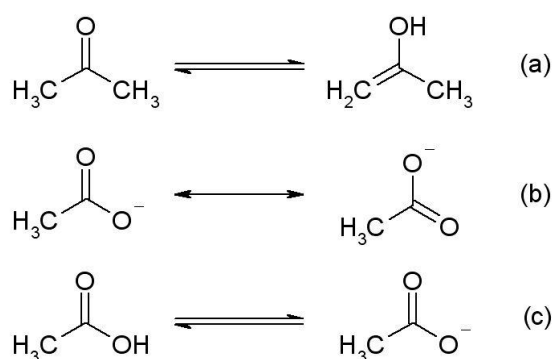


Figure 14. Exemple de deux formes tautomères (a), mésomère (b) ou états d'ionisation différents (c).

Rationnellement, le composé devrait être inclus dans la chimiothèque dans sa forme la plus probable, mais en pratique, choisir entre les différents états d'un composé peut s'avérer très difficile et très lourd en calculs. Pour certains groupements ce choix est aisé et un seul état est généralement présent dans la chimiothèque. C'est par exemple le cas des acides carboxyliques présents dans les conditions physiologiques sous forme anionique uniquement. Pour les autres, tous les états d'ionisation, de mésomérie et de tautomérie pertinents doivent être générés (à l'aide de logiciels tels que LigPrep⁵⁹ ou QuacPac⁷¹) et inclus dans la chimiothèque⁷².

2.2.3.2 Génération des conformations 3D

A l'instar de la prise en compte des différents états d'ionisation, de mésomérie et de tautomérie, il est essentiel de prendre en compte la flexibilité des molécules. En effet, les ligands sont généralement générés dans leur conformation de plus basse énergie. Cependant, la conformation bioactive d'une molécule ne correspond que rarement à la conformation de plus basse énergie puisque même si la liaison au récepteur se fait par un mécanisme de

sélection de conformation, un ajustement du récepteur et du ligand l'un à l'autre se produit également ⁷³ (Figure 15), ce qui entraîne une augmentation de l'énergie du ligand ⁷⁴.

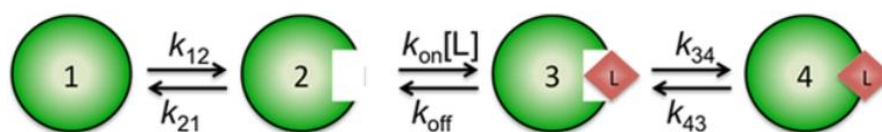


Figure 15. Mécanisme de liaison d'un ligand à son récepteur par sélection de la meilleure conformation du récepteur par le ligand (passage de l'état 1 à 2) suivi d'un ajustement de la protéine et du ligand l'un à l'autre selon un mécanisme induced fit (passage de l'état 3 à l'état 4) ⁷³.

Pour prendre en compte cette flexibilité, deux options sont valables. La première est de générer un ensemble de conformations avant le criblage ⁷⁵. La seconde consiste à utiliser un logiciel qui prendra en compte cette flexibilité au cours du criblage, que ce soit lors de criblages basés sur les pharmacophores 3D (voir paragraphe 3.2.2) ou sur la structure (voir paragraphe 4).

2.2.3.3 Filtres ADME-Tox

Dans les années 1990, l'échec des candidats médicaments lors des phases cliniques était principalement imputable à de mauvaises performances pharmaco-cinétiques et de biodisponibilité (Figure 16).

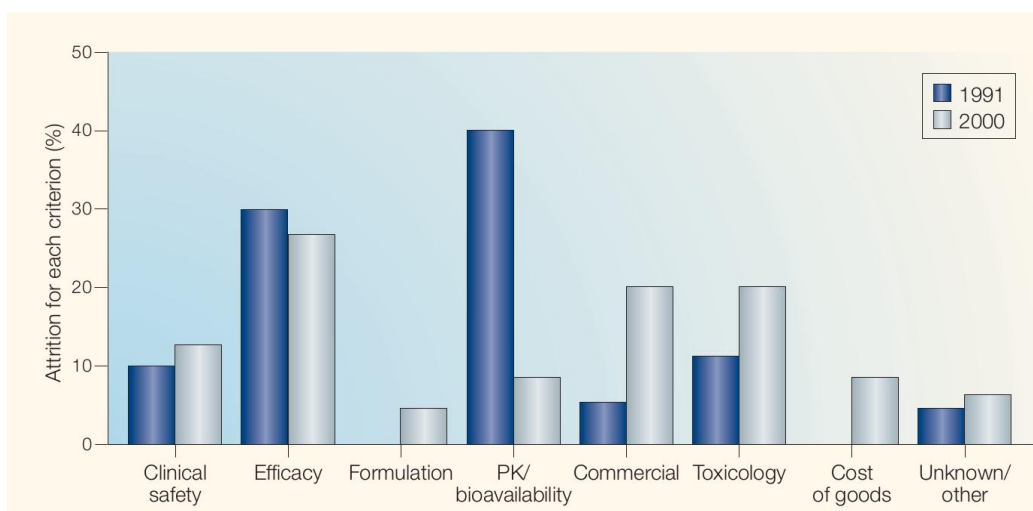


Figure 16. Evolution des causes d'attrition des candidats médicaments lors des phases cliniques entre 1991 et 2000 ³¹

Il est alors apparu nécessaire de prendre en compte ces critères dès les phases précoces de R&D pour tenter de diminuer le taux d'échec des candidats-médicaments. La mise en place de filtres ADME (Absorption Distribution Métabolisme et Elimination)-Tox (Toxicité) des chimiothèques avant tout processus de criblage a donné d'excellent résultats. En effet, même si des tentatives d'amélioration de la prédiction des caractères ADME et de la toxicité sont encore à l'étude, le taux d'échec attribuable à une mauvaise pharmacocinétique ou biodisponibilité a chuté de 40 % à moins de 10 % en 10 ans. Actuellement, le manque d'efficacité et la toxicologie sont les principales causes d'arrêt de développement de candidats-médicaments³¹.

2.2.3.3.1 Règles de Lipinski

Les chimiothèques étant utilisées dans les processus de R&D de nouveaux médicaments, il semble raisonnable de ne tester que les molécules possédant des caractères susceptibles de faire d'elles des médicaments (autrement dénommé « drug-likeness »). Cependant, l'évaluation de ces caractéristiques n'est pas aisée, et ainsi au fil des ans, différents critères de sélection, la plupart du temps basés sur des propriétés physico-chimiques, ont été proposés. Ainsi, en 1997, Lipinski et ses collègues de Pfizer⁴⁶ proposent ce qu'on appelle couramment la « règle de Lipinski » ou la « règle de 5 » (Tableau 5). La définition de cette règle repose sur la recherche de propriétés communes à 2245 molécules extraites du WDI (World Drug Index) ayant au moins atteint la phase II des essais cliniques et donc présentant *a priori* de bonnes valeurs de solubilité et de perméabilité intestinale. La « règle de Lipinski » propose ainsi des valeurs seuils pour le poids moléculaire (PM), le nombre d'accepteurs et de donneurs de liaisons hydrogènes (ALH et DLH) et le coefficient de partition eau/octanol (logP). Selon cette règle, un composé est plus susceptible de posséder une mauvaise biodisponibilité orale lorsque'il présente plus d'une violation de cette règle.

Propriétés physico-chimiques	« Drug-likeness »		« Lead-likeness »	
	Règle de Lipinski ⁴⁶	Veber et al. ⁷⁶	Hann et Oprea ³⁹	Règle des 3 ⁴⁷
Poids moléculaire (PM) en Daltons	≤ 500		≤ 460	< 300
Lipophilie (logP)	≤ 5		-4/4.2	≤ 3
Nombre de donneurs de liaisons hydrogène (DLH)	≤ 5	DLH + ALH ≤ 12 *	≤ 5	≤ 3
Nombre d'accepteurs de liaisons hydrogène (ALH)	≤ 10		≤ 9	≤ 3
Aire de la surface polaire (ASP)		≤ 140 Å ² *	≤ 170 Å ²	(≤ 60)
Nombre de liaisons rotatives (nrot)		≤ 10	≤ 10	(≤ 3)
Nombre de cycles aromatiques			≤ 4	
Perméabilité membranaire (Caco-2)			≥ 100	
Solubilité dans l'eau (logS)			-5/0.5	

Tableau 5. Valeurs seuils des différentes propriétés physico-chimiques définissant le caractère « drug-like » ou « lead-like » d'un composé. Pour la règle de Veber et al., le symbole « * » signifie que l'un ou l'autre de ces critères doit être respecté en plus du nombre de liaisons rotatives

Depuis, des tentatives pour affiner cette prédiction ont été menées, notamment par étude de données de pharmaco-cinétiques chez le rat ^{77, 76}, ou par assouplissement de ces règles en appliquant un poids différent à chacune des propriétés physico-chimiques étudiées ⁷⁸. Ainsi, Veber et ses collègues ⁷⁶ proposent une simplification des règles de « Lipinski » après analyse des données pharmaco-cinétiques chez le rat pour 1100 candidats médicaments issus de la base de données de biodisponibilité orale SmithKline Beecham (Tableau 5). Dans cette nouvelle méthode de prédiction de la biodisponibilité orale, seul le nombre de liaisons rotatives et soit l'aire de la surface polaire ou soit l'addition du nombre de donneurs et d'accepteurs de liaisons hydrogènes sont pris en compte. Une autre approche, suggérée par Hann et Oprea ³⁹, s'intéresse à définir le caractère « lead-like » des composés (Tableau 5). En effet, lors d'un criblage virtuel, ce ne sont pas des médicaments qui sont recherchés mais des hits et des leads qui seront ensuite optimisés pour obtenir des candidats-médicaments. Si ces leads se conforment aux règles de Lipinski, les optimisations ultérieures risquent de donner des composés qui eux ne seront plus « drug-like ». Pour proposer de nouveaux critères

permettant d'évaluer le caractère « lead-like » des composés, Hann et Oprea ont étudié les différences existantes entre les leads et les médicaments par comparaison de la distribution des propriétés de 176 leads et de 532 médicaments ⁷⁹ (Figure 17). Les critères proposés précédemment pour définir le caractère « drug-like » sont conservés mais avec des valeurs seuils ajustées, et de nouveaux sont à prendre en compte tels que le nombre de cycles aromatiques, la perméabilité intestinale mesurée sur des cellules Caco-2, et la solubilité dans l'eau (logS). De même, une règle de 3 (ou « rule of three ») a été définie ⁴⁷ pour construire des chimiothèques de fragments présentant un caractère « lead-like » (voir paragraphe 2.2.1.1.2). Les valeurs seuils de poids moléculaire, de lipophilie, et du nombre d'accepteurs et de donneurs de liaisons hydrogènes (Tableau 5) ont été déduites de l'analyse des propriétés physico-chimiques de fragments ayant été identifiés comme hits lors de criblages virtuels. Le nombre de liaisons rotatives et l'aire de la surface polaire ont aussi été mis en évidence comme des critères intéressants de sélection de fragments « lead-like ».

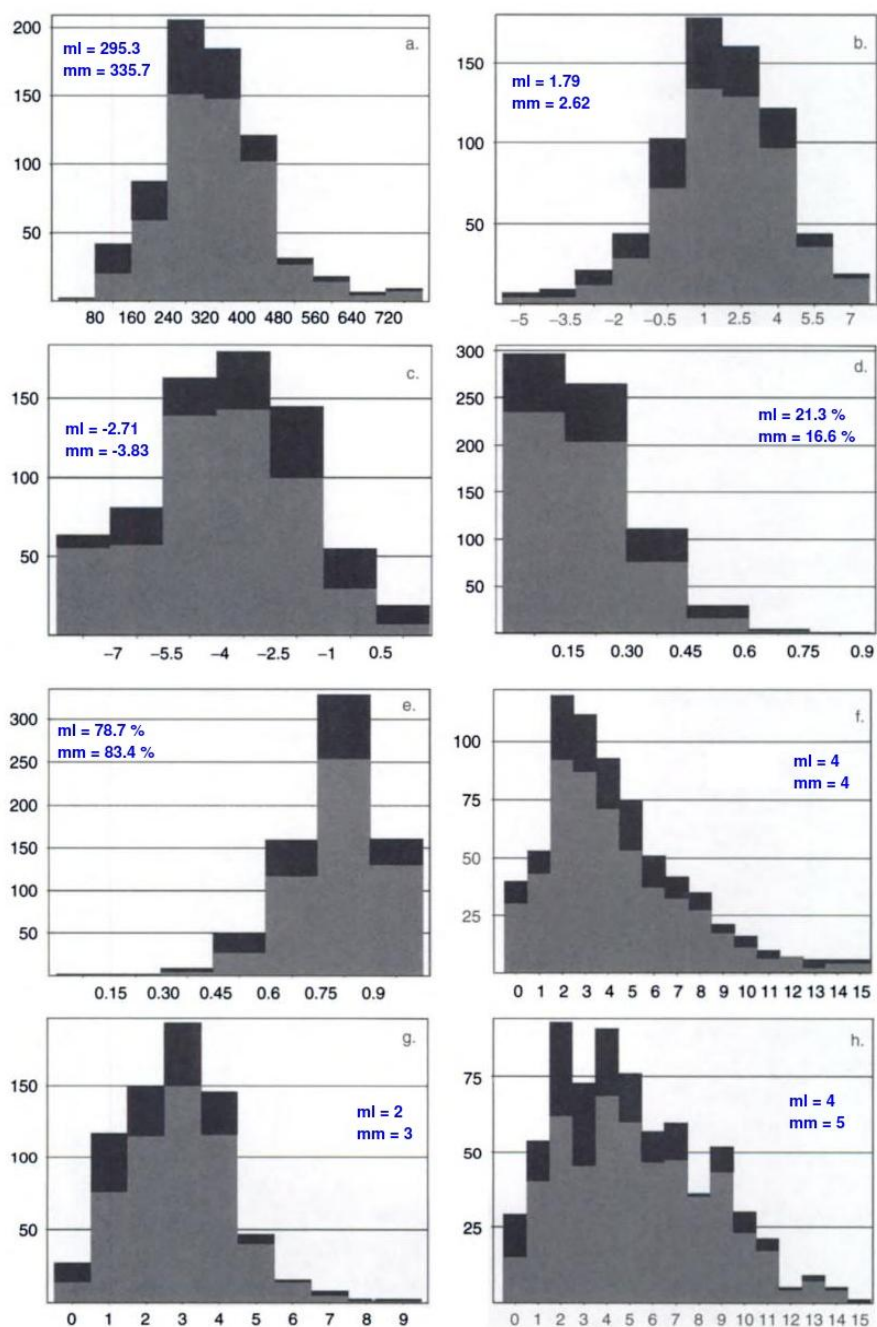


Figure 17. Comparaison de la distribution de huit propriétés entre les leads (en noir) et les médicaments (en gris) : le poids moléculaire (a), le logP calculé (b), la solubilité logS (c), la fraction d'aire de surface polaire (d), et non polaire (e), la somme du nombre de donneurs et d'accepteurs de liaison (f), le nombre de cycles aromatiques (g) et le nombre de liaison rotatives (h). L'axe vertical représente le nombre de composés. Les valeurs ml et mm indiquent la valeur médiane de chaque propriété pour les leads et les médicaments respectivement (d'après ⁷⁹).

2.2.3.3.2 Toxicologie

La toxicité d'un candidat-médicament est l'un des paramètres les plus redoutés par les développeurs, surtout lorsqu'il entre en phases cliniques. En effet, un candidat-médicament présentant des effets secondaires graves lors des phases cliniques est immédiatement abandonné, et ce généralement de manière définitive. Il est donc très important de tenter de prédire la toxicité d'un composé à partir de sa structure lors des phases précoces de développement R&D. La toxicité peut être classifiée selon l'organe affecté (hépatotoxicité, néphrotoxicité, cardiotoxicité, ...) ou selon le mécanisme de toxicité (généotoxicité, mutagénicité, carcinogénicité, ...). De nombreuses méthodes *in silico* ont donc été mises en place pour tenter de prédire la toxicité, les effets secondaires et le métabolisme des composés⁸⁰, c'est ce qu'on appelle la toxicologie *in silico* ou toxicologie computationnelle : construction de bases de données toxicologiques, établissement de modèles QSAR (Quantitative Structure-Activity Relationship ou relation quantitative structure-activité), et méthodes basées sur la structure⁸¹.

Bases de données toxicologiques

Les bases de données toxicologiques regroupent des informations obtenues d'études de toxicité provenant de diverses sources (recherche académique, industries pharmaceutique, chimique, cosmétique, alimentaire, ...) pour permettre la recherche et la création de modèle de prédiction de toxicité pour des composés similaires⁸¹. De nombreuses bases de données publiques (Tableau 6) et privées ont donc été construites dans ce but.

Database	Site web	Type de données
CCRIS (Chemical Carcinogenesis Research Information System)	http://toxnet.nlm.nih.gov/cgi-bin/sis/htmlgen?CCRIS	documentation relative à la carcinogénicité pour 8000 composés
FAERS (FDA Adverse Effects Reporting System)	http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/default.htm	documentation de pharmacovigilance relative à la toxicité
IRIS (Integrated Risk Information System)	http://cfpub.epa.gov/ncea/iris/index.cfm	documentation relative à la toxicité de composés chimiques environnementaux
ITER (International Toxicity Estimates for Risk Assessment)	http://www.tera.org/iter/	documentation relative à la carcinogénicité pour 600 composés chimiques environnementaux
JECDB (Japanese Ministry of Health, Labour and Welfare Chemical Toxicity Database)	http://dra4.nihs.go.jp/mhlw_data/jsp/SearchPageENG.jsp	documentation relative à la toxicité pour 369 composés
NPIC (National Pesticide Information Center)	http://npic.orst.edu	documentation relative à la toxicité de pesticides
PAN Pesticide (Pesticide Action Network North America)	http://www.pesticideinfo.org/	documentation relative à la toxicité et l'écotoxicité de 6500 pesticides, insecticides et herbicides
STITCH (Search Tool for Interactions of Chemicals)	http://stitch.embl.de/	documentation relative à la toxicité de 68000 composés
ToxRefDB (Toxicity Reference Database)	http://www.epa.gov/ncct/toxrefdb/	documentation relative à la toxicité
ACToR (Aggregated Computational Toxicology Ressource)	http://actor.epa.gov/actor/faces/ACToRHome.jsp	données numériques expérimentales de toxicité pour 500000 composés chimiques environnementaux
CPDB (University of California, Berkeley, Carcinogenic Potency Database)	http://potency.berkeley.edu/	données numériques expérimentales de carcinogénicité pour 1547 composés
Drugs@FDA	http://www.fda.gov/Drugs/InformationOnDrugs/ucm135821.htm	données numériques expérimentales de toxicité
DSSTox (Distributed Structure-Searchable Toxicity)	http://www.epa.gov/ncct/dsstox/index.html	données numériques expérimentales de toxicité
EXTOXNET (Extension TOXicology NETwork)	http://extoxnet.orst.edu/ghindex.html	données numériques expérimentales de toxicité de pesticides
Gene-Tox	http://toxnet.nlm.nih.gov/cgi-bin/sis/htmlgen?GENETOX	données numériques expérimentales de toxicité pour 3000 composés
LAZAR (LAZY structure-activity Relationships database)	http://www.in-silico.de/	données numériques prédites sur la néphrotoxicité, mutagénicité et carcinogénicité
Riskline, Kemi	http://apps.kemi.se/riskline/	données numériques prédites sur la toxicité
TEXTRATOX	http://www.vet.utk.edu/TETRATOX/index.php	données numériques prédites sur la toxicité aquatique de 2400 composés

Tableau 6. Présentation de quelques bases de données toxicologiques publiques en fonction de leur contenu (d'après ⁸²)

Prédiction de la toxicité à l'aide de modèles QSAR

L'utilisation de modèles QSAR (voir 3.3) est l'une des méthodes les plus populaires pour prédire la toxicité d'un composé. L'utilisation de modèles QSAR prédictif fait même partie des exigences de l'autorité de régulation européenne des composés chimiques (REACH : Registration, Evaluation, Authorisation and Restriction of Chemicals) dans l'Union Européenne ⁸³. Les modèles QSAR utilisés sont des équations mathématiques permettant d'estimer la toxicité de nouveaux composés en utilisant les valeurs de toxicité obtenues précédemment avec d'autres composés ⁸¹. Très souvent, les modèles QSAR sont construits à partir des données extraites des bases de données de toxicologie précédemment présentées. Les études QSAR de toxicologie sont particulièrement populaires pour étudier la carcinogénicité et la génotoxicité, comme en témoigne les 78 modèles QSAR non commerciaux correspondant recensés en 2008 par Benigni et Bossa ⁸⁴. Les modèles QSAR sont aussi très utiles pour la prédiction de toxicité affectant un organe précis, telle que l'hépatotoxicité et la néphrotoxicité ⁸⁵. Pour permettre d'obtenir ces modèles prédictifs, de nombreux logiciels publics ou commerciaux ont été développés (Tableau 7), les logiciels DRAGON et CODESSA étant les plus cités ⁸².

Logiciel	Site web	Type
CODESSA (Comprehensive Descriptors for Structural and Statistical Analysis)	http://www.codessa-pro.com/	commerciale
DRAGON	http://www.talete.mi.it/products/dragon_description.htm	commerciale
GOLPE (Generating Optimal Linear PLS Estimations)	http://www.miasrl.com/golpe.htm	commerciale
PRECLAV (PRoperty Evaluation by CLAss Variables)	http://www.softpedia.com/get/Science-CAD/PRECLAV.shtml	commerciale
ChemProp (Chemical Properties Estimation Software System)	http://www.ufz.de/index.php?en=10684	public
CORAL (CORrelation And Logic)	http://www.insilico.eu/coral	public
OEBCD ToolBox	http://www.oecd.org/env/ehs/	public
RmSquare	http://aptsoftware.co.in/rmsquare/	public
T.E.S.T. (Toxicity Estimation Software Tool)	http://www.epa.gov/nrmrl/std/qsar/qsar.html	public
Virtual Computational Chemistry Laboratory	http://www.vcclab.org/	public

Tableau 7. Liste de quelques programmes publics et commerciaux permettant d'obtenir des modèles QSAR de toxicologie prédictifs (d'après ⁸²).

Prédiction de la toxicité par des méthodes basées sur la structure

La prédiction de la toxicité par des méthodes basées sur la structure consiste à prédire par docking (voir paragraphe 4.5) la liaison des candidats-médicaments à des cibles connues pour leur implication dans des mécanismes de toxicité. C'est notamment le cas des canaux potassiques hERG (human Ether-a-go-go Related Gene) dont l'inhibition provoque une cardiotoxicité potentiellement mortelle, les torsades de pointes ⁸⁶. Cette toxicité par liaison aux canaux potassiques hERG est notamment à l'origine du retrait du marché de plusieurs médicaments, parmi lesquels le cisapride (PREPULSID®) ou encore la terfénadine (SELDANE®) ⁸⁷. De nombreuses tentatives ont donc été menées pour tenter de prédire le potentiel d'inhibition des canaux potassiques hERG des composés dès les phases précoces de R&D ^{88, 89, 90, 91, 92, 93}. Du et ses collègues ⁹⁴ ont proposé une approche alternative, basée sur la structure des canaux potassiques. En effet, par docking de ligands dans un modèle construit par homologie des canaux potassiques hERG humains (aucune structure expérimentale n'étant disponible), les valeurs de pIC50 prédites sont concordantes avec les valeurs

expérimentales. Ce modèle peut donc être utilisé pour filtrer une chimiothèque avant toute étape de criblage, et ainsi identifier et éliminer de potentiels composés cardiotoxiques.

3 Criblage virtuel « ligand-based »

Lorsqu'au moins un ligand de la cible étudiée est connu, un criblage virtuel basé sur les ligands ou « ligand-based » peut être mis en œuvre. Le principe de base commun à toutes les méthodes basées sur les ligands est que des molécules similaires vont avoir tendance à présenter des profils d'activité similaires⁹⁵. La similarité des molécules peut se mesurer par recherche de propriétés communes, qui seront utilisées comme descripteurs de similarité. En fonction du nombre de ligands de référence pour la cible et du type de descripteurs, différentes méthodes peuvent être employées : la recherche de similarité, le criblage à l'aide de pharmacophore et les méthodes QSAR.

3.1 Recherche de similarité

La recherche de similarité est la méthode à employer lorsque très peu de ligands ont été rapportés pour la cible biologique choisie. En effet, une recherche de similarité peut être menée dès lors qu'un ligand actif est connu⁹⁶. Cette méthode repose donc sur l'utilisation de descripteurs et de métriques de similarité permettant de comparer des molécules à cribler à un ou plusieurs ligands de référence pour prédire leur profil d'activité.

3.1.1 Descripteurs de similarité

Les descripteurs de similarité sont des nombres ou des vecteurs qui représentent des caractéristiques clés de composés, dérivant la plupart du temps de leur structure⁹⁶. Ils peuvent être obtenus de deux façons : soit par une procédure mathématique logique qui transforme l'information chimique d'une molécule, codée par une représentation symbolique, en un nombre utile, soit par une procédure expérimentale, dont le résultat est utilisé comme descripteur⁹⁷. Ils permettent de définir au sein d'une base de données quels ligands sont les plus ressemblants du ou des ligands actifs connus.

Les descripteurs sont souvent classés selon leur dimension⁹⁸ : 1D, 2D et 3D (Figure 18). Les descripteurs 1D tels que le poids moléculaire ou le nombre d'atomes lourds ne sont généralement pas utilisés pour des mesures de la similarité⁹⁶.

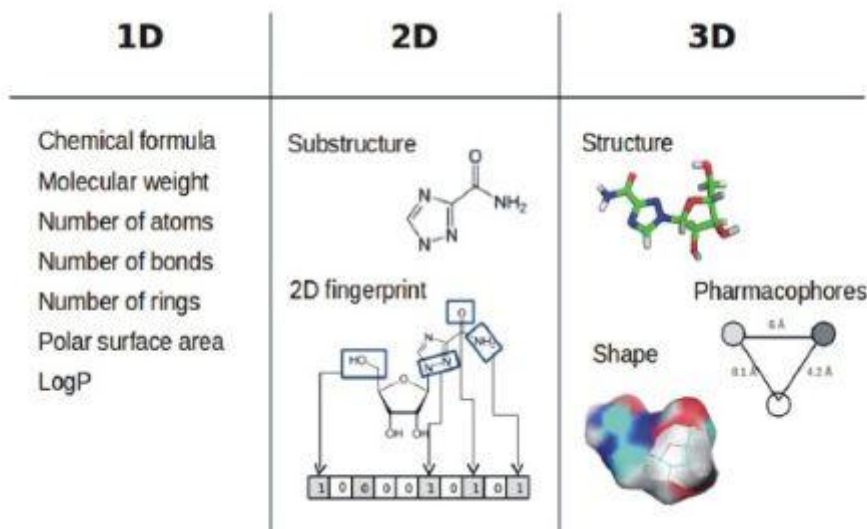


Figure 18. Classification des descripteurs selon leurs dimensions⁹⁹

3.1.1.1 Descripteurs 2D

La recherche de similarité à l'aide de sous-structures consiste à identifier au sein de la chimiothèque les composés dont la structure contient la sous-structure définie à partir du ou des ligands de référence. Pratiquement, les graphes topologiques des molécules à cribler sont analysés pour détecter l'existence d'un isomorphisme avec le graphe du ou des composés actifs de référence.¹⁰⁰ La recherche de la sous-structure commune maximale (MCS ou Maximum Common Substructure) est l'une des toutes premières méthodes de recherche de sous-structure à avoir été utilisée.¹⁰¹

Les empreintes 2D ont été développées pour accélérer la recherche de sous-structures. Dans cette approche, les molécules sont représentées par un vecteur binaire. Chaque bit du vecteur binaire représente un fragment moléculaire, et peut présenter deux valeurs : « 1 » pour indiquer la présence ou « 0 » pour l'absence de ce fragment (Figure 19). La similarité entre deux molécules est donc évaluée par comparaison du nombre de bits communs entre les deux structures.¹⁰² Les empreintes 2D dites « keyed fingerprints », telles que les empreintes MACCS¹⁰³ ou BCI¹⁰⁴, utilisent un dictionnaire de fragments de référence pour déterminer quels sont les fragments à rechercher.

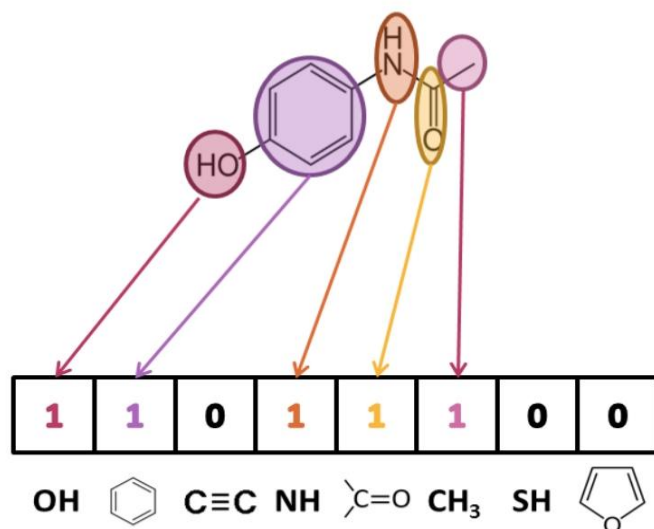


Figure 19. Exemple d'empreinte 2D possible pour le paracétamol

A l'opposé, les fragments des empreintes 2D dites « hashed fingerprints », telles que les empreintes Daylight ¹⁰⁵, sont générés automatiquement à l'aide d'un algorithme qui parcourt toutes les liaisons à partir d'un atome jusqu'au septième rang ¹⁰². Chaque fragment créé par cette procédure est alors représenté dans le vecteur par un bit de valeur 1 (Figure 20). L'avantage majeur de cette méthode par rapport aux « keyed fingerprints » est de proposer une meilleure exploration de la structure, puisque tous les fragments possibles sont générés.

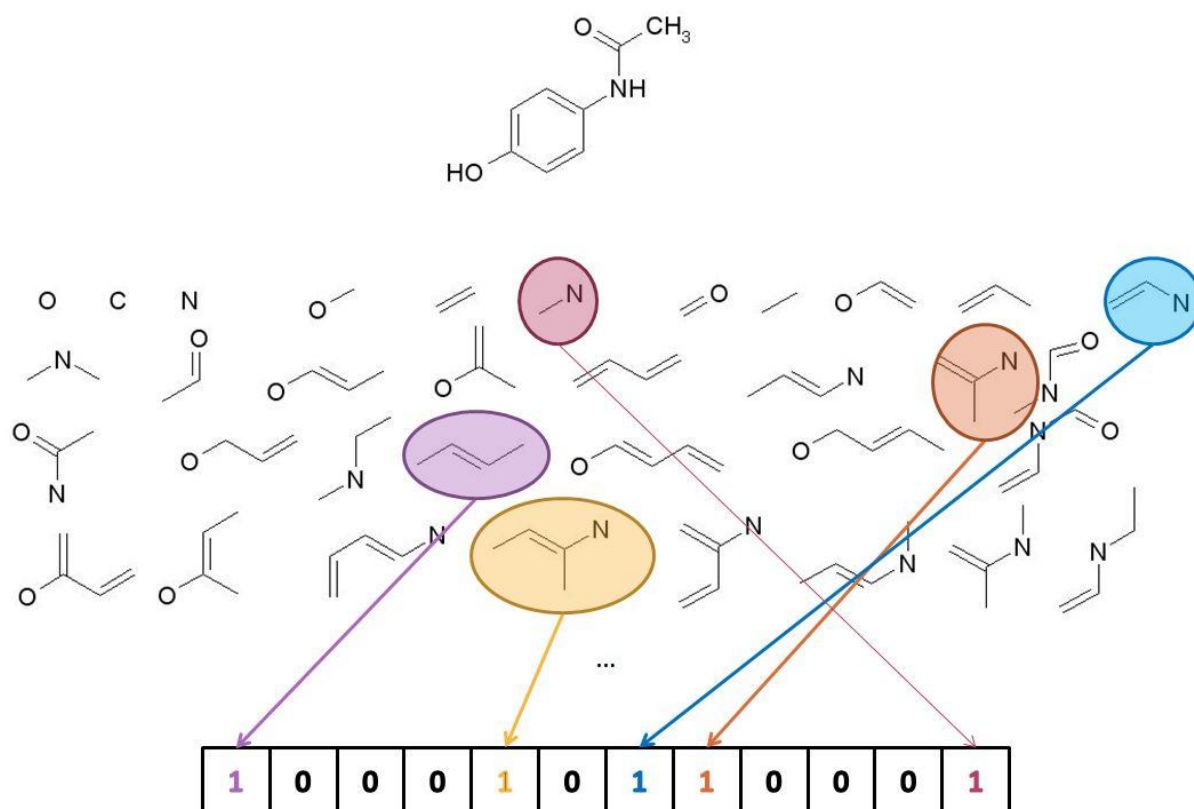


Figure 20. Exemples de quelques fragments (rang de liaison de 4) créés par un algorithme permettant d'obtenir des empreintes de type « hashed fingerprints »

La recherche de similarité en utilisant des empreintes 2D est efficace pour obtenir de nouveaux composés dont la structure est très proche de la ou des molécules de référence. Cependant, les descripteurs 2D ne permettent pas de comparer efficacement des composés très différents structurellement, même si ils ont le même profil de liaison sur une cible donnée¹⁰⁶. Hors, la tendance actuelle est à la recherche, notamment à l'aide de méthodes de « scaffold hopping »¹⁰⁷, de nouveaux composés actifs, par modification du squelette de base d'une molécule de référence. Les avantages d'une telle approche sont nombreux. Tout d'abord, ces méthodes permettent de découvrir de nouvelles séries de composés qui ne seront pas couvertes par un éventuel brevet. Ensuite, elles peuvent permettre d'améliorer des propriétés ADME défavorables par exemple par remplacement d'un squelette lipophile par un plus polaire pour améliorer la solubilité ou encore par substitution d'un squelette métaboliquement instable par un plus stable et moins toxique... Enfin, l'obtention de nouveaux composés avec des squelettes de base différents peut faciliter une synthèse auparavant difficile voire impossible. Au contraire des empreintes 2D qui ne constituent pas de bons descripteurs pour les méthodes de « scaffold hopping », les descripteurs 3D peuvent eux être employés.

3.1.1.2 Descripteurs 3D

Il existe deux grands types de descripteurs 3D : les empreintes 3D ou « 3D fingerprints » et les descripteurs basés sur la surface au sein desquels les descripteurs basés sur la forme sont les plus populaires.

3.1.1.2.1 Les empreintes 3D

Les empreintes 3D (Figure 21) décrivent la présence ou l'absence de caractéristiques géométriques telles que des paires ou des triplets d'atomes ou de points pharmacophores, ou des valeurs d'angles de valence ou de torsion spécifiques.

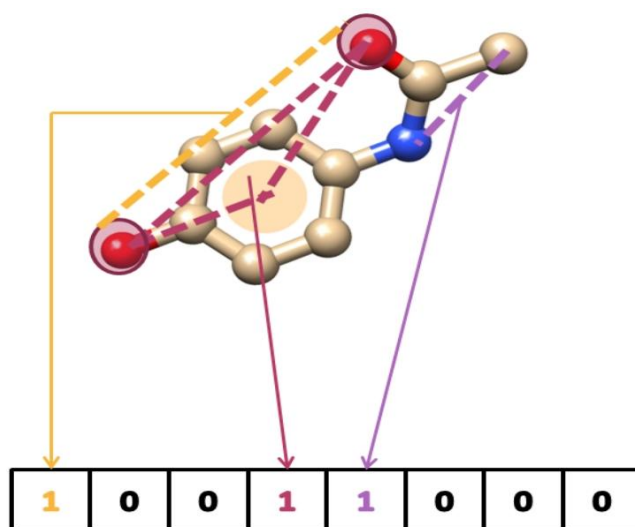


Figure 21. Exemple d'empreinte 3D pour le paracétamol codant une paire de pharmacophores (jaune), un triplet de pharmacophores (grenat) et une paire d'atomes (violet)

3.1.1.2.2 Les descripteurs basés sur la forme

L'idée que des molécules similaires ont tendance à présenter des profils d'activité similaires est ici affinée en précisant que des molécules présentant des formes similaires devraient posséder des propriétés communes^{108,109}. Les méthodes qui utilisent la comparaison de formes peuvent être divisées en 2 catégories : les méthodes de non-superposition et les méthodes de superposition¹¹⁰.

Les méthodes de non-superposition sont indépendantes de l'orientation et de la position des atomes. Certaines méthodes utilisent des triplets d'atomes pour décrire la forme de la molécule^{111, 112, 113}. La méthode des « Signatures de formes » ou « Shape Signatures » proposée par Zauhar et ses collègues¹¹⁴ est une méthode innovante de non-superposition.

Dans cette méthode, la forme est assimilée à la surface moléculaire accessible au solvant. Pour obtenir une représentation détaillée de la surface, l'algorithme SMART (SMooth molecularAR surface Triangulator) qui divise la surface en triangles réguliers est utilisé. Le volume de la surface est ensuite exploré à l'aide d'une méthode de traçage de rayons (« ray-tracing ») par suivi du parcours d'un rayon lumineux à partir du milieu de la surface d'un des triangles et de sa propagation selon les lois de la réflexion optique. Les différents points de réflexion sont ensuite enregistrés et permettent de représenter la forme (Figure 22). A partir de ces traces, des distributions de probabilité sont calculées et utilisées comme descripteurs.

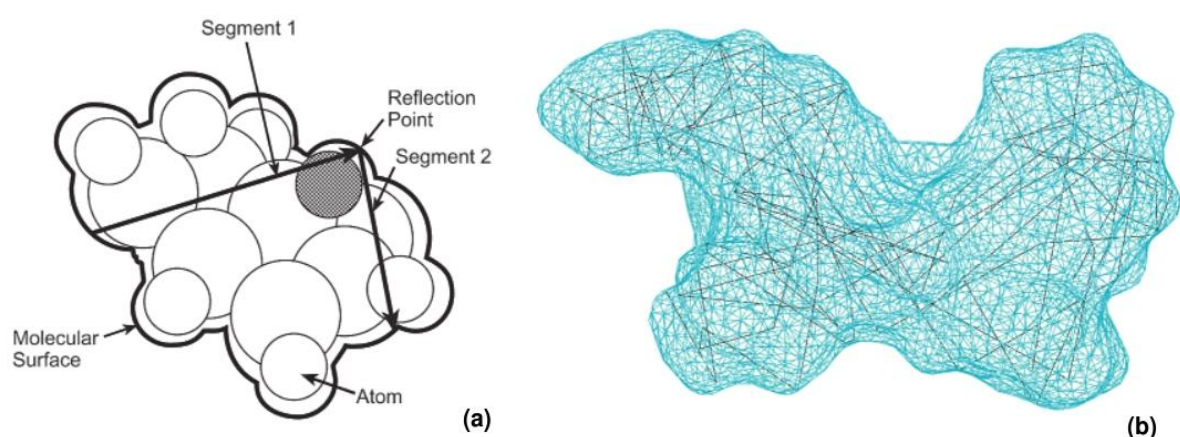


Figure 22. (a) Principe du suivi d'un rayon lumineux et de sa réflexion pour décrire la forme d'une molécule. (b) Traces des rayons pour l'indinavir à basse (100 réflexions) résolution ¹¹⁴

Une autre méthode, l'USR (Ultrafast method for Shape Recognition) ¹¹⁰ repose sur le dogme que la forme d'une molécule est uniquement déterminée par la position relative de ces atomes. Chaque molécule est associée à un vecteur unique de douze descripteurs géométriques représentant les 3 moments statistiques (moyenne, variance et asymétrie) des distances atomiques par rapport à quatre positions de références stratégiques (centroïde (ctd), atome le plus près du centroïde (cst), atome le plus éloigné du centroïde (fst) et atome le plus éloigné de l'atome le plus éloigné du centroïde (fft)) (Figure 23). Une fonction de score normalisée permet ensuite de mesurer la similarité entre des molécules à partir de ces douze descripteurs.

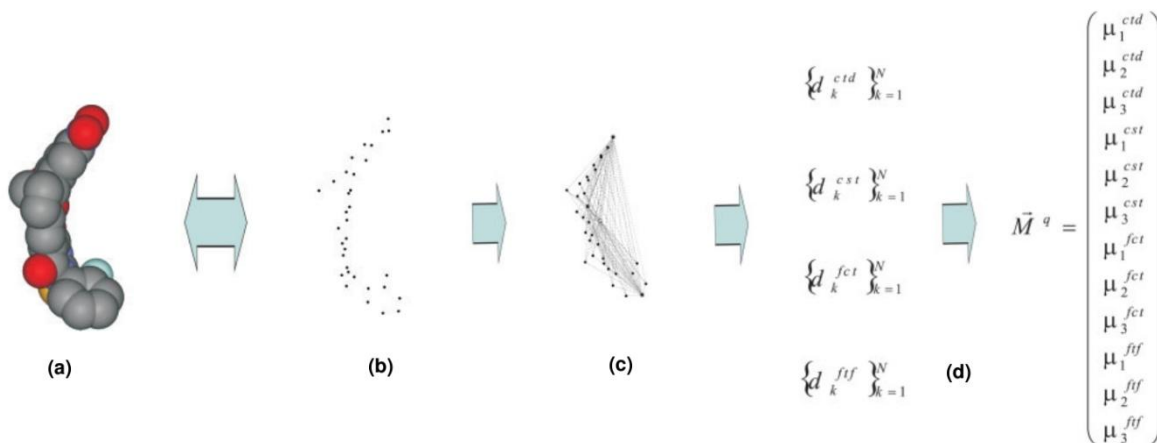


Figure 23. Représentation de (a) la forme à l'aide du modèle CPK (Corey-Pauling-Koltun) (b) la position de tous les atomes. (c) toutes les distances atomiques par rapport aux quatre points de références (d) la définition des douze descripteurs géométriques¹¹⁰

A l'opposé, les méthodes dites de superposition utilisent une superposition optimale des molécules pour pouvoir ensuite comparer leur forme. Certaines de ces méthodes utilisent une grille pour aligner les molécules et évaluer leur forme^{115, 116}. D'autres, telles que proposées par Good et Richards¹¹⁷, emploient une fonction gaussienne comme approximation de la densité électronique des atomes permettant ainsi de décrire la forme de chaque atome. Enfin, certaines méthodes comparent la forme des molécules en fonction du chevauchement de leur volume à l'aide d'une représentation de la molécule sous forme de sphères centrées sur les atomes (Figure 24). Ces sphères peuvent modéliser la surface atomique de van der Waals¹¹⁸ ou une densité gaussienne comme proposé par Grant et Pickup¹¹⁹ et dans une implémentation plus récente par le logiciel ROCS (Rapid Overlay of Chemical Structures)¹²⁰. ROCS, considéré comme la référence des méthodes de similarité 3D¹⁰⁹, permet notamment de préciser des types atomiques (cycles, donneurs ou accepteurs de liaisons hydrogènes...) en représentant des atomes ou des groupes d'atomes comme des gaussiennes de différents « types » ou « couleurs ».

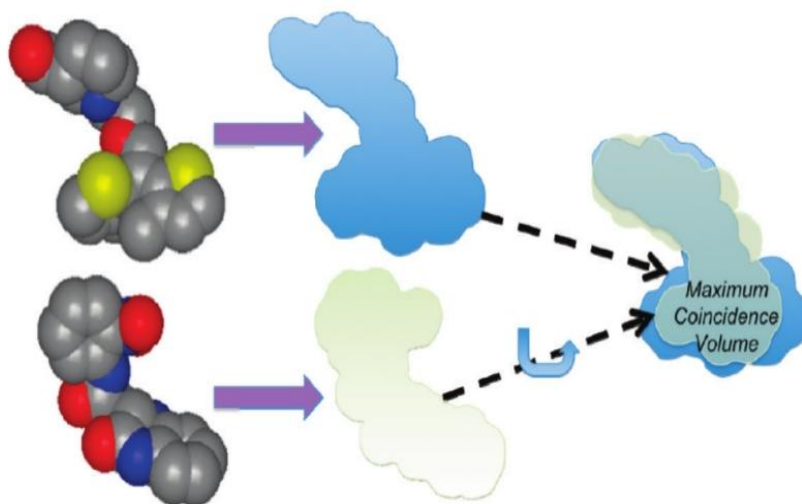


Figure 24. Illustration de la comparaison des formes de deux molécules en fonction du chevauchement de leur volume ¹⁰⁹

Les recherches de similarité basées sur la forme ont obtenu plusieurs succès dans la découverte de nouvelles molécules actives ^{121, 122}.

3.1.2 Métriques de similarité

Lors d'un criblage par similarité, la ou les molécules actives de références ainsi que l'ensemble des molécules de la chimiothèque sont décrites à l'aide d'un ou plusieurs descripteurs permettant d'obtenir des vecteurs de n éléments pour chaque molécule (Équation 1) ¹²³.

$$X_A = \{x_{1A}, x_{2A}, x_{3A}, \dots, x_{jA}, \dots, x_{nA}\}$$

Équation 1. Description d'une molécule A par un vecteur de n attributs x

Ceci permet ensuite de comparer ces molécules à l'aide de métriques de similarité. Ces métriques peuvent être classées en trois catégories : les mesures directes de similarité, les mesures de distances ou de dissimilitude et les mesures de corrélation. Les coefficients sont dits « d'association » lorsque qu'ils prennent chacun leur valeur dans l'intervalle $[0 ; 1]$, permettant ainsi par une simple soustraction, de convertir une métrique de distance en métrique de similarité et inversement. C'est par exemple le cas du coefficient de Tanimoto.

Lorsque l'on utilise des métriques de dissimilitude, plus des composés seront similaires, plus la valeur de leur coefficient de distance sera proche de 0. C'est notamment le cas des

distances euclidienne et de Hamming (autrement appelée distance de Manhattan) (Équation 2) qui sont couramment utilisées.

$$D_{A,B} = \left[\sum_{j=1}^{j=n} (|x_{jA} - x_{jB}|)^t \right]^{1/t}$$

Équation 2. Mesures de distance de Hamming (t=1) et Euclidienne (t=2) entre 2 composés A et B ¹²³

Ces deux distances sont dites monotoniques, c'est-à-dire qu'elles permettent d'obtenir la même classification de similarité de différents composés par rapport à une référence, même si leurs valeurs diffèrent.

Lorsque des descripteurs dichotomiques sont utilisés, c'est-à-dire que les deux seules valeurs possibles du descripteur sont 0 ou 1, marquant respectivement l'absence ou la présence d'une caractéristique particulière, la formule de ces distances peut être simplifiée (Équation 3).

$$D_{A,B} = [a + b - 2c]^{1/t}$$

Équation 3. Mesures de distance de Hamming (t=1) et Euclidienne (t=2) entre 2 composés A et B lorsque des descripteurs dichotomiques sont employés, avec a et b les nombres de bits des vecteurs des descripteurs égaux à 1 pour respectivement la molécule A et la molécule B, et c le nombre de bits des descripteurs égaux à 1 communs aux deux molécules A et B ¹²³

Au contraire des métriques de mesure directe de similarité, ces coefficients de distance considèrent comme élément de similarité l'absence commune d'une caractéristique donnée chez deux composés.

Parmi les différentes métriques de mesure directe de similarité, les coefficients de Tanimoto T_c (aussi appelé coefficient de Jacquard), Dice (ou coefficient de Czekanowski ou de Sørensen) et Cosine (possiblement connu sous le nom de coefficient de Ochiai) sont les plus populaires. Les valeurs de coefficients les plus élevées sont associées aux molécules les plus similaires.

Le coefficient de Tanimoto (T_c) est le plus couramment utilisé pour la comparaison de descripteurs de type « empreinte » qui sont dichotomiques (Équation 4) ⁹⁶.

$$T_{c(A,B)} = \frac{c}{a + b - c}$$

Équation 4. Formule du coefficient de Tanimoto pour deux molécules A et B en utilisant des empreintes 2D ou 3D comme descripteurs, avec a et b les nombres de bits des vecteurs des descripteurs égaux à 1 pour respectivement la molécule A et la molécule B, et c le nombre de bits des descripteurs égaux à 1 communs aux deux molécules A et B ¹²³

Cependant, le T_c peut aussi être employé pour des variables continues (Équation 5). Il a démontré son utilité et sa compétitivité non seulement dans la recherche de similarité mais aussi dans d'autres applications, ce qui fait de cette métrique la référence à laquelle les nouvelles méthodes de mesure de similarité sont comparées ¹²⁴.

$$T_{c(A,B)} = \frac{\sum_{j=1}^{j=n} x_{jA} x_{jB}}{\sum_{j=1}^{j=n} (x_{jA})^2 + \sum_{j=1}^{j=n} (x_{jB})^2 - \sum_{j=1}^{j=n} x_{jA} x_{jB}}$$

Équation 5. Formule du coefficient de Tanimoto pour deux molécules A et B en utilisant des variables continues

Le seul bémol à l'utilisation du T_c concerne les cas où le descripteur de type empreinte de la structure de référence ne présente pas beaucoup de bits positifs (marquant la présence d'un élément caractéristique), et pour lesquels le coefficient de Tanimoto attribue des faibles valeurs de similarité ¹²⁵.

Le coefficient de Dice est monotonique du coefficient de Tanimoto. Ce coefficient présente la particularité d'être un équivalent de l'index d'Hodgkin ¹²⁶ mesurant le chevauchement des fonctions de densité électronique. Il peut aussi bien être utilisé pour des descripteurs dichotomiques (Équation 6) ou des variables continues (Équation 7).

$$S_{(A,B)} = \frac{2c}{a + b}$$

Équation 6. Formule du coefficient de Dice pour deux molécules A et B en utilisant des empreintes 2D ou 3D comme descripteurs, avec a et b les nombres de bits des vecteurs des descripteurs égaux à 1 pour respectivement la molécule A et la molécule B, et c le nombre de bits des descripteurs égaux à 1 communs aux deux molécules A et B ¹²³

$$S_{(A,B)} = \frac{2 \sum_{j=1}^{j=n} x_{jA} x_{jB}}{\sum_{j=1}^{j=n} (x_{jA})^2 + \sum_{j=1}^{j=n} (x_{jB})^2}$$

Équation 7. Formule du coefficient de Dice pour deux molécules A et B en utilisant des variables continues ¹²³

Le coefficient de Cosine est quant à lui hautement corrélé avec le coefficient de Tanimoto, sans toutefois lui être strictement monotonique. Il s'agit d'un équivalent de l'index Carbo ¹²⁷ lui aussi utilisé pour mesurer le chevauchement des fonctions de densité électronique. A l'instar des coefficients de Tanimoto et Dice, il peut être employé pour des descripteurs dichotomiques (Équation 8) ou des variables continues (Équation 9).

$$S_{(A,B)} = \frac{c}{\sqrt{ab}}$$

Équation 8. Formule du coefficient de Cosine pour deux molécules A et B en utilisant des empreintes 2D ou 3D comme descripteurs, avec a et b les nombres de bits des vecteurs des descripteurs égaux à 1 pour respectivement la molécule A et la molécule B, et c le nombre de bits des descripteurs égaux à 1 communs aux deux molécules A et B ¹²³

$$S_{(A,B)} = \frac{\sum_{j=1}^{j=n} x_{jA} x_{jB}}{\sqrt{\sum_{j=1}^{j=n} (x_{jA})^2 \sum_{j=1}^{j=n} (x_{jB})^2}}$$

Équation 9. Formule du coefficient de Cosine pour deux molécules A et B en utilisant des variables continues ¹²³

Les coefficients de corrélation, tels que le coefficient de Pearson mesurent la corrélation entre deux variables. Les valeurs de coefficient de Pearson s'inscrivent dans l'intervalle [-1 ;1], plus la valeur absolue de ce coefficient s'approche de 1, plus les variables seront corrélées alors que des valeurs proches de 0 témoignent d'une absence de corrélation. Ce coefficient exprime le ratio entre la covariance et la déviation standard des variables étudiées, qu'elles soient dichotomiques (Équation 10) ou continues (Équation 11).

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

Équation 10. Formule du coefficient de Pearson pour deux molécules A et B en utilisant des empreintes 2D ou 3D comme descripteurs, avec a et b les nombres de bits des vecteurs des descripteurs égaux à 1 pour respectivement la molécule A et la molécule B, c le nombre de bits des descripteurs égaux à 1 communs aux deux molécules A et B et n la longueur de la chaîne de bits ¹²⁸.

$$Pearson_{(A,B)} = \frac{\sum_{j=1}^{j=n} x_{jA}x_{jB} - \frac{1}{n} \sum_{j=1}^{j=n} x_{jA} \sum_{j=1}^{j=n} x_{jB}}{\sqrt{\left(\sum_{j=1}^{j=n} x_{jA}^2 - \frac{1}{n} \sum_{j=1}^{j=n} x_{jA}^2\right) \left(\sum_{j=1}^{j=n} x_{jB}^2 - \frac{1}{n} \sum_{j=1}^{j=n} x_{jB}^2\right)}}$$

Équation 11. Formule du coefficient de Pearson pour deux molécules A et B lorsque les descripteurs constituent des variables continues

3.2 Modèles pharmacophoriques « ligand- based »

Le concept de pharmacophore a été développé par Ehrlich à la fin du XIX^e siècle ¹²⁹. A cette époque, même si le terme pharmacophore n'est pas employé, Ehrlich développe l'idée que certains groupes chimiques dans une molécule sont responsable de l'action biologique ou pharmacologique ¹³⁰. La première définition moderne du pharmacophore, utilisant le terme de « caractéristiques abstraites » plutôt que « groupes chimiques » date de 1960 ¹³¹. Le premier modèle de pharmacophore identifiant des ordres de grandeur de distance entre les caractéristiques constituant le pharmacophore (Figure 25a) a été publié en 1963 pour des agents muscariniques ¹³². Kier publie quant à lui avec son « modèle proposé de récepteur » (« proposed receptor pattern ») ¹³³ le premier modèle de pharmacophore avec des distances précises mesurées entre les différents groupements constituant le pharmacophore (Figure 25b).

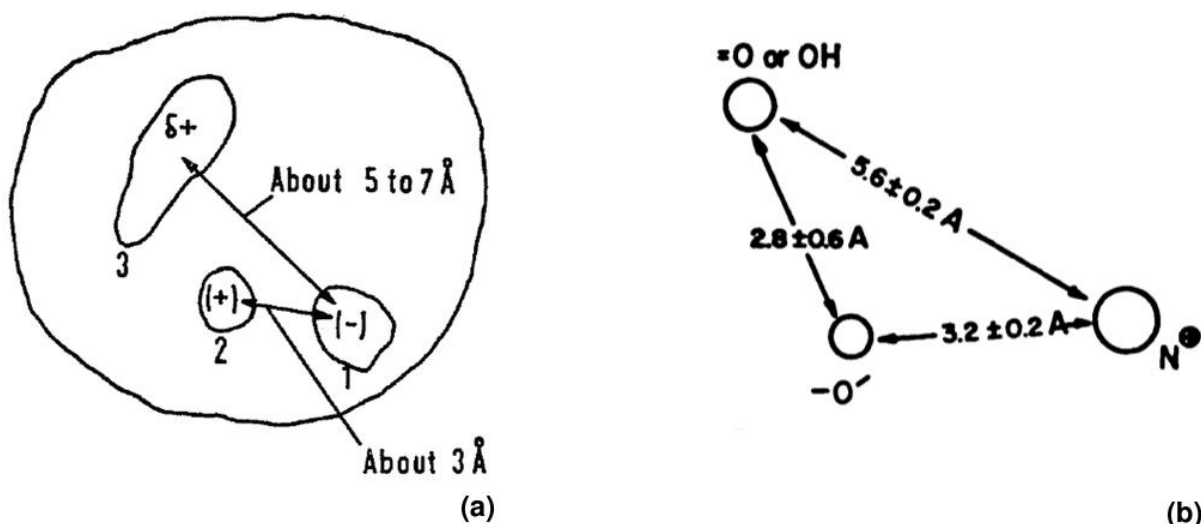


Figure 25. Présentation des premiers pharmacophores à avoir été publié, définissant un pharmacophore pour les agonistes muscariniques. Le modèle de Beckett (a) datant de 1963 définit des ordres de grandeur de distances entre la zone 1, une cavité anionique chargée négativement pour accueillir une amine quaternaire, la zone 2, un point chargé positivement permettant la liaison ou de l'acétylcholine et ses analogues et la zone 3, chargée pour interagir avec le OH de la muscarine, le C-O de l'acétylcholine et ses analogues ou la double liaison des analogues furaniques de la muscarine¹³². Le modèle de Kier (b) propose quant à lui des distances calculées entre 3 atomes clés communs à l'acétylcholine, la muscarine et la muscarone¹³³.

La définition officielle de l'IUPAC de 1998 indique que l'ensemble des propriétés stériques et électroniques d'une molécule, nécessaire pour assurer l'établissement d'interactions supramoléculaires optimales avec une cible biologique spécifique et engendrer ou bloquer une réponse biologique, constitue un pharmacophore¹³⁴.

D'après cette définition, des molécules partageant le même pharmacophore pour une cible donnée devraient donc se lier de manière identique à ce récepteur et présenter des profils d'activité similaires. Le pharmacophore généré est donc utilisé pour cribler la chimiothèque à la recherche de molécules se superposant à ce pharmacophore.

L'une des caractéristiques majeures de ce type de méthodes est qu'un pharmacophore est défini par des points pharmacophoriques complémentaires les uns des autres, qui sont des groupes fonctionnels et non plus des groupes d'atomes. Les différents points pharmacophoriques recherchés sont les donneurs et les accepteurs de liaisons hydrogènes, les groupements chargés positivement qui forment des interactions électrostatiques avec ceux chargés négativement et vice versa, et les groupements aromatiques, considérés distinctement

de la classe plus large des groupements hydrophobes dont ils sont issus, et qui sont tous deux complémentaires d'autres groupement hydrophobes.¹³⁵

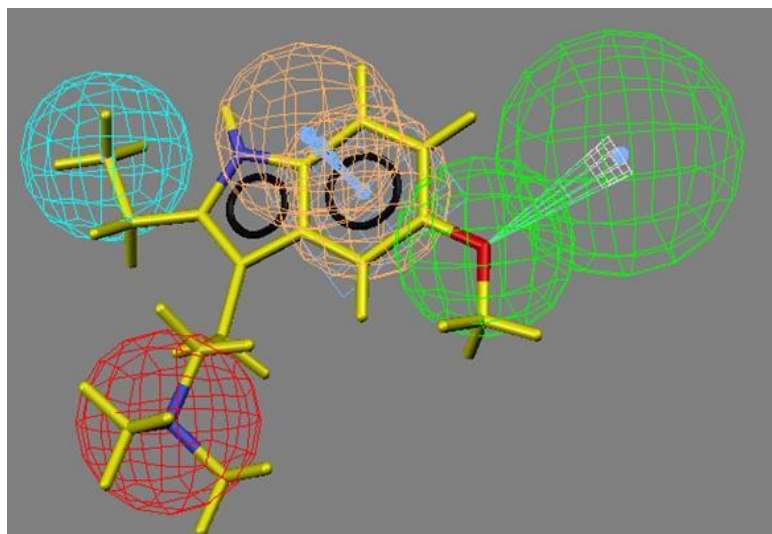


Figure 26. Pharmacophore généré à l'aide du logiciel CATALYST à partir de 5 antagonistes du récepteur de la sérotonine 5-HT_{5A} dont l'EMDT (2-Ethyl-5-methoxy-N,N-dimethyltryptamine) ici représenté. Les différents types de points pharmacophoriques sont illustrés par des sphères de couleur bleue pour les groupements hydrophobes, rouge pour les groupements chargés positivement, orange pour les aromatiques et verte pour les accepteurs de liaison hydrogène.

Un pharmacophore est dit basé sur le ligand (ou « ligand-based ») lorsqu'il est déterminé à partir de la structure de composés actifs de référence, sans connaître ou sans prendre en compte la structure du récepteur. Lorsque la structure du récepteur est utilisée pour construire le pharmacophore, celui-ci est dit basé sur la structure ou « structure-based » (voir paragraphe 4.2).

On distingue classiquement les approches pharmacophoriques 2D et 3D selon le format dans lequel sont présentés les ligands utilisés pour rechercher le pharmacophore.

3.2.1 Approches pharmacophoriques 2D

Même si les approches pharmacophoriques 2D ne contiennent pas d'informations sur les coordonnées atomiques des ligands, elles peuvent être très intéressantes, notamment du point des moindres temps de calculs associés. De nombreuses approches pharmacophoriques 2D ont été développées, parmi lesquelles les « Similog keys », l'ErG (Extended reduced Graph), les descripteurs CATS 2D (Chemical Advanced Template Search) et les « feature trees ».

Les « Similog keys »¹³⁶ sont à l'interface entre la recherche de similarité utilisant des descripteurs 2D et les approches pharmacophoriques 2D. Des triplets d'atomes forment ce qu'on appelle des « Similog keys » et sont caractérisés par le nombre de liaisons séparant les atomes et par les propriétés de l'atome : donneur (codé 1000) ou accepteur de liaisons hydrogène (codé 0100), encombrement (codé 0010) et électropositivité (codée 0001) (Figure 27). La molécule est alors représentée par un vecteur indiquant les occurrences des différents « Similog keys » dans la structure. La comparaison des vecteurs permet ensuite d'évaluer la similarité entre composés.

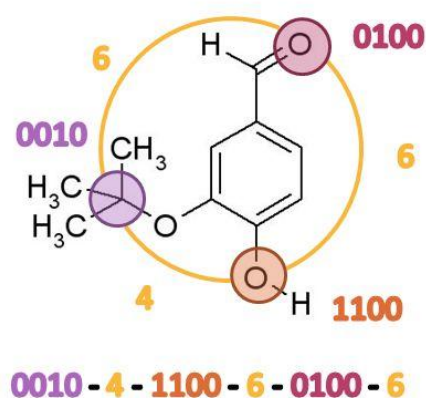


Figure 27. Exemple de « Similog keys » (d'après¹³⁶)

La méthode ErG (Extended reduced Graph)¹³⁷ représente une autre approche pharmacophorique 2D. Il s'agit d'une variante de la méthode des graphes réduits¹³⁸ dans laquelle seules les parties de la structure supposées importantes pour la liaison et l'activité sont conservées et dont les nœuds des graphes représentent des points pharmacophoriques. L'obtention des graphes ErG nécessite quatre étapes (Figure 28).

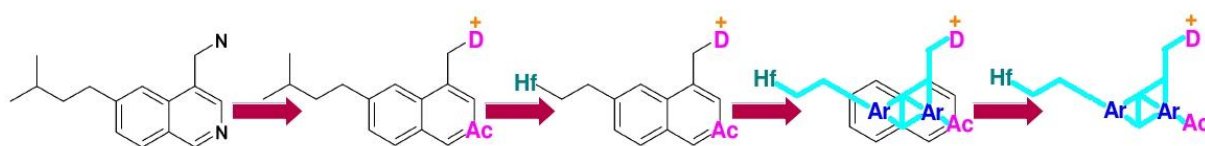


Figure 28. Conversion de la structure chimique en graphe réduit en 4 étapes (D : donneur de liaisons hydrogènes, Ac : accepteur de liaison hydrogène, Hf : groupement hydrophobe, Ar : groupement aromatique, + : charge positive) (d'après¹³⁷)

La première étape s'attache à ioniser la molécule selon les conditions physiologiques et à identifier les groupements donneurs et accepteurs de liaisons hydrogènes qui seront annotés respectivement D et Ac. L'étape suivante s'intéresse aux groupements hydrophobes terminaux constitués de 3 atomes conservés sous le terme Hf. La dernière étape consiste à

décrire les groupements aromatiques dans une forme abstraite. Ce graphe réduit est ensuite converti en descripteur qui est un vecteur formé de triplets de la forme PP1-DT-PP2 avec PP1 et PP2 les points pharmacophoriques 1 et 2 et DT la distance topologique séparant ces deux points. Ce vecteur permettra ensuite de rechercher dans la chimiothèque des molécules similaires à la molécule de référence.

Les descripteurs CATS 2D (Chemical Advanced Template Search) ¹³⁹ représentent une autre approche pharmacophorique 2D très proche de l'ErG. Cette méthode nécessite 4 étapes (Figure 29) ¹⁴⁰. Tout d'abord, la structure chimique est réduite à un graphe moléculaire. Ensuite chaque atome est remplacé par le point pharmacophorique correspondant, selon sa nature lipophile, aromatique, donneur ou accepteur de liaisons hydrogènes ou encore ionisable positivement ou négativement. L'étape suivante consiste à relier les paires de points pharmacophoriques selon le chemin le plus court en termes de liaisons. Le nombre normalisé d'occurrences de chaque paire de points pharmacophoriques en fonction des distances constitue un vecteur qui est ensuite utilisé pour comparer des molécules à l'aide de métriques telle que la distance euclidienne ¹⁴¹.

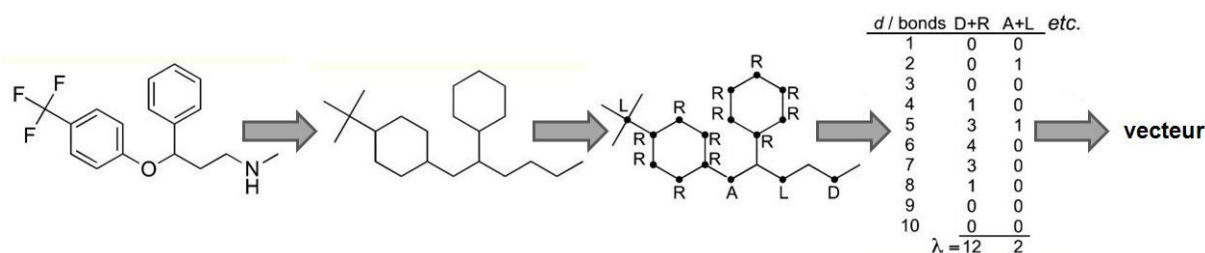


Figure 29. Principe de l'utilisation des descripteurs CATS 2D (R : Aromatique, L : Lipophile, A : Accepteur de liaisons hydrogènes, D : Donneurs de liaisons hydrogènes) (d'après ¹⁴⁰)

Enfin, la méthode dite des « feature trees » ¹⁴² (ou arbre des caractéristiques) représente la molécule sous la forme d'un arbre dont les nœuds décrivent les propriétés chimiques de groupes d'atomes leur permettant d'établir des interactions (Figure 30).

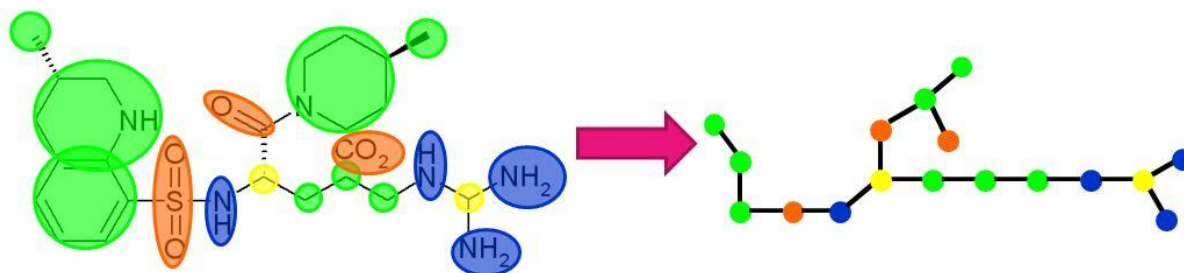


Figure 30. Exemple de construction d'un « feature tree » (en vert les nœuds principalement hydrophobes, en orange les accepteurs de liaisons hydrogènes, en bleu les donneurs de liaisons hydrogène et en jaune les atomes ne formant pas d'interactions directes) (d'après ¹⁴²)

Les arbres sont ensuite comparés les uns aux autres en les alignant de façon à maximiser le nombre de nœuds superposés et le score de similarité, $sim(m)$, est une moyenne pondérée des valeurs de similarité de toutes les parties superposées, calculées en prenant en compte la similarité chimique et stérique (Equation).

$$sim(m) = s(similarité\ stérique) + (1-s)(similarité\ chimique)$$

Équation 12. Calcul de la similarité des parties communes entre deux molécules ($sim(m)$) avec s un facteur de pondération approprié (d'après ¹⁴²)

3.2.2 Approches pharmacophoriques 3D

Les pharmacophores 3D décrivent l'arrangement spatial des propriétés chimiques nécessaires pour l'activité biologique ¹⁴¹ à partir d'un ensemble de ligands actifs de référence. Des pharmacophores 3D peuvent aussi être obtenus à partir de la structure 3D d'un complexe entre un ligand et son site de liaison ou de la structure 3D du récepteur, les pharmacophores générés sont alors basés sur la structure ou « structure-based » (voir paragraphe 4.2)

Une fois le pharmacophore obtenu, il est utilisé pour cribler la chimiothèque à la recherche de molécules potentiellement actives.

3.2.2.1 Elucidation du pharmacophore

L'élucidation d'un pharmacophore est un processus complexe divisé en plusieurs étapes ¹⁴³ (Figure 31).

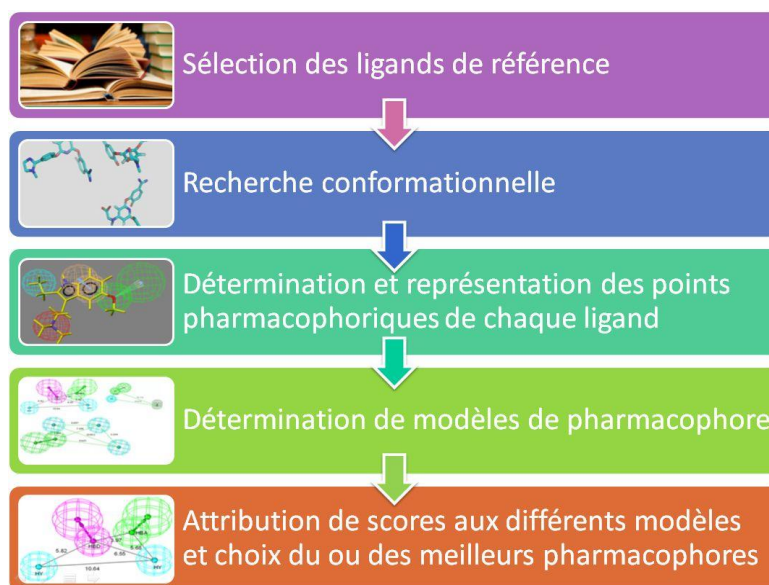


Figure 31. Principales étapes de l'élucidation d'un pharmacophore

De nombreux logiciels ont été développés présentant différentes approches notamment en termes de traitement de la flexibilité des ligands, de définition des points pharmacophoriques et de recherche des modèles de pharmacophore (Tableau 9).

3.2.2.1.1 Sélection des ligands de référence

Dans la plupart des méthodes, les ligands servant de référence pour construire un pharmacophore sont des composés actifs dont l'activité sur la cible biologique étudiée est renseignée et comparable pour tous, ce qui permet d'omettre les valeurs numériques d'activité pour la suite ¹⁴³. Cependant, ce n'est pas toujours le cas. En effet, certaines méthodes telles que DISCO (DIStance COmparison) ¹⁴⁴ ou CLEW ¹⁴⁵ utilisent des ligands inactifs pour rechercher des caractéristiques structurales néfastes pour l'activité. La méthode HypoGen (Hypothesis Generator) ¹⁴⁶ met à profit les valeurs d'activité des ligands de référence pour construire un pharmacophore permettant de prédire l'activité de nouvelles molécules. Généralement, moins de cent ligands de référence et les plus divers possible sont sélectionnés pour générer le pharmacophore afin d'identifier les caractéristiques les plus critiques pour la liaison au récepteur (tout en prenant garde à ce qu'ils se lient tous sur le même site de liaison).

3.2.2.1.2 Recherche conformationnelle

Les ligands utilisés pour créer le pharmacophore doivent être dans leurs conformations bioactives, c'est-à-dire la conformation avec laquelle ils se lient au récepteur. Cependant, lorsque celle-ci n'a pas été identifiée, une recherche conformationnelle doit être menée pour

pouvoir inclure toutes les conformations des ligands dans l'étude. Dans la majorité des cas, pour limiter les ressources computationnelles nécessaires, cette recherche conformationnelle a lieu séparément et préalablement à la génération du pharmacophore, comme dans les méthodes DISCO¹⁴⁷ et RAPID (RANdomized Pharmacophore Identification for Drug design)¹⁴⁸. D'autres méthodes procèdent à la recherche conformationnelle parallèlement à l'identification du pharmacophore comme par exemple dans les logiciels SCAMPI (Statistical Classification of Activities of Molecules for Pharmacophore Identification)¹⁴⁹, GASP (Genetic Algorithm Superposition Program)¹⁵⁰ ou GAMMA (Genetic Algorithm for Multiple Molecule Alignment)¹⁵¹.

3.2.2.1.3 Détermination et représentation des points pharmacophoriques de chaque ligand

Les points pharmacophoriques peuvent être de trois types¹⁴³ : basés sur les atomes, basés sur les groupes topologiques et basés sur les propriétés chimiques de groupes d'atomes.

La première et plus simple définition des points pharmacophoriques est basée sur les atomes, en s'intéressant à leurs coordonnées dans l'espace associées à leur type atomique, comme dans les méthodes MPHIL (Mapping Pharmacophores in Ligands)¹⁵², GAMMA¹⁵¹ et RAPID¹⁴⁸.

D'autres méthodes préfèrent regrouper les atomes en groupes topologiques (phenyle, carbonyle...) ¹⁴⁹.

Enfin, les points pharmacophoriques peuvent décrire les propriétés chimiques de groupes d'atomes, essentielles pour établir des interactions avec le site de liaison. Ces propriétés sont principalement accepteurs et donneurs de liaisons hydrogène, groupements chargés positivement ou négativement, aromatiques, groupements hydrophobes. Cette approche est largement employée dans de nombreux logiciels tels que SCAMPI¹⁴⁹, les deux méthodes d'identification de pharmacophores du logiciel CATALYST¹⁵³, HipHop¹⁵⁴ et HypoGen¹⁴⁶, MOE¹⁵⁵ ou encore PHASE (Pharmacophore Alignment and Scoring Engine)¹⁵⁶. Pour chaque ligand, les propriétés sélectionnées sont dessinées (Tableau 8) et combinées pour former une représentation de l'ensemble de la molécule.

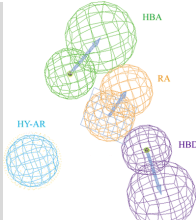
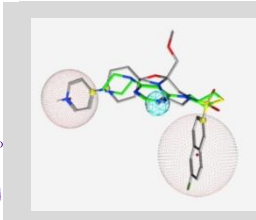
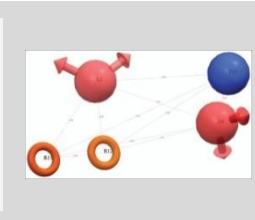
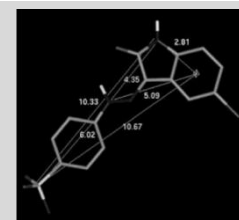
Propriétés	CATALYST	MOE	PHASE	SCAMPI
Interaction par liaison hydrogène	Sphères centrées sur les atomes lourds	Sphères centrées sur les atomes lourds (contraintes géométriques dépendent du schéma de pharmacophore choisi)	Vecteurs centrés sur l'atome d'hydrogène (donneur) ou sur l'atome lourd (accepteur) correspondant ou par leur point de projection	Atomes lourds
Groupelements hydrophobes	Sphères	Sphères (aromatiques ne sont pas considérés comme des hydrophobes)	Sphères (aromatiques ne sont pas considérés comme des hydrophobes)	Atomes lourds
Aromatiques	Sphères centrées sur le centroïde et anneaux plans indiquant l'orientation de l'aromatique	Sphères ou anneaux plans selon le schéma de pharmacophore choisi	Anneaux	Centroïdes
Interactions électrostatiques	Sphères	Sphères en précisant la charge	Sphères	Atomes lourds
Exemple				

Tableau 8. Mode de représentation des points pharmacophoriques et exemple de modèles de pharmacophore obtenu avec les logiciels CATALYST¹⁵⁷, MOE¹⁵⁸, PHASE¹⁵⁹ et SCAMPI¹⁶⁰ (d'après¹⁶¹)

Le choix du type de point pharmacophorique n'est pas anodin. En effet, si les points pharmacophoriques d'un modèle sont basés sur les atomes ou sur les groupes topologiques (par exemple un oxygène ou un carbonyle) seules les molécules possédant exactement ces atomes ou ces groupes topologiques peuvent être identifiées en tant que hits lors d'un criblage. A l'opposé, si les points pharmacophoriques décrivent des propriétés chimiques, le nombre de molécules pouvant satisfaire les critères du pharmacophore augmente puisque différents éléments peuvent représenter une même fonction chimique (par exemple un azote et un oxygène peuvent être tous deux accepteurs de liaisons hydrogène)¹⁴³.

3.2.2.1.4 Détermination de modèles de pharmacophore

Lorsque les points pharmacophoriques ont été définis pour chaque ligand, il est nécessaire de rechercher ceux qui sont communs aux différents ligands de référence afin d'obtenir un ou plusieurs modèles de pharmacophore. Pour cela, les ligands doivent être alignés et des algorithmes permettant de rechercher la sous-structure commune maximale (MCS) sont généralement employés, que ce soit des algorithmes de « Clique-detection » comme dans les méthodes DISCO¹⁴⁴ et MPHIL¹⁵², des algorithmes permettant une recherche exhaustive tels que ceux associés aux logiciels HipHop¹⁵⁴, HypoGen¹⁴⁶ et SCAMPI¹⁶⁰ ou encore des algorithmes génétiques comme implémentés dans les approches de GASP¹⁵⁰ et GAMMA¹⁵¹. Il est à noter que les méthodes HipHop¹⁵⁴ et MPHIL¹⁵² utilisent une approche de MCS dite assouplie. Ces méthodes ne pénalisent pas les modèles pour lesquels un nombre défini de ligands ne répondent pas aux propriétés du pharmacophore.

3.2.2.1.5 Attribution de scores aux différents modèles et choix du ou des meilleurs pharmacophores

Dans la dernière étape, un score est attribué aux différents modèles de pharmacophore ce qui permet ensuite de les classer. Les différentes fonctions de score utilisées se basent sur le nombre et la qualité de la superposition des points pharmacophoriques entre les ligands de référence, l'énergie conformationnelle, le volume de recouvrement entre les différents ligands¹⁶², mais aussi la rareté du pharmacophore. En effet, certains arrangements de points pharmacophoriques ne sont pas spécifiques de l'interaction des ligands de référence avec la cible biologique étudiée puisqu'ils sont retrouvés dans un très large nombre de molécules de profils d'activité divers. Les modèles de pharmacophores correspondant sont donc associés à un score faible. Au contraire, les arrangements de points pharmacophoriques présents dans de nombreux modèles générés pour les ligands de référence de l'étude mais rarement observés lors d'études précédentes avec divers jeux de données de ligands se voient attribuer un score très élevé. La rareté est ainsi prise en compte dans le score lors de l'utilisation des méthodes HipHop¹⁵⁴ ou Phase¹⁵⁶. Il est à noter que pour les méthodes qui utilisent des algorithmes génétiques pour générer les pharmacophores (GASP¹⁵⁰, GAMMA¹⁵¹), la phase d'attribution de score est incluse dans la génération des pharmacophores eux-mêmes.

Logiciel	Ligands de référence	Recherche conformationnelle	Points pharmacophoriques	Méthode de détermination des pharmacophores	Algorithme de détermination des pharmacophores	Score
DISCO et DISCOtech	Faible nombre (<100), actifs et inactifs	Avant la génération du pharmacophore	Basés sur les propriétés chimiques	MCS	Clique-détection	Nombre de ligands et de propriétés et distance entre propriétés
GASP	Faible nombre (<100), actifs	Concomittante à la génération du pharmacophore	Basés sur les propriétés chimiques	MCS	Algorithme génétique	Inclus dans la génération du pharmacophore
GAMMA	Faible nombre (<100), actifs	Concomittante à la génération du pharmacophore	Basés sur les atomes	MCS	Algorithme génétique	Inclus dans la génération du pharmacophore
CATALYST	Faible nombre (<100), actifs, valeurs d'activité (HypoGen)	Avant la génération du pharmacophore	Basés sur les propriétés chimiques	MCS (HypoGen) et MCS assouplie (HipHop)	Recherche exhaustive	Rareté (HipHop)
PHASE	Faible nombre (<100), actifs	Avant la génération du pharmacophore	Basés sur les propriétés chimiques	MCS assouplie	Recherche exhaustive	Superposition des propriétés et des volumes, rareté
MOE	Faible nombre (<100), actifs	Avant la génération du pharmacophore	Basés sur les propriétés chimiques	MCS	Aucun	Aucun
RAPID	Faible nombre (<100), actifs	Avant la génération du pharmacophore	Basés sur les atomes	MCS assouplie	Fusion linéaire	Superposition des propriétés
SCAMPI	nombre élevé (1000-2000), actifs	Concomittante à la génération du pharmacophore	Basés sur les propriétés chimiques	MCS	Recherche exhaustive	Inclus dans la génération du pharmacophore à l'aide un test de Student
CLEW	Faible nombre (<100), actifs et inactifs	Avant la génération du pharmacophore	Basés sur les propriétés chimiques	MCS	"Machine Learning" (algorithme génétique et programme évolutionnaire)	Superposition des propriétés
MPHIL	Faible nombre (<100), actifs	Avant la génération du pharmacophore	Basés sur les atomes	MCS assouplie	Clique-détection et algorithme génétique	Inclus dans la génération du pharmacophore à l'aide d'un test de Student

Tableau 9. Résumé des caractéristiques de quelques approches de pharmacophores 3D

3.2.2.2 Criblage de chimiothèques

Le pharmacophore ainsi généré peut ensuite être utilisé dans le processus de R&D pour cribler une chimiothèque à la recherche de nouveaux hits ¹⁴³, c'est ce qu'on appelle du « pharmacophore searching » ¹⁶³.

Pour s'assurer de la réussite d'un criblage à l'aide d'un pharmacophore, il est très important de prendre en compte la flexibilité des molécules composant cette chimiothèque, sous peine d'éliminer une molécule qui aurait pu satisfaire les critères du pharmacophore si elle avait été proposée dans la bonne conformation ¹⁴³. Pour cela, deux approches peuvent être utilisées. La première réalise un échantillonnage de conformations, en choisissant un ensemble de conformations pour chaque molécule (méthode « Fast » de CATALYST ¹⁶⁴, Chem-X ^{165, 166}). Cependant, pour couvrir l'espace conformationnel de chaque molécule, le nombre de conformations requis peut être très important et des conformations pertinentes peuvent être omises. La seconde approche consiste donc à traiter la flexibilité des molécules au moment du criblage de la chimiothèque en utilisant le pharmacophore pour guider cette recherche (UNITY, CFS ¹⁶⁷, 3DFS ¹⁶⁸). La méthode « Best Flexible Search » de CATALYST ¹⁶⁴ permet de combiner les deux approches, c'est-à-dire que la chimiothèque utilisée présente les molécules dans diverses conformations et lors du criblage, la flexibilité des molécules est mise à profit pour tenter de s'adapter au pharmacophore.

Le criblage de chimiothèques à l'aide d'un pharmacophore comporte généralement deux étapes :

- La première étape consiste en un pré-filtrage rapide de la chimiothèque pour éliminer les composés qui ne possèdent pas les propriétés de base du pharmacophore (comparaison des types de propriétés et de leur nombre, empreintes...) ¹⁶⁹.
- La seconde étape consiste à identifier au sein des molécules issues du pré-filtrage celles répondant aux critères du pharmacophore. Pour cela, la méthode la plus populaire est la recherche de sous-structures ¹⁰⁰, basée sur la théorie des graphes ^{170, 171}. Dans cette approche, des graphes sont construits à la fois pour le pharmacophore de référence et pour les conformères étudiés, les noeuds représentant les points pharmacophoriques reliés par des arêtes représentant les distances entre les points pharmacophoriques. Lorsqu'un conformère satisfait les contraintes spatiales du pharmacophore, il est considéré comme un hit. Dans le cas contraire, il peut être éliminé de la recherche lorsque la flexibilité des molécules a été prise en compte avant le criblage, ou il peut être le point de départ d'une recherche conformationnelle pour tenter de satisfaire les critères du pharmacophore ¹⁴³.

Ainsi, le criblage de chimiothèques basé sur les pharmacophores permet de découvrir de nouveaux hits, qui peuvent être très similaires de ligands de référence déjà connus ou au contraire constitués de nouvelles classes de composés. Les pharmacophores permettent d'identifier les propriétés importantes pour l'activité d'une molécule, sans évaluer le poids de chacune d'entre elles, au contraire des méthodes QSAR.

3.3 Modèles de relations quantitatives structure-activité (QSAR)

L'analyse quantitative des relations existantes entre les structures d'un ensemble de composés et leurs activités permet d'identifier et d'évaluer l'impact des propriétés influençant l'activité biologique. Cette relation peut être décrite par une équation qui corrèle mathématiquement les influences réciproques des paramètres concernés. L'extrapolation de ces résultats peut servir de base à la prédiction de l'activité de nouveaux composés.¹⁷²

Lors d'une étude QSAR (Figure 32), il faut calculer à partir de la structure moléculaire de plusieurs actifs tous les descripteurs moléculaires possibles. Après élimination des descripteurs dont la valeur ne varie pas ou peu sur l'ensemble des molécules, une analyse multivariée suivie d'une évaluation statistique sont menées pour obtenir le modèle QSAR. Une dernière étape de validation du modèle est nécessaire pour s'assurer de sa fiabilité.

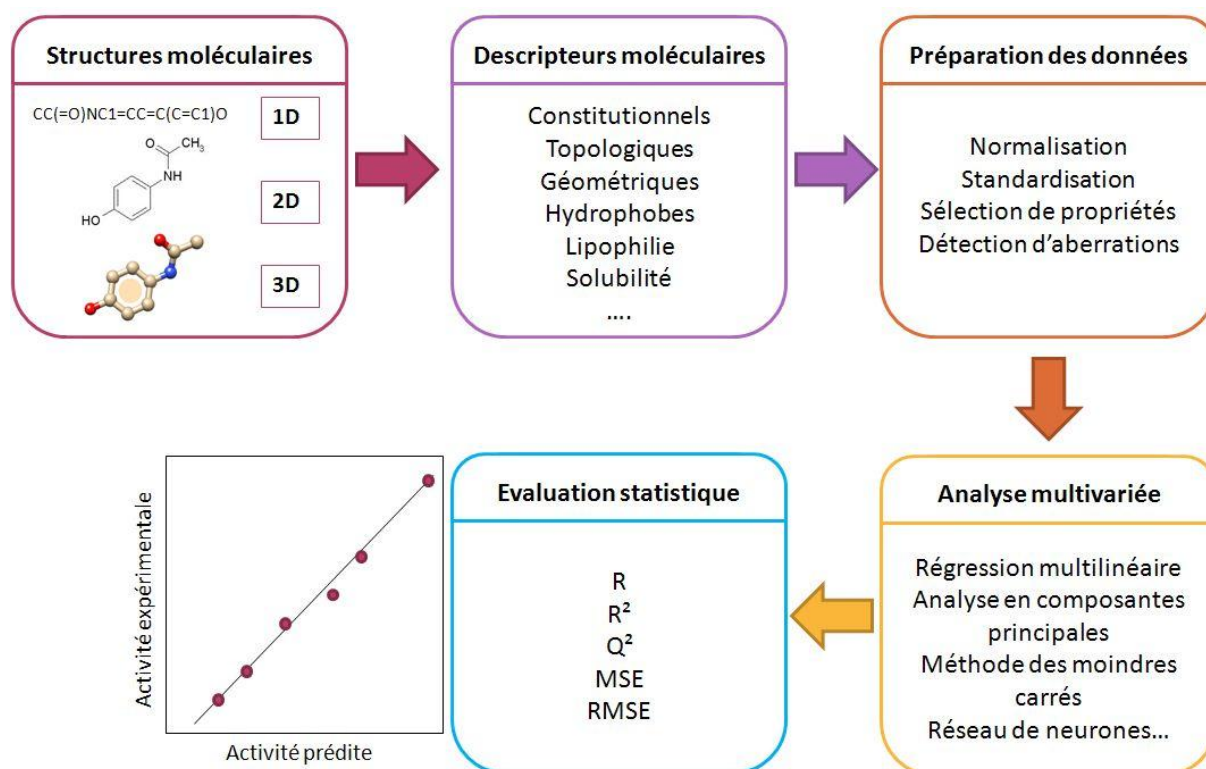


Figure 32. Schéma général des différentes étapes d'une étude QSAR (d'après ¹⁷³)

Il existe plus de 3000 descripteurs moléculaires pouvant être utilisés lors d'études QSAR ⁹⁷, provenant de différentes théories et approches, et qui peuvent être obtenus soit directement à partir de la structure chimique, soit en utilisant des logiciels appropriés. Les descripteurs les plus utilisés lors des études QSAR peuvent être des descripteurs topologiques (2D-QSAR), des descripteurs géométriques ou basés sur des grilles (3D-QSAR). De nouvelles approches plus récentes, QSAR-4D, -5D et -6D ont été développées pour améliorer les approches QSAR-3D, en prenant en compte différentes conformations des ligands (4D ¹⁷⁴), et l'adaptation structurale de la cible au ligand (5D ¹⁷⁵) et les effets de solvants (6D ¹⁷⁶).

3.3.1 QSAR-2D

Le premier modèle QSAR-2D a été proposé par Hansch dans les années 60 ¹⁷⁷ qui tentait de relier l'activité biologique de régulateurs de croissance de plantes à des paramètres d'hydrophobie, électroniques et stériques. Depuis, de nombreux modèles ont été proposés, basés sur différents descripteurs qui peuvent être classifiés en constitutionnels, topologiques, de la chimie quantique, et fragmentaux.

Les descripteurs constitutionnels représentent les propriétés d'un composé sans prendre en compte sa connectivité ou sa géométrie. Le poids moléculaire, le nombre d'atomes (et en particulier d'hydrogène, de carbone et d'halogènes), le nombre de cycles aromatiques, le nombre de liaisons simples, doubles, triples ou aromatiques, etc... font partie des descripteurs constitutionnels et sont fréquemment utilisés dans les études QSAR ¹⁷³.

Les descripteurs topologiques décrivent les orientations des liaisons, la taille de la molécule, la forme, les ramifications et la présence d'hétéroatomes. Ainsi, l'indice de Wiener qui est la somme de toutes les liaisons chimiques existantes entre toutes les paires d'atomes autres que les hydrogènes dans une molécule est le plus ancien et l'un des descripteurs les plus utilisés lors des études QSAR. De nombreux autres descripteurs topologiques tels que les indices de Randic, Balaban, Scultz, Kier-Hall sont basés sur la connectivité des molécules. L'indice de charge topologique de Galvez décrit quant à lui le transfert de charge ayant lieu au sein d'une molécule. Les descripteurs BCUT (Burden-CAS-University of Texas) sont aussi très intéressants puisqu'ils décrivent les propriétés atomiques cruciales pour les interactions telles que la charge atomique, la polarisabilité et la capacité à établir des liaisons hydrogènes ¹⁷⁸.

Les descripteurs de la chimie quantique s'intéressent aux propriétés électroniques et géométriques des molécules. Les plus fréquemment utilisés sont les charges atomiques, les énergies d'orbitales moléculaires, les densités d'orbitales frontières, la polarisabilité, les indices de moments dipolaires et de polarité,...

Les descripteurs fragmentaux sont employés pour analyser les effets de certains fragments d'une molécule sur son activité. C'est notamment le cas de la constante de Hammett qui s'intéresse aux effets de polarisation de différents substituants sur le pKa des groupements phénols correspondants ¹⁷⁹, de la constante de Hansch qui peut être utilisée pour estimer la lipophilie d'une molécule ¹⁸⁰ ou encore du facteur stérique de Taft ¹⁸¹.

Une fois les descripteurs obtenus, une régression linéaire est calculée pour chaque descripteur et l'analyse statistique des résultats permet d'éliminer un certain nombre de descripteurs en fonction de leur absence d'influence sur l'activité et de leur intercorrélacion. Le nombre de descripteurs dans le modèle final doit être aussi faible que possible et peut être déterminé par la méthode du point de rupture. ¹⁸²

3.3.2 QSAR 3D

Les approches 3D-QSAR ont été développées pour corréler l'activité biologique d'une série de composés actifs de référence avec l'arrangement spatial de nombreuses propriétés de la molécule telles que les propriétés stériques, lipophiliques et électroniques.¹⁸³ L'analyse d'un modèle 3D-QSAR permet donc de fournir des indications pour l'optimisation par phamacomodulation et la conception de nouveaux composés avec des profils d'activité améliorés.

La première approche 3D-QSAR, proposée en 1979, décrivait des propriétés de champs moléculaires de composés, calculées sur une grille régulière¹⁸⁴ puis corrélées à leur activité biologique par analyse en composante principale (PCA). Cette méthode, plus tard dénommée DYLOMMS (Dynamic Lattice-Oriented Molecular Modeling System) ne prend en réalité son essor que grâce à l'application de la méthode des moindres carrés partiels (ou Partial Least Squares PLS) à la corrélation des propriétés à l'activité biologique.

Actuellement différentes méthodes 3D-QSAR sont utilisées parmi lesquelles CoMFA (Comparative Molecular Field Analysis), CoMSIA (Comparative Molecular Similarity Indices Analysis), GRID/GOLPE et Phase. Il est à noter que toutes ses méthodes nécessitent un alignement minutieux des ligands de référence. Lorsque la structure 3D de la cible biologique est résolue, des modèles 3D-QDAR dit « receptor dependent » (RD-QSAR) peuvent être mises en œuvre (voir paragraphe 4.3).

3.3.2.1 Analyse comparative des champs moléculaires : CoMFA

Les méthodes CoMFA¹⁸⁵ ont été les premières approches 3D-QSAR développées. L'idée de base des études CoMFA repose sur le fait que les différences d'activité biologique entre les molécules s'expliquent souvent par des différences dans la forme et la force des champs d'interactions non covalentes entourant les molécules¹⁸⁶. Autrement dit, les champs stériques et électroniques suffiraient pour comprendre les propriétés biologiques d'un ensemble de composés¹⁸³. Ainsi, lors d'une étude CoMFA (Figure 33), des molécules dont l'activité biologique est connue et comparable, sont virtuellement alignées et placées dans une grille cubique de résolution usuelle de 2 Å. Les énergies d'interactions entre les molécules et des atomes « sondes » sont calculées pour chaque point de la grille par les potentiels de Lennard-Jones et Coulomb. Les sondes sont des atomes possédant les propriétés de Van der Waals d'un carbone sp³ et une charge de + 1,0¹⁸⁵. Les valeurs d'énergie d'interaction mesurées sont notées dans un tableau, dans lequel chaque ligne correspond à une des molécules de référence,

et qui possède des milliers de colonnes qui doivent être corrélées à l'activité biologique. Pour cela, la méthode la plus appropriée est l'analyse partial least square (PLS) ou régression des moindres carrés partiels permettant d'obtenir une équation de régression avec des milliers de coefficients.

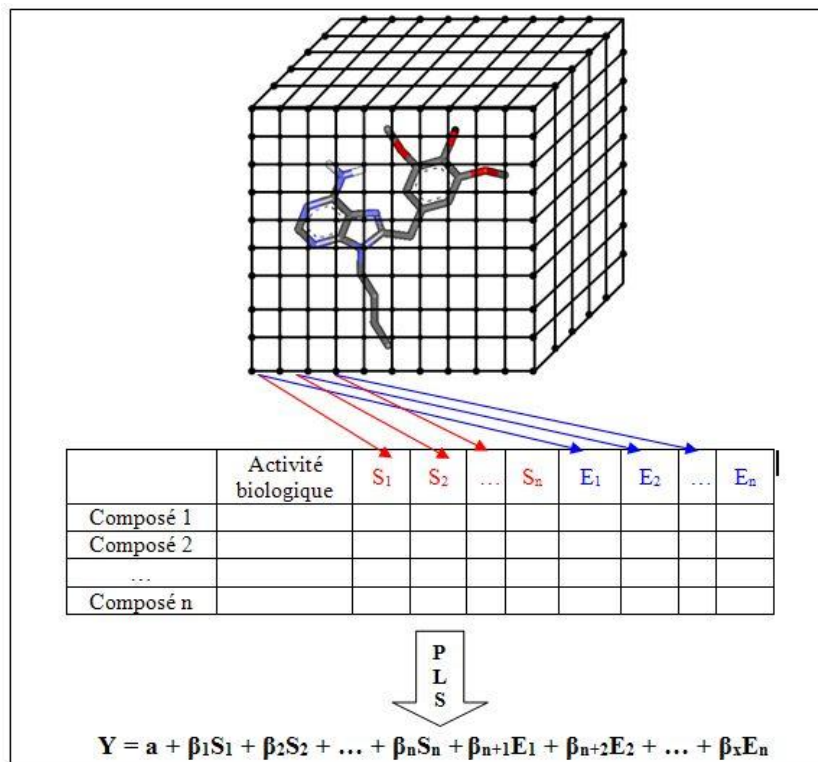


Figure 33. Déroulement d'une étude CoMFA

Le plus souvent, ces résultats sont représentés sous la forme d'aires de contour et indiquent les régions stériques et électrostatiques favorables ou défavorables autour des molécules (Figure 34).

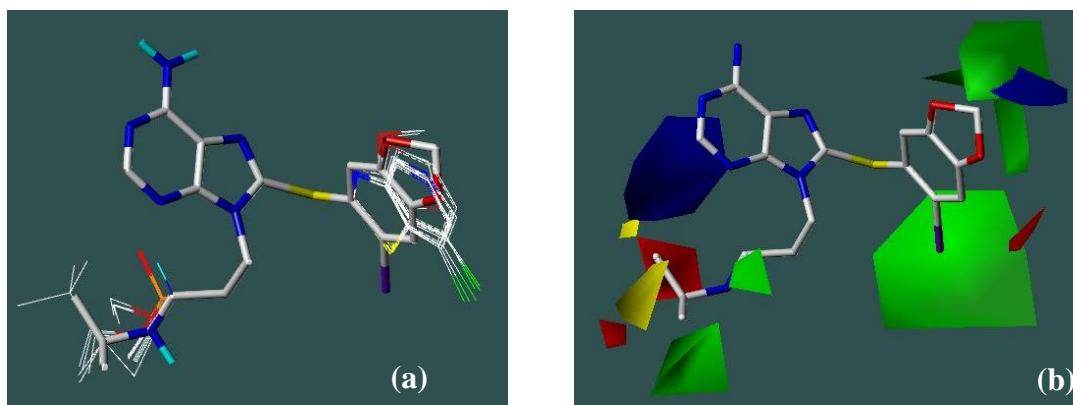


Figure 34. (a) Alignement des 14 molécules sur la structure cristallisée du PU-H71 (code PDB : 2FWZ) et (b) Représentation du modèle CoMFA permettant de visualiser la contribution stérique (région où des substituants encombrants sont favorables (vert) ou défavorable (jaune) à l'activité, et la contribution électrostatique (région où un potentiel négatif est favorable (bleu) ou défavorable (rouge) à l'activité) (source : rapport de M2 de Nathalie Lagarde, laboratoire BioCIS, 2010)

3.3.2.2 Analyse comparative d'indices de similarité moléculaire : CoMSIA

Le problème majeur rencontré lors des études CoMFA réside dans la forme (pentes abruptes) des champs électrostatiques mais aussi et surtout stériques calculés par les potentiels de Coulomb et de Lennard-Jones (Figure 35). Ainsi, les variables pour le champ stérique peuvent parfois se réduire à des valeurs proches de 0 (lorsqu'il n'y a aucun atome autour) ou atteignant la valeur limite définie (atomes à l'intérieur de la molécule). Ceci conduit à des aires de contour fragmentées et difficilement interprétables.

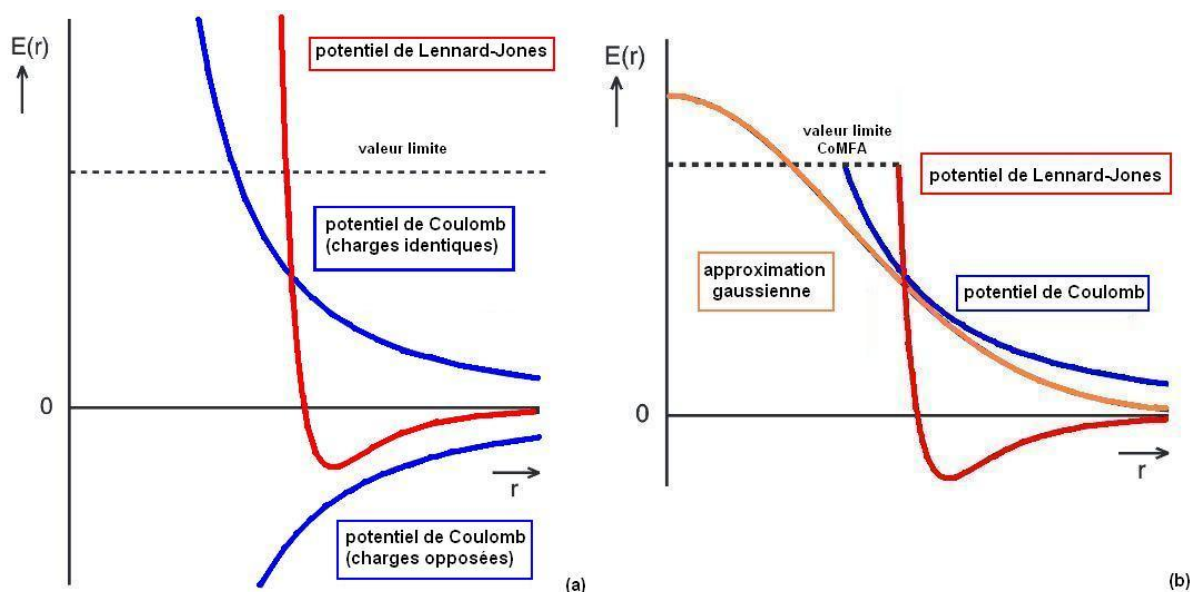


Figure 35. (a) Courbes des potentiels de Lennard-Jones (rouge) et Coulomb (bleu) utilisées dans les études CoMFA et définition d'une valeur limite supérieure. (b) La fonction gaussienne en forme de cloche des champs SEAL (orange) utilisée lors des études CoMSIA est une bonne approximation des potentiels de Lennard-Jones et Coulomb tout en présentant l'avantage d'être plus lissée (d'après ¹⁸⁷)

Pour pallier ce problème, Klebe et ses collègues ¹⁸⁸ proposent une méthode CoMFA basée sur des indices de similarité qu'ils nomment CoMSIA. Les indices de similarité stérique, électrostatique et hydrophobe sont obtenus à partir d'une forme modifiée de l'algorithme SEAL (Steric and Electrostatic Alignment) ¹⁸⁹. La fonction gaussienne utilisée pour décrire les molécules représente non seulement une bonne approximation des potentiels de Lennard-Jones et Coulomb (Figure 35) mais permet aussi d'obtenir des aires de contour facilitant l'interprétation des résultats ¹⁸³. Tout comme dans les études CoMFA, les molécules sont placées dans une grille cubique et des sondes sont situées sur chaque point de la grille. Les indices de similarité sont calculés à chaque point de la grille, portant principalement sur les propriétés stériques, électrostatiques, d'hydrophobie et de liaisons hydrogène. La corrélation de ces propriétés à l'activité biologique est ici encore représentée par des aires de contour (Figure 36).

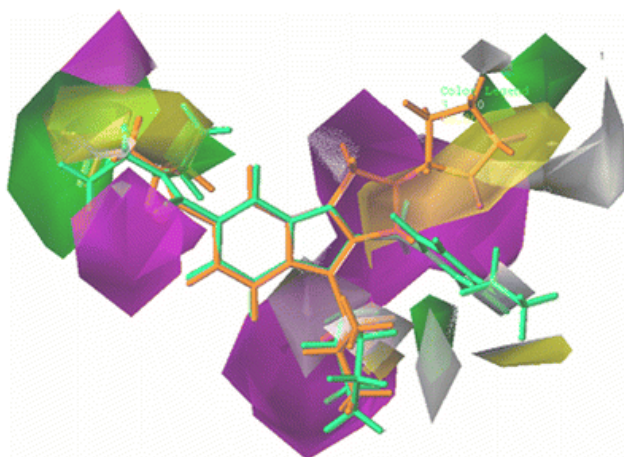


Figure 36. Aires de contour d'un modèle CoMSIA permettant de visualiser la contribution hydrophobique favorable (jaune) ou défavorables (gris), et la contribution des accepteurs de liaisons hydrogène favorable (violet) ou défavorable (vert) à l'activité ¹⁹⁰

3.3.2.3 GRID/GOLPE

Le programme GRID ¹⁹¹ peut être utilisé en association avec le programme GOLPE (Generating Optimal Linear PLS Estimations) ¹⁹² pour réaliser une étude 3D-QSAR. En effet, le programme GRID permet de calculer des champs d'interactions avec une fonction plus facilement interprétable que celles utilisées dans la méthode CoMFA et propose différents types de sondes (méthyle, carbonyle, hydroxyle, amines...). Le programme GOLPE se charge du déroulement de l'analyse statistique des résultats en proposant des fonctionnalités permettant de sélectionner les variables significatives pour la prédiction de l'activité biologique des composés (test de multiples combinaisons de variables, prise en compte d'ensemble de variables plutôt que des variables uniques séparées).

Comme dans les approches précédentes, les résultats peuvent être visualisés sous forme d'aires de contour autour des ligands (Figure 37).

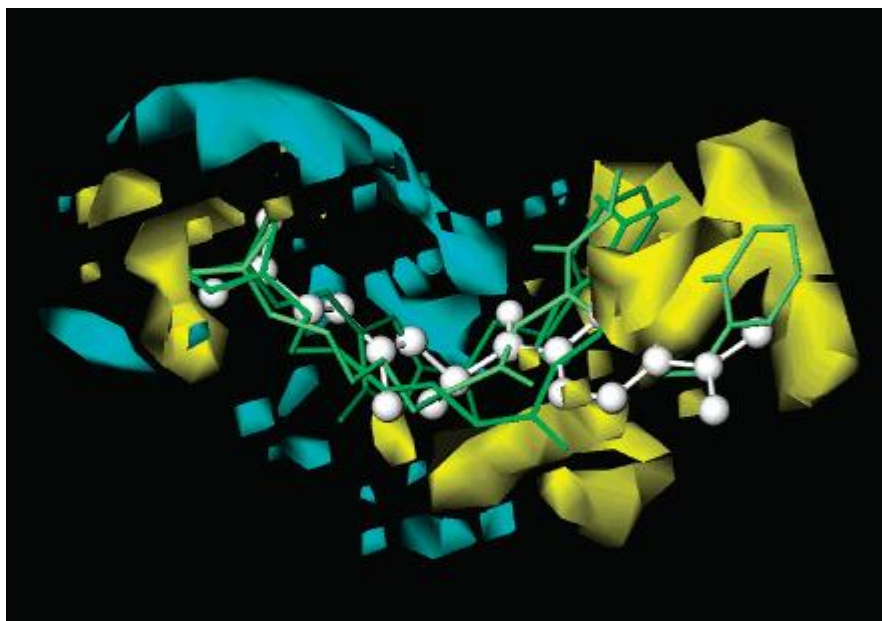


Figure 37. Aires de contour obtenues par une approche GRID/GOLPE pour des inhibiteurs des histones déacetylases (en blanc la trichostatine A et en vert le composé TAA_5A). Les aires de contours colorées en cyan illustrent les zones de contribution négative à l'activité alors que celles colorées en jaune indiquent les zones favorables à l'activité¹⁹³

3.3.2.4 Phase

Le logiciel Phase¹⁵⁶ est un logiciel permettant non seulement de générer des pharmacophores mais aussi de réaliser des études 3D-QSAR. L'approche (Figure 38) est relativement similaire à celle utilisée dans la méthode CoMFA. Un pharmacophore des molécules de référence est généré. Ces molécules sont alignées par rapport à ce pharmacophore et sont placées dans une grille cubique, de 1 Å de résolution qui divise l'espace en cubes de même taille. Chaque point de la grille est occupé par des sphères de Van der Waals, dont le rayon dépend du type atomique.

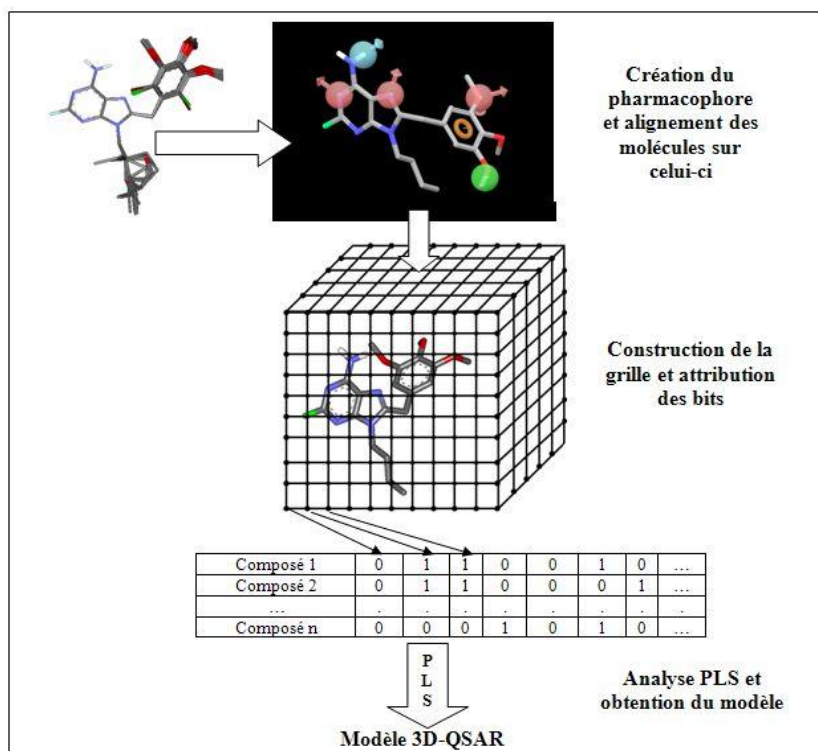


Figure 38. Principales étapes du déroulement d'une étude 3D-QSAR à l'aide du logiciel Phase

Un « bit » est attribué à chaque classe d'atome qui occupe un cube donné correspondant à une caractéristique de pharmacophore (donneur de liaison hydrogène, hydrophobe/non polaire, ionisation négative ou positive, électroattracteur et miscellanés). Chaque molécule peut donc être représentée par une série d'une centaine valeurs de « bits » (0 ou 1) qui indique quels cubes sont occupés par quels atomes de chaque classe. L'analyse PLS est ensuite utilisée pour générer le modèle QSAR. Il est représenté par un ensemble de cubes de couleur bleue ou rouge. La couleur bleue représente un coefficient positif, indiquant une augmentation de l'activité et à l'inverse la couleur rouge symbolise un coefficient négatif témoignant d'une diminution de l'activité (Figure 39).

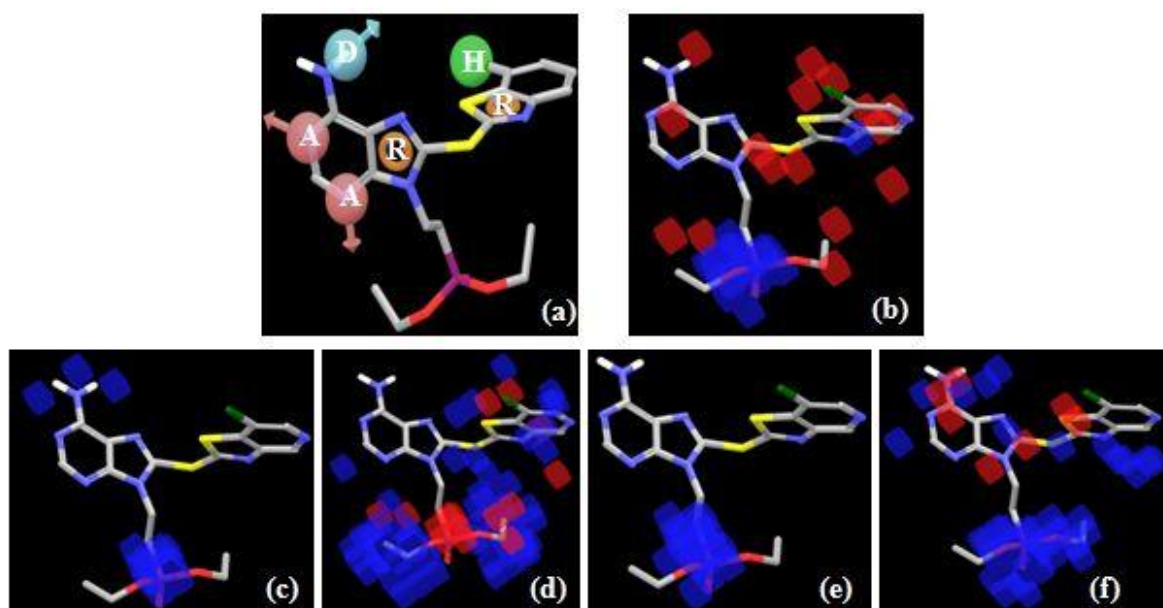


Figure 39. (a) Hypothèse de pharmacophore AADHRR pour des inhibiteurs de la HSP90 (A : accepteur de liaison hydrogène, D : donneur de liaisons hydrogènes, H : hydrophobe, R : aromatique) et (b) Représentation du modèle 3D-QSAR obtenu (cubes bleus : coefficient positif, rouges : coefficient négatif) et contribution de chaque classe d'atomes : (c) donneur de liaisons hydrogène, (d) hydrophobe, (e) ionisation positive et (f) électroattracteur (source : rapport de M2 de Nathalie Lagarde, laboratoire BioCIS, 2010)

Grâce au développement de programmes faciles à prendre en main avec interprétation visuelle des résultats et à la diversité des méthodes disponibles, l'intérêt pour les approches 3D-QSAR a explosé au cours des années 2000. Fin 2007, plus de 2500 études 3D-QSAR étaient recensées dans le Chemical Abstract Service. Cependant, ces méthodes ne sont pas applicables sur des jeux de données comportant un très grand nombre de molécules, tels que ceux typiquement utilisés lors des criblages.¹⁸³

3.4 Succès du criblage virtuel « ligand-based »

Les méthodes de criblages virtuels basés sur les ligands sont très populaires puisqu'elles permettent de rationaliser les processus de découverte de nouveaux composés, lorsque des données sont disponibles sur un ou plusieurs actifs de référence. Ces méthodes peuvent à la fois être employées dans la phase d'identification de nouveaux hits mais aussi dans les phases d'optimisation des hits et des leads. De nombreux succès ont déjà été obtenus pour découvrir de nouveaux composés actifs en utilisant ces méthodes, que ce soit les approches de recherche de similarité, pharmacophoriques ou QSAR (Tableau 10).

Méthodes	Approche	Auteur	Sujet
Recherche de similarité 2D	CATS	Schneider et al. ¹³⁹	Antagonistes Ca ²⁺ (bloqueurs des canaux T)
Recherche de similarité 2D	CATS	Naerum et al. ¹⁹⁴	Inhibiteurs de la glycogène synthase kinase-3
Approche pharmacophorique 3D	CATALYST	Singh et al. ¹⁹⁵	Antagonistes de l'antigène alpha4beta1 (VLA-4)
Approche pharmacophorique 3D	CATALYST	Flohr et al. ¹⁹⁶	Antagonistes non peptidiques de l'urotensine II
Approche pharmacophorique 3D	GALAHAD	Mustata et al. ¹⁹⁷	Inhibiteurs de la formation de l'hétérodimère cMyc-Max
Recherche de similarité 3D basée sur la forme	ROCS	Rush et al. ¹⁹⁸	Inhibiteurs de l'interaction protéine-protéine ZipA-FtsZ
Recherche de similarité 3D basée sur la forme	ROCS	Boström et al. ¹⁹⁹	Antagonistes des récepteurs cannabinoïdes CB1
Combinaison d'approches 2D (empreintes 2D) et 3D (similarité de forme, pharmacophore)	Empreintes 2D: Daylight et MDL ; Similarité de forme: ROCS ; Pharmacophore: ALMOND	Bologa et al. ²⁰⁰	Agonistes GPR30
Combinaison d'approches 2D (2D-QSAR) et 3D (3D-QSAR et recherche de similarité de forme)	2D-QSAR : descripteurs fragmentaux ; 3D-QSAR : CoMFA ; Similarité de forme : ROCS	Freitas et al. ²⁰¹	Inhibiteurs de la cruzaine (<i>Trypanosoma cruzi</i> cysteine protease)
Combinaison d'approches 2D (2D-QSAR) et 3D (pharmacophore)	2D-QSAR: algorithme génétique, Pharmacophore: CATALYST	Al-Sha'er et al. ²⁰²	Inhibiteurs de la Heat Shock Protein 90 alpha

Tableau 10. Quelques succès obtenus dans la découverte de nouveaux composés actifs en utilisant des méthodes LBVLS

Les méthodes LBVLS ont notamment contribué à la mise sur le marché d'un médicament anti-cancéreux, le gefitinib (Iressa®). En effet, au cours du processus de R&D, une approche pharmacophorique 3D suivie d'une approche pharmacophorique 2D a permis l'identification d'un hit, le CAQ, dont l'optimisation a donné naissance au gefitinib ^{203, 204}.

4 Criblage virtuel « structure-based »

Lorsque la structure 3D de la cible biologique d'intérêt est disponible, des méthodes dites basées sur la structure peuvent être employées pour réaliser le criblage virtuel. Il existe deux types de méthodes expérimentales pour obtenir la structure 3D d'une cible : la cristallographie aux rayons X et la résonance magnétique nucléaire (RMN). Lorsque la structure 3D expérimentale n'a pas encore été résolue, des méthodes de prédiction de la structure 3D par homologie de séquence peuvent être mises en œuvre (Figure 40).

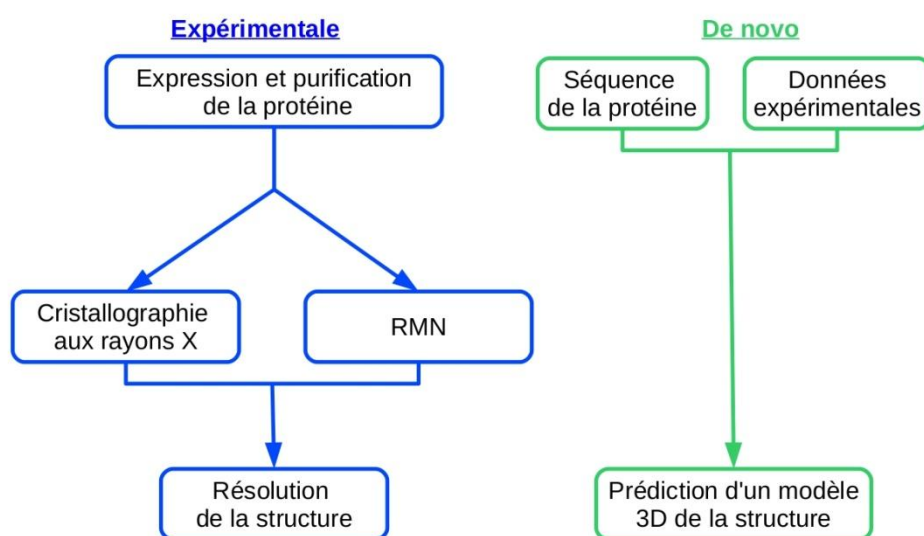


Figure 40. Méthodes pour obtenir la structure 3D d'une cible biologique d'intérêt: expérimentale (cristallographie aux rayons X et RMN) ou de novo (prédiction par homologie)

Comme leur nom l'indique, ces méthodes utilisent la structure de la cible pour découvrir de nouveaux composés actifs. Pour cela, différentes approches peuvent être employées : la construction de modèles pharmacophoriques basés sur la structure, l'établissement de modèles RD-QSAR, la conception *de novo* (ou *de novo* design) et les méthodes de docking qui sont les plus populaires. Toutes ces techniques nécessitent l'identification préalable du site de liaison.

4.1 Identification du site de liaison

L'identification du site de liaison est nécessaire et cruciale pour pouvoir réaliser un criblage virtuel basé sur la structure. Pour cela, deux solutions sont possibles : obtenir une structure 3D de la protéine co-cristallisée avec un ligand, ou utiliser des outils de prédiction²⁰⁵.

4.1.1 Structure co-cristallisée avec un ligand

L'analyse de la structure cristallisée d'une protéine avec un ligand permet de rechercher les interactions s'établissant entre ces deux partenaires²⁸. Les acides aminés clés, c'est-à-dire impliqués dans les interactions ligand-protéine, permettent ainsi de guider la définition du site de liaison. Il s'agit de la méthode la plus sûre pour identifier un site de liaison puisqu'elle utilise des données expérimentales et non pas des données prédites.

4.1.2 Outils de prédiction de site de liaison

Les outils de prédiction de site de liaison peuvent être divisés en trois catégories, ceux basés sur la géométrie, ceux basés sur les énergies et ceux basés sur la connaissance²⁰⁵.

4.1.2.1 Outils de prédiction basés sur la géométrie

Les outils de prédiction basés sur la géométrie utilisent comme hypothèse de départ l'assertion que les poches et les cavités sont souvent associées aux sites de liaison. En effet, diverses études ont suggéré que les sites de liaison sont souvent situés dans la plus grande poche de la protéine^{206, 207, 208}. De nombreux logiciels s'attachent donc à identifier celles-ci au sein de la structure protéique. Pour cela, la plupart utilise une grille tridimensionnelle pour définir la surface moléculaire (MOLCAD²⁰⁹, POCKET²¹⁰, LIGSITE²¹¹ et son implémentation Pocket-Finder²¹², VolSite²¹³ ...). Ainsi, le programme MOLCAD (MOlecular Computer Aided Design)²⁰⁹ utilise l'algorithme de Connolly²¹⁴ permettant de calculer la surface exclue au solvant ou surface de Connolly pour identifier au sein d'une structure les canaux et les cavités. Pour cela, la surface de Connolly est générée pour la protéine entière et est placée dans une grille cartésienne. Tous les points de la grille qui se trouvent dans l'espace défini par la surface de Connolly sont nommés « in » et tous les autres points de la grille « out ». Pour chaque point de la grille « out », les points voisins de moins de 12 Å sont étudiés. Si un point de la grille « out » possède des points voisins « in » dans au moins 2 directions de l'espace, le point « out » est défini comme étant un « point de la cavité ». Tous les « points de la cavité » sont combinés ensemble pour former des clusters. Deux opérations de « logique cellulaire » sont effectuées sur ces clusters : la « contraction » et « l'expansion ». La contraction définit que tout point de la cavité avec un voisin qui n'est pas un point de la cavité est supprimé alors que l'expansion affirme que tout point de la cavité avec au moins un voisin défini comme étant un point de la cavité est ajouté au cluster correspondant. A travers ces opérations, les petits clusters sont éliminés et ceux de taille plus

importante sont subdivisés en plus petits clusters. Ceci permet d'obtenir un ou plusieurs clusters représentant des régions concaves de la protéine. Des algorithmes de surface moléculaire utilisant des sondes parcourant la surface de la molécule peuvent aussi être employés (Automatic PROtein POcket Search APROPOS ²¹⁵, CASTp ²¹⁶, SurfNet ²¹⁷). Le logiciel CASTp (Computed Atlas of Surface Topography of proteins) ²¹⁶ utilise un système de triangulation (dit de Delaunay) pour identifier et mesurer le volume et la surface des poches accessibles mais aussi les cavités intérieures inaccessibles (Figure 41). Les ouvertures de la molécule permettant d'accéder aux poches et cavités sont aussi recherchées.

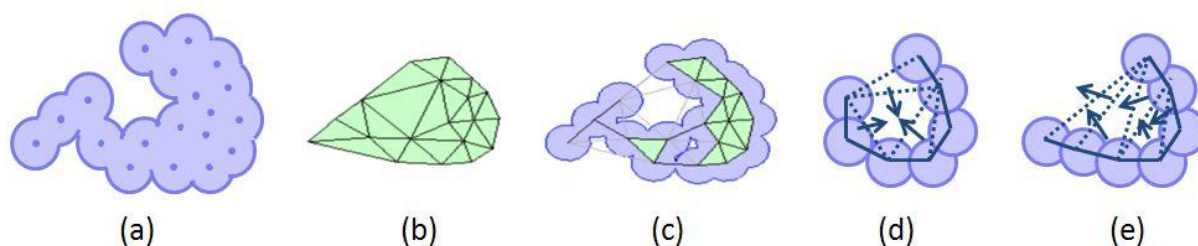


Figure 41. Schématisation de la triangulation de Delaunay pour un modèle simplifié d'atomes possédant tous le même rayon (a). Lorsque l'on relie le centre de tous les atomes, un polygone est formé (b) qui peut être triangulé de manière à ce que tout le polygone soit couvert sans superposition de triangles (c). Selon la méthode « discrete flow », un triangle agit comme un « puit » pour les triangles voisins et la poche est définie (d). Dans certains cas, ce « puit » ne peut pas être créé et CASTp ne considère donc pas cette partie comme une poche. (d'après ²¹⁸)

4.1.2.2 Outils de prédiction basés sur les énergies

Les outils de prédiction basés sur les énergies tentent d'estimer les énergies d'interaction entre une sonde (qui peut être un groupement méthyle, hydroxyle ou amine) et un point donné de la protéine pour définir des zones favorables d'interaction. Différentes techniques emploient cette approche (GRID ²¹⁹, méthode de Ruppert et al ²²⁰), parmi lesquelles le logiciel Q-SiteFinder. La recherche de sites de liaison avec Q-SiteFinder ²¹² est réalisée par liaison de sondes hydrophobes (CH₃) à la protéine puis génération de clusters en regroupant les zones où les sondes se lient avec les énergies de liaison les plus favorables (Figure 42).

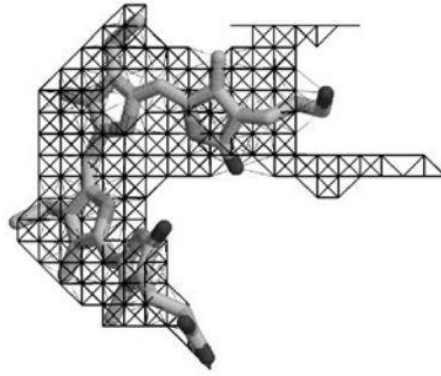


Figure 42. Site de liaison (code PDB: 1BBP) pour la protéine liant la biline (Biling Binding Protein BBP) prédit avec Q-SiteFinder. Les sondes utilisées pour prédire ce site correspondent aux nœuds de la grille. (²¹²)

Les clusters sont rangés par ordre de probabilité d'être un site de liaison en se basant sur la somme totale des énergies de liaison pour chaque cluster. Contrairement aux outils de prédiction basés sur la géométrie, les volumes des sites prédits ne sont que très faiblement corrélés aux volumes des sites de liaison de la protéine.

4.1.2.3 Outils de prédiction basés sur la connaissance

Les outils de prédiction basés sur la connaissance utilisent notamment des données biochimiques, de mutagenèse dirigée et de similarité de séquence ou structurale ²⁰⁵. La recherche de similarité de séquence, reposant sur la haute conservation de séquence des sites de liaison, est employée par différents outils de prédiction (Rate4Site ²²¹, ConSurf ²²², l'algorithme de Dai et al. ²²³) réalisant des études d'homologie avec des protéines similaires. La recherche de similarité structurale avec un complexe connu protéine/ligand peut aussi permettre d'identifier un site de liaison, et ce particulièrement lorsque l'on considère des enzymes catalysant la même réaction ²⁰⁵. Pour cela, différentes bases de données regroupant des informations sur les sites de liaisons et permettant leur comparaison sont disponibles, parmi lesquelles CavBase ²²⁴, PINTS (Patterns In Non-homologous Tertiary Structures) ²²⁵, SiteEngine ²²⁶, eF-site ²²⁷, ProFunc ²²⁸, SitesBase ²²⁹. WebFEATURE ²³⁰ est un outil d'analyse structurale qui permet aux utilisateurs de scanner des structures à la recherche de sites fonctionnels. Le logiciel fournit des modèles de sites précédemment générés et testés (Figure 43), à utiliser lors de la recherche de sites de liaison, par exemple pour les protéines : des sites de liaison pour le calcium, des sites de liaison pour le chlore et des sites de liaison de

l'ATP. Le modèle choisi est utilisé pour scanner la structure à tester à la recherche d'un site similaire, c'est-à-dire présentant un environnement physico-chimique identique.

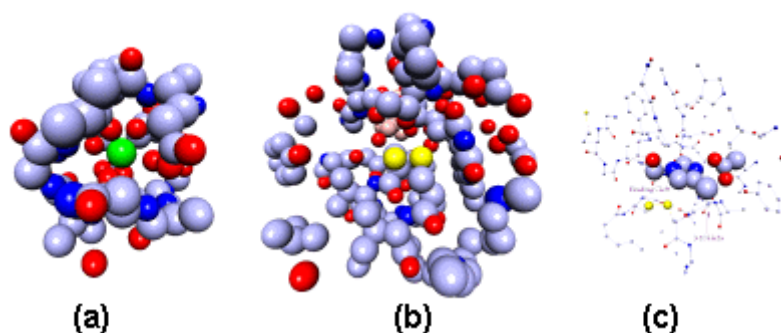


Figure 43. Exemples de modèles de sites proposés par WebFEATURE représentant l'environnement 3D en utilisant différentes propriétés physico-chimiques : (a) site de liaison pour le calcium, (b) site de pont disulfure et (c) site actif de sérine protéase. ⁽²³¹⁾ (en bleu : carbones, en bleu foncé : azotes, en rouge : oxygènes, en vert : calcium, jaune : soufre)

Parmi tous les logiciels utilisables pour identifier un site de liaison, certains sont disponibles librement sur internet. C'est notamment le cas de CASTp, Pocket-Finder, Q-SiteFinder et WebFeature qui proposent des interfaces graphiques pour visualiser le résultat de leur prédiction (Figure 44).

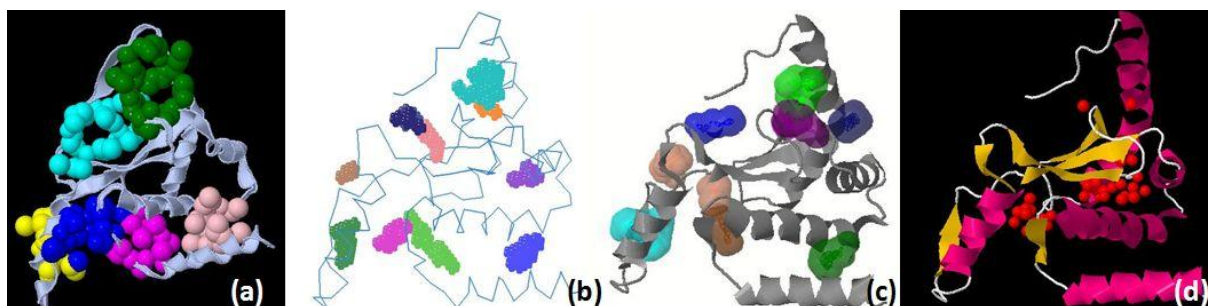


Figure 44. Résultats de la recherche d'un site de liaison sur le domaine C-terminal de la Hsp90 d'un modèle construit par homologie à l'aide des logiciels Pocket-Finder (a), CASTp (b) Q-SiteFinder(c) et WebFEATURE (d)

4.2 Modèles pharmacophoriques basés sur la structure du récepteur

Comme expliqué précédemment (voir paragraphe 3.2), les modèles pharmacophoriques basés sur les ligands permettent d'identifier de nouveaux composés à partir d'actifs de référence. Cependant, lorsqu'elle est disponible, prendre en compte la structure 3D du site de liaison

permet d'obtenir de nouveaux modèles pharmacophoriques incorporant des informations cruciales et inexploitées dans les modèles basés sur les ligands. Deux types d'approches pour construire un modèle pharmacophorique basé sur la structure du récepteur peuvent être distinguées : les approches basées sur le récepteur (« receptor-based approach ») et celles basées sur le complexe récepteur-ligand (« complex-based approach »)¹⁴³.

4.2.1 Approche basée sur le récepteur

L'approche basée sur le récepteur nécessite comme point de départ la structure 3D du site de liaison de la cible d'intérêt pour décrire ses propriétés et leurs relations spatiales. Contrairement à l'approche basée sur le complexe récepteur-ligand, cette structure 3D ne doit pas présenter de ligand co-cristallisé dans le site de liaison. Cependant, la connaissance de structures de ligands de référence actifs est généralement cruciale pour pouvoir discriminer les propriétés importantes pour la liaison des ligands parmi toutes celles identifiées. La définition du pharmacophore nécessite généralement cinq étapes (Figure 45).

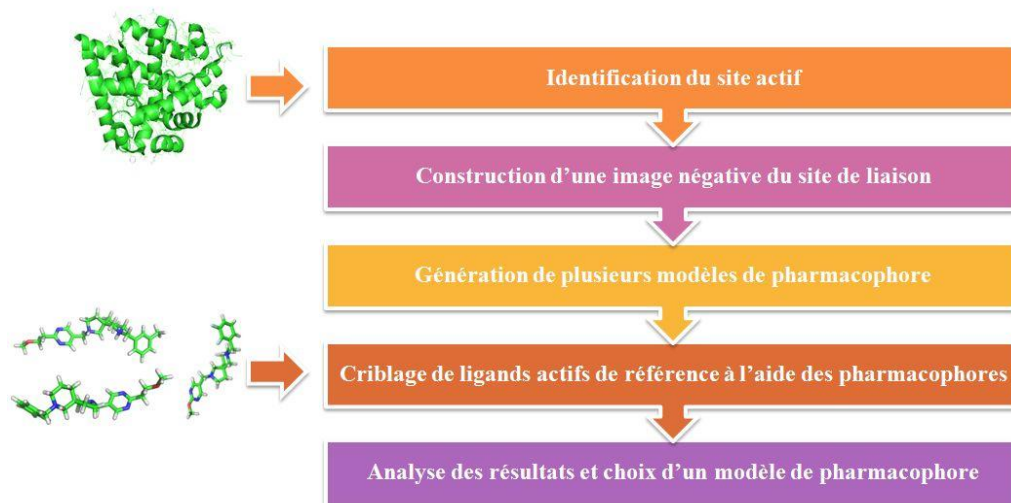


Figure 45. Schématisation des différentes étapes jalonnant la construction d'un modèle de pharmacophore basé sur la structure du récepteur (d'après¹⁴³)

Après avoir identifié le site de liaison, une image négative du récepteur est créée. Pour cela, plusieurs méthodes peuvent être utilisées. Ainsi, le logiciel LUDI^{232, 233}, initialement développé pour les approches de conception *de novo*, permet de générer une carte d'interaction du site de liaison. Cette carte regroupe les propriétés complémentaires de celles identifiées dans le site de liaison. Par exemple, une propriété de type donneur de liaison identifiée dans le site de liaison sera représentée par une propriété de type accepteur de liaison sur la carte d'interaction. D'autres méthodes^{234, 235} prennent en compte la flexibilité du

récepteur en utilisant des simulations de dynamique moléculaire pour générer des images du récepteur dans différentes conformations. Des sondes, représentant des groupes fonctionnels, sont ensuite placées sur chacune de ces images par complémentarité avec la surface du récepteur. Les sondes communes à un grand nombre d'images sont incorporées dans l'image négative du site de liaison. Quelque soit la méthode utilisée, les images négatives du site de liaison obtenues comportent souvent un trop grand nombre de propriétés pour pouvoir générer un modèle unique de pharmacophore, et cela même après regroupement des propriétés de même type. La solution à ce problème consiste à utiliser la structure des molécules actives de référence pour guider la sélection des propriétés à utiliser dans le pharmacophore. Pour cela, différents modèles de pharmacophore composés d'un plus petit nombre de propriétés et représentant toutes leurs combinaisons possibles sont donc générés. Ces modèles sont ensuite testés sur un ensemble de composés actifs de référence et l'évaluation de leurs performances permet de choisir le modèle de pharmacophore à utiliser pour cribler la chimiothèque d'intérêt.

4.2.2 Approche basée sur le complexe récepteur-ligand

L'analyse de la structure 3D d'un complexe ligand-récepteur permet d'obtenir des informations directes et cruciales sur les interactions s'établissant entre les deux. Certains logiciels, tels que ZINCPharmer²³⁶, permettent d'obtenir à partir d'un seul complexe des modèles de pharmacophore pouvant ensuite être utilisés pour cribler la base de données ZINC. Cependant, pour s'assurer de la robustesse et de la fiabilité des pharmacophores générés, il est souvent indispensable de regrouper les informations obtenues à partir de différents complexes. Pour cela, un alignement des différentes structures 3D des complexes est nécessaire. Les groupes fonctionnels impliqués dans les interactions sont généralement assez bien alignés et peuvent donc être extraits pour construire le pharmacophore. Le logiciel LigandScout²³⁷ permet ainsi de générer des pharmacophores à partir de complexes ligand-récepteur (Figure 46).

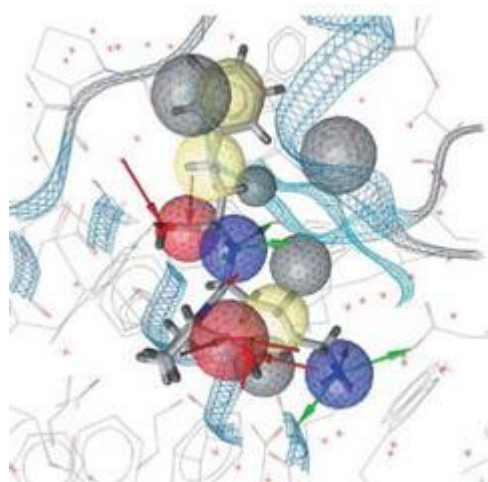


Figure 46. Modèle de pharmacophore généré pour le complexe Angiotensin Converting Enzyme (ACE) lisinopril, code PDB 1O86 (flèche verte : donneur de liaison hydrogène, flèche rouge : accepteur de liaison hydrogène, sphère jaune : hydrophobe, sphère noire : volume exclus, sphère rouge : groupe ionisable négativement, sphère bleue : groupe ionisable positivement) ²³⁸

4.3 RD-QSAR (Receptor Dependent-Quantitative Structure-Activity Relationship)

Le principe des méthodes 3D-QSAR dépendantes du récepteur consiste à établir une corrélation entre l'activité biologique représentée par l'énergie libre de liaison et des interactions ligand-récepteur déduites de la structure 3D d'un complexe correspondant ²³⁹. Le logiciel VALIDATE ²⁴⁰ est l'une des premières approches développées dans ce sens. A partir de 12 propriétés physico-chimiques, parmi lesquelles les interactions stériques et électrostatiques récepteur-ligand, une régression PLS permet de prédire l'activité biologique de nouveaux composés. Comme dans les méthodes 3D-QSAR non dépendantes du récepteur, les résultats sont visibles comme des aires de contour autour des ligands (Figure 47).

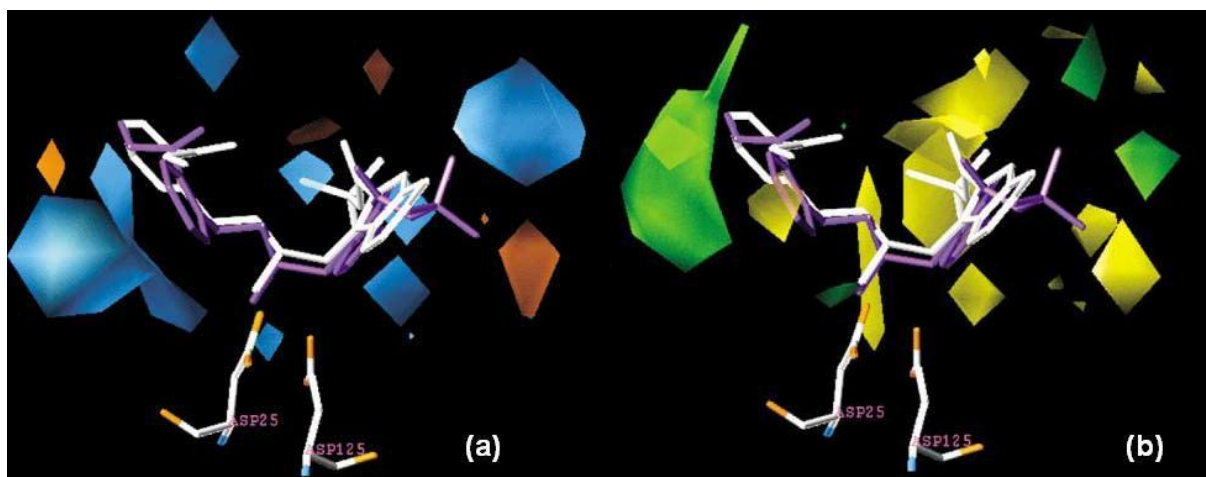


Figure 47. Exemple d'application de la méthode VALIDATE pour la recherche d'inhibiteurs de la HIV-1 protease (les contours oranges et bleus représentent des zones où des groupes chargés respectivement positivement et négativement diminuent l'activité, les contours jaune et vert symbolisent des zones dans lesquelles la présence de groupements encombrants est respectivement défavorable et favorable) ²⁴¹

De même, la méthode COMBINE (Comparative BINDing Energy) ²³⁹ se base sur des calculs d'interactions entre récepteur et ligand pour tenter de prédire l'activité biologique. Cependant, dans cette approche, les ligands et les récepteurs ne sont pas considérés dans leur ensemble mais sont fragmentés et les interactions entre ces différents fragments sont renseignées. Ces interactions sont les interactions stériques et électroniques pouvant survenir entre chaque fragment du ligand et chaque fragment du récepteur, les changements d'énergie des termes non liés (stérique et électrostatique) et liés (liaison, angle, torsion) des fragments de ligands et des fragments de récepteur survenant lors de liaison du ligand au récepteur. L'analyse des résultats se fait à l'aide d'une régression PLS telle qu'implémentée dans le programme GOLPE. Le logiciel gCOMBINE ²⁴² permet de visualiser les résultats (Figure 48).

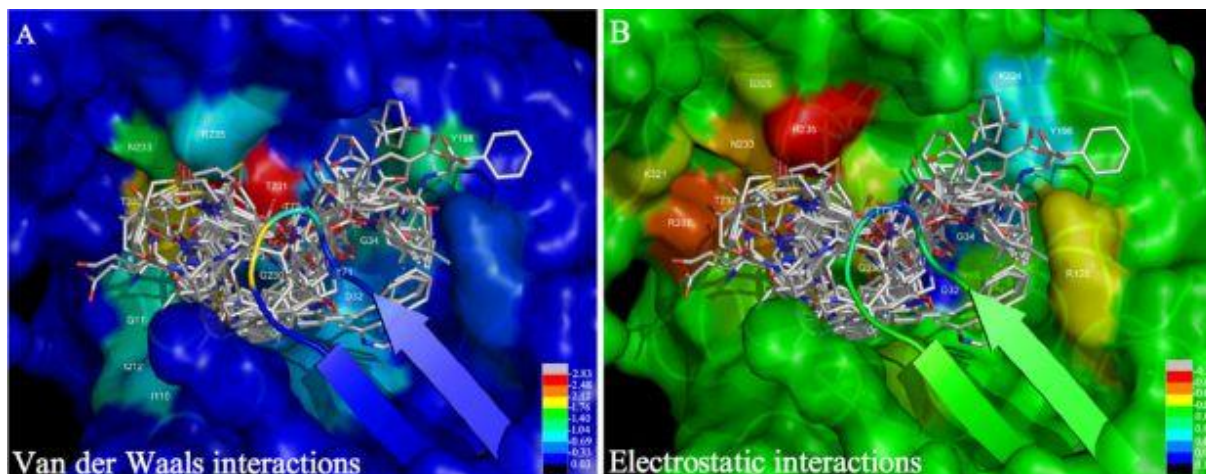


Figure 48. Coloration de la surface du récepteur BACE-1 (code PDB 1W51) avec 46 inhibiteurs superposés dans le site de liaison, à l'aide des coefficients PLS de van der Waals (A) et électrostatiques (B) ²⁴³

D'autres approches RD-QSAR utilisent la structure 3D du complexe récepteur-ligand pour réaliser l'alignement initial de tous les ligands actifs de référence. A partir de cet alignement, une étude 3D-QSAR classique peut ensuite être menée. ¹⁸³

4.4 Conception de novo

La conception *de novo* désigne l'approche consistant à obtenir des composés bioactifs par une construction incrémentale de ligands dans le site actif d'un récepteur (ou d'un enzyme) ²⁴⁴. La structure du site de liaison est utilisée comme point de départ de la recherche de nouveaux actifs. Cependant, lorsque celle-ci n'est pas disponible, des méthodes basées sur la structure du ligand peuvent aussi être employées (Tableau 11).

Nom	Blocs de construction		Basée sur		Stratégie de recherche					Construction des ligands					Fonctions de score	
	Atomes	Fragments	Rc	Li	DFS	BFS	Rnd	MC	AE	Gr	Lk	Lat	DM	Sto		
HSITE/2D Skeletons ²⁴⁵ (1989)		X	X			X					Adaptation et accrochage squelettes plans					Contraintes stériques et liaisons hydrogènes
3D Skeletons ²⁴⁶ (1990)		X	X		X					X					Contraintes stériques et liaisons hydrogènes	
Diamond Lattice ²⁴⁷ (1990)	X		X		X							X			Contraintes stériques et liaisons hydrogènes	
BUILDER v1 ²⁴⁸ (1992)		X	X		X	X						X			Contraintes stériques et sites d'interactions clés	
LEGEND ²⁴⁹ (1991)	X		X					X		X					Champs de force	
LUDI ^{232, 233} (1992)		X	X			X				X	X				Fonction de score empirique	
NEWLEAD ²⁵⁰ (1993)	X	X	X			X					X				Contraintes stériques	
SPLICE ²⁵¹ (1993)		X	X		X						X				Pharmacophore et contraintes stériques	
GenStar ²⁵² (1993)	X		X		X					X					Contraintes stériques	
GroupBuild ²⁵³ (1993)		X	X		X					X					Champs de force	
CONCEPTS ²⁵⁴ (1993)	X		X					X							Fonction de score empirique	
SPROUT ²⁵⁵ (1993)		X	X		X	X							X		Surface accessible au solvant, liaisons hydrogène, interactions électrostatiques et hydrophobes	
MCSS&HOOK ^{256, 257} (1994)		X	X			X					X				Potentiel de van der Waals simplifié des interactions non polaires	
GrowMol ²⁵⁸ (1994)	X	X	X					X			X				Fonction de score empirique	
MCDNLG ²⁵⁹ (1995)	X		X					X		X				X	Energie potentielle	
Chemical Genesis ²⁶⁰ (1995)		X	X	X					X					X	Contraintes de forme, basé sur la grille et scalaire	
DLD ²⁶¹ (1995)	X		X					X						X	Fonction énergie potentielle sans interactions électrostatiques	
PRO_LIGAND ²⁶² (1995)		X	X	X	X					X	X				Fonction de score empirique	
SMoG ^{263, 264}		X						X		X					Fonction de score basée sur la connaissance	
BUILDER v2 ²⁶⁵ (1995)	X		X			X						X			Contraintes stériques	
CONCERTS ²⁶⁶ (1996)		X	X					X					X		Champ de force	
RASSE ²⁶⁷ (1996)	X		X			X				X					Champ de force et règles chimiques	
PRO_SELECT ²⁶⁸ (1997)		X	X			X				X					Fonction de score empirique	
SkelGen ^{269, 270} (1997)		X	X	X				X						X	Contraintes géométriques, de connectivité et chimiques	
Nachbar ²⁷¹ (1998)		X		X					X					X	QSAR (descripteurs topologiques de connectivité)	
Globus ²⁷² (1999)		X		X					X					X	Similarité moléculaire basée sur descripteur 2D	
DycoBlock ²⁷³		X	X		X								X		Champs de force et surface accessible au solvant	

(1999)										
LEA ²⁷⁴ (2000)		X	X				X		X	Modèle QSAR basé sur des descripteurs 3D
LigBuilder ²⁷⁵ (2000)		X	X				X	X	X	Fonction de score empirique
TOPAS ²⁷⁶ (2000)		X		X			X			Similarité basée sur un pharmacophore et les sous-structures
F-DycoBlock ²⁷⁷ (2001)		X	X		X				X	Champs de force et surface accessible au solvant
ADAPT ²⁷⁸ (2001)		X	X				X			DOCK score, clogP, poids moléculaire, nombre de liaisons rotatives, liaisons hydrogènes
Pellegrini & Field ²⁷⁹ (2003)		X	X	X		X				Modèle QSAR
SYNOPSIS ²⁸⁰ (2003)		X	X				X			Exemples: moment dipole électrique, score HIVRT empirique
CoG ²⁸¹ (2004)	X	X		X			X			Similarité moléculaire basée sur descripteur 2D
BREED ²⁸² (2004)		X		X	X					Pas de fonction de score interne
LEA3D ²⁸³ (2005)		X	X				X			Score docking FlexX avec différentes propriétés d'évaluation
Nikitin ²⁸⁴ (2005)		X	X		X				X	Liaisons hydrogène, interactions électrostatiques basées sur une grille
LCT ²⁸⁵		X	X						X	Fonction de score basée sur les champs de force
FlexNoVo ²⁸⁶ (2006)		X	X		X				X	Score docking FlexX
BOMB ²⁸⁷ (2006)		X	X		X				X	Fonction de score basée sur les champs de force
GANDI ²⁸⁸ (2008)		X	X				X			Fonction de score basée sur les champs de force
MED-Hybridise ²⁸⁹ (2009)		X	X		X				X	Score MED-SuMo
E-Novo ²⁹⁰ (2009)		X	X		X				X	Fonction de score basée sur la physique (CHARMm)
Fragment Shuffling ²⁹¹ (2009)		X	X		X				X	Fonction de score basée sur les champs de force
Hecht et Fogel ²⁹² (2009)		X	X			X			X	Fonction GOLD
AutoGrow ²⁹³ (2009)		X	X			X			X	Score AutoDock
MEGA ²⁹⁴ (2009)	X	X	X				X			MOFit (MultiObjective Fitness)
PhDD ²⁹⁵ (2010)		X	X		X				X	Score d'accessibilité synthétique
Contour ²⁹⁶ (2012)		X	X		X				X	Fonction de score empirique

Tableau 11. Caractéristiques des principaux logiciels, classés par ordre chronologique et utilisés pour la conception de novo (DFS : Depth-First-Search, BFS : Breadth-First-Search, Rnd : Random, MC : Monte-Carlo, AE : Algorithme Evolutionnaire ; Gr : Growth, Lk : Link, Lat : Lattice, DM : dynamique moléculaire, Sto : Stochastique^{297, 298})

4.4.1 Identification des sites d'interactions dans le site de liaison

La première étape lors d'études de conception *de novo* consiste à extraire du site de liaison toutes les informations concernant les interactions ligand-récepteur en déterminant les sites d'interaction. Les interactions étudiées sont de type liaisons hydrogène, électrostatiques et hydrophobes.²⁹⁷ Différentes méthodes peuvent être employées pour déterminer ces sites d'interactions. Les méthodes basées sur les règles (HSITE²⁹⁹, HIPPO²⁵⁵) s'intéressent principalement aux liaisons hydrogènes. D'autres méthodes, basées sur des grilles (GRID²¹⁹, LigBuilder²⁷⁵), déterminent les sites par des calculs d'interactions entre des sondes et des atomes ou des fragments du récepteur. Ces deux types de méthodes permettent seulement l'obtention de cartes du site de liaison mettant en évidence les zones d'interaction favorables et constituent une étape préalable à la conception *de novo*. La méthode MCSS (Multiple Copy Simultaneous Search)²⁵⁷ intègre quant à elle cette étape d'identification des sites d'interaction au placement de fragments dans le site de liaison.

4.4.2 Assemblage des blocs de construction

Les blocs utilisés pour la conception *de novo* peuvent être soit des atomes uniques, soit des fragments. Chacune de ces deux approches possède ses avantages. L'approche basée sur les atomes permet d'obtenir plus de diversité structurale que celle basée sur les fragments. Cependant, le nombre de solutions possibles est aussi beaucoup plus important et il peut être très compliqué d'extraire de cette masse de composés ceux les plus prometteurs. L'approche basée sur les fragments permet donc de réduire la taille de l'espace chimique à étudier, et ce de manière rationnelle lorsque les fragments sont correctement choisis. En conséquence, même si les approches basées sur les atomes ont été beaucoup utilisées lors des premières études de conception *de novo*, les approches basées sur les fragments se sont depuis généralisées.²⁹⁷

Les fragments ainsi obtenus sont ensuite assemblés pour créer de nouveaux composés. Cet assemblage peut se faire selon différents procédés : par liaison (ou « link »), par croissance (ou « growth »), en utilisant des treillis (ou « lattice »), par des mutations aléatoires (« random »), par des transitions guidées par dynamique moléculaire, et par des approches stochastiques (Tableau 11).

La construction des ligands par **liaison** (Figure 49) consiste à placer des fragments au niveau des sites d'interaction préalablement identifiés et de les lier les uns aux autres pour obtenir une molécule capable d'établir toutes les interactions nécessaires. La nature de la liaison entre

les fragments est choisie de manière à participer aux interactions favorables ligand-récepteur.

297

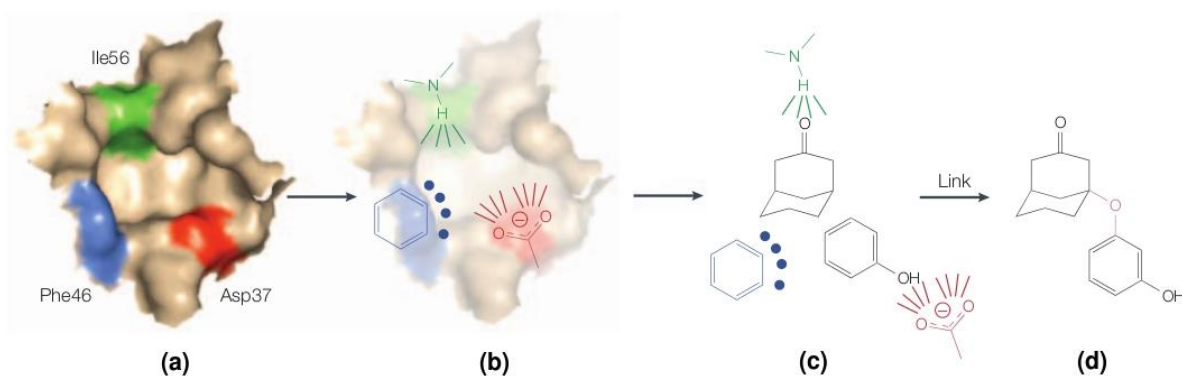


Figure 49. Illustration de la conception de novo de ligands par construction incrémentale par liaison après identification du site actif (a) et des sites d'interaction (b). Les différents fragments sont placés dans le site actif (c) puis lier les uns aux autres pour obtenir la molécule finale (d)²⁹⁷

Lors de la construction des ligands par **croissance** (Figure 50), un premier fragment est placé au niveau d'un site d'interaction. Les autres fragments sont ensuite ajoutés les uns après les autres de manière à maximiser les interactions avec les sites prédéfinis ainsi qu'avec les autres zones du site de liaison.

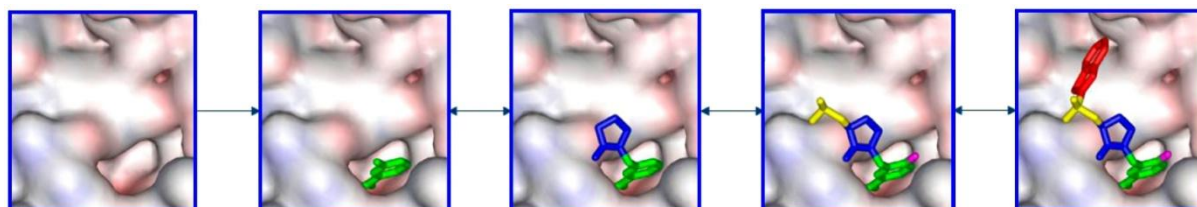


Figure 50. Conception de novo de ligands par croissance (d'après²⁹⁶)

L'approche utilisant des **treillis de points** (Figure 51) consiste à placer dans le site de liaison des atomes distribués régulièrement ou au hasard (carbones sp^3 ²⁴⁷, atomes choisis au hasard²⁶⁵, ou fragments pré-dockés²⁴⁸). Les atomes du treillis se trouvant dans les zones des sites d'interactions sont ensuite reliés par le plus court chemin passant par des points du treillis. Les atomes qui font partie de ce chemin sont ensuite reliés par des liaisons chimiques.²⁹⁷

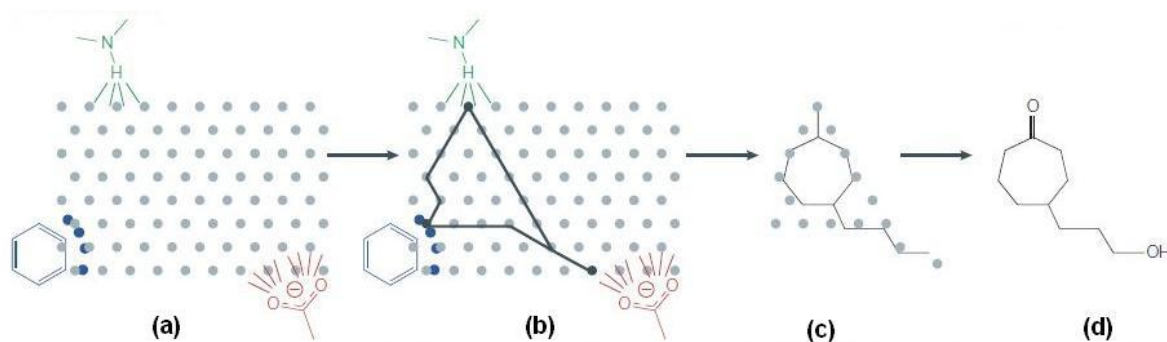


Figure 51. Conception de ligands à l'aide d'un treillis de points qui remplit le site de liaison (a). Les sites d'interaction sont ensuite reliés en suivant le plus court chemin passant par des points de la grille (b). Le chemin ainsi obtenu est ensuite transformé en un squelette moléculaire adéquat (c) puis en une molécule finale (d).²⁹⁷

Enfin, certaines méthodes de conception *de novo* utilisent des simulations de **dynamique moléculaire** pour construire les ligands. Dans cette approche, les fragments sont placés aléatoirement dans le site actif et des liaisons covalentes sont créées entre les différents fragments de manière stochastique et réversible pour pouvoir évoluer en fonction des simulations de dynamique moléculaire. A chaque étape, un fragment est choisi au hasard et toutes ses liaisons sont clivées (certaines méthodes utilisent comme point de départ plus d'un fragment) et de nouvelles liaisons sont formées pour relier les blocs situés à proximité du fragment.²⁹⁷

4.4.3 Recherche combinatoire

L'un des principaux problèmes des méthodes de conception *de novo* est de gérer l'explosion combinatoire des ligands. Pour cela, plusieurs approches peuvent être utilisées. La première appelée « depth-first search » (ou « recherche en profondeur d'abord ») ne conserve qu'une seule solution parmi toutes celles possibles à chaque niveau du processus de construction des ligands (Figure 52). A l'inverse, l'approche « breadth-first search » (ou « recherche en largeur d'abord ») permet à chaque niveau à toutes les solutions possibles de passer au niveau suivant jusqu'à la fin de la construction (Figure 52). Cependant, à chaque étape, toutes les solutions sont examinées pour identifier la plus optimale.

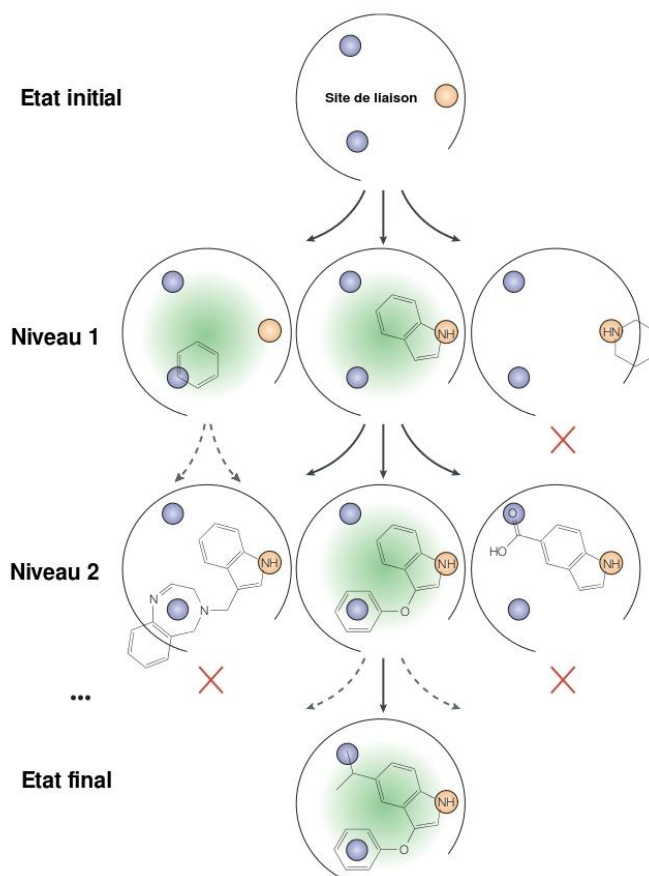


Figure 52. Modèle d'exploration de l'espace de recherche dans les méthodes de conception de novo. L'approche « depth-first search » ne conserve qu'une seule solution à chaque niveau (ici, le chemin du milieu) alors que l'approche « breadth-first search » conserve toutes les solutions plausibles à chaque niveau (ici, les chemins de gauche et du milieu pour le niveau 1, le chemin du milieu pour le niveau 2 etc...) (d'après ²⁹⁷)

Certaines méthodes emploient l'approche de Monte Carlo éventuellement combinée au critère de Metropolis. A chaque modification aléatoire apportée à la structure du ligand, une évaluation de la nouvelle structure obtenue est réalisée pour accepter ou rejeter ce changement. Si la modification permet d'obtenir un meilleur composé, elle est immédiatement acceptée. Dans le cas contraire, elle peut quand même être acceptée selon le critère de Metropolis.

Enfin certains programmes utilisent des algorithmes évolutionnaires. C'est notamment le cas du logiciel LEA ²⁷⁴ et de sa nouvelle version LEA3D ²⁸³ qui considère les codes SMILES des ligands en tant que chromosomes sur lesquels les opérateurs génétiques (reproduction, mutation et recombinaison) sont appliqués (Figure 53).

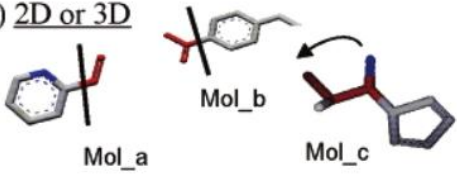
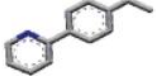

Molecular representation	Genetic operators (crossover, mutation)	
a) <u>2D or 3D</u> 	Result of the crossover operation between Mol_a and Mol_b 	Result of the 'cyclisation' mutation operation on Mol_c 
b) <u>SMILES line notation:</u> Mol_a <chem>N1=CC=CC=C1C=O</chem> Mol_b <chem>C1(C([O-])[O-])=CC=C(CC)C=C1</chem> Mol_c <chem>C1=CC=CC1C(N)C(C)C</chem>	<chem>N1=CC=CC=C1C1=CC=C(CC)C=C1</chem>	<chem>C1=CC=CC1C1=NC=CC=C1C</chem>
c) <u>Fragment-based notation:</u> Mol_a 106-2 Mol_b 37-40 Mol_c 56	106-40	Mutation by 'cyclisation' is prohibited

Figure 53. Le logiciel LEA3D utilise la représentation 3D (a), SMILES (b) et basée sur les fragments (c). Les molécules Mol_a et Mol_b sont combinées par une opération de recombinaison ou (crossover) et la molécule Mol_c est modifiée par cyclisation à l'aide de l'opérateur de mutation²⁸³

4.4.4 Attribution de scores

L'utilisation de fonctions de score pour pouvoir distinguer les composés les plus prometteurs parmi tous ceux générés est primordiale. Cette étape d'attribution de score (ou « scoring ») a lieu à la fois lors de la construction incrémentale des molécules pour guider la conception et à la fin du processus.²⁹⁷ Dans les premières approches de conception *de novo*, le score était basé uniquement sur des contraintes stériques (BUILDER^{248, 265}, NEWLEAD²⁵⁰). Depuis, le calcul du score s'est complexifié et de nombreuses fonctions de score ont été développées, principalement divisées en trois types : les fonctions de score basées sur les champs de force, les fonctions de score empiriques et les fonctions de score basées sur la connaissance (Tableau 11). Ces différents types de fonctions de score seront détaillés ultérieurement (voir paragraphe 4.5.2.2).

4.5 Méthodes de docking

L'utilisation des méthodes de docking dans le processus de conception de médicaments a débuté il y a plus de 30 ans²⁰⁶. Leur objectif est de prévoir la capacité ou non d'une molécule à se lier au site actif d'une protéine en se basant pour cela sur la prédiction de la conformation et de l'orientation de la molécule lors de sa liaison au récepteur³⁰⁰. A cet effet, les méthodes de docking combinent l'utilisation d'un algorithme de recherche, permettant de générer des modes de liaisons putatifs du ligand dans le récepteur, ou « poses », et d'une fonction de score, employée pour classer les différentes poses selon un score prédit d'affinité³⁰¹. Les méthodes de docking s'attachent donc, d'une part, à identifier les molécules qui sont des ligands véritables du récepteur parmi toutes celles étudiées, mais aussi d'autre part, à déterminer les poses correctes soit les conformations adoptées par les ligands lors de la liaison au récepteur.

4.5.1 Docking avec ligand rigide

Pendant longtemps, le mécanisme de liaison d'un ligand à son récepteur a été envisagé comme un processus statique dans lequel le ligand constituait une clé de forme complémentaire à celle de la serrure qu'il était capable « d'ouvrir », le récepteur (modèle « clé-serrure » ou « lock-and-key » model)³⁰². Pour tenter de reproduire ce modèle, les premiers logiciels de docking, parmi lesquels le logiciel DOCK²⁰⁶ considéraient donc le ligand et le site de liaison comme deux entités rigides. C'est ce qu'on appelle le docking avec ligand rigide. Dans cette approche, le positionnement des ligands dans le site de liaison se fait par translation et rotation. C'est notamment le cas du logiciel FRED³⁰³ dont la première étape consiste à énumérer toutes les rotations et translations possibles pour un ligand à l'intérieur du site de liaison (Figure 54). Ensuite une image négative du site de liaison est utilisée pour éliminer toutes les poses incompatibles avec le site actif (« clash », distance). Enfin, les poses sélectionnées se voient attribuées un score et les meilleures sont optimisées.

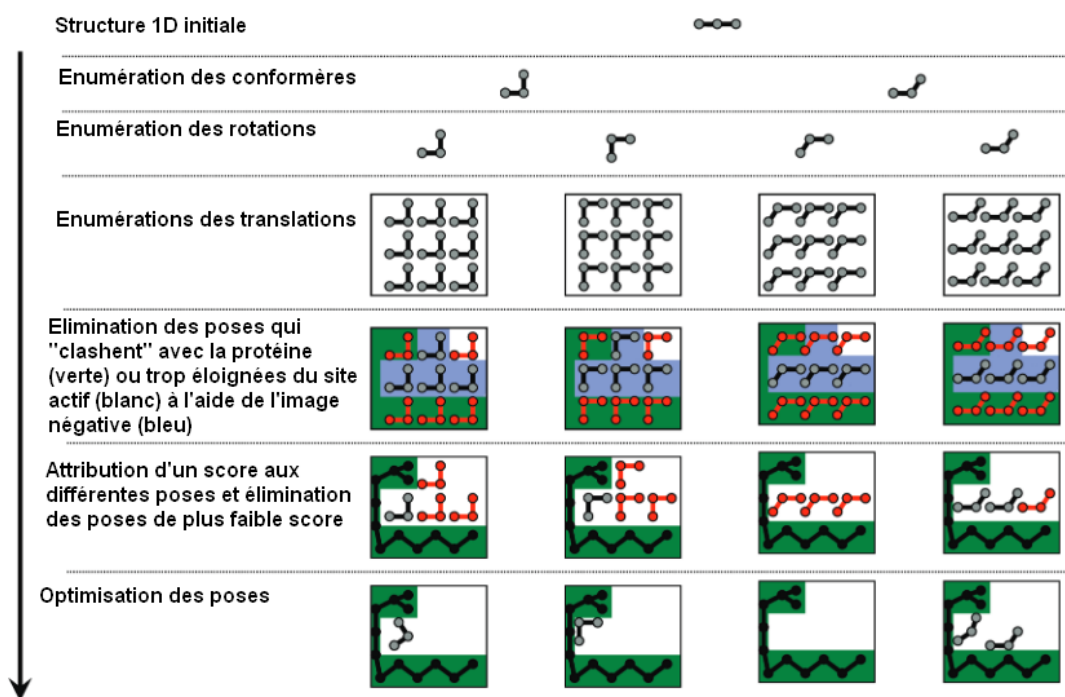


Figure 54. Schématisation du protocole de docking ligand rigide à l'aide du logiciel FRED (d'après³⁰³)

L'intérêt des méthodes de docking ligand rigide réside dans leur rapidité et elles peuvent donc être employées lors des criblages virtuels comme un premier filtre permettant de ne pas retenir des molécules aberrantes (trop grandes, mauvaise complémentarité avec le site de liaison, ...).³⁰⁴

4.5.2 Docking avec ligand flexible

Comme souligné précédemment (2.2.3.2), la liaison du ligand à un site de liaison se fait en réalité dans la majorité des cas par un processus de sélection de conformation qui se termine par une étape d'adaptation du ligand et du récepteur l'un à l'autre⁷³. Il est donc important pour les méthodes de docking de considérer cette flexibilité du ligand et du récepteur. La flexibilité du ligand, permettant d'explorer les conformations adoptées par celui-ci lors de la fixation au site de liaison, est maintenant classiquement prise en compte dans la plupart des études de docking. A l'inverse, la prise en compte de la flexibilité du récepteur, proposée par de plus en plus de logiciels de docking^{305, 306} est encore trop coûteuse en temps de calcul pour être systématiquement envisagée dans les protocoles de docking.

4.5.2.1 Algorithmes de recherche

Les algorithmes de recherche permettant de traiter la flexibilité du ligand peuvent être classés en trois grandes catégories : les algorithmes de recherche systématique, les algorithmes de recherche aléatoire ou stochastique et les algorithmes de recherche déterministe ou de simulation.³⁰⁷

4.5.2.1.1 Recherche systématique

Les algorithmes de recherche systématique ont pour objectif d'explorer tous les degrés de liberté des ligands par rotation de 0 à 360° de toutes les liaisons rotatives à l'aide d'un pas incrémental choisi. En conséquence, le nombre de conformations ainsi générées peut être très grand (Équation 13). C'est ce qu'on appelle l'explosion combinatoire.

$$N_{Conformations} = \prod_{i=1}^N \prod_{j=1}^{n_{inc}} \frac{360}{\theta_{i,j}}$$

Équation 13. Calcul du nombre de conformations possibles ($N_{conformations}$) pour un ligand avec N le nombre de liaisons rotatives, n_{inc} le nombre d'incrémentations et $\theta_{i,j}$ la valeur de l'angle incrémental rotationnel j pour la liaison i ³⁰⁰

Deux types de méthodes peuvent être utilisés : des méthodes de recherche exhaustive et des méthodes de reconstruction incrémentale.

Les **méthodes de recherche exhaustive** permettent de mener un docking avec ligand flexible par des rotations systématiques de toutes les liaisons rotatives possibles du ligand à un intervalle donné. L'explosion combinatoire sous-jacente est limitée par définition de contraintes géométriques et chimiques à imposer aux conformations initiales des ligands à docker et les conformations ainsi sélectionnées sont finement affinées et optimisées. La première étape du protocole de filtres hiérarchiques de GLIDE³⁰⁸ emploie un algorithme de recherche systématique exhaustive.³⁰⁹

Les **méthodes de reconstruction incrémentale** quant à elles débutent par la fragmentation des ligands qui sont par la suite reconstruit de façon incrémentale dans le site actif. En effet, à partir d'une molécule avec 7 liaisons rotatives, en considérant 6 rotamères pour chaque liaison, il existe 6^7 , soit 279936, conformations possibles. Si une liaison rotative de cette molécule est rompue pour obtenir deux fragments, cela élimine une liaison pour la recherche conformationnelle et les conformations des deux fragments sont indépendantes, ce qui réduit le nombre de conformations à $6^3 + 6^3$ soit 432 conformations.³¹⁰ Ces approches sont

généralement divisées en 3 étapes : (1) identification de fragments rigides et de fragments flexibles, (2) les fragments rigides sont dockés dans le site actif et (3), les fragments rigides sont ajoutés incrémentalement.^{311, 312}. De nombreux logiciels utilisent des algorithmes de fragmentation reconstruction, tels que par exemple, DOCK 6³¹³, FlexX³¹², ou encore Surflex-Dock³¹⁰. Ce dernier utilise un algorithme de Hammerhead³¹⁴ modifié pour réaliser le docking flexible des ligands dans le site de liaison. La première étape du protocole de docking du logiciel Surflex-Dock consiste à générer un ligand idéal pour le site de liaison étudié, communément appelé « protomol ». Pour cela, trois types de fragments (CH₄, C=O et N-H) sont placés dans de multiples positions dans le site actif et optimisés pour permettre l'établissement d'interactions avec le site de liaison. Les fragments les mieux scorés sont assemblés pour former le protomol qui couvre l'intégralité du site actif. (Figure 55).

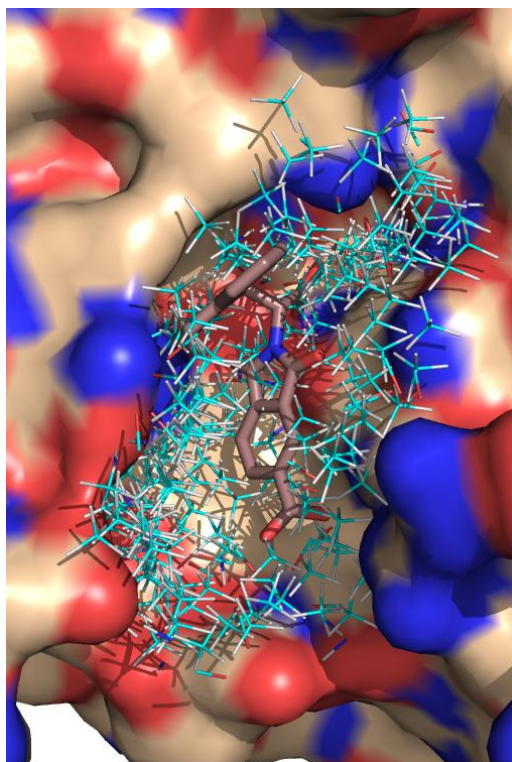


Figure 55. Protomol généré pour la beta lactamase AMPc par extension de 4 Å autour du ligand (code PDB: 2R9W)

Le docking en lui-même débute par la fragmentation des ligands en 1 à 10 parties par rupture de liaisons non aromatiques. Pour chaque fragment, une recherche conformationnelle est menée. Chaque conformation de chaque fragment est ensuite alignée au protomol par une fonction de similarité morphologique³¹⁵. Les fragments ayant obtenu les meilleurs scores sont utilisés comme des têtes (ou « heads ») et sont localement optimisés pour s'adapter au mieux au site de liaison. Ensuite, un autre fragment, appelé queue (ou « tail ») est ajouté par

alignement de toutes les conformations de ce fragment sur le protomol tout en respectant les contraintes de distance imposées par le fragment tête pour permettre la connexion entre les deux. L'attribution de scores permet ensuite de distinguer les poses entre elles.

4.5.2.1.2 Recherche aléatoire ou stochastique

Les algorithmes de recherche aléatoire ou algorithmes stochastiques procèdent à des changements aléatoires en termes de translations, rotations et torsions permettant de générer de nouvelles conformations de ligands. Les changements sont acceptés ou rejetés à l'aide d'une fonction de probabilité³⁰⁰. Quatre grands types d'algorithmes appartiennent à cette classe : les méthodes de Monte-Carlo, les algorithmes génétiques, les algorithmes de recherche tabou et les algorithmes d'optimisation en essaim³⁰⁹.

Les **méthodes de Monte-Carlo**³¹⁶ débutent par la génération aléatoire d'une conformation initiale d'un ligand. Toujours de manière aléatoire, un changement conformationnel est opéré, que ce soit par une translation, une rotation ou une torsion. L'impact de ce changement sur l'énergie du ligand est évalué et une nouvelle conformation est retenue si elle présente une énergie plus basse que la précédente ou dans le cas contraire selon une fonction de probabilité dérivée de Boltzmann, appelée critère de Metropolis (Équation 14).

$$P = e^{\frac{E_{new} - E_{old}}{kT}}$$

Équation 14. Probabilité P d'acceptation d'une nouvelle conformation selon le critère de Metropolis (E_{new} : énergie de la nouvelle conformation, E_{old} : énergie de l'ancienne conformation, k : constante de Boltzmann et T : température de simulation)

A partir de la nouvelle conformation ou de l'ancienne si la nouvelle a été rejetée, un changement aléatoire est à nouveau réalisé. Ces étapes sont ainsi répétées un certain nombre de fois, défini par le nombre de pas de Monte-Carlo. ICM (Internal Coordinate Mechanics)³¹⁷ est un exemple de logiciel utilisant des simulations de Monte Carlo pour optimiser les coordonnées spatiales d'un ligand. Le protocole de docking débute par la génération de différentes conformations des ligands à l'extérieur du site de liaison et la sélection de celles de plus basses énergies, utilisées comme point de départ pour le docking. Le site de liaison est représenté par des cartes de grilles de potentiels de liaisons hydrogènes, de van der Waals, hydrophobe et électrostatiques pour réduire les temps de calculs. Ces cartes sont générées dans une boîte rectangulaire centrée sur le site de liaison. Un changement aléatoire (rotation, translation ou conformation) est imposé au ligand placé dans le site de liaison. Après minimisation, la conformation obtenue est soumise au critère de Metropolis pour décider de

son acceptation ou de son rejet. Les logiciels Glide³⁰⁸, QXP³¹⁸, PRODOCK³¹⁹ et MCDOCK³²⁰ utilisent eux aussi des méthodes de Monte-Carlo lors de la phase de recherche de poses.

Les **algorithmes génétiques** miment l'évolution biologique telle que décrite par la théorie de l'évolution de Darwin³²¹ à l'aide d'opérateurs génétiques que sont la sélection, la mutation et la recombinaison (ou « crossover »). Pour cela, plusieurs conformations initiales d'un ligand sont générées. Les combinaisons d'un ensemble de variables, par exemple angles de torsion, donneurs de liaisons hydrogène, accepteurs de liaisons hydrogène pour le logiciel GOLD³²², sont assimilées à des chromosomes. Une fonction de score appelée fonction de survie (ou « fitness function ») permet de déterminer dans la population initiale les meilleures conformations qui constituent donc les chromosomes « parents ». Par application de mutation ou de recombinaison aux chromosomes, des changements respectivement ponctuels ou par échange de groupes de variables sont appliqués et de nouvelles conformations « enfants » sont obtenues (Figure 56). La fonction de survie permet de sélectionner les meilleures d'entre elles qui constituent une nouvelle population « parents » auxquelles seront appliqués les opérateurs génétiques, jusqu'à l'obtention d'une population « enfants » satisfaisant la fonction de survie choisie³⁰⁵. Ce type d'algorithme dirige le processus de docking des logiciels GOLD³²² et AutoDock³²³ notamment.

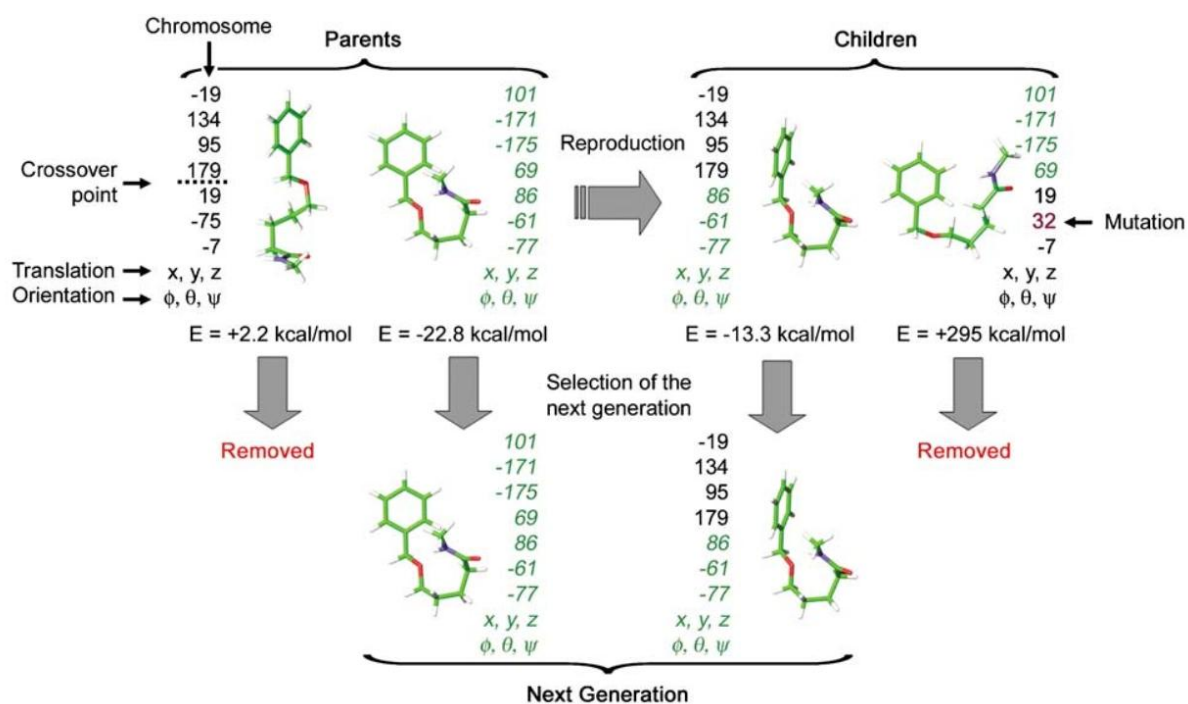


Figure 56. Schématisation du processus de docking à l'aide d'un algorithme génétique²³

Les **algorithmes de recherche tabou** ³²⁴ débutent par la détermination d'une position et d'une orientation aléatoire du ligand dans le site de liaison. A partir de cette conformation initiale, un nombre défini de changements vont être réalisés, par une procédure mimant des mutations par application de nombres aléatoires aux variables du docking. Ces changements sont ensuite évalués et triés à l'aide d'une fonction de score. Si le changement permet d'obtenir une énergie plus basse, il est immédiatement accepté et intègre la « liste tabou » regroupant les 25 meilleures solutions. Dans le cas contraire, la conformation est comparée aux conformations précédemment acceptées et si la valeur de moyenne quadratique (ou Root Mean Square RMS) obtenue est supérieure à un seuil défini (par exemple 0,75 pour le logiciel PRO_LEADS ³²⁴), le changement est accepté et intégré à la « liste tabou ». Dans le cas contraire, le changement est considéré comme tabou et éliminé. La recherche se termine lorsque plus aucun changement acceptable n'est obtenu ou lorsque la limite d'itération est atteinte. Les algorithmes de recherche tabou sont exploités dans les logiciels PRO_LEADS ³²⁴ et PSI-DOCK ³²⁵ notamment.

Les **algorithmes d'optimisation par essaims particuliers** ou (Swarm Optimization algorithm) considèrent un ensemble de conformations d'un ligand. Chaque conformation est appelée une particule et l'ensemble, un essaim. Chaque particule subit des changements et retient en mémoire la meilleure position qu'elle a adoptée au cours du docking. Les mouvements du ligand à travers l'espace conformationnel sont ainsi guidés par les meilleures positions adoptées par ses voisins ³⁰⁹. Parmi les logiciels de docking utilisant un algorithme d'optimisation en essaim, il est possible de citer SODOCK ³²⁶, Tribe-PSO ³²⁷, pso@autodock ³²⁸ et FIPSDock ³²⁹.

4.5.2.1.3 Recherche déterministe ou de simulation

Les méthodes de recherche déterministe sont divisées en deux grandes classes : les méthodes de dynamique moléculaire et les méthodes de minimisation.

Les **méthodes de dynamique moléculaire** tentent de résoudre l'équation du mouvement d'un système atomique de Newton (Équation 15).

$$F_i = m_i a_i$$

Équation 15. Equation du mouvement de Newton pour un système atomique (F: force, m: masse, a : accélération)

Pour cela, la force de chaque atome est calculée à partir d'un changement dans l'énergie potentielle entre une nouvelle position et l'actuelle (Équation 16).

$$F_i = -\frac{dE}{r_i}$$

Équation 16. Calcul de la force (F) de chaque atome par changement dans l'énergie potentielle (dE) entre deux positions de distance r_i

Les positions atomiques de chaque atome sont ensuite calculées à très courts intervalles de temps (Équation 17) et permettent d'obtenir la trajectoire des changements de positions atomiques au cours du temps.

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

Équation 17. Calcul des positions atomiques à chaque intervalle de temps (d^2r_i/dt^2) en fonction de la force F et de la masse atomique m

Malheureusement, avec des temps de calculs raisonnables, les méthodes de dynamique moléculaire ne sont généralement pas capables de franchir les barrières de hautes énergies et les ligands sont donc obtenus dans des conformations correspondant à des minimums locaux d'énergie³⁰⁷. Cependant, différentes stratégies ont été développées pour pallier à ce problème. La première consiste à simuler différentes parties du complexe récepteur/ligand à différentes températures³³⁰. Une autre stratégie réalise plusieurs calculs de dynamique moléculaire à partir de différentes conformations initiales³⁰⁰. Enfin, d'autres approches manipulent la surface d'énergie potentielle pour la rendre plus lisse³³¹. Malgré cela, les temps de simulation restent encore trop élevés pour pouvoir utiliser ces méthodes à large échelle lors d'un criblage virtuel³⁰⁷. Cependant, les approches de dynamique moléculaire peuvent être employées sur un faible nombre de molécules à la fin d'un criblage virtuel pour optimiser la sélection des meilleurs composés.

Les **méthodes de minimisation d'énergie** ne permettent d'atteindre que des minimums locaux d'énergie. Elles ne sont donc pas utilisées seules mais plutôt en complément d'autres méthodes de recherche.

4.5.2.2 Scoring

Les fonctions de score sont utilisées pour estimer mathématiquement l'affinité de liaison entre un récepteur et chacune des poses générées pendant le docking³³². L'efficacité de ces fonctions de score est au moins tout aussi importante que celle des algorithmes de recherche conformationnelle. En effet, même si la conformation bioactive du ligand a été obtenue lors

docking, si les fonctions de score ne permettent pas de différencier les poses correctes de celles incorrectes, les composés les plus prometteurs pour la cible ne pourront pas être identifiés.³⁰⁰

4.5.2.2.1 Aspects théoriques de la liaison d'un ligand dans un site actif

En solution, la liaison d'un composé au site de liaison est dirigée par des effets enthalpiques et entropiques affectant à la fois le ligand, la cible et les molécules de solvant. L'évaluation de l'enthalpie libre de liaison ΔG_{bind} (ou « free energy of binding ») constitue un moyen efficace de déterminer les meilleures poses parmi les résultats de docking³³³. Cette enthalpie de liaison libre mesure la stabilité d'un complexe en évaluant sa constante d'équilibre K_{eq} (Équation 18)

$$\Delta G_{bind} = -RT \ln K_{eq} = -RT \ln \frac{k_a}{k_d}$$

Équation 18. L'enthalpie libre de liaison (ΔG_{bind}) peut être exprimée en fonction de la constante des gaz parfaits (R), de la température (T) et de la constante d'équilibre du complexe ligand-récepteur, égale au ratio constante d'association (k_a) constante de dissociation (k_d)

A pression et température constante, l'enthalpie libre de liaison peut aussi être calculée par les variations d'enthalpie et d'entropie du système.

$$\Delta G_{bind} = \Delta H - T\Delta S$$

Équation 19. A température (T) et pression constante, l'enthalpie libre de liaison (ΔG_{bind}) peut être reliée aux variations d'enthalpie (ΔH) et d'entropie (ΔS)

Cependant, l'évaluation directe de l'enthalpie libre de liaison nécessite des temps de calcul incompatibles avec le criblage d'un grand nombre de composés. Les fonctions de score utilisées lors des études de docking utilisent donc des approximations et de simplifications lors de l'évaluation des complexes récepteur / ligand. Certains phénomènes physiques guidant la liaison du composé au site actif ne sont donc que partiellement pris en compte.³⁰⁰

4.5.2.2.2 Paramètres utilisés dans les fonctions de score

Les critères qui dirigent la liaison d'un ligand à son récepteur sont d'une part la complémentarité de forme ³⁰², et d'autre part l'établissement d'interactions entre le ligand et le site de liaison, que ce soit des interactions hydrogènes, électrostatiques et de van der Waals ³³⁴ ou encore des interactions hydrophobes ³³⁵. Les interactions hydrogènes et électrostatiques constituent des interactions spécifiques entre ligands et sites de liaison et permettent de contrebalancer partiellement le coût énergétique de désolvatation des parties polaires et chargées de ces deux partenaires. Les interactions hydrophobes constituent quant à elles le principal facteur de stabilisation des complexes ligands récepteur et résultent de la désolvatation des parties apolaires engendrée par la liaison du ligand. Ces critères sont mesurés directement ou indirectement dans les fonctions de score par différents paramètres que sont la complémentarité géométrique, le chevauchement inter- et intra-moléculaire, les liaisons hydrogènes, les aires de contact, les contacts de paires d'acides aminés ou de paires d'atomes, les interactions électrostatiques et l'énergie de solvatation ³³⁶. Ces paramètres peuvent être utilisés de différentes manières. Ainsi leur contribution au score peut être positive ou négative formant ainsi ce qu'on appelle une pénalité. Ils peuvent aussi être utilisés comme des filtres d'exclusion. Par exemple les paramètres géométriques servent usuellement de filtres d'exclusion primaires étant donné leur rapidité de calcul, et les paramètres d'énergie sont ensuite appliqués aux solutions restantes ³³⁶.

4.5.2.2.3 Fonctions de score

Classiquement, trois types de fonctions de score sont distingués : les fonctions de scores basées sur les champs de force, les fonctions de scores empiriques, et les fonctions de scores basées sur les connaissances. ³⁰⁰

Les **fonctions de score basées sur les champs de force** déterminent la somme de l'énergie d'interaction ligand-récepteur, de l'énergie interne du ligand et de l'énergie interne de la protéine. Cependant, dans la plupart des fonctions de score basées sur les champs de force implémentées dans les logiciels de docking, une seule conformation de la protéine est prise en compte, ce qui permet de négliger l'énergie libre de la protéine et ainsi de simplifier les calculs de score. Les champs de force de mécanique moléculaire permettent de calculer l'énergie potentielle du système E_{MM} et sont généralement exprimés sous la forme d'une somme de quatre termes (Équation 20). Les trois premiers représentent les termes liés par des pénalités en cas d'écart par rapport aux valeurs de références de longueurs et d'angles de

liaison. Le dernier terme quantifie la contribution des termes non liés par le potentiel de Lennard-Jones pour les interactions de Van der Waals et le potentiel de Coulomb pour les interactions électrostatiques.

$$E_{MM} = \sum_{bond} K_r (r - r_{ref})^2 + \sum_{angles} K_\vartheta (\vartheta - \vartheta_{ref})^2 + \sum_{dihedrals} K_\Phi [1 + \cos(n\Phi - \Phi_0)]^2 + \sum_{i < j} \left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^6} + \frac{q_i q_j}{\epsilon R_{ij}} \right]$$

Équation 20. Calcul de l'énergie potentielle d'un système (E_{MM}) par les champs de force de la mécanique moléculaire (K_r , K_ϑ et K_Φ : facteurs de pénalité pour les liaisons, les angles et les angles dièdres ; r et r_{ref} : longueurs des liaisons mesurées et de référence ; ϑ et ϑ_{ref} : valeurs des angles mesurées et de référence ; Φ et Φ_{ref} : valeurs des angles dièdres mesurées et de référence ; A_{ij} et B_{ij} : constantes attractives et répulsives mesurées expérimentalement ; R_{ij} : distance entre les atomes i et j ; q_i et q_j : charges des atomes i et j ; ϵ : constante diélectrique)

Les fonctions de score basées sur les champs de force se limitent généralement aux interactions stériques et électrostatiques (Équation 21).

$$\Delta G_{bind} = \sum_{i=1}^{ligand} \sum_{j=1}^{protéine} \left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^6} + \frac{q_i q_j}{\epsilon R_{ij}} \right]$$

Équation 21. Evaluation de l'enthalpie libre de liaison ΔG_{bind} par les potentiels de Lennard-Jones et Coulomb

De nombreux logiciels de docking utilisent des fonctions de score basées sur les champs de force. Cependant, les valeurs des champs de force ne sont pas toujours les mêmes. Ainsi, les fonctions de score de DOCK 6³¹³, Autodock³²³ dérivent du champ de force AMBER^{337, 338}, la fonction Goldscore³³⁹ de GOLD³²² est basée sur le champ de force Tripos³²² et la fonction de score d'ICM, hybride « champs de force-empirique », utilise le champ de force ECEPP/3³⁴⁰.

Le défaut majeur des fonctions de score basées sur les champs de force standard réside donc dans l'absence des termes de solvation et d'entropie. La cause en est historique puisque ces fonctions de score ont initialement été développées pour quantifier les contributions enthalpiques en phase gazeuse sur les structures et les énergies.³⁰⁰ Pour tenter de pallier à ce problème, des améliorations des fonctions de score basées sur les champs de force ont été proposées. Ainsi la fonction de score hybride d'ICM (Équation 22) a ajouté un terme empirique d'entropie de torsion pour les ligands et différents termes de liaisons hydrogènes en fonction de la nature et de la géométrie de l'interaction pour le complexe protéine-ligand. Les

fonctions de score de GOLD³²² et d'Autodock³²³ présentent elles aussi des termes pour les liaisons hydrogène protéine-ligand.

$$\begin{aligned}
 E &= (E_{vdw} + E_{HB} + E_{torsion} + E_{el}) + E_{solv} + E_{bonds} + E_{angles} + \\
 &E_{phases} + E_{dist.restr.} + E_{tethers} + E_{var.restr.} \\
 &= \left(\frac{FA}{d^{12}} - \frac{C}{d^6} + \frac{A'}{d^{12}} - \frac{B}{d^{10}} + K(1 \pm \cos(n\varphi)) + 332 \frac{q_\alpha q_\beta}{\epsilon d} \right) + K_s \sum (AAS) + \\
 &0.5K_b(b - b_0)^2 + 0.5K_\omega(\omega - \omega_0)^2 + 0.5K_\Phi(\Phi - \Phi_0)^2 + 0.25W_p \frac{(d^2 - D_U^2)^2}{D_U^2} + \\
 &0.25W_p \frac{(d^2 - D_L^2)^2}{D_L^2} + W_\tau d^2 + \frac{U(\delta^2 - 1)^2(2\delta^2 - 3F'^2 + 1)}{(1 - F'^2)^3}
 \end{aligned}$$

Équation 22. Equation de la fonction de score d'ICM sommant des termes de van der Waals (E_{vdw}) de liaisons hydrogène (E_{HB}), de torsion ($E_{torsion}$), électrostatique (E_{el}) de solvation (E_{solv}), de déformation de liaisons (E_{bond}), de flexion d'angle (E_{angles}), de déformation de l'angle de phase (E_{phases}) de restrictions de distances ($E_{dist.restr.}$), d'attaches ($E_{tethers}$) et de pénalité de restrictions de variables ($E_{var.restr.}$) (avec $F=0,5$ pour les atomes séparés par 3 liaisons covalentes et $F=1$ dans tous les autres cas, A et C : constantes de van der Waals, d : distance entre les atomes, A' et B : constantes de liaisons hydrogènes, K , K_b , K_ω , K_Φ : constantes de force pour la torsion, les déformations de liaisons, les flexions d'angles et les déformations de l'angle de phase dièdre ; φ : angle de torsion, q_α et q_β : charges des atomes ; ϵ : constante diélectrique ; K_s : paramètres de solvation, AAS : surface atomique accessible au solvant ; b_0 , ω_0 et Φ_0 : valeurs de référence de distances de liaisons, d'angles et d'angles de phase dièdres ; W_p et W_τ : coefficients de restrictions de distance et d'attaches ; D_U et D_L : valeurs hautes et basses pour les restrictions de liaisons ; U : constante de restriction de variable ; δ : distance normalisée ; F' : F' : fraction de la dimension occupée par le fond plat de la fonction)³¹⁷

D'autre part, les modèles PB/SA (Poisson-Boltzman/Surface Area)³⁴¹ et GB/SA (Generalized Born/ Surface Area)³⁴² permettent de simuler le solvant de façon implicite mais les temps de calcul associés ne sont pas compatibles avec des études de docking à large échelle. Cependant, ces modèles peuvent être utilisés pour affiner la prédiction d'énergie de liaison ou le classement des ligands sur un petit nombre de composés, en combinaison avec une fonction de score ou après l'étape d'attribution de score³⁰⁶.

Le principe des **fonctions de score empiriques** est de tenter de reproduire des données expérimentales d'énergies de liaison en utilisant des termes individuels non corrélés pondérés par des coefficients adéquats calculés par des analyses de régression sur les données expérimentales d'affinité et de données structurales de cristallographie (contacts non covalents)³⁰⁰. L'intérêt de ses fonctions de score empiriques réside dans leur double simplicité : leur forme est généralement plus simple que celles basées sur les champs de force et les termes constituant ces fonctions sont aisément évaluables. Cependant, la combinaison de différentes fonctions de score empirique est peu praticable puisque les coefficients étant déterminés à partir de données expérimentales, généralement de jeux de données d'une centaine de complexes tout au plus. Selon les données utilisées, leurs valeurs peuvent donc varier.³⁰⁵ Cependant, l'augmentation considérable du nombre de structures de complexes ligands-récepteurs disponibles et des données d'affinité correspondantes, devrait permettre de concevoir une fonction de score empirique généraliste obtenue à partir d'un large nombre de complexes³⁴³. La nature et la forme des termes utilisés varient selon les fonctions de score considérées. Ces termes permettent notamment de quantifier la contribution des termes non liés comme les liaisons hydrogène (LUDI²³³, ChemScore³⁴⁴), les interactions hydrophobes (LUDI²³³, ChemScore³⁴⁴) et les interactions aromatiques (F-Score³¹²) mais aussi une approximation incomplète de contributions non enthalpiques comme l'entropie (ChemScore³⁴⁴) ou la solvation (fonction de score Fresno³⁴⁵). La fonction de score Surfex-score (Équation 23) du logiciel Surfex³¹⁰ appartient elle aussi à la classe des fonctions de score empiriques³⁴⁶.

$$\begin{aligned}
-\log(K_d) &= \text{steric_score} + \text{polar_score} + \text{polar_repulsion_score} + \text{entropy_score} \\
&= l_1 \exp \frac{-(r+n_1)^2}{n_2} + \frac{l_2}{1 + \exp^{n_3(r+n_4)}} + l_3 \max(0, r + n_5)^2 \\
&+ \left[l_4 \exp \frac{-(r+n_6)^2}{n_7} + \frac{l_5}{1 + \exp^{n_3(r+n_8)}} + l_3 \max(0, r + n_9)^2 \right] \left[\frac{1}{1 + \exp^{n_3(-(b_{ij} \cdot v_i)(b_{ij} \cdot v_j)) - n_{10}}} \right] [(1 + n_{11} c_i)(1 + n_{11} c_j)] \\
&+ \left[l_6 \exp \frac{-(r+n_{12})^2}{n_{13}} \right] \left[\frac{1}{1 + \exp^{n_3(-(b_{ij} \cdot v_i)(b_{ij} \cdot v_j)) - n_{10}}} \right] \\
&+ (l_7 \cdot n_{rot})(l_8 \log(\text{molweight}))
\end{aligned}$$

Équation 23. Equation de la fonction de score Surflex-dock tentant de prédire l'affinité de liaison $-\log(K_d)$ d'un complexe à l'aide de termes stérique, polaire et d'entropie (l_1 :facteur stérique gaussien d'attraction, l_2 :facteur stérique sigmoïde de répulsion, l_3 : facteur stérique de pénétration, l_4 :facteur polaire gaussien d'attraction, l_5 : facteur polaire sigmoïde de répulsion, l_6 : facteur polaire d'inadéquation, l_7 : facteur du nombre de liaisons rotatives, l_8 : facteur du poids moléculaire, n_1 : position gaussienne stérique, n_2 : propagation gaussienne stérique ; n_3 :raideur de la pente sigmoïde, n_4 : point d'inflexion sigmoïde stérique, n_5 : tolérance de van der Waals aux clashes stériques, n_6 :position polaire gaussienne, n_7 : propagation gaussienne stérique, n_8 : point d'inflexion sigmoïde polaire, n_9 : tolérance de van der Waals pour les clashes polaires, n_{10} :point d'inflexion de la sigmoïde de direction polaire, n_{11} : facteur de charge, n_{12} : position gaussienne de la répulsion polaire, n_{13} : propagation gaussienne de la répulsion polaire) ³⁴⁷

Les **fonctions de score basées sur les connaissances ou « knowledge-based »** reposent sur le même principe de reproduction de données expérimentales que les fonctions de score empirique. Cependant, là s'arrête la comparaison puisque les fonctions de score basées sur les connaissances ne s'intéressent pas à l'énergie de liaison mais à la structure expérimentale du complexe ligand-récepteur. L'hypothèse de base de ces fonctions est que dans un complexe cristallographique les atomes du ligand sont placés en position optimale par rapport aux atomes de la protéine. L'analyse statistique de complexes résolus avec succès permet de déterminer la fréquence et la distribution des paires d'atomes ligand/protéine et ainsi quantifier les potentiels d'interaction constituant la fonction de score. Les avantages et les inconvénients de ce type de fonction de score sont tout à fait similaires à ceux des fonctions empiriques. En effet, leur calcul est simple et ne nécessite pas de grandes ressources computationnelles ^{300, 305}. Bien qu'elles dépendent toujours des structures des complexes utilisés pour leur calibration, l'augmentation du nombre de complexes résolus devrait contribuer là encore à l'amélioration de ces fonctions ^{348, 349}. Une autre amélioration à apporter consiste à prendre en compte des données « négatives » sous forme de contacts

défavorables pour la liaison du ligand pour permettre d'éliminer plus facilement les poses incorrectes ³⁵⁰. Les fonctions de score PMF (Potential of Mean Force) ³⁵¹ (Équation 24), FlexX ³¹², DrugScore ³⁵² et SMOG ²⁶³ sont trois exemples de fonctions de score basées sur les connaissances.

$$PMF_score = \varepsilon \cdot \Delta G_{bind} = \sum_{\substack{kl \\ r < r_{cut-off}^{ij}}} A_{ij}(r) = -k_B T \ln \left[f_{Vol_corr}^j(r) \frac{\rho_{seg}^{ij}(r)}{\rho_{bulk}^{ij}} \right]$$

Équation 24. La fonction de score basée sur les connaissances PMF peut être reliée à l'enthalpie de liaison libre ΔG_{bind} à l'aide d'un facteur d'échelle (ε) représentant tous les termes traités implicitement dans le modèle. Elle est définie comme étant la somme de toutes les interactions de paires d'atomes du complexe protéine-ligand (avec $r_{cut-off}^{ij}$ la valeur seuil de distance pour les paires kl d'atomes de type ij ; r : distance entre paires d'atomes ; k_B : constante de Boltzmann, T : température absolue ; $f_{Vol_corr}^j(r)$: facteur de correction du volume du ligand ; $\rho_{seg}^{ij}(r)$: densité de paires d'atomes ij pour une certaine distance « seg » ; ρ_{bulk}^{ij} : distribution de i et j lorsqu'il n'y a pas d'interaction entre i et j) ³⁵¹

Le choix d'une fonction de score est critique, et actuellement aucune fonction de score existante n'est parfaite puisque toutes ces fonctions sont conçues à partir d'un certain nombre d'approximations. L'utilisation de **fonctions de score consensus** a été proposée pour tenter de résoudre ce problème. Ces fonctions permettent de combiner les informations de différentes fonctions de score pour compenser leurs imperfections individuelles et ainsi améliorer la qualité de la prédiction des poses correctes. ³⁰⁰. Il existe deux hypothèses à cela. La première est que les différentes fonctions de score permettent de prendre en compte différents aspects des interactions protéine-ligand. L'utilisation d'une fonction de score consensus appropriée permettrait donc de mieux décrire l'ensemble des interactions protéine-ligand et donc le processus de liaison ³⁵³. La seconde hypothèse suggère que la probabilité pour des composés d'être de réels ligands augmente s'ils sont associés aux meilleurs scores en utilisant différentes fonctions de score plutôt qu'une seule. ³⁵⁴ Les fonctions de score peuvent être combinées de diverses manières pour créer une fonction consensus ³⁵⁴. Ainsi, dans l'approche de classement par le nombre (« rank by number »), une moyenne des scores fournis par chaque fonction de score à combiner est réalisée. Cependant, ceux-ci ne sont pas toujours comparables ce qui rend cette stratégie difficilement utilisable (par exemple les

fonctions de score empiriques prédisent généralement l'enthalpie libre de liaison alors que les fonctions de score basées sur les champs de force permettent d'évaluer des énergies de champs de force). Une autre approche est le classement par le rang (« rank by rank ») qui permet de pallier aux problèmes associés à la combinaison des scores en calculant la moyenne des rangs obtenus pour chaque composé avec les différentes fonctions. Enfin, la méthode de classement par vote (« rank by vote ») consiste à classer tous les composés d'une manière semi-quantitative. Dans cette approche, pour chaque fonction de score, un point est attribué à un composé s'il se trouve dans le premiers n pourcents du classement avec n une valeur à définir. Ainsi, plus un composé obtiendra de points, la valeur maximale étant égale au nombre de fonctions de score utilisées, plus la probabilité qu'il soit un réel ligand du récepteur étudié est élevée.

La supériorité des fonctions de score consensus (Figure 57) a été illustrée dans diverses études³⁵⁵⁻³⁵⁸. Cependant, le choix des fonctions de score à combiner n'est pas anodin puisque si les fonctions de score utilisées sont corrélées le résultat risque d'amplifier les erreurs plutôt que de les compenser.

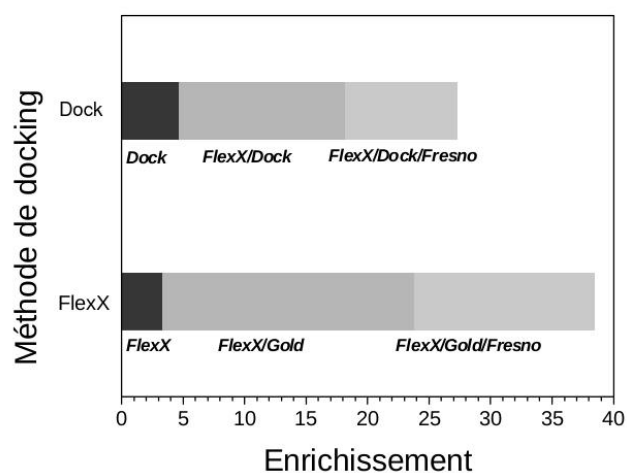


Figure 57. Influence de l'utilisation de fonctions de score consensus sur l'enrichissement en vrais actifs (en noir : FlexX ou Dock seuls ; en gris foncé : combinaison FlexX/Dock ou FlexX/Gold, en gris clair : combinaison FlexX/Dock/Fresno ou FlexX/Gold/Fresno)³⁵⁹

4.5.3 Principaux logiciels de docking

De nombreux logiciels de docking (Tableau 12) présentant des algorithmes de recherche et des fonctions de score variées ont été développés au cours du temps.

Programme	Traitement du ligand	Recherche conformationnelle	Algorithme de recherche	Fonction de score
AutoDock ³²³	Flexible	Stochastique	Génétique	Basée sur les champs de force
Dock ³¹³	Flexible	Systématique	Fragmentation / reconstruction	Basée sur les champs de force
FlexX ³¹²	Flexible	Systématique	Fragmentation / reconstruction	Basée sur les connaissances
FRED ³⁰³	Rigide	Systématique	Recherche exhaustive	Basée sur les connaissances
Glide ³⁰⁸	Flexible	Stochastique	Monte Carlo	Empirique
Gold ³²²	Flexible	Stochastique	Génétique	Basée sur les champs de force
ICM ³¹⁷	Flexible	Stochastique	Monte Carlo	Basée sur les champs de force
Surflex-Dock ³⁶⁰	Flexible	Systématique	Hammerhead modifié	Empirique
PRO_LEADS ³²⁴	Flexible	Stochastique	Tabou	Empirique

Tableau 12. Quelques exemples de logiciels de docking classés selon leur gestion de la flexibilité du ligand, leur approche de la recherche conformationnelle des ligands, leurs algorithmes de recherche et leurs fonctions de score

AutoDock, Glide et GOLD font partie des logiciels de docking les plus populaires (Figure 58).

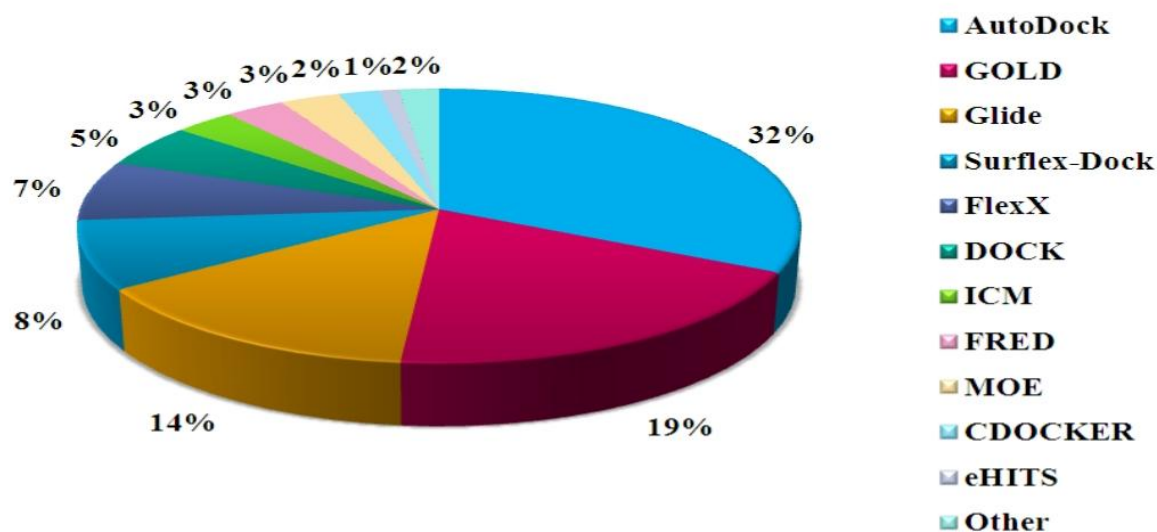


Figure 58. Utilisation des différents logiciels de docking en 2010-2011 (fréquence de citation des noms des logiciels dans les titres et les résumés dans PubMed) (d'après ³⁶¹)

4.5.4 Problématique lié aux méthodes de docking

De nombreux paramètres peuvent conduire à l'échec d'un criblage virtuel réalisé à l'aide d'une méthode de docking (Tableau 13). La disponibilité et la qualité de la structure utilisée pour réaliser l'étude sont notamment primordiales. Pour certains de ces problèmes, aucune solution totalement satisfaisante n'a encore été trouvée. C'est le cas par exemple de la sélection, lorsque plusieurs sont disponibles, de la structure à utiliser ou encore de la prise en compte de la solvatation et de la flexibilité de la protéine.

Cause d'erreur	Traitement
Site actif dénué de cavité	Impossible ?
Flexibilité de la protéine	Très difficile
Influence de l'eau	Très difficile
Imprécision des fonctions d'évaluations	Très difficile
Flexibilité du ligand	Difficile
Mauvais jeux de coordonnées (protéine)	Facile
Mauvais types atomiques (ligand, protéine)	Facile

Tableau 13. Principales sources d'erreur lors d'un docking, classé par complexité décroissante de la solution à apporter pour éviter cette erreur ³⁵⁹

4.5.4.1 Disponibilité des structures des protéines

La base de données de structures la plus exhaustive, et donc logiquement la plus populaire, est la Protein Data Bank (PDB) ⁶⁸. Le nombre de structures disponibles dans cette base de données n'a cessé d'augmenter au cours du temps (Figure 59) et devrait bientôt atteindre la barre symbolique des 100000. Les structures présentes dans cette base de données sont majoritairement des protéines et des acides nucléiques, essentiellement résolues par cristallographie aux rayons X et résonance magnétique nucléaire. Les structures de protéines obtenues par cristallographie aux rayons X décrivent un état cristallin de la protéine présentant un ligand lié dans son site actif (structures « holo ») ou non (structure « apo »). A l'inverse, la résonance magnétique nucléaire ne permet pas d'obtenir une image figée de la protéine mais plutôt un ensemble de conformations illustrant la dynamique de la protéine.

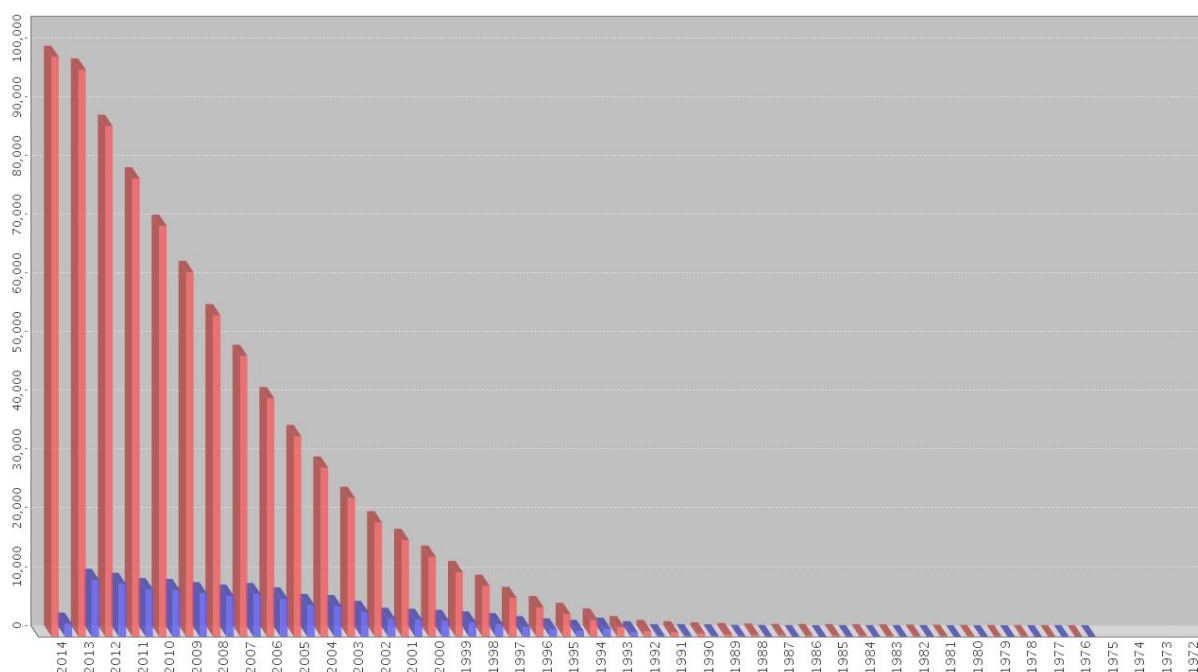


Figure 59. Evolution du nombre de structures disponibles dans la PDB depuis sa création à 2014 (en bleu: nombre de nouvelles structures par an, en rose: nombre total de structures cumulées) (18/03/2014) ³⁶²

La plus grande disponibilité des structures expérimentales a largement contribué au développement des méthodes de criblage virtuel basé sur les structures. Cependant, lorsque de nombreuses structures expérimentales ont été résolues pour une même protéine et que l'on veut n'en retenir qu'une seule pour le criblage virtuel, il n'est pas aisé de décider laquelle choisir. Différentes études ³⁶³⁻³⁶⁸ ont montré que ce choix est loin d'être anodin puisqu'il peut influencer les performances du docking. Différentes pistes ont été proposées, notamment de

préférer les structures holo aux apo ³⁶⁵. J. Liebeschuetz ³⁶⁹ suggère de superposer et d'inspecter toutes les structures disponibles pour vérifier que la structure choisie ne présente pas de caractéristiques non usuelles mais aussi de choisir une structure dont le profil pharmacologique (agoniste ou antagoniste) du ligand co-cristallisé est en adéquation avec celui des ligands à cribler. Enfin, nous avons proposé dans une récente étude des critères de sélection basés sur les propriétés du site de liaison ³⁶⁸. Même si de plus en plus de structures expérimentales de protéines sont disponibles, certaines restent encore non résolues. Dans ce cas, la construction d'un modèle par homologie de séquence constitue une solution. Cette méthode est relativement populaire puisqu'une étude rétrospective montre que près d'un quart des criblages basés sur la structure réalisés entre 2000 et 2009 ont été menés sur un modèle construit par homologie de séquence ³⁷⁰.

4.5.4.2 Importance du solvant

Dans la grande majorité des criblages virtuels basés sur la structure, le solvant n'est pas pris en compte, ni de façon implicite ni de façon explicite. Cependant, en réalité, les interactions ligand-récepteur se déroulent dans un solvant, qui contribue par le biais de liaisons hydrogène et de phénomènes de désolvatation au processus de liaison. Des solutions ont donc été proposées pour tenter de prendre en compte le solvant. La première consiste, comme vu précédemment (4.5.2.2) à utiliser une fonction de score qui simule la solvation de façon implicite comme par exemple le modèle GB/SA ³⁴². Les temps de calculs associés à cette méthode ne permettent cependant pas de l'utiliser en routine dans les criblages virtuels, et d'autres solutions doivent donc être envisagées. Une autre approche consiste à ne pas considérer le solvant dans son ensemble mais de se focaliser sur les molécules de solvant présentes dans le site actif et qui jouent le rôle de relais de liaison hydrogène entre le ligand et le site actif ³⁷¹ (Figure 60).

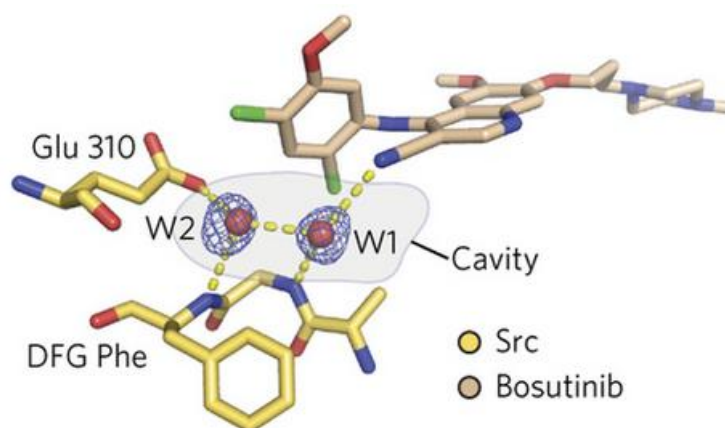


Figure 60. Liaison du bosutinib (en beige) dans le site actif de la Src kinase (en jaune) mettant en jeu des liaisons hydrogènes médiées par des molécules d'eau (W1 et W2) ³⁷²

Dans une étude de 2009 menée sur 1300 structures ³⁷³, Søndergaard et ses collègues dénombrent au moins une molécule d'eau dans le site de liaison placée de manière à pouvoir établir une liaison hydrogène relai entre le ligand et le site de liaison (c'est-à-dire située à moins de 3,5 Å à la fois du ligand et des résidus du site actif) dans 88% des protéines. Pour essayer de prendre en compte ces molécules d'eau, essentielles pour la liaison, différentes approches ont été proposées. La première consiste à conserver une molécule d'eau du site de liaison identifiée comme critique pour la liaison, en l'intégrant comme une partie du récepteur ^{374 369}. Cependant, le point de faible de cette approche réside dans la non prise en compte de la mobilité de la molécule d'eau. Le logiciel SLIDE ³⁷⁵ résout partiellement ce problème en utilisant une approche basée sur les connaissances pour déterminer les molécules d'eau susceptibles d'être conservées après la liaison du ligand et associe une pénalité lorsque la formation du complexe ligand-récepteur provoque le déplacement de ces molécules d'eau. Un autre logiciel, FlexX ³¹², propose un algorithme dénommé le « particle concept » ³⁷⁶ qui intègre des molécules d'eau pour guider la reconstruction incrémentale des ligands lors du docking. Pour cela, avant le processus de docking, les positions favorables des molécules d'eau dans le site actif sont enregistrées en tant que « particules fantômes ». Elles sont ensuite utilisées au cours de la reconstruction incrémentale des ligands, en repositionnant des molécules précédemment enregistrées lorsqu'elles permettent l'établissement de liaisons hydrogène avec le ligand. L'orientation du ligand est alors fonction des contraintes stériques liées aux molécules d'eau et de la géométrie des liaisons hydrogènes entre celles-ci et le ligand.

4.5.4.3 Gestion de la flexibilité de la protéine

La très grande majorité des logiciels considèrent la protéine et donc le site de liaison comme rigide au cours du processus de docking. Ces logiciels négligent donc le phénomène de flexibilité de la protéine lors de la liaison, facilement observable par comparaison de la conformation d'une même protéine co-cristallisée avec deux ligands différents. Les changements conformationnels observés sont souvent limités aux chaînes latérales des résidus du site de liaison mais peuvent aussi constituer une restructuration complète d'une partie de la protéine. C'est notamment le cas lors de la liaison d'un antagoniste dans le site actif du récepteur nucléaire PPAR_alpha (Peroxisome Proliferator Activated Receptor alpha) qui conduit au repositionnement de l'hélice 12 (Figure 61).

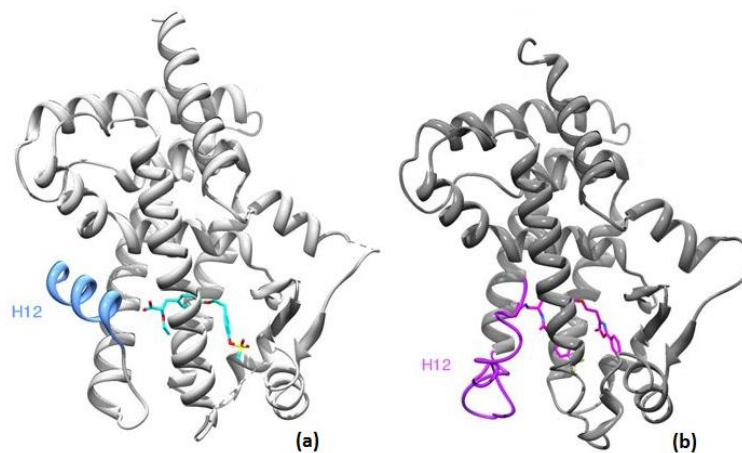


Figure 61. Différence de positionnement de l'hélice H12 du récepteur PPAR_alpha lors de la liaison (a) d'un agoniste (code I17G) ou (b) d'un antagoniste (code 1KKQ)

La prise en compte de la flexibilité de la protéine semble donc essentielle pour la réussite de criblages virtuels avec des méthodes de docking. Cependant, considérer l'ensemble de la protéine comme flexible serait déraisonnable aux vues des temps de calculs associés. Différentes approches, classées en fonction du moment de la prise en compte de la flexibilité, permettent de tenir compte des changements conformationnels de la protéine. La première approche consiste à choisir **avant** le docking un ensemble de conformations de la protéine étudiée. Cet ensemble de conformations peut résulter d'une sélection de structures cristallographiques de la protéine présentant des ligands co-cristallisés différents ou encore de structures résolues par RMN³⁷⁷. Les méthodes de dynamique moléculaire et de minimisation peuvent aussi produire différentes conformations d'une protéine utilisables pour le docking³⁷⁸. La méthode IFREDA (ICM Flexible REceptor Docking Algorithm)³⁷⁹ réalise un premier criblage virtuel à petite échelle sur différentes structures de la cible (cristallographiques ou

généérées *de novo* par l'algorithme IFREDA). L'analyse des performances obtenues (facteurs d'enrichissement et écarts quadratiques moyens) en fusionnant les résultats des différentes cibles et en ne conservant que le meilleur résultat pour chaque ligand permet de sélectionner un petit ensemble de structures sur lequel réaliser le criblage virtuel. De même, le logiciel FlexE superpose un ensemble de structures de la protéine et les regroupe dans une seule structure « unifiée ». Dans cette structure « unifiée », les chaînes latérales des résidus ne présentant pas de changements conformationnels entre les différentes structures étudiées sont représentées par une seule et même chaîne latérale. A l'inverse, si différentes positions sont observées pour une même chaîne latérale, ces différentes positions seront représentées dans la structure « unifiée » (Figure 62) ³⁸⁰.

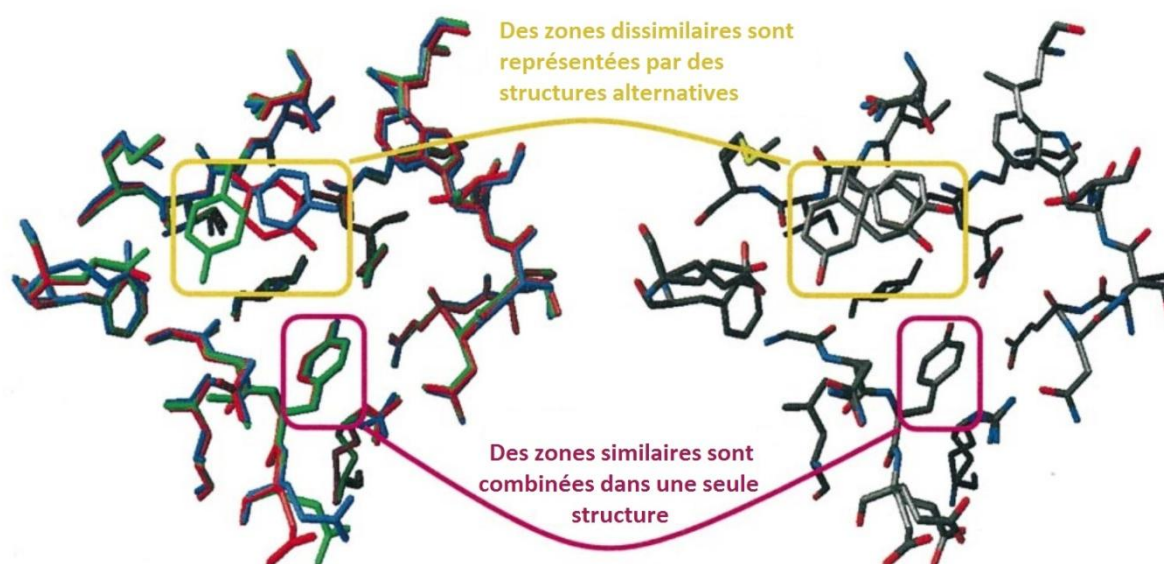


Figure 62. Illustration de la création de la protéine "unifiée" de FlexE ³⁸⁰

La deuxième approche consiste à tenir compte de la flexibilité de la protéine **au cours** du docking. Pour cela, des banques de rotamères sont utilisées pour explorer les degrés de liberté conformationnels des résidus du site de liaison en se limitant aux valeurs discrètes correspondant aux conformations des chaînes latérales des acides aminés les plus fréquemment observées au niveau expérimental ³⁸¹. Enfin la dernière approche consiste à traiter la flexibilité de la protéine **après** le processus de docking, lors de la phase d'attribution de score. C'est le principe du « soft-docking » ³⁸², l'une des premières approches à avoir été développé pour résoudre le problème de la prise en compte de la flexibilité de la protéine. Dans cette méthode, le potentiel de Lennard-Jones est remplacé par une fonction plus permissive autorisant des clashes stériques mineurs qui simulent les mouvements des atomes de la protéine pour s'éloigner du ligand qui se fixe dans le site de liaison ³⁷⁸.

4.6 Succès du criblage virtuel basé sur la structure

Les criblages virtuels basés sur la structure sont actuellement très populaires pour tenter d'identifier de nouveaux hits lors des processus de R&D de nouveaux médicaments. De très nombreux exemples de réussite de criblages virtuels de chimiothèques sont disponibles dans la littérature scientifique ^{32, 33, 359}. Une revue ³⁸³ recense en 2009 31 molécules commercialisées pour lesquelles la structure 3D a rationnellement été mise à profit, et ce de manière significative, pour aboutir à la mise sur le marché du médicament (Tableau 14). Ces molécules sont majoritairement issues de l'analyse rationnelle de structures co-cristallisées plutôt que de criblages virtuels basés sur la structure à proprement dit. En effet, peu de données sont disponibles sur les méthodes mises en œuvre dans les laboratoires de R&D pour découvrir de nouveaux médicaments. Il est cependant possible de citer l'exemple du composé LY-517717, un inhibiteur du facteur Xa issu d'un criblage virtuel basé sur la structure qui a atteint les phases II des essais cliniques ³⁸⁴. Dans les années à venir, le nombre de médicaments issus du criblage virtuel basé sur la structure devrait considérablement augmenter.

DCI	Nom Commercial	Année	Cible	Indication	Stratégie de découverte	Référence
Dorzolamide	Trusopt	1995	Anhydrase Carbonique	Glaucome	RSA à partir de la structure co-cristallisée de la HCA-II avec 2 énantiomères	385, 386
Saquinavir	Invirase	1995	HIV protéase	SIDA	RSA à partir de la structure co-cristallisée du composé Ro 31-8558 et de HIV-1 Protéase	387, 388
Ritonavir	Norvir	1996	HIV protease	SIDA	RSA à partir de la structure cristallisée de HIV-1 Protéase	389
Indinavir	Crixivan	1996	HIV protease	SIDA	RSA à partir de la structure cristallisée de HIV-1 Protéase	390
Brinzolamide	Azopt	1999	Anhydrase Carbonique	Glaucome	RSA à partir du dorzolamide	383
Nelfinavir	Viracept	1999	HIV protease	SIDA	RSA à partir de la structure cristallisée de HIV-1 Protéase	391
Amprenavir	Agenerase	1999	HIV protease	SIDA	RSA à partir de la structure cristallisée de HIV-1 Protéase	391
Lopinavir	Aluviran	1999	HIV protease	SIDA	RSA à partir de la structure cristallisée de HIV-1 Protéase	383
Zanamivir	Relenza	1999	Neuraminidase	Influenza	Analyse du site de liaison (GRID), RSA à partir de la structure co-cristallisée de la neuraminidase avec un ligand mimant la partie acide sialique	392, 393, 394
Oseltamivir	Tamiflu	1999	Neuraminidase	Influenza	RSA à partir du zanamivir	395
Atazanavir	Reyataz	2003	HIV protéase	SIDA	RSA à partir de la structure cristallisée de HIV-1 Protéase	396
Fosamprenavir	Telzir	2003	HIV protéase	SIDA	RSA à partir de la structure cristallisée de HIV-1 Protéase	397
Ximelagatran	Exanta	2004	Thrombine	Embolie veineuse	Etudes de docking guidant la synthèse	398
Tipranavir	Aptivus	2005	HIV protéase	SIDA	RSA à partir de la structure cristallisée de HIV-1 Protéase	399, 400
Sunitinib	Sutent	2006	Kinase	Cancer	RSA à partir de la structure co-cristallisée de la FGFR1 kinase et du SU5402	401
Darunavir	Prezista	2006	HIV protéase	SIDA	RSA à partir amprenavir	402
Nilotinib	Tasigna	2006	BCR-ABL kinase	LMC	RSA à partir de la structure co-cristallisée de la BCR-ABL kinase avec l'imatinib	403
Aliskiren	Tekturna	2007	Rénine	Hypertension	RSA incluant des étapes de docking	404
Dabigatran	Pradaxa	2008	Thrombine	Embolie veineuse	RSA à partir de la structure co-cristallisée de la thrombine bovine avec le NAPAP	405
Pazopanib	Votrient	2009	Kinase	Cancer de l'ovaire	RSA à partir de modèle par homologie de VEGFR2	406
Boceprevir	Victrilis	2011	HCV protease	Hépatite C	RSA à partir de la structure cristallisée	407

Tableau 14. Quelques exemples de médicaments pour lesquels la structure 3D a été rationnellement et significativement mise à profit (RSA : Relations Structure-Activité, HIV : Human Immunodeficiency Virus, LMC : Leucémie Myéloïde Chronique) (d'après ³⁸³)

5 Evaluation des méthodes de criblage virtuel

L'intérêt porté aux méthodes de criblage virtuel réside dans le gain potentiel de temps et d'argent que représente leur utilisation, en combinaison de HTS dans le processus de R&D du médicament. Cependant, les résultats obtenus par criblage virtuel ne sont que des prédictions et leur fiabilité est donc un sujet d'interrogation. L'évaluation des méthodes de criblage virtuel est donc essentielle pour s'assurer de leurs performances et ainsi valider leurs résultats. Cette évaluation permet aussi, de manière tout aussi importante, de guider les bioinformaticiens dans le choix du meilleur outil, ou tout du moins de l'outil le mieux adapté, pour réaliser un criblage virtuel sur une cible donnée.

Deux types d'évaluation sont possibles : les évaluations prospectives et les évaluations rétrospectives. Les évaluations prospectives permettent de vérifier expérimentalement les prédictions d'affinité de ligands d'un criblage virtuel pour une cible ⁴⁰⁸. Cependant, de telles études sont très coûteuses et ne sont possibles que par mise à disposition des résultats expérimentaux de HTS des laboratoires de l'industrie pharmaceutique qui sont très réticents à partager ces résultats aux vues de leur politique de protection de propriété intellectuelle ³⁵⁰. Les évaluations des méthodes de docking se font donc de manière rétrospective, en s'intéressant à deux critères principalement, la précision du positionnement et l'enrichissement d'une chimiothèque ³⁵⁰.

5.1 Précision du positionnement

La précision du positionnement peut être évaluée de différentes manières. La première et la plus populaire, s'intéresse à l'écart quadratique moyen ou RMSD (Root Mean Square Deviation) entre une position prédite et une position observée dans un co-cristal. Parmi les autres méthodes, l'erreur relative de déplacement (Relative Displacement Error RDE), l'espace réel du facteur R (Real Space R-factor RSR) et l'évaluation visuelle du positionnement par classification de précision basée sur les interactions (Interactions-Based Accuracy Classification IBAC) peuvent être citées.

5.1.1 Ecart quadratique moyen ou RMSD

Le RMSD entre deux poses est une mesure géométrique de la distance entre les positions atomiques de la structure expérimentale et celles de la structure prédite du complexe ligand / site de liaison ³⁶⁴ (Équation 25).

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

Équation 25. Calcul du RMSD (avec v_i et w_i les atomes identiques de la structure expérimentale et de la structure prédite respectivement ; x , y et z les coordonnées cartésiennes ; n le nombre total d'atomes)

Plus la prédiction du positionnement est précise, plus les différences entre les deux structures sont faibles, plus la valeur du RMSD est basse.

Pour réaliser un calcul de RMSD, le ligand co-cristallisé d'une cible donnée est extrait, et repositionné par docking dans le site actif. C'est ce qu'on appelle en anglais le « cognate re-docking » ou encore « self docking » ³⁶⁴. Pour évaluer la précision du positionnement, seuls les atomes lourds sont généralement considérés ⁴⁰⁹. Les avantages présentés par le calcul du RMSD tels que l'objectivité, la sensibilité et la simplicité d'automatisation ⁴⁰⁹ ont permis à cette métrique de s'imposer comme la référence dans l'évaluation de la précision du positionnement. Toutefois, le RMSD ne constitue pas un critère d'évaluation parfait et présente certaines faiblesses. Ainsi, le RMSD permet d'évaluer les distances entre les positions atomiques mais ne fournit aucune information sur la conservation des interactions entre la structure prédite et la structure cristallisée ⁴⁰⁹. En conséquence, des différences dans les champs de force utilisés pour le docking, des variations de chaînes latérales flexibles non impliquées dans la liaison au récepteur, et une molécule presque symétrique retournée dans le site de liaison peuvent être associées à de grandes valeurs de RMSD traduisant un mauvais positionnement alors que les interactions clés sont bel et bien conservées ⁴¹⁰. La dépendance vis-à-vis du poids moléculaire reste tout aussi problématique, notamment pour les petits composés qui sont fréquemment associés à de faibles valeurs de RMSD ⁴⁰⁹. Enfin, l'analyse des valeurs de RMSD en elle-même n'est pas totalement aisée, puisqu'il faut définir une valeur seuil au-dessus de laquelle la précision du positionnement est considérée comme mauvaise.

5.1.2 L'erreur relative de déplacement RDE

L'utilisation de la valeur moyenne du RMSD peut conduire à des erreurs d'interprétation, surtout lorsque quelques valeurs individuelles sont extrêmes ³⁶⁴. L'erreur relative de déplacement (Équation 1), RDE, a été développée pour réduire l'impact de ces grandes divergences sur la valeur moyenne à analyser.

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

Équation 26. Calcul de l'erreur relative de déplacement (RDE) pour un ligand de N atomes i par rapport au ligand co-cristallisé de N atomes i' (avec L : le paramètre d'échelle définissant l'échelle de précision et dont la valeur est couramment située entre 1,5 et 3 Å ; et $D_{ii'}$ la déviation de l'atome i prédit par rapport à l'atome i' de référence) ⁴¹¹

Lorsque le positionnement prédit est correct, les valeurs de déviations sont nulles ou proches de 0 et la valeur correspondante de RDE est égale à 0% ou presque. Une valeur de RDE de 50% peut être obtenue lorsque le positionnement des atomes est tel que la valeur de déviation D est égale à la valeur du paramètre d'échelle L ou encore lorsque le positionnement de la moitié des atomes est correct (D très inférieure à L) et celui de l'autre moitié est incorrect (D très supérieure à L) ⁴¹¹.

Le RDE ne permet cependant pas de pallier les autres faiblesses du RMSD.

5.1.3 L'espace réel du facteur R RSR

Les valeurs de RSR ⁴¹² sont calculées à partir des cartes de densité électronique théoriques du modèle (le ligand de référence ou la pose prédite) et de la densité électronique expérimentale (Équation 27 et Figure 63).

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

Équation 27. Calcul du RSR à l'aide de la densité électronique observée expérimentalement (ρ_{obs}) et de la densité électronique calculée à partir des coordonnées atomiques du modèle : le ligand co-cristallisé ou la pose prédite (ρ_{calc}) ⁴¹²

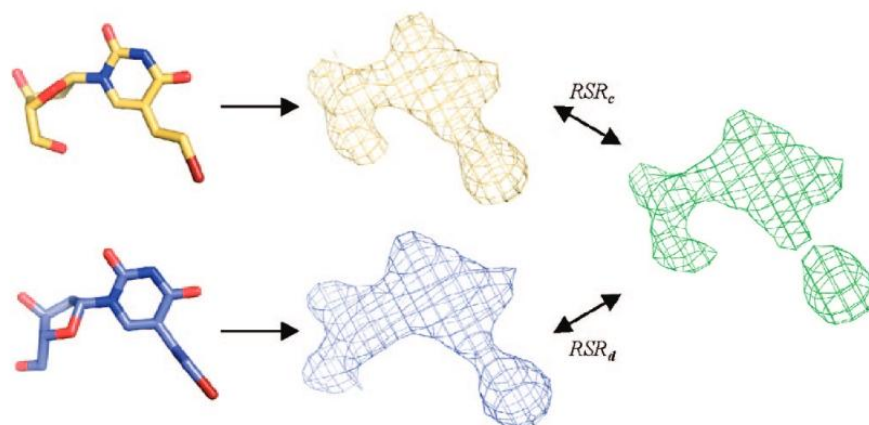


Figure 63. Calcul des valeurs de RSR pour le ligand co-cristallisé (RSR_c) et la pose prédite (RSR_d) à partir des cartes de densité électronique théoriques du ligand de référence (jaune) ou de la pose prédite (bleu) et de la densité électronique mesurée expérimentalement (vert) ⁴¹²

Lorsque la structure étudiée est un modèle fidèle de la réalité, les densités électroniques observées expérimentalement et calculées à partir du modèle sont similaires et la valeur de RSR correspondante est faible. Dans le cas contraire, si le modèle n'est pas correct, la valeur de RSR est élevée. Cependant, comme les valeurs de RSR dépendent fortement de la résolution de la structure cristallographique, la comparaison des valeurs de RSR_d , calculées pour les poses prédites, ne permet pas à elle seule de différencier les poses correctes des poses incorrectes. Pour s'affranchir de ce problème, le ratio des valeurs de RSR des poses prédites (RSR_d) et du ligand co-cristallisé (RSR_c) est utilisé pour évaluer la qualité du positionnement (Équation 28).

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

Équation 28. Calcul du ratio RSR_n à partir des valeurs de RSR des poses prédites (RSR_d) et du ligand co-cristallisé (RSR_c) ⁴¹²

Une valeur de RSR_n égale à 1 indique que les modèles prédit et cristallisé représentent de la même manière la carte de densité électronique alors qu'une valeur de RSR_n supérieure à 1 dénote une moins bonne représentation de la carte de densité électronique par le modèle prédit. Il est à noter que des modèles prédit et cristallisé peuvent présenter une valeur de RSR_n égale à 1 tout en possédant des coordonnées atomiques différentes. ⁴¹²

5.1.4 Classification de précision basée sur les interactions IBAC

La classification de précision basée sur les interactions repose sur une inspection visuelle des structures co-cristallisées et prédites⁴¹⁰. L'inconvénient majeur de cette technique est qu'il s'agit d'un protocole non automatisé. La classification débute par une inspection visuelle du complexe co-cristallisé ligand / site de liaison pour déterminer précisément l'ensemble des interactions. Ensuite, différents protocoles sont possibles, dont deux sont proposés dans la présentation de l'IBAC⁴¹⁰. Le premier consiste à diviser le ligand en deux parties : le squelette et le reste de la molécule. Lorsque le positionnement du squelette du ligand prédit est comparable à celui du ligand co-cristallisé et que toutes les interactions observées dans la structure co-cristallisée (pour le squelette et le reste de la molécule) sont retrouvées avec la structure prédite, le positionnement est accepté comme étant correct. Dans le cas contraire, il est déclaré incorrect. Dans le second protocole, le ligand est considéré dans son ensemble et la pose prédite peut être classée comme correcte, presque correcte ou incorrecte. Ainsi, une pose est dite correcte si sa conformation et son orientation ainsi que toutes les interactions clés sont similaires à celle de la structure de référence. Une pose presque correcte possède elle aussi une conformation et une orientation adéquates mais certaines interactions clés (jusqu'à un quart du nombre total) ne sont pas retrouvées. Toute autre pose est incorrecte. Le grand avantage de cette classification par rapport au RMSD est la prise en compte des interactions, cependant, le protocole de l'IBAC n'est ni standardisé, ni automatisé, ce qui rend son utilisation encore marginale.

5.2 Enrichissement d'une chimiothèque

Une méthode de docking doit, en plus de prédire le mode de liaison correct d'un ligand dans le site actif (précision du positionnement), permettre de différencier les molécules réellement capables de se lier au récepteur pour provoquer un effet biologique, ce qu'on appelle des actifs, du reste des molécules de la chimiothèque, les inactifs. Lorsque la fraction d'actifs retrouvée grâce aux méthodes de docking est largement supérieure à celle retrouvée en sélectionnant des composés au hasard dans la chimiothèque, la méthode de docking peut être considérée comme efficace⁴⁰⁹. Pour évaluer la capacité d'une méthode de docking à enrichir une chimiothèque en actifs, c'est-à-dire de leur attribuer les meilleurs scores afin qu'ils se concentrent dans la première fraction de la chimiothèque classée, des banques d'évaluation et des métriques de performances ont été développées.

5.2.1 Les banques d'évaluation

Les banques d'évaluation (ou « benchmarking data sets ») rassemblent, pour une ou plusieurs cibles, deux types de composés : des actifs, c'est-à-dire des molécules dont l'activité est connue et prouvée pour une cible donnée et des inactifs. Idéalement, les inactifs devraient être choisis, tout comme les actifs, sur la base de données expérimentales. Cependant, les composés inactifs pour une cible donnée ne sont généralement pas renseignés, ou dans une proportion trop faible pour permettre une évaluation. Des molécules leurres ou « decoys », présumées inactives pour une cible donnée sont alors utilisées comme inactifs.⁴⁰⁹

La première banque d'évaluation a été proposée par Didier Rognan et ses collègues du département des Biosciences Appliquées de Zürich³⁵⁵. Cette banque, utilisée pour évaluer 3 programmes de docking (Dock, FlexX et GOLD) et 7 fonctions de score (Dock, FlexX, GOLD, PMF, Chemscore, Fresno, Score), regroupe deux cibles (le récepteur aux œstrogènes alpha ER α et la Thymidine Kinase TK). Pour chacune de ces cibles, le jeu de données d'évaluation est constitué de 10 actifs, 990 decoys choisis au hasard dans la base de données ACD (Advanced Chemical Directory) préfiltrée pour éliminer les composés réactifs, inorganiques et de poids moléculaire inadéquat (inférieur à 250 ou supérieur à 500 Da) et une structure extraite de la PDB. La publication de cette banque a donné le coup d'envoi au développement d'autres bases de données, avec pour objectif de tenter d'améliorer la qualité de l'évaluation des méthodes de docking. Ainsi, dans un travail de 2006, Pham et Jain⁴¹³ intègrent cette base de données initiale dans une banque avec 27 nouveaux jeux de données. Les actifs pour ces nouvelles cibles sont issus de travaux précédents pour 2 d'entre elles (Poly ADP-Ribose Polymerase PARP⁴¹⁴ et Protein Tyrosine Phosphatase 1b PTP1b⁴¹⁵) et générés à partir de la PDBbind database⁴¹⁶ pour les 25 restantes. Les decoys de la banque de Rognan sont conservés mais filtrés pour présenter un nombre de liaisons rotatives inférieur à 15 et pour les 27 nouvelles cibles, 1000 decoys sont extraits aléatoirement de la fraction des composés de la ZINC⁴³ présentant un caractère « drug-like ». D'autres équipes préfèrent mettre à profit des données internes plutôt que d'utiliser des bases de données publiques, notamment pour la sélection des actifs⁴¹⁷. Ainsi de nombreuses banques d'évaluation ont été construites avec différents protocoles^{304, 365, 418, 419, 420, 421}. Cependant, il s'est vite avéré que l'un des points clés influençant la qualité de ces banques réside dans l'étape de sélection des actifs et des inactifs. Si la sélection des actifs est relativement aisée grâce aux données de la littérature, il faut cependant se méfier des bases de données de bioactivités qui peuvent contenir des erreurs⁴²² et préférer une revue manuelle des données (voir Deuxième Partie

Résultats). La sélection des decoys est plus complexe. En effet, la sélection aléatoire des decoys précédemment décrite est une source de biais dans l'évaluation des méthodes puisque les propriétés physiques des decoys peuvent être très dissimilaires de celles des ligands. Pour résoudre ce problème, la DUD (Directory of Useful Decoys)⁴²³ a été créée et est rapidement devenue la banque d'évaluation de référence en alliant une sélection rationnelle des decoys avec 40 larges jeux de données pour permettre une évaluation complète et robuste. Ainsi, pour s'affranchir du biais précédemment évoqué, les decoys sont choisis de manière à présenter des propriétés physicochimiques comparables à celle des ligands, mais à être structurellement différents afin de tenter de ne sélectionner que de réels inactifs non capables de se lier au site de liaison de la cible étudiée. (Figure 64). Malgré cela, des améliorations possibles ont été proposées pour pallier le manque de diversité, que ce soit dans les structures des ligands⁴²⁴ ou dans le choix des cibles, ou encore pour améliorer la sélection des decoys. Ainsi, le défaut de prise en compte de la charge nette pour les decoys de la DUD a été mis en avant^{364, 408, 425}. Une autre étude intéressante a proposé l'utilisation de decoys virtuels dont la construction ne tient pas en compte de la faisabilité synthétique pour obtenir des composés dont les propriétés physicochimiques sont encore plus similaires à celles des actifs⁴²⁶.

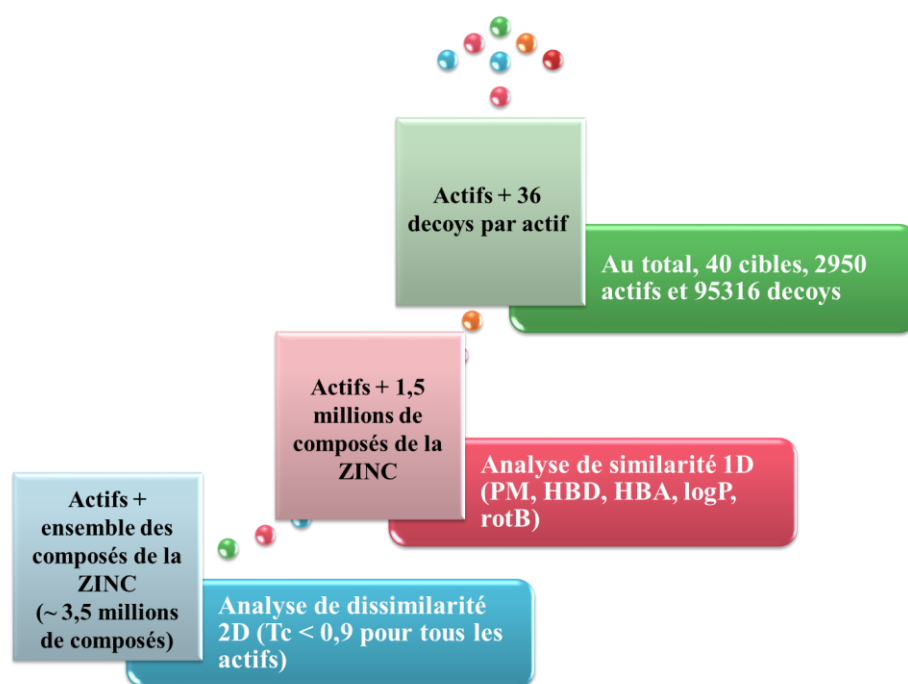


Figure 64. Protocole de sélection des decoys de la DUD dans la ZINC (avec PM: Poids Moléculaire, HBD: donneur de liaison hydrogène, HBA: accepteur de liaison hydrogène, logP: coefficient de partage eau-octanol, rotB, nombre de liaisons rotatives) d'après⁴²³

Pour répondre aux faiblesses précédemment évoquées, une nouvelle version de la DUD, la DUD_E⁴²⁷ a été présentée en 2012. Cette banque d'évaluation présente pas moins de 102 cibles appartenant à 8 grandes familles de protéines, pour un total de 66695 ligands (et 22886 après clusterisation), 9219 decoys expérimentaux et 1411214 decoys extraits de la ZINC à l'aide d'un protocole assez similaire à celui de la DUD mais en ajoutant notamment la charge nette en tant que critère supplémentaire de sélection des decoys et en durcissant le critère de dissimilitude topologique.

5.2.2 Les métriques de performance

Pour évaluer les méthodes de criblage virtuel, des métriques de performance sont utilisées pour chiffrer, au sein de la fraction étudiée de la chimiothèque classée, le taux d'actifs aussi appelés vrais positifs, VP, illustrant la sélectivité (Équation 29), et de decoys (somme des vrais négatifs VN et des faux positifs FP) témoin de la spécificité (Équation 30)⁴⁰⁹.

$$Se = \frac{N_{actifs[fraction]}}{N_{actifs[total]}} = \frac{VP}{VP + FN}$$

Équation 29. La sélectivité (Se) est définie comme le ratio du nombre d'actifs retrouvés par la méthode de docking dans une fraction donnée de la chimiothèque classée ($N_{actifs[fraction]}$ ou VP : vrais positifs) sur le nombre d'actifs total de la chimiothèque ($N_{actifs[total]}$) (avec FN les faux négatifs, c'est-à-dire les actifs non reconnus en tant que tels par la méthode de criblage virtuel)⁴⁰⁹

$$Sp = \frac{N_{inactifs\ non\ présents[fraction]}}{N_{inactifs[total]}} = \frac{VN}{VN + FP}$$

Équation 30. La spécificité (Sp) représente le ratio du nombre de decoys non classés dans la fraction de la chimiothèque ($N_{inactifs\ non\ présents[fraction]}$ ou VN : vrais négatifs) sur le nombre de decoys total ($N_{inactifs[total]}$) (avec FP : faux positifs)⁴⁰⁹

A partir de ces valeurs, des métriques de performances peuvent être calculées. On distingue généralement les métriques dites « classiques », telles que les facteurs d'enrichissement et les courbes de ROC (Receiver Operating Characteristic) et les métriques dites « avancées », telles que « l'amélioration robuste initiale » (ou Robust Initial Enhancement RIE) et la discrimination de Boltzmann améliorée des courbes de ROC (Boltzmann-enhanced discrimination of ROC BEDROC), qui ne considèrent pas seulement la capacité des méthodes

de criblage virtuel à retrouver un certain nombre d'actifs mais aussi leur capacité à classer ces actifs au début de la fraction de la chimiothèque triée.

5.2.2.1 Facteurs d'enrichissement

Le facteur d'enrichissement (ou enrichment factor EF) est une métrique de performance couramment utilisée. Il est calculé pour des fractions choisies de la chimiothèque et compare la capacité de la méthode à retrouver des actifs dans une fraction donnée de la chimiothèque par rapport à une sélection aléatoire (Équation 31).⁴⁰⁹

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

*Équation 31. Calcul du facteur d'enrichissement (EF) pour la fraction des 100*n/N premiers pourcents de la chimiothèque (avec n: nombre de composés dans la fraction de la chimiothèque étudiée, N : nombre total de composés dans la chimiothèque, VP : vrais positifs, et FN : faux négatifs)*

Cependant, cette métrique présente deux faiblesses majeures. La première réside dans la dépendance du facteur d'enrichissement au ratio d'actifs de la chimiothèque⁴²⁸. En conséquence, le facteur d'enrichissement permet de comparer les performances de différentes méthodes sur un même jeu de données mais son utilisation est plus délicate lorsqu'il s'agit de comparer des performances obtenues sur des jeux de données présentant des ratios d'actifs / decoys différents. Le second problème associé au facteur d'enrichissement est commun à toutes les métriques « classiques » d'évaluation des performances. En effet, le facteur d'enrichissement chiffre le nombre d'actifs présents dans la fraction étudiée de la chimiothèque sans tenir compte de leur distribution⁴²⁸. Ainsi, deux méthodes retrouvant le même nombre d'actifs dans la même fraction de la chimiothèque présentent une valeur de facteur d'enrichissement identique, et ce, même dans les cas où dans la fraction considérée, les actifs sont tous classés premiers avec une méthode et derniers avec l'autre (Figure 65).

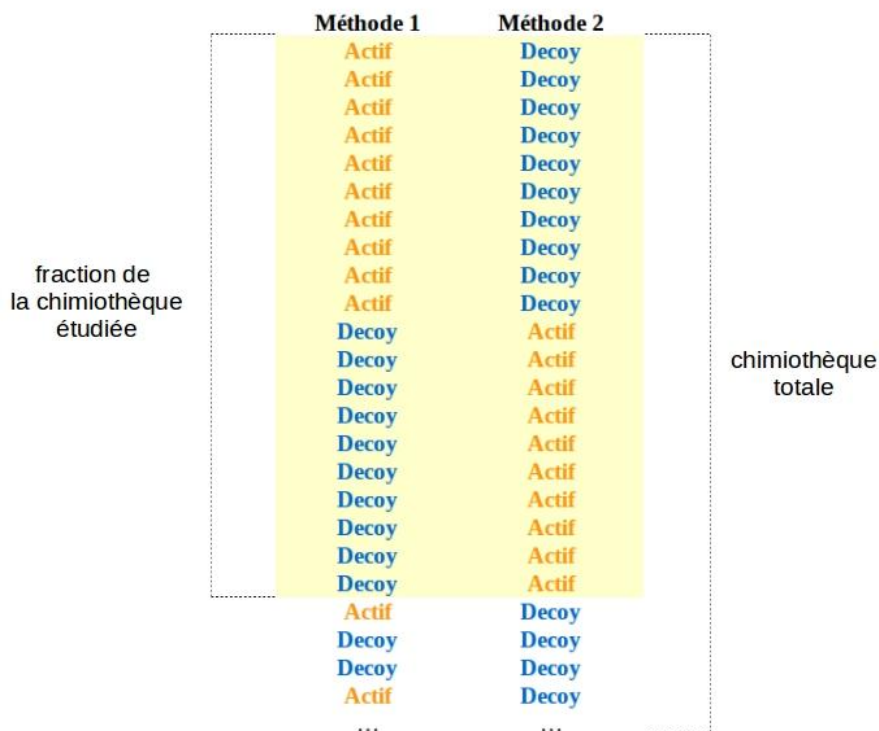


Figure 65. Les méthodes 1 et 2 présentent le même taux d'actifs retrouvés pour la fraction considérée de la chimiothèque et donc des facteurs d'enrichissement similaires tout en affichant des distributions d'actifs et de decoys tout à fait dissemblables. Les valeurs d'EF indiquent que les deux méthodes considérées possèdent des performances identiques, alors qu'en réalité la méthode 1 semble plus performante.

Les courbes d'enrichissement représentant le taux d'actifs en fonction du pourcentage de la chimiothèque classée permettent de visualiser les performances globales des méthodes ⁴²⁹, même si pour cela l'AUC (Area Under the Curve) des courbes de ROC est généralement préférée.

5.2.2.2 Courbes de ROC (Receiver Operating Characteristic)

Les courbes de ROC (Figure 66) illustrent l'évolution de la sensibilité (Se) en fonction de un moins la spécificité (1-Sp) ou en d'autres termes le pourcentage d'actifs en fonction du pourcentage de decoys pour chaque fraction de la chimiothèque ⁴²⁹.

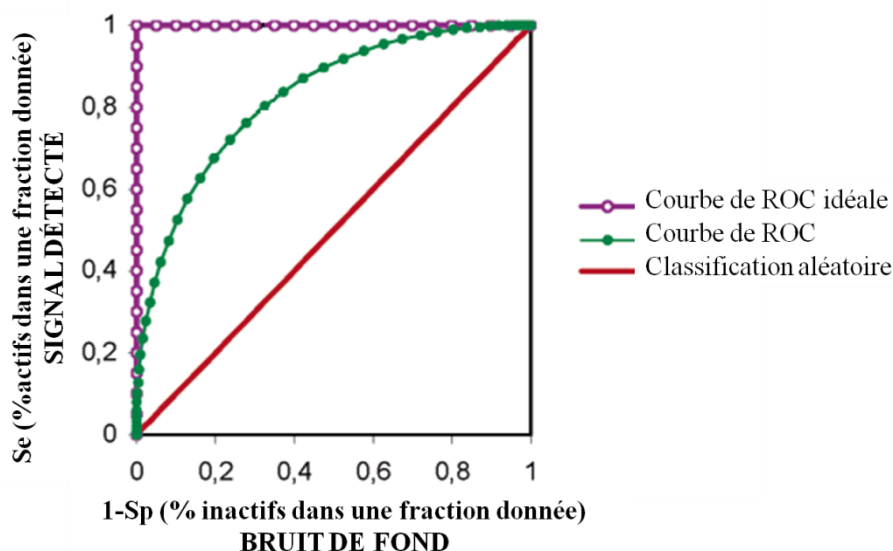


Figure 66. Les courbes de ROC représentent l'évolution, pour chaque fraction de la chimiothèque, de la sensibilité (Se) et de la spécificité (Sp), avec Se fonction de $1 - Sp$. Une classification aléatoire des composés de la chimiothèque est représentée par une diagonale allant du point (0,0) au point (1,1) du graphique. Si une méthode permet de discriminer les actifs et les decoys de manière plus efficace que le hasard, la majorité des points de la courbe de ROC est située au-dessus de la diagonale. Pour une distribution idéale, c'est-à-dire que les actifs sont tous classés en premier, la courbe monte en ligne droite jusqu'au point (0,1) ($Se = Sp = 1$ pour tous les actifs) puis viennent ensuite les inactifs ($Se = 1, Sp = 0$ pour tous les inactifs) représentés par une ligne horizontale entre les points (0,1) et (1,1) du graphique (d'après ⁴²⁹)

A partir d'une courbe de ROC, le calcul de l'aire sous la courbe (Équation 32), couramment notée AUC (Area Under the Curve) permet de quantifier la performance globale d'enrichissement de la méthode étudiée.

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

Équation 32. L'AUC est calculée en sommant tous les rectangles formés par les valeurs de Se et $1-Sp$ de chaque point de la courbe de ROC correspondant à la molécule classée à la position i ⁴⁰⁹

Pour une classification aléatoire des composés de la chimiothèque, l'AUC est égale à 0,5. Les performances d'une méthode conduisant à une AUC inférieure ou égale 0,5 sont dites mauvaises puisque la méthode ne permet pas d'obtenir une discrimination entre les actifs et les decoys meilleure que celle générée par le hasard. La valeur maximale d'AUC est de 1, elle

est obtenue pour les cas idéaux où tous les actifs sont associés à un score supérieur à ceux de tous les decoys. De manière plus générale, plus la valeur d'AUC associée à une méthode est élevée, plus la capacité de la méthode à discriminer les actifs des inactifs est forte. Ces valeurs d'AUC peuvent donc être utilisées pour comparer les performances de plusieurs méthodes de docking, ce qui est aussi possible par inspection visuelle des courbes.

Les courbes de ROC présentent l'avantage sur les facteurs d'enrichissement de ne pas dépendre du ratio d'actifs / decoys de la chimiothèque à cribler⁴²⁹. Cependant, tout comme les facteurs d'enrichissement, les valeurs d'AUC ne reflètent pas des différences dans le classement des actifs (Figure 67). Pour une évaluation complète des performances de méthode de docking, il est donc important de corrélérer la valeur d'AUC à l'inspection visuelle des courbes de ROC et aux valeurs des facteurs d'enrichissement précoces.

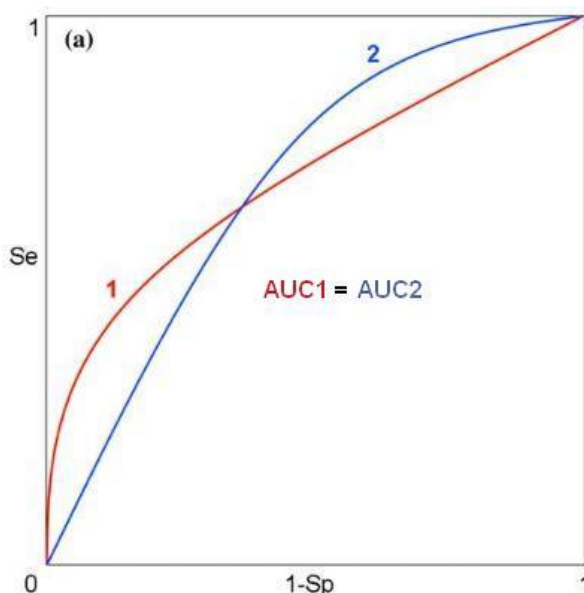


Figure 67. Les AUC des courbes 1 et 2 présentent des valeurs similaires, indiquant que les deux méthodes de criblages virtuels présentent des performances globales identiques. Cependant, la méthode associée à la courbe rouge est plus efficace pour discriminer de façon précoce les actifs des decoys⁴⁰⁹

5.2.2.3 Robust Initial Enhancement (RIE)

Le RIE, développé par Sheridan et ses collègues⁴³⁰, permet de corrélérer le rang d'un actif dans la chimiothèque classée au nombre de composés étudiés (Équation 33).

$$RIE = \frac{S}{\langle S \rangle} = \sum_{i=1}^{actives} \exp\left(\frac{-rank(i)}{a}\right) \times \frac{1}{\langle S \rangle}$$

Équation 33. Le RIE est calculé comme le ratio de la somme des poids de tous les actifs (S) sur la moyenne de la somme ($\langle S \rangle$) obtenue par 1000 tests aléatoires de classification des actifs (avec i le $i^{\text{ème}}$ actif de la chimiothèque et a : le nombre de composés sélectionnés) ⁴³⁰

Le poids des actifs placés au début de la chimiothèque classée est donc proche de 1, et diminue doucement avec l'augmentation des rangs ⁴³⁰.

Le RIE permet donc de quantifier combien de fois la distribution d'actifs obtenue avec une méthode de criblage virtuel est meilleure qu'une distribution aléatoire. Ainsi, une valeur de RIE supérieure à 1 indique que la méthode étudiée permet de scorer les molécules actives à de plus hauts rangs qu'une distribution aléatoire. Au contraire, une valeur de RIE égale à 1 montre que la méthode étudiée ne permet pas d'obtenir de meilleures performances que lors de la sélection aléatoire de composés dans la chimiothèque. ⁴²⁹

Le RIE permet donc, contrairement aux descripteurs classiques de prendre en compte le rang des actifs dans la chimiothèque classée. Cependant, tout comme les facteurs d'enrichissement, les valeurs sont dépendantes du ratio d'actifs / decoys dans la chimiothèque, ce qui rend difficile la comparaison de valeurs de RIE obtenues sur des jeux de données de ratios différents ⁴²⁹. D'autre part, la valeur de RIE représente une performance pour un seuil donné (a dans l'Équation 33) et ne permet donc pas de quantifier la performance globale de la méthode.

5.2.2.4 Boltzmann-enhanced discrimination of ROC (BEDROC)

La métrique de performance BEDROC (Équation 34) a été créée par Truchon et ses collègues ⁴²⁸ pour allier la prise en compte du rang des actifs du RIE à l'évaluation des performances globales de l'AUC. Les valeurs possibles de la métrique BEDROC sont comprises entre 0 et 1.

$$\begin{aligned}
 BEDROC &= \frac{RIE - RIE_{min}}{RIE_{max} - RIE_{min}} \\
 &= RIE \times \frac{R_a \sinh \frac{\alpha}{2}}{\cosh \frac{\alpha}{2} - \cosh \left(\frac{\alpha}{2} - \alpha R_a \right)} + \frac{1}{1 - e^{\alpha(1-R_a)}}
 \end{aligned}$$

Équation 34. Calcul de la métrique de performance BEDROC à partir du descripteur RIE (avec R_a : le ratio d'actifs et α : le paramètre de reconnaissance précoce calculé à partir du pourcentage θ du score total à z pourcent du rang par l'équation $0 = \theta(1 - e^{-\alpha}) + 1 - e^{-\alpha z} - 1$. La valeur de α usuellement utilisée est de 20 pour que les 8 premiers pourcents (z) des rangs relatifs contribuent pour 80 pourcents (θ) de la valeur de BEDROC)⁴²⁸

La valeur de BEDROC ainsi calculée est donc encore dépendante du ratio actifs / decoys de la chimiothèque criblée. Cependant, lorsque la valeur αR_a est très inférieure à 1, une approximation de la valeur de BEDROC s'affranchissant du ratio actifs / decoys est obtenue (Équation 35).

$$\text{Pour } \alpha R_a \ll 1 \text{ et } \alpha \neq 0, \text{ BEDROC} \approx \frac{RIE}{\alpha} + \frac{1}{1 - e^{\alpha}}$$

Équation 35. Approximation de la valeur de BEDROC lorsque la valeur de αR_a est très inférieure à 1.

La valeur de BEDROC ainsi obtenue peut donc être utilisée pour comparer les performances de différentes méthodes de criblage virtuel. Elle indique la probabilité qu'un actif sélectionné aléatoirement dans la chimiothèque classée soit retrouvé avant un composé sélectionné aléatoirement dans une fonction de distribution de probabilité hypothétique.

6 Objectifs de thèse

Dans le cadre de ce travail, nous nous sommes intéressés aux méthodes de criblage virtuel basées sur les structures de deux façons différentes.

Dans un premier temps, nous nous sommes consacrés à l'évaluation rétrospective de ces méthodes. Pour cela, nous avons étudié la façon de traiter la flexibilité de la protéine lors d'un criblage virtuel, avec comme problème sous-jacent le choix de la (des) structure(s) de référence. Dans le but d'améliorer encore l'évaluation des méthodes, nous avons aussi construit une banque d'évaluation des méthodes de criblage virtuel dédiée aux récepteurs nucléaires présentant de nouvelles lignes directrices pour guider la construction de futures banques d'évaluation.

Grâce à l'expertise ainsi acquise, des méthodes de docking ont ensuite été mises en œuvre dans un criblage virtuel prospectif à la recherche de nouveaux inhibiteurs de l'IL-6 utilisables notamment dans le traitement de la polyarthrite rhumatoïde.

Deuxième partie
Résultats

1 Evaluation des méthodes de criblage virtuel

1.1 SBVLS : Définition de critères basés sur les propriétés du site de liaison pour optimiser la sélection de la (ou des) structure(s) de référence

1.1.1 Introduction

Comme présentées dans l'introduction, les techniques de criblage virtuel sont maintenant couramment intégrées dans les processus de R&D, et notamment les approches basées sur les structures telles que les méthodes de docking. Cependant, le traitement de la flexibilité de la protéine, inhérent à l'utilisation de méthodes basées sur les structures, reste problématique. Les solutions développées peuvent être divisées en deux, en fonction du moment où va intervenir la prise en compte de cette flexibilité. Les premières utilisent une structure de la protéine unique comme point de départ et la flexibilité est prise en compte au cours du docking ou après le docking. Les secondes préfèrent réaliser un docking rigide sur un ensemble de conformations de la protéine étudiée ⁴³¹⁻⁴³⁸, dans ce cas la flexibilité est donc traitée avant même le criblage virtuel. Pour réaliser un criblage virtuel prospectif, il est donc possible d'utiliser une structure unique correctement choisie ou un ensemble de conformations. Il n'existe actuellement aucun consensus en faveur de l'une ou l'autre des approches ^{431, 439, 440}. En plus du choix de l'approche à utiliser, la sélection de la(des) structure(s) de départ est essentielle pour la bonne conduite du criblage virtuel ^{363, 364, 367}. Là encore, il n'existe pas de directive claire pour guider cette sélection lorsque plusieurs structures sont disponibles pour une même protéine, et ce, quel que soit l'approche choisie. Nous avons donc décidé, dans cette étude, de rechercher des critères de sélection de la ou des structures de départ pour le docking, Nous avons voulu que ces critères soient simples et ne nécessitent pas de longs temps de calcul pour être définis. Notre choix s'est donc porté vers des critères basés sur les propriétés du site de liaison et notamment l'ouverture et le volume de celui-ci. Pour réaliser cette étude, les jeux de données des ligands et des decoys de la version 2 de la DUD, la banque d'évaluation de référence des méthodes de docking, ont été utilisés. Par contre, nous n'avons pas toujours utilisé les structures des protéines fournies dans

la DUD. Nous avons recherché pour chacune des 40 cibles, toutes les structures existantes dans la PDB. La cible PDGFR beta (Platelet-Derived Growth Factor Receptor beta) a été éliminée puisque la seule structure existante pour celle-ci est un modèle construit par homologie. Pour les 39 cibles restantes, les structures extrêmes en termes de volume et d'ouverture ont été sélectionnées, permettant ainsi d'obtenir de 2 à 4 structures de départ pour le docking avec l'approche structure unique et ensemble docking (Figure 68).

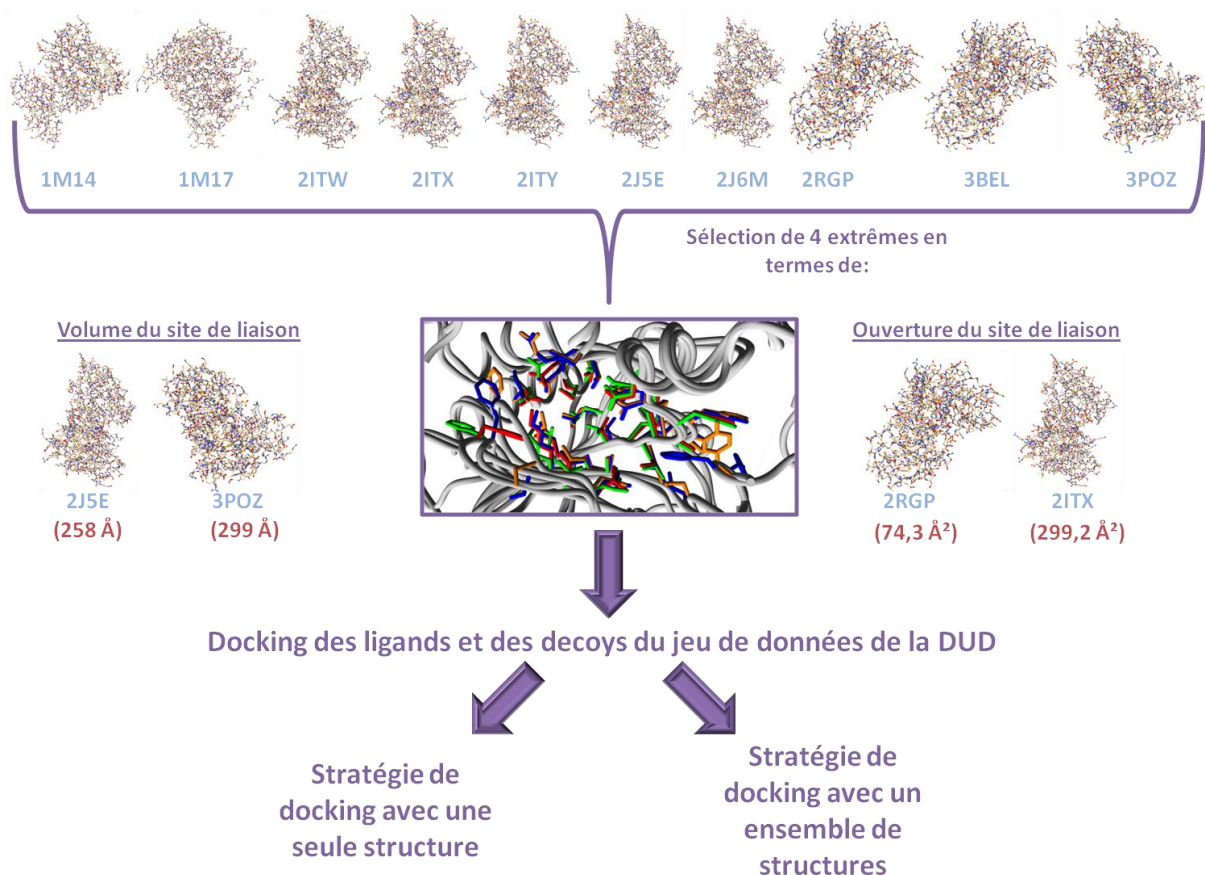


Figure 68. Protocole de sélection des structures de l'EGFR pour l'étude, en fonction du volume et de l'ouverture du site de liaison et utilisation de ces structures pour le docking des ligands et des decoys du jeu de données correspondant de la DUD

Deux logiciels de docking, présentant des caractéristiques différentes en termes de recherche conformationnelle et de fonctions de score, Surflex-Dock et ICM, ont été utilisés pour s'assurer que les tendances observées lors de l'analyse des performances ne sont pas spécifique d'un logiciel et ainsi renforcer la portée de ces résultats. Lors de cette étude nous nous sommes attachés à : (1) confirmer l'impact du choix de la structure sur les résultats du criblage virtuel, (2) identifier des critères de sélection de structures basés sur les propriétés du site de liaison et (3) optimiser la construction du meilleur ensemble possible pour les approches dites d'ensemble docking.

1.1.2 Publication

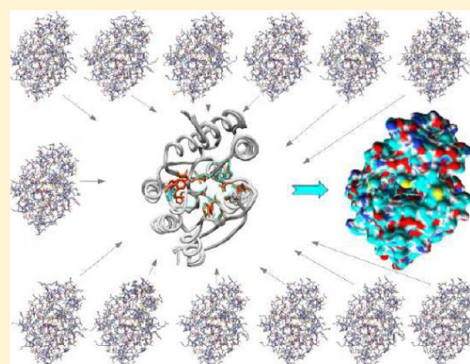
Multiple Structures for Virtual Ligand Screening: Defining Binding Site Properties-Based Criteria to Optimize the Selection of the Query

Nesrine Ben Nasr,[†] H el ene Guillemain,[†] Nathalie Lagarde,[†] Jean-Fran ois Zagury, and Matthieu Montes*

Laboratoire G enomique Bioinformatique et Applications,  Equipe d'accueil EA 4627, Conservatoire National des Arts et M etiers, 292 rue Saint Martin, 75003 Paris, France

Supporting Information

ABSTRACT: Structure based virtual ligand screening (SBVLS) methods are widely used in drug discovery programs. When several structures of the target are available, protocols based either on single structure docking or on ensemble docking can be used. The performance of the methods depends on the structure(s) used as a reference, whose choice requires retrospective enrichment studies on benchmarking databases which consume additional resources. In the present study, we have identified several trends in the properties of the binding sites of the structures that led to the optimal performance in retrospective SBVLS tests whatever the docking program used (Surflex-dock or ICM). By assessing their hydrophobicity and comparing their volume and opening, we show that the selection of optimal structures should be possible with no requirement of prior retrospective enrichment studies. If the mean binding site volume is lower than 350 Å³, the structure with the smaller volume should be preferred. In the other cases, the structure with the largest binding site should be preferred. These optimal structures may be either selected for a single structure docking strategy or an ensemble docking strategy. When constructing an ensemble, the opening of the site might be an interesting criterion additionally to its volume as the most closed structures should not be preferred in the large systems. These "binding site properties-based" guidelines could be helpful to optimize future prospective drug discovery protocols when several structures of the target are available.



INTRODUCTION

In silico screening of compound collections has been extensively used to reduce the number of compounds going into high throughput screening procedures.¹ Different strategies can be used, depending on the data available. If the structure of the target is known, a structure-based virtual ligand screening (SBVLS) protocol that generally includes docking methods can be applied.

Conformational changes occur in protein binding sites upon ligand binding. Besides the classical rigid single structure docking methods, several approaches have been developed in order to handle the plasticity of the binding site during docking like induced-fit docking² or Monte Carlo simulations.³ Another way to handle the flexibility of the target is to perform a rigid receptor docking experiment on multiple structures, i.e. ensemble docking. An ensemble can comprise experimental structures^{4–7} or theoretical structures derived from simulation studies.^{8–11} Compared with the classical single structure docking that consider the binding site as rigid, these approaches are time-consuming and can sometimes increase the rate of false positive results due to scoring errors and/or nonoptimal choice of the structures comprising the ensembles.^{4,5,12–14} Despite the previously described advances, most of the available docking methods handle receptor flexibility implicitly by softening the contact repulsive term in the scoring function

(i.e., soft docking¹³) or by using Gaussian potentials for contacts.¹⁵

With the recent developments of high throughput crystallography, several structures of a given target are available in the Protein Data Bank (PDB).¹⁶ Thus, for a prospective virtual screening, two options are possible: (1) choose the most appropriate structure in the PDB or (2) use an ensemble of structures altogether. Although many studies have already been performed on this matter, there is currently no consensus favoring one approach over the other.^{5,17,18} Different research groups have concluded that an ensemble should comprise a relatively small number of structures to be optimal (up to five depending on the studies).^{4,18–20} Recently, Bottegoni et al.²¹ performed an ensemble docking study on 36 targets from the DUD data set with ICM. They concluded that ensemble docking outperforms average single structure runs and suggested that prior to a prospective SBVLS, preliminary docking studies should be performed to identify the best structures. Similar conclusions arose from very recent studies of Korb et al.²² and Rueda et al.²³ Hence, without preliminary virtual screening studies, it seems, at the moment, very challenging to identify the optimal structure(s) for a "real-life" SBVLS experiment.

Received: September 24, 2012

Published: January 14, 2013

As shown in the literature,^{19,24–26} the choice of the starting conformation for docking studies is critical. We thus decided to search for “binding site properties-based” guidelines to optimize the selection of the conformations that will compose an ensemble when several structures of the target are available. To design our study, we used the targets of the DUD data set for which several experimental structures were available in the PDB (all but PDGFR β). After a careful inspection of the properties of the binding sites of all the conformations available for a given target, we selected up to four “extreme” experimental structures for which the binding site properties in terms of volume and opening were significantly different. For each structure, we evaluated the performance in terms of enrichment of two different SBVLS methods, Surflex-dock (SF) and ICM using the corresponding DUD-own data sets in order to (1) confirm the impact of the choice of the structure on the results of the screening; (2) identify potential trends within the structural properties shared by the structure(s) that resulted into the best enrichments; and (3) reduce the possible bias in the trends that could be observed by using a single docking method. We also assessed the performance of SF and ICM on all the possible ensembles composed by the structures selected for each system in order to optimize the selection of the structures composing the best ensembles and compare their structural properties.

In the present study, we describe our attempt to develop “low-cost” binding site properties-based criteria to identify the optimal structure(s) of the target of a structure-based drug discovery program.

MATERIAL AND METHODS

Directory of Useful Decoys (DUD) Data Set. DUD is a public benchmarking data set designed for docking methods evaluation containing known active compounds for 40 targets, including 36 decoys for each active compound. We selected for each target its corresponding DUD-own data set that comprises only its associated active compounds and decoys. A large random-drug-like data set seeded with known active compounds for each target could have been used to be closer to a prospective VLS, but we decided to use the DUD data set as it is the current standard for fair results comparison between different groups.^{27–29} For our study, the DUD release 2 data set was obtained from the Web site <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 two experimental structures were available; the PDGFR β target was thus excluded. For each target, all the wild-type experimental holo structures available in the Protein Data Bank (PDB) were downloaded; their binding sites were superimposed and visually inspected. After a careful examination of their corresponding binding site properties, we selected up to four “extreme” structures in order to cover the experimentally available flexibility of the binding sites in terms of volume and opening (i.e.: most open, most closed, smallest volume, largest volume). The area of the opening of the binding sites was assessed using CASTp.³⁰ The volume of the binding sites was computed with POVME.³¹ POVME allows an accurate manual edition of the probes selected for the calculation of the volume. Thus, we edited the points manually to cover, for a given target, only the binding site and not its surroundings (Supporting Information Figure S2). The hydrophobicity of the binding site was defined as the ratio of its apolar solvent accessible surface area (ASASA) over its solvent accessible surface area (SASA). ASASA and SASA were

computed with GetArea.³² Hydrogen atoms were added using Chimera.³³

Computational Methods. Surflex-dock. SF is based on a modified Hammerhead fragmentation/reconstruction algorithm to dock compounds flexibly into the binding site.³⁴ 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 are evaluated by an empirical scoring function. For each structure, the binding site has been defined at 4 Å around the cocrystallized ligand for the protomol generation step. In this study, Surflex-dock version 2.5 has been used for all calculations with the options +premin, +remin.

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.³⁵ The docking poses were evaluated using ICM-VLS empirical scoring function.³⁶ The binding site defined for docking has been adjusted to be similar to the Surflex protomol. In this study, ICM version 3.6 has been used for all calculations.

Ensemble Docking. For each target, all possible ensembles (up to eleven for a given system) have been constructed from the experimental structures selected in the PDB. Ensemble docking experiments were performed with Surflex-dock and ICM. For a given compound on a given target, the best score obtained within the different structures composing the ensemble has been retained. Thus, for each ensemble, the ensemble docking result consists in a ranked list of the best scoring poses.

Performance Metrics. All graphics were produced with the statistical and graphical tool R (<http://www.r-project.org/>). The ROCR package was used to plot rate of change (ROC) curves and the Wilcoxon–Mann–Whitney algorithm was used for the ROC area under curve (AUC) calculations.³⁷ Enrichment factors were computed as follows:

$$EF_{x\%} = \frac{N_{\text{experimental}}^{x\%}}{N_{\text{actives}}^{x\%}}$$

$N_{\text{experimental}}^{x\%}$: number of active compounds retrieved in the top $x\%$ of the sorted database. $N_{\text{actives}}^{x\%}$: total number of active compounds in the compound collection.

When comparing the impact of the choice of the structure on enrichment, in the case of similar AUCs, EF1% that reflects early enrichment was considered. Raw tendencies have been compared to an estimation of the random case (Rd) computed as follows:

$$Rd = \sum \frac{t_n}{n^k}$$

t_n : number of targets for which n conformations are compared. n : number of conformations selected for the study for a given target. k : number of docking softwares used (SF and/or ICM).

RESULTS

Presentation of the 39 Targets—Selection of the Different Experimental Conformations. As shown in the literature,^{19,24–26} the choice of the starting conformation for docking studies is of major importance. For each of the 39 targets explored in this study, we examined the different experimental structures available in the PDB (from 2 to 300

Table 1. Physicochemical Properties of the Binding Sites of the Structures Selected for the Study for the (a) Small (i.e., Mean Binding Site Volume under 350 Å³) and “Less-Hydrophobic” (i.e., Mean Percent Hydrophobicity Lower than 70%), (b) Small and “More-Hydrophobic” (i.e., Mean Percent Hydrophobicity over 70%), (c) Large (i.e., Mean Binding Site Volume over 350 Å³) and Less-Hydrophobic, and (d) Large and More-Hydrophobic DUD Targets^a

	mean volume (Å ³)	mean percent hydrophobicity	no. DUD actives	no. available PDB structures	PDB ID	resolution (Å)	RMSD	volume (Å ³)	opening (Å ²)	SASA (Å ²)	ASASA (Å ²)	percent hydrophobicity
(a) small and less-hydrophobic												
NA	243	45.0	49	5	1A4G	2.20		227	12.1	810.94	356.16	43.9
					1A4Q	1.90	0.471	258	118.3	854.53	394.47	46.2
HSP90	337	60.7	37	62	2CDD	1.90		228	33.7	595.76	326.53	54.8
					3K99	2.10	3.279	233	92.4	565.48	353	62.4
					1UYF	2.00	2.299	549	70.3	775.09	503.33	64.9
PNP	159	60.7	50	7	1B8O	1.50		98	0	398.64	211.62	53.1
					2QPPL	2.10	0.579	148	52.8	824.75	513.52	62.3
					1V48	2.20	0.734	230	27.3	886.51	591.93	66.8
TRP	138	63.6	49	300	1BTY	1.50		60	2.9	441.9	272.75	61.7
					3AAU	1.80	0.196	171	12.1	773.09	484.91	62.7
					1V2O	1.62	0.951	183	8.9	810.34	538.76	66.5
VEGFR2	322	63.8	88	3	2XIR	1.50		221	119.1	561.89	387.11	68.9
					1VR2	2.40	2.864	422	253.5	658.96	386.68	58.7
EGFR	279	64.5	475	10	2JSE	3.10		258	140.7	590	381.75	64.7
					2ITX	2.98	0.794	278	299.2	598.66	390.75	65.3
					2RGP	2.00	2.333	282	74.3	466.08	314.1	67.4
					3POZ	1.50	2.404	299	176.3	462.56	281.29	60.8
SAHH	259	66.3	33	2	1A7A	2.80		177	0	522.19	329.13	63.0
					1LI4	2.01	0.349	341	189.2	363.05	252.56	69.6
TK	146	67.4	22	4	1E2P	2.50		114	0	275.68	200.36	72.7
					1E2I	1.90	7.991	128	86.1	467.39	306.21	65.5
					1E2N	2.20	2.568	195	0	510.22	326.11	63.9
(b) small and more-hydrophobic												
COX2	264	71.0	426	6	3PGH	2.50		248	102.3	757.73	523.27	69.1
					1CX2	3.00	0.832	264	375.6	802.35	562.03	70.1
					1DDX	3.00	1.221	280	371.3	893.57	660.19	73.9
GART	320	71.0	40	6	1C2T	2.10		269	54.6	797.69	561.75	70.4
					1GAR	1.96	2.561	293	39.5	1610.82	1163.58	72.2
					1JKX	1.60	2.524	399	126.4	1095.99	771.79	70.4
PR	302	72.7	27	12	1A28	1.80		207	0	395.43	328.24	83.0
					2OVH	2.00	5.861	396	130.6	793.22	495.17	62.4
COMT	171	73.5	11	12	2CL5	1.60		84	20.5	1046.81	755.33	72.2
					1H1D	2.00	3.456	157	10.7	1028.2	749.55	72.9
					3A7D	2.40	0.905	272	27.0	1132.68	854.69	75.5
AR	280	74.5	79	48	1T7T	1.70		240	0	376.4	284.97	75.7
					3B66	1.65	1.006	320	0	404.38	296.45	73.3
GR	315	75.3	78	7	1M2Z	2.50		228	0.65	387.61	299.76	77.3
					3CLD	2.84	1.325	262	0	413.39	323.81	78.3
					1NHZ	2.30	1.524	317	647.7	703.95	494.52	70.3
					3K22	2.10	1.059	452	5.7	493.64	371.79	75.3
MR	173	76.5	15	3	2A3I	1.95		172	0	622.61	472.22	75.9
					2AA2	1.95	0.798	174	0	713.29	549.85	77.1
COX1	243	77.8	25	22	1PTH	3.40		200	336.9	347.91	269.87	77.6
					3KK6	2.75	5.736	256	370.7	431.26	334.67	77.6
					1PGG	4.50	0.527	272	327.2	400.29	312.88	78.2
HIVRT	182	80.1	43	52	2JLE	2.90		133	1742.4	547.7	410.53	74.9
					1DTQ	2.80	1.937	177	1613.3	549	478.49	87.2
					3LAM	2.76	1.751	236	1910.6	774.32	604.6	78.1
RXR	253	80.1	20	17	1FBY	2.25		228	0	649.47	509.52	78.5
					3DZY	3.10	1.207	278	0	649.62	531.41	81.8
ER ago	259	84.9	67	10	3ERD	2.03		245	0	291.99	248.15	85.0
					2P15	1.94	0.972	272	0	458.98	389.22	84.8
INHA	331	86.4	86	25	2X22	2.10		147	143.2	751.32	653.26	86.9
					1P44	2.60	3.275	515	91.4	895.1	769.33	85.9
(c) large and less-hydrophobic												
AMPC	664	58.6	21	44	1FSW	1.90		592	92.8	748.23	397.21	53.1
					2HDR	2.20	0.527	672	238.8	866.63	527.39	60.9

Table 1. continued

	mean volume (Å ³)	mean percent hydrophobicity	no. DUD actives	no. available PDB structures	PDB ID	resolution (Å)	RMSD	volume (Å ³)	opening (Å ²)	SASA (Å ²)	ASASA (Å ²)	percent hydrophobicity
(c) large and less-hydrophobic												
FGFR1	632	59.7	120	4	1XGJ	1.97	2.722	728	226.9	929.97	573.72	61.7
					1FGI	2.40		355	145.7	809.6	462.71	57.2
					2FGI	2.50	0.511	909	213.1	1104.25	685.99	62.1
HIVPR	513	59.9	62	87	1NH0	1.03		451	37.4	1003.31	612.23	61.0
					2PWC	1.78	0.761	465	22.6	880.66	523.98	59.5
					1XL2	1.50	0.322	642	50.7	932.73	552.85	59.3
EXA	471	60.0	146	100	2RA0	2.30		415	10.8	797.61	460.25	57.7
					2D1J	2.20	0.829	422	33.8	827.86	492.57	59.5
					1IQF	3.20	1.054	577	7.0	879.35	551.15	62.7
ACE	595	60.4	49	4	3BKL	2.18		448	86.4	657.23	407.95	62.1
					2OC2	2.25	0.323	664	81.9	697.22	415.67	59.6
					1O86	2.00	0.275	672	135.5	684.48	407.95	59.6
SRC	515	60.6	159	4	2H8H	2.20		278	64.5	842.3	523.48	62.1
					1Y57	1.91	2.017	698	771.2	864.17	510.5	59.1
GPB	366	61.3	52	52	6GPB	2.86		174	24.0	269.17	181.18	67.3
					1A8I	1.78	1.281	293	0	335.56	198.76	59.2
					1K06	1.80	1.772	630	96.3	414	237.28	57.3
HMGR	928.5	67.0	35	8	1DQA	2.00		737	62.7	842.41	576.33	68.4
					1HWJ	2.26	0.507	1120	139.9	1157.2	757.43	65.5
					2C5F	2.60		378	16.4	569.06	392.2	68.9
ACHE	516	68.7	107	75	2CEK	2.20	0.685	558	32.1	717.6	493.22	68.7
					1OCE	2.70	0.711	612	29.8	627.45	430.46	68.6
					1DOJ	1.70		568	0.1	696.97	470.88	67.6
THR	608	68.7	72	4	1JWT	2.50	13.717	648	25.51	782.92	547.52	69.9
					1EFY	2.20		506	90.2	640.37	450.63	70.4
					4PAX	2.80	0.637	509	31.3	664.49	459.34	69.1
(d) large and more-hydrophobic												
CDK2	456	70.0	72	143	3PY1	2.05		304	62.2	675.73	497.63	73.6
					2I40	2.8	2.6	504	131.3	999.26	635.47	63.6
					2B55	1.85	1.792	560	81.5	972.25	708.62	72.9
P38	367	71.1	454	96	3BX5	2.4		185	7.5	542.64	373.81	68.9
					3FSF	2.1	1.221	442	388.7	848.18	581.67	68.6
					1KV2	2.8	2.559	473	297.7	974.64	737.97	75.7
ADA	507	72.3	39	12	1NDW	2		360	15.2	544.4	382.81	70.3
					1NDV	2.3	1.357	575	20.9	770.89	570.77	74.0
					2E1W	2.5	1.424	585	20.0	796.15	578.72	72.7
PPAR	474	74.4	85	83	3OSW	2.55		332	57.0	873.94	639.04	73.1
					1WM0	2.9	1.272	380	1.5	716.24	515.46	72.0
					1FM9	2.1	1.698	710	74.6	899.83	702.78	78.1
DHFR	723	74.7	410	7	3DFR	1.7		542	16.2	696.07	544.38	78.2
					1BZF	NMR_19 ^b	1.185	904	111.6	761.21	541.65	71.2
					1TBF	1.3		519	55.2	1039.01	789.05	75.9
PDE5	530	75.3	88	3	1XOZ	1.37	0.999	540	30.6	1045.72	780.25	74.6
					1AH3	2.3		272	16.7	624.68	493.17	78.9
ALR2	356	76.2	26	5	1EK0	1.48	0.618	440	34.3	616.48	452.66	73.4
					3ERT	1.9		358	81.8	794.89	599.04	75.4
					2IOG	1.6	2.025	483	4.5	785.79	611.62	77.8
ER antago	421	76.6	39	18								

^aBinding site RMSD has been computed using ICM on all heavy atoms towards the coordinates of the smallest conformer. Binding site volumes and opening were computed using respectively POVME and CASTp. ASASA and SASA were calculated with GetArea. Percent hydrophobicity is the ratio of ASASA over SASA. ^bFor 1BZF, the conformation selected for the study is the 19th model from the NMR structure.

depending on the system). We selected up to four structures (2 for 18 targets, 3 for 19 targets, and 4 for 2 targets) in order to cover the experimentally available flexibility of the binding sites in terms of volume and opening. For a given structure, the volume of the binding site was calculated with POVME and its opening was computed with CASTp. The mean volume value among the conformers selected for the study, 350 Å³, was used as a cutoff to classify the targets in two equilibrated groups of 20 small targets and 19 large targets. Similarly, we split the 39

targets in two equilibrated groups of 20 more-hydrophobic targets harboring a cutoff of more than 70% of the solvent accessible surface area of the binding site being apolar, and of 19 less-hydrophobic targets. For four small and more-hydrophobic targets (AR, ER ago, MR, and RXR), the opening of the binding site was very limited and did not vary significantly between the structures available in the PDB. They were considered as closed and invariable in terms of binding site opening. The binding sites properties (volume,

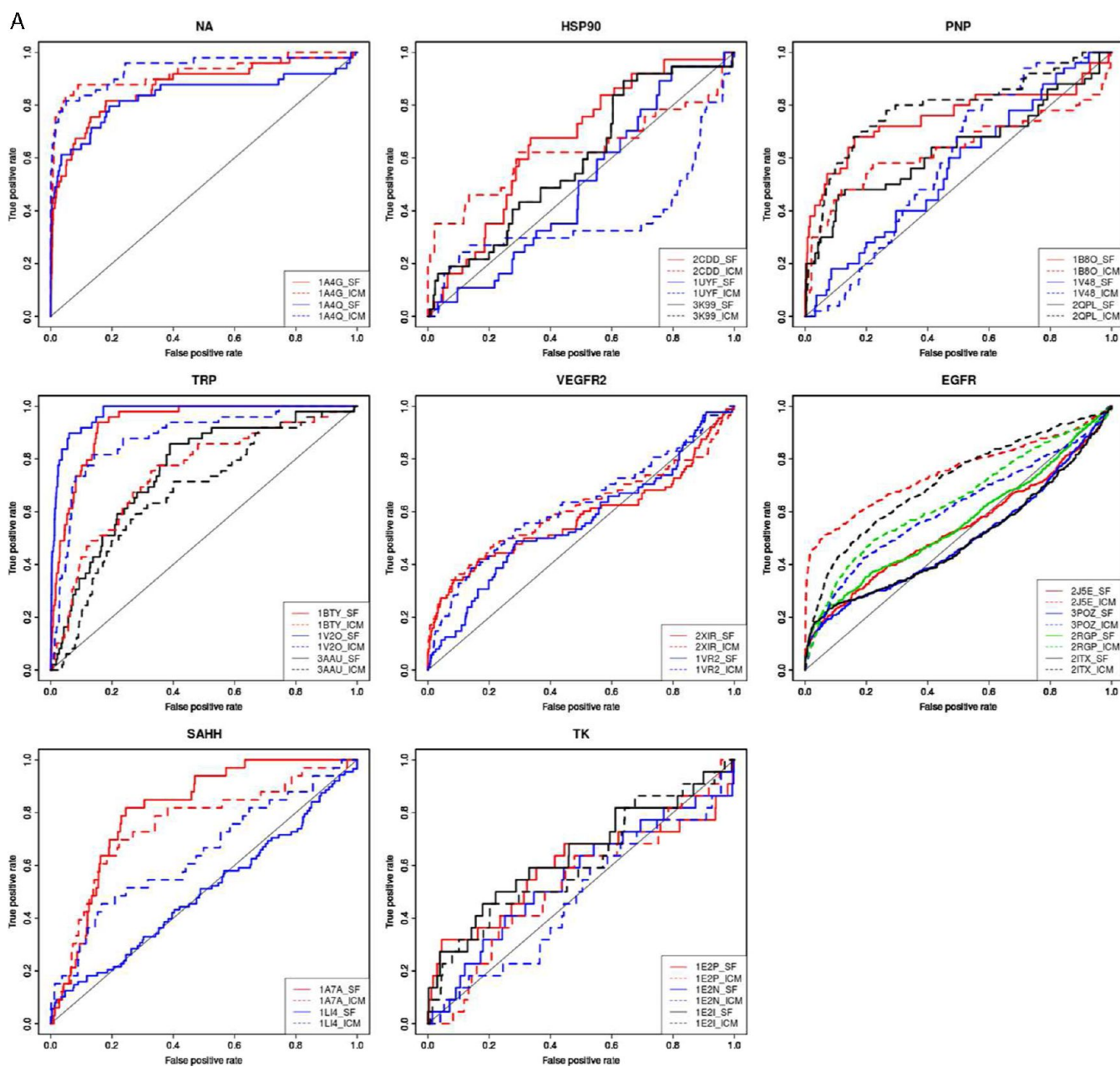


Figure 1. continued

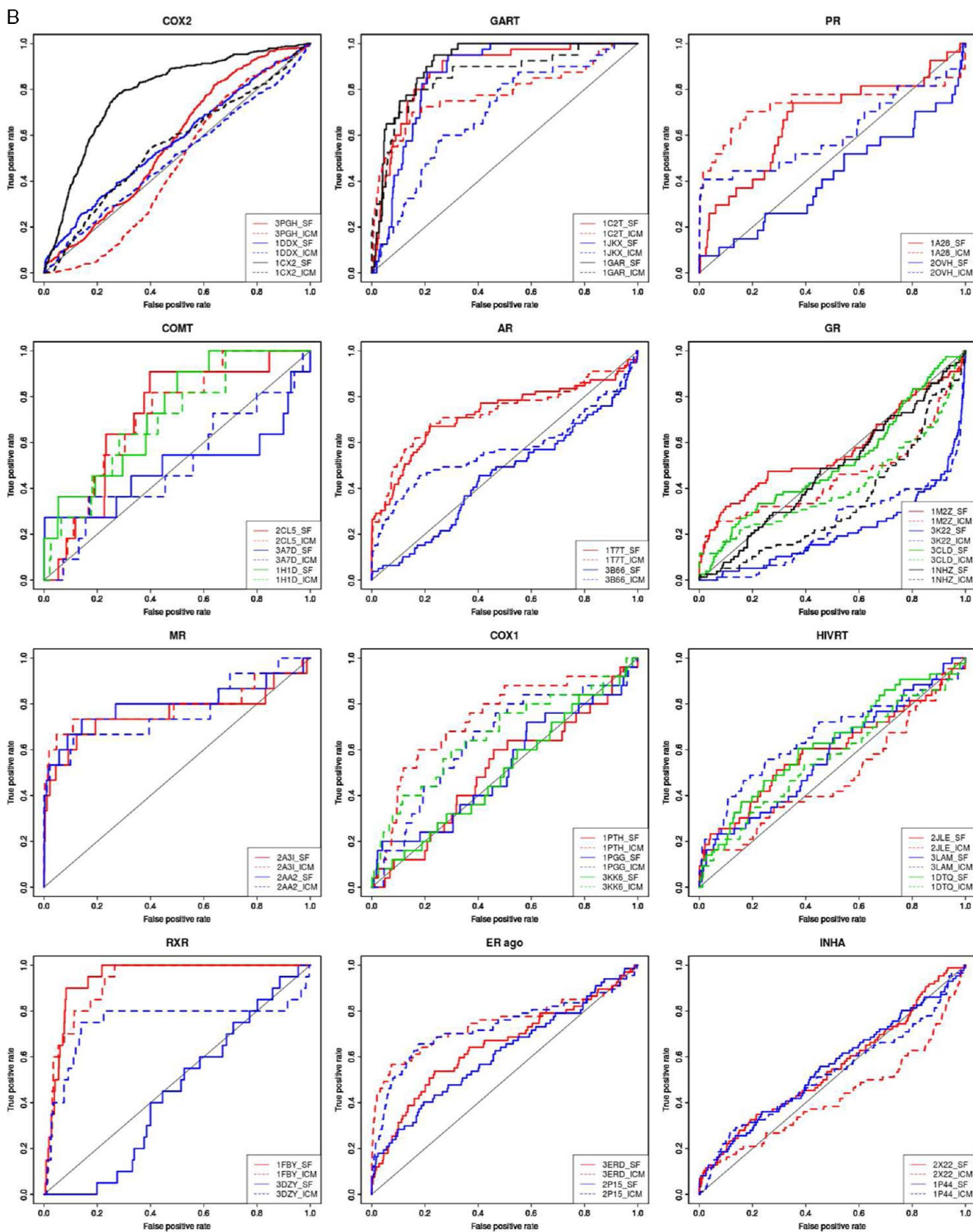


Figure 1. continued

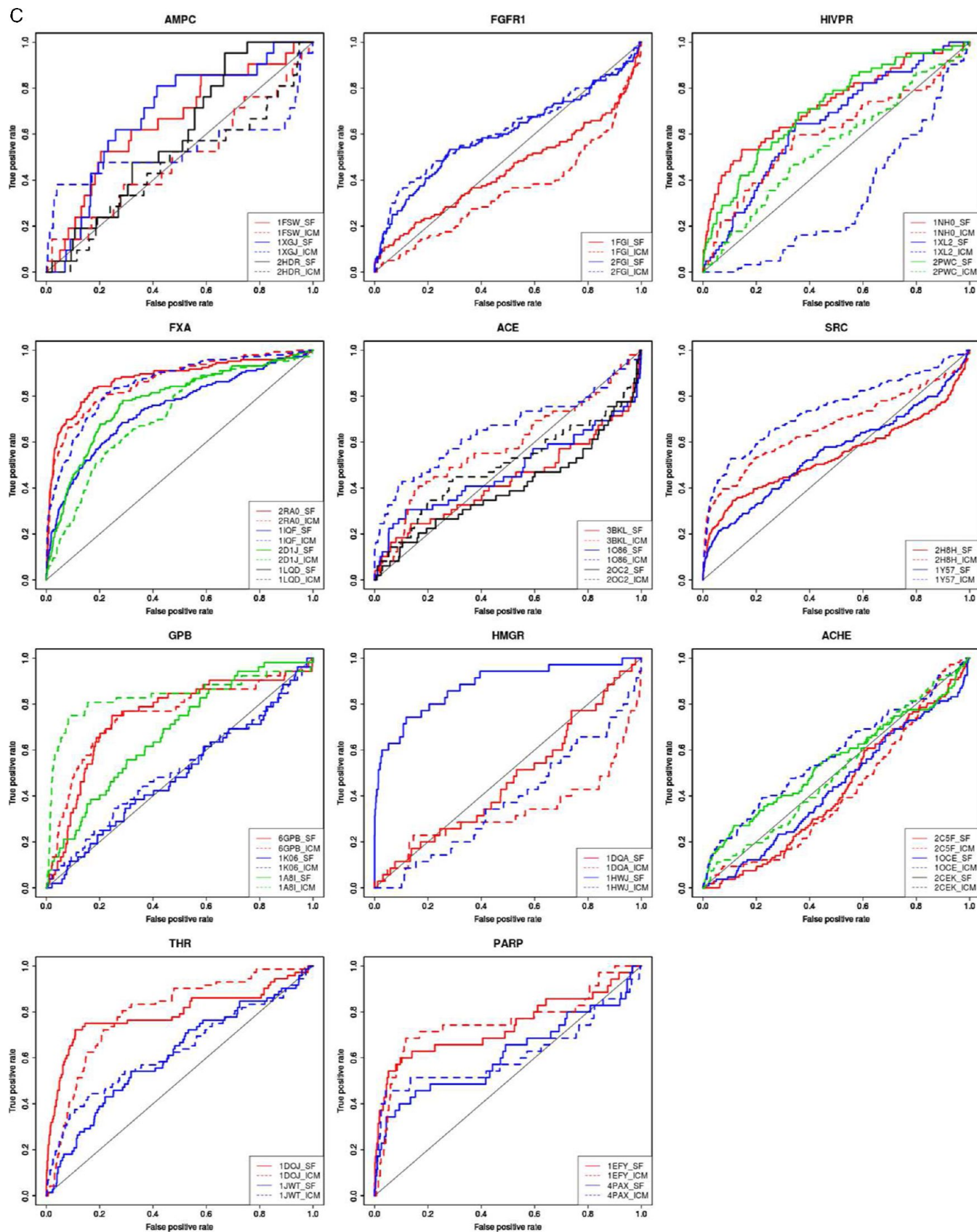


Figure 1. continued

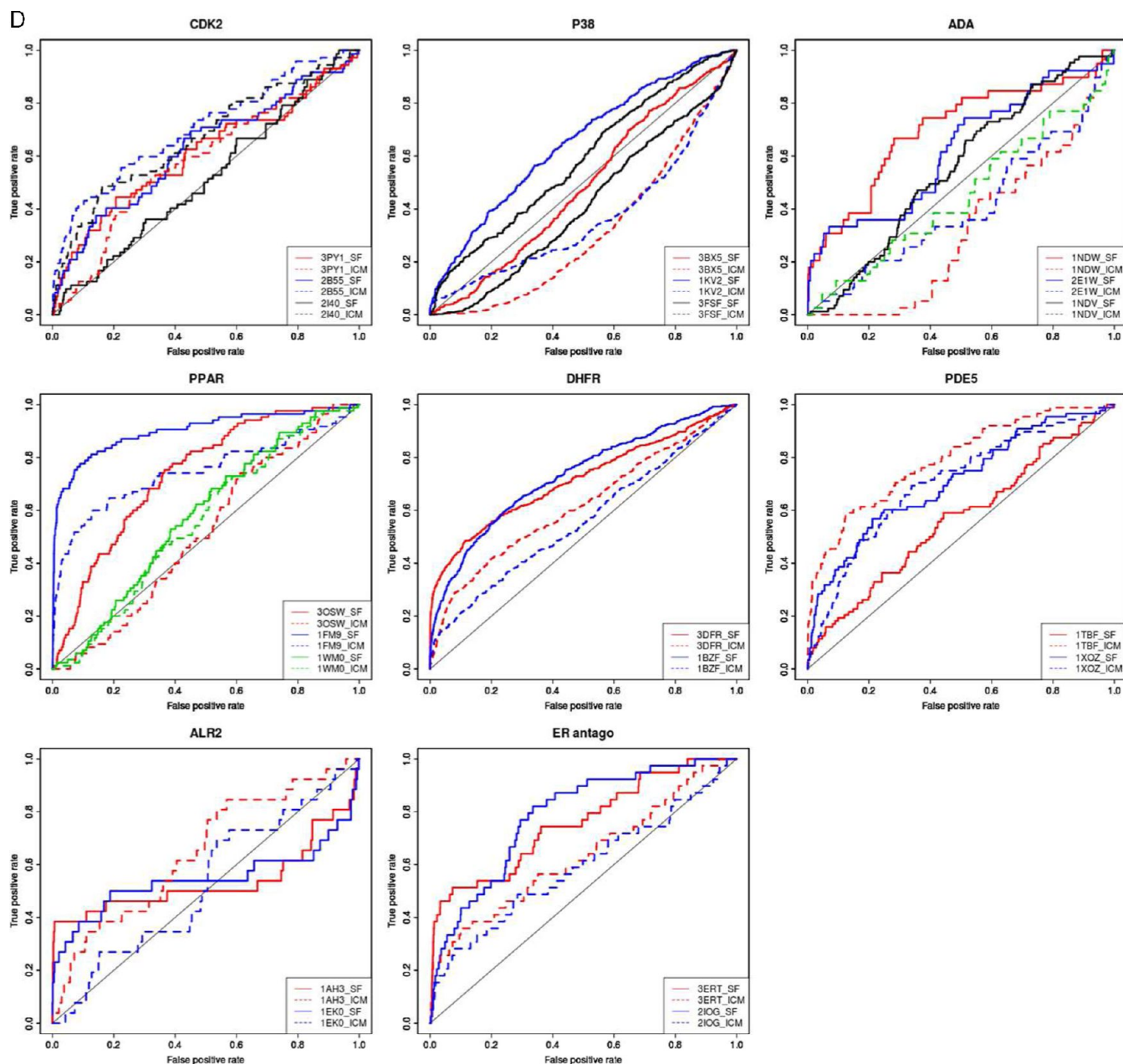


Figure 1. ROC curves with single structure docking protocols on the (A) small and less-hydrophobic, (B) small and more-hydrophobic, (C) large and less-hydrophobic, and (D) large and more-hydrophobic DUD systems using SF (plain lines) and ICM (dotted lines). For a given target, the ROC curves associated to the “extreme” structures in terms of binding site properties are represented in red for the smallest binding site, blue for the largest, green for the most closed and black for the most open. When the structure was considered as extreme in two categories, the volume of the binding site has been prioritized for the color of the ROC curve.

opening, and hydrophobicity) of the structures selected for the 39 DUD targets and their corresponding DUD-own sets are summarized in Table 1.

Single Structure Docking Strategy. The first part of our study was to identify possible trends in the structural properties of the conformations used as a reference for docking that gave the best results in terms of enrichment. We analyzed the performance of two different docking methods, SF and ICM, on all the conformers selected for this study taken separately and compared the structural properties of the structures that led to the best enrichments. The results are presented using ROC curves in Figure 1 and ROC AUC and enrichment factors in Table 2

Overall, for 20 out of the 39 systems used in this study, the conformer that resulted into the best performances was common to SF and ICM ($R_d = 6.74/39$). Within these 20 systems, in 6 out of the 10 systems for which the pocket volume was inferior to 350 \AA^3 (small systems) the optimal conformer was the smallest ($R_d = 3.5/10$). This trend was confirmed within the seven small and more-hydrophobic systems, for which the conformer displaying the best performances was the smallest in five cases ($R_d = 3.2/7$).

In the systems for which the pocket volume was superior to 350 \AA^3 (large systems), the conformer that gave the best performances in enrichment was the largest for 7 systems out of 10 ($R_d = 4.2/10$).

Table 2. Volume, Binding Site Opening, Area under the ROC Curve (AUC), Enrichment Factors at 1% and 10%, and Average Calculation Time Per Compound (CPU) Using Single Structure Docking Protocols with Surflex-dock (SF) or ICM with the Different Structures Selected for the Study on the (a) Small and Less-Hydrophobic, (b) Small and More-Hydrophobic, (c) Large and Less-Hydrophobic, and (d) Large and More-Hydrophobic Data Sets

	PDB ID	SF				ICM				volume	opening	
		AUC	EF1%	EF10%	CPU	PDB ID	AUC	EF1%	EF10%			CPU
(a) small and less-hydrophobic systems												
NA	1A4G	0.869	24.79	6.75	11.60	1A4G	0.923	26.85	8.58	17.80	227	12.1
	1A4Q	0.836	30.98	6.34	13.79	1A4Q	0.938	39.24	8.18	22.73	258	118.3
HSP90	2CDD	0.665	0.00	1.63	14.53	2CDD	0.627	19.22	3.53	14.77	228	33.7
	3K99	0.593	2.75	1.90	17.66	3K99	0.544	13.73	2.99	16.10	233	92.4
	1UYF	0.498	2.75	1.09	13.76	1UYF	0.631	0.00	1.90	14.69	549	70.3
PNP	1B8O	0.755	15.20	5.43	7.60	1B8O	0.635	8.69	4.02	9.90	98	0
	2QPL	0.638	15.20	3.02	5.64	2QPL	0.788	10.86	5.23	7.19	148	52.8
	1V48	0.561	0.00	1.81	7.07	1V48	0.590	0.00	0.40	7.16	230	27.3
TRP	1BTY	0.933	16.45	7.16	27.92	1BTY	0.747	2.06	3.68	47.07	60	2.9
	3AAU	0.755	2.06	3.27	25.81	3AAU	0.675	0.00	2.04	45.25	171	12.1
	1V2O	0.973	26.73	9.00	26.76	1V2O	0.882	4.11	7.16	50.71	183	8.9
VEGFR2	2XIR	0.576	11.73	3.41	18.08	2XIR	0.609	14.08	3.53	15.76	221	119.1
	1VR2	0.560	4.69	1.48	10.00	1VR2	0.625	3.52	2.84	17.40	422	253.5
EGFR	2JSE	0.536	8.03	2.23	9.70	2JSE	0.750	26.01	5.22	11.60	258	140.7
	2ITX	0.493	9.30	2.44	10.02	2ITX	0.713	7.40	4.00	12.25	278	299.2
	2RGP	0.558	8.46	2.48	11.88	2RGP	0.641	5.50	3.14	10.93	282	74.3
	3POZ	0.494	8.88	2.08	11.88	3POZ	0.609	4.02	2.82	13.12	299	176.3
SAHH	1A7A	0.811	0.00	3.05	8.90	1A7A	0.751	0.00	3.97	11.30	177	0
	1LI4	0.497	5.87	1.59	13.00	1LI4	0.643	9.64	3.05	10.00	341	189.2
TK	1E2P	0.593	9.22	3.19	8.87	1E2P	0.534	0.00	0.46	6.52	114	0
	1E2I	0.643	13.83	2.74	8.28	1E2I	0.607	4.61	2.74	6.80	128	86.1
	1E2N	0.555	0.00	1.37	7.49	1E2N	0.478	0.00	0.91	6.42	195	0
(b) small and more-hydrophobic systems												
COX2	3PGH	0.573	3.77	1.15	11.56	3PGH	0.530	0.00	0.16	12.58	248	102.3
	1CX2	0.785	3.76	3.50	12.20	1CX2	0.560	0.24	1.29	11.16	264	375.6
	1DDX	0.574	5.87	1.71	11.80	1DDX	0.500	0.71	1.41	12.21	280	371.3
GART	1C2T	0.880	2.55	5.55	25.60	1C2T	0.780	10.21	5.30	40.15	269	54.6
	1GAR	0.916	2.55	6.56	21.87	1GAR	0.861	17.87	6.06	43.12	293	39.5
	1JKX	0.861	0.00	4.29	23.31	1JKX	0.692	2.55	2.02	46.91	399	126.4
PR	1A28	0.667	7.91	2.99	8.80	1A28	0.738	31.64	5.22	8.80	207	0
	2OVH	0.419	7.91	1.12	8.31	2OVH	0.609	27.69	4.10	9.73	396	130.6
COMT	2CL5	0.716	0.00	1.85	4.26	2CL5	0.712	0.00	1.85	6.33	84	20.5
	1H1D	0.733	21.77	3.71	4.50	1H1D	0.700	0.00	2.78	5.99	157	10.7
	3A7D	0.491	32.66	2.78	6.39	3A7D	0.501	0.00	0.93	5.84	272	27.0
AR	1T7T	0.726	26.88	4.18	1.43	1T7T	0.731	21.76	4.94	7.98	240	0
	3B66	0.452	3.84	0.63	1.92	3B66	0.572	12.80	3.42	7.93	320	0
GR	1M2Z	0.564	9.05	3.08	12.60	1M2Z	0.450	10.34	2.57	11.70	228	0.65
	3CLD	0.511	2.59	1.67	13.51	3CLD	0.591	11.63	1.67	12.96	262	0
	1NHZ	0.478	1.29	0.77	11.30	1NHZ	0.340	2.59	0.51	11.72	317	647.7
	3K22	0.209	0.00	0.39	14.10	3K22	0.230	0.00	0.13	11.49	452	5.7
MR	2A3I	0.757	28.93	5.34	12.40	2A3I	0.788	36.17	6.68	12.90	172	0
	2AA2	0.796	36.17	6.00	13.90	2AA2	0.767	43.40	5.34	13.10	174	0
COX1	1PTH	0.507	0.00	1.21	8.72	1PTH	0.734	0.00	3.22	8.08	200	336.9
	3KK6	0.503	4.16	1.21	8.10	3KK6	0.644	0.00	1.61	9.43	256	370.7
	1PGG	0.521	4.16	2.01	8.53	1PGG	0.670	4.16	3.22	8.03	272	327.2
HIVRT	2JLE	0.593	9.69	2.56	11.22	2JLE	0.515	9.69	1.63	12.45	133	1742.4
	1DTQ	0.628	7.27	1.86	10.95	1DTQ	0.559	0.00	1.86	12.20	177	1613.3
	3LAM	0.578	2.42	2.33	10.45	3LAM	0.657	7.27	3.03	11.62	236	1910.6
RXR	1FBY	0.941	5.50	8.50	26.49	1FBY	0.922	5.50	7.00	27.00	228	0
	3DZY	0.444	0.00	0.00	16.29	3DZY	0.750	0.00	5.00	27.28	278	0
ER ago	3ERD	0.646	7.57	2.69	8.10	3ERD	0.746	22.71	5.69	8.51	245	0
	2P15	0.613	9.08	2.54	7.00	2P15	0.730	12.11	5.24	7.36	272	0
INHA	2X22	0.550	4.72	1.86	13.84	2X22	0.575	8.27	1.51	11.77	515	143.2
	1P44	0.550	4.07	2.57	10.90	1P44	0.525	1.16	2.33	12.80	533	91.4

Table 2. continued

	PDB ID	SF				PDB ID	ICM				volume	opening
		AUC	EF1%	EF10%	CPU		AUC	EF1%	EF10%	CPU		
(c) large and less-hydrophobic systems												
AMPC	1FSW	0.656	0.00	1.92	6.60	1FSW	0.490	4.80	1.44	6.74	592	92.8
	2HDR	0.593	0.00	0.96	5.72	2HDR	0.530	0.00	0.48	7.22	672	238.8
FGFR1	1XGJ	0.688	0.00	1.44	5.70	1XGJ	0.530	14.41	3.84	6.62	728	226.9
	2FGI	0.444	5.08	1.58	19.45	2FGI	0.400	2.54	1.33	23.08	355	145.7
HIVPR	1NH0	0.728	9.68	4.35	35.14	1NH0	0.602	3.23	2.26	55.71	451	37.4
	2PWC	0.713	10.16	2.92	34.09	2PWC	0.550	1.61	1.45	55.21	465	22.6
FXA	1XL2	0.649	3.23	1.61	35.51	1XL2	0.685	0.00	0.00	58.95	642	50.7
	2RA0	0.877	18.09	6.99	21.87	2RA0	0.865	14.61	6.65	42.04	415	10.8
ACE	2D1J	0.778	8.35	4.25	22.04	2D1J	0.714	5.57	2.47	37.80	422	33.8
	1IQF	0.743	9.74	4.18	22.14	1IQF	0.855	22.96	5.41	37.79	577	7.0
SRC	3BKL	0.434	4.19	1.84	15.00	3BKL	0.580	2.09	1.64	18.98	448	86.4
	2OC2	0.402	2.09	1.43	13.39	2OC2	0.502	4.19	1.64	18.84	664	81.9
GPB	1O86	0.476	2.09	2.66	12.90	1O86	0.660	16.74	4.09	17.23	672	135.5
	2H8H	0.542	8.91	3.15	21.21	2H8H	0.668	14.64	3.97	19.14	278	64.5
HMGR	1Y57	0.555	8.91	2.33	15.94	1Y57	0.742	12.10	4.72	19.47	698	771.2
	6GPB	0.747	2.00	3.27	11.00	6GPB	0.752	4.01	4.43	17.56	174	24.0
ACHE	1A8I	0.674	4.01	2.12	9.60	1A8I	0.835	8.03	7.51	13.64	293	0
	1K06	0.492	0.00	1.15	9.60	1K06	0.513	2.01	1.35	14.44	630	96.3
THR	1DQA	0.470	0.00	1.15	25.11	1DQA	0.332	2.89	0.86	36.76	737	62.7
	1HWJ	0.880	34.63	6.31	20.40	1HWJ	0.381	0.00	0.29	35.99	1120	337.9
PARP	2CSF	0.424	0.00	0.37	15.66	2CSF	0.582	0.00	0.94	20.38	378	16.4
	2CEK	0.549	3.83	2.15	16.68	2CEK	0.505	0.96	1.22	21.76	558	32.1
ER antago	1OCE	0.436	0.96	0.47	14.96	1OCE	0.581	3.83	2.15	22.85	612	29.8
	1DOJ	0.793	15.45	6.27	31.28	1DOJ	0.797	1.40	4.18	41.65	568	0.1
CDK2	1JWT	0.611	1.40	1.95	28.81	1JWT	0.631	7.02	3.62	47.44	648	25.51
	1EFY	0.738	18.28	5.74	4.70	1EFY	0.756	3.05	5.74	5.10	506	90.2
P38	4PAX	0.615	9.14	3.73	4.80	4PAX	0.615	12.18	4.59	5.82	509	31.3
	3PY1	0.608	2.84	2.65	15.38	3PY1	0.583	1.42	1.25	14.54	304	62.2
ADA	2I40	0.515	0.00	1.11	11.66	2I40	0.669	5.68	3.34	14.48	504	131.3
	2B55	0.626	5.68	2.37	11.52	2B55	0.705	14.19	4.18	14.52	560	81.5
ER antago	3BX5	0.488	0.67	0.60	13.75	3BX5	0.318	0.00	0.04	11.38	185	7.5
	3FSF	0.585	1.78	1.90	12.71	3FSF	0.414	0.00	0.13	11.00	442	388.7
PPAR	1KV2	0.650	5.12	2.20	11.00	1KV2	0.367	4.45	0.95	10.80	473	297.7
	1NDW	0.700	16.51	3.10	8.80	1NDW	0.320	0.00	0.00	10.77	360	15.2
DHFR	1NDV	0.580	5.50	1.81	8.07	1NDV	0.549	0.00	1.29	11.24	575	20.9
	2E1W	0.624	13.76	3.35	7.70	2E1W	0.595	0.00	0.77	11.05	585	20.0
ALR2	3OSW	0.737	1.18	3.06	37.98	3OSW	0.516	0.00	0.59	54.50	332	57.0
	1WM0	0.579	0.00	0.71	41.73	1WM0	0.556	1.18	0.59	57.12	380	1.5
PDE5	1FM9	0.901	27.16	7.65	37.90	1FM9	0.748	16.53	5.18	51.10	710	74.6
	3DFR	0.716	18.45	4.32	12.50	3DFR	0.620	4.19	2.94	15.36	542	16.2
ER antago	1BZF	0.731	11.32	3.59	11.01	1BZF	0.552	0.49	0.71	15.82	904	111.6
	1TBF	0.565	4.70	1.82	20.56	1TBF	0.791	17.61	4.56	28.87	519	55.2
ER antago	1XOZ	0.707	5.87	3.53	23.06	1XOZ	0.705	4.70	2.62	26.07	540	30.6
	1AH3	0.537	27.49	3.85	5.20	1AH3	0.650	3.93	2.69	6.00	272	16.7
ER antago	1EK0	0.545	19.63	3.85	4.88	1EK0	0.525	0.00	0.77	6.47	440	34.3
	3ERT	0.757	16.34	5.15	25.50	3ERT	0.630	13.62	3.09	32.00	358	81.8
ER antago	2IOG	0.785	5.45	3.86	25.62	2IOG	0.591	5.45	2.83	34.53	483	4.5

Overall, in the small systems, the conformer that gave the worst enrichment was the one displaying the largest binding site for 15 out of 20 targets with SF ($R_d = 8/20$) and for 14 out of 20 targets with ICM ($R_d = 8/20$).

There was a noticeable difference in the performance of SF and ICM depending on the opening of the conformation chosen for docking. Using SF, the strongest signal was obtained with small targets for which the most closed conformation

resulted in the best performance (10 out of 16; $R_d = 6/16$) whereas the most open conformation resulted in the worst performance (10 out of 16; $R_d = 6/16$). Using ICM, the most closed conformations of the large targets were associated with the worst performances (11 out of 19; $R_d = 7.75/19$).

Ensemble Docking Strategy. Ensemble docking can be an alternative when different structures are available. Thus, for each target, we constructed all possible ensembles with the

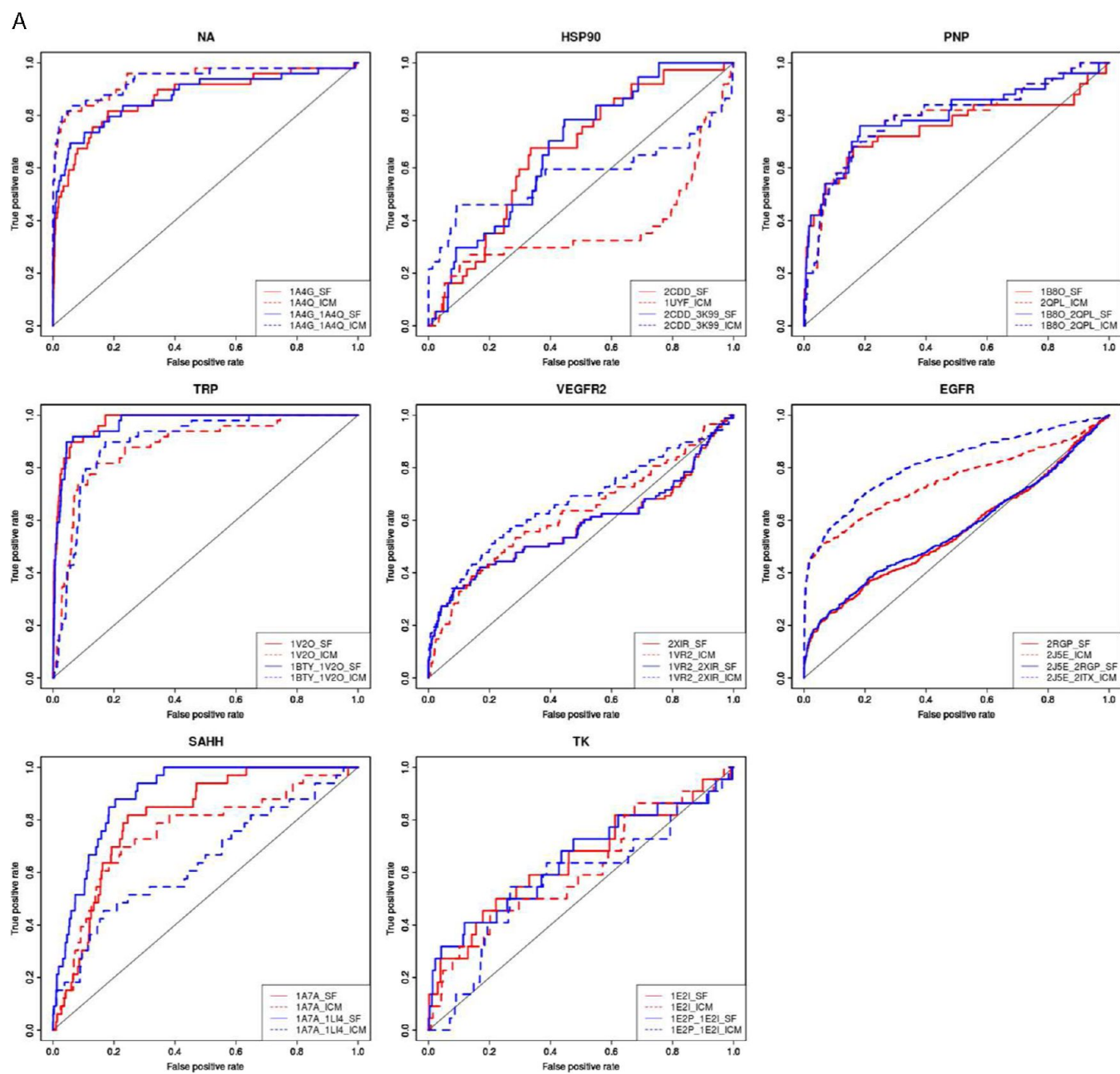


Figure 2. continued

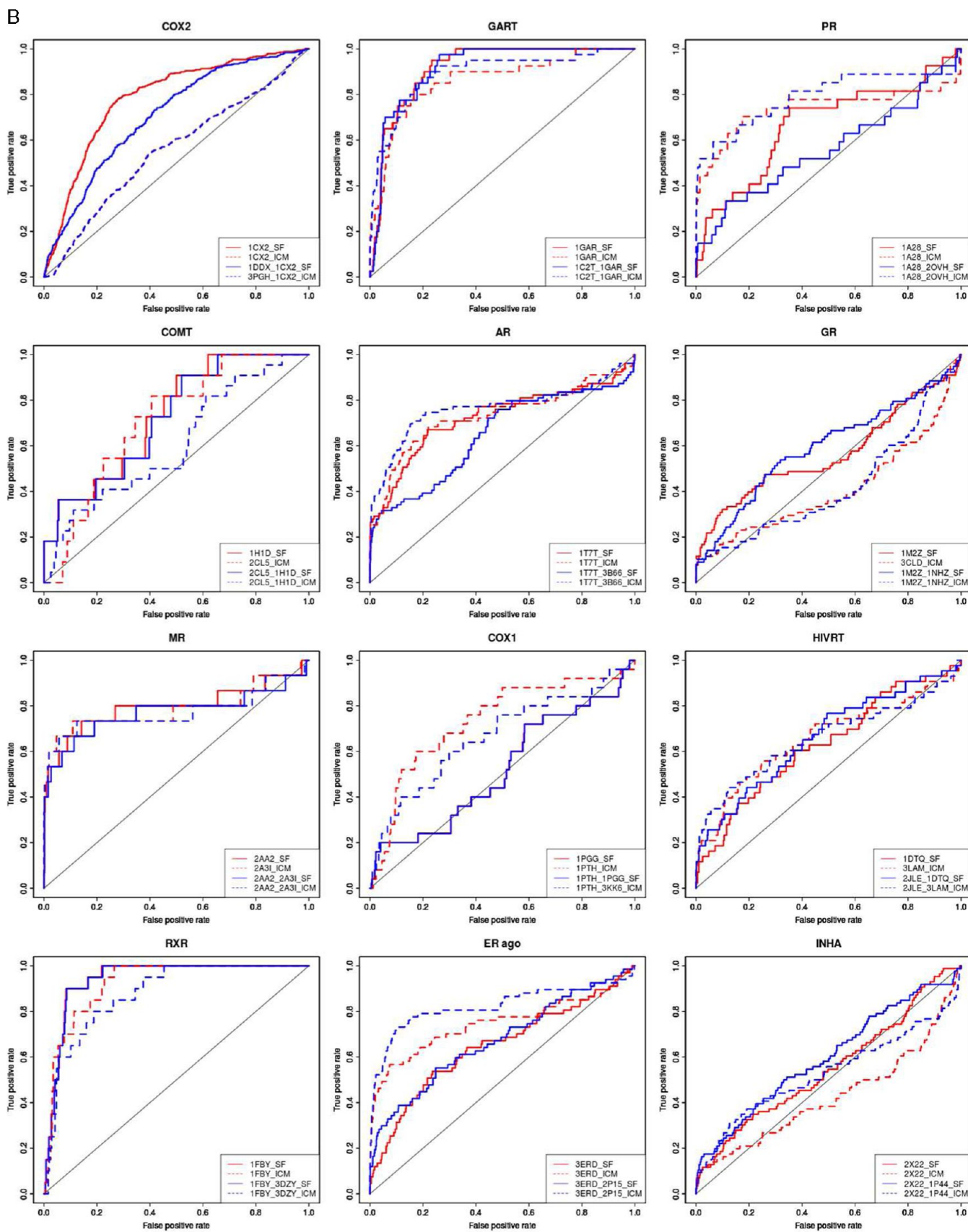


Figure 2. continued

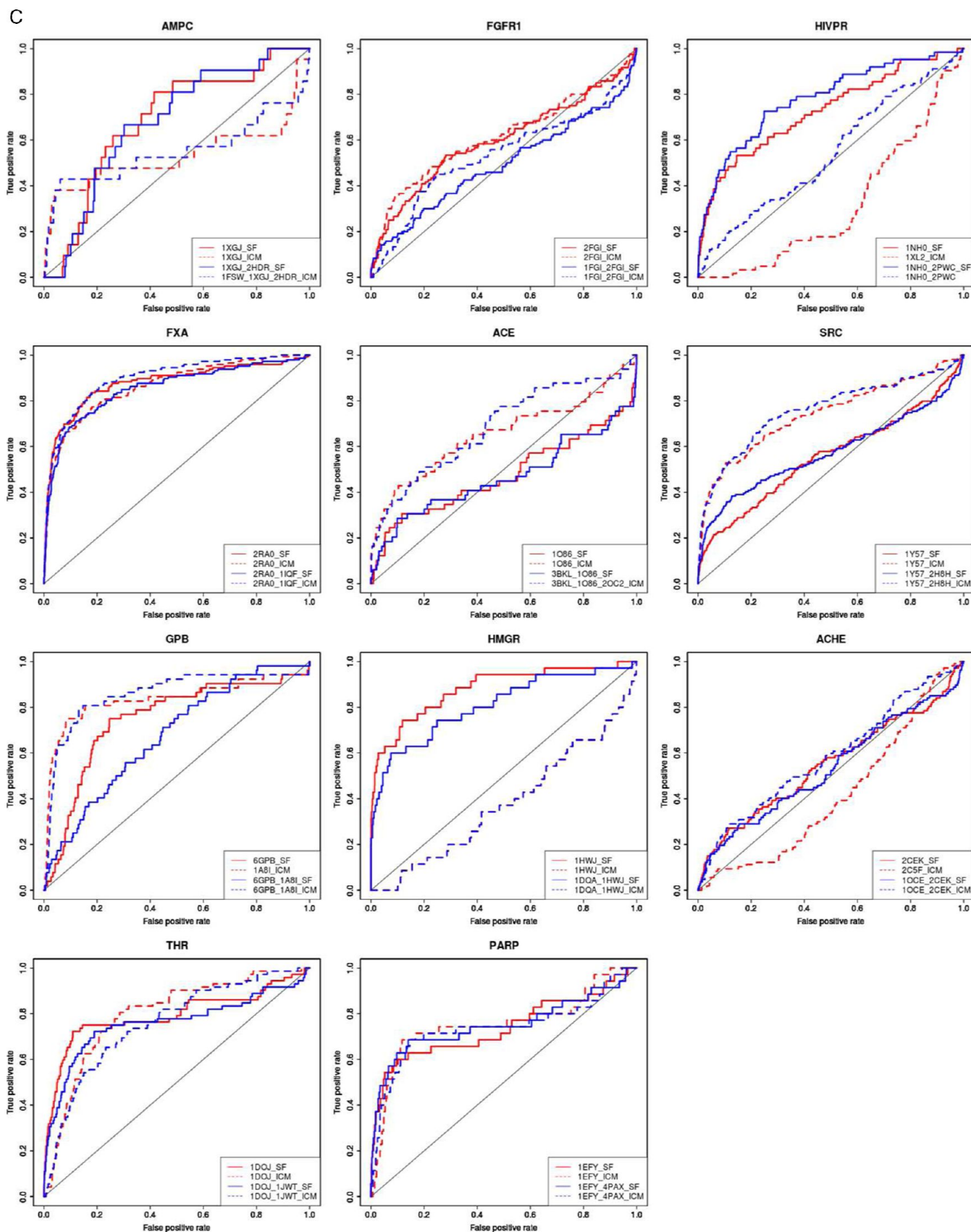


Figure 2. continued

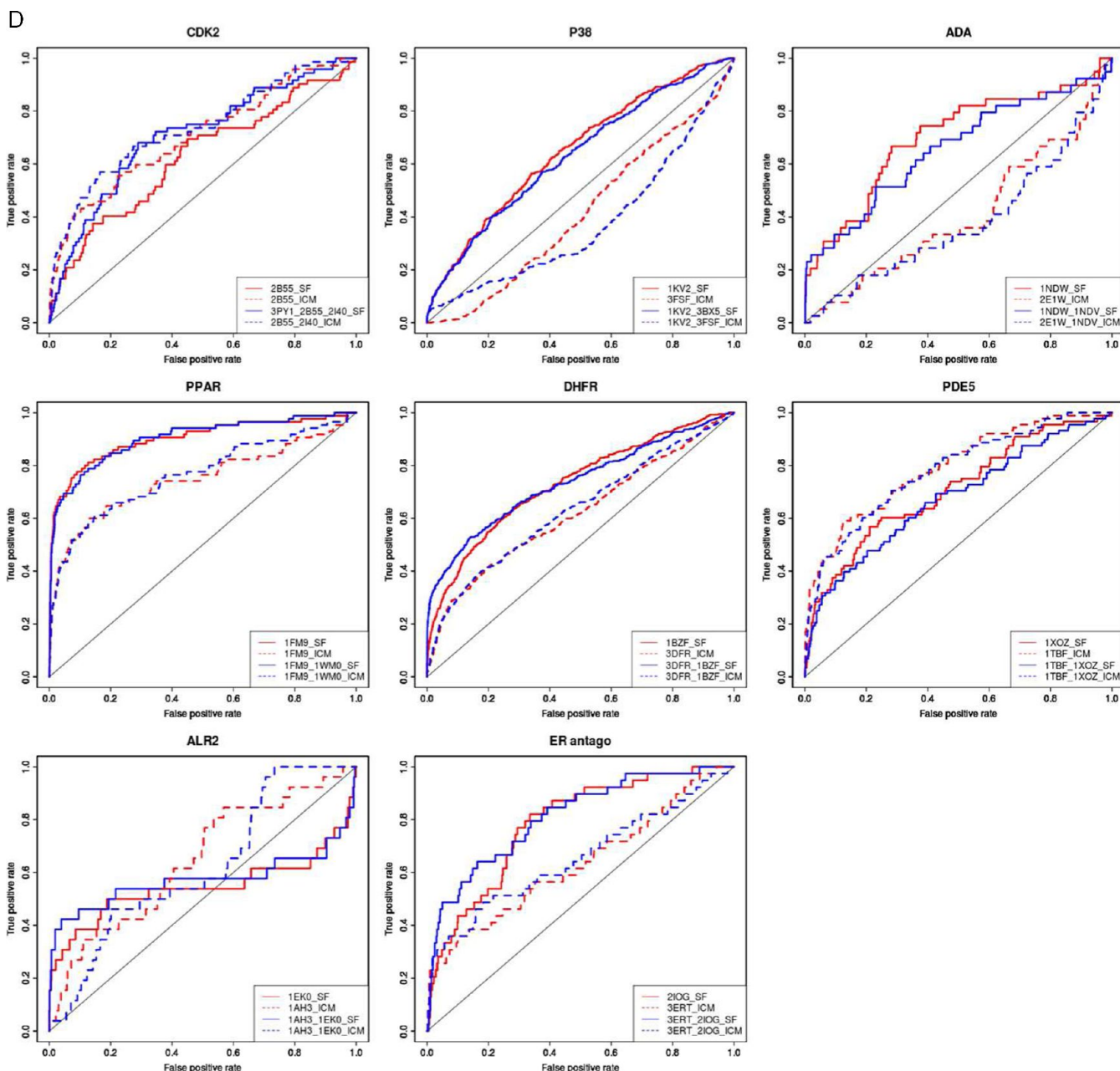


Figure 2. ROC curves with the best single docking (red) and the best ensemble (blue) docking protocols on the (A) small and less-hydrophobic, (B) small and more-hydrophobic, (C) large and less-hydrophobic, and (D) large and more-hydrophobic DUD systems with SF (plain-lines) and ICM (dotted lines).

structures selected for single structure docking. For a given compound on a given target, the best score obtained within the different structures composing the ensemble has been retained. Thus, for each ensemble, the ensemble docking result consists in a ranked list of the best scoring poses. Best ensemble docking ROC curves are presented in Figure 2. Ensemble docking enrichment factors and ROC AUC are presented in supplementary online material in Supporting Information Table S1.

In order to analyze the impact of the choice of the structures composing the ensembles, we excluded the 18 targets for which there was only 1 possible ensemble.

For respectively 20 with SF and 19 with ICM out of the 21 remaining targets, the highest performance was obtained with ensembles comprising only 2 conformations ($R_d = 15.3/21$).

For respectively 21 with SF and 17 with ICM out of the 21 targets, the ensemble that led to the best performances comprised the structure associated with the best performance by single structure docking ($R_d = 14.4/21$). In respectively 17 with SF and 16 with ICM out of 21 targets, this best ensemble encompassed the conformer displaying the smallest binding site ($R_d = 14.4/21$). Concerning the worst performances, for respectively 21 with SF and 20 with ICM out of the 21 targets, the structure associated with the worst performance by single structure docking is part of the ensemble providing the worst performances ($R_d = 14.4/21$). In 17 out of 21 targets with SF, the conformer displaying the largest binding site is always in the worst ensemble ($R_d = 14.4/21$). Using ICM, no signal was obtained when focusing on the ensemble associated with the worst performance.

For the large systems, the structure with the most closed binding site was in five out of nine cases not present in the best ensemble with SF and ICM ($R_d = 2.5/9$). For 6 out of the 11 small systems, the most open structure was not part of the best ensemble using SF ($R_d = 3/11$).

Comparison of Ensemble Docking and Single Docking Strategies. Using SF and ICM, single docking protocol outperformed the ensemble one for 18 targets whereas the ensemble docking performances are better than the single ones for 19 out of 39 targets. For two targets, no noticeable difference was noted between single and ensemble docking results (Figure 2)

DISCUSSION

Single Structure Docking Strategy. In the present work, we intended to identify possible trends in the structural properties of the conformations used as a reference for docking that led to the best enrichments in active compounds. The best conformation was the same with SF and ICM in 19 out of 39 targets and the worst for 21 out of 39 targets ($R_d = 6.74/39$). This observation confirms that the conformation of the query impacts the quality of the enrichment and is in agreement with what was described in the literature.^{19,25,26} It also shows that the performance of different programs can be dependent on common trends regarding the properties of the structure used as a reference. We thus looked for these common properties shared by the optimal conformations. To our knowledge, no binding site properties-based guidelines have been proposed to identify the optimal conformation within a set of available structures. We thus examined the physicochemical properties of the binding sites of the structures that led to the best enrichments.

For the small systems, the conformers that provided the best enrichments, either using SF or ICM, were in majority the conformers presenting the smallest binding site volume. Interestingly, this trend is strengthened when considering the conformers associated to the worst performances as they were at the opposite the ones that presented the largest binding sites. For the large systems, the conformers that provided the best enrichments were in majority the conformers presenting the largest binding site volume.

No strong trends could be observed when trying to link the quality of the performances to the opening of the binding site using both SF and ICM, even if slight tendencies could be observed with SF. Indeed, we observed that the most closed structures were associated with better performances compared to the most open structures.

Overall, it seemed that the volume of the binding site could be a critical criterion to select the optimal structure when different structures are available for a given target for a docking protocol.

Ensemble Docking. When different structures of the target are available, ensemble docking strategies are feasible, even with limited computational resources. We thus built with the experimental structures selected for each system, all the possible ensembles to conduct ensemble docking experiments using SF and ICM.

The first observation was that, in agreement with the literature,^{4,18,20} ensembles limited in size and in particular dual structure ensembles led to better performance than larger sets and thus seemed more appropriate for ensemble docking studies. Indeed with SF and ICM, respectively 20 and 19 out of the 21 best ensembles comprise 2 structures ($R_d = 15.3$).

However to our knowledge, no binding site properties-based guidelines have ever been defined to identify these optimal structures that should be part of the optimal ensemble within a set of known and available structures. An extensive work has been performed by Yoon et al.,¹⁷ Bottegoni et al.,²¹ and Rueda^{18,23} to remove the structures that should not be comprised in ensembles. They performed systematic preliminary docking and enrichment studies on all the structures available and removed from the ensembles the structures that performed poorly in enrichment with the benchmarking data sets. Despite giving excellent enrichment in known active compounds using the resulting ensembles (even better than each single structure approach), this strategy is very time-consuming as it requires a huge amount of preliminary calculation and analysis before performing the desired virtual screening on the “real-life” target. That is why the aim of our study was to identify “low-cost” guidelines to optimize this selection step and conduct the optimal SBVLS strategy on the best structure(s). In the 21 targets for which there was more than 1 possible ensemble, the best ensembles that we obtained comprised most often the structure that gave the best performance in a single docking strategy, using SF (20/21) and almost always when using ICM (19/21). This explains why the results obtained with ensemble docking protocol are directly correlated with those of single docking strategy. Indeed, both using SF (17/21) and ICM (16/21), the best ensemble comprised the conformer displaying the smallest binding site while; when using SF, the conformers presenting the largest binding site were present in the ensemble associated with the worst performances (17/21). The opening of the binding site could also be an interesting criterion for selecting the structure that should be part of the optimal ensemble as we observed that, in the large systems, the structure with the most closed binding site was in five out of nine cases not present in the best ensemble with SF and ICM ($R_d = 2.5/9$). For 6 out of the 11 small systems, the most open structure was not part of the best ensemble using SF ($R_d = 3/11$). This latter result was not replicated using ICM and may be due to the better treatment of solvation in the ICM scoring function compared to SF’s scoring function. Indeed, the opening of the binding site could impact its solvent accessibility.

Comparison of Ensemble Docking and Single Docking Strategies. Different opinions are expressed in the literature about the strategy to use when different structures of the target are available. Using prior SBVLS studies on all the structures available to define optimized ensembles, Yoon et al.,¹⁷ Bottegoni et al.,²¹ and Rueda et al.^{18,23} outperformed single structure approaches. However, our results, similarly to numerous other studies,^{4,5,18,38} underline that the best ensemble docking strategies that we used do not systematically outperform the best single structure docking approach (19 systems out of 39; see Figure 2). Since the computational cost of ensemble docking approaches increases linearly with the number of structures composing the ensemble, we estimated, at least for groups that have access to limited computational resources, that ensemble docking approaches should not be systematically preferred as they do not always perform better than single structure approaches.

Caveats within the Benchmarking Database. In addition to the caveats that were recently observed by Schneider et al.³⁹ in the DUD data set, we identified several points that should be taken into account in future versions of the DUD and that have not been corrected in the very recently released DUD-E.⁴⁰

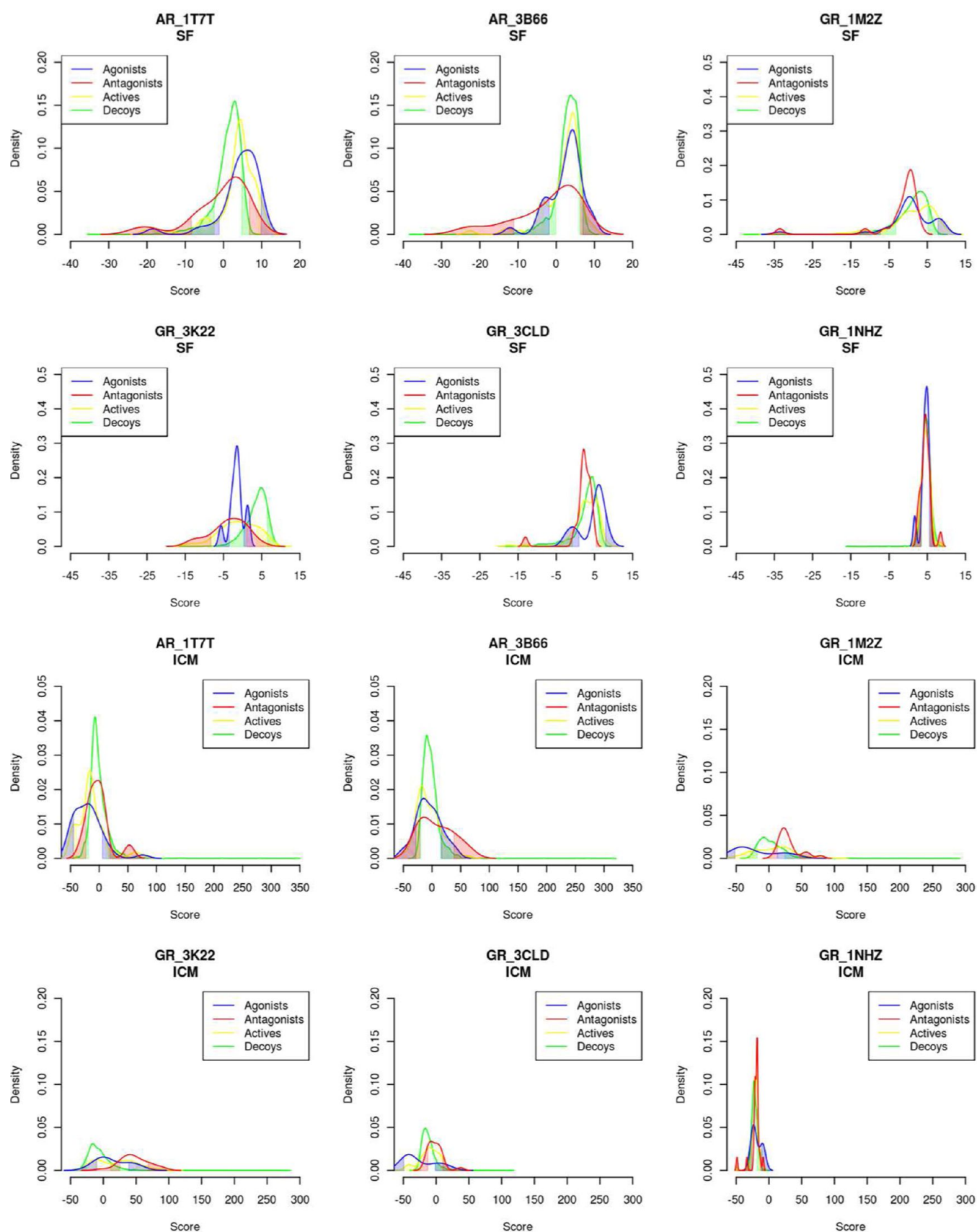


Figure 3. Score distribution curves of the agonist compounds (blue), antagonist compounds (red) within the active compounds (yellow) and their associated decoys (green) of the DUD data sets AR and GR using Surflex-dock or ICM on the different structures of AR and GR used in the study. For AR, 1T7T is agonist-bound; 3B66 is modulator-bound. For GR, 1M2Z, 3K22, and 3CLD are agonist-bound; 1NHZ is antagonist-bound.

Except for ER where the distinction between the pharmacological profile of the ligands (agonist, antagonist, or modulator) has already been made in DUD2, there are 6 other NRs, i.e. AR, GR, MR, PPAR, PR, and RXR, for which this distinction has not been performed. It is known that the pharmacological profile of the bound ligands influences the structure of the target,⁴¹ but this distinction has only been performed for ER in the DUD2. For GR and AR which comprise a sufficient number of agonists, antagonists, or modulators (12 antagonists and 36 agonists for AR; 23 antagonists and 11 agonists for GR), there are significant differences between the score distribution of the agonists and the score distribution of the antagonists when the reference structure is agonist-bound (1T7T for AR, 1M2Z, 3K22, and 3CLD for GR) or antagonist/modulator-bound (3B66 modulator-bound for AR, 1NHZ antagonist-bound for GR). This separation illustrated in Figure 3 is purely based on the pharmacological profile (whatever the docking method used) and demonstrates its influence on the resulting enrichments. Interestingly, this separation for agonists and antagonists of ER that existed in the DUD2 is no longer present in the DUD-E. Our results for GR and AR highlight the need to create separate DUD-own sets for NRs based on the pharmacological profiles of the known active ligands and the bound-structures used as a reference.

CONCLUSION

In the present work, we aimed to define the optimal protocol for a SBVLS experiment when different structures of the target are available. Our main objective was to assess guidelines based on the binding site properties to identify the optimal structure without performing time-consuming preliminary enrichment studies on benchmarking databases. We thus selected up to four structures for each system within the DUD data set into the PDB based on their differences of binding site properties (volume, opening). We evaluated the performance of two docking softwares, ICM and Surflex-dock, on these targets in order to identify the structures that led to optimal enrichments, and we compared their binding site properties. We identified several trends: (1) There was a structure that led to optimal enrichment which is in majority the same whatever the docking program used (in our case SF or ICM). (2) For small systems, the best enrichments are in majority obtained with the conformers presenting the smallest binding site volume whereas the worst performances are associated with the structure with the largest binding sites. Opposite conclusions are driven for the large systems for which the best structures displayed the largest binding sites.

We then constructed, for each system, all the possible ensembles with the structures selected in the previous step, compared the performances of ICM and SF on all these ensembles, and analyzed the composition of the resulting best ensembles. In our results, the best ensembles generally comprised only two structures and always comprised the best structure from single docking. Similarly to the results obtained in single structure docking, in the small systems, the conformers presenting the smallest binding site were in majority present in the best ensemble and the one with the largest binding site in the worst ensemble. For the large systems, the best structures displayed the largest binding sites. Concerning the opening of the binding site, in large systems with both programs, the most closed structures were not part of the best ensembles whereas with SF, in the small systems, the most open structures were

not part of the best ensembles. This information we got with two distinct docking programs could constitute helpful guidelines for the docking and screening community as there is a strong need for standardized protocols for selecting the right query especially when few ligands are known.

In addition to the issues that have been raised recently by Schneider et al.³⁹ in the DUD data set, we also identified several points that should be taken into account to help the construction of even more robust data sets.

In conclusion, we have been able to establish useful guidelines for the simple completion of a real-life SBVLS project, based on the binding site properties observed in the structures that have led to the optimal performance in our retrospective SBVLS tests.

ASSOCIATED CONTENT

Supporting Information

Docking accuracy, chemotype enrichments, and detailed ensemble docking results for SF and ICM. Species information for each structure selected for the study and details about the calculation of the volume of the binding sites using POVME. This material is available free of charge via the Internet at <http://pubs.acs.org>.

AUTHOR INFORMATION

Corresponding Author

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

Author Contributions

†These authors contributed equally to this work.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank Prof Jain for generously providing the Surflex package and Molsoft LLC for providing academic licenses for the ICM suite. H.G. and N.L. are recipients of a CIFRE fellowship from ANRT. N.B.N. is the recipient of a MNRT fellowship.

ABBREVIATIONS

SBVLS, structure-based virtual ligand screening; 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; COX1, cyclooxygenase-1; COX2, cyclooxygenase-2; DHFR, dihydrofolate reductase; EGFR, epidermal growth factor receptor kinase; ER ago, estrogen receptor agonist; ER antago, estrogen 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; HIVRT, HIV 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

REFERENCES

- (1) Alvarez, J. C. High-throughput docking as a source of novel drug leads. *Curr. Opin. Chem. Biol.* **2004**, *8* (4), 365–370.
- (2) Sherman, W.; Beard, H. S.; Farid, R. Use of an induced fit receptor structure in virtual screening. *Chem. Biol. Drug. Des.* **2006**, *67* (1), 83–84.
- (3) Cavasotto, C. N.; Abagyan, R. A. Protein flexibility in ligand docking and virtual screening to protein kinases. *J. Mol. Biol.* **2004**, *337* (1), 209–225.
- (4) Barril, X.; Morley, S. D. Unveiling the full potential of flexible receptor docking using multiple crystallographic structures. *J. Med. Chem.* **2005**, *48* (13), 4432–4443.
- (5) Craig, I. R.; Essex, J. W.; Spiegel, K. Ensemble docking into multiple crystallographically derived protein structures: an evaluation based on the statistical analysis of enrichments. *J. Chem. Inf. Model.* **2010**, *50* (4), 511–524.
- (6) Huang, S. Y.; Zou, X. Ensemble docking of multiple protein structures: considering protein structural variations in molecular docking. *Proteins* **2007**, *66* (2), 399–421.
- (7) Bolstad, E. S.; Anderson, A. C. In pursuit of virtual lead optimization: pruning ensembles of receptor structures for increased efficiency and accuracy during docking. *Proteins* **2009**, *75* (1), 62–74.
- (8) Broughton, H. B. A method for including protein flexibility in protein-ligand docking: improving tools for database mining and virtual screening. *J. Mol. Graph. Model.* **2000**, *18* (3), 247–257 302–304.
- (9) Cavasotto, C. N.; Kovacs, J. A.; Abagyan, R. A. Representing receptor flexibility in ligand docking through relevant normal modes. *J. Am. Chem. Soc.* **2005**, *127* (26), 9632–9640.
- (10) Frimurer, T. M.; Peters, G. H.; Iversen, L. F.; Andersen, H. S.; Moller, N. P.; Olsen, O. H. Ligand-induced conformational changes: improved predictions of ligand binding conformations and affinities. *Biophys. J.* **2003**, *84* (4), 2273–2281.
- (11) Sperandio, O.; Mouawad, L.; Pinto, E.; Villoutreix, B. O.; Perahia, D.; Miteva, M. A. How to choose relevant multiple receptor conformations for virtual screening: a test case of Cdk2 and normal mode analysis. *Eur. Biophys. J.* **2010**, *39* (9), 1365–1372.
- (12) Armen, R. S.; Chen, J.; Brooks, C. L. An Evaluation of Explicit Receptor Flexibility in Molecular Docking Using Molecular Dynamics and Torsion Angle Molecular Dynamics. *J. Chem. Theory Comput.* **2009**, *5* (10), 2909–2923.
- (13) Ferrari, A. M.; Wei, B. Q.; Costantino, L.; Shoichet, B. K. Soft docking and multiple receptor conformations in virtual screening. *J. Med. Chem.* **2004**, *47* (21), 5076–5084.
- (14) Totrov, M.; Abagyan, R. Flexible ligand docking to multiple receptor conformations: a practical alternative. *Curr. Opin. Struct. Biol.* **2008**, *18* (2), 178–184.
- (15) McGann, M. R.; Almond, H. R.; Nicholls, A.; Grant, J. A.; Brown, F. K. Gaussian docking functions. *Biopolymers* **2003**, *68* (1), 76–90.
- (16) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28* (1), 235–242.
- (17) Yoon, S.; Welsh, W. J. Identification of a minimal subset of receptor conformations for improved multiple conformation docking and two-step scoring. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (1), 88–96.
- (18) Rueda, M.; Bottegoni, G.; Abagyan, R. Recipes for the selection of experimental protein conformations for virtual screening. *J. Chem. Inf. Model.* **2010**, *50* (1), 186–193.
- (19) Thomas, M. P.; McInnes, C.; Fischer, P. M. Protein structures in virtual screening: a case study with CDK2. *J. Med. Chem.* **2006**, *49* (1), 92–104.
- (20) Rao, S.; Sanschagrin, P. C.; Greenwood, J. R.; Repasky, M. P.; Sherman, W.; Farid, R. Improving database enrichment through ensemble docking. *J. Comput. Aided Mol. Des.* **2008**, *22* (9), 621–627.
- (21) Bottegoni, G.; Rocchia, W.; Rueda, M.; Abagyan, R.; Cavalli, A. Systematic exploitation of multiple receptor conformations for virtual ligand screening. *PLoS One* **2011**, *6* (5), e18845.
- (22) Korb, O.; Olsson, T. S.; Bowden, S. J.; Hall, R. J.; Verdonk, M. L.; Liebeschuetz, J. W.; Cole, J. C. Potential and limitations of ensemble docking. *J. Chem. Inf. Model.* **2012**, *52* (5), 1262–1274.
- (23) Rueda, M.; Totrov, M.; Abagyan, R. ALiBERO: Evolving a team of complementary pocket conformations rather than a single leader. *J. Chem. Inf. Model.* **2012**, *52* (10), 2705–2714.
- (24) Giganti, D.; Guillemin, H.; Spadoni, J. L.; Nilges, M.; Zagury, J. F.; Montes, M. Comparative evaluation of 3D virtual ligand screening methods: impact of the molecular alignment on enrichment. *J. Chem. Inf. Model.* **2010**, *50* (6), 992–1004.
- (25) Hawkins, P. C.; Warren, G. L.; Skillman, A. G.; Nicholls, A. How to do an evaluation: pitfalls and traps. *J. Comput. Aided Mol. Des.* **2008**, *22* (3–4), 179–190.
- (26) McGovern, S. L.; Shoichet, B. K. Information decay in molecular docking screens against holo, apo, and modeled conformations of enzymes. *J. Med. Chem.* **2003**, *46* (14), 2895–2907.
- (27) McGann, M. FRED and HYBRID docking performance on standardized datasets. *J. Comput. Aided Mol. Des.* **2012**, *26* (8), 897–906.
- (28) Spitzer, R.; Jain, A. N. Surflex-Dock: Docking benchmarks and real-world application. *J. Comput. Aided Mol. Des.* **2012**, *26* (6), 687–699.
- (29) Brozell, S. R.; Mukherjee, S.; Balius, T. E.; Roe, D. R.; Case, D. A.; Rizzo, R. C. Evaluation of DOCK 6 as a pose generation and database enrichment tool. *J. Comput. Aided Mol. Des.* **2012**, *26* (6), 749–773.
- (30) Dundas, J.; Ouyang, Z.; Tseng, J.; Binkowski, A.; Turpaz, Y.; Liang, J. CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. *Nucleic Acids Res.* **2006**, *34* (Web Server issue), W116–W118.
- (31) Durrant, J. D.; de Oliveira, C. A.; McCammon, J. A. POVME: an algorithm for measuring binding-pocket volumes. *J. Mol. Graph. Model.* **2011**, *29* (5), 773–776.
- (32) Fraczkiwicz, R.; W, B. Exact and Efficient Analytical Calculation of the Accessible Surface Areas and Their Gradients for Macromolecules. *J. Comput. Chem.* **1998**, *19* (3), 319–333.
- (33) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **2004**, *25* (13), 1605–1612.
- (34) Jain, A. N. Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine. *J. Med. Chem.* **2003**, *46* (4), 499–511.
- (35) Abagyan, R.; Totrov, M.; Kuznetsov, D. ICM - a new method for protein modelling and design. Applications to docking and structure prediction from the distorted native conformation. *J. Comput. Chem.* **1994**, *15*, 488–506.
- (36) Schapira, M.; Abagyan, R.; Totrov, M. Nuclear hormone receptor targeted virtual screening. *J. Med. Chem.* **2003**, *46* (14), 3045–3059.
- (37) Sing, T.; Sander, O.; Beerewinkel, N.; Lengauer, T. ROCRC: visualizing classifier performance in R. *Bioinformatics* **2005**, *21* (20), 3940–3941.
- (38) Birch, L.; Murray, C. W.; Hartshorn, M. J.; Tickle, I. J.; Verdonk, M. L. Sensitivity of molecular docking to induced fit effects in influenza virus neuraminidase. *J. Comput. Aided Mol. Des.* **2002**, *16* (12), 855–869.
- (39) Schneider, N.; Hindle, S.; Lange, G.; Klein, R.; Albrecht, J.; Briem, H.; Beyer, K.; Claussen, H.; Gastreich, M.; Lemmen, C.; Rarey, M. Substantial improvements in large-scale redocking and screening using the novel HYDE scoring function. *J. Comput. Aided Mol. Des.* **2012**, *26* (6), 701–723.

(40) 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.* **2012**, *55* (14), 6582–6594.

(41) Bourguet, W.; Germain, P.; Gronemeyer, H. Nuclear receptor ligand-binding domains: three-dimensional structures, molecular interactions and pharmacological implications. *Trends Pharmacol. Sci.* **2000**, *21* (10), 381–388.

1.1.3 Discussion

Le premier objectif de cette étude était de confirmer que, comme suggéré dans des publications antérieures³⁶³⁻³⁶⁵ le choix de la structure de départ influe sur les performances obtenues lors du docking. En effet, lors de l'étude de docking utilisant une structure unique comme point de départ, les meilleures performances d'une part, et les pires d'autre part, ont été obtenues dans la majorité des cas avec la même structure avec les logiciels Surflex-Dock et ICM. Ces structures doivent donc vraisemblablement posséder des propriétés particulières communes qui leur confèrent leur statut de structure la plus adaptée ou au contraire la moins adaptée pour le docking. Nous nous sommes donc mis en quête de celles-ci et nous avons posé l'hypothèse de départ, jusqu'à présent jamais proposée à notre connaissance, que les propriétés physico-chimiques du site de liaison (volume, ouverture et hydrophobie) pouvaient influencer les performances du docking. Cependant, les 39 cibles de la DUD représentent un ensemble hétérogène de structures, ce qui est une source possible de biais car les propriétés nécessaires à la bonne conduite du docking peuvent différer si les structures sont trop différentes. Nous avons donc décidé de diviser les 39 cibles de la DUD en fonction des propriétés moyennes de volume et d'hydrophobie des structures choisies pour chaque cible (Tableau 15).

	Volume		
Hydrophobie	< 350 Å ³	> 350 Å ³	Total
< 70%	8 cibles « petite et moins hydrophobe »	11 cibles « grande et moins hydrophobe »	19 cibles « moins hydrophobes »
> 70%	12 Cible « petite et plus hydrophobe »	8 cibles « grande et plus hydrophobe »	20 cibles « plus hydrophobes »
Total	20 cibles « petites »	19 cibles « grandes »	39 cibles

Tableau 15. Classification des 39 cibles de la DUD en fonction du volume moyen et de l'hydrophobie moyenne des sites de liaison de leurs différentes structures

A l'intérieur des différentes classes ainsi créées, des propriétés communes intéressantes ont pu être mises en évidence en comparant les résultats obtenus avec Surflex-Dock et ICM dans les approches structure unique et ensemble docking.

Lors du docking à partir d'une structure unique, pour les cibles classées comme « petite », le conformère associé aux meilleures performances était dans la majorité des cas celui présentant le volume de site de liaison le plus faible. Parallèlement et de façon très intéressante, la structure associée aux performances les moins bonnes pour cette même classe

était celle dont le volume du site de liaison était le plus élevé. Similairement, pour les cibles appartenant à la catégorie « grande », les meilleures performances ont été obtenues à l'aide de la conformation dont le volume du site de liaison est le plus grand. Il semble donc bel et bien exister une corrélation entre le volume du site de liaison et les performances du docking, et ce critère pourrait être utilisé pour guider la sélection des structures lorsque plusieurs sont disponibles. Malheureusement, il n'a pas été possible d'établir une conclusion similaire par analyse des résultats du docking obtenus à l'aide des 2 logiciels en fonction de l'ouverture du site de liaison.

Lorsque l'approche ensemble docking est utilisée, des lignes directrices guidant le nombre et le type de structures à inclure sont là encore nécessaires. Pour tenter d'apporter des réponses, nous avons construits, pour chaque cible, tous les ensembles possibles à partir des structures précédemment sélectionnées. Les ensembles pouvaient donc être constitués de 2 à 4 structures et pour chaque cible de 1 à 11 ensembles étaient possibles. Similairement à ce qui avait déjà été observé par d'autres équipes ^{431, 440, 441}, les ensembles de taille limitée, et plus particulièrement dans notre cas, les ensembles constitués de 2 structures ont permis d'obtenir de meilleures performances que les ensembles incluant un plus grand nombre de structures. Même si cette information était très précieuse pour guider la sélection du meilleur ensemble possible, il restait encore à déterminer comment choisir de manière simple les structures à inclure ou à exclure, sans que de lourds calculs préliminaires soient nécessaires ^{439, 440, 442, 443}. Pour tenter de découvrir des propriétés communes dans les ensembles présentant les meilleures performances, seules les 21 cibles pour lesquelles plus d'un ensemble était possible ont été sélectionnées. Dans la quasi-totalité des cas, la structure permettant d'obtenir les meilleures performances lors du criblage à l'aide d'une structure unique était incluse dans le meilleur ensemble. Les résultats obtenus avec l'approche ensemble docking étaient donc directement corrélés avec ceux obtenus avec l'approche structure unique. En effet, quel que soit le logiciel utilisé, la structure associée au site de liaison de plus faible volume était présente dans la grande majorité des ensembles associés aux meilleurs résultats et que pour Surflex-Dock, les ensembles présentant les pires performances contenait dans la majorité des cas la structure avec le plus grand site de liaison. Cependant, et contrairement à l'approche structure unique, nous avons pu mettre en évidence que l'ouverture du site de liaison pouvait être utilisée comme critère des structures à ne pas inclure dans un ensemble. Ainsi, pour les « grandes » cibles, les structures présentant les sites de liaison fermés étaient dans la majorité des cas absentes de l'ensemble dont les performances étaient optimales pour les deux logiciels

utilisés. A l'inverse, pour les « petites » cibles, les structures les plus ouvertes n'étaient pas incluses dans la majorité des meilleurs ensembles obtenus avec Surflex-Dock.

Nous avons donc réussi à définir des critères de sélection de structures basés sur les propriétés du site de liaison permettant d'aider quiconque souhaitant réaliser un criblage virtuel à choisir la ou les structure(s) de départ à utiliser. Cependant, il reste une dernière question restée sans réponse : faut-il préférer l'approche structure unique ou celle ensemble docking ? Dans notre étude, nous avons constaté que, comme cela a déjà été décrit dans des études précédentes ^{431, 432, 440, 444}, l'approche ensemble docking, plus coûteuse en temps de calcul que l'approche structure unique, n'était pas systématiquement supérieure à cette dernière. Pour les groupes à capacités computationnelles limitées, le docking peut donc être réalisé à partir d'une structure unique optimale sans altérer significativement les performances par rapport à celles qui auraient été obtenues avec un ensemble de structures.

1.1.4 Analyse critique de l'étude

1.1.4.1 La banque d'évaluation DUD

La base de données DUD permet une évaluation des méthodes de criblage virtuel basées sur la structure de grande qualité. Cependant, depuis sa création, des problèmes ont été soulevés comme le manque de diversité des cibles et des structures des ligands ⁴²⁴ ou la sélection des decoys ^{364, 408, 425, 426}. Une nouvelle version de la DUD, la DUD_E a alors été proposée pour prendre en compte les critiques précédentes. Au cours de notre étude, nous avons identifié deux nouveaux points dont la prise en compte, jusque là ignorée, devrait permettre d'améliorer la qualité des banques d'évaluation.

Le premier s'intéressait au profil pharmacologique des ligands. En effet, dans la version 2 de la DUD, le récepteur nucléaire des œstrogènes alpha (ER_alpha) possède un jeu de données agoniste et un jeu de données antagonistes. Cette distinction, très intéressante aux vues des différences de conformations du site de liaison en fonction du profil pharmacologique des ligands co-cristallisés, n'est cependant pas réalisée pour les six autres récepteurs nucléaires de la DUD (récepteurs des androgènes AR, des glucocorticoïdes GR, des minéralocorticoïdes MR, activé par les proliférateurs de peroxyosomes gamma PPAR_gamma, de la progestérone PR et des rétinoïdes X alpha RXR_alpha) et est même totalement ignorée dans la DUD_E, même pour le récepteur ER_alpha. Pour étudier l'impact de la prise en compte du profil pharmacologique des ligands, nous avons choisi deux récepteurs nucléaires de la DUD, AR et GR, car ils présentaient tous les deux un nombre suffisants d'agonistes et d'antagonistes (et

de modulateurs). L'analyse des distributions de score a montré des différences significatives pour chaque jeu de données séparé, en fonction notamment du profil pharmacologique du ligand co-cristallisé dans la structure étudiée. Nous avons donc suggéré que la qualité des banques d'évaluation pourrait être amélioré par création de jeux de données séparés pour les agonistes et les antagonistes et par prise en compte du profil pharmacologique du ligand dans le site de liaison étudié.

Le second problème concernait la sélection des actifs. En effet, les données pour les actifs de la DUD ont été obtenues à partir de différentes bases de bioactivité (KiBank ⁴⁴⁵, PDBbind database ⁴⁴⁶, PubChem ⁴¹ ou encore ChEMBL ⁴² pour la DUD_E) ou d'études précédentes ^{355, 358, 419, 447-451, 452, 453-464} sans aucune vérification. La fiabilité des données ainsi collectées est donc à remettre en question. Nous avons notamment pu mettre en évidence que des ligands de la GPB (Figure 69) étaient systématiquement associés à des mauvaises valeurs d'enrichissement (données non présentées dans la publication). Une recherche bibliographique nous a permis d'expliquer simplement cette observation. En effet, les ligands mis en causes étaient bien des ligands de la GPB mais ne se fixaient pas au site de liaison étudié (le site catalytique de type purine) mais à un deuxième site de liaison allostérique ^{465, 466}. Cet exemple précis et probablement non isolé, illustre parfaitement le besoin de vérification des ligands proposés comme actifs dans les banques d'évaluation.

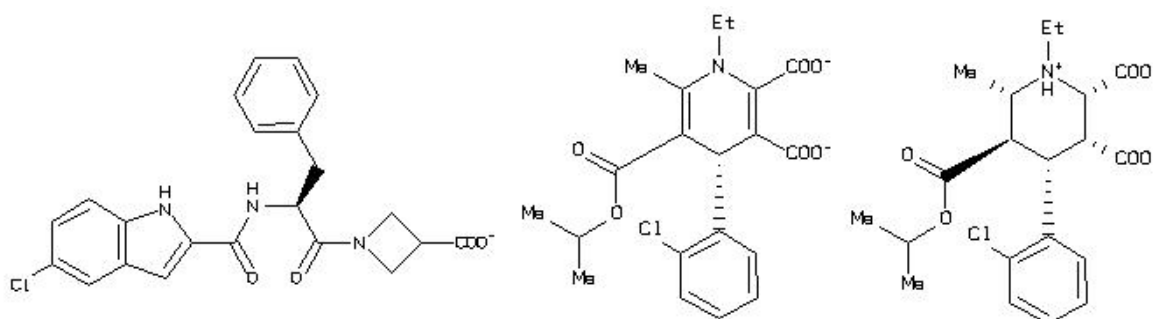


Figure 69. Ligands de la GPB inclus dans la DUD mais ne se fixant pas au site catalytique étudié mais à un site allostérique.

1.1.4.2 Déroulement de l'étude

Pour pouvoir mener cette étude, plusieurs stratégies pouvaient être adoptées à différents niveaux. Il est donc possible de discuter des différents choix que nous avons faits, et notamment de leur adéquation et des alternatives que nous aurions pu utiliser.

Ainsi, nous avons décidé de prendre en compte la flexibilité de la protéine par l'intermédiaire de la sélection de plusieurs conformations. Cependant, ceci représente une option parmi de nombreuses autres, et nous n'avons pas encore testé les conclusions de ce travail avec d'autres stratégies de prise en compte de la flexibilité.

Lors de la phase de sélection des structures des protéines à inclure dans l'étude nous avons décidé d'éliminer les structures apo, c'est-à-dire non co-cristallisées avec un ligand. En effet, si le choix de la structure de départ pour des études de docking est encore difficile et complexe, il est cependant recommandé de préférer des structures holo³⁶⁵ lorsque celles-ci sont disponibles. Nous avons choisi de ne conserver parmi toutes ses structures que les extrêmes en termes de volume et d'ouverture du site de liaison pour évaluer l'impact de ses propriétés sur les performances du docking. En effet, nos capacités computationnelles limitées ne nous permettaient pas de réaliser le docking sur toutes les structures afin de rechercher *a posteriori* les propriétés communes favorables. Néanmoins, force nous est de reconnaître que la sélection effectuée de 2 à 4 structures selon les cibles peut être une source de biais d'interprétation. En effet, lorsque la structure associée aux meilleures performances représente la structure extrême à la fois en termes de volume et d'ouverture du site de liaison, il est possible de s'interroger sur la contribution respective de ses propriétés sur les performances.

Pour réaliser une étude robuste, nous avons pris le parti d'utiliser deux logiciels de docking différents et les 39 cibles de la DUD pour lesquelles il existait au moins une structure expérimentale. Cependant, pour valider les tendances observées dans notre étude, le même travail pourrait être mené à l'aide de logiciels de docking supplémentaires (et complémentaires en termes d'algorithme de recherche et de fonctions de score) sur les jeux de données de la DUD_E qui présente à la fois un plus grand nombre de cibles à évaluer mais aussi de ligands et de decoys.

Pour analyser les résultats obtenus, nous avons là encore réalisé deux choix déterminants. Le premier, comme expliqué précédemment, a été de séparer les cibles en quatre catégories en fonction de leurs valeurs moyennes de volume et d'ouverture de site de liaisons. En effet, il est peu probable que les prérequis pour assurer le succès d'un criblage virtuel soient les mêmes lorsque les protéines sont très différentes. Nous avons donc fixé des seuils arbitraires à 350 Å pour le volume et 70 % pour l'hydrophobie qui nous permettaient de diviser de manière homogène les cibles dans les différentes classes. Une autre solution possible aurait été de considérer les grandes familles de protéines pour analyser les performances.

Malheureusement, les groupes ainsi constitués étaient trop petits (Figure 70) pour permettre d'interpréter statistiquement les résultats. Enfin, nous avons choisi l'AUC comme métrique de performance de notre étude. Cette métrique évaluant les performances globales d'une méthode aurait pu être complétée par d'autres métriques telles que l'EF1% et l'EF10% pour prendre en compte les performances précoces, ce qui a cependant été le cas en cas d'égalité des valeurs d'AUC.

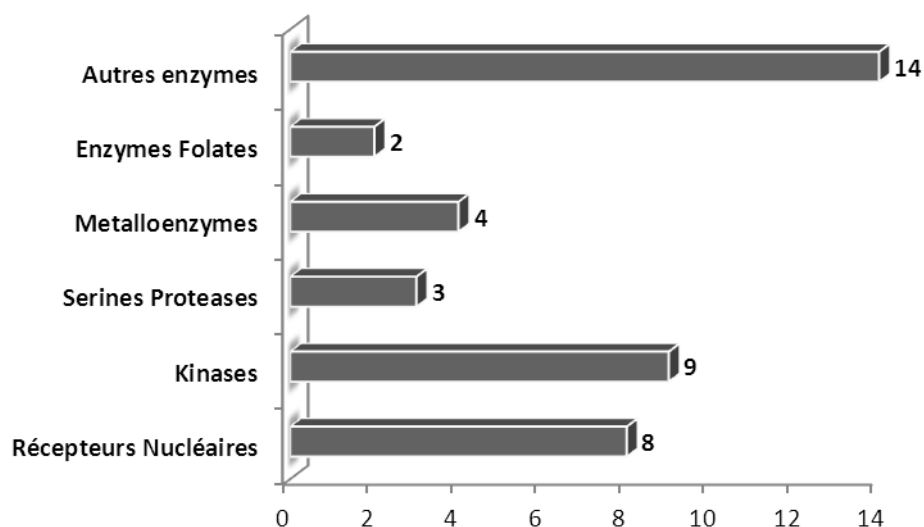


Figure 70. Classification des cibles de la DUD selon les grandes familles de protéines

1.1.5 Conclusion

Le but de cette étude était de définir le protocole optimal pour un criblage virtuel basé sur la structure et notamment dans la sélection de la ou des structures optimales de la protéine à l'aide des propriétés du site de liaison.

Après avoir confirmé l'importance de la structure de départ sur les performances du docking, en constatant qu'il y avait toujours une structure meilleure que les autres et que bien souvent elle était la même quel que soit le logiciel utilisé, nous avons recherché ce qui conférait à cette structure ce statut particulier.

Nous avons pu ainsi identifier le volume du site de liaison comme critère de sélection des structures, que ce soit pour une structure unique ou un ensemble de structures. En effet, pour les cibles dont le volume moyen ne dépasse pas 350 Å³, les meilleurs et les pires enrichissements en structure unique ont été en majorité obtenus avec les structures dont le

volume du site de liaison était, respectivement, le plus faible et le plus grand. Au contraire, pour les « grandes » cibles, les conformères présentant le site de liaison le plus grand ont été associés aux meilleurs résultats. Le meilleur ensemble (généralement composé de deux structures seulement) incluait toujours la structure identifiée comme la plus performante dans l'approche structure unique. Les tendances observées pour la première approche ont donc été retrouvées pour l'approche ensemble docking, à savoir que, pour les « petites » cibles, le conformère de plus petit site de liaison était en majorité présent dans le meilleur ensemble et celui de plus grand site de liaison dans le pire ensemble et que pour les « grandes » cibles le meilleur ensemble incluait la structure associée au plus grand site de liaison. L'ouverture du site de liaison a aussi été mise en évidence, mais cette fois comme critère d'exclusion pour constituer les ensembles les plus optimaux, pour les « grandes cibles » des structures les plus fermées et pour les « petites » cibles des structures les plus ouvertes.

Cette étude a aussi permis, à travers une analyse critique de la DUD, de soulever l'importance de créer des jeux de données séparés « agonistes » et « antagonistes » dans les banques d'évaluation, ainsi que de vérifier les données d'activités des actifs. Cette étude a donc été le point de départ de la création d'une nouvelle base de données dans laquelle les améliorations ainsi suggérées seraient prises en compte, la NRLiSt BDB (voir paragraphe 1.2).

En conclusion, nous avons donc réussi à mettre en évidence des critères de sélection de structures, simples et non dépendants de calculs préliminaires, basés sur les propriétés du site de liaison, et qui peuvent être utilisés comme lignes directrices pour les criblages virtuels prospectifs basés sur la structure.

1.2 La NRLiSt BDB : une banque d'évaluation validée manuellement dédiée aux ligands et aux structures des récepteurs nucléaires

1.2.1 Introduction

1.2.1.1 Les récepteurs nucléaires

Le clonage, il y a maintenant près de 30 ans, des récepteurs des glucocorticoïdes GR⁴⁶⁷ et des œstrogènes (ER)^{468, 469} a conduit à la découverte d'une nouvelle super-famille de facteurs de transcriptions: les récepteurs nucléaires (RN). Cette super-famille est composée de 48 membres, identifiés par le séquençage du génome humain et divisés en deux groupes égaux selon l'identification ou non d'un ligand endogène : les RNs endocrins et les RNs orphelins respectivement (Figure 71).

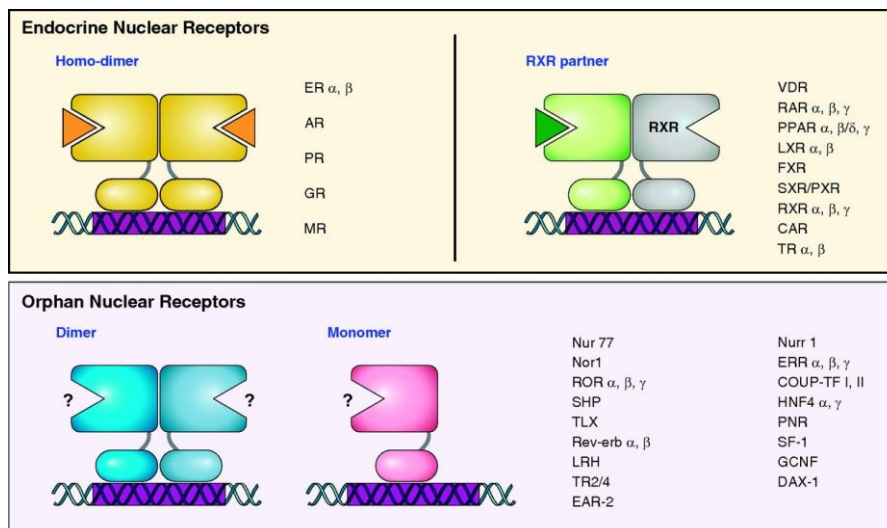


Figure 71. Classification des récepteurs nucléaires en « endocrins » ou « orphelins » selon l'identification préalable d'un ligand endogène. Les RNs « endocrins » peuvent former des homodimères ou des hétérodimères pour se lier à des parties spécifiques de l'ADN⁴⁷⁰

Les récepteurs nucléaires sont des facteurs de transcription naturellement activés ou inhibés par des petites molécules hormonales⁴⁷¹. Ils reconnaissent et se lient à des éléments spécifiques de l'ADN pour contrôler transcriptionnellement l'expression de gènes cibles (Figure 72). En général, l'action biologique des ligands se liant aux récepteurs nucléaires résulte de la régulation positive ou négative de la transcription du gène cible selon la nature de ce ligand (agoniste ou antagoniste). Cependant, les RNs orphelins agissent comme des

régulateurs transcriptionnels constitutifs, et peuvent, selon le cas, être des activateurs ou des répresseurs.⁴⁷⁰

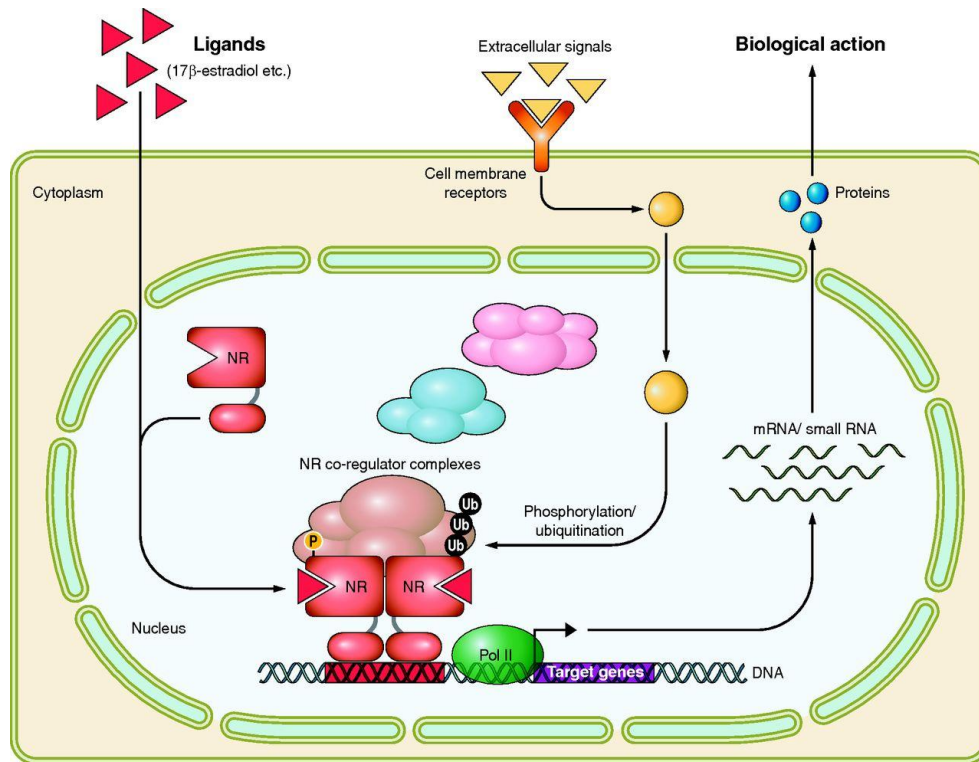


Figure 72. Illustration du contrôle transcriptionnel ligand dépendant exercé par les récepteurs nucléaires. La liaison du ligand provoque la dimérisation du récepteur et sa translocation dans le noyau cellulaire. Le dimère ainsi formé reconnaît et se lie à des éléments spécifiques de la séquence de l'ADN situés dans les régions régulatrices des gènes cibles pour pouvoir contrôler leur transcription grâce au recrutement de co-régulateurs⁴⁷⁰

Les récepteurs nucléaires permettent de moduler de nombreuses fonctions physiologiques clés, parmi lesquelles, la croissance, la différenciation, la reproduction, le métabolisme, l'homéostasie électrolytique, les réponses au stress et les fonctions immunitaires et sont impliqués dans une grande variété de maladies (Tableau 16). Ceci explique le grand intérêt porté à cette super-famille, en particulier pour le développement de nouveaux ligands à visée thérapeutique. Ainsi, aux Etats-Unis, 13% des médicaments recevant une autorisation de mise sur le marché sont des ligands des récepteurs nucléaires, et 15 d'entre eux font partie des 200 médicaments les plus prescrits⁴⁷².

Récepteurs nucléaires	Maladie / Fonction
TR α,β	Hypothyroïdisme, obésité
RAR α,β,γ	Maladies inflammatoires de la peau, leucémie
PPAR α,β,γ	Diabète, maladies coronaires, obésité
ROR α,β,γ	Athérosclérose, troubles immunologiques et neurologiques, ostéoporose
LXR α,β	Athérosclérose
FXR	Dyslipidémie, maladie du rein
VDR	Ostéoporose, homéostasie du calcium, prévention des cancers
PXR	<i>Métabolisme des xénobiotiques</i>
CAR	<i>Métabolisme des xénobiotiques</i>
HNF4 α,γ	Diabète, hémophilie, métabolisme des lipides
RXR α,β,γ	Leucémie, maladies coronaires
ER α,β	Cancer du poumon, ostéoporose, athérosclérose
ERR α,β,γ	<i>Développement embryogénique précoce</i>
GR	Troubles immunologiques et métaboliques
MR	Hypertension, hypertrophie myocardique
PR	Cancer du poumon, infertilité, maintien de la grossesse
AR	Cancer de la prostate, insensibilité androgénique liée au X, atrophie spinale/musculaire
NGFI-B, Nurr1	Troubles neurologiques et immunologiques, cancer
GCNF	<i>Fertilité/contraception</i>

Tableau 16. Les récepteurs nucléaires sont des cibles potentielles pour un grand nombre de maladies (indiquées en gras) et de dysfonctionnements physiologiques (indiqués en italique) (d'après ⁴⁷³)

Les médicaments actuellement commercialisés peuvent agir par activation (agonistes) ou inhibition (antagoniste) d'un ou plusieurs récepteurs nucléaires. C'est le cas du bicalutamide, un antagoniste des récepteurs des androgènes utilisé dans le traitement du cancer de la prostate ou de la prednisone, un agoniste des récepteurs des glucocorticoïdes, prescrit pour lutter contre les infections ou les allergies. Le mécanisme d'action de ces molécules repose sur leur liaison sur l'un des domaines constituant les récepteurs nucléaires, le domaine de liaison des ligands (ou Ligand Binding Domain LBD) (Figure 73).

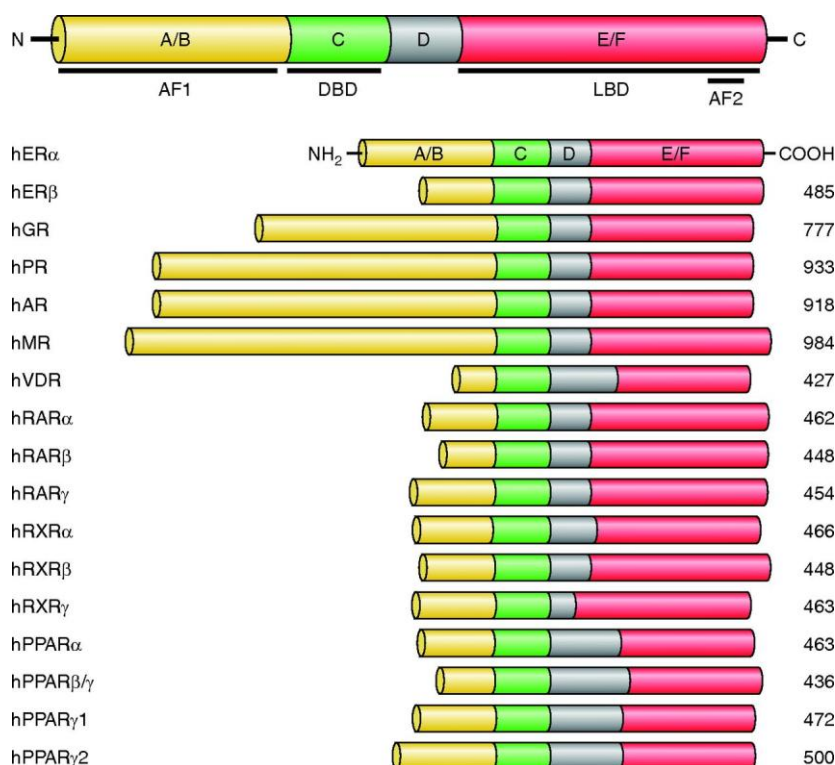


Figure 73. Représentation schématique des différents domaines constituant un récepteur nucléaire. Les RNs partagent une très grande similarité structurale et de fonctions. Cependant, le domaine A/B est la région la moins conservée au sein des RNs et contient une fonction d'activation transcriptionnelle (AF1) indépendante de la liaison d'un ligand. Le domaine C est hautement conservé et constitue le domaine de liaison à l'ADN (DNA Binding Domain DBD) sous la forme de deux doigts de zinc. Le domaine de liaison des ligands est quant à lui inclus dans la partie C-terminale E/F et est composé de 12 hélices formant en son centre une poche hydrophobe pouvant accueillir un ligand. Ce LBD contient aussi une fonction d'activation transcriptionnelle (AF2), qui est cette fois, ligand dépendante.⁴⁷⁰

1.2.1.2 Récepteurs nucléaires et évaluation des méthodes de criblage virtuel

L'évaluation des méthodes de criblage virtuel, le plus souvent réalisée de façon rétrospective, est une étape cruciale pour s'assurer de leur fiabilité et pour valider les résultats obtenus lors de criblages virtuels prospectifs. Cette évaluation n'est possible que grâce à la disponibilité de banques d'évaluation de haute qualité, telles que la DUD⁴²³ et sa nouvelle version la DUD_E⁴²⁷, actuellement considérées comme les meilleures dans ce domaine. Cependant, comme nous l'avons vu dans notre travail précédent (1.1.4.1), la qualité de cette banque d'évaluation peut, et doit, encore être améliorée.

La banque d'évaluation la plus exhaustive dédiée aux RNs était jusqu'à présent celle construite par Park et ses collègues ⁴⁷¹ qui proposait de 3 à 33 actifs pour 13 RNs associés à des ensembles de conformations du site de liaison. En plus de cette banque d'évaluation, de nombreuses bases de données consacrées aux récepteurs nucléaires sont disponibles et permettent d'étudier leur biologie (NURSA Nuclear Receptor Signaling Atlas ^{474, 475}, IUPHAR International Union of Basic and Clinical Pharmacology database ⁴⁷⁶), d'accéder aux données de séquences (NucleaRDB ⁴⁷⁷, NUREBASE ⁴⁷⁸, NRMD ⁴⁷⁹) ... Malheureusement, il n'existe actuellement aucune base de données regroupant de façon exhaustive l'ensemble des ligands et des structures disponibles pour les récepteurs nucléaires. Nous avons donc décidé de créer une banque d'évaluation exhaustive dédiée aux récepteurs nucléaires, optimisée pour l'évaluation des méthodes de criblages virtuels basées sur les structures mais aussi pour celles basées sur les ligands. Ce travail a donné naissance à la NRLiSt BDB (Nuclear Receptors Ligands and Structures Benchmarking DataBase) qui regroupe tous les ligands agonistes et antagonistes, identifiés lors d'une recherche bibliographique manuelle et exhaustive, pour les récepteurs nucléaires possédant plus d'un agoniste et un antagoniste et au moins une structure expérimentale. Nous avons aussi suggéré dans notre publication précédente, que mise à part la vérification manuelle des données des actifs inclus dans la banque d'évaluation, la création de jeux de données séparés « agonistes » et « antagonistes » devrait permettre d'améliorer la qualité des banques d'évaluation. Pour cela, nous nous étions basés sur les différences conformationnelles existantes dans les sites de liaison des RNs selon le profil pharmacologique du ligand lié ⁴⁸⁰. Nous avons donc décidé d'appliquer notre propre précepte dans la NRLiSt BDB et de fournir, pour chaque RN, un jeu de données agoniste et un antagoniste. Chacun de ces jeux contient toutes les structures expérimentales humaines « holo » du RN, les structures et les propriétés des ligands dont l'activité est documentée dans la littérature (agonistes ou antagonistes selon le jeu de données) et les structures des decoys correspondants générés à l'aide de l'outil automatique de la DUD_E ⁴²⁷ (Figure 74).

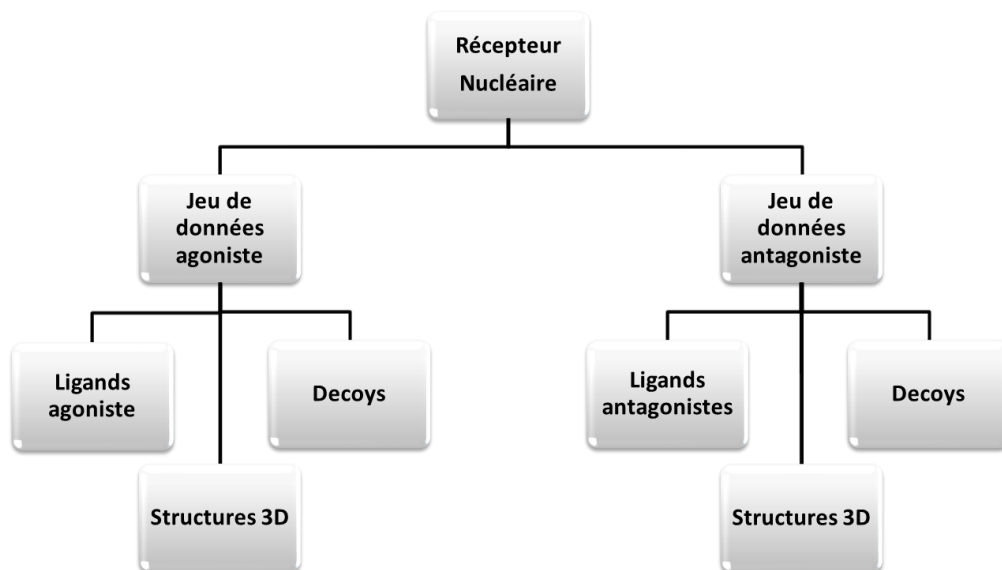


Figure 74. Contenu des jeux de données agoniste et antagoniste de la NRLiSt BDB

Dans cette étude, nous avons donc présenté le protocole employé pour générer la NRLiSt BDB. Au cours de notre recherche bibliographique approfondie, nous avons utilisé la base de données ChEMBL⁴² pour nous aider dans l'identification des agonistes et antagonistes présents dans la littérature. Cependant, nous avons rencontré quelques déconvenues, ici décrites, qui nous ont confortées dans notre volonté de vérifier manuellement les données fournies par les bases de bioactivités. Au final, la NRLiSt BDB constitue la banque d'évaluation la plus complète en ce qui concerne les ligands et les structures des RNs. Grâce à l'inclusion des données de profils pharmacologiques et d'activités récoltées manuellement dans la littérature scientifique, la NRLiSt BDB peut dépasser son rôle de banque d'évaluation et être utilisée pour aider à la compréhension des mécanismes biologiques fonctionnels et de modulation des RNs mais aussi et surtout à la découverte de nouveaux médicaments ciblant les RNs.

1.2.2 Publication

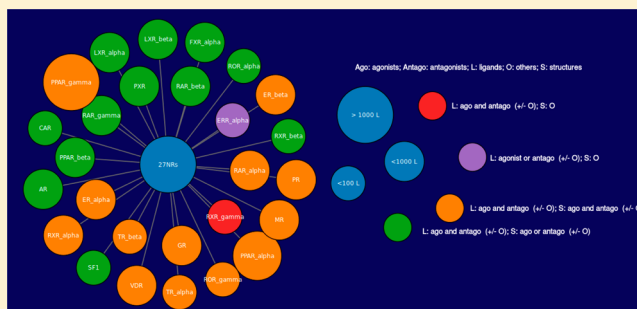
NRLiSt BDB, the Manually Curated Nuclear Receptors Ligands and Structures Benchmarking Database

Nathalie Lagarde, Nesrine Ben Nasr, Aurore Jérémie, H el ene Guillemain, Vincent Laville, Taoufik Labib, Jean-Fran ois Zagury, and Matthieu Montes*

Laboratoire G enomique, Bioinformatique et Applications, EA 4627, Conservatoire National des Arts et M etiers, 292 Rue Saint Martin, 75003 Paris, France

Supporting Information

ABSTRACT: Nuclear receptors (NRs) constitute an important class of drug targets. We created the most exhaustive NR-focused benchmarking database to date, the NRLiSt BDB (NRs ligands and structures benchmarking database). The 9905 compounds and 339 structures of the NRLiSt BDB are ready for structure-based and ligand-based virtual screening. In the present study, we detail the protocol used to generate the NRLiSt BDB and its features. We also give some examples of the errors that we found in ChEMBL that convinced us to manually review all original papers. Since extensive and manually curated experimental data about NR ligands and structures are provided in the NRLiSt BDB, it should become a powerful tool to assess the performance of virtual screening methods on NRs, to assist the understanding of NR's function and modulation, and to support the discovery of new drugs targeting NRs. NRLiSt BDB is freely available online at <http://nrlist.drugdesign.fr>.



INTRODUCTION

The nuclear receptors (NRs) are transcription factors naturally switched on and off by small-molecule hormones.¹ The NRs are involved in a wide range of physiological key functions and constitute an important class of drug targets. The evaluation of virtual screening methods, often achieved retrospectively, is necessary to ensure their reliability and relies on the availability of high quality benchmarking data sets. Among those developed over the years,^{2–5} the directory of useful decoys (DUD)⁶ and its enhanced version (DUD-E)⁷ are considered to date as the current gold standard benchmarking database. In the DUD, 40 structurally different targets are available, including 8 NRs: AR, ER_alpha agonist and ER_alpha antagonist, GR, mineralocorticoid receptor (MR), peroxisome proliferative activated receptor γ (PPAR_gamma), progesterone receptor (PR), and retinoid X receptor α (RXR_alpha). Recently, an enhanced version of the DUD has been proposed. This new benchmarking data set has taken into account several “weaknesses” pointed out in the construction of the DUD^{8–11} and consists of 102 targets including 11 NRs, the 7 previously available in the DUD (ER_alpha agonist and antagonist were merged into a unique data set), and 4 new NRs: estrogen receptor β (ER_beta), peroxisome proliferative activated receptor α (PPAR_alpha), peroxisome proliferative activated receptor δ also known as peroxisome proliferative activated receptor β (PPAR_beta), and thyroid hormone receptor β (TR_beta). To date, the most exhaustive benchmarking data set that includes NRs has been constructed in 2010 by Park et

al¹ to perform a study on NR ligands profiling and includes 3 to 33 active compounds on 13 nuclear receptors (NRs) with an ensemble of different pocket conformations. Benchmarking databases apart, several databases dedicated to NRs have been developed over the time. Some of them provide tools to elucidate NRs biology, for example, the NURSA (Nuclear Receptor Signaling Atlas)¹² which furnishes transcriptomic, proteomic, and metabolomic resources to study the expression, the organization, and the function of NRs.¹³ Similarly, the IUPHAR-DB (International Union of Basic and Clinical Pharmacology database) is focused on the NRs pharmacology.¹⁴ Other databases are more focused on NR sequences data, like the NucleaRDB which contains sequences, structures, and mutations of the NRs,¹⁵ the NUREBASE^{16,17} which provides the protein and DNA sequences, protein alignments and phylogenies, taxonomy and annotations of NRs, and the NRMD (nuclear receptor mutation database) which contains an atlas of NR mutations.¹⁸ Finally, the NURBS (nuclear receptor binding site database)¹⁹ includes data on experimental and predicted murine NRs. Unfortunately, none of them provides exhaustive data on ligands and structures of NRs, with annotated information about NR ligands activities and pharmacological profile.

We thus decided to create the most exhaustive NR-focused benchmarking database to date, the NRLiSt BDB (nuclear

Received: January 23, 2014

Published: March 17, 2014

Table 1. Pharmacological Profile of the Ligands Bound in the 339 Structures and of the 9905 NR Ligands Presented in the NRLiSt BDB

	no. of ligands						no. of structures		
	agonists			antagonists			agonists-bound	antagonists-bound	others-bound
	ligands	clusters	decoys	ligands	clusters	decoys			
AR	179	11	8746	226	23	11586	29	0	7
CAR	33	16	1499	2	2	146	2	0	0
ER_alpha	434	26	21642	137	13	6555	11	4	17
ER_beta	392	28	18953	70	12	3547	20	2	3
ERR_alpha	13	2	1043	3	2	150	0	0	2
FXR_alpha	320	13	16559	28	4	1172	7	0	0
GR	295	12	15207	369	16	19664	6	1	0
LXR_alpha	259	18	14149	50	5	2485	4	0	0
LXR_beta	374	18	18743	38	4	1865	7	0	0
MR	9	3	495	146	18	7467	5	2	0
PPAR_alpha	1401	27	67058	7	3	420	11	1	0
PPAR_beta	906	17	44650	11	4	597	11	0	2
PPAR_gamma	1820	51	50760	9	6	494	62	1	17
PR	269	15	13803	531	26	26539	8	2	3
PXR	100	24	6272	7	6	327	7	0	0
RAR_alpha	133	11	6550	66	3	3292	2	1	1
RAR_beta	130	9	6446	31	3	1583	1	0	0
RAR_gamma	132	11	6683	57	3	2855	8	0	0
ROR_alpha	3	2	147	13	2	649	2	0	0
ROR_gamma	7	2	348	4	1	200	3	1	0
RXR_alpha	210	13	11790	135	3	7151	24	2	3
RXR_beta	65	6	3874	7	3	348	2	0	0
RXR_gamma	71	6	4124	6	2	300	0	0	1
SF1	19	1	837	20	2	991	2	0	0
TR_alpha	69	8	3974	17	7	841	6	0	0
TR_beta	78	8	3648	15	5	795	10	0	0
VDR	132	4	6626	47	3	2336	16	0	0
total	7853	362	354626	2052	181	104355	266	17	56

receptors ligands and structures benchmarking database), optimized for the evaluation of structure-based and ligand-based virtual screening methods and for assisting the discovery of new NR ligands by providing comprehensive manually curated data on NR ligands and bound-structures pharmacological profiles. This new benchmarking database provides for the NRs having more than one agonist and one antagonist ligand and at least one experimental structure available, all available agonist and antagonist ligands identified in the literature. In order to assist more efficiently the discovery of new ligands of NRs, experimentally characterized agonist and antagonist ligands are provided in separated data sets, since evidence was reported that not only the ligand binding could induce conformational changes in the nuclear receptor structure but also these changes could vary according to the pharmacological profile of the ligand (agonist or antagonist).²⁰ Each data set of the NRLiSt BDB for a given NR comprises all the experimental human holo structures of the protein, the structures, and properties of all known active ligands (agonists or antagonists) and the structures of their corresponding decoys as provided by the DUD_E decoy generation tool.⁷

In the present study, we present the protocol used to generate the NRLiSt BDB. We also describe some of the errors found in ChEMBL that convinced us to manually review all original papers before including corresponding ligands into our database. NRLiSt BDB is the most comprehensive benchmarking data set on NR ligands and structures including their pharmacological profile and activity retrieved manually from

the literature. Additionally to its natural usage for benchmarking virtual screening methods, this database could also be used to assist the understanding of NRs function and modulation and the discovery of new drugs targeting NRs.

RESULTS

Nuclear Receptors Ligands and Structures Benchmarking DataBase (NRLiSt BDB). The NRLiSt BDB is a manually curated benchmarking database dedicated to the NR ligands and structures pharmacological profiles. We focused the database on the NRs for which more than one agonist and one antagonist ligand and at least one experimental structure were available. Only 27 NRs out of the 48 identified to date satisfied these conditions. We retrieved from the PDB all experimentally resolved holo human structures of the 27 NRs except for RXR_gamma for which only one apo structure was available. Of note, GR and MR structures presented only 99% of identity with the human sequences since mutations were produced to enable the protein crystallization.^{21–28} The holo structures were classified according to the pharmacological profile of the ligand that was cocrystallized in their binding site. An exhaustive review of the scientific literature using the ChEMBL database²⁹ as a starting point was conducted. We decided to focus only on the agonist and antagonist ligands and thus to ignore all ligands with alternative pharmacological profiles: inverse agonists, modulators, agonists/antagonists, and weak to partial agonists and antagonists. The structures of these ligands and their

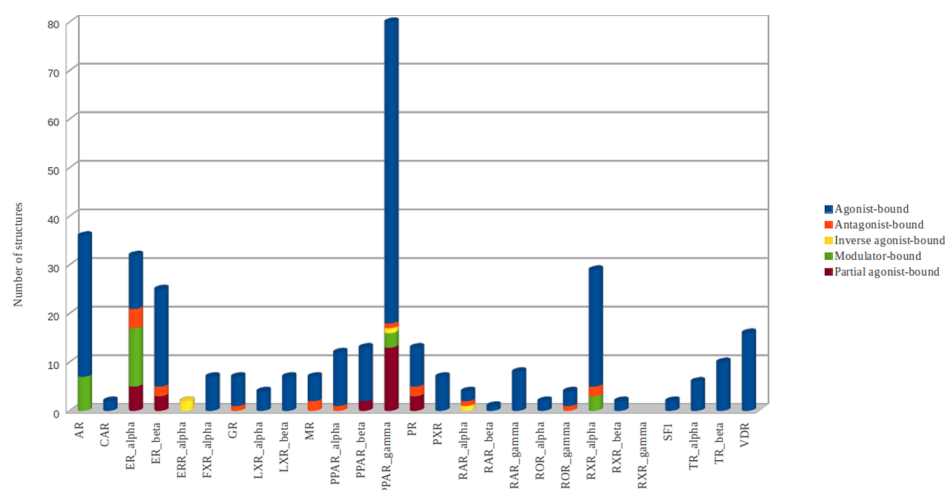


Figure 1. Classification of the holo structures included in the NRLiSt BDB according to the pharmacological profile of the bound ligand in the binding site. For RXR_gamma, only one apo structure was available.

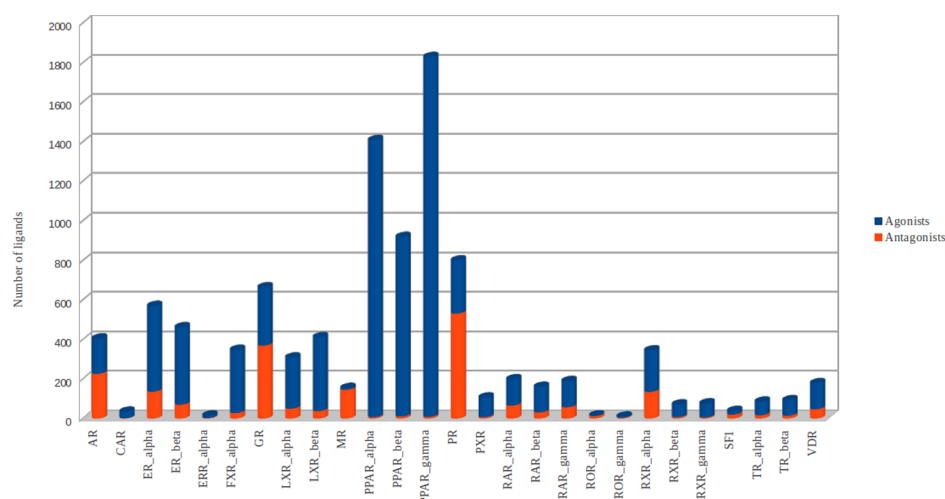


Figure 2. Classification of the ligands included in the NRLiSt BDB according to their pharmacological profiles.

corresponding activity and pharmacological profile are described in Supporting Information Table S1.

Hence, the NRLiSt BDB is constituted of 27 NRs, divided into 54 data sets comprising 339 structures (266 agonist-bound, 17 antagonist-bound, 56 others-bound) and 9905 ligands (7853 agonists, 2052 antagonists) (Table 1 and Figures 1 and 2). The number of structures available per NR varies from 1 (RAR_beta, RXR_gamma) to 80 (PPAR_gamma). The smallest set is CAR_antagonist containing only 2 ligands, and the largest is PPAR_gamma_agonist containing 1820 ligands. The structural diversity of the ligands in each data set was investigated using a Tanimoto similarity distance threshold ($T_d < 0.5$) (see Table 1). For each data set, 1 (SF1_agonist) to 51 (PPAR_gamma_agonist) clusters were obtained, with an average value of 10 clusters per data set. For each ligand, the DUD_E decoy generation tool⁷ was used to generate appropriate decoys. 458 981 decoys were obtained, leading to an average rate of 51 decoys per ligand. We also reported the mean values of 6 structural descriptors of the ligands computed with Dragon 6³⁰ (see Table 2 and Figure 3): the molecular weight (MW), the number of rotatable bonds (nrotB), the number of hydrogen bond donors (HBD) and acceptors (HBA), the Moriguchi octanol–water partition coefficient

(mlogP), and the topological polar surface area using N, O, S, and P polar contributions (TPSA).

By analyzing the differences between the values of these six descriptors obtained with the two pharmacological classes of ligands for each NR (agonists and antagonists), we observed that the mean value of the MW was significantly different for 16 out of the 27 NRs studied according to a Wilcoxon test³¹ (Supporting Information Table S2). Focusing on the NRs presenting at least 10 ligands in each class, this rate raised to 13 out of 17. Similarly, when considering the flexibility of ligands, for 13 out of the 27 NRs studied, the mean value of the nrotB was also significantly different between the two classes according to a Wilcoxon test³¹ (Supporting Information Table S2). As observed with the MW descriptor, when focusing on the NRs presenting at least 10 ligands in each class, this rate raised to 11 out of 17. No significant difference in the mean value of HBD, HBA, mlogP, and TPSA was noticed between agonists and antagonists.

DISCUSSION

In the present study, we have presented an exhaustive NR-focused benchmarking database, the NRLiSt BDB, dedicated to a family of targets widely studied for their therapeutical

Table 2. Mean Values of Six Descriptors Computed with Dragon 6.0 for the 54 Data Sets of the NRLiST BDB^a

	MW	nrotB	HBD	HBA	mlogP	TPSA
AR_agonist	337.323	1.724	1.149	6.326	2.883	60.728
AR_antagonist	386.177	3.659	0.889	6.478	3.088	67.529
CAR_agonist	347.318	5.212	0.758	3.667	3.504	62.520
CAR_antagonist	341.860	3.500	0.000	2.000	4.624	25.510
ER_alpha_agonist	323.653	2.765	2.012	3.825	3.169	60.660
ER_alpha_antagonist	458.390	8.110	1.971	5.360	3.451	65.916
ER_beta_agonist	303.690	2.362	1.906	3.778	2.961	60.246
ER_beta_antagonist	382.938	4.443	2.071	4.300	3.432	60.332
ERR_alpha_agonist	264.158	1.692	0.769	3.538	2.996	49.362
ERR_alpha_antagonist	238.293	2.000	0.667	0.667	4.965	13.487
FXR_alpha_agonist	506.109	6.959	0.966	6.066	4.221	82.750
FXR_alpha_antagonist	452.890	5.750	1.107	3.857	4.940	59.740
GR_agonist	426.141	4.414	2.227	5.973	3.593	73.831
GR_antagonist	490.209	5.144	1.277	5.448	4.347	71.244
LXR_alpha_agonist	466.515	6.436	0.884	7.301	4.450	62.959
LXR_alpha_antagonist	378.834	3.440	0.660	3.700	4.711	50.322
LXR_beta_agonist	475.726	6.283	0.735	6.794	4.591	66.174
LXR_beta_antagonist	416.367	3.763	0.737	4.711	4.712	58.997
MR_agonist	365.658	2.889	1.889	4.889	1.993	79.309
MR_antagonist	380.065	3.521	0.849	5.288	3.557	69.381
PPAR_alpha_agonist	449.990	9.592	0.673	6.453	3.766	84.637
PPAR_alpha_antagonist	465.703	8.286	1.000	5.429	3.530	122.473
PPAR_beta_agonist	476.844	9.413	0.522	7.134	4.006	89.735
PPAR_beta_antagonist	448.175	6.727	1.091	8.364	3.146	89.213
PPAR_gamma_agonist	455.864	9.878	0.777	6.331	3.845	84.219
PPAR_gamma_antagonist	336.536	6.667	1.000	5.444	2.832	85.980
PR_agonist	347.317	1.851	0.680	3.450	3.921	53.948
PR_antagonist	350.452	2.512	0.859	4.079	3.657	54.665
PXR_agonist	420.947	5.790	1.320	5.270	3.785	72.750
PXR_antagonist	443.186	3.143	1.714	6.571	2.865	88.894
RAR_alpha_agonist	364.844	3.812	0.504	3.308	4.739	56.010
RAR_alpha_antagonist	444.993	4.182	0.212	3.288	5.511	54.711
RAR_beta_agonist	354.842	4.277	0.277	3.062	4.948	52.585
RAR_beta_antagonist	422.596	3.219	0.469	3.281	5.679	51.220
RAR_gamma_agonist	367.609	3.556	0.331	3.128	4.913	54.436
RAR_gamma_antagonist	445.899	3.684	0.193	3.123	5.581	52.540
ROR_alpha_agonist	427.933	5.000	1.000	5.667	5.693	46.127
ROR_alpha_antagonist	416.732	7.462	1.231	5.615	3.473	71.082
ROR_gamma_agonist	404.523	4.857	1.857	3.286	5.506	38.837
ROR_gamma_antagonist	402.225	5.000	1.750	2.000	5.485	39.670
RXR_alpha_agonist	381.307	4.900	0.271	3.733	4.815	50.206
RXR_alpha_antagonist	514.234	6.200	1.022	6.096	5.169	78.311
RXR_beta_agonist	370.459	4.785	0.323	2.969	4.739	49.351
RXR_beta_antagonist	486.984	4.857	0.429	6.000	5.045	85.721
RXR_gamma_agonist	372.868	4.845	0.296	2.915	4.881	48.414
RXR_gamma_antagonist	495.543	4.333	0.333	6.333	5.022	89.408
SF1_agonist	367.253	8.211	0.579	0.789	6.312	9.717
SF1_antagonist	425.356	9.200	1.150	7.750	2.138	93.778
TR_alpha_agonist	487.943	5.841	1.986	5.652	3.371	85.939
TR_alpha_antagonist	491.622	7.824	1.882	5.824	3.979	106.041
TR_beta_agonist	482.584	5.846	1.949	5.679	3.376	86.554
TR_beta_antagonist	461.425	7.467	1.733	5.333	3.266	90.214
VDR_agonist	457.805	8.579	3.075	4.165	4.330	72.571
VDR_antagonist	505.089	6.702	2.553	4.596	4.641	75.961

^a(MW: Molecular Weight, nrotB: number of rotatable bonds, HBD: number of hydrogen bond donors, HBA: number of hydrogen bond acceptors; MlogP: Moriguchi octanol-water partition coefficient, TPSA: topological polar surface area using N, O, S, P polar contributions).

potential, the NRs. This benchmarking database exhibits three important features: (1) we did not just look for putative NR ligands but for compounds for which the pharmacological

profile on a given NR was determined experimentally; (2) all information about the compound activity was manually curated from the literature; (3) given (1) and (2), despite being a

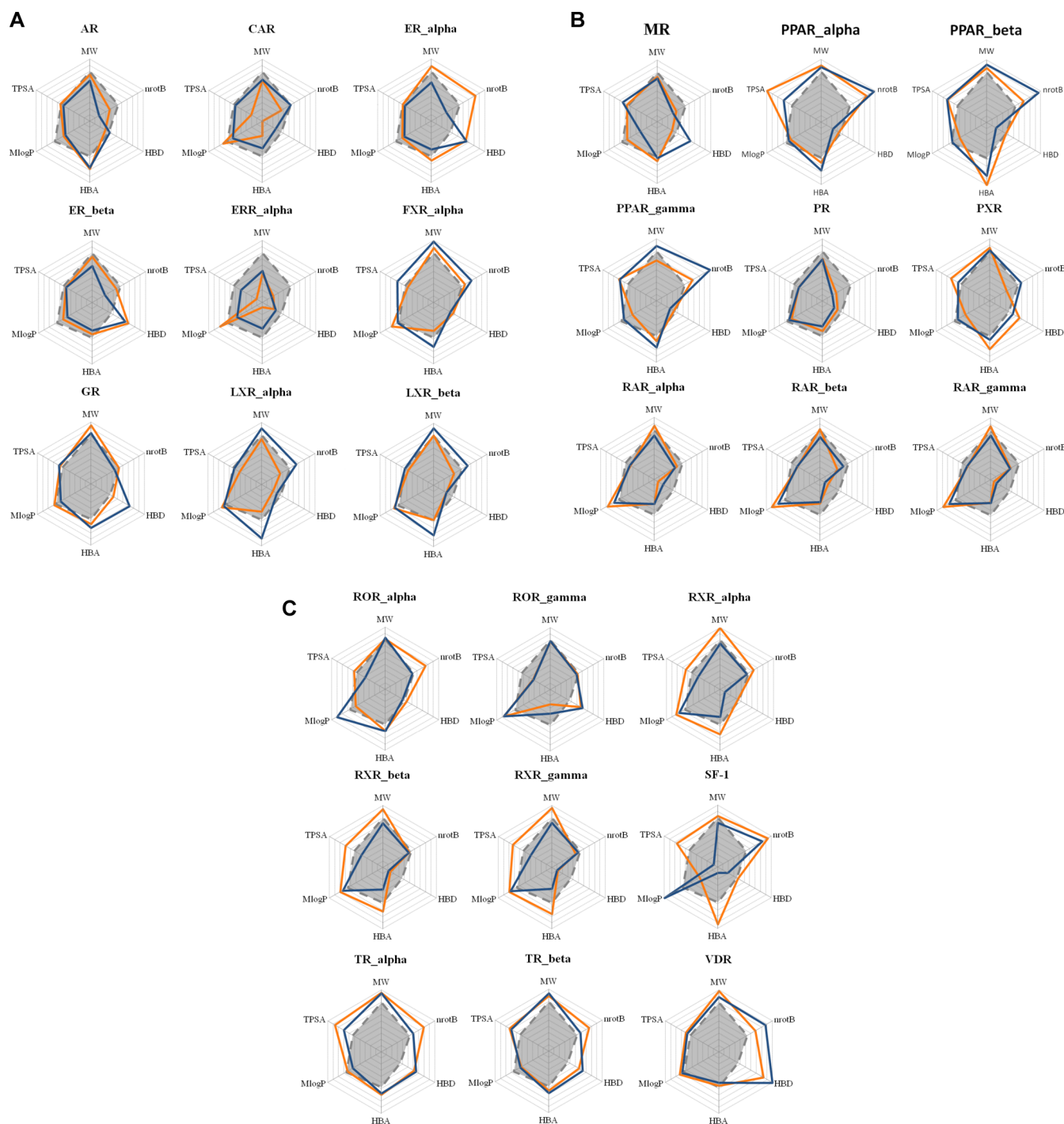


Figure 3. (A–C) Radiochart representation of mean values of the six descriptors computed with Dragon 6.0 for the agonist data sets (blue) and the antagonist data sets (orange) for each of the 27 NRs of the NRLiSt BDB: molecular weight (MW), number of rotatable bonds (nrotB), number of hydrogen bond donors (HBD) and acceptors (HBA), Moriguchi octanol–water partition coefficient (MlogP), and topological polar surface area using N, O, S, and P polar contributions (TPSA).

benchmarking database, the NRLiSt BDB can additionally be a precious source of information for assisting the discovery of new compounds targeting NRs.

The first benchmarking database was initiated by Bissantz et al.² more than 10 years ago. Several other benchmarking data sets have been proposed since then,^{3–7,32} among which are the DUD⁶ and DUD_E⁷ databases that are considered as the gold standard to date. The initial DUD, contained 8 NRs and its enhanced version contained 11 NRs. Interestingly, for

ER_alpha, the authors of the DUD decided to merge the agonist and antagonist data sets in a single one in the enhanced version of the DUD. Since we pointed out in a previous study³³ that this separation could be crucial for the quality of enrichment with structure-based virtual screening methods, we decided to create a benchmarking database specially focused on NRs with distinct data sets according to the pharmacological profile of the ligands. A similar distinction was previously

performed with GPCR ligands in the recently published GLL and GLD databases.³²

To make our database suitable to benchmark structure-based virtual screening methods, we decided to focus on the 27 NRs for which at least one experimental structure was available. Contrary to the DUDs, we decided to provide, for each NR, all holo structures available in the PDB to avoid bias in future benchmarking studies by choosing a priori a “best” structure given its use in docking studies, resolution or feedback from previous use in the DUD.⁷ In order to provide access to comprehensive data, for each holo structure, we have retrieved from the literature the information about the pharmacological profile of the bound ligand (agonist-bound, antagonist-bound, or other-bound). We decided to focus the NRLiSt BDB on agonist and antagonist ligands and thus eliminate all ligands with alternative pharmacological profiles: modulators, inverse agonist and partial agonists and antagonists.

To retrieve the ligands of each NR, we first used the ChEMBL database.²⁹ For each target of interest, the ChEMBL database proposed compounds that were classified according to their activity available in the literature. However, the information on the pharmacological profile of the ligands was not necessarily available, and we thus decided to manually retrieve all information about ligand binding and pharmacological profiles. During this extensive literature screen, we pointed out numerous false positive results in the ChEMBL database, which confirmed the legitimacy of our decision to review manually all data. For example, there are two compounds (namely, CHEMBL1961797 and CHEMBL1961794) that are denoted as bioactive on all NRs. By manual checking of the associated references in the literature,³⁴ these compounds have been determined experimentally as active with an agonist profile on Rev-erb alpha and Rev-erb beta only. “Neither compound exhibited activity at the other NR.”³⁴ For LXR_alpha, two compounds (CHEMBL209145 also known as DPPF-01 and CHEMBL210372 also known as DPHK-01) are proposed as active compounds. However, as claimed in the associated publication,³⁵ “DPPF-01 did not activate VDR, PPAR_alpha, PPAR_gamma, PPAR_delta, LXR_alpha, RAR_alpha, or RXR_alpha at 10 μM under these experimental conditions”. Conversely, “DPKH-01 did not activate VDR, FXR, PPAR_gamma, PPAR_delta, LXR_alpha, RAR_alpha, or RXR_alpha at 10 μM under these experimental conditions.”³⁵ Another example among numerous ones was about a factor Xa inhibitor classified as an ER_alpha ligand, but no trace of such activity could be found in the associated reference (ER_alpha is not even mentioned in the publication).³⁶

To complete the benchmarking database, we added decoys for each ligand based on the procedure to generate the DUD_E benchmarking data sets.⁷

We ultimately obtained a large benchmarking database, comprising 27 NRS, divided into 54 data sets, with a total of 9905 ligands including 7853 agonists and 2052 antagonists, 458 981 decoys and 339 experimental structures including 266 agonist-bound, 17 antagonist-bound, and 56 others-bound. There is a large disequilibrium between and within data sets, in terms of number of ligands and number of structures since several NRs have been more widely explored. For example, the PPARs are the NRs for which the largest number of ligands and structures were available, underlying the pharmaceutical potential of targeting such receptors.³⁷ Indeed, the PPARs mediate the action of several commercially available drugs

active for diabetes (thiazolidinediones like rosiglitazone, PPARγ agonist still used in the U.S.³⁸), hypolipidemia (bezafibrate, PPARα agonist³⁹), or inflammation.⁴⁰ Conversely, many NRs are still orphan receptors, and for some of them no experimental structure was even available.

Structural Diversity of the NRLiSt. As discussed by J. Irwin,¹⁰ sampling a sufficient part of the chemical space is of high importance for a benchmarking database. The NRs form a family of highly conserved receptors, which means that the absolute part of the chemical space that will be covered by the NRLiSt BDB is necessarily small. However, the NRLiSt BDB comprises the most exhaustive list of NRs agonists and antagonists and thus covers the most extensive chemical space related to NR ligands. By analysis of the cluster diversity as defined with a Td cutoff of 0.5, the NRLiSt BDB contains an average rate of 10 clusters per data set. Only 10 out of the 54 data sets contain less than 3 clusters, illustrating the lack of sufficient data for some receptors. For example, there are only 19 SF1 agonists that constitute a single cluster, which were extracted from the only two publications available to date. On the opposite, there were some very diverse data sets, for instance, the PPAR_gamma agonists data set that contains 51 clusters with a large amount of ligands, illustrating the extensive work performed on this particular class of NR ligands.

Structural Features of NR Ligands, Leads for Their Profiling. Simple structural descriptors could be used in an attempt to profile NR ligands. In many data sets, agonists and antagonists appeared to be significantly different in terms of molecular weight (MW) and number of rotatable bonds (nrotB). This observation is strengthened with the “large and equilibrated” NR data sets in terms of number of compounds in the agonist and antagonist data sets (like AR, ER, GR, LXR, PR, RAR, RXR_alpha, and VDR). No significant signal was found with hydrogen bond donors and acceptors, log P, or TPSA. These findings can be rationalized with previous knowledge about differences in the binding of agonists and antagonists. Indeed, many antagonists can be perfectly superimposed with their corresponding agonists, sharing similar hydrogen bonds (explaining why there is no significant difference between agonists and antagonists HBD and HBA mean values). However, as it is known for TR⁴¹ and RAR_alpha,⁴² some antagonists present bulky extensions (explaining the difference in MW and nrotB) that could prevent the H12 positioning in its agonist-bound form. All these data, correlated with the biological data that were also provided whenever available, constitute a robust support for structure–activity relationship studies and/or structure–properties relationship studies on NR ligands and could be used to assist medicinal chemists in prioritizing the synthesis of new NR ligands. Indeed, in the literature, such studies already enabled the successful discovery of new NRs ligands.^{43–45}

NRLiSt Web Site. We decided to provide freely all data collected during our bibliographic investigation and constituting the NRLiSt BDB on our Web site (<http://nrlist.drugdesign.fr>). All 9905 compounds identified to be NR agonists or antagonists with a confirmed biological activity in the literature are presented as tables, with their identification (ZINC, ChEMBL, or CID ID), their pharmacological profile (agonist or antagonist), their name (common name or name attributed in the associated reference in the literature), binding and/or activity data, the corresponding references from the literature, and finally their cross-reactivity data (cross-reactivity statistics are also presented in Supporting Information Table S3).

Similarly, all 339 holo human experimental structures of the NRs collected in the PDB are indicated, with their resolution, the name, and pharmacological profile of the bound ligand, and the organism from which the NR gene was extracted for expression. The 54 agonist and antagonist data sets corresponding to the 27 NRs selected for the NRLiSt BDB are available for download. The data sets are formed with three directories named “ligands”, “decoys”, and “targets”. The “ligands” directory contains the ligands in MOL2 and SMILES format, and their corresponding six structural descriptors are computed with Dragon 6.0. The “decoys” directory contains the decoys identified in the ZINC database for each compound of the data set in MOL2 and SMILES format. The “targets” directory contains all experimental structures for a given NR. For each structure the data provided are the original PDB format (for example, 1E3G.pdb), the same structure without ligand (1E3G_WL.pdb), the structure prepared for docking using CHIMERA (1E3G_prot.mol2), and the ligand protonated with adequate charges using Gasteiger partial charges (1E3G_ligand_prot.mol2). We hope this Web site will be useful for cheminformatics studies, since our downloadable database, the NRLiSt BDB, could be used for benchmarking virtual screening softwares but also for pharmaceutical chemistry studies, since the binding and activity data are provided together with the ligands and receptors structures.

CONCLUSION

The NRLiSt BDB is, to date, the database that contains the most comprehensive experimental data about NR ligands and structures. Since its data have been entirely reviewed manually from the literature, it is also, to our knowledge, one of the most reliable benchmarking database and NR-focused database publicly available. Our choices regarding the construction of the NRLiSt BDB can provide new insights about the building of better benchmarking data sets in particular by taking into account the final phenotype (i.e. pharmacological profile) resulting from the binding of ligands in addition to their sole affinity. NRLiSt BDB should become the database of reference to assess the performance of either structure or ligand-based virtual screening methods on NRs, to assist the understanding of NR's function and modulation, and to support the discovery of new drugs targeting NRs.

EXPERIMENTAL SECTION

NRLiSt BDB Generation. Protein Target Selection and Ligand Collection. The NRLiSt BDB was created by selecting from the 48 known NRs those for which more than one agonist, one antagonist ligand, and at least one experimental structure were available. For each of the 27 NRs corresponding to these criteria, all the human experimental holo structures available in the Protein Data Bank (PDB)⁴⁶ were downloaded. The pharmacological profile of the cocrystallized ligand was carefully monitored in the literature to define three classes of NR structures: agonist-bound, antagonist-bound, and others-bound (others corresponding to inverse agonists, modulators, agonists/antagonists, weak to partial agonists, and weak to partial antagonists). An extensive review of the scientific literature was achieved to collect all identified ligands for each NR, using the ChEMBL database²⁹ as a starting point. We manually curated all molecules proposed for a given NR to ascertain whether they were actually ligands or not and to classify them according to their pharmacological profile: agonist or antagonist. All inverse agonists, modulators, agonists/antagonists, weak to partial agonists, and weak to partial antagonists were eliminated. All ligands whose pharmacological profile was not informed were also eliminated. In addition to this qualitative activity data, quantitative data (K_d , K_i , EC_{50} , for example)

for each compound for each NR were also retrieved. Of note, we also gathered the original reference from which the data were retrieved to allow rigorous activity comparisons between compounds tested in the same conditions.

Three data sets were created for each NR: “agonist”, “antagonist”, and a “total” data set including merged agonist and antagonist data sets. Each data set was formed of three elements: all the previously selected holo PDB structures (except for NRs for which only apo structures were available), all the ligands found to be agonists or antagonists, and their corresponding computed decoys using the DUD_E decoy generation tool.⁷ The data set total is constituted, similar to the other data sets of all the previously selected PDB structures, of all ligands agonist and antagonist and the corresponding decoys.

Target Preparation. The structures were prepared to be usable for docking studies. The ligand bound in the active site was first removed from the protein. Then we used the DockPrep tool of UCSF CHIMERA⁴⁷ to delete the cocrystallized water molecules, repair truncated side chains, protonate the protein, and assign partial charges using the AMBER force field.

Ligands Preparation. All the ligands available in the ZINC database⁴⁸ were downloaded in MOL2 format, immediately usable for docking, and also in SMILES format. The remaining ligands were downloaded in SMILES format from the ChEMBL database²⁹ or the Pubchem database,⁴⁹ depending on their availability. The 3D structures of the ligands were generated with Corina online demo.⁵⁰ We used OpenBabel⁵¹ to calculate Gasteiger's partial atomic charges.

Decoys Preparation. We used the DUD_E decoy generation tool⁷ to select appropriate decoys for each ligand. These decoys were selected to have physical properties similar to those of their corresponding ligands (molecular weight, estimated water–octanol partition coefficient, rotatable bonds, hydrogen bond acceptors, hydrogen bond donors, and net charge) but to be topologically distant (computed with the Tanimoto coefficient, Tc). For each ligand and their alternative protonation states at pH 6–8 computed with Schrodinger's Epik, the DUD_E automated tool proposed, whenever possible, 50 decoys. The redundant decoys within a data set were removed, and the remaining decoys were then downloaded from the ZINC database in MOL2 and SMILES formats.

Ligand Cluster Definition. ICM, version 3.6, has been used to define ligand clusters in the NRLiSt BDB. Chemical descriptors fingerprints and Tanimoto similarity distance (Td) as implemented in ICM were used to classify the compounds in each data set. A Td cutoff was defined at 0.5 to obtain at least two equilibrated clusters in each data set, with the exception of SF1_agonist constituted by only one cluster.

Descriptors Calculation. Dragon 6. Dragon 6 enabled calculation of up to 4885 descriptors, from simple molecular descriptors to 2D and 3D descriptors, divided into 29 logical blocks.³⁰ We used the 3D structure of each compound to calculate 6 of their 4885 descriptors as proposed by Dragon 6.

Performance Metrics. All graphics were produced with the statistical and graphical tool R (<http://www.r-project.org/>).

ASSOCIATED CONTENT

Supporting Information

Details of the structure and inhibitory activity values for the 27 NR modulators, partial agonists and antagonists, and reverse agonists; *p*-values with a Wilcoxon test for comparing the mean values in four structural descriptors as computed by Dragon 6.0 for each data set; and cross-reactivity data about the NR ligands among the different NRs. This material is available free of charge via the Internet at <http://pubs.acs.org>.

AUTHOR INFORMATION

Corresponding Author

*Phone: +33140272809. E-mail: matthieu.montes@cnam.fr.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank Molsoft LLC for providing academic licenses for the ICM suite and Chemaxon for providing academic licenses for the Marvin suite. N.L. and V.L. are recipients of a CIFRE fellowship from ANRT. N.B.N. is recipient of a MNRT fellowship. H.G. is recipient of an ANSM fellowship.

■ ABBREVIATIONS USED

AR, androgen receptor; CAR, constitutive androstane receptor; DUD, directory of useful decoys; DUD-E, directory of useful decoys enhanced; EF, enrichment factor; ER_alpha, estrogen receptor α ; ER_beta, estrogen receptor β ; ERR_alpha, estrogen related receptor α ; FXR_alpha, farnesoid X receptor α ; GR, glucocorticoid receptor; HBA, hydrogen bond acceptor; HBD, hydrogen bond donor; LXR_alpha, liver X receptor α ; LXR_beta, liver X receptor β ; mlogP, Moriguchi octanol-water partition coefficient; MR, mineralocorticoid receptor; MW, molecular weight; NRLiSt BDB, nuclear receptors ligands and structures benchmarking database; nrotB, number of rotatable bonds; NR, nuclear receptor; PDB, Protein Data Bank; PPAR_alpha, peroxisome proliferator activated receptor α ; PPAR_beta, peroxisome proliferator activated receptor β ; PPAR_gamma, peroxisome proliferator activated receptor γ ; PR, progesterone receptor; PXR, pregnane X receptor; RAR_alpha, retinoic acid receptor α ; RAR_beta, retinoic acid receptor β ; RAR_gamma, retinoic acid receptor γ ; ROR_alpha, retinoic acid receptor related orphan receptor α ; ROR_gamma, retinoic acid receptor related orphan receptor γ ; RXR_alpha, retinoid X receptor α ; RXR_beta, retinoid X receptor β ; RXR_gamma, retinoid X receptor γ ; SF1, steroidogenic factor 1; Tc, Tanimoto coefficient; TPSA, topological polar surface area; TR_alpha, thyroid hormone receptor α ; TR_beta, thyroid hormone receptor β ; VDR, vitamin D receptor

■ REFERENCES

(1) Park, S. J.; Kufareva, I.; Abagyan, R. Improved docking, screening and selectivity prediction for small molecule nuclear receptor modulators using conformational ensembles. *J. Comput.-Aided Mol. Des.* **2010**, *24*, 459–471.

(2) Bissantz, C.; Folkers, G.; Rognan, D. Protein-based virtual screening of chemical databases. I. Evaluation of different docking/scoring combinations. *J. Med. Chem.* **2000**, *43*, 4759–4767.

(3) Miteva, M. A.; Lee, W. H.; Montes, M. O.; Villoutreix, B. O. Fast structure-based virtual ligand screening combining FRED, DOCK, and Surflex. *J. Med. Chem.* **2005**, *48*, 6012–6022.

(4) Montes, M.; Miteva, M. A.; Villoutreix, B. O. Structure-based virtual ligand screening with LigandFit: pose prediction and enrichment of compound collections. *Proteins* **2007**, *68*, 712–725.

(5) Pham, T. A.; Jain, A. N. Parameter estimation for scoring protein–ligand interactions using negative training data. *J. Med. Chem.* **2006**, *49*, 5856–5868.

(6) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking sets for molecular docking. *J. Med. Chem.* **2006**, *49*, 6789–6801.

(7) 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.* **2012**, *55*, 6582–6594.

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

(9) Hawkins, P. C.; Warren, G. L.; Skillman, A. G.; Nicholls, A. How to do an evaluation: pitfalls and traps. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 179–190.

(10) Irwin, J. J. Community benchmarks for virtual screening. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 193–199.

(11) Mysinger, M. M.; Shoichet, B. K. Rapid context-dependent ligand desolvation in molecular docking. *J. Chem. Inf. Model.* **2010**, *50*, 1561–1573.

(12) Margolis, R. N.; Evans, R. M.; O'Malley, B. W.; Consortium, N. A. The nuclear receptor signaling atlas: development of a functional atlas of nuclear receptors. *Mol. Endocrinol.* **2005**, *19*, 2433–2436.

(13) McKenna, N. J.; Cooney, A. J.; DeMayo, F. J.; Downes, M.; Glass, C. K.; Lanz, R. B.; Lazar, M. A.; Mangelsdorf, D. J.; Moore, D. D.; Qin, J.; Steffen, D. L.; Tsai, M. J.; Tsai, S. Y.; Yu, R.; Margolis, R. N.; Evans, R. M.; O'Malley, B. W. Minireview: Evolution of NURSA, the nuclear receptor signaling atlas. *Mol. Endocrinol.* **2009**, *23*, 740–746.

(14) Sharman, J. L.; Mpamhanga, C. P. IUPHAR-DB: an open-access, expert-curated resource for receptor and ion channel research. *ACS Chem. Neurosci.* **2011**, *2*, 232–235.

(15) Vroiling, B.; Thorne, D.; McDermott, P.; Joosten, H. J.; Attwood, T. K.; Pettifer, S.; Vriend, G. NucleaRDB: information system for nuclear receptors. *Nucleic Acids Res.* **2012**, *40*, D377–380.

(16) Duarte, J.; Perriere, G.; Laudet, V.; Robinson-Rechavi, M. NUREBASE: database of nuclear hormone receptors. *Nucleic Acids Res.* **2002**, *30*, 364–368.

(17) Ruau, D.; Duarte, J.; Ourjidal, T.; Perriere, G.; Laudet, V.; Robinson-Rechavi, M. Update of NUREBASE: nuclear hormone receptor functional genomics. *Nucleic Acids Res.* **2004**, *32*, D165–D167.

(18) Van Durme, J. J.; Bettler, E.; Folkertsma, S.; Horn, F.; Vriend, G. NRMD: nuclear receptor mutation database. *Nucleic Acids Res.* **2003**, *31*, 331–333.

(19) Fang, Y.; Liu, H. X.; Zhang, N.; Guo, G. L.; Wan, Y. J.; Fang, J. NURBS: a database of experimental and predicted nuclear receptor binding sites of mouse. *Bioinformatics* **2013**, *29*, 295–297.

(20) Bourguet, W.; Germain, P.; Gronemeyer, H. Nuclear receptor ligand-binding domains: three-dimensional structures, molecular interactions and pharmacological implications. *Trends Pharmacol. Sci.* **2000**, *21*, 381–388.

(21) Bledsoe, R. K.; Madauss, K. P.; Holt, J. A.; Apolito, C. J.; Lambert, M. H.; Pearce, K. H.; Stanley, T. B.; Stewart, E. L.; Trump, R. P.; Willson, T. M.; Williams, S. P. A ligand-mediated hydrogen bond network required for the activation of the mineralocorticoid receptor. *J. Biol. Chem.* **2005**, *280*, 31283–31293.

(22) Hasui, T.; Matsunaga, N.; Ora, T.; Ohyabu, N.; Nishigaki, N.; Imura, Y.; Igata, Y.; Matsui, H.; Motoyaji, T.; Tanaka, T.; Habuka, N.; Sogabe, S.; Ono, M.; Siedem, C. S.; Tang, T. P.; Gauthier, C.; De Meese, L. A.; Boyd, S. A.; Fukumoto, S. Identification of benzoxazin-3-one derivatives as novel, potent, and selective nonsteroidal mineralocorticoid receptor antagonists. *J. Med. Chem.* **2011**, *54*, 8616–8631.

(23) Bledsoe, R. K.; Montana, V. G.; Stanley, T. B.; Delves, C. J.; Apolito, C. J.; McKee, D. D.; Consler, T. G.; Parks, D. J.; Stewart, E. L.; Willson, T. M.; Lambert, M. H.; Moore, J. T.; Pearce, K. H.; Xu, H. E. Crystal structure of the glucocorticoid receptor ligand binding domain reveals a novel mode of receptor dimerization and coactivator recognition. *Cell* **2002**, *110*, 93–105.

(24) Biggadike, K.; Bledsoe, R. K.; Coe, D. M.; Cooper, T. W.; House, D.; Iannone, M. A.; Macdonald, S. J.; Madauss, K. P.; McLay, I. M.; Shipley, T. J.; Taylor, S. J.; Tran, T. B.; Uings, I. J.; Weller, V.; Williams, S. P. Design and X-ray crystal structures of high-potency nonsteroidal glucocorticoid agonists exploiting a novel binding site on the receptor. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 18114–18119.

(25) Biggadike, K.; Bledsoe, R. K.; Hassell, A. M.; Kirk, B. E.; McLay, I. M.; Shewchuk, L. M.; Stewart, E. L. X-ray crystal structure of the novel enhanced-affinity glucocorticoid agonist fluticasone furoate in the glucocorticoid receptor–ligand binding domain. *J. Med. Chem.* **2008**, *51*, 3349–3352.

(26) Kauppi, B.; Jakob, C.; Farnegardh, M.; Yang, J.; Ahola, H.; Alarcon, M.; Calles, K.; Engstrom, O.; Harlan, J.; Muchmore, S.; Ramqvist, A. K.; Thorell, S.; Ohman, L.; Greer, J.; Gustafsson, J. A.

Carlstedt-Duke, J.; Carlquist, M. The three-dimensional structures of antagonistic and agonistic forms of the glucocorticoid receptor ligand-binding domain: RU-486 induces a transconformation that leads to active antagonism. *J. Biol. Chem.* **2003**, *278*, 22748–22754.

(27) Suino-Powell, K.; Xu, Y.; Zhang, C.; Tao, Y. G.; Tolbert, W. D.; Simons, S. S., Jr.; Xu, H. E. Doubling the size of the glucocorticoid receptor ligand binding pocket by deacetylcortivazol. *Mol. Cell. Biol.* **2008**, *28*, 1915–1923.

(28) Madauss, K. P.; Bledsoe, R. K.; McLay, I.; Stewart, E. L.; Uings, I. J.; Weingarten, G.; Williams, S. P. The first X-ray crystal structure of the glucocorticoid receptor bound to a non-steroidal agonist. *Bioorg. Med. Chem. Lett.* **2008**, *18*, 6097–6099.

(29) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012**, *40*, D1100–D1107.

(30) Talete SRL. Dragon 6. http://www.talete.mi.it/products/dragon_description.htm.

(31) Wilcoxon, F. Individual comparisons by ranking methods. *Biom. Bull.* **1945**, *1*, 80–83.

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

(33) 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.

(34) Solt, L. A.; Wang, Y.; Banerjee, S.; Hughes, T.; Kojetin, D. J.; Lundasen, T.; Shin, Y.; Liu, J.; Cameron, M. D.; Noel, R.; Yoo, S. H.; Takahashi, J. S.; Butler, A. A.; Kamenecka, T. M.; Burris, T. P. Regulation of circadian behaviour and metabolism by synthetic REV-ERB agonists. *Nature* **2012**, *485*, 62–68.

(35) Kainuma, M.; Kasuga, J.; Hosoda, S.; Wakabayashi, K.; Tanatani, A.; Nagasawa, K.; Miyachi, H.; Makishima, M.; Hashimoto, Y. Diphenylmethane skeleton as a multi-template for nuclear receptor ligands: preparation of FXR and PPAR ligands. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 3213–3218.

(36) Shi, Y.; Zhang, J.; Shi, M.; O'Connor, S. P.; Bisaha, S. N.; Li, C.; Sitkoff, D.; Pudzianowski, A. T.; Chong, S.; Klei, H. E.; Kish, K.; Yanchunas, J., Jr.; Liu, E. C.; Hartl, K. S.; Seiler, S. M.; Steinbacher, T. E.; Schumacher, W. A.; Atwal, K. S.; Stein, P. D. Cyanoguanidine-based lactam derivatives as a novel class of orally bioavailable factor Xa inhibitors. *Bioorg. Med. Chem. Lett.* **2009**, *19*, 4034–4041.

(37) Desvergne, B.; Wahli, W. Peroxisome proliferator-activated receptors: nuclear control of metabolism. *Endocr. Rev.* **1999**, *20*, 649–688.

(38) Vázquez, M.; Silvestre, J. S.; Prous, J. R. Experimental approaches to study PPAR gamma agonists as antidiabetic drugs. *Methods Find. Exp. Clin. Pharmacol.* **2002**, *24*, 515–523.

(39) Krey, G.; Braissant, O.; L'Horsset, F.; Kalkhoven, E.; Perroud, M.; Parker, M. G.; Wahli, W. Fatty acids, eicosanoids, and hypolipidemic agents identified as ligands of peroxisome proliferator-activated receptors by coactivator-dependent receptor ligand assay. *Mol. Endocrinol.* **1997**, *779*–791.

(40) Straus, D. S.; Glass, C. K. Anti-inflammatory actions of PPAR ligands: new insights on cellular and molecular mechanisms. *Trends Immunol.* **2007**, *28*, 551–558.

(41) Togashi, M.; Borngraeber, S.; Sandler, B.; Fletterick, R. J.; Webb, P.; Baxter, J. D. Conformational adaptation of nuclear receptor ligand binding domains to agonists: potential for novel approaches to ligand design. *J. Steroid Biochem. Mol. Biol.* **2005**, *93*, 127–137.

(42) Schapira, M.; Raaka, B. M.; Samuels, H. H.; Abagyan, R. Rational discovery of novel nuclear hormone receptor antagonists. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 1008–1013.

(43) Stauffer, S. R.; Coletta, C. J.; Tedesco, R.; Nishiguchi, G.; Carlson, K.; Sun, J.; Katzenellenbogen, B. S.; Katzenellenbogen, J. A. Pyrazole ligands: structure–affinity/activity relationships and estrogen receptor- α -selective agonists. *J. Med. Chem.* **2000**, *43*, 4934–4947.

(44) Meyers, M. J.; Sun, J.; Carlson, K. E.; Marriner, G. A.; Katzenellenbogen, B. S.; Katzenellenbogen, J. A. Estrogen receptor-beta potency-selective ligands: structure–activity relationship studies of diarylpropionitriles and their acetylene and polar analogues. *J. Med. Chem.* **2001**, *44*, 4230–4251.

(45) Greschik, H.; Moras, D. Structure–activity relationship of nuclear receptor–ligand interactions. *Curr. Top. Med. Chem.* **2003**, *3*, 1573–1599.

(46) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.

(47) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **2004**, *25*, 1605–1612.

(48) Irwin, J. J.; Shoichet, B. K. ZINC—a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **2005**, *45*, 177–182.

(49) Sayers, E. W.; Barrett, T.; Benson, D. A.; Bolton, E.; Bryant, S. H.; Canese, K.; Chetvernin, V.; Church, D. M.; Dicuccio, M.; Federhen, S.; Feolo, M.; Geer, L. Y.; Helmberg, W.; Kapustin, Y.; Landsman, D.; Lipman, D. J.; Lu, Z.; Madden, T. L.; Madej, T.; Maglott, D. R.; Marchler-Bauer, A.; Miller, V.; Mizrachi, I.; Ostell, J.; Panchenko, A.; Pruitt, K. D.; Schuler, G. D.; Sequeira, E.; Sherry, S. T.; Shumway, M.; Sirotkin, K.; Slotta, D.; Souvorov, A.; Starchenko, G.; Tatusova, T. A.; Wagner, L.; Wang, Y.; John Wilbur, W.; Yaschenko, E.; Ye, J. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **2010**, *38*, D5–16.

(50) Gasteiger, J.; Rudolph, C.; Sadowski, J. Automatic generation of 3D-atomic coordinates for organic molecules. *Tetrahedron Comput. Methodol.* **1990**, *3*, 537–547.

(51) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: an open chemical toolbox. *J. Cheminf.* **2011**, *3*, 33.

1.2.3 Discussion

Dans cette étude, nous avons présenté la NRLiSt BDB, une banque d'évaluation dédiée à une famille de protéines très importante du point de vue biologique, les récepteurs nucléaires. Par rapport aux précédentes banques d'évaluations, la NRLiSt BDB présente trois caractéristiques majeures :

- (1) Lors de notre investigation bibliographique, pour qu'un composé soit retenu, il ne devait pas être seulement un ligand des RNs mais son profil pharmacologique (agoniste ou antagoniste) devait en plus être précisé ;
- (2) Toutes les informations sur les composés et leur activité ont été minutieusement inspectées manuellement par lecture des publications scientifiques concernées ;
- (3) Grâce aux deux points précédents, la NRLiSt BDB peut à la fois être utilisée comme une banque d'évaluation des méthodes de criblage virtuel et comme une source unique d'informations pour la découverte de nouveaux ligands des récepteurs nucléaires.

1.2.3.1 Séparation des jeux de données « agoniste » et « antagoniste »

Le choix de proposer des jeux de données séparés « agoniste » et « antagoniste » a été rationnellement décidé grâce aux constatations de notre étude précédente³⁶⁸. Ce parti pris est original par rapport à d'autres banques d'évaluation mais pas unique (jeux de données ER_alpha agoniste et antagoniste de la DUD⁴²³, bases de données GPCR Ligands Library, GLL, et GPCR Decoys Database GDD⁴⁸¹).

1.2.3.2 Sélection des RNs, des structures et des ligands à inclure dans la NRLiSt BDB

Pour pouvoir prétendre au titre de banque d'évaluation à la fois des méthodes de criblage virtuel basées sur les ligands et sur les structures, les jeux de données de la NRLiSt BDB devaient répondre à un critère simple : la disponibilité des ligands et des structures. Nous avons donc posé comme conditions d'inclusion pour les récepteurs nucléaires dans la NRLiSt BDB qu'ils possèdent au moins une structure expérimentale et plus d'un ligand agoniste et plus d'un ligand antagoniste. Seuls 27 RNs sur les 48 humains existant remplissaient ces critères.

Etant donné qu'il n'existe toujours pas de consensus pour déterminer quelle structure choisir pour un criblage virtuel lorsque plusieurs sont disponibles, nous avons décidé, afin d'éviter de possibles biais, de fournir toutes les structures humaines « holo » des RNs inclus dans la NRLiSt BDB, extraites de la PDB. Les conformations des sites de liaison étant sujet à de grandes variations selon la nature du ligand co-cristallisé⁴⁸⁰, nous avons annoté et classifié chaque structure selon le profil pharmacologique des ligands co-cristallisés (agoniste, antagoniste ou autre).

La sélection des ligands a représenté la partie la plus complexe, longue et fastidieuse de ce travail. Pour cela, différentes bases de données (PDB⁶⁸, IUPHAR⁴⁷⁶, pubmed, ChEMBL⁴²...), ont été utilisées non pas comme sources de ligands, mais comme sources de références bibliographiques pour pouvoir vérifier manuellement les données contenues dans les bases de données. Tous les ligands dont l'activité biologique en tant qu'agoniste ou antagoniste était renseignée ont été inclus dans la NRLiSt BDB. Nous avons donc décidé de ne pas conserver, d'une part, les ligands dont le profil pharmacologique n'avait pas été étudié (c'était notamment le cas dans les publications les plus anciennes) et, d'autre part, les ligands de profils pharmacologiques « alternatifs » : modulateurs, agonistes ou antagonistes partiels...

Enfin, le troisième élément présent dans chaque jeu de données est le decoy. Ces decoys ont été obtenus à l'aide de l'outil de génération automatique de decoys de la DUD_E⁴²⁷. Cet outil prend en compte le poids moléculaire, le coefficient de partition eau/octanol, le nombre de liaisons rotatives, le nombre de donneurs et d'accepteurs de liaisons hydrogènes et la charge nette, calculés à partir de tous les états de protonation du ligand dans des rangs de pH 6-8. Ces six critères sont utilisés de manière plus ou moins souple pour extraire de 3000 à 9000 composés de la ZINC⁴³. Ensuite, une recherche de similarité à l'aide du coefficient de Tanimoto permet de ne conserver qu'un quart de ces composés, et uniquement ceux les plus dissimilaires du ligand de référence, parmi lesquels 50 molécules sont choisies au hasard pour servir de decoys.⁴²⁷

Au final, la NRLiSt BDB inclus 27 RNs, divisés en 54 jeux de données, totalisant 9905 ligands (7853 agonistes et 2052 antagonistes), 458981 decoys et 339 structures expérimentales (266 liées à un agoniste, 17 liées à un antagoniste, 56 liées à un autre type de ligand) (Figure 75).

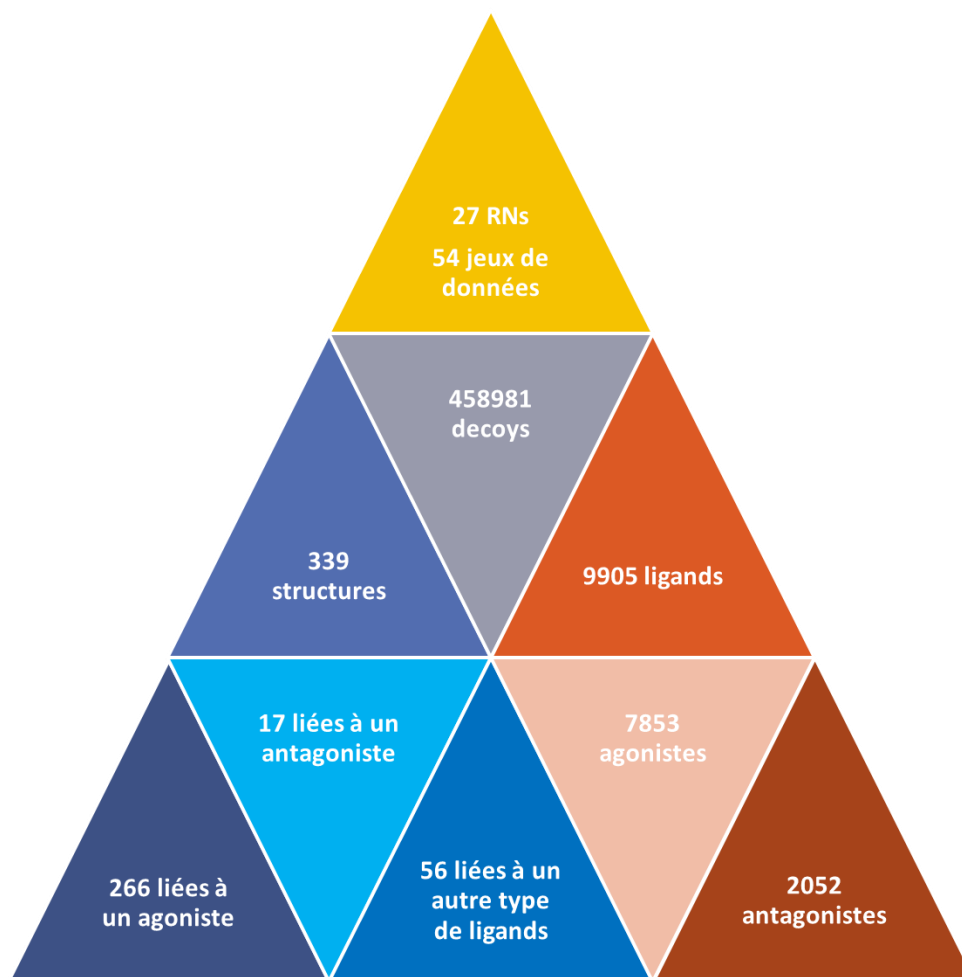


Figure 75. Représentation schématique des données contenues dans la NRLiSt BDB

1.2.3.3 Tentative de profilage des ligands agonistes et antagonistes à l'aide de descripteurs structuraux

Nous avons utilisé des descripteurs structuraux simples, calculés à l'aide du logiciel Dragon 6.0⁴⁸², pour rechercher des différences significatives expliquant le caractère agoniste ou au contraire antagoniste d'un ligand. C'était notamment le cas du poids moléculaire (PM) et du nombre de liaisons rotatives (nrotB), dont les valeurs moyennes différaient de manière significative entre les jeux de données agoniste et antagoniste pour de nombreux RNs. Ces deux descripteurs illustrant la flexibilité des ligands, celle-ci apparaît donc comme un critère potentiel de discrimination des ligands agonistes et antagonistes. Malheureusement, des résultats similaires n'ont pas pu être obtenus à l'aide des descripteurs du nombre de donneurs et d'accepteurs de liaisons hydrogène (HBD et HBA), du coefficient de partition eau/octanol (logP) ou de l'aire de surface polaire topologique (TPSA). L'étude des différences de

conformations lors de la liaison des agonistes et des antagonistes montre que de nombreux antagonistes peuvent être parfaitement superposés aux agonistes, et peuvent donc établir globalement les mêmes liaisons hydrogènes (ce qui explique l'absence de différence significative dans les valeurs moyennes de HBD et HBA). Cependant, comme cela a déjà été montré pour les récepteurs nucléaires TR⁴⁸³ et RAR_{alpha}⁴⁸⁴, certains antagonistes présentent des extensions volumineuses (expliquant les différences trouvées de PM et de nrotB) interdisant le repositionnement de l'hélice 12 dans sa conformation agoniste (Figure 76).

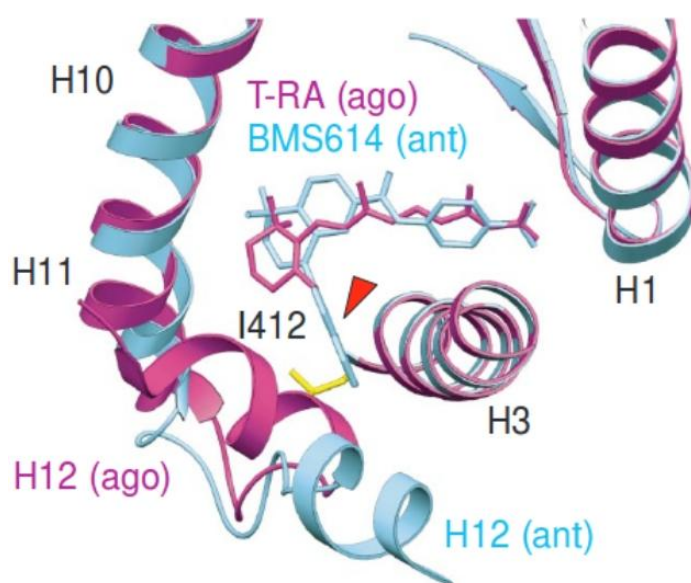


Figure 76. Superposition des sites de liaison de RAR α co-cristallisé avec un ligand antagoniste (bleu) et du RAR γ co-cristallisé avec un agoniste (rose) illustrant le clash stérique (flèche rouge) se produisant entre l'extension du ligand antagoniste et l'hélice 12 en position agoniste⁴⁸⁰

Ces descripteurs peuvent aussi être utilisés en corrélation avec les données d'activité et de liaison (renseignées chaque fois que possible). La NRLiSt BDB constitue donc un support robuste pour des études de relations structure-activités ou structure-propriétés sur les ligands des récepteurs nucléaires. Notre base de données peut notamment être employée par les chimistes médicaux pour guider le choix des ligands à synthétiser en priorité.

1.2.3.4 Présentation du site web de la NRLiSt BDB

Nous avons décidé de rendre accessible gratuitement et sans aucune condition notre banque d'évaluation et l'ensemble des données biologiques collectées au cours de la construction de la NRLiSt BDB. Pour cela, nous avons conçu un site internet, disponible à l'adresse suivante : <http://nrlist.drugdesign.fr>.

Dans ce site internet, les données biologiques des ligands et des structures sont présentées sous forme de tableaux. Ainsi, pour les 9905 ligands de la NRLiSt BDB, les identifiants (ZINC, ChEMBL, or CID ID), leurs profils pharmacologiques (agoniste ou antagoniste), leurs noms (nom commun ou nom attribué dans la publication de référence), les données de liaison et d'activité, les références bibliographiques correspondantes et les réactivités croisées entre les différents récepteurs nucléaires sont renseignés. Similairement, pour les 339 structures constituant la NRLiSt BDB, l'identifiant PDB, la résolution, les informations sur les ligands co-cristallisés (profil pharmacologique et nom) et l'organisme d'origine de la protéine sont rassemblés dans un tableau pour chaque RN.

Enfin, l'intégralité des 54 jeux de données constituant la NRLiSt BDB sont disponibles au téléchargement. Pour chaque jeu de données, 3 répertoires sont proposés : « ligands », « decoys », et « targets » (« cibles »). Ainsi dans les répertoires « ligands » et « decoys », les ligands agonistes et antagonistes ou les decoys peuvent être téléchargés au format MOL2 ou SMILES. Le répertoire « ligand » contient en plus le fichier des descripteurs 2D et 3D calculés à l'aide du logiciel Dragon 6.0⁴⁸² pour chaque ligand. Le dossier des structures (« targets ») regroupe toutes les structures expérimentales pour chaque RN. Pour chacune d'entre elles, la structure du RN originale telle que téléchargée dans la PDB est associée à celle sans ligand (X_WL.pdb) et préparée pour le docking avec l'outil Dock-prep de CHIMERA⁴⁸⁵ (X_prot.mol2). La structure du ligand protoné et chargé correctement à l'aide des charges partielles de Gasteiger (X_ligand_prot.mol2) ainsi que le protomol généré pour le docking à l'aide du logiciel Surflex sont aussi fournis.

Nous espérons que ce site web sera utile et utilisé par le plus grand nombre de personnes, que ce soit des bioinformaticiens souhaitant réaliser des évaluations de leurs méthodes de criblage virtuel ou encore par toute personne souhaitant mettre à profit les données de liaison et d'activité présentées dans la NRLiSt BDB pour mener des projets de chimie pharmaceutique.

1.2.4 Analyse critique de l'étude

1.2.4.1 Critique de la base de données ChEMBL

Afin d'avoir accès à tous les ligands des récepteurs nucléaires ayant été décrit dans la littérature nous avons utilisé différentes bases de données. La ChEMBL s'est révélée être particulièrement intéressante, notamment grâce à la grande quantité de données intégrée dans cette base de données et la simplicité d'obtention des informations recherchées. Ainsi, au mois d'avril 2014, la ChEMBL regroupait des informations sur plus de 1,5 millions de composés et plus de 9000 cibles biologiques. Pour chaque cible, les composés sont classifiés selon le type d'activité référencée dans la littérature (par exemple : EC50, IC50, Ki, activité, inhibition, ...). Cependant, le profil pharmacologique du ligand en tant qu'agoniste, antagoniste ou autre n'étant pas systématiquement renseigné, il nous était impossible d'intégrer directement les composés proposés pour chaque RN par la ChEMBL dans la NRLiSt BDB. Nous avons donc vérifié dans les références bibliographiques correspondantes, pour chaque ligand proposé par la ChEMBL, s'il s'agissait réellement d'un ligand pour le RN concerné et lorsque c'était le cas, si son profil pharmacologique était renseigné. Même si cette procédure s'est avérée très longue, elle s'est aussi révélée très utile et bien plus précise. En effet, nous avons rencontré de nombreux ligands assimilables à des faux positifs. Par exemple, deux composés (CHEMBL1961797 et CHEMBL1961794) étaient systématiquement proposés comme ligands de tous les récepteurs nucléaires. Or, la publication associée ⁴⁸⁶, ne fait état que de leur caractère agoniste pour les récepteurs Rev-erb α et β et précise même que « aucun composé n'a démontré d'activité sur les autres récepteurs nucléaires » ⁴⁸⁶. Un autre exemple tout aussi frappant concerne un inhibiteur du facteur Xa présenté dans la ChEMBL comme ligand du récepteur ER α alors que dans la référence correspondante ⁴⁸⁷, aucune activité de ce ligand pour ER α n'est rapportée et celui-ci n'est même pas mentionné dans la publication. Ces erreurs sont donc très dommageables lorsque les données de la ChEMBL sont utilisées sans vérification manuelle et ne sont malheureusement pas isolées (Tableau 17).

Récepteur	Nombre de publications	Nombre de composés	Cause de l'erreur		
			Isoforme différent	Récepteur différent	Autre
CAR	2	5		X	
ER α	1	4	X		
	61	410		X	
	14	61			X
ER β	30	191		X	
	8	34			X
ERR α	1	4	X		
	5	66		X	
FXR α	10	40		X	
	2	2			X
GR	41	218		X	
	23	30			X
LXR α	10	21		X	
	1	4			X
LXR β	1	9	X		
	7	14		X	
MR	20	96		X	
	1	3			X
PPAR α	10	198	X		
	6	12		X	
	6	39			X
PPAR β	32	328	X		
	12	34		X	
	9	16			X
PPAR γ	13	108	X		
	16	93		X	
	10	52			X
PR	19	56		X	
	4	4			X
PXR	3	5		X	
	28	65			X
ROR α	2	2		X	
ROR γ	1	2		X	
RXR α	10	32		X	
	2	37			X
RXR β	3	13		X	
RXR γ	3	13		X	
TOTAL	427	2321			

Tableau 17. Classification du nombre de publications et de composés associés à tort à un récepteur nucléaire en fonction de la cause de l'erreur, pour 19 des 27 RNs de la NRLiSt BDB

1.2.4.2 Déséquilibre inter- et intra- jeux de données

Les jeux de données de la NRLiSt BDB présentent un large déséquilibre du nombre de ligands et de structures, illustrant les différences d'intérêt porté à chacun des récepteurs nucléaires. Ainsi, les ligands et les structures des récepteurs PPARs représentent respectivement plus de 40% et plus de 30% du nombre total de ligands et de structures de la NRLiSt BDB, car il s'agit de récepteurs très étudiés pour leur potentiel thérapeutique⁴⁸⁸ en tant qu'anti-diabétiques⁴⁸⁹, hypolipémiants⁴⁹⁰ ou encore anti-inflammatoires⁴⁹¹. Cependant, au sein même de ses récepteurs, le nombre d'agonistes et d'antagonistes décrits dans la littérature est largement inégal. A l'opposé, beaucoup de récepteurs nucléaires restent inexplorés, et l'absence de données suffisantes pour 21 de ces récepteurs nous a notamment obligé de les exclure, pour le moment, de la NRLiSt BDB.

1.2.4.3 Diversité structurale de la NRLiST BDB

Un des critères utilisé pour noter la qualité d'une banque d'évaluation est la diversité structurale proposée par celle-ci⁴⁰⁸ et constitue l'un des reproches qui pourrait être adressé à l'encontre de la NRLiSt BDB. En effet, les récepteurs nucléaires forment une famille de protéines hautement conservée et donc la NRLiSt BDB est focalisée sur une petite partie de l'ensemble de l'espace chimique possible. Cependant, comme la recherche de ligands a été menée de la manière la plus exhaustive possible, la NRLiSt BDB couvre l'espace chimique le plus large possible en ce qui concerne les récepteurs nucléaires. Ainsi, en utilisant une valeur seuil de coefficient de Tanimoto fixée à 0,5, en moyenne 10 clusters sont obtenus pour chaque jeu de données. Pour seulement 10 jeux de données sur 54 moins de trois clusters sont formés, car le nombre de ligands décrits dans la littérature n'est pas suffisant.

1.2.4.4 Améliorations

La construction de la NRLiSt BDB a représenté un travail très ambitieux mais aussi complexe et long, aux vues de la grande quantité de données à traiter et à intégrer dans la banque d'évaluation. Nous pensons avoir proposé une banque de qualité pouvant être utile pour la communauté scientifique. Cependant, des améliorations restent possibles et devraient être intégrées dans la mise à jour de la banque que nous prévoyons d'effectuer au début de l'année 2015. Ces améliorations concernent d'une part les ligands actifs et d'autre part les decoys.

En effet, nous avons choisi de ne pas inclure les ligands présentant un profil pharmacologique alternatif (modulateurs, agonistes et antagonistes partiels). Nous avons bien conscience que ces données représentent aussi une source d'informations essentielles pour la découverte de nouveaux composés ciblant les récepteurs nucléaires. Nous sommes donc en pleine réflexion pour décider comment intégrer ces données dans la NRLiSt BDB.

Les decoys de la NRLiSt BDB ont été obtenus avec l'outil de génération automatique de la DUD_E. Cependant, le caractère réel de non ligands ou decoys de ces composés reste toujours soumis à suspicion. Il nous a donc été suggéré d'intégrer des composés que nous avons précédemment exclus de la NRLiSt BDB à cause de leur absence d'activité sur un RN donné en tant que decoys de cette cible. Ceci va représenter un nouveau travail bibliographique important, et d'après les connaissances acquises lors de la construction de la NRLiSt BDB, nous prévoyons que le ratio actifs / decoys ainsi obtenu sera beaucoup plus faible que l'actuel (1/50). Cependant, l'utilisation de decoys expérimentaux devrait permettre d'améliorer significativement la qualité de la NRLiSt BDB, et nous sommes donc prêts à nous engager dans cette perspective.

1.2.5 Conclusion

La NRLiSt BDB est actuellement la base de données la plus exhaustive en ce qui concerne les ligands et les structures des récepteurs nucléaires mais aussi l'une des banques d'évaluation les plus fiables grâce à la vérification manuelle des données et la prise en compte du profil pharmacologique des ligands lors de sa construction. Ceci devrait conférer à la NRLiSt BDB le statut de base de données de référence pour l'évaluation des méthodes basées sur les ligands ou sur les structures des récepteurs nucléaires mais aussi pour la compréhension des mécanismes contrôlant la modulation et l'action de ses récepteurs ainsi que pour guider la découverte de nouveaux composés ciblant les RNs.

1.3 Importance du profil pharmacologique du ligand co-cristallisé et utilisation de « decoys ligands »

1.3.1 Introduction

Parmi les banques d'évaluation développées pour garantir une évaluation robuste des méthodes de criblage virtuel^{304, 418, 492, 493}, couramment acceptées en complément de criblages expérimentaux dans les processus de R&D de nouveaux médicaments³², la DUD⁴²³ et la DUD_E⁴²⁷ incluent respectivement 8 et 11 récepteurs nucléaires. Cependant, seul le récepteur nucléaire ER_alpha dans la DUD possède 2 jeux de données séparés selon le profil pharmacologique des ligands : ER_alpha agoniste et ER_alpha antagoniste. Nous avons alors formulé l'hypothèse selon laquelle la prise en compte du profil pharmacologique du ligand, réalisée pour le récepteur ER_alpha dans la DUD, devrait améliorer la qualité des banques d'évaluation. En effet, au niveau conformationnel, des changements dans la structure du récepteur nucléaire sont observés lorsqu'un ligand se lie dans le site actif de celui-ci, et ces changements sont différents selon le profil pharmacologique de ce ligand⁴⁸⁰. Dans la base de données NRLiSt BDB construite dans notre laboratoire⁴⁹⁴, 2 jeux de données, un agoniste et un antagoniste, sont disponibles pour les 27 récepteurs nucléaires présentant plus d'un ligand agoniste, plus d'un ligand antagoniste et au moins une structure expérimentale. Ces jeux de données regroupent des ligands agonistes et antagonistes, expérimentalement décrits, toutes les structures holo humaines disponibles dans la PDB et les decoys correspondants obtenus avec l'outil de génération automatique de la DUD_E⁴²⁷.

Nous avons donc décidé d'étudier l'impact sur les performances d'une méthode de criblage virtuel basée sur les structures, Surflex-Dock, des différents partis-pris lors de la construction de la NRLiSt BDB. Nous avons notamment étudié l'influence de l'utilisation de jeux de données séparés en fonction du profil pharmacologique, en comparant les enrichissements obtenus lorsque les ligands agonistes et les ligands antagonistes sont séparés dans deux jeux de données distincts ou regroupés dans un seul. Nous avons aussi étudié l'influence du profil pharmacologique du ligand co-cristallisé dans la structure de référence sur l'enrichissement, puisque le choix de la structure de référence est reconnu comme étant de haute importance dans les méthodes de docking, et peut influencer sur l'enrichissement et la précision du docking⁴⁹⁵⁻⁴⁹⁸. Enfin, pour les récepteurs nucléaires pour lesquels suffisamment de données étaient disponibles, nous avons exploré l'utilisation de « ligands decoys », c'est-à-dire des composés

capable de se lier aux récepteurs (« ligands ») mais incapable de produire l'activité biologique recherchée (« decoy »), à la place de decoys putatifs comme ceux de la DUD ou de la DUD_E.

Ce travail permet donc d'illustrer différentes lignes directrices qui peuvent être utilisées pour améliorer la qualité des études d'évaluation en prenant en compte le profil pharmacologique des ligands, et ce à différents niveaux, et en utilisant des ligands decoys expérimentaux.

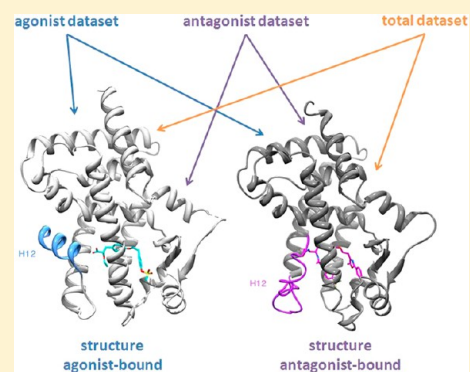
1.3.2 Publication

Importance of the Pharmacological Profile of the Bound Ligand in Enrichment on Nuclear Receptors: Toward the Use of Experimentally Validated Decoy Ligands

Nathalie Lagarde, 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

ABSTRACT: The evaluation of virtual ligand screening methods is of major importance to ensure their reliability. Taking into account the agonist/antagonist pharmacological profile should improve the quality of the benchmarking data sets since ligand binding can induce conformational changes in the nuclear receptor structure and such changes may vary according to the agonist/antagonist ligand profile. We indeed found that splitting the agonist and antagonist ligands into two separate data sets for a given nuclear receptor target significantly enhances the quality of the evaluation. The pharmacological profile of the ligand bound in the binding site of the target structure was also found to be an additional critical parameter. We also illustrate that active compound data sets for a given pharmacological activity can be used as a set of experimentally validated decoy ligands for another pharmacological activity to ensure a reliable and challenging evaluation of virtual screening methods.



INTRODUCTION

Virtual screening is now widely accepted as a complement to bioactivity screening,¹ and structure-based (SBVLS) and ligand-based (LBVLS) virtual ligand screening methods are commonly integrated into drug discovery processes.² High-quality benchmarking data sets are needed to warrant robust evaluation of these methods and ensure their reliability. Different benchmarking data sets have been developed over the years,^{3–6} among which the Directory of Useful Decoys (DUD)⁷ and its enhanced version (DUD·E)⁸ are considered as the current gold standard benchmarking databases. These databases include respectively eight and 11 nuclear receptors (NRs), but only estrogen receptor alpha (ER_alpha) in the DUD is presented with two distinct data sets according to the ligand agonist or antagonist pharmacological profile. Taking into account this profile information should improve the quality of the benchmarking data sets since ligand binding can induce conformational changes in the nuclear receptor structure and such changes may vary according to the agonist or antagonist ligand profile.⁹ The recently released Nuclear Receptors Ligands and Structures Benchmarking DataBase (NRLiSt BDB)¹⁰ comprises 27 NRs for which more than one agonist and one antagonist ligand and at least one experimental structure are available, as well as two separate agonist and antagonist data sets. These data sets regroup experimentally characterized agonist and antagonist ligands, all corresponding to human holo structures available in the Protein Data Bank (PDB), and their corresponding decoys as provided by the DUD·E decoy generation tool.⁸

In the present work, we studied the impact on enrichment of different features of the NRLiSt BDB using an SBVLS method, Surflex-Dock (SF). We evaluated the influence of

using separate data sets by comparing SF performances in terms of enrichment when considering agonist ligands and antagonist ligands split into two distinct data sets or mixed altogether in a single one as in the DUD·E. Since the choice of the structure of reference is of major importance in terms of enrichment and docking accuracy,^{11–14} we decided to assess the influence of the pharmacological profile (agonist or antagonist) of the ligand bound in the binding site of the reference structure on enrichment. Also, for those NRs for which sufficient data were available, we explored the impact of using experimentally confirmed decoy ligands instead of putative DUD-like decoys on enrichment. This work brings up several guidelines that could be used to enhance the quality of benchmarking studies by taking into account the pharmacological profiles of the ligands and using experimentally validated decoy ligands.

EXPERIMENTAL METHODS

NRLiSt BDB. The NRLiSt BDB is a public benchmarking data set designed for evaluation of both structure-based and ligand-based methods. The NRLiSt BDB is dedicated to the NRs and contains an agonist data set and an antagonist data set for the 27 targets (out of the 48 known NRs) for which more than one agonist ligand, one antagonist ligand, and at least one experimental structure are available. Each data set consists of three elements: all of the available human holo PDB structures (except for RXR_gamma, for which only one apo structure was available), all of the ligands found to be agonists or antagonists

Received: May 21, 2014

Published: September 24, 2014

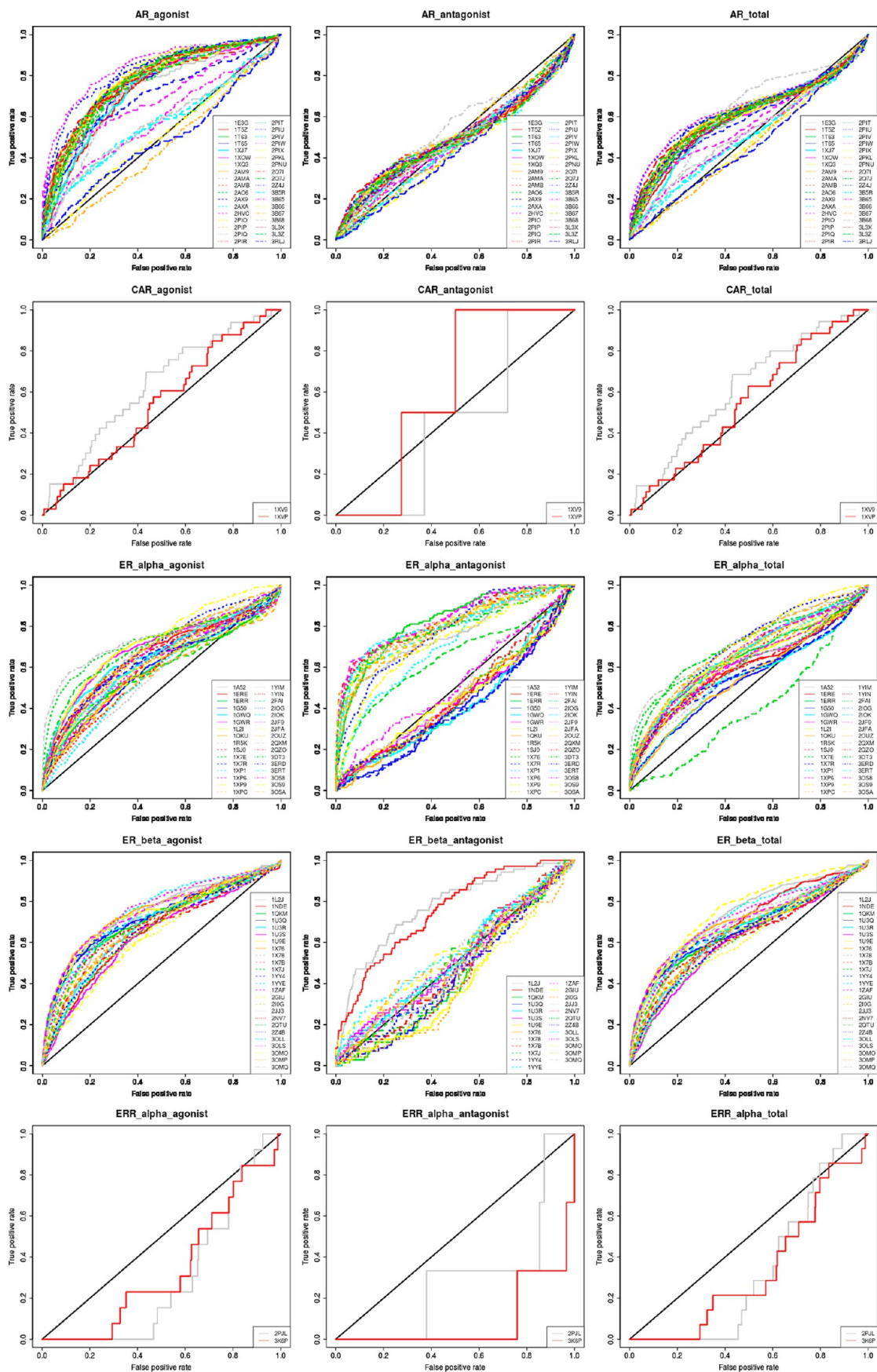


Figure 1. continued

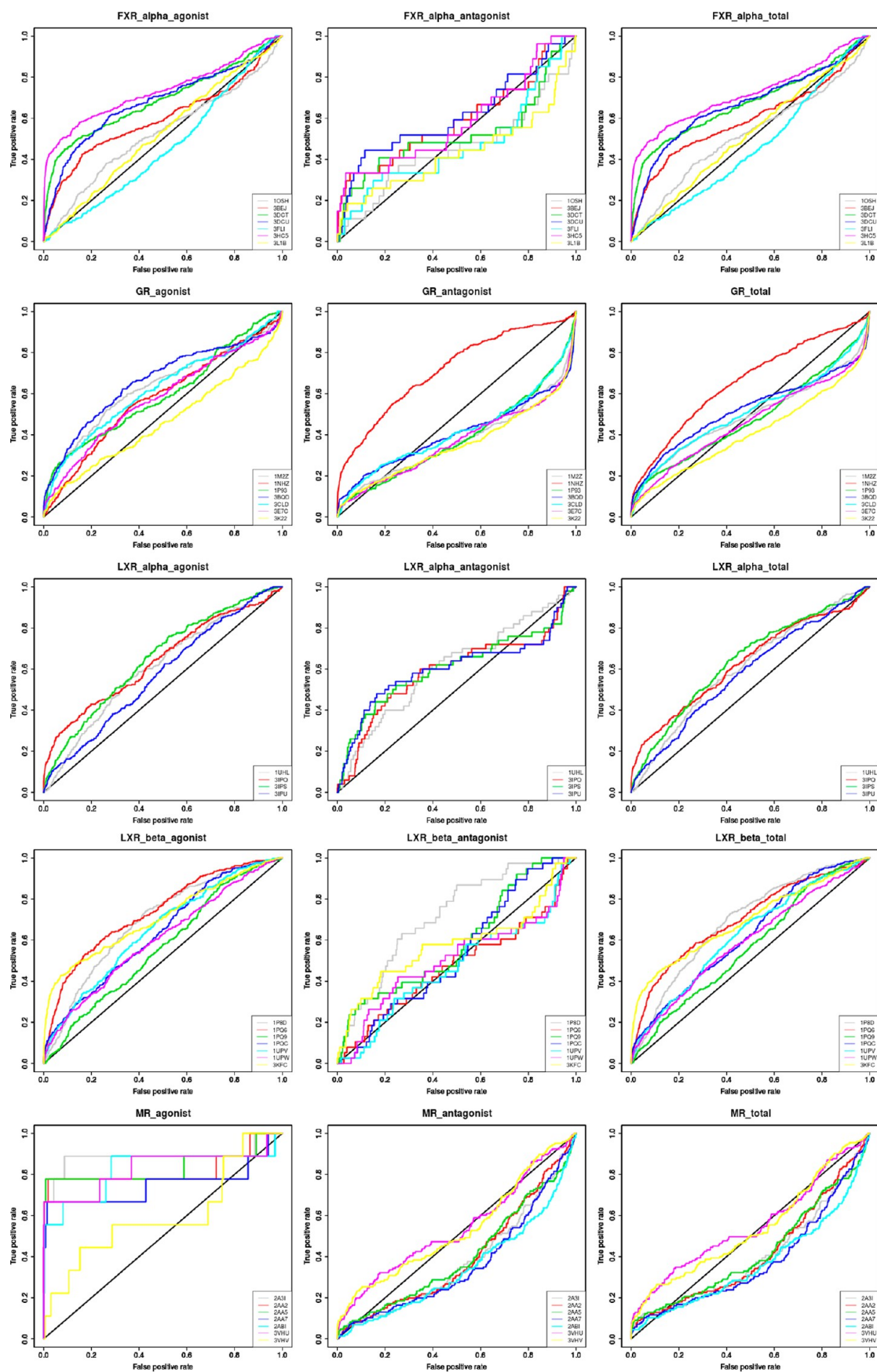


Figure 1. continued

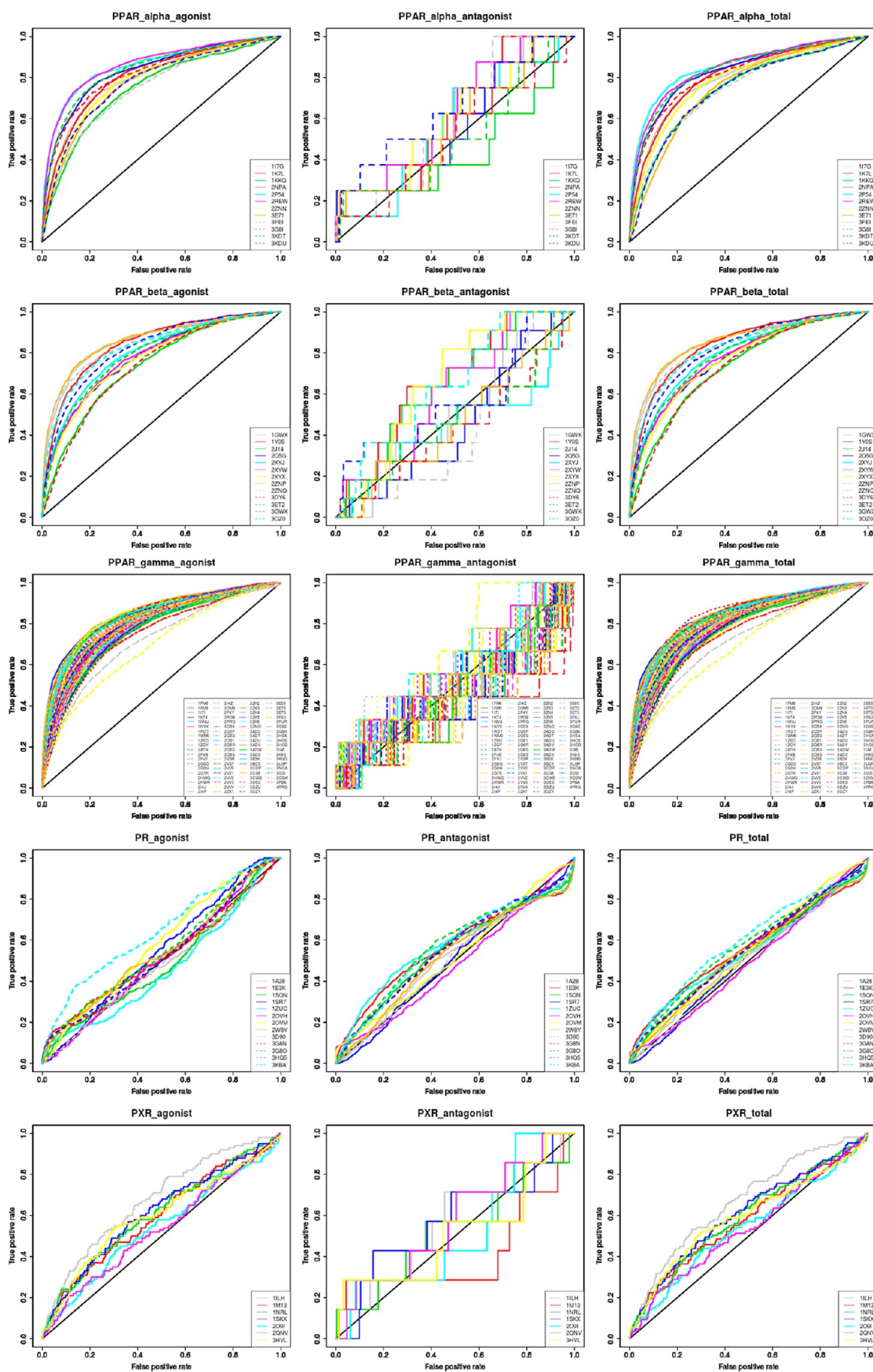


Figure 1. continued

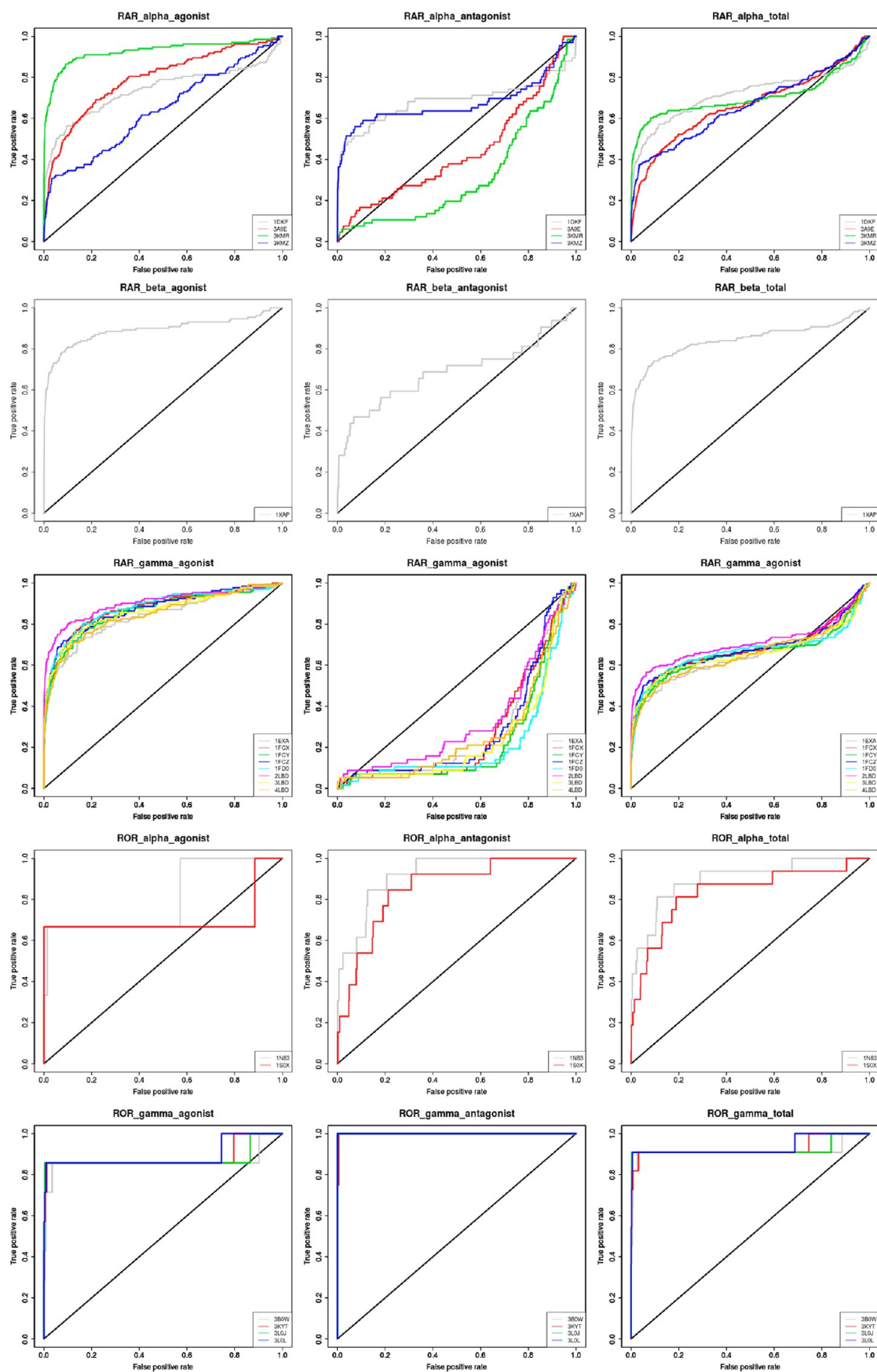


Figure 1. continued

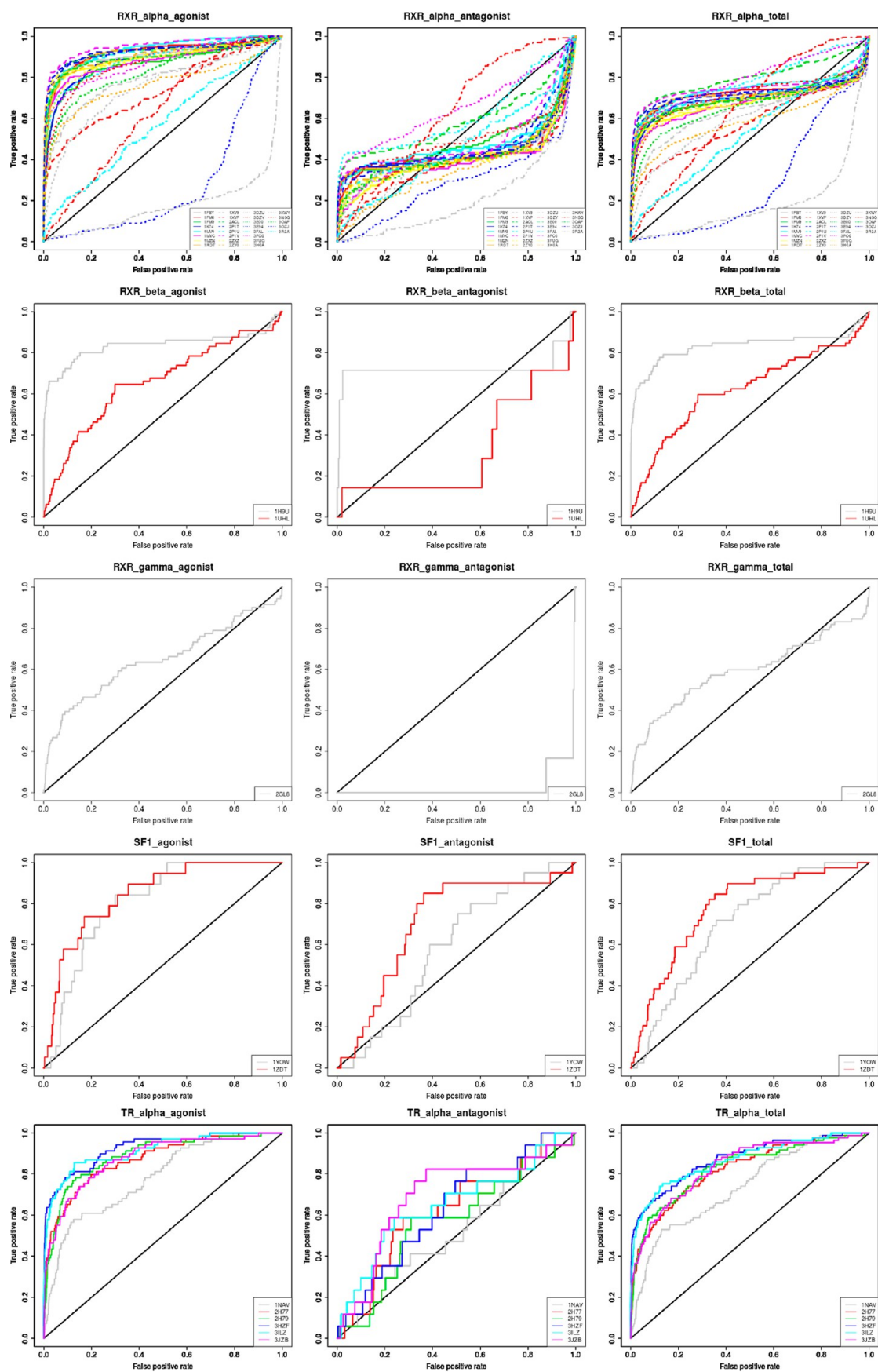


Figure 1. continued

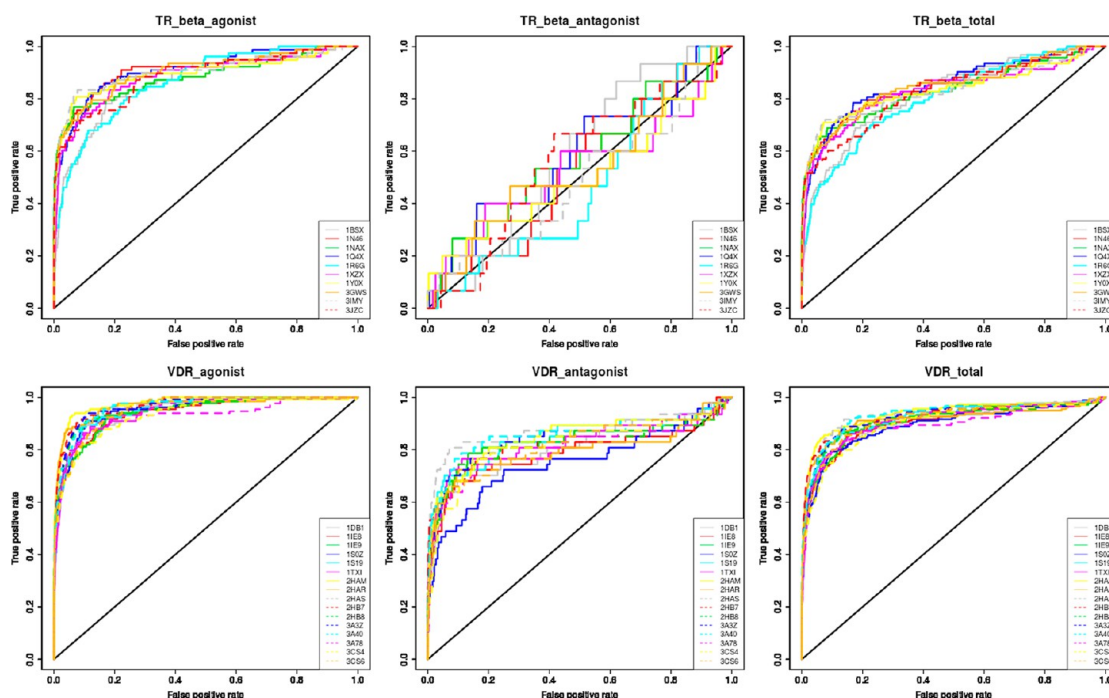


Figure 1. ROC curves obtained using Surflex-Dock with the agonist data set, the antagonist data set, or the total data set of a given NR on all structures composing that NR data set.

in the scientific literature, and their corresponding computed decoys obtained using the DUD-E decoy generation tool.⁸ The NRLiSt BDB comprises 7853 actives, 458 981 decoys, and 339 structures divided into 54 data sets of various sizes in terms of active ligands (ranging from two ligands for CAR antagonists to 1820 ligands for PPAR_gamma agonists) and available structures (from one for RAR_beta and RXR_gamma to 80 for PPAR_gamma). The NRLiSt BDB was downloaded from the Web site <http://nrlist.drugdesign.fr>.

Preparation of Ligands and Decoys. All of the ligands and decoys provided in MOL2 format in the NRLiSt BDB were used in this study. Agonists and antagonists and their corresponding decoys were considered in two separate data sets. A third data set, named the “total” data set, was constructed for each NR by gathering together all of ligands (either agonist or antagonist) and their corresponding decoys.

Surflex-Dock. SF is based on a modified Hammerhead fragmentation/reconstruction algorithm to dock compounds flexibly into the binding site.¹⁵ The query molecule is decomposed into rigid fragments that are superimposed on the Surflex-protomol, i.e., molecular fragments covering the entire binding site. The docking poses are evaluated by an empirical scoring function. For each structure, the binding site has been defined as 4 Å around the cocrystallized ligand. Surflex minimization options were used before docking (+premin) on the ligands and after docking (+remin) on all of the atoms of the binding site. In this study, Surflex-Dock version 2.5 was used for all of the calculations.

Performance Metrics. All of the enrichment graphs were produced with the statistical and graphical tool R (<http://www.r-project.org/>). The ROCR package¹⁶ was used to plot receiver operating characteristic (ROC) curves, and the Wilcoxon–Mann–Whitney algorithm was used for the ROC area under the curve (AUC) calculations. Enrichment factors (EFs) were computed as follows:

$$EF_{x\%} = \frac{N_{\text{experimental}}^{x\%}}{N_{\text{actives}}^{x\%}}$$

RESULTS

Impact of Using Separate Data Sets on Overall Enrichment. DUD and its enhanced version DUD-E are the current standard benchmarking databases. Interestingly, one constitutional change between these two versions is that the two original ER_alpha data sets (ER_alpha_agonist and ER_alpha_antagonist) of the first release have now been merged into a single ER_alpha data set. The NRLiSt BDB is composed of two separate data sets for each NR, one agonist data set and one antagonist data set. To evaluate the relevance of this choice, for each NR we analyzed the performance on enrichment of a structure-based virtual screening method, Surflex-Dock (SF), on all of the experimental structures composing the data set, using the two separated agonist and antagonist data sets and the “total” data set gathering both the agonist and antagonist data sets together (Figures 1 and 2 and Table 1).

For 10 out of the 27 NRs (AR, ER_beta, GR, MR, RAR_beta, RAR_gamma, ROR_gamma, RXR_alpha, RXR_beta, and VDR), the mean AUC obtained by screening a separate data set on all of the structures available for a given NR (the agonist data set for all but VDR and the antagonist data set for VDR) was significantly superior to the mean AUC obtained with the total data set according to a Wilcoxon test.¹⁷ This trend was strongly confirmed when focusing on single AUC values. For every NR, the best signal was always associated with a separate data set (the agonist data set for 21 NRs and the antagonist data set for six NRs). For 17 out of the 25 NRs for which more than one experimental structure of the protein was available, the structure that provided the best AUC was the same using separate data sets or the total data set. The separate data sets provided

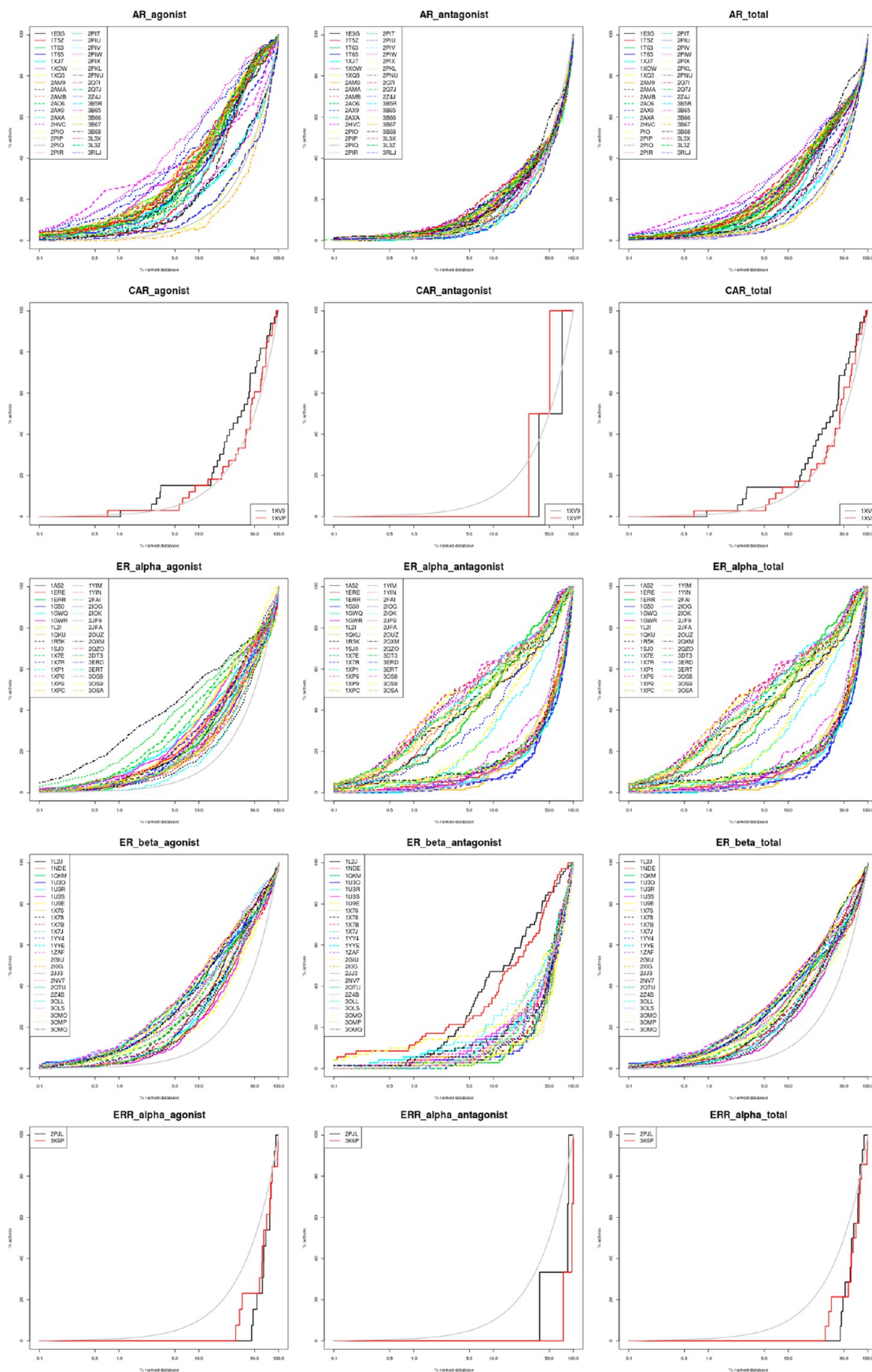


Figure 2. continued

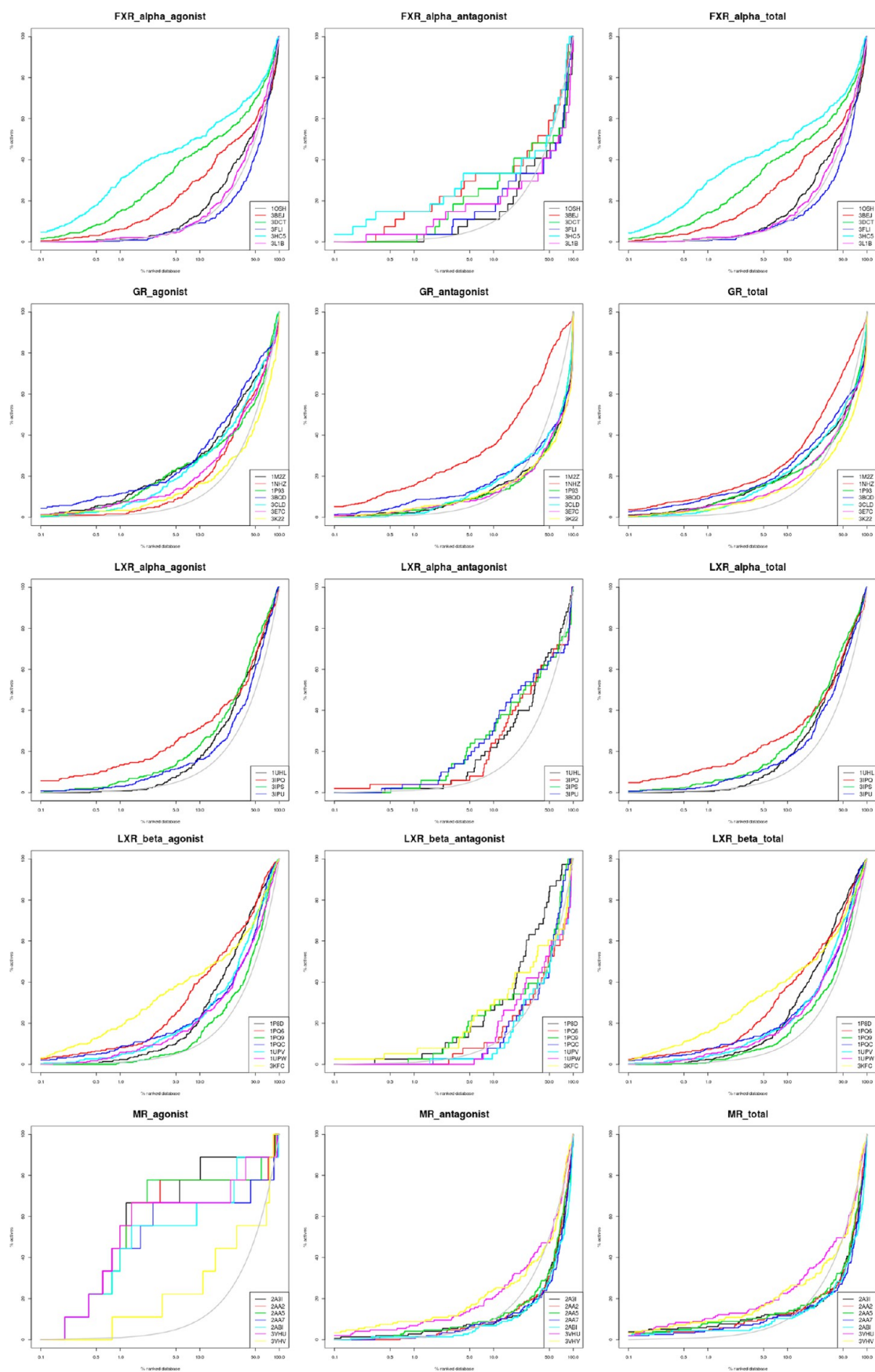


Figure 2. continued

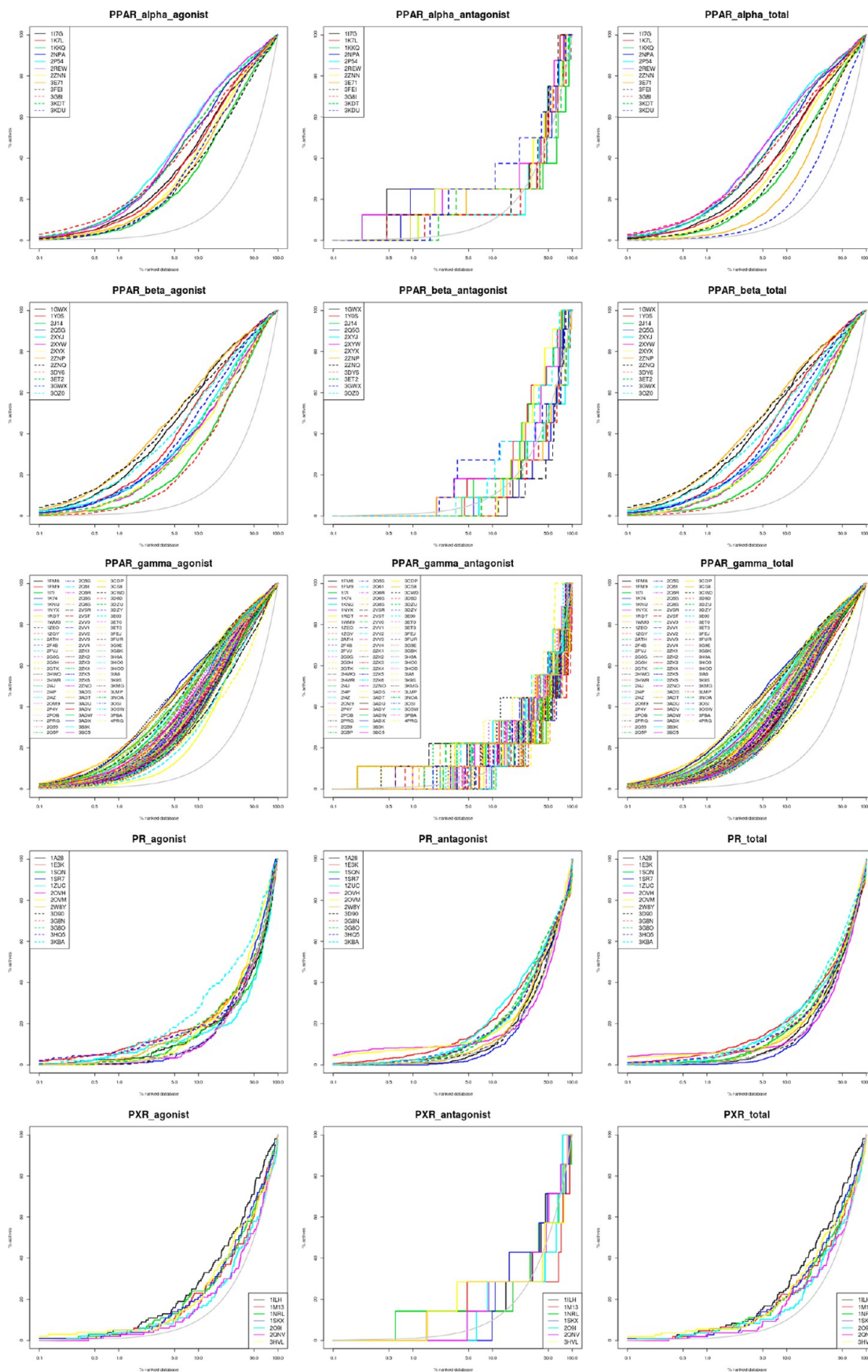


Figure 2. continued

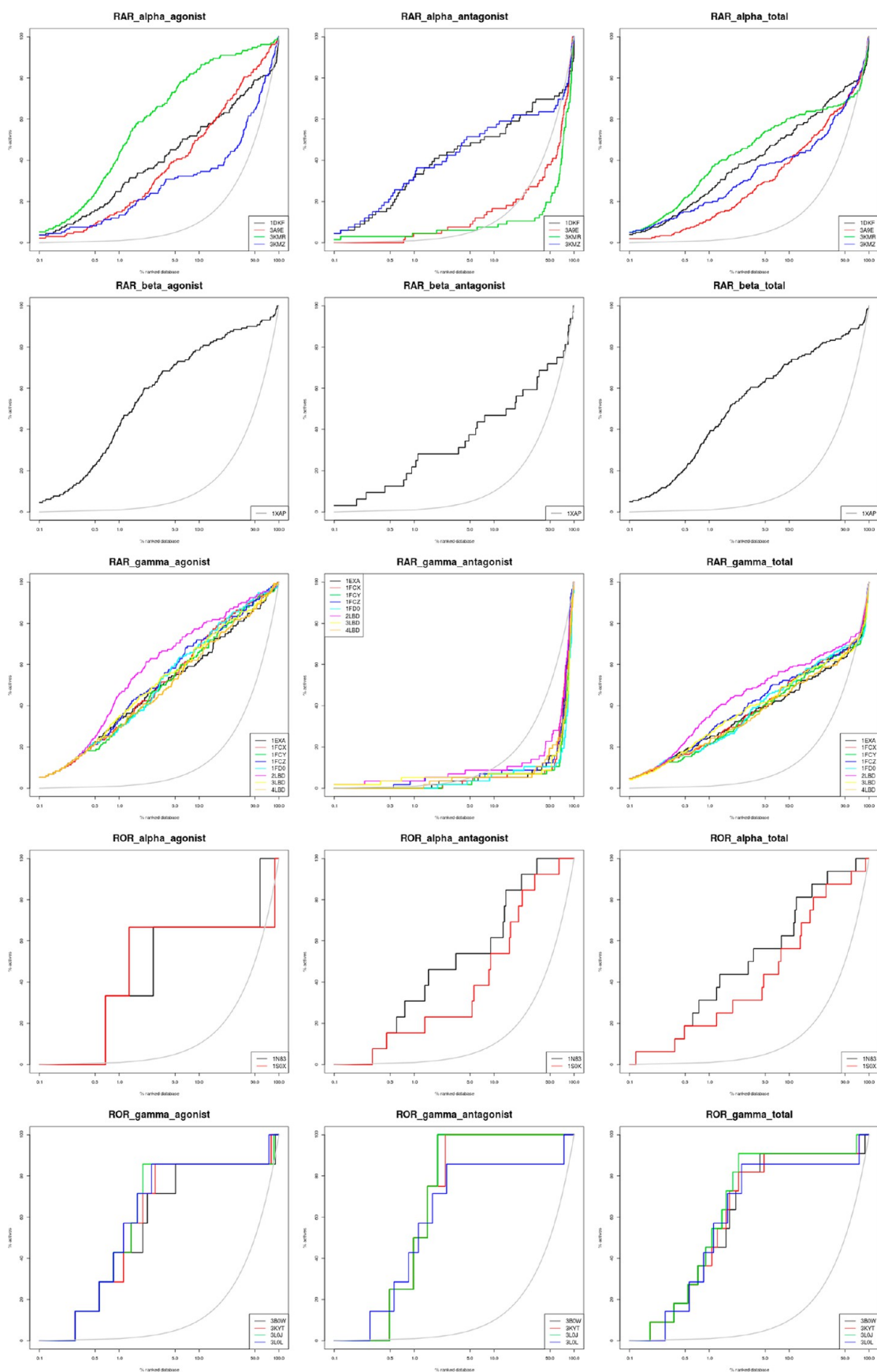


Figure 2. continued

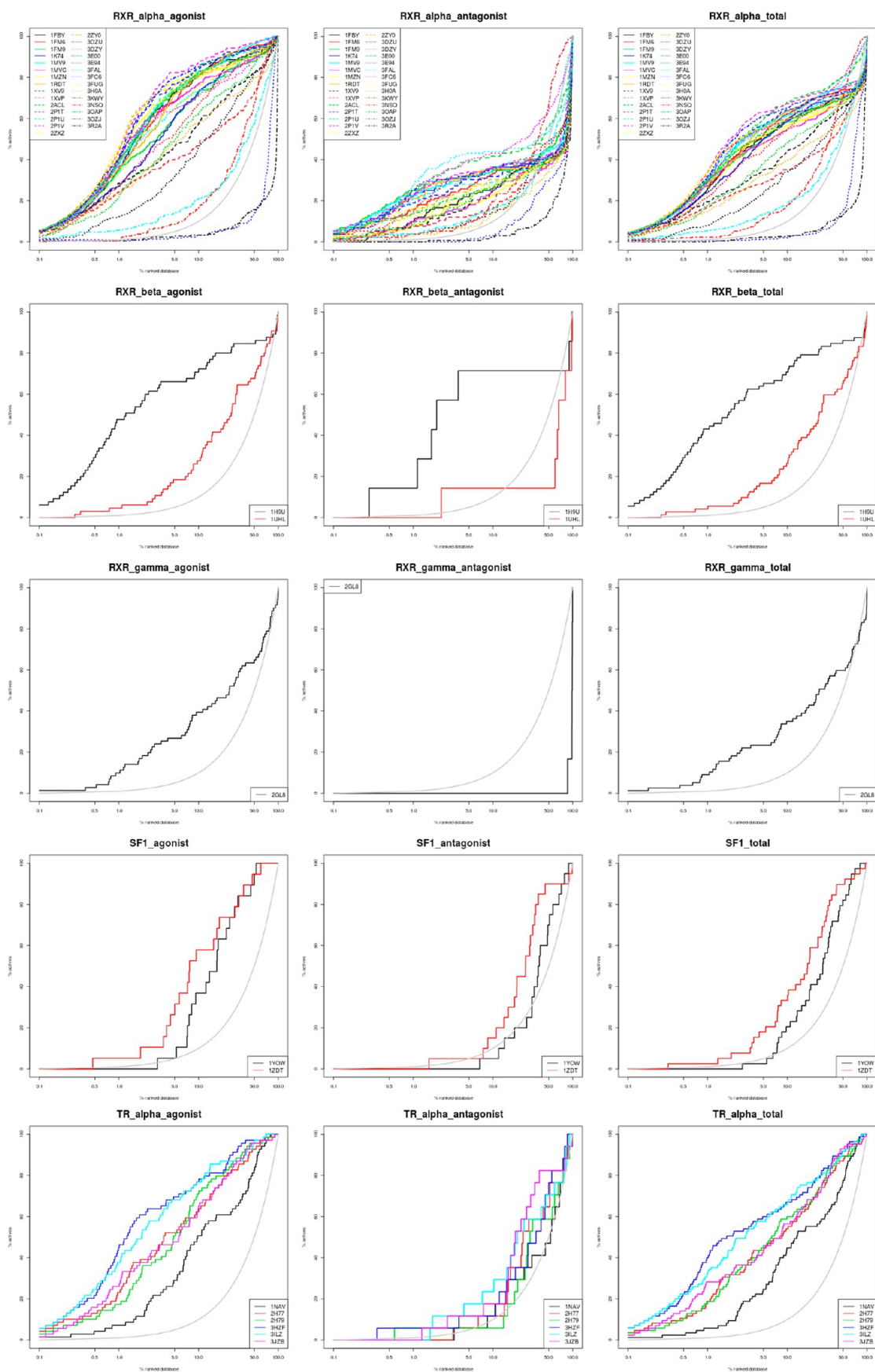


Figure 2. continued

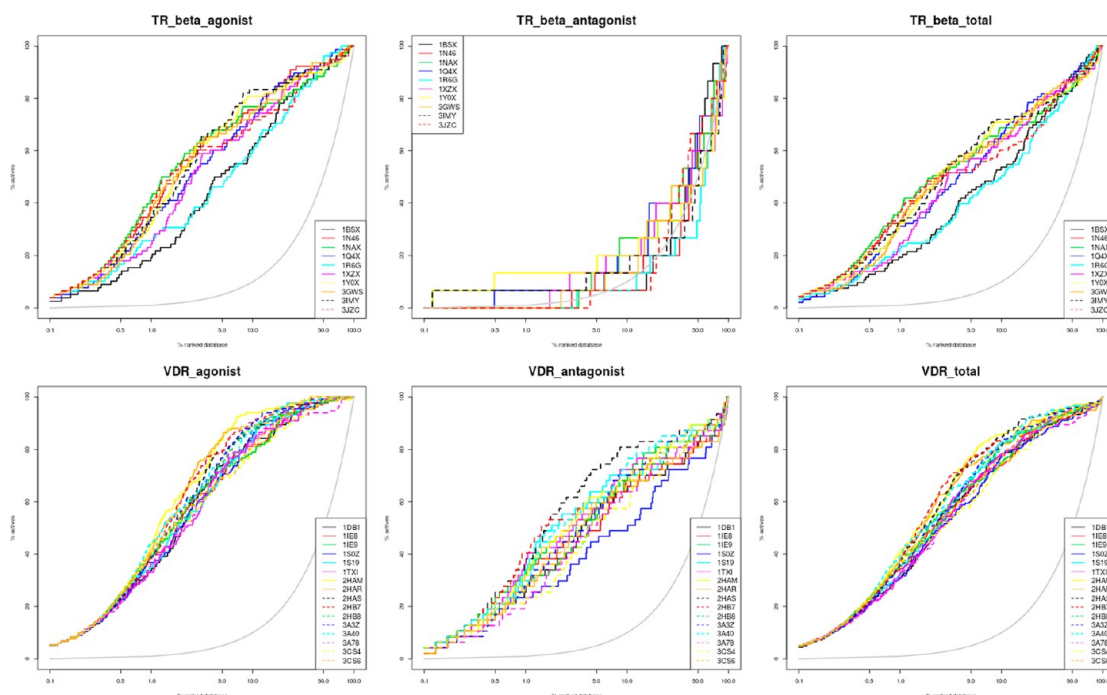


Figure 2. Enrichment graphs obtained using Surflex-Dock with the agonist data set, the antagonist data set, or the total data set of a given NR on all structures composing that NR data set.

the best AUC for 317 out of the 339 structures of the NRLiSt BDB. For seven out of the 27 NRs, all of the AUC values obtained with one of the separate data sets were superior to the best AUC obtained with the total data set. For 18 out of the 27 NRs, for at least 50% of the structures, the AUC values obtained with one of the separate data sets were superior to the best AUC obtained with the total data set for that target.

Impact of Using Separate Data Sets on Early Enrichment. Similarly, for 24 out of the 27 NRs, the best early enrichment was obtained using a separate data set (the agonist data set for 18 NRs and the antagonist data set for six NRs). For only one NR (ERR_alpha) was there no early enrichment with either the separate data sets or the total data set. For 16 out of the 25 NRs with more than one structure, the structure that provided the best early enrichment was the same using one of the separate data sets or using their combination. Over the 339 experimental structures tested in this study, the data set leading to the best early enrichment was one of the two separate data sets in 289 cases and their combination in 46 cases.

Importance of the Bound Ligand. We wanted to investigate the importance of the pharmacological profile of the cocrystallized ligand in the binding site of the structure on the performances in enrichment using SF with separate data sets. For this part of the interpretation, we focused on the NRs for which at least one agonist-bound and one antagonist-bound structure were available (ER_alpha, ER_beta, GR, MR, PPAR_alpha, PPAR_gamma, PR, RAR_alpha, ROR_gamma, RXR_alpha). We thus compared the performances obtained with SF in enrichment using the agonist data set and the antagonist data set separately on each structure. We compared the enrichments associated with the best agonist-bound structure and the best antagonist-bound structure for these 10 NRs (see Figures 3 and 4). For all of the NRs, the structure that gave the best enrichment was an agonist-bound structure when the agonist data set was used. In regard to the antagonist data sets, the structure that gave the best performance in

enrichment was an antagonist-bound structure for six out of the 10 NRs tested. As shown in Figure 5, there is a clear separation of the score distribution profiles between the agonist-bound structures and the antagonist-bound ones using the agonist data sets (for seven NRs out of 10) and using the antagonist data sets (for seven NRs out of 10).

Impact of Using Experimentally Confirmed Decoy Ligands. The decoys provided in the NRLiSt BDB were generated using the DUD-E automated tool. As it is often the case in benchmarking databases, these decoys are assumed to be inactive with no experimental confirmation. Therefore, we decided to use the antagonist data sets as decoy ligand data sets for the agonist data sets and, reciprocally, to use the agonist data sets as decoy ligand data sets for the antagonist data sets if possible (i.e., when a sufficient number of ligands and the appropriate bound structure were available). We thus evaluated the enrichment obtained with the agonist data sets of AR, GR, MR, and PR and with the antagonist data sets of ER_alpha, ER_beta, PPAR_alpha, PPAR_gamma, RAR_alpha, and RXR_alpha using the corresponding antagonist or agonist data sets as their decoys (i.e., using experimentally confirmed decoy ligands). The results are shown in Table 2.

For the agonist data sets (Figure 6A), overall good performances were obtained, except on PR, for which the ROC curve obtained was close to the random one. Similarly, on the antagonist data sets (Figure 6B), overall good performances were associated with ER_alpha, ER_beta, and RXR_alpha, while poor performances were obtained on PPAR_alpha and PPAR_gamma and, to a lesser extent, on RAR_alpha. Similar trends were found when focusing on early enrichments (Figure 7). These results could be correlated with the ratio of the number of active compounds to the number of decoy ligands in the data set. Indeed, when the number of decoy ligands was too large (PPAR_alpha and PPAR_gamma) relative to the number of active compounds, the corresponding performances in enrichment were impacted. On the targets for which the number of

Table 1. Areas under the ROC Curve (AUC) and Enrichment Factors (EF) at 1% and 10% Obtained by Docking with Surflex-Dock with the 339 NR Structures Contained in the NRLiSt BDB and Used in This Study

structure	agonist data set			antagonist data set			total data set		
	AUC	EF _{1%}	EF _{10%}	AUC	EF _{1%}	EF _{10%}	AUC	EF _{1%}	EF _{10%}
AR									
1E3G	0.742	9.76	3.78	0.488	2.20	1.98	0.594	6.07	3.10
1TSZ	0.727	3.80	2.76	0.494	2.65	1.41	0.592	2.67	2.23
1T63	0.777	8.14	3.89	0.495	2.21	1.67	0.612	5.34	2.74
1T65	0.746	7.05	3.30	0.483	1.32	1.54	0.593	4.37	2.50
1XJ7	0.744	8.68	2.70	0.485	0.88	0.75	0.592	3.64	1.87
1XOW	0.769	11.93	4.16	0.491	1.76	1.89	0.605	8.01	2.89
1XQ3	0.770	11.93	4.11	0.494	1.32	1.76	0.609	7.53	2.99
2AM9	0.782	8.68	4.49	0.497	2.65	1.81	0.615	5.83	3.01
2AMA	0.752	8.68	3.46	0.485	1.76	1.67	0.597	4.86	2.60
2AMB	0.763	9.76	3.95	0.510	1.76	1.63	0.612	6.07	2.91
2AO6	0.748	7.59	3.51	0.501	1.76	1.67	0.601	4.37	2.70
2AX9	0.698	5.97	3.73	0.500	4.41	2.07	0.585	6.31	2.94
2AXA	0.567	4.34	2.27	0.453	2.20	1.06	0.504	3.16	1.60
2HVC	0.696	25.49	3.95	0.470	4.85	1.59	0.568	14.32	2.74
2PIO	0.770	7.59	3.95	0.492	0.88	1.59	0.609	3.16	2.74
2PIP	0.778	11.93	3.73	0.507	2.65	1.54	0.617	6.31	2.50
2PIQ	0.759	5.97	3.57	0.495	0.88	1.41	0.606	2.67	2.40
2PIR	0.775	12.48	3.89	0.478	1.76	1.37	0.600	6.56	2.57
2PIT	0.766	13.02	3.68	0.477	0.88	1.15	0.596	5.10	2.45
2PIU	0.824	18.98	5.51	0.509	3.53	2.11	0.638	11.17	3.79
2PIV	0.759	5.42	3.57	0.484	1.76	1.37	0.599	3.40	2.55
2PIW	0.845	21.15	5.68	0.519	3.97	1.89	0.650	12.14	3.79
2PIX	0.781	4.34	3.46	0.504	0.88	1.76	0.614	2.67	2.69
2PKL	0.782	9.22	4.27	0.488	2.21	1.94	0.611	7.28	3.16
2PNU	0.739	4.34	2.32	0.521	2.21	1.01	0.609	2.19	1.38
2Q7I	0.765	10.85	4.11	0.490	2.20	1.50	0.603	6.55	2.84
2Q7J	0.760	7.59	3.68	0.501	3.97	1.76	0.608	6.07	2.79
2Z4J	0.804	16.27	4.54	0.502	1.76	1.63	0.627	8.98	3.11
3B5R	0.560	4.34	2.32	0.458	0.88	0.75	0.503	2.43	1.46
3B65	0.579	3.25	2.27	0.463	0.88	1.15	0.513	1.70	1.72
3B66	0.500	2.71	0.76	0.441	2.21	0.75	0.469	2.43	0.78
3B67	0.453	0.00	0.59	0.483	3.97	1.10	0.475	2.67	0.87
3B68	0.566	2.71	2.27	0.489	3.97	1.50	0.523	3.64	1.80
3L3X	0.782	9.22	4.43	0.502	3.53	2.38	0.619	5.83	3.37
3L3Z	0.785	10.85	4.16	0.493	3.53	2.25	0.615	7.28	3.23
3RLJ	0.483	2.17	1.35	0.439	0.88	0.88	0.461	0.97	1.07
CAR									
1XV9	0.630	0.00	1.52	0.545	0.00	0.00	0.622	3.00	1.44
1XVP	0.544	3.09	1.52	0.613	0.00	0.00	0.548	3.00	1.44
ER_alpha									
1A52	0.634	5.08	2.41	0.755	17.02	5.04	0.657	9.40	2.93
1ERE	0.656	5.77	2.97	0.438	14.06	5.04	0.592	5.41	2.59
1ERR	0.586	3.23	2.41	0.822	2.98	1.10	0.637	7.14	3.13
1G50	0.596	3.46	2.12	0.410	0.74	0.66	0.543	3.13	1.86
1GWQ	0.647	7.84	3.56	0.423	0.74	1.10	0.577	7.31	3.06
1GWR	0.690	8.07	3.56	0.445	1.48	1.10	0.618	6.79	3.25
1L2I	0.700	5.77	3.61	0.457	5.92	1.24	0.630	5.75	3.18
1QKU	0.594	4.38	2.00	0.448	0.74	1.02	0.551	3.66	1.68
1R5K	0.668	5.08	2.71	0.812	25.90	5.62	0.693	10.45	3.11
1SJO	0.604	3.92	2.32	0.823	28.12	6.35	0.650	13.06	3.06
1X7E	0.701	10.29	4.44	0.559	5.18	1.39	0.627	10.10	3.84
1X7R	0.639	6.46	3.17	0.401	0.74	0.66	0.568	5.57	2.73
1XP1	0.615	4.15	2.00	0.799	25.90	5.48	0.648	11.49	2.71
1XP6	0.596	2.77	2.14	0.837	31.07	6.43	0.650	11.67	3.00
1XP9	0.594	3.39	2.00	0.806	31.08	5.55	0.638	12.89	2.83
1XPC	0.557	3.69	1.98	0.792	22.94	5.26	0.608	10.62	2.74
1YIM	0.555	3.23	1.68	0.811	26.64	5.91	0.611	13.06	2.59

Table 1. continued

structure	agonist data set			antagonist data set			total data set		
	AUC	EF _{1%}	EF _{10%}	AUC	EF _{1%}	EF _{10%}	AUC	EF _{1%}	EF _{10%}
	ER_alpha								
1YIN	0.580	2.77	1.84	0.813	17.76	5.99	0.623	10.62	2.80
2FAI	0.731	14.86	4.97	0.556	3.73	1.25	0.647	13.61	4.23
2IOG	0.674	3.69	2.83	0.771	12.58	3.87	0.691	6.44	2.81
2IOK	0.572	1.15	1.43	0.704	4.44	3.14	0.597	3.66	1.93
2JF9	0.611	3.92	2.32	0.836	27.38	6.28	0.659	11.15	3.09
2JFA	0.615	5.77	2.71	0.810	22.20	5.77	0.653	10.80	3.11
2OUZ	0.615	4.15	2.16	0.796	20.72	5.26	0.648	10.27	2.87
2QXM	0.738	24.23	5.44	0.500	6.05	1.42	0.661	21.35	4.54
2QZO	0.655	7.84	3.20	0.466	2.22	1.24	0.594	6.79	2.69
3DT3	0.699	9.46	4.02	0.792	25.16	5.62	0.704	12.02	3.89
3ERD	0.627	4.15	3.12	0.415	1.48	1.31	0.568	3.66	2.78
3ERT	0.630	4.84	2.69	0.810	14.80	5.91	0.660	10.27	3.06
3OS8	0.661	7.15	2.85	0.528	3.70	1.97	0.617	5.40	2.57
3OS9	0.688	3.23	2.71	0.717	4.44	3.58	0.676	4.18	2.40
3OSA	0.680	4.84	3.01	0.774	13.32	5.04	0.696	8.53	3.32
	ER_beta								
1L2J	0.710	2.81	3.08	0.773	5.74	4.72	0.717	3.67	3.11
1NDE	0.656	3.32	2.32	0.752	11.48	3.44	0.667	4.75	2.44
1QKM	0.711	9.18	4.15	0.443	0.00	0.29	0.658	9.51	3.78
1U3Q	0.723	8.16	4.15	0.454	4.31	0.57	0.670	7.81	3.80
1U3R	0.672	3.06	2.55	0.472	5.74	1.43	0.644	3.25	2.51
1U3S	0.634	2.30	2.32	0.500	2.87	1.72	0.610	2.16	2.25
1U9E	0.728	7.65	4.30	0.417	1.44	0.57	0.669	6.92	3.74
1X76	0.748	10.46	4.56	0.504	4.31	1.15	0.701	9.72	4.13
1X78	0.685	5.10	3.05	0.486	0.00	0.72	0.645	4.54	2.64
1X7B	0.661	4.34	3.11	0.491	2.91	0.87	0.627	3.90	2.84
1X7J	0.698	6.38	3.36	0.512	2.87	0.86	0.662	5.62	3.13
1YY4	0.715	10.71	4.53	0.469	0.00	1.43	0.669	9.29	4.02
1YYE	0.707	10.46	4.07	0.520	0.00	1.29	0.670	10.15	3.65
1ZAF	0.745	12.24	4.56	0.507	1.44	1.29	0.698	11.45	4.02
2GIU	0.640	2.55	1.99	0.561	10.05	2.00	0.751	4.43	3.10
2IOG	0.672	4.34	2.85	0.529	1.44	1.15	0.647	3.89	2.64
2JJ3	0.688	2.30	2.62	0.523	1.44	0.86	0.655	2.16	2.40
2NV7	0.688	8.67	3.77	0.452	0.00	0.86	0.642	7.78	3.39
2QTU	0.653	3.06	2.52	0.454	0.00	1.00	0.617	2.59	2.29
2Z4B	0.657	4.85	2.77	0.468	0.00	0.86	0.623	4.32	2.51
3OLL	0.745	8.16	3.77	0.568	5.74	2.29	0.712	7.78	3.50
3OLS	0.745	9.95	3.67	0.524	4.31	1.57	0.703	9.07	3.39
3OMO	0.664	6.38	3.54	0.419	1.44	0.72	0.621	6.48	3.30
3OMP	0.680	4.85	3.21	0.408	0.00	0.44	0.631	4.33	3.01
3OMQ	0.739	8.93	3.97	0.489	2.87	1.43	0.692	7.99	3.69
	ERR_alpha								
2PJL	0.297	0.00	0.00	0.298	0.00	0.00	0.339	0.00	3.42
3K6P	0.342	0.00	0.00	0.091	0.00	0.00	0.337	0.00	2.37
	FXR_alpha								
1OSH	0.524	1.26	1.38	0.455	0.00	1.12	0.518	1.15	1.36
3BEJ	0.598	6.28	3.06	0.593	16.13	3.36	0.596	6.92	3.14
3DCT	0.688	15.07	4.50	0.525	0.00	2.61	0.675	14.42	4.35
3DCU	0.679	4.52	3.13	0.615	4.03	3.73	0.667	4.90	3.78
3FLI	0.460	0.63	0.94	0.471	4.04	1.49	0.461	0.58	1.01
3HCS	0.735	30.13	5.06	0.585	16.15	3.36	0.722	29.98	4.93
3L1B	0.526	2.20	1.09	0.434	4.04	1.87	0.519	2.31	1.15
	GR								
1M2Z	0.636	7.84	2.98	0.376	2.18	1.41	0.494	5.89	2.05
1NHZ	0.579	1.71	1.73	0.717	15.32	3.50	0.655	10.30	2.72
1P93	0.596	6.82	2.95	0.387	3.54	1.17	0.483	5.28	2.02
3BQD	0.661	11.53	3.15	0.407	8.16	1.77	0.520	9.21	2.40
3CLD	0.619	4.41	2.88	0.419	1.91	1.68	0.510	3.32	2.20

Table 1. continued

structure	agonist data set			antagonist data set			total data set		
	AUC	EF _{1%}	EF _{10%}	AUC	EF _{1%}	EF _{10%}	AUC	EF _{1%}	EF _{10%}
GR									
3E7C	0.578	7.16	2.10	0.369	4.35	1.17	0.464	5.13	1.61
3K22	0.464	3.39	1.66	0.357	4.08	1.25	0.410	3.62	1.45
LXR_alpha									
1UHL	0.612	0.77	1.66	0.611	2.03	2.20	0.608	0.65	1.68
3IPO	0.643	13.52	3.13	0.591	4.06	2.40	0.631	11.99	2.85
3IPS	0.651	5.43	2.28	0.608	2.03	3.01	0.640	4.86	2.20
3IPU	0.574	3.09	1.62	0.604	4.06	3.01	0.584	3.24	1.65
LXR_beta									
1P8D	0.693	2.41	2.27	0.717	2.64	2.64	0.691	2.44	2.28
1PQ6	0.736	8.25	4.16	0.478	0.00	1.05	0.702	7.32	3.80
1PQC	0.637	8.59	2.27	0.535	0.00	0.79	0.624	7.93	2.13
1PQ9	0.564	1.03	1.31	0.591	2.64	2.90	0.566	0.91	1.43
1UPV	0.648	4.47	2.20	0.470	0.00	0.53	0.623	3.97	1.98
1UPW	0.604	5.84	2.27	0.508	0.00	1.05	0.592	5.18	2.01
3KFC	0.708	18.57	4.40	0.580	5.27	2.90	0.694	15.58	4.14
MR									
2A3I	0.887	56.00	7.84	0.380	2.74	0.75	0.401	6.47	1.23
2AA2	0.821	44.80	7.84	0.395	1.39	0.89	0.413	4.53	1.29
2AA5	0.834	44.80	7.84	0.401	4.11	1.03	0.420	8.40	1.36
2AA7	0.750	44.80	6.72	0.361	0.69	0.82	0.378	4.53	1.16
2ABI	0.822	44.80	6.72	0.345	1.37	0.69	0.367	4.53	1.03
3VHU	0.829	56.00	6.72	0.541	6.94	2.06	0.552	10.34	2.26
3VHV	0.600	11.20	2.24	0.534	8.92	2.47	0.537	9.05	2.45
PPAR_alpha									
1I7G	0.797	10.44	4.81	0.602	26.75	2.55	0.795	10.53	4.80
1K7L	0.807	9.37	4.70	0.595	13.38	2.55	0.805	9.39	4.69
1KKQ	0.748	4.94	3.56	0.447	0.00	2.55	0.746	5.06	3.55
2NPA	0.833	13.74	5.58	0.556	26.75	2.55	0.831	13.81	5.56
2P54	0.858	14.59	6.23	0.565	13.38	1.27	0.855	14.51	6.20
2REW	0.863	12.54	6.10	0.624	13.38	2.55	0.847	15.52	5.97
2ZNN	0.786	6.80	4.39	0.602	0.00	2.55	0.784	6.90	4.38
3ET1	0.777	7.23	4.04	0.556	13.38	2.55	0.749	2.58	2.44
3FEI	0.746	5.94	3.83	0.599	13.38	1.27	0.745	5.98	3.82
3G8I	0.821	15.81	5.40	0.517	0.00	1.27	0.819	15.80	5.37
3KDT	0.839	13.81	5.75	0.507	0.00	2.55	0.838	13.52	5.76
3KDU	0.776	5.72	4.28	0.653	0.00	2.55	0.742	1.54	1.63
PPAR_beta									
1GWX	0.857	16.85	6.06	0.516	0.00	0.00	0.852	16.99	6.02
1Y0S	0.844	12.89	5.68	0.667	0.00	1.84	0.842	12.85	5.65
2J14	0.736	4.29	3.31	0.645	0.00	1.84	0.735	4.35	3.32
2Q5G	0.777	10.68	4.29	0.457	0.00	0.92	0.774	40.67	4.24
2XYJ	0.798	11.45	4.57	0.446	0.00	0.92	0.794	11.54	4.53
2XYW	0.779	7.82	4.29	0.614	0.00	1.84	0.776	7.73	4.28
2XYX	0.771	8.70	4.17	0.662	0.00	0.00	0.768	8.93	3.98
2ZNP	0.861	21.04	6.39	0.476	0.00	0.92	0.856	20.91	6.35
2ZNQ	0.842	21.50	6.27	0.416	0.00	0.00	0.837	21.37	6.19
3DY6	0.736	3.63	2.97	0.428	0.00	0.00	0.733	3.70	2.96
3ET2	0.788	8.59	4.45	0.470	0.00	0.92	0.785	8.49	4.41
3GWX	0.823	11.79	5.09	0.570	0.00	2.76	0.820	11.76	5.09
3OZ0	0.832	17.09	5.56	0.651	0.00	1.84	0.829	16.89	5.54
PPAR_gamma									
1FM6	0.795	12.10	4.79	0.474	0.00	2.24	0.793	11.93	4.78
1FM9	0.848	18.64	5.87	0.529	0.00	2.24	0.846	18.60	5.87
1I7I	0.807	16.17	5.38	0.460	11.18	2.24	0.807	15.87	5.38
1K74	0.851	17.49	6.12	0.443	0.00	2.24	0.850	17.73	6.08
1KNU	0.836	12.43	5.70	0.555	0.00	2.23	0.835	12.42	5.68
1NYX	0.808	7.64	4.44	0.450	0.00	1.12	0.807	7.82	4.47
1RDT	0.853	18.42	5.94	0.520	0.00	1.12	0.850	18.77	5.93

Table 1. continued

structure	agonist data set			antagonist data set			total data set		
	AUC	EF _{1%}	EF _{10%}	AUC	EF _{1%}	EF _{10%}	AUC	EF _{1%}	EF _{10%}
	PPAR _γ								
1WM0	0.764	8.47	3.81	0.538	0.00	1.12	0.760	7.88	3.78
1ZEO	0.841	13.97	5.77	0.454	0.00	2.24	0.839	13.95	5.76
1ZGY	0.807	9.68	4.96	0.413	11.18	2.24	0.807	9.68	4.96
2ATH	0.802	7.15	4.66	0.441	0.00	1.12	0.801	7.22	4.68
2F4B	0.785	5.33	4.10	0.529	0.00	2.24	0.784	5.20	4.09
2FVJ	0.776	5.50	3.91	0.473	0.00	2.24	0.798	4.38	3.60
2G0G	0.790	9.13	4.32	0.453	0.00	1.12	0.788	9.14	4.30
2G0H	0.812	9.02	4.67	0.468	11.18	1.12	0.811	9.90	4.79
2GTK	0.833	14.85	5.52	0.449	0.00	2.24	0.831	14.83	5.51
2HWQ	0.765	5.00	3.69	0.521	0.00	2.24	0.764	4.98	3.68
2HWR	0.836	9.13	5.33	0.462	0.00	1.12	0.846	5.56	4.35
2I4J	0.820	9.62	5.28	0.517	11.18	2.24	0.819	9.63	5.27
2I4P	0.802	8.52	4.68	0.480	0.00	1.12	0.800	8.54	4.67
2I4Z	0.835	10.34	5.44	0.506	0.00	2.24	0.834	10.29	5.45
2OM9	0.820	9.40	5.15	0.599	0.00	3.35	0.819	9.30	5.13
2P4Y	0.763	3.90	3.33	0.460	0.00	1.12	0.761	3.89	3.32
2POB	0.838	9.67	5.48	0.513	0.00	1.12	0.836	9.57	5.47
2PRG	0.765	5.83	3.89	0.464	0.00	2.24	0.764	5.80	3.89
2Q59	0.801	10.17	4.85	0.491	0.00	1.12	0.801	10.18	4.86
2Q5P	0.787	6.60	4.10	0.420	11.18	1.12	0.785	6.62	4.12
2Q5S	0.774	6.65	3.84	0.465	0.00	1.12	0.773	6.73	3.85
2Q61	0.771	5.39	3.65	0.533	0.00	1.12	0.770	5.09	3.60
2Q6R	0.796	6.87	4.43	0.493	0.00	1.12	0.794	6.84	4.42
2Q6S	0.794	7.26	4.34	0.458	11.18	2.24	0.793	7.28	4.35
2Q8S	0.798	8.63	4.74	0.448	0.00	2.24	0.797	8.59	4.73
2VSR	0.717	6.93	3.33	0.564	11.18	2.24	0.715	6.95	3.31
2VST	0.771	5.66	3.83	0.356	0.00	1.12	0.769	5.64	3.82
2VV0	0.836	12.98	5.47	0.485	0.00	2.24	0.834	12.91	5.45
2VV1	0.788	6.66	4.05	0.448	0.00	1.12	0.792	7.11	4.25
2VV2	0.801	6.65	4.40	0.437	0.00	2.24	0.799	6.62	4.39
2VV3	0.795	6.27	4.17	0.542	11.18	2.24	0.798	6.67	4.32
2VV4	0.794	6.43	4.20	0.570	0.00	1.12	0.796	6.84	4.44
2ZK1	0.774	5.73	3.79	0.509	0.00	1.12	0.777	6.19	3.93
2ZK2	0.760	6.01	3.84	0.507	0.00	2.24	0.776	10.01	4.40
2ZK3	0.751	8.52	3.90	0.454	0.00	1.12	0.749	8.48	3.89
2ZK4	0.803	7.26	4.41	0.444	11.18	1.12	0.801	7.11	4.39
2ZK5	0.764	6.70	3.90	0.545	0.00	1.12	0.768	6.95	4.01
2ZK6	0.772	6.10	3.80	0.507	0.00	1.12	0.770	6.07	3.78
2ZNO	0.797	4.57	3.69	0.498	0.00	2.24	0.776	4.92	3.85
3ADS	0.815	9.67	4.68	0.469	0.00	2.24	0.815	10.78	4.83
3ADT	0.779	5.39	4.05	0.408	0.00	0.00	0.777	5.36	4.03
3ADU	0.777	6.27	3.88	0.530	0.00	2.24	0.779	6.29	3.92
3ADV	0.797	4.90	3.56	0.464	0.00	1.12	0.773	6.29	3.86
3ADW	0.761	4.56	3.53	0.410	0.00	1.12	0.767	5.47	3.91
3ADX	0.767	5.28	3.70	0.470	0.00	1.12	0.765	5.25	3.71
3B3K	0.841	7.75	4.89	0.560	11.18	2.24	0.837	13.51	5.52
3BC5	0.810	10.73	4.95	0.560	11.18	2.24	0.821	13.13	5.27
3CDP	0.831	2.57	2.60	0.451	11.18	2.24	0.820	8.65	5.03
3CDS	0.810	7.37	4.76	0.475	11.18	1.12	0.814	8.32	4.94
3CS8	0.797	6.05	3.88	0.447	11.18	1.12	0.790	9.08	4.25
3CWD	0.836	10.12	5.35	0.461	0.00	1.12	0.834	10.18	5.34
3D6D	0.807	8.74	4.64	0.507	0.00	2.24	0.817	10.72	4.91
3DZU	0.797	8.14	4.34	0.460	0.00	1.12	0.795	7.99	4.32
3DZY	0.786	3.50	3.03	0.538	0.00	2.24	0.814	12.25	4.80
3E00	0.794	6.21	4.23	0.468	0.00	1.12	0.793	6.18	4.22
3ET0	0.689	8.30	3.12	0.607	11.18	1.12	0.688	8.21	3.11
3ET3	0.827	12.15	5.45	0.460	0.00	2.24	0.826	12.09	5.43
3FEJ	0.823	17.92	5.85	0.431	0.00	2.24	0.822	17.45	5.91

Table 1. continued

structure	agonist data set			antagonist data set			total data set		
	AUC	EF _{1%}	EF _{10%}	AUC	EF _{1%}	EF _{10%}	AUC	EF _{1%}	EF _{10%}
PPAR _{gamma}									
3FUR	0.764	4.40	3.58	0.454	0.00	1.12	0.763	4.38	3.57
3G9E	0.845	13.97	6.00	0.504	0.00	2.24	0.843	13.84	5.98
3GBK	0.819	7.20	4.87	0.433	11.18	1.12	0.817	7.11	4.85
3H0A	0.829	8.86	5.42	0.546	0.00	1.12	0.827	8.81	5.41
3H0O	0.810	7.92	4.64	0.501	0.00	2.24	0.809	7.88	4.62
3HOD	0.808	8.69	4.93	0.543	0.00	3.35	0.807	8.65	4.92
3IA6	0.824	15.78	5.89	0.402	11.18	2.24	0.820	15.60	5.83
3K8S	0.777	7.04	4.01	0.468	0.00	2.24	0.778	7.22	4.01
3KMG	0.787	7.48	4.23	0.392	0.00	1.12	0.785	7.44	4.21
3LMP	0.764	7.09	3.80	0.473	0.00	0.00	0.763	7.00	3.81
3NOA	0.807	8.91	4.73	0.434	0.00	1.12	0.806	8.87	4.72
3OSI	0.788	5.61	3.81	0.530	0.00	0.00	0.786	5.58	3.80
3OSW	0.780	5.78	3.91	0.484	0.00	2.24	0.779	5.69	3.93
3PBA	0.782	4.40	3.78	0.466	0.00	2.24	0.781	4.38	3.77
4PRG	0.788	7.53	4.31	0.538	0.00	2.24	0.786	7.33	4.32
PR									
1A28	0.494	2.61	1.41	0.511	0.76	0.91	0.509	1.38	1.15
1E3K	0.524	7.07	1.85	0.566	4.74	2.17	0.555	5.52	2.09
1SQN	0.488	0.37	1.44	0.550	2.46	1.66	0.533	1.88	1.64
1SR7	0.547	0.74	0.96	0.498	0.00	0.70	0.513	0.25	0.81
1ZUC	0.463	5.58	1.48	0.574	1.14	2.38	0.541	2.13	2.24
2OVH	0.509	0.37	0.85	0.486	8.33	1.27	0.492	5.77	1.11
2OVM	0.579	1.12	1.70	0.526	7.57	1.51	0.542	5.52	1.45
2W8Y	0.545	2.98	1.85	0.522	0.57	1.12	0.532	1.13	1.34
3D9O	0.486	0.37	0.89	0.500	1.33	0.93	0.498	1.00	0.90
3G8N	0.526	5.58	1.82	0.554	2.27	1.64	0.547	3.26	1.69
3G8O	0.549	1.86	1.96	0.570	1.14	1.66	0.566	1.25	1.83
3HQ5	0.528	5.96	1.82	0.557	2.46	1.66	0.550	3.64	1.71
3KBA	0.644	5.21	2.78	0.565	2.27	1.64	0.592	3.26	1.98
PXR									
1ILH	0.677	5.06	2.80	0.568	0.00	1.45	0.671	4.67	2.52
1M13	0.594	3.03	2.30	0.412	0.00	2.89	0.581	4.67	2.24
1NRL	0.606	3.03	2.40	0.529	15.90	1.45	0.592	3.74	2.15
1SKX	0.613	2.02	2.10	0.577	0.00	1.45	0.608	1.87	2.15
2O9I	0.543	4.05	1.60	0.516	0.00	2.89	0.540	3.74	1.50
2QNV	0.535	2.02	1.60	0.573	0.00	2.89	0.540	2.80	1.96
3HVL	0.599	5.06	1.90	0.518	0.00	2.89	0.591	5.61	2.06
RAR _{alpha}									
1DKF	0.738	25.12	5.42	0.686	32.38	5.16	0.717	25.14	5.23
3A9E	0.792	15.23	5.12	0.439	4.63	1.67	0.663	11.06	3.87
3KMR	0.927	43.40	8.42	0.317	4.63	0.76	0.705	34.19	6.03
3KMZ	0.639	12.94	3.39	0.678	32.38	5.62	0.660	19.61	4.12
RAR _{beta}									
1XAP	0.892	42.02	7.85	0.687	22.08	4.70	0.846	38.30	7.29
RAR _{gamma}									
1EXA	0.836	34.17	6.22	0.274	0.00	0.53	0.650	24.40	4.55
1FCX	0.875	29.61	6.82	0.270	0.00	0.53	0.674	21.21	5.19
1FCY	0.862	30.37	6.59	0.233	0.00	0.70	0.659	21.74	4.82
1FCZ	0.873	33.41	7.20	0.283	3.52	0.70	0.680	27.05	5.24
1FD0	0.878	29.61	6.97	0.217	0.00	0.70	0.667	21.21	5.08
2LBD	0.902	45.55	7.73	0.329	3.52	0.88	0.711	34.47	5.82
3LBD	0.859	34.93	6.82	0.241	5.28	0.70	0.660	25.99	5.03
4LBD	0.847	29.61	6.37	0.261	0.00	0.53	0.655	22.27	4.60
ROR _{alpha}									
1N83	0.805	50.00	6.67	0.921	33.95	6.17	0.901	31.72	6.27
1SOX	0.705	50.00	6.67	0.852	16.97	5.40	0.836	19.03	5.64
ROR _{gamma}									
3B0W	0.864	50.71	8.69	1.000	51.00	10.20	0.915	50.64	9.21
3KYT	0.883	33.81	8.69	0.999	51.00	10.20	0.928	40.51	9.21

Table 1. continued

structure	agonist data set			antagonist data set			total data set		
	AUC	EF _{1%}	EF _{10%}	AUC	EF _{1%}	EF _{10%}	AUC	EF _{1%}	EF _{10%}
ROR_gamma									
3LOJ	0.875	50.71	8.69	1.000	51.00	10.20	0.923	40.51	9.21
3LOL	0.892	50.71	8.69	1.000	51.00	10.20	0.936	50.64	9.21
RXR_alpha									
1FBY	0.902	35.71	8.05	0.426	9.96	2.58	0.696	25.79	6.00
1FM6	0.919	36.67	8.29	0.433	16.09	3.33	0.717	27.26	6.23
1FM9	0.865	35.24	7.24	0.480	25.28	3.49	0.701	29.60	5.85
1K74	0.874	27.37	7.34	0.451	22.22	3.56	0.697	24.03	5.88
1MV9	0.933	38.10	8.29	0.476	22.99	3.56	0.739	30.48	6.38
1MVC	0.878	40.81	7.67	0.390	9.96	2.50	0.669	29.31	5.47
1MZN	0.894	42.73	7.96	0.398	7.66	2.27	0.677	30.48	5.67
1RDT	0.912	35.24	8.10	0.436	21.45	3.56	0.713	28.43	6.35
1XV9	0.793	28.57	5.76	0.416	7.66	2.50	0.632	19.64	4.42
1XVP	0.727	24.49	4.77	0.376	7.66	1.67	0.574	17.00	3.51
2ACL	0.922	38.41	8.48	0.579	25.28	4.17	0.796	30.48	6.79
2P1T	0.928	40.48	8.43	0.457	19.16	3.18	0.721	31.07	6.44
2P1U	0.918	33.81	8.14	0.462	21.45	2.96	0.712	28.72	6.14
2P1V	0.951	44.76	8.71	0.494	23.76	3.64	0.749	34.00	6.73
2ZXZ	0.906	46.57	8.20	0.415	16.86	2.80	0.695	34.29	5.97
2ZYO	0.904	45.13	8.15	0.428	13.79	3.03	0.698	33.41	5.97
3DZU	0.749	12.38	4.62	0.443	3.83	1.44	0.601	7.91	3.42
3DZY	0.870	27.14	6.71	0.496	15.32	3.56	0.712	21.69	5.53
3E00	0.824	22.09	6.29	0.402	8.43	2.58	0.648	17.58	4.83
3E94	0.277	0.95	0.33	0.287	0.00	0.30	0.285	0.59	0.20
3FAL	0.888	38.10	7.33	0.614	27.58	4.39	0.785	30.78	6.17
3FC6	0.873	39.85	7.10	0.647	22.21	4.32	0.791	29.30	5.94
3FUG	0.880	34.09	7.58	0.404	11.49	2.73	0.674	23.45	5.50
3H0A	0.764	23.33	5.86	0.361	3.06	1.44	0.586	14.65	4.01
3KWY	0.177	0.48	0.33	0.207	0.00	0.23	0.183	0.00	0.29
3NSQ	0.634	0.48	1.67	0.631	3.83	1.97	0.647	2.34	1.76
3OAP	0.896	40.81	8.15	0.424	13.79	2.65	0.692	29.59	5.94
3OZJ	0.923	40.95	8.52	0.445	25.28	3.64	0.725	31.94	6.55
3R2A	0.569	5.76	2.10	0.448	3.83	1.59	0.529	5.28	1.87
RXR_beta									
1H9U	0.836	48.17	7.09	0.725	16.90	7.24	0.833	43.74	7.10
1UHL	0.656	4.66	2.78	0.326	0.00	1.45	0.622	4.23	2.64
RXR_gamma									
2GL8	0.646	10.09	3.95	0.027	0.00	0.00	0.596	9.27	3.51
SF1									
1YOW	0.807	0.00	3.71	0.585	0.00	0.50	0.701	0.00	2.06
1ZDT	0.844	5.63	5.83	0.707	0.00	1.50	0.776	2.66	3.35
TR_alpha									
1NAV	0.781	7.32	5.08	0.541	0.00	1.19	0.739	5.99	4.48
2H77	0.872	23.44	6.24	0.618	0.00	1.19	0.823	19.18	5.41
2H79	0.888	17.58	7.11	0.570	6.30	0.59	0.829	17.97	6.01
3HZF	0.926	45.40	7.69	0.635	6.31	1.19	0.874	40.74	6.72
3ILZ	0.915	38.09	7.69	0.651	0.00	2.38	0.868	32.36	6.71
3JZB	0.877	26.36	6.53	0.695	0.00	1.78	0.836	28.76	5.66
TR_beta									
1BSX	0.865	19.37	6.16	0.585	6.75	0.67	0.821	19.48	5.38
1N46	0.910	37.73	7.58	0.527	0.00	1.33	0.848	33.56	6.45
1NAX	0.883	42.61	7.70	0.575	0.00	2.67	0.832	36.80	6.88
1Q4X	0.906	34.85	7.19	0.583	6.75	2.00	0.854	31.38	6.68
1R6G	0.867	25.82	6.04	0.471	0.00	0.67	0.803	22.73	5.06
1XZX	0.885	24.52	7.19	0.532	0.00	2.00	0.830	24.89	6.36
1Y0X	0.897	32.26	8.09	0.509	13.50	2.00	0.835	32.46	7.11
3GWS	0.904	37.42	7.44	0.539	0.00	1.33	0.847	33.54	6.36
3IMY	0.905	34.86	8.35	0.491	6.75	1.33	0.837	32.47	7.20
3JZC	0.885	40.02	7.19	0.565	0.00	0.67	0.828	36.80	6.02

Table 1. continued

structure	agonist data set			antagonist data set			total data set		
	AUC	EF _{1%}	EF _{10%}	AUC	EF _{1%}	EF _{10%}	AUC	EF _{1%}	EF _{10%}
	VDR								
1DB1	0.945	34.13	8.36	0.810	26.44	6.39	0.902	30.84	7.84
1IE8	0.942	39.44	8.13	0.789	33.05	6.81	0.901	34.77	7.39
1IE9	0.944	40.20	7.91	0.831	33.05	6.81	0.910	35.89	7.50
1S0Z	0.952	37.17	8.51	0.753	24.24	4.90	0.891	30.84	7.73
1S19	0.951	35.65	8.36	0.849	35.24	7.02	0.919	32.52	8.23
1TXI	0.944	33.37	8.28	0.830	37.46	7.24	0.909	32.53	7.78
2HAM	0.976	43.99	9.41	0.850	33.05	6.81	0.936	40.94	8.56
2HAR	0.965	40.95	8.88	0.797	28.65	6.81	0.914	38.13	8.23
2HAS	0.962	40.19	8.73	0.874	37.46	8.09	0.935	37.57	8.45
2HB7	0.968	38.68	9.03	0.821	41.87	6.60	0.923	35.89	8.34
2HB8	0.951	39.44	8.81	0.818	28.65	6.81	0.910	36.45	8.12
3A3Z	0.965	39.44	9.03	0.834	28.65	7.03	0.923	36.45	8.17
3A40	0.965	43.99	8.73	0.853	30.84	7.45	0.931	39.25	8.28
3A78	0.930	36.40	8.21	0.800	22.03	6.39	0.889	31.40	7.72
3CS4	0.933	40.96	8.06	0.821	22.03	5.75	0.897	37.01	7.45
3CS6	0.955	40.20	8.06	0.818	28.65	6.60	0.914	37.57	7.73

decoy ligands outmatched the number of active compounds (MR_agonist, ER_alpha_antagonist, ER_beta_antagonist, and RXR_alpha_antagonist), the best performances in enrichment were observed in terms of both AUC and EF_{1%}.

DISCUSSION

In the present study, we used an exhaustive NR-focused benchmarking database, the NRLiSt BDB, to study on the one hand the importance of distinguishing agonist and antagonist ligand data sets for the quality of benchmarking databases and on the other hand whether the pharmacological profile of the cocrystallized ligand could guide the query structure choice for docking methods.

Impact of Using Separate Data Sets on the Overall Enrichment. Numerous benchmarking databases have been proposed over time,^{3–8,18} but the DUD⁷ and DUD-E⁸ databases offer the highest-quality features for evaluation methods to date and are thus considered as the gold standard. In light of a previous study,¹¹ we assumed that the quality of benchmarking data sets could be further improved, especially by providing separate data sets according to ligand pharmacological profiles, as already made in a GPCR ligands database.¹⁸

To test this hypothesis, for each of the 27 NRs of the NRLiSt BDB, we used three data sets, the agonist and antagonist data sets as provided in the NRLiSt BDB and a total data set constructed by combining the agonist and antagonist data sets. We wanted to study the impact of this choice on the performance in enrichment after a retrospective structure-based virtual screening using Surflex-Dock.

For each NR, the data set that provided the best enrichment was always the agonist data set or the antagonist data set, regardless of the structure used for screening. Moreover, for 97% and 89% of the structures used in this study, respectively, the best AUC and the best early enrichment were obtained using one of the two separate data sets. It thus appears that the choice to build separate data sets is relevant in view of the enhanced performances obtained. Another point is that the structure associated with the best AUC or best early enrichment for a given NR was in the majority of cases the same when one of the separate data sets or the total data set was used. This observation confirms that the choice of a more

appropriate query impacts the docking performance.^{11–14} However, for 26% of the targets, the AUC value associated with the agonist or antagonist data set for each structure was always superior to the AUC value obtained for the best-performing structure with the total data set. Furthermore, for 67% of the NRs, for at least 50% of their structures the AUC value obtained with one of the separate data sets was superior to the best AUC value obtained with the total data set for that target. It seems that when agonist ligands and antagonist ligands are pooled for a given NR (i.e., in the case of the total data set), there could be a bias with the evaluation of structure-based methods due to the pharmacological profiles of the ligands. Hence, separating the data sets depending on their pharmacological activity (agonist or antagonist) seems to be a better option to ensure a benchmarking database of the best quality.

Importance of the Bound Ligand in the Structure Used for Screening. It is currently accepted that the quality of the enrichment depends on the query conformation,^{11–14} and the results presented here again confirm this point. We studied the influence of the pharmacological profile of the ligand cocrystallized in the binding site on the performance in enrichment of Surflex-Dock, since conformational changes occur in protein binding sites upon ligand binding. For all of the NRs studied, the structure that led to the best enrichment was always agonist-bound when the agonist data set was used, and an antagonist-bound structure was most often better with the antagonist data set. This observation was confirmed with the score distributions of agonist ligands and antagonist ligands depending on the bound ligand in the original structure. The structural basis of agonist and antagonist action was previously reported,⁹ underlying the different conformational changes occurring upon agonist or antagonist binding. The NR apo structures present a fold constituted of the association of 12 α -helices and a short β -turn in an antiparallel “ α -helical sandwich”. Agonist binding induces the repositioning of helices H11, H12, and H3 of the receptor and generates a surface requisite for coactivator recruitment.⁹ Conformational changes induced by antagonist ligand binding affect the conformation of helix H12, which is no longer able to adapt its holo position, preventing the formation of the coactivator recruitment surface.

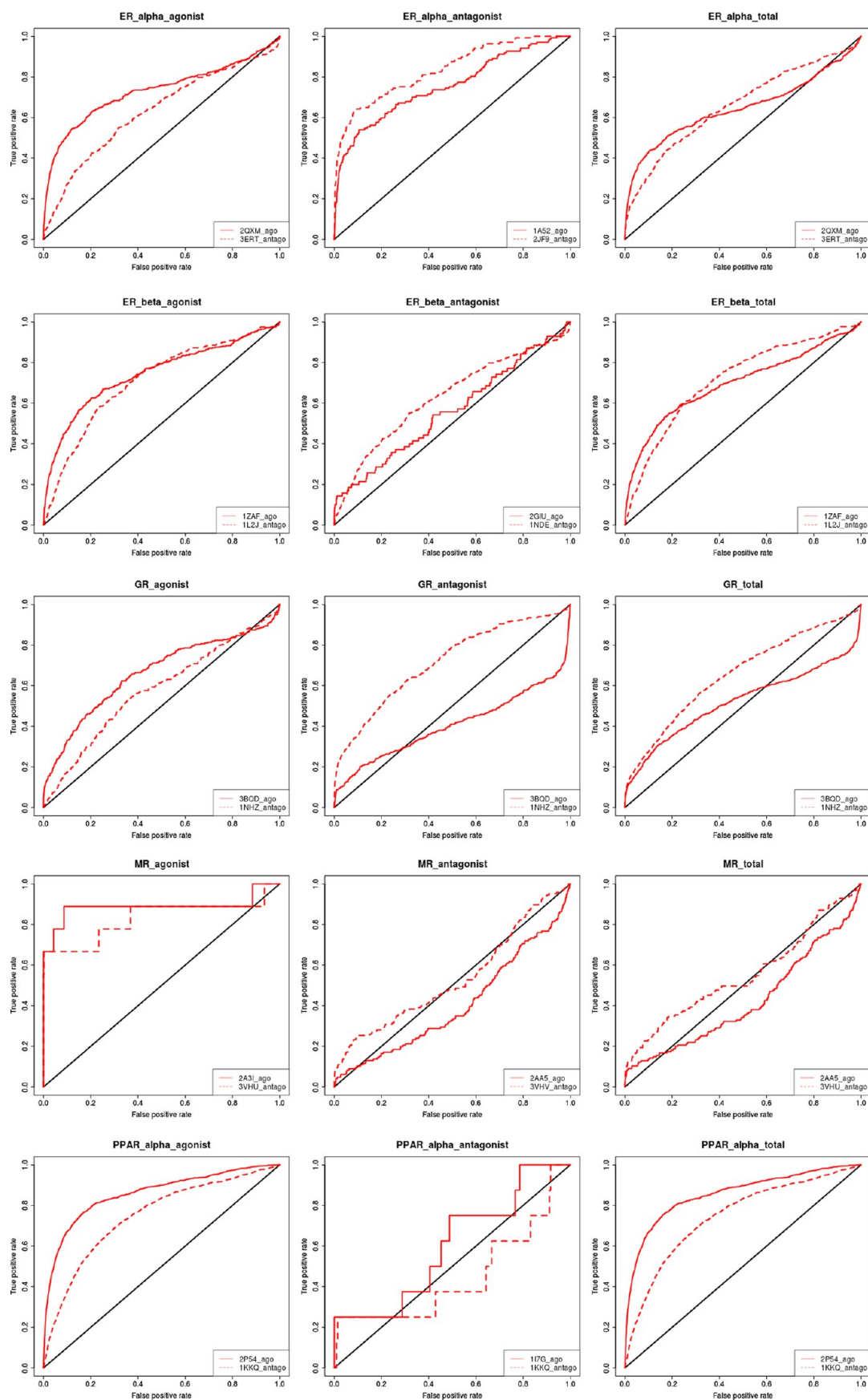


Figure 3. continued

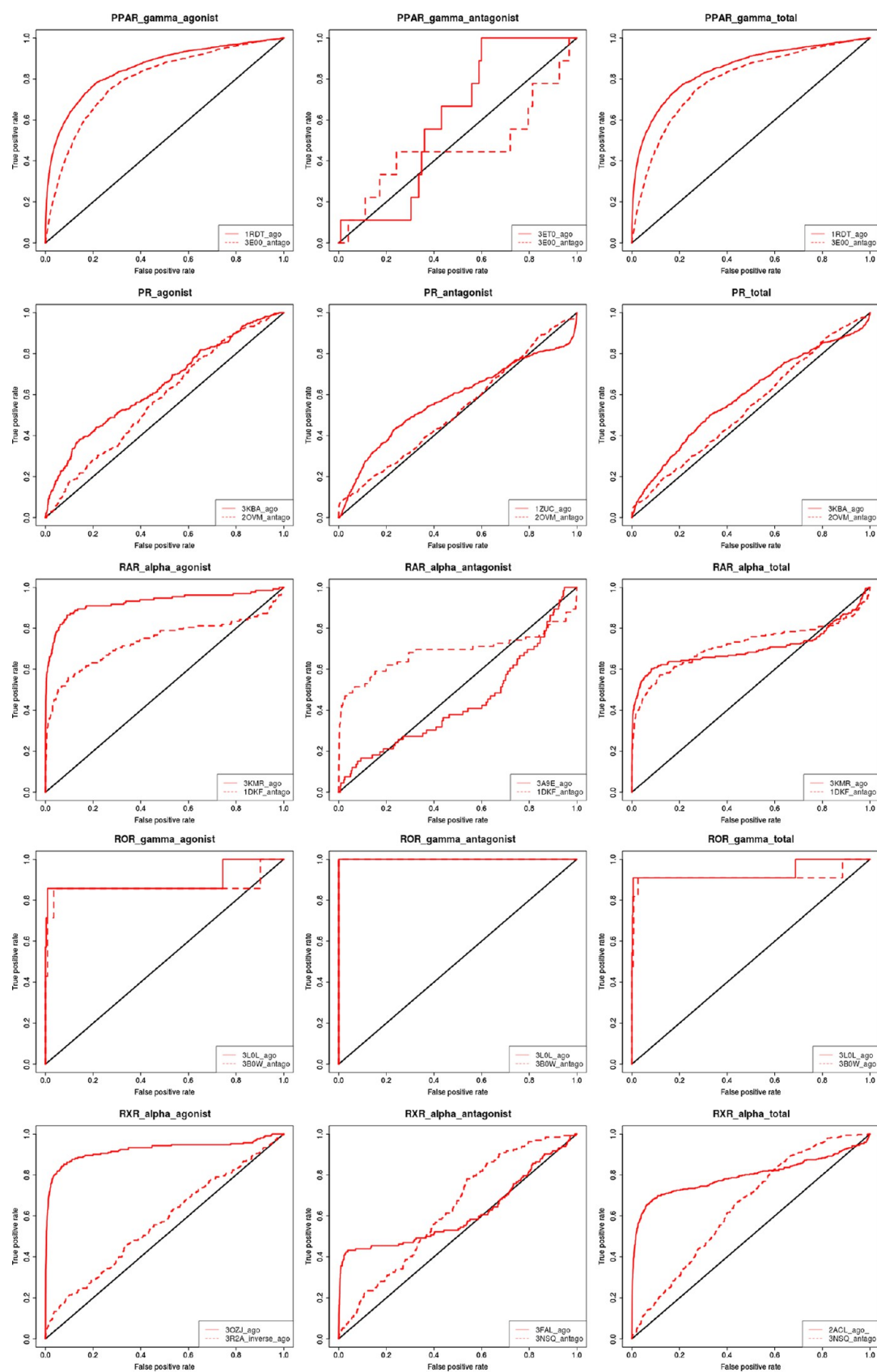


Figure 3. ROC curves obtained using Surflex-Dock with the agonist data set, the antagonist data set, or the total data set of a given NR on the best agonist-bound structure (solid lines) and the best antagonist-bound structure (dashed lines) for the 10 NRs presenting at least one agonist-bound structure and one antagonist-bound structure.

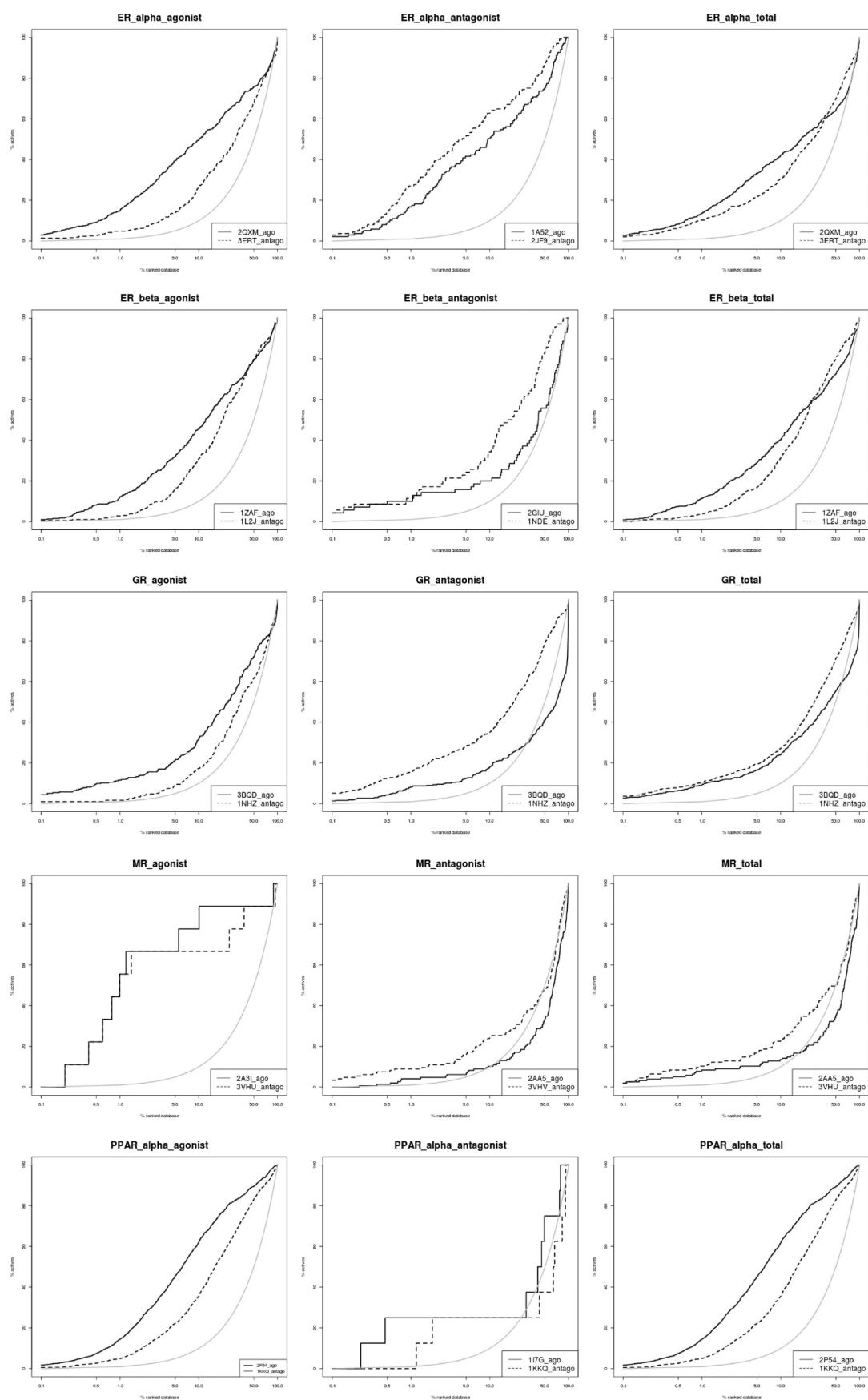


Figure 4. continued

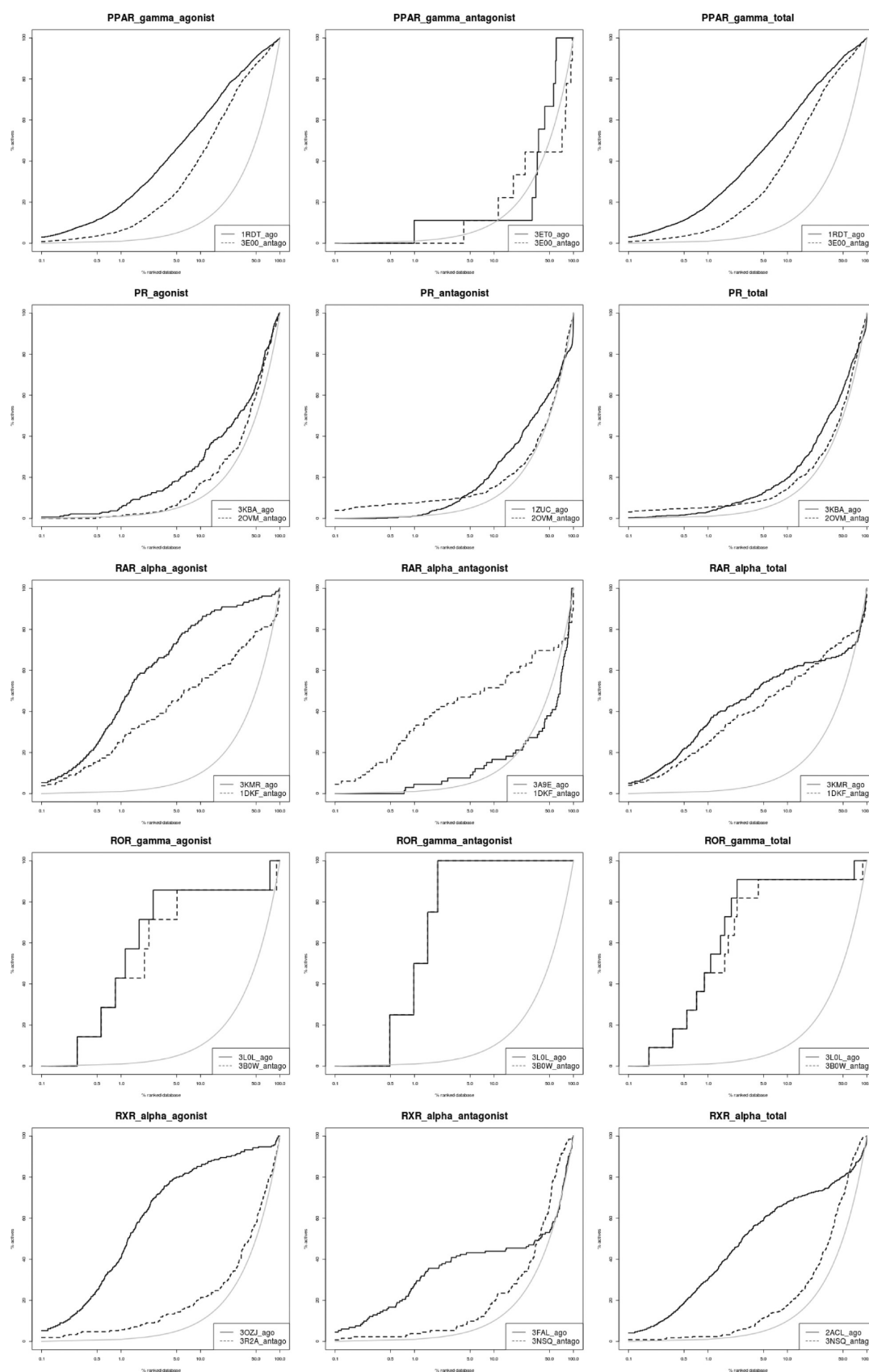


Figure 4. Enrichment graphs obtained using Surflex-Dock with the agonist data set, the antagonist data set, or the total data set of a given NR on the best agonist-bound structure (solid lines) and the best antagonist-bound structure (dashed lines) for the 10 NRs presenting at least one agonist-bound structure and one antagonist-bound structure.

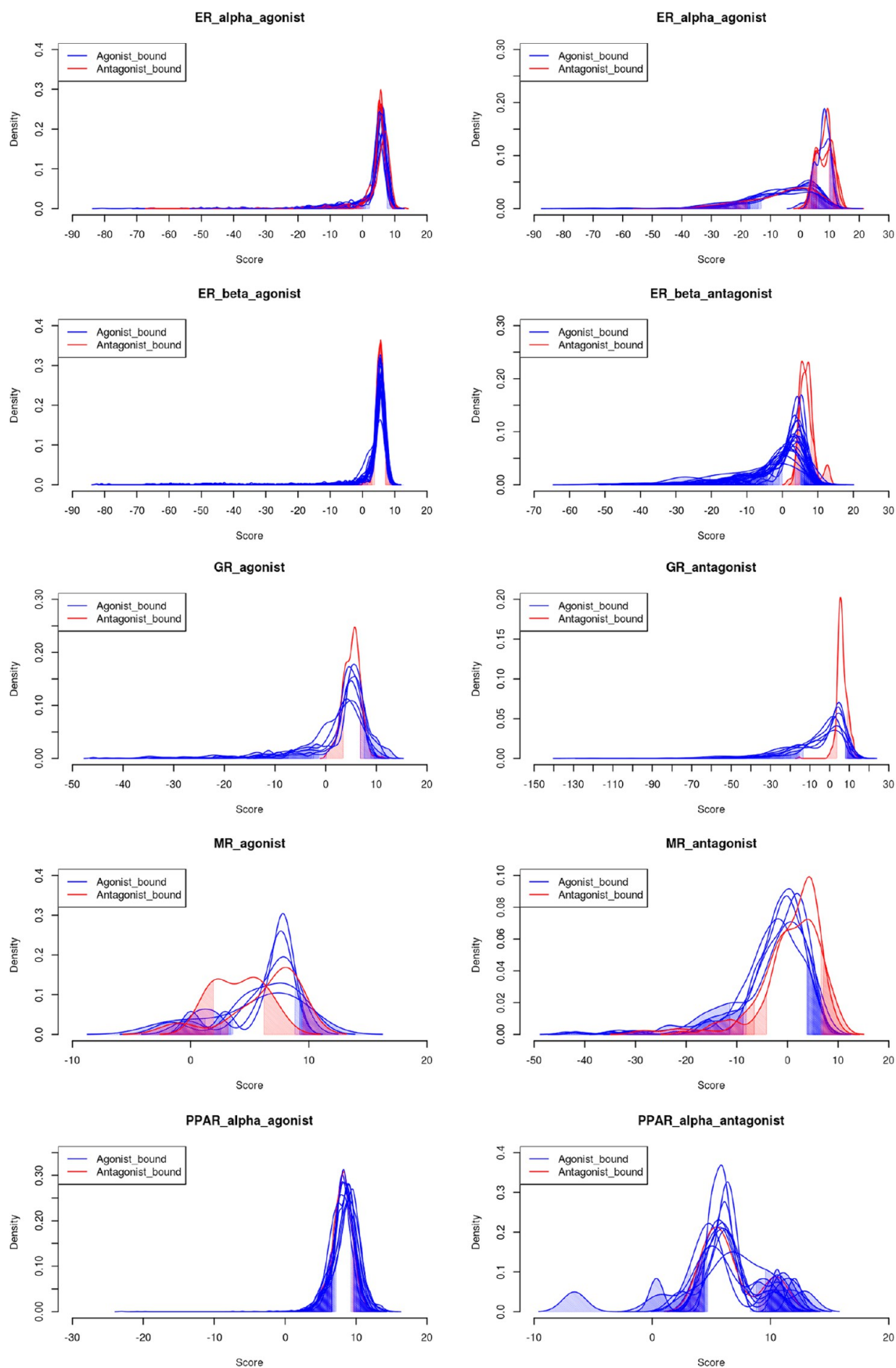


Figure 5. continued

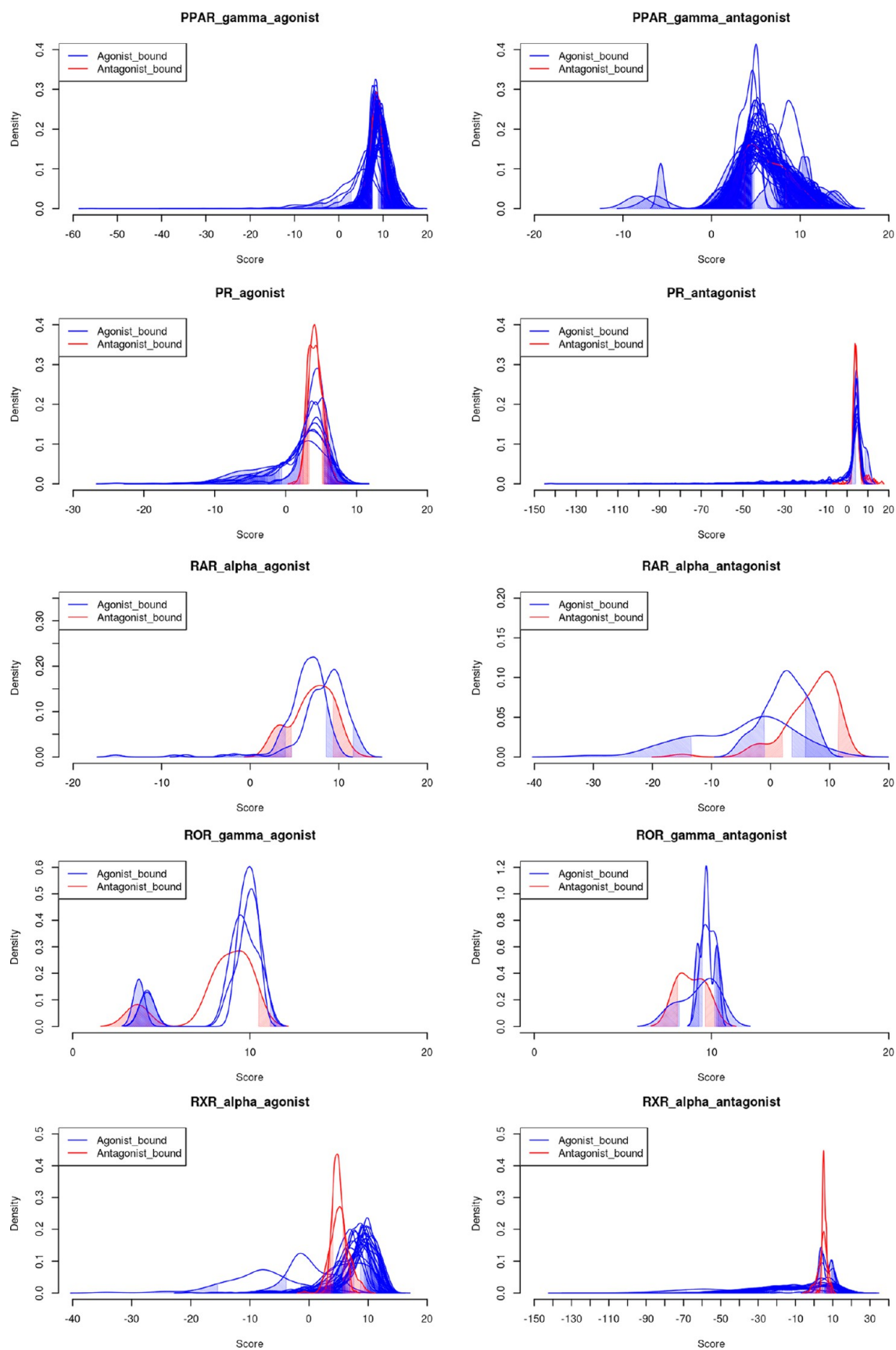


Figure 5. Score distribution curves obtained using Surflex-Dock with the agonist data set and the antagonist data set on all agonist-bound structures (blue) and all antagonist-bound structures (red) for the 10 NRs presenting at least one agonist-bound structure and one antagonist-bound structure.

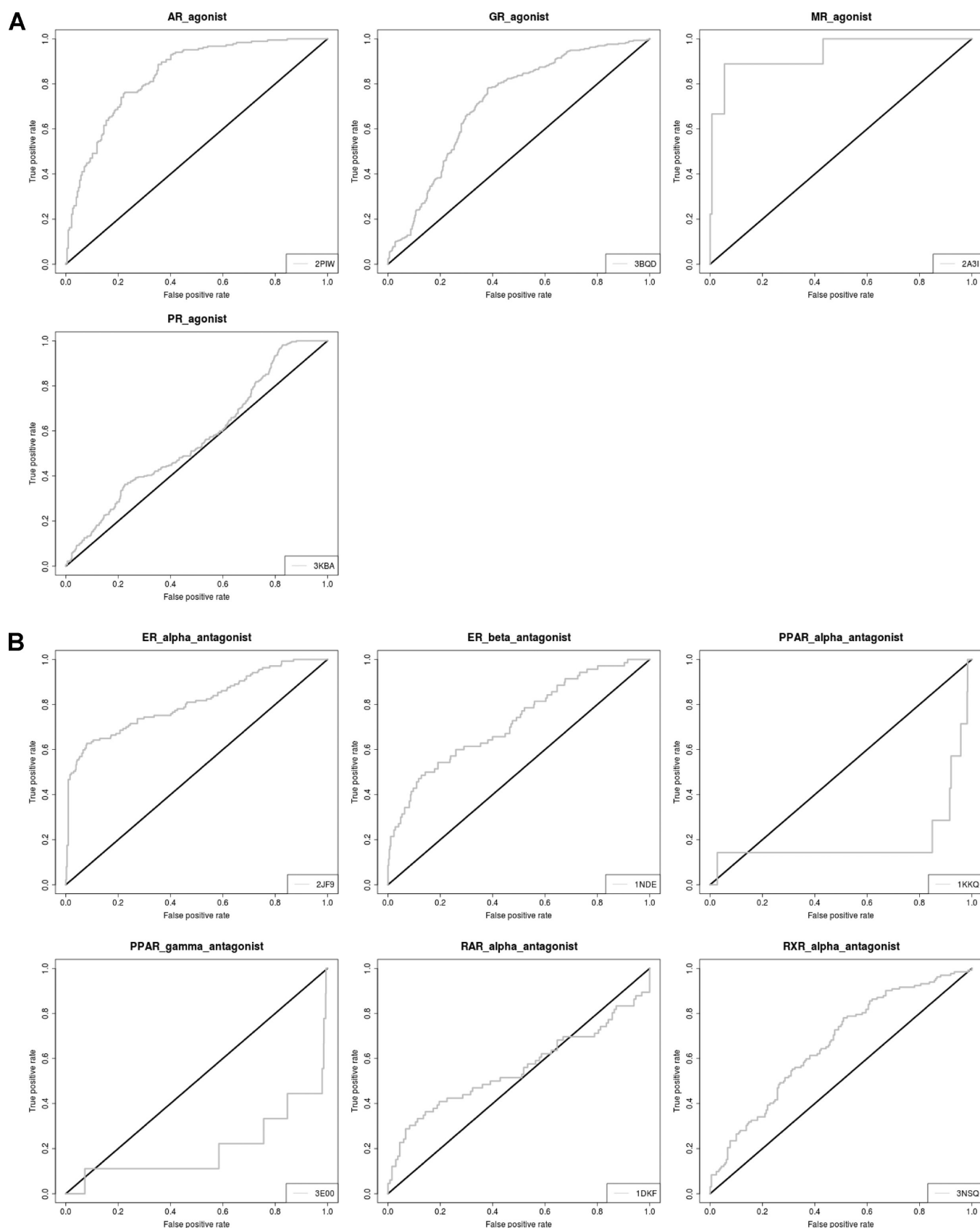


Figure 6. ROC curves obtained with (A) the agonist data sets of AR, GR, MR, and PR using the corresponding antagonist data sets as experimentally confirmed decoys and (B) the antagonist data sets of ER_alpha, ER_beta, PPAR_alpha, PPAR_gamma, RAR_alpha, and RXR_alpha using the corresponding agonist data sets as experimentally confirmed decoys.

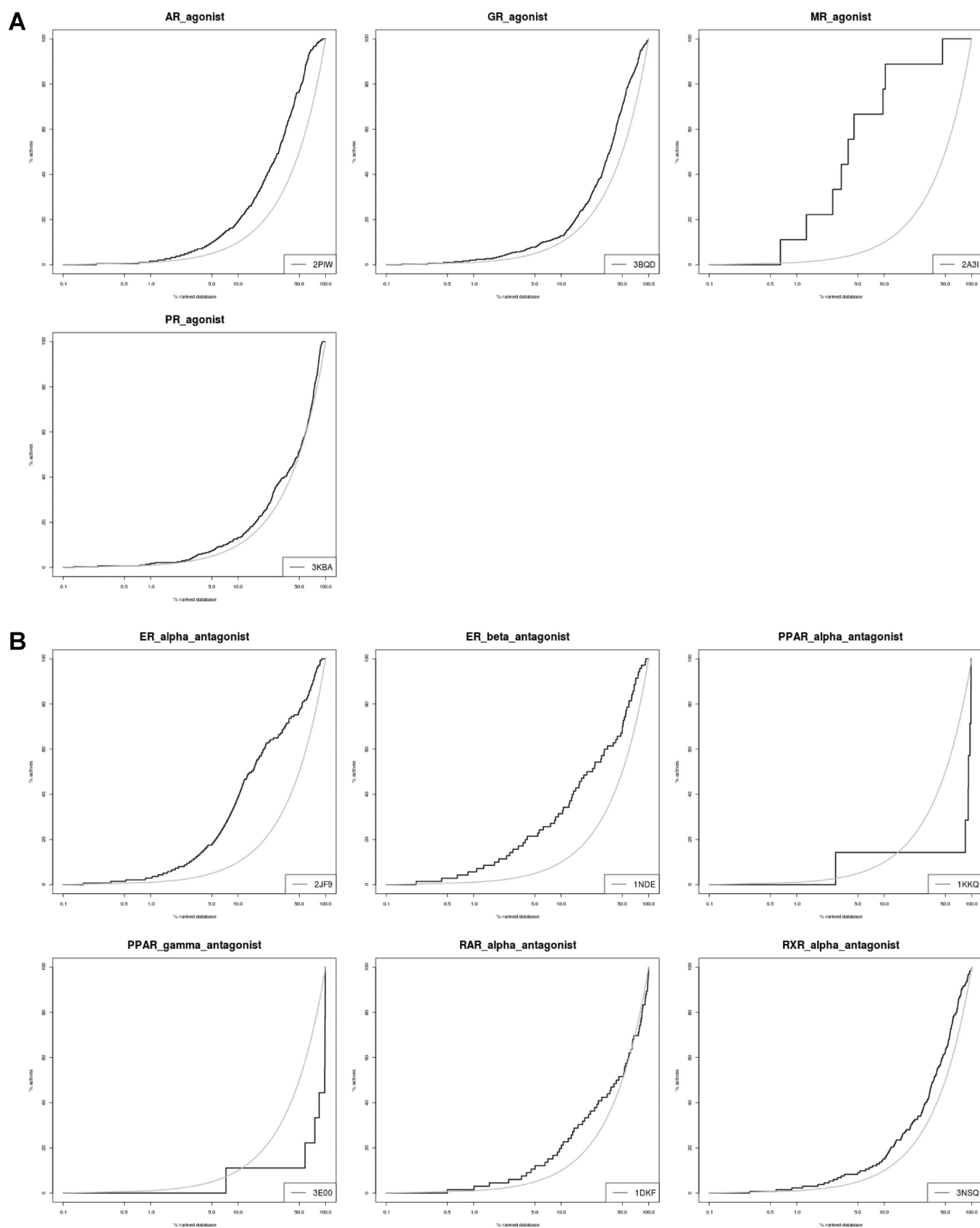


Figure 7. Enrichment graphs obtained with (A) the agonist data sets of AR, GR, MR, and PR using the corresponding antagonist data sets as experimentally confirmed decoys and (B) the antagonist data sets of ER_alpha, ER_beta, PPAR_alpha, PPAR_gamma, RAR_alpha, and RXR_alpha using the corresponding agonist data sets as experimentally confirmed decoys.

Table 2. Areas under the ROC Curve (AUC) and Enrichment Factors (EF) at 1% and 10% Obtained with Surflex-Dock on (1) the Agonist-Bound Structures Previously Found To Be Associated with the Best Performances for AR, GR, MR, and PR Using the Agonist Data Sets as Active Compounds and the Antagonist Data Sets as Decoy Ligands and (2) the Antagonist-Bound Structures Previously Found To Be Associated with the Best Performances for ER_alpha, ER_beta, PPAR_alpha, PPAR_gamma, RAR_alpha, and RXR_alpha Using the Antagonist Data Sets as Actives and the Agonist Data Sets as Ligand Decoys (A/D Is the Ratio of the Number of Active Compounds to the Number of Decoy Ligands for a Given Data Set)

NR (structure)	agonist data set				antagonist data set			
	A/D	AUC	EF _{1%}	EF _{10%}	A/D	AUC	EF _{1%}	EF _{10%}
AR (2PIW)	1/1.23	0.840	1.67	1.96				
ER_alpha (2JF9)					1/3.17	0.809	3.36	3.91
ER_beta (1NDE)					1/5.60	0.724	6.61	3.16
GR (3BQD)	1/1.25	0.716	2.25	1.29				
MR (2A3I)	1/16.22	0.937	17.22	8.04				
PPAR_alpha (1KKQ)					1/200.14	0.195	0.00	1.43
PPAR_gamma (3E00)					1/202.22	0.200	0.00	1.11
PR (3KBA)	1/1.96	0.564	1.69	1.35				
RAR_alpha (1DKF)					1/2.02	0.554	3.02	2.06
RXR_alpha (3NSQ)					1/11.83	0.664	2.59	1.52

These clear conformational differences may explain the diverse performances obtained using agonist- or antagonist-bound structures. It appears wise to use agonist-bound structures for the screening of agonist ligands and antagonist-bound structures for the screening of antagonist ligands, if available. In a previous study using only one receptor (ER), Liebeschuetz obtained similar results and conclusions.¹⁹

Exploring Better Decoys. The quality of a benchmarking database depends not only on the good selection of actives and structures: decoys play a central role in the evaluation process. Ideally, inactive compounds should be included, just as active compounds, on the basis of experimental data. Unfortunately, the compounds found to be experimentally inactive for a given target, often called “negative” data, are seldom documented or are insufficient to constitute benchmarking data sets. Thus, compounds presumed to be inactive for a given target, called decoys, have been selected on the basis of their nonsimilarity with known active compounds.²⁰ However, there is no evidence that some compounds selected as decoys are inactive, and the use of true inactive compounds instead of presumed decoys should enhance the quality of the evaluation. Using the NRLiSt BDB, we performed such a study by using antagonist compounds as decoy ligands for agonist activity and reciprocally. A compound that can be both an agonist and an antagonist is a modulator, and modulators were not included in the NRLiSt BDB.¹⁰ Since we wanted to have at least as many decoy ligands as active ligands, we performed the analyses on the 10 targets of the NRLiSt BDB that presented a sufficient number of ligands in each data set and an appropriate structure. We found that despite the difficulty due to experimental decoys, the performances in enrichment obtained using Surflex-Dock were still acceptable for a majority of the selected NRs. It is worthy of note that the observed performance depended on the ratio of the number of active compounds to the number of decoy ligands, as highlighted in the literature for putative decoys.^{7,13}

CONCLUSION

In the present work, we used the NRLiSt BDB, a new benchmarking database dedicated to NRs, to study the impact of the choices regarding its construction on enrichment with a SBVLS method, Surflex-Dock. We found that distinguishing agonists from antagonists actually enhances the quality of the

evaluation since the best docking performances were obtained with the separate data sets rather than with the total data sets. Furthermore, the choice of the structure used for both docking method evaluation and virtual ligand screening studies is of high importance, and the pharmacological profile of the ligand bound in the binding site was found to be a critical parameter. We finally have shown that the NRLiSt BDB active compound data set for a given target/pharmacological activity can be used as a set of decoy ligands for the other pharmacological activity to ensure a reliable and challenging evaluation of virtual screening methods. On the basis of this work, we have shown that (1) the NRLiSt BDB constitutes a high-quality and reliable benchmarking data set that can be used for the evaluation of virtual screening methods, (2) the rationale used to construct databases such as the NRLiSt BDB could become a reference for developing better benchmarking data sets, and (3) the use of active ligand data sets for a given target and for a given pharmacological activity as experimentally validated decoy ligand data sets for the same target but for another pharmacological activity could ensure a more robust and challenging evaluation.

AUTHOR INFORMATION

Corresponding Author

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

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank Prof. Jain for generously providing the Surflex package and Chemaxon for generously providing the Marvin suite. N.L. is a beneficiary of a CIFRE Fellowship from ANRT.

REFERENCES

- Mestres, J. Virtual screening: A real screening complement to high-throughput screening. *Biochem. Soc. Trans.* **2002**, *30*, 797–799.
- Oprea, T. I.; Matter, H. Integrating virtual screening in lead discovery. *Curr. Opin. Chem. Biol.* **2004**, *8*, 349–358.
- Bissantz, C.; Folkers, G.; Rognan, D. Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. *J. Med. Chem.* **2000**, *43*, 4759–4767.

(4) Miteva, M. A.; Lee, W. H.; Montes, M. O.; Villoutreix, B. O. Fast structure-based virtual ligand screening combining FRED, DOCK, and Surflex. *J. Med. Chem.* **2005**, *48*, 6012–6022.

(5) Montes, M.; Miteva, M. A.; Villoutreix, B. O. Structure-based virtual ligand screening with LigandFit: Pose prediction and enrichment of compound collections. *Proteins* **2007**, *68*, 712–725.

(6) Pham, T. A.; Jain, A. N. Parameter estimation for scoring protein–ligand interactions using negative training data. *J. Med. Chem.* **2006**, *49*, 5856–5868.

(7) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking sets for molecular docking. *J. Med. Chem.* **2006**, *49*, 6789–6801.

(8) 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.* **2012**, *55*, 6582–6594.

(9) Bourguet, W.; Germain, P.; Gronemeyer, H. Nuclear receptor ligand-binding domains: Three-dimensional structures, molecular interactions and pharmacological implications. *Trends Pharmacol. Sci.* **2000**, *21*, 381–388.

(10) Lagarde, N.; Ben Nasr, N.; Jeremie, A.; Guillemain, H.; Laville, V.; Labib, T.; Zagury, J. F.; Montes, M. NRLiSt BDB, the Manually Curated Nuclear Receptors Ligands and Structures Benchmarking Database. *J. Med. Chem.* **2014**, *57*, 3117–3125.

(11) 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.

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

(13) Hawkins, P. C.; Warren, G. L.; Skillman, A. G.; Nicholls, A. How to do an evaluation: Pitfalls and traps. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 179–190.

(14) Thomas, M. P.; McInnes, C.; Fischer, P. M. Protein structures in virtual screening: A case study with CDK2. *J. Med. Chem.* **2006**, *49*, 92–104.

(15) Jain, A. N. Surflex: Fully automatic flexible molecular docking using a molecular similarity-based search engine. *J. Med. Chem.* **2003**, *46*, 499–511.

(16) Sing, T.; Sander, O.; Beerenwinkel, N.; Lengauer, T. ROCRC: Visualizing classifier performance in R. *Bioinformatics* **2005**, *21*, 3940–3941.

(17) Wilcoxon, F. Individual Comparisons by Ranking Methods. *Biom. Bull.* **1945**, *1*, 80–83.

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

(19) Liebeschuetz, J. W. Evaluating docking programs: Keeping the playing field level. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 229–238.

(20) Kirchmair, J.; Markt, P.; Distinto, S.; Wolber, G.; Langer, T. Evaluation of the performance of 3D virtual screening protocols: RMSD comparisons, enrichment assessments, and decoy selection—What can we learn from earlier mistakes? *J. Comput.-Aided Mol. Des.* **2008**, *22*, 213–228.

1.3.3 Discussion

Pour réaliser cette étude, nous avons mis à contribution notre banque d'évaluation maison, la NRLiSt BDB, pour évaluer l'impact de :

- La construction des jeux de données distincts pour les ligands agonistes et les ligands antagonistes sur la qualité des banques d'évaluation ;
- L'utilisation du profil pharmacologique du ligand co-cristallisé comme guide dans le choix de la structure de référence pour les méthodes de docking ;
- L'inclusion de ligands decoys expérimentaux.

1.3.3.1 Influence de l'utilisation de jeux de données séparés sur les performances du docking

De nombreuses banques d'évaluation ont été développées au cours du temps^{304, 418, 423, 427, 481, 492, 493} et la qualité d'évaluation proposée par certaines d'entre elles-elles (la DUD⁴²³ et la DUD_E⁴²⁷) leur a permis d'acquérir le statut de banques d'évaluation de référence. Cependant, la qualité de ces banques peut et doit encore être améliorée, notamment par construction de jeux de données séparés selon le profil pharmacologique des ligands, comme récemment réalisé pour une base de données dédiées aux récepteurs couplés aux protéines G⁴⁸¹ ou dans la NRLiSt BDB⁴⁰.

Dans notre étude, nous avons donc comparé les performances obtenues pour les 27 cibles de la NRLiSt BDB avec une méthode de criblage virtuel basée sur les structures, Surflex-Dock, lors de l'utilisation des jeux de données séparés agonistes et antagonistes et d'un jeu de données « total » réunissant les deux jeux de données précédent en un seul. Il est alors très rapidement et très clairement apparu des différences significatives en termes d'enrichissement. En effet, pour chacun des 27 RNs, le jeu de données associé au meilleur enrichissement était, dans tous les cas, soit le jeu de données agoniste soit le jeu de données antagoniste. De plus, en analysant les résultats structure par structure, nous avons pu remarquer que dans respectivement 97% et 89% des cas, le meilleur AUC et le meilleur facteur d'enrichissement précoces ont été obtenus avec l'un des deux jeux de données séparés. Ces résultats sont encore renforcés puisque pour 26% des cibles de notre étude, les valeurs individuelles d'AUC obtenues avec toutes les structures en utilisant un des jeux de données séparé sont supérieures à la meilleure valeur d'AUC obtenue avec le jeu de données total. Ce

pourcentage atteint 67% lorsque l'on considère les cibles pour lesquelles, en utilisant le jeu de données agoniste ou antagoniste, au moins 50% des structures présentent une valeur d'AUC supérieure à la plus haute valeur d'AUC obtenue avec le jeu de données total. Mélanger les ligands agonistes et antagonistes semble donc créer un biais dans l'évaluation des méthodes de criblage virtuel basées sur les structures et séparer les jeux de données selon le profil pharmacologique des ligands inclus semble donc être un moyen très intéressant pour améliorer la qualité des banques d'évaluation.

Nous avons aussi pu constater à nouveau⁴⁹⁵⁻⁴⁹⁸, que le choix de la structure de départ est très important. En effet, nous avons observé que, dans la grande majorité des cas, pour une cible donnée, la structure associée au meilleur AUC ou au meilleur enrichissement précoce était la même en utilisant un des jeux de données séparé ou le jeu de données total. Cependant, déterminer quelle(s) structure(s) choisir pour réaliser un criblage virtuel basé sur les structures reste incertain. Nous nous sommes alors intéressés au profil pharmacologique du ligand co-cristallisé en tant que critère potentiel de sélection de la structure de référence.

1.3.3.2 Importance du ligand co-cristallisé dans la structure utilisée pour le criblage

Des changements conformationnels surviennent dans la structure de la protéine lors de la liaison d'un ligand. Dans le cas des récepteurs nucléaires, ces changements sont dépendants du profil pharmacologique du ligand, et nous avons donc décidé d'étudier l'influence de ce critère sur l'enrichissement obtenu avec Surflex-Dock.

Pour chaque récepteur nucléaire présentant au moins une structure liée à un agoniste et une structure liée à un antagoniste, la structure associée au meilleur enrichissement était toujours une structure liée un agoniste lorsque le jeu de données agoniste était utilisé, et dans la majorité des cas, une structure liée à un antagoniste en utilisant le jeu de données antagoniste. L'analyse de la distribution des scores des ligands agonistes et des ligands antagonistes en fonction du profil pharmacologique du ligand de la structure de référence a aussi permis de mettre en évidence des différences significatives. Ces résultats peuvent facilement être rationalisés au niveau moléculaire. En effet, comme déjà présenté précédemment (1.2.1.1), les différences conformationnelles observables après liaison de ligands agonistes et antagonistes ont été décrit⁴⁸⁰. Lors de la liaison d'un agoniste, l'assemblage en « sandwich d'hélices alpha anti-parallèle » est modifié avec repositionnement des hélices H11, H12 et H3 permettant la création d'une surface de recrutement de co-activateurs. A l'opposé, le

repositionnement de l'hélice H12 induit par la liaison d'un antagoniste ne permet plus la création de cette surface de recrutement de co-activateurs.⁴⁸⁰

Le profil pharmacologique du ligand co-cristallisé peut donc être utilisé avec confiance comme critère de sélection rationnel de la structure de départ pour réaliser un criblage virtuel basé sur les structures. Ainsi, pour rechercher de nouveaux agonistes, une structure liée à un agoniste devrait être préférée, et à l'inverse, lorsque la prospection concerne de nouveaux antagonistes, une structure liée à un antagoniste. Dans une étude datant de 2008 menée sur le seul récepteur aux œstrogènes, des résultats et conclusions similaires avaient été émis par Liebeschuetz³⁶⁹.

1.3.3.3 Recherche de nouveaux decoys

Au-delà des jeux de données d'actifs et de la sélection rationnelle d'une structure de départ, l'évaluation des méthodes de criblage virtuel peut être améliorée du point de vue de l'obtention de decoys plus appropriés. En effet, idéalement, des composés inactifs devraient être inclus sur la base de données expérimentales, à l'instar des composés actifs. Cependant, peu de données négatives de ce type sont disponibles dans la littérature, et il est donc difficile de construire des banques d'évaluation avec de réels inactifs. En effet, lorsqu'une série de composés synthétisés se révèle inactive, cette série est généralement abandonnée, sans faire l'objet de publication. Les données négatives les plus fréquemment disponibles sont celles de composés peu ou pas actifs à l'intérieur d'une série de molécules dont certaines présentent, au contraire, des valeurs d'activité intéressantes. Par défaut, des « decoys », c'est-à-dire des composés présumés inactifs pour une cible, sont donc employés. Cependant, l'utilisation de ces decoys est toujours sujette à caution puisqu'il est impossible de savoir si ces decoys sont réellement inactifs. Malgré les différents efforts réalisés pour obtenir les decoys les plus adaptés à l'évaluation des méthodes de criblage virtuel, l'inclusion de vrais inactifs devrait permettre d'améliorer la qualité des banques d'évaluation. A l'aide des jeux de données de la NRLiSt BDB, nous avons donc testé l'utilisation des composés antagonistes comme ligands decoys du jeu de données agoniste et inversement des agonistes comme ligands decoys du jeu de données antagoniste. En effet, même si les decoys que nous proposons d'utiliser sont capable de se lier à la cible étudiée, ce sont de réels inactifs pour l'activité biologique considérée puisqu'un composé à la fois agoniste et antagoniste possède un profil pharmacologique « modulateur » et ce type de ligands n'a pas été inclus dans la NRLiSt BDB⁴⁹⁴. Pour les 10 récepteurs nucléaires avec suffisamment de données pour réaliser cette étude,

nous avons pu constater que, même si l'évaluation des méthodes résultante était très exigeante, puisqu'il ne s'agit plus seulement d'identifier un ligand mais bien un ligand avec une activité biologique précise, pour la majorité des cibles, les performances obtenues avec Surflex-Dock étaient acceptables. Nous avons aussi pu constater, et ce assez logiquement, une corrélation entre l'enrichissement et le ratio du nombre d'actifs et de decoys. En effet, les performances du docking étaient médiocres pour les jeux de données avec un nombre de decoys très largement supérieur au nombre d'actifs, mais tout à fait correctes lorsque le nombre de decoys était plus adapté. Ce problème n'est cependant pas spécifique aux ligands decoys puisque des conclusions similaires ont précédemment été rapportés pour les decoys putatifs^{423, 497}.

1.3.4 Analyse critique de l'étude

Pour réaliser cette étude, nous avons utilisé la NRLiSt BDB, une banque d'évaluation construite dans notre laboratoire, que nous maîtrisons donc parfaitement. Cette banque répondait à tous les critères que nous recherchions, puisque le profil pharmacologique agoniste ou antagoniste de chaque ligand est renseigné et est mis à profit pour créer 2 jeux de données et classer les structures des récepteurs.

La masse de données disponibles était largement suffisante pour la première et la deuxième partie de notre travail, qui concernaient l'influence sur les performances d'une méthode de docking de, respectivement, l'utilisation de jeux de données d'actifs séparés en fonction de la nature de leurs profils pharmacologiques et du profil pharmacologique du ligand co-cristallisé. Malheureusement, ce n'était pas le cas pour la troisième partie de notre étude consacrée à l'utilisation de ligands decoys expérimentaux. En effet, peu de récepteurs présentaient des jeux de données dont le déséquilibre entre le nombre d'agonistes et d'antagonistes permettait d'utiliser les uns comme les ligands decoys des autres. Nous projetons donc de poursuivre ce travail en incluant des decoys expérimentaux supplémentaires. Pour cela, une nouvelle investigation de la littérature scientifique sera nécessaire pour rechercher des composés décrits comme étant inactifs pour une cible donnée, et que nous avons précédemment exclus lors de la construction de la NRLiSt BDB (Figure 77).

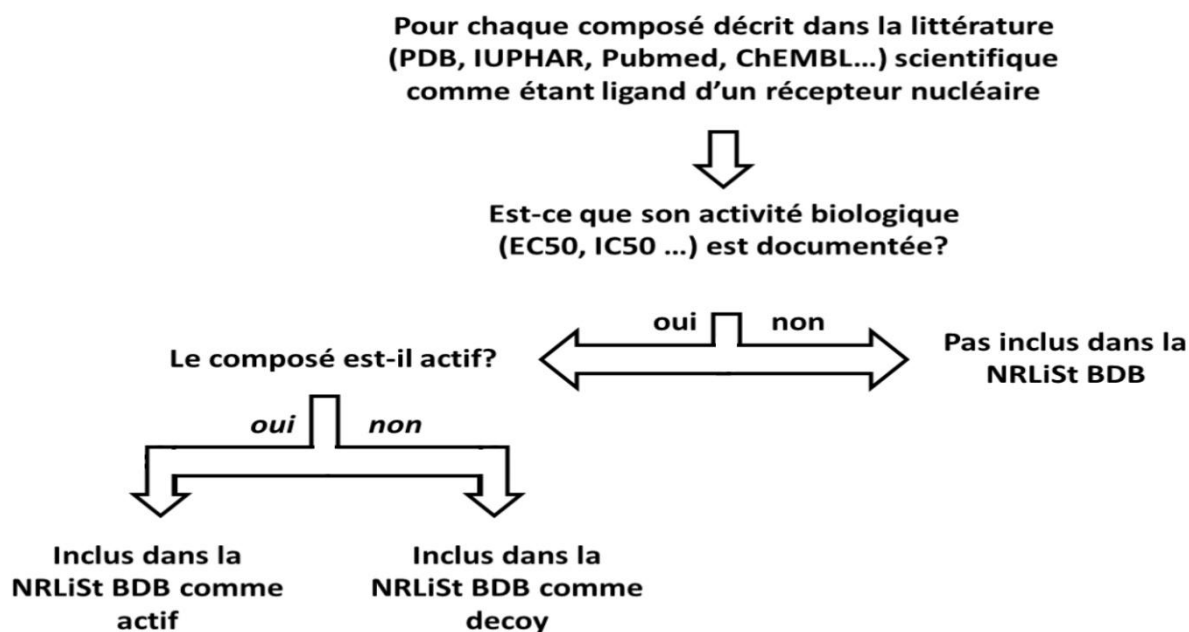


Figure 77. Stratégie prévisionnelle d'intégration de nouveaux decoys expérimentaux dans la version mise à jour de la NRLiSt BDB

Cette étude a été réalisée avec un seul logiciel de docking, Surflex-Dock. Ce logiciel utilise un algorithme de recherche conformationnelle de fragmentation / reconstruction des ligands et une fonction de score empirique. Il serait donc utile de renouveler cette étude avec d'autres logiciels de docking, mettant en œuvre des algorithmes de recherche conformationnelle et des fonctions de score différents pour valider les conclusions présentées ici.

Pour analyser les performances obtenues lors du docking, nous avons utilisé à la fois un descripteur de performances globales, l'AUC, et des descripteurs de performances précoces, EF1% et EF10%. Les résultats obtenus avec ces deux types de descripteurs étaient similaires et nous ont permis de dégager des tendances communes.

Enfin, la NRLiSt BDB est une banque d'évaluation construite pour l'évaluation des méthodes de criblage virtuel basées sur les structures mais aussi sur les ligands. Dans notre étude, nous n'avons considéré qu'une méthode de criblage virtuel basée sur les structures. Il serait donc intéressant d'utiliser la NRLiSt BDB avec différentes méthodes basées sur les ligands (pharmacophore, QSAR, etc...) pour valider l'utilisation possible de la NRLiSt BDB pour évaluer ce type de méthodes.

1.3.5 Conclusion

La NRLiSt BDB, une banque d'évaluation publiée très récemment et dédiée aux récepteurs nucléaires, a donc été utilisée dans ce travail pour analyser les conséquences des partis pris

durant la construction de cette banque sur les performances d'une méthode de criblage virtuel basée sur les structures, Surflex-Dock. Nous avons confirmé que, comme prédit dans la publication présentant la NRLiSt BDB, construire des jeux de données séparés selon le profil pharmacologique agoniste ou antagoniste des ligands permet réellement d'améliorer la qualité de l'évaluation. En effet, les performances obtenues lors du docking des jeux de données séparés de la NRLiSt BDB étaient meilleures que celles obtenues avec un seul jeu de données mélangeant tous les ligands pour un récepteur donné. Nous avons donc proposé que, pour la construction de futures banques d'évaluation, le profil pharmacologique des ligands soient précisément recherchés et pris en compte lors de la constitution des jeux de données. Nous avons aussi étudié l'influence du profil pharmacologique du ligand lié dans la structure de référence, puisque celle-ci est très importante, que ce soit lors de l'évaluation des méthodes ou lors des criblages virtuels prospectifs eux-mêmes. Il est clairement apparu que celui-ci était un paramètre critique, et que lorsque cela était possible, une structure avec un ligand co-cristallisé du même profil pharmacologique que les ligands étudiés devrait être utilisée. Ceci constitue donc une deuxième ligne directrice pour la construction de banques d'évaluation mais aussi, et peut être de façon plus importante encore, pour la réalisation de criblage virtuel à la recherche de nouveaux médicaments. Finalement, après avoir présenté des améliorations possibles de la qualité des banques d'évaluations concernant les actifs puis les structures, nous avons étudié si il était possible de proposer des decoys plus adaptés. Nous avons ainsi montré que les actifs d'un certain profil pharmacologique pour un récepteur donné pouvaient être utilisés comme decoys pour le profil pharmacologique opposé, lorsque leur nombre était suffisant. L'évaluation des méthodes résultante était alors robuste mais aussi plus stringente puisque la méthode de criblage virtuel doit être capable de distinguer non plus les ligands des non ligands mais les molécules actives pour un profil d'activité donné des non actives.

En résumé, ce travail a donc permis de montrer que :

- (1) La NRLiSt BDB est une banque d'évaluation fiable et de haute qualité, utilisable pour l'évaluation des méthodes de criblage virtuel ;
- (2) Les partis pris originaux utilisés durant la construction de la NRLiSt BDB devraient servir de référence pour le développement de nouvelles banques d'évaluation pour leur assurer une qualité d'évaluation la meilleure possible ;
- (3) Des efforts doivent être menés pour proposer des decoys plus adaptés à l'évaluation des méthodes tels que les ligands decoys expérimentalement validés ici présentés et ainsi remplacer les decoys putatifs actuellement utilisés.

2 Réalisation d'un criblage virtuel à la recherche de composés inhibant l'interleukine IL-6

2.1 La polyarthrite rhumatoïde

La polyarthrite rhumatoïde est une maladie chronique inflammatoire auto-immune caractérisée par une inflammation de la membrane synoviale (synovite) persistante et des destructions osseuses et cartilagineuses des articulations (Figure 78).

La polyarthrite rhumatoïde, comme son nom l'indique, atteint plusieurs articulations (typiquement plus de 6), les mains, les pieds et les genoux étant généralement les plus touchés⁴⁹⁹. Cette maladie, dont l'origine est toujours inconnue même si l'implication d'agents infectieux est suspectée, touche entre 0,5 et 1% de la population, majoritairement des femmes⁵⁰⁰.

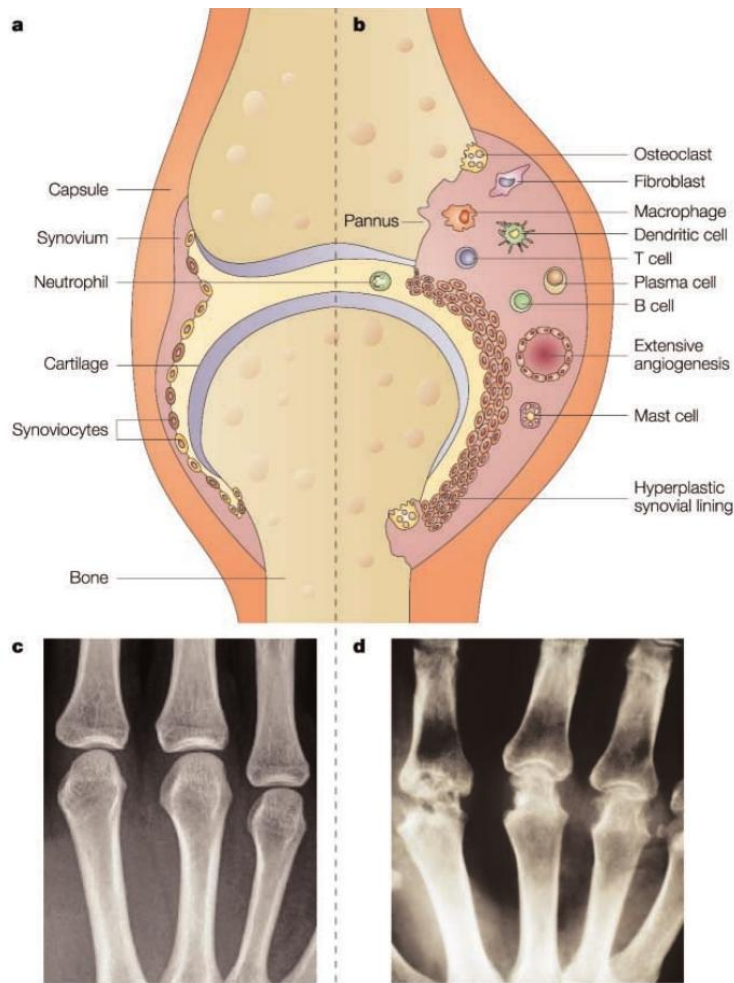


Figure 78. Représentation schématique d'une articulation normale (a) et d'une articulation atteinte dans un cas de polyarthrite rhumatoïde (b). Dans une articulation, deux extrémités osseuses se font face, couvertes par une couche de cartilage, séparées par un espace et entourées de la membrane synoviale et de la capsule articulaire. L'atteinte articulaire observée en cas de polyarthrite rhumatoïde est initialement caractérisée par une inflammation de la membrane synoviale résultant de l'afflux et l'activation locale de cellules variées (lymphocytes B et T, cellules plasmatiques et dendritiques, macrophages, mastocytes). Les ostéoclastes sont responsables des destructions survenant au niveau articulaire observées dans la polyarthrite rhumatoïde (c et d : radiographies respectives d'une main saine et atteinte de polyarthrite rhumatoïde) et qui débutent généralement à la jonction membrane synoviale-cartilage-os (la portion détruite de la membrane synoviale est appelé « pannus »).

501

Le protocole thérapeutique employé pour traiter la polyarthrite rhumatoïde combine l'utilisation d'anti-inflammatoires non-stéroïdiens et d'Anti-Rhumatismaux Modificateurs de

la Maladie ARMM (ou DMARDs : Disease-Modifying AntiRheumatic Drugs). Les anti-inflammatoires non stéroïdiens s'opposent aux symptômes de la maladie, alors que les ARMM tentent de s'opposer aux processus inflammatoires et destructifs de la polyarthrite rhumatoïde.⁵⁰¹ Ces derniers peuvent être classifiés en deux catégories : les ARMM conventionnels (ARMMc) et les agents biologiques (ARMMb) et ciblent différents acteurs du processus biologique complexe de la polyarthrite rhumatoïde (Tableau 18).

ARMMc			
Médicament	Cible	Effet biologique	
Méthotrexate	DHFR, TS, ATIC, AICAR	Inhibition de la prolifération lymphocytaire, effets anti-inflammatoires	
Léflunomide	DHODH	Inhibition de la prolifération lymphocytaire	
Sulfasalazine	Enzymes folate-dépendants	Inhibition de la prolifération lymphocytaire, effets anti-inflammatoires, induction de l'apoptose (neutrophiles et macrophages)	
Sels d'or	Acides aminés soufrés	Inhibition du signal de transduction et de la présentation de l'antigène	
(Hydro)chloroquine	Lysosomes, enzymes lysosomales, TLR-9	Inhibition de la présentation de l'antigène et de l'activation du système immunitaire inné	
Tofacitinib	JAK1 et JAK3	Inhibition de la maturation et de l'activation des leucocytes, inhibition de la production de cytokines et d'immunoglobulines	
ARMMb			
Médicament	Cible	Mécanisme d'action	Structure
Infliximab	TNF	Liaison de l'Ac au TNF	Ac monoclonal chimérique murin-humain IgG1
Etanercept	TNF	Récepteur leurre pour la liaison du TNF	Protéine de fusion recombinante soluble humaine
Adalimumab	TNF	Liaison de l'Ac au TNF	Ac monoclonal humain
Golimumab	TNF	Liaison de l'Ac au TNF	Ac monoclonal humain
Certolizumab pegol	TNF	Liaison du fragment Fab' au TNF	Fragment Fab' anti-TNF pegylée humanisée
Anakinra	IL-1	Liaison au récepteur IL-1 type 1	Antagoniste recombinant du récepteur de l'IL-1 humain
Tocilizumab	IL-6	Liaison au récepteur de l'IL-6	Ac monoclonal recombinant humanisé du récepteur de l'IL-6 humain
Abatacept	Lymphocyte T	Blocage de la co-stimulation des lymphocytes T	Protéine de fusion CTLA-4 IgG1
Rituximab	Lymphocyte B	Liaison et déplétion des cellules B-CD20+	Ac monoclonal IgG1K chimérique murin-humain

Tableau 18. Présentation des différents AMRR actuellement disponible (Ac: Anticorps, AICAR: 5-aminoimidazole-4-carboxamide ribonucleotide, ATIC : 5-aminoimidazole-4-carboxamide ribonucléotide transformylase, DHFR : Dihydrofolate reductase, DHPDH : Dehydroorotate dehydrogenase, TLR : Toll-like receptor, TS : Thymidilate synthase) (d'après ⁵⁰²)

Ainsi, de façon assez surprenante, les ARMMc et les ARMMb actuellement commercialisés ne ciblent pas les mêmes entités biologiques. Cependant, obtenir un traitement d'efficacité similaire aux ARMMb mais sans les effets indésirables inhérents aux produits biologiques représenterait une avancée thérapeutique. En effet, l'utilisation des ARMMb et leur fabrication présentent de nombreux inconvénients. Ainsi, le protocole thérapeutique d'utilisation des ARMMb est contraignant pour les patients puisqu'ils sont injectés par voie parentérale à l'inverse des ARMMc généralement administrable par voie orale, préférée par les patients. De plus, les risques d'infections lors de l'utilisation de produits biologiques sont importants et les demi-vies plasmatiques élevées ont comme conséquence néfaste la persistance des effets après l'arrêt du traitement. En ce qui concerne la préparation industrielle et la distribution des ARMMb, la nécessité d'obtenir et de conserver un produit stérile complique ces processus, ce qui explique en partie les prix très élevés des ARMMb. Ces conclusions ont conduit différents groupes de recherche, parmi lesquels l'équipe GBA (Génomique Bioinformatique et Applications, EA4627) du CNAM (Conservatoire National des Arts et Métiers) de Paris au sein duquel ce travail de thèse a été réalisé, à rechercher des ARMMc capable d'interagir avec les cibles biologiques des ARMMb. Ainsi, différents inhibiteurs du TNF α « non biologiques » ont déjà été proposés⁵⁰³⁻⁵⁰⁸ parmi lesquels un dérivé benzène-sulfonamide, découvert par notre équipe à l'aide d'un criblage virtuel basé sur la structure, capable d'inhiber *in vitro* le TNF α au niveau micromolaire et constituant le premier inhibiteur du TNF α actif *in vivo* dans un modèle animal agissant directement sur le TNF α ⁵⁰⁹. Après ce premier succès, nous avons décidé de nous intéresser à une autre cytokine impliquée dans la polyarthrite rhumatoïde, déjà ciblée par un ARMMb, l'IL-6.

2.2 Interleukine IL-6

L'interleukine 6 (IL-6) est une glycoprotéine pléiotropique (Figure 79) de 21 kDa⁵¹⁰.

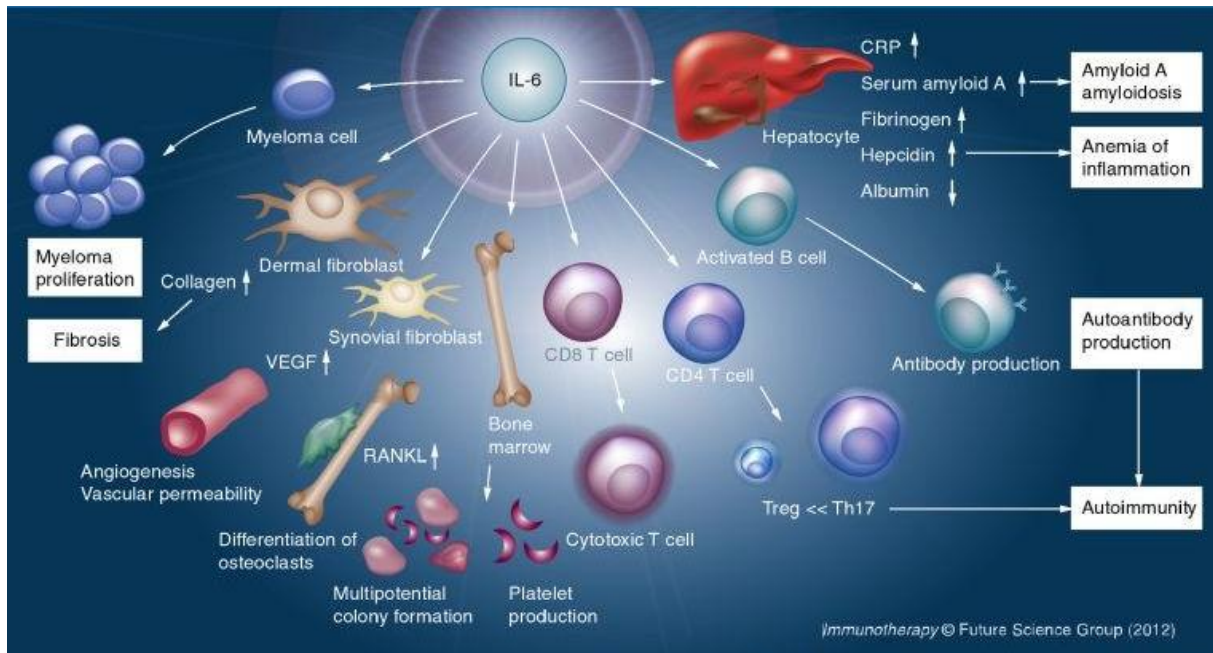


Figure 79. L'IL-6 régule l'action de nombreux partenaires cellulaires, soit par induction de leur production, inhibition de leur synthèse ou stimulation de leur différenciation.⁵¹¹

L'IL-6 intervient notamment dans les phénomènes inflammatoires, par induction de la production de la protéine C réactive, CRP, des protéines de la famille sérum amyloïde A, de l'haptoglobine, du fibrinogène et réduction de la synthèse d'albumine, de fibronectine et de transferrine⁵¹². L'IL-6 permet aussi de réguler l'immunité en étant un facteur de différenciation des lymphocytes B en cellule productrice d'anticorps, mais aussi des lymphocytes T CD4 en Th17⁵¹³ et des lymphocytes T CD8 en cellules T cytotoxiques⁵¹⁴. L'hématopoïèse est elle aussi l'une des cibles de l'IL-6, et plus particulièrement la maturation des mégacaryocytes en plaquettes et l'activation des cellules souches hématopoïétiques. Parmi les nombreux autres rôles de l'IL-6, il est possible de citer la régulation de l'ostéogenèse par promotion de la formation des cellules « osteoclast-like »⁵¹⁵ et induction de la protéine RANKL (Receptor Activator of NF-κB Ligand)⁵¹⁶ impliquée dans la différenciation et l'activation des ostéoclastes, ainsi que l'induction de la production de VEGF (Vascular Endothelium Growth Factor)⁵¹⁷, de kératinocytes⁵¹⁸ et de collagène⁵¹⁹.

L'IL-6 agit par liaison au récepteur de l'IL-6 (IL-6R)^{520, 521}, présent physiologiquement sous deux formes, une transmembranaire et une soluble, puis formation d'un complexe hexamérique (Figure 80) avec la protéine gp130 (composé de deux molécules de chacun des trois partenaires IL-6, IL-6R et gp130)⁵²².

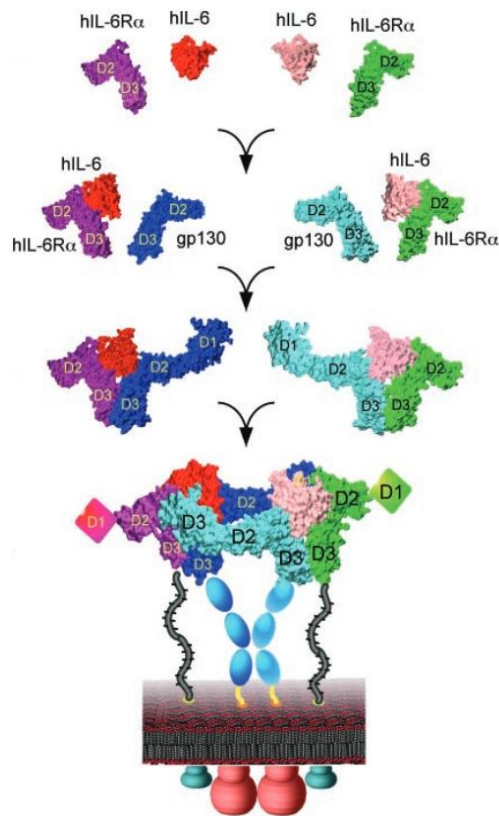


Figure 80. Formation du complexe hexamérique activé IL-6/IL-6R/gp130 par assemblage séquentiel et coopératif des 3 partenaires. Après liaison de l'IL-6 à l'IL-6R, un complexe trimérique est formé par addition de la gp130. Deux complexes ainsi formés peuvent ensuite interagir pour former le complexe hexamérique fonctionnel (avec D1 : domaine d'activation « immunoglobulin-like » de la gp130 et D2D3 : région homologue de liaison des cytokines CHR (Cytokine binding Homology Region) de la gp130 et de l'IL-6R) ⁵²²

Le rôle de dérégulations dans la production de l'IL-6 dans l'apparition de différentes maladies auto-immunes et inflammatoires a été mis en évidence ⁵²³⁻⁵²⁵. C'est notamment le cas de la polyarthrite rhumatoïde dont un des traitements actuellement commercialisés, le tocilizumab ou Actemra®, est un anticorps anti-IL-6R (Figure 81) ⁵²⁶.

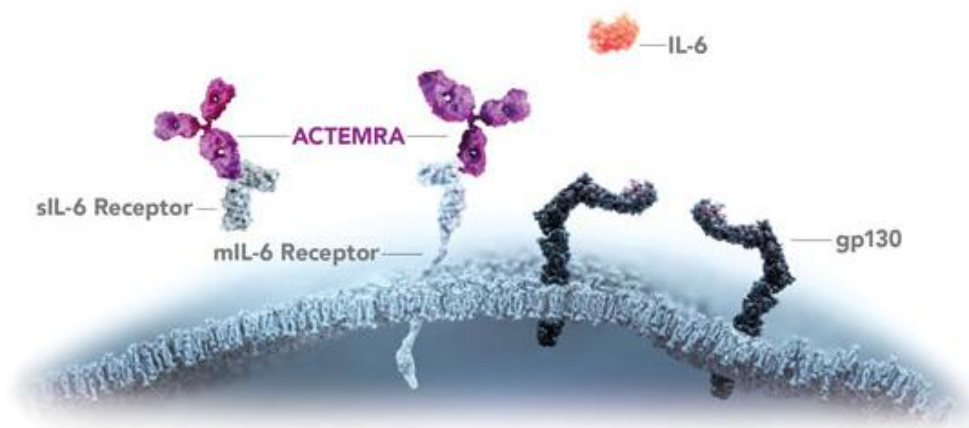


Figure 81. Le tocilizumab, commercialisé sous le nom Actemra® est un anticorps anti-IL-6R (avec sIL-6 Receptor : la forme soluble de l'IL-6R et mIL-6 Receptor : la forme transmembranaire de l'IL-6R) ⁵²⁷

2.3 Protocole du criblage

Afin de découvrir des inhibiteurs « conventionnels » de l'IL-6, nous avons utilisé un protocole combinant un criblage virtuel pour sélectionner les 1500 composés les plus prometteurs, suivi d'un criblage expérimental de ces 1500 molécules (Figure 82).

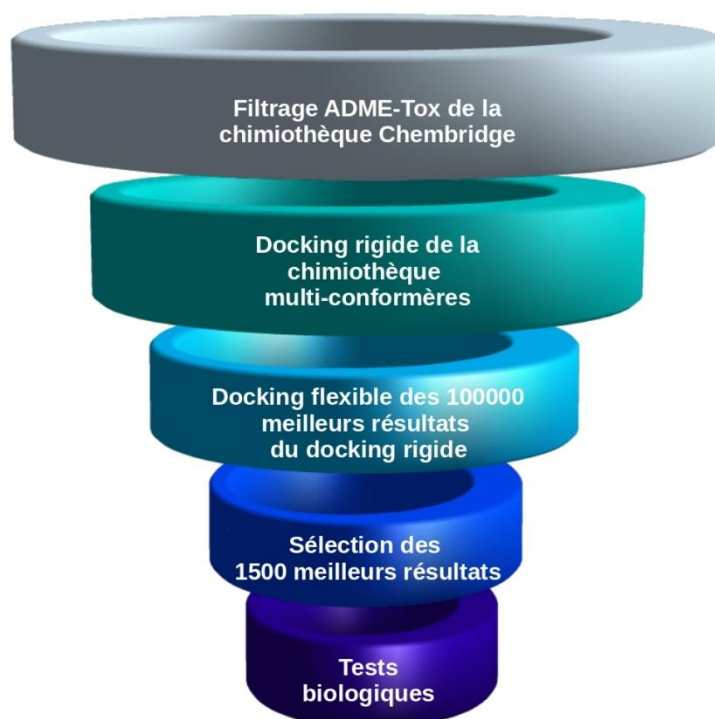


Figure 82. Représentation schématique du protocole mis en place au laboratoire GBA pour la découverte d'inhibiteurs de l'IL-6

2.3.1 Chimiothèque de criblage

La chimiothèque utilisée pour réaliser le criblage est la même que celle utilisée lors de la recherche d'inhibiteurs du TNF α . Il s'agit de la chimiothèque ChemBridge, filtrée à l'aide du logiciel FAF-drugs2⁵²⁸ selon des critères ADME-tox, permettant de ne conserver que 524891 composés sur les 900000 de départ. De multiples (jusqu'à 50) conformères ont été générés pour chaque composé de la chimiothèque.

2.3.2 Sélection de la structure et identification du site actif

Nous avons utilisé la structure du trimère IL-6/IL-6R/gp130 disponible dans la PDB avec le code 1P9M (Figure 83).

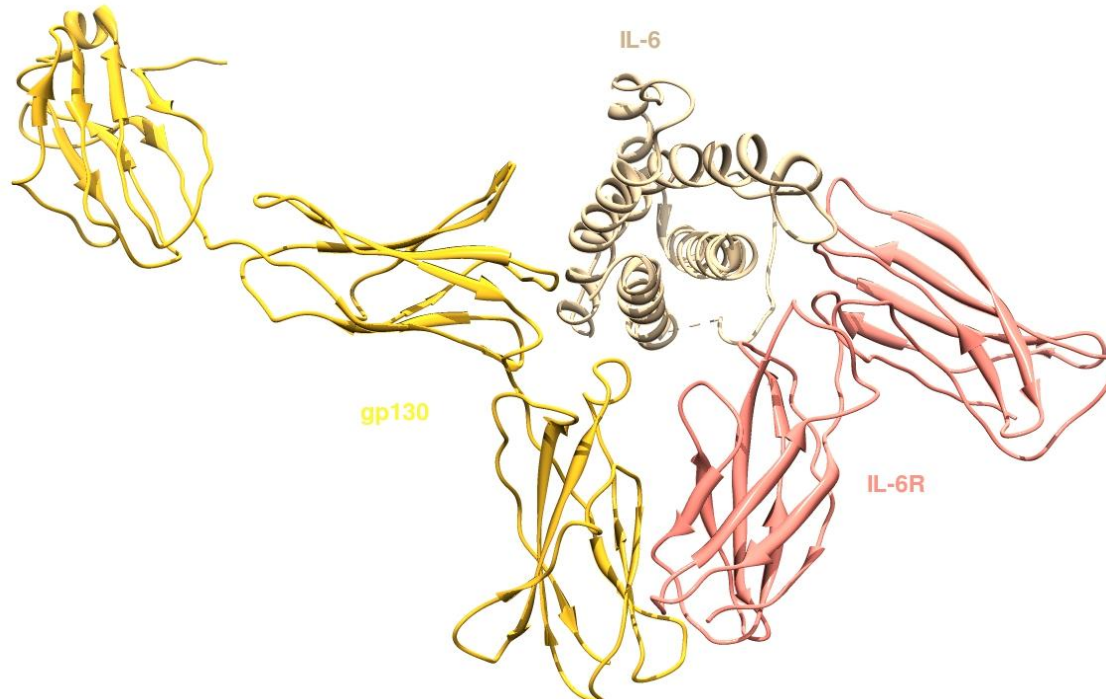


Figure 83. Structure du complexe IL-6 (beige) / IL-6R (rose) / gp130 (jaune) décrits dans la PDB sous le code 1P9M par Boulanger et ses collègues⁵²²

Après inspection visuelle de cette structure, nous avons défini une poche à l'interface entre l'IL-6 et la gp130 comme site de liaison potentiel d'inhibiteurs de l'IL-6 (Figure 84).

La structure a été ensuite préparée pour le docking, en éliminant la chaîne correspondant à l'IL-6R, et en utilisant l'outil DockPrep du logiciel Chimera⁴⁸⁵ qui supprime les molécules d'eau co-cristallisées, répare les chaînes latérales tronquées, protonne la protéine et assigne les charges partielles à l'aide du champ de force AMBER.

2.3.3 Réalisation du criblage virtuel

A partir de la chimiothèque et de la structure ainsi préparées, le criblage virtuel a été réalisé en deux étapes, à l'aide du logiciel de docking Surflex-Dock ³¹⁰ version 2.5. Après avoir défini le site de liaison en éditant manuellement le protomol (Figure 84), la première étape a consisté à procéder à un premier criblage rigide de la chimiothèque (commande **+rigid** de Surflex-Dock) en utilisant les options de minimisation avant le docking des ligands (**+premin**), et après le docking de tous les atomes du site de liaison (**+remin**).

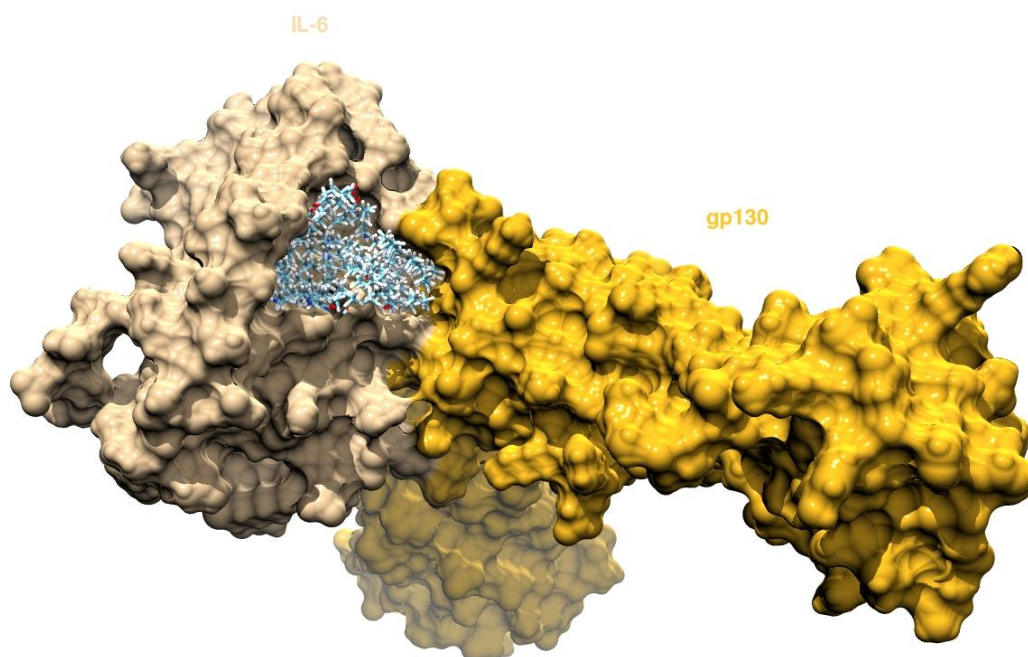


Figure 84. Structure du complexe IL-6/gp130 utilisée pour le docking avec le protocole généré avec le logiciel Surflex-Dock et édité manuellement pour couvrir le site de liaison choisi

Les résultats obtenus ont ensuite été analysés pour ne conserver que le meilleur conformère pour chaque composé de la chimiothèque, c'est-à-dire celui auquel la fonction de score du logiciel Surflex-Dock a attribué le meilleur score et triés pour sélectionner les 100000 composés associés aux scores les plus élevés (entre 8,17 et 4,03).

A partir de ces 100000 composés, un nouveau criblage virtuel a été réalisé, en conservant la même structure et le même structure de départ mais, cette fois, les ligands n'étaient plus considérés comme rigides mais comme flexibles. Les résultats ont ensuite été triés pour ne conserver que les 1500 molécules les plus prometteuses (dont le score associé était compris entre 10,81 et 7,84) dont le potentiel d'inhibiteurs de l'IL-6 a ensuite été testé expérimentalement.

2.3.4 Tests biologiques

Les tests biologiques des 1500 molécules identifiées par criblage virtuel ont été réalisés par l'équipe de pharmacologie du laboratoire GBA, et en particulier par Julie Perrier, doctorante au sein de ce laboratoire.

Trois types de tests biologiques ont été utilisés pour étudier l'activité des composés précédemment sélectionnés :

- un criblage expérimental par un test sur cellule à dose unique ensuite confirmé à plusieurs doses ;
- un essai de spécificité sur cellule des produits confirmés ;
- un test biochimique de liaison IL-6/IL6R.

2.3.4.1 Criblage expérimental par test cellulaire HEK-BLUE™ IL-6

Pour mener à bien le criblage expérimental, la lignée cellulaire HEK-Blue™ IL-6 commercialisée par la société InvivoGen⁵²⁹ a été utilisée.

Les cellules HEK-Blue™ IL-6 permettent de suivre l'activation de la voie JAK-STAT par l'IL-6 (Figure 85).

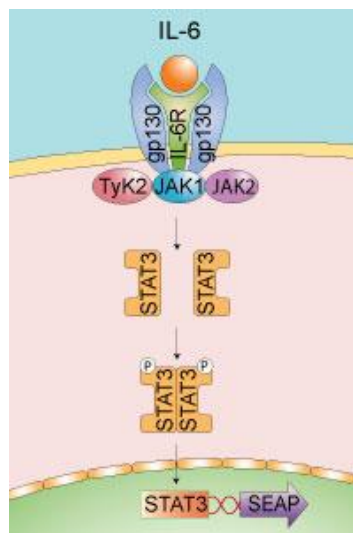


Figure 85. La formation du complexe actif IL-6/IL-6R/gp130 permet l'activation des protéines de la famille Janus : Janus Kinase 1 (JAK1) et 2 (JAK2) et Tyrosine kinase 2 (Tyk2). Ces kinases ainsi activées induisent la phosphorylation, la dimérisation et la translocation dans le noyau du facteur de transcription STAT3 (Signal transducer and activator of transcription 3) qui se lie à des éléments activateurs des gènes inducibles par l'IL-6. Dans les cellules HEK-Blue™ IL-6, l'activation de la voie JAK/STAT3 provoque la sécrétion de SEAP (Secreted Embryonic Alkaline Phosphatase)⁵²⁹

L'activation de cette voie induit la production de la protéine SEAP (Secreted Embryonic Alkaline Phosphatase), facilement détectable avec un test colorimétrique enzymatique QUANTI-Blue™, également fourni par la société InvivoGen⁵³⁰. Il est alors possible de calculer le pourcentage de neutralisation (Équation 36).

$$\%neutralisation = 100 - \frac{DO_{produit} - DO_{cellule}}{DO_{IL6} - DO_{cellule}}$$

Équation 36. Calcul du pourcentage de neutralisation à l'aide des densités optiques du test QUANTI-Blue™ des puits contenant les produits à tester et des différents puits contrôles ($DO_{produit}$: Densité Optique du puit contenant le produit à tester, des cellules et de l'IL-6 ; $DO_{cellule}$: Densité Optique du puit contenant des cellules uniquement ; DO_{IL6} : Densité Optique du puit contenant des cellules et de l'IL-6).

Les 1382 produits disponibles commercialement, sur les 1500 identifiés par criblage virtuel, ont tout d'abord été testés à 25 µg/mL pour déterminer leur capacité à neutraliser l'activité de l'IL-6 sur la lignée cellulaire HEK-Blue™ IL-6. Pour les composés les plus actifs, les résultats obtenus ont été confirmés à plusieurs doses (12,5 ; 25 ; 50 µg/mL).

Un test XTT (2,3-Bis(2-Methoxy-4-Nitro-5-Sulphophenyl)-2H-Tetrazolium-5-carboxanilide) de survie cellulaire⁵³¹ a aussi été utilisé. Ce test permet d'évaluer la cytotoxicité des composés à tester. Le XTT tetrazolium est un sel jaune clivé en XTT formazan, un produit orange soluble, par le système succinate deshydrogenase des mitochondries (Figure 86). Cette activité n'est présente que dans les cellules vivantes dont les membranes cellulaires et mitochondriales sont intactes.

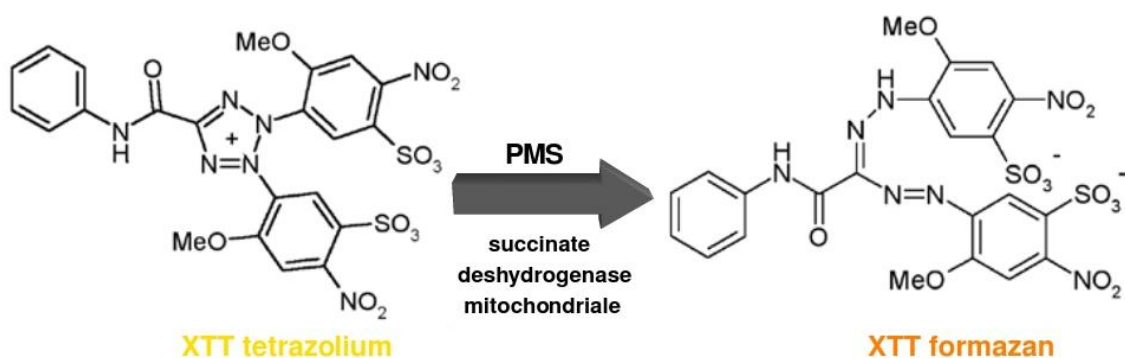


Figure 86. Clivage du XTT tetrazolium en XTT formazan par l'activité succinate deshydrogenase mitochondriale en présence de phenazine methosulfate (PMS).

Les produits perturbant les membranes et détruisant les chaînes respiratoires mitochondriales inhibent la succinate deshydrogenase et donc la formation de la réaction colorée orange. Le potentiel cytotoxique des produits peut donc être évalué par le pourcentage de survie, mesuré à l'aide des valeurs d'absorbance spectrophotométrique (Équation 37).

$$\%survie = \frac{DO_{produit} - DO_{blanc}}{DO_{IL6} - DO_{blanc}}$$

Équation 37. Calcul du pourcentage de neutralisation à l'aide des densités optiques du test XTT des puits contenant les produits à tester et des différents puits contrôles ($DO_{produit}$: Densité Optique du puit contenant le produit à tester, des cellules et de l'IL-6; DO_{blanc} : Densité Optique du puit contenant du milieu de culture uniquement, sans produit ni cellule ni IL-6; DO_{IL6} : Densité Optique du puit contenant des cellules et de l'IL-6).

A partir du pourcentage de neutralisation et du pourcentage de survie, le pourcentage de neutralisation corrigé peut être calculé (Équation 38).

$$\%neutralisation\ corrigé = 100 - \frac{DO_{produit} * (100/\%survie) - DO_{cellule}}{DO_{IL6} - DO_{cellule}}$$

Équation 38. Calcul du pourcentage de neutralisation à l'aide des densités optiques du test QUANTI-BLUE™ des puits contenant les produits à tester et des différents puits contrôles et du pourcentage de survie ($DO_{produit}$: Densité Optique du puit contenant le produit à tester, des cellules et de l'IL-6; $DO_{cellule}$: Densité Optique du puit contenant des cellules uniquement; DO_{IL6} : Densité Optique du puit contenant des cellules et de l'IL-6).

2.3.4.2 Essai de spécificité des produits confirmés

La spécificité pour l'IL6 des produits confirmés à l'encontre du TNF α , de l'IL-1 et de l'IL-4 a été évaluée à 25 μ g/mL. Pour cela les lignées cellulaires HEK-Blue™ TNF α /IL-1 β et HEK-Blue™ IL-4, de fonctionnement similaire au test HEK-Blue™ IL-6, ont été utilisées (Figure 87).

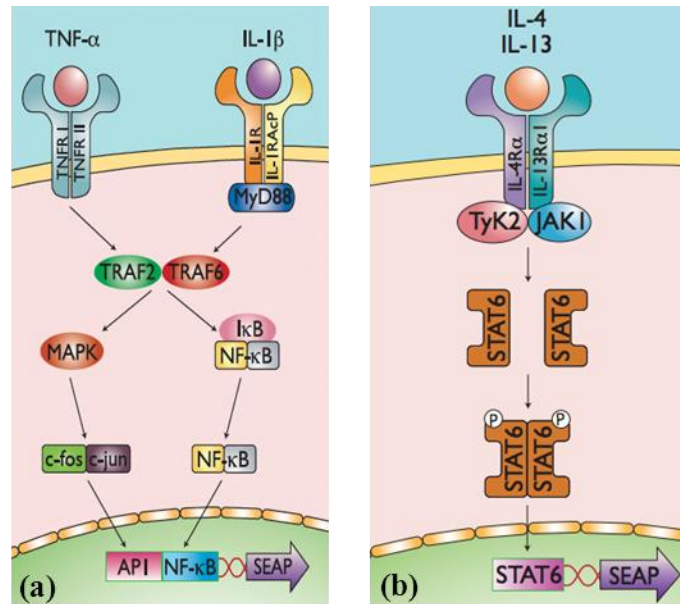


Figure 87. Les tests *HEK-BlueTM TNFα/IL-1β* (a) et *HEK-BlueTM IL-4* (b) reposent sur l'activation de la production de la protéine SEAP ensuite détectée par le test colorimétrique *QUANTI-BlueTM*. (a) Les cellules *HEK-BlueTM TNFα/IL-1β* permettent de mesurer le potentiel du TNFα et de l'IL-1β à se fixer sur leurs récepteurs et plus particulièrement à induire l'activation de la voie NF-κB⁵³². (b) Les cellules *HEK-BlueTM IL-4* mettent à profit la voie STAT6 pour évaluer le potentiel de l'IL-4 à se fixer sur son récepteur⁵³³.

2.3.4.3 Test de liaison IL-6/IL-R

Les produits confirmés ont aussi été soumis à un test de liaison IL-6/IL-6R. Ce test permet de s'assurer que l'action observée des produits est bien imputable à son action sur l'IL-6. Le test de liaison utilisé est un test biochimique modélisant le complexe IL-6/IL-6R/gp130 (Figure 88).

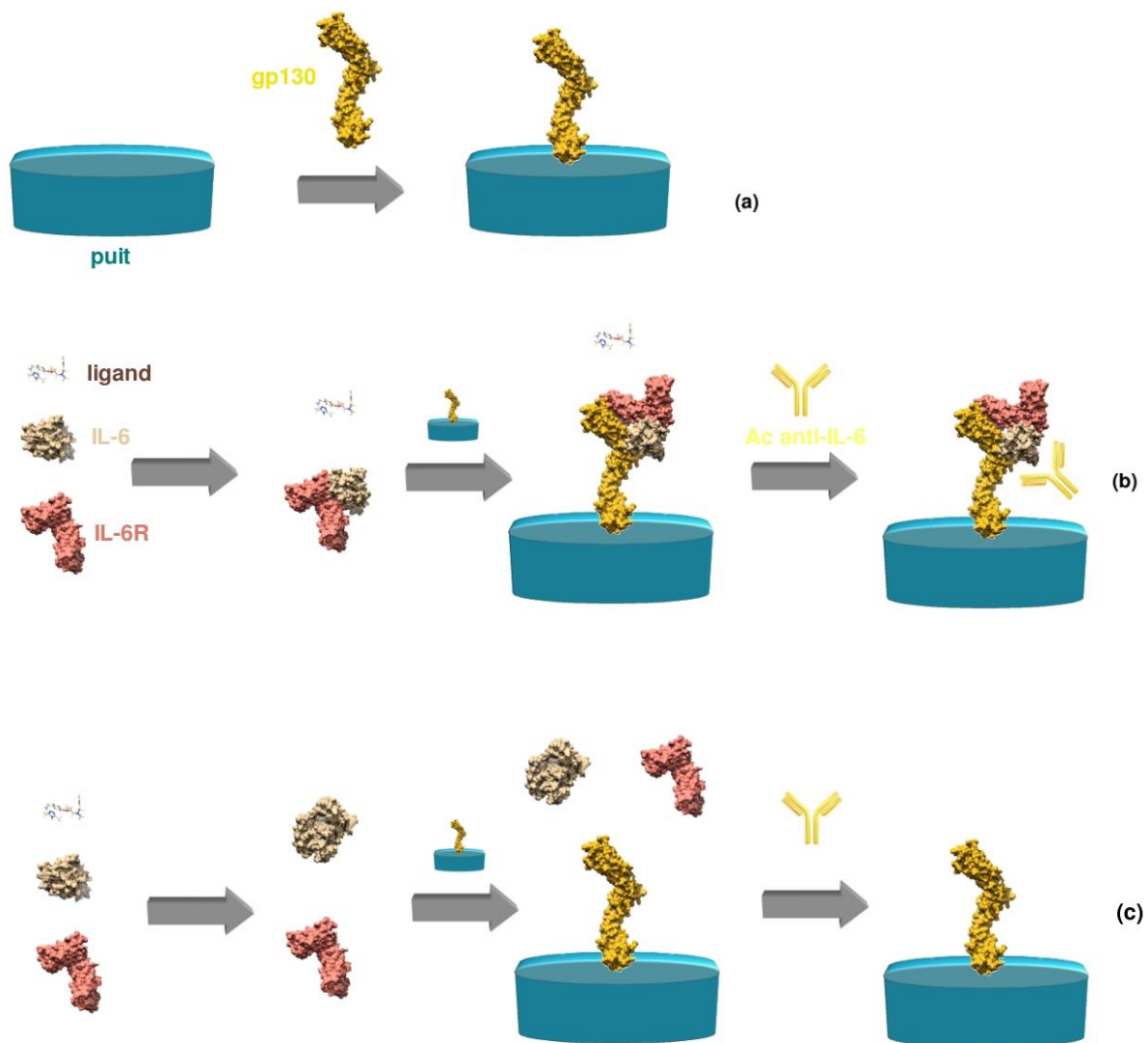


Figure 88. Le test de liaison IL-6/IL-6R se déroule en plusieurs étapes successives séparées par des phases de lavage des puits. (a) Fixation de la gp130 au fond des puits, et saturation des puits avec de l'albumine de sérum bovin (BSA). (b) Incubation de l'IL-6 avec son récepteur l'IL-6R et avec le ligand à tester. Le ligand ne se fixe pas à l'IL-6, celle-ci est donc libre d'interagir avec son récepteur IL-6R pour ensuite former un trimère avec la gp130 fixée dans le puits. Après lavage, au cours duquel le ligand qui ne s'est pas fixé est éliminé, un anticorps anti-IL6 associé à la biotine est ajouté. Celui-ci peut se lier à l'IL-6 du complexe fixé dans le puits et n'est pas éliminé par lavage. La révélation de la présence des anticorps anti-IL-6 (et donc de la présence de l'IL-6) se fait avec de l'avidine-HRP (Horseradish Peroxidase) qui va se coller à la biotine sur l'anticorps et va générer un produit coloré, en réaction avec son substrat le TMB (3,3',5,5'-tétraméthylbenzidine). La réaction est stoppée avec du H_2SO_4 et l'absorbance est mesurée par spectrophotométrie à 450 nm. (c) Incubation

de l'IL-6 avec son récepteur l'IL-6R et avec le ligand à tester. Le ligand se fixe à l'IL-6 qui ne peut donc pas interagir avec son récepteur IL-6R. L'IL-6 et l'IL-6R ne peuvent donc pas se fixer dans les puits et sont éliminés par lavage des puits. Lorsque l'anticorps anti-IL6 est ajouté, il ne trouve pas d'IL-6 auquel se lier et est donc éliminé par lavage des puits. Aucune réaction colorée n'est alors observée.

2.3.5 Résultats préliminaires

Sur les 1382 produits testés, 353 possédaient une valeur de survie supérieure à 70% et un pourcentage de neutralisation supérieur à 20% dans le test cellulaire HEK-BlueTM IL-6 (Tableau 19). Cette neutralisation a été reconfirmée pour 208 d'entre eux. Parmi ces 208 produits, 18 présentaient des pourcentages de neutralisation inférieurs à 10 % dans les tests de spécificité HEK-BlueTM TNF α /IL-1 β et HEK-BlueTM IL-4. Lors des tests de binding, 7 produits ont montré des neutralisations supérieures à 15%.

Les résultats obtenus lors de ces tests cellulaires devront maintenant être confirmés sur un modèle animal.

D'autre part, nous prévoyons aussi de rechercher et de tester expérimentalement des analogues des composés les plus prometteurs.

Mol.	h-IL6 neutra.		h-IL6 neutra. RECONF		h-TNFa neutra.		h-IL1b neutra.		h-IL4 neutra.		Liaison IL-6
	Neutra .	Neutra. COR.	Neutra.	Neutra. COR.	Neutra .	Neutra .COR.	Neutra .	Neutra .COR.	Neutra .	Neutra .COR.	Neutra.
1	107	106	103	104	97	98	95	97	97	98	0,4
2	105	104	101	102	87	89	78	81	78	83	-12,6
3	104	101	47	50	9	5	11	4	18	9	-1,6
4	102	99	86	85	47	38	37	29	31	28	4,2
5	89	95	77	80	9	11	44	41	38	41	3,3
6	97	93	64	66	33	39	24	26	31	40	4,8
7	97	92	49	51	7	13	19	24	10	23	2,7
8	95	90	39	39	25	18	25	13	41	40	3,9
9	93	88	66	69	67	65	49	49	49	46	2,2
10	91	88	81	76	28	-1	25	3	32	28	7,3
11	91	84	76	70	65	53	30	1	55	49	-7,7
12	87	84	32	39	50	50	24	24	27	36	4,1
13	89	84	64	58	93	59	83	58	86	17	11,9
14	90	84	93	88	-2	-42	14	-11	23	-1	1,0
15	88	83	33	19	27	24	24	20	53	47	18,2
16	90	82	34	27	NT	NT	NT	NT	NT	NT	NT
17	83	78	16	20	-5	0	-1	-8	14	17	3,0
18	88	78	33	37	85	80	58	45	51	44	3,7
19	86	78	56	42	36	29	12	9	20	10	-0,4
20	86	78	38	36	44	21	35	24	21	25	4,3
21	85	78	52	44	11	9	27	28	16	18	3,4
22	83	77	68	71	22	29	46	49	41	44	-3,9
23	91	77	49	48	43	44	20	20	32	29	1,8
24	83	74	79	67	66	53	56	42	70	58	24,6
25	81	73	15	9	61	54	50	42	32	30	-1,4
26	76	72	35	38	-8	-12	9	8	13	14	1,0
27	76	71	75	64	19	6	32	28	9	18	0,0
28	80	70	20	19	14	-2	21	16	23	21	-0,3
29	74	68	36	35	7	7	12	18	19	18	29,2
30	80	68	10	-7	50	38	21	9	31	23	5,6
31	75	66	86	77	60	52	31	7	47	34	8,1
32	74	66	24	33	2	0	25	25	20	25	-1,5
33	79	66	12	1	15	0	8	-12	33	22	4,4
34	77	65	97	91	NT	NT	NT	NT	NT	NT	NT
35	75	65	17	9	NT	NT	NT	NT	NT	NT	NT
36	75	64	40	35	-25	-21	32	25	28	25	2,9
37	67	62	50	47	51	48	45	36	45	35	5,4
38	71	62	49	44	12	4	37	26	21	23	12,3

Mol.	h-IL6 neutra.		h-IL6 neutra. RECONF		h-TNFa neutra.		h-IL1b neutra.		h-IL4 neutra.		Liaison IL-6
	Neutra .	Neutra. COR.	Neutra.	Neutra. COR.	Neutra .	Neutra .COR.	Neutra .	Neutra .COR.	Neutra .	Neutra .COR.	Neutra.
39	76	61	61	52	39	26	40	30	24	11	1,4
40	64	61	51	53	6	10	0	10	19	27	1,4
41	71	61	63	49	84	82	34	20	23	14	7,9
42	71	61	44	37	6	-2	-10	-9	22	12	-1,7
43	63	60	36	43	-33	-17	5	16	3	15	6,6
44	68	59	37	33	7	4	6	4	-1	-8	-12,1
45	66	58	79	73	75	67	43	23	61	57	-4,4
46	67	58	47	42	71	66	42	35	27	28	2,2
47	64	57	45	43	21	22	40	40	21	21	-0,4
48	64	56	18	15	6	1	15	8	-7	-1	2,1
49	61	54	28	24	8	6	23	17	33	31	0,8
50	60	54	27	16	9	13	0	12	13	39	4,3
51	71	54	54	43	22	9	15	-11	45	40	8,6
52	66	53	30	31	12	6	-6	-2	22	18	0,1
53	63	53	40	41	33	-1	29	27	-7	-1	-1,3
54	64	53	52	63	-5	6	22	31	12	25	-1,5
55	68	53	41	34	14	12	26	17	47	40	-0,3
56	62	52	38	34	-2	-10	18	10	13	10	-2,7
57	69	52	42	10	NT	NT	NT	NT	NT	NT	NT
58	63	52	94	87	57	53	32	21	24	13	4,2
59	52	52	30	34	-1	-17	15	12	15	28	4,4
60	57	51	15	20	-1	-6	8	0	18	24	7,9
61	66	51	79	56	NT	NT	NT	NT	NT	NT	NT
62	50	50	67	69	65	67	46	49	40	45	4,7
63	54	50	10	21	11	6	15	12	18	26	2,4
64	59	50	73	66	11	-3	35	19	27	14	-1,8
65	57	50	69	59	54	54	19	12	28	25	3,7
66	65	49	26	6	NT	NT	NT	NT	NT	NT	NT
67	65	47	68	51	6	-4	-14	-24	14	3	10,5
68	46	46	2	-2	NT	NT	NT	NT	NT	NT	NT
69	61	45	41	33	41	32	19	11	30	17	2,9
70	53	45	53	47	17	11	3	6	15	15	1,6
71	59	45	1	3	7	-2	14	8	-3	-20	8,4
72	59	45	40	33	20	22	17	12	16	4	4,1
73	46	44	43	37	15	8	17	12	27	21	6,1
74	46	44	24	27	-18	-12	-9	0	-16	-9	0,4
75	56	44	35	30	NT	NT	NT	NT	NT	NT	NT
76	44	44	12	18	-10	-14	-3	-2	22	26	5,2

Mol.	h-IL6 neutra.		h-IL6 neutra. RECONF		h-TNFa neutra.		h-IL1b neutra.		h-IL4 neutra.		Liaison IL-6
	Neutra .	Neutra. COR.	Neutra. .	Neutra. COR.	Neutra .	Neutra .COR.	Neutra .	Neutra .COR.	Neutra .	Neutra .COR.	Neutra.
77	59	44	72	53	NT	NT	NT	NT	NT	NT	NT
78	58	43	12	-6	-9	-26	5	4	3	-7	9,9
79	57	43	40	34	21	-45	-12	-21	4	-55	3,9
80	38	43	-10	-14	NT	NT	NT	NT	NT	NT	NT
81	53	43	-2	-20	NT	NT	NT	NT	NT	NT	NT
82	56	42	39	28	NT	NT	NT	NT	NT	NT	NT
83	61	42	44	35	29	10	5	-10	18	11	1,7
84	51	42	22	16	NT	NT	NT	NT	NT	NT	NT
85	54	42	30	44	15	-235	35	35	28	21	0,6
86	48	41	13	11	0	-35	39	34	22	24	-2,9
87	41	41	24	-14	NT	NT	NT	NT	NT	NT	NT
88	41	41	40	51	7	24	18	30	4	27	-3,6
89	46	41	67	64	35	23	12	5	26	21	0,8
90	41	40	39	34	-11	-27	-1	-7	25	4	2,4
91	29	40	-1	3	NT	NT	NT	NT	NT	NT	NT
92	29	40	-5	-4	NT	NT	NT	NT	NT	NT	NT
93	24	40	14	21	NT	NT	NT	NT	NT	NT	NT
94	52	39	31	24	6	-66	22	20	-15	-37	4,6
95	52	39	2	-19	NT	NT	NT	NT	NT	NT	NT
96	57	39	60	44	28	6	-3	-22	32	19	-0,5
97	51	38	21	18	NT	NT	NT	NT	NT	NT	NT
98	34	37	24	26	NT	NT	NT	NT	NT	NT	NT
99	36	37	32	29	22	21	4	10	20	22	-1,6
100	53	37	45	45	25	16	24	16	26	28	2,2
101	50	37	29	27	13	3	20	12	16	13	10,5
102	47	36	79	56	NT	NT	NT	NT	NT	NT	NT
103	53	36	12	1	NT	NT	NT	NT	NT	NT	NT
104	31	36	34	32	13	-6	26	19	20	23	2,2
105	40	36	44	44	12	8	0	-4	4	5	1,6
106	44	35	43	35	-8	-31	13	-1	23	19	7,3
107	26	35	14	33	13	-4	4	25	9	-15	2,8
108	33	35	6	9	NT	NT	NT	NT	NT	NT	NT
109	40	35	52	54	15	13	16	23	17	19	-0,4
110	46	34	59	55	51	44	29	19	20	16	1,5
111	53	34	26	35	30	28	5	6	23	21	-0,9
112	45	34	63	58	52	34	19	-14	32	30	7,9
113	50	34	13	18	NT	NT	NT	NT	NT	NT	NT
114	26	34	10	12	NT	NT	NT	NT	NT	NT	NT
115	40	34	12	-1	NT	NT	NT	NT	NT	NT	NT

Mol.	h-IL6 neutra.		h-IL6 neutra. RECONF		h-TNFa neutra.		h-IL1b neutra.		h-IL4 neutra.		Liaison IL-6
	Neutra .	Neutra. COR.	Neutra.	Neutra. COR.	Neutra .	Neutra .COR.	Neutra .	Neutra .COR.	Neutra .	Neutra .COR.	Neutra.
116	27	33	4	14	NT	NT	NT	NT	NT	NT	NT
117	44	33	25	30	12	5	-3	1	13	8	3,2
118	43	33	23	16	-5	-21	7	-4	13	4	5,2
119	55	33	38	21	NT	NT	NT	NT	NT	NT	NT
120	43	33	32	27	11	0	25	14	24	9	1,9
121	40	32	21	16	9	4	-3	-1	3	-2	7,9
122	24	32	-2	1	NT	NT	NT	NT	NT	NT	NT
123	52	32	31	24	18	-12	26	10	2	-8	7,1
124	38	32	68	55	-5	14	-11	-15	3	29	-2,5
125	43	32	8	8	NT	NT	NT	NT	NT	NT	NT
126	45	31	85	74	-15	-10	-18	-15	-2	-31	-2,4
127	38	31	41	46	19	14	9	7	21	35	1,9
128	32	31	12	8	NT	NT	NT	NT	NT	NT	NT
129	43	31	52	48	55	47	34	21	16	10	0,8
130	48	31	63	45	89	88	17	5	33	11	1,0
131	31	31	18	10	NT	NT	NT	NT	NT	NT	NT
132	46	31	30	30	-17	-18	-1	-6	24	22	1,6
133	29	31	1	0	NT	NT	NT	NT	NT	NT	NT
134	32	31	25	19	NT	NT	NT	NT	NT	NT	NT
135	32	31	15	20	3	6	-8	-9	-3	4	-1,4
136	35	31	10	-12	NT	NT	NT	NT	NT	NT	NT
137	38	30	16	-13	NT	NT	NT	NT	NT	NT	NT
138	36	30	35	25	11	-46	15	-2	26	-1	2,6
139	33	30	21	24	2	6	20	20	-14	-4	8,0
140	30	30	16	6	NT	NT	NT	NT	NT	NT	NT
141	43	30	-1	-7	NT	NT	NT	NT	NT	NT	NT
142	40	30	39	39	26	25	29	29	32	37	10,0
143	31	30	-6	-20	NT	NT	NT	NT	NT	NT	NT
144	49	30	37	24	34	15	20	-2	29	19	4,6
145	34	30	18	15	-22	-36	6	1	20	21	17,5
146	29	29	12	19	NT	NT	NT	NT	NT	NT	NT
147	25	29	35	40	46	47	-9	-8	13	23	4,2
148	46	29	48	47	11	7	5	-3	11	0	0,8
149	38	29	30	34	0	1	14	15	27	30	6,4
150	25	29	19	20	10	17	-25	-8	-2	-2	-1,4
151	50	29	42	47	21	28	21	21	18	26	-2,5
152	43	29	44	40	13	3	15	5	4	1	0,4
153	26	29	4	-6	NT	NT	NT	NT	NT	NT	NT

Mol.	h-IL6 neutra.		h-IL6 neutra. RECONF		h-TNFa neutra.		h-IL1b neutra.		h-IL4 neutra.		Liaison IL-6
	Neutra .	Neutra. COR.	Neutra.	Neutra. COR.	Neutra .	Neutra .COR.	Neutra .	Neutra .COR.	Neutra .	Neutra .COR.	Neutra.
154	39	28	42	32	3	-18	15	-11	42	35	6,8
155	33	28	5	15	-11	-3	32	40	26	36	8,3
156	41	28	57	43	-6	-20	-7	-19	0	20	-1,5
157	26	28	18	17	NT	NT	NT	NT	NT	NT	NT
158	41	28	41	40	10	-1	26	18	20	16	2,6
159	31	28	53	41	93	86	52	30	38	6	4,5
160	33	28	33	5	NT	NT	NT	NT	NT	NT	NT
161	25	27	15	25	-25	-20	16	20	16	26	1,5
162	38	27	73	62	5	2	8	-6	23	17	-0,9
163	33	27	2	-3	NT	NT	NT	NT	NT	NT	NT
164	33	27	12	4	NT	NT	NT	NT	NT	NT	NT
165	46	27	25	8	NT	NT	NT	NT	NT	NT	NT
166	29	27	26	24	5	11	8	9	33	38	4,2
167	35	27	5	-1	NT	NT	NT	NT	NT	NT	NT
168	29	26	33	36	9	1	1	5	12	19	6,3
169	27	26	15	11	NT	NT	NT	NT	NT	NT	NT
170	27	26	3	-8	NT	NT	NT	NT	NT	NT	NT
171	38	26	26	17	NT	NT	NT	NT	NT	NT	NT
172	44	26	10	-2	NT	NT	NT	NT	NT	NT	NT
173	31	26	7	6	NT	NT	NT	NT	NT	NT	NT
174	25	26	19	18	NT	NT	NT	NT	NT	NT	NT
175	23	26	13	11	NT	NT	NT	NT	NT	NT	NT
176	26	26	33	33	9	5	5	5	26	22	2,0
177	31	25	23	17	NT	NT	NT	NT	NT	NT	NT
178	24	25	1	-2	NT	NT	NT	NT	NT	NT	NT
179	41	25	21	13	49	34	24	6	9	-12	3,3
180	39	25	16	25	NT	NT	NT	NT	NT	NT	NT
181	23	25	17	8	NT	NT	NT	NT	NT	NT	NT
182	23	25	43	37	43	9	28	16	20	22	1,4
183	34	25	26	29	-12	-4	-5	0	36	41	0,1
184	29	25	19	18	NT	NT	NT	NT	NT	NT	NT
185	29	25	5	1	NT	NT	NT	NT	NT	NT	NT
186	30	25	6	-1	NT	NT	NT	NT	NT	NT	NT
187	30	25	13	-7	NT	NT	NT	NT	NT	NT	NT
188	31	25	9	-11	62	54	16	0	-1	-12	11,4
189	42	25	17	-22	NT	NT	NT	NT	NT	NT	NT
190	26	25	13	12	NT	NT	NT	NT	NT	NT	NT
191	31	25	10	4	NT	NT	NT	NT	NT	NT	NT
192	44	25	5	6	32	34	-20	-15	7	5	4,7
193	29	24	22	14	NT	NT	NT	NT	NT	NT	NT

Mol.	h-IL6 neutra.		h-IL6 neutra. RECONF		h-TNFa neutra.		h-IL1b neutra.		h-IL4 neutra.		Liaison IL-6
	Neutra .	Neutra. COR.	Neutra.	Neutra. COR.	Neutra .	Neutra .COR.	Neutra .	Neutra .COR.	Neutra .	Neutra .COR.	Neutra.
194	34	24	52	48	-6	-46	10	-6	17	27	3,8
195	30	24	17	5	10	12	-1	14	16	23	6,1
196	20	24	1	2	NT	NT	NT	NT	NT	NT	NT
197	24	24	5	0	NT	NT	NT	NT	NT	NT	NT
198	21	24	14	9	23	19	9	9	-4	-1	1,4
199	23	24	2	3	NT	NT	NT	NT	NT	NT	NT
200	21	24	-3	-9	NT	NT	NT	NT	NT	NT	NT
201	29	24	-19	-49	NT	NT	NT	NT	NT	NT	NT
202	23	24	6	8	NT	NT	NT	NT	NT	NT	NT
203	22	24	22	18	NT	NT	NT	NT	NT	NT	NT
204	44	23	46	40	27	17	19	-3	35	20	-1,4
205	27	23	6	1	NT	NT	NT	NT	NT	NT	NT
206	28	23	28	36	-5	-20	21	26	-8	1	4,9
207	28	23	53	49	16	15	17	11	18	8	5,0
208	27	23	25	17	NT	NT	NT	NT	NT	NT	NT
209	26	23	3	0	NT	NT	NT	NT	NT	NT	NT
210	23	23	19	21	-21	-26	19	22	28	33	17,8
211	48	23	42	28	NT	NT	NT	NT	NT	NT	NT
212	39	23	30	30	-15	15	-8	-1	-3	22	1,6
213	26	23	5	-6	NT	NT	NT	NT	NT	NT	NT
214	25	23	28	29	NT	NT	NT	NT	NT	NT	NT
215	28	23	-3	-12	NT	NT	NT	NT	NT	NT	NT
216	46	22	32	34	8	-4	21	11	22	14	-2,4
217	26	22	25	24	2	6	11	15	7	15	1,9
218	22	22	7	7	NT	NT	NT	NT	NT	NT	NT
219	30	22	12	3	NT	NT	NT	NT	NT	NT	NT
220	32	22	10	12	NT	NT	NT	NT	NT	NT	NT
221	31	22	32	30	-13	-3	-11	-12	4	17	11,5
222	27	22	-5	-3	NT	NT	NT	NT	NT	NT	NT
223	34	22	23	23	-12	-14	15	8	20	24	1,2
224	32	22	17	12	NT	NT	NT	NT	NT	NT	NT
225	29	22	24	27	31	33	4	9	1	7	0,4
226	23	21	1	-6	NT	NT	NT	NT	NT	NT	NT
227	36	21	10	11	NT	NT	NT	NT	NT	NT	NT
228	33	21	13	-5	NT	NT	NT	NT	NT	NT	NT
229	29	21	-16	-23	NT	NT	NT	NT	NT	NT	NT
230	26	21	18	5	NT	NT	NT	NT	NT	NT	NT
231	37	21	26	2	NT	NT	NT	NT	NT	NT	NT

Mol.	h-IL6 neutra.		h-IL6 neutra. RECONF		h-TNFa neutra.		h-IL1b neutra.		h-IL4 neutra.		Liaison IL-6
	Neutra .	Neutra. COR.	Neutra.	Neutra. COR.	Neutra .	Neutra .COR.	Neutra .	Neutra .COR.	Neutra .	Neutra .COR.	Neutra.
232	38	21	28	19	NT	NT	NT	NT	NT	NT	NT
233	28	21	25	24	NT	NT	NT	NT	NT	NT	NT
234	28	21	18	23	NT	NT	NT	NT	NT	NT	NT
235	24	21	77	66	28	-44	69	63	58	-19	-2,5
236	24	20	0	-1	NT	NT	NT	NT	NT	NT	NT
237	26	20	28	34	0	-59	6	8	7	17	3,8
238	25	20	35	29	48	44	8	1	21	20	3,3
239	29	20	23	22	NT	NT	NT	NT	NT	NT	NT
240	27	20	29	27	36	27	6	-11	26	30	1,6
241	25	20	10	10	NT	NT	NT	NT	NT	NT	NT
242	22	20	21	18	NT	NT	NT	NT	NT	NT	NT
243	41	20	21	20	-2	0	24	21	8	8	4,6
244	33	20	58	48	27	-26	17	-3	39	-31	14,9
245	26	20	31	27	-16	-7	-13	-9	-3	23	-3,7
246	23	20	13	13	-4	-8	0	-5	7	8	-4,9
247	33	20	17	1	NT	NT	NT	NT	NT	NT	NT
248	35	19	48	34	75	75	26	26	29	26	7,6
249	23	19	-3	-6	NT	NT	NT	NT	NT	NT	NT
250	23	19	-3	-8	NT	NT	NT	NT	NT	NT	NT
251	44	19	61	57	-3	-38	27	1	21	-1	0,1
252	25	19	12	14	15	7	14	14	14	15	6,7
253	24	19	-4	-8	NT	NT	NT	NT	NT	NT	NT
254	41	19	53	30	NT	NT	NT	NT	NT	NT	NT
255	29	19	8	11	NT	NT	NT	NT	NT	NT	NT
256	25	19	25	18	18	16	15	13	14	13	4,7
257	22	19	11	-45	NT	NT	NT	NT	NT	NT	NT
258	38	19	12	2	NT	NT	NT	NT	NT	NT	NT
259	29	19	43	38	-18	-21	-11	-2	4	14	0,1
260	23	19	0	-7	NT	NT	NT	NT	NT	NT	NT
261	21	18	25	23	NT	NT	NT	NT	NT	NT	NT
262	23	18	49	36	-26	-16	-21	-10	-16	-48	-2,0
263	24	18	0	-10	NT	NT	NT	NT	NT	NT	NT
264	22	18	35	38	-12	-14	0	-1	9	8	2,6
265	21	18	5	0	NT	NT	NT	NT	NT	NT	NT
266	43	18	22	21	15	-35	24	20	6	-36	1,7
267	25	18	13	9	NT	NT	NT	NT	NT	NT	NT
268	25	17	-4	-13	NT	NT	NT	NT	NT	NT	NT
269	28	17	11	6	NT	NT	NT	NT	NT	NT	NT
270	35	17	42	30	4	-8	32	33	24	27	10,1

Mol.	h-IL6 neutra.		h-IL6 neutra. RECONF		h-TNFa neutra.		h-IL1b neutra.		h-IL4 neutra.		Liaison IL-6
	Neutra .	Neutra. COR.	Neutra.	Neutra. COR.	Neutra .	Neutra .COR.	Neutra .	Neutra .COR.	Neutra .	Neutra .COR.	Neutra.
271	35	17	11	-26	NT	NT	NT	NT	NT	NT	NT
272	32	17	32	27	3	4	5	4	16	20	2,4
273	28	16	14	4	NT	NT	NT	NT	NT	NT	NT
274	35	16	56	53	60	60	38	32	28	22	3,0
275	26	16	60	60	8	4	18	15	12	11	2,9
276	25	16	4	-1	NT	NT	NT	NT	NT	NT	NT
277	23	16	49	45	-5	-13	9	-1	8	-2	-0,2
278	29	16	-9	-31	NT	NT	NT	NT	NT	NT	NT
279	20	16	8	9	-11	-20	-24	-23	3	4	2,9
280	23	15	14	8	38	40	7	11	6	24	5,0
281	25	15	13	10	NT	NT	NT	NT	NT	NT	NT
282	22	15	11	11	NT	NT	NT	NT	NT	NT	NT
283	23	15	38	31	-15	-240	0	-9	-2	-10	0,4
284	21	15	-1	7	17	12	7	5	27	34	10,6
285	21	14	33	39	18	23	21	28	12	17	3,5
286	23	14	16	16	NT	NT	NT	NT	NT	NT	NT
287	22	14	-3	-12	NT	NT	NT	NT	NT	NT	NT
288	22	14	9	7	NT	NT	NT	NT	NT	NT	NT
289	25	14	7	-1	NT	NT	NT	NT	NT	NT	NT
290	26	14	-8	-13	NT	NT	NT	NT	NT	NT	NT
291	26	14	52	44	84	79	27	6	22	11	2,2
292	21	13	10	4	NT	NT	NT	NT	NT	NT	NT
293	33	13	27	15	NT	NT	NT	NT	NT	NT	NT
294	27	13	25	21	4	6	22	17	-17	-11	4,3
295	23	13	6	17	NT	NT	NT	NT	NT	NT	NT
296	20	13	28	32	16	10	21	22	22	29	5,9
297	22	13	14	9	NT	NT	NT	NT	NT	NT	NT
298	28	12	36	26	NT	NT	NT	NT	NT	NT	NT
299	24	12	4	2	NT	NT	NT	NT	NT	NT	NT
300	25	12	17	11	19	10	30	30	23	27	0,0
301	31	12	21	21	NT	NT	NT	NT	NT	NT	NT
302	24	12	16	14	NT	NT	NT	NT	NT	NT	NT
303	28	12	4	-10	NT	NT	NT	NT	NT	NT	NT
304	26	12	22	7	NT	NT	NT	NT	NT	NT	NT
305	32	12	4	5	NT	NT	NT	NT	NT	NT	NT
306	20	12	-2	-1	NT	NT	NT	NT	NT	NT	NT
307	20	11	-4	3	NT	NT	NT	NT	NT	NT	NT
308	21	11	24	26	NT	NT	NT	NT	NT	NT	NT

Mol.	h-IL6 neutra.		h-IL6 neutra. RECONF		h-TNFa neutra.		h-IL1b neutra.		h-IL4 neutra.		Liaison IL-6
	Neutra .	Neutra. COR.	Neutra. .	Neutra. COR.	Neutra .	Neutra .COR.	Neutra .	Neutra .COR.	Neutra .	Neutra .COR.	Neutra.
309	23	11	6	5	NT	NT	NT	NT	NT	NT	NT
310	33	11	59	43	1	-25	-10	-12	1	-29	5,7
311	22	11	18	16	NT	NT	NT	NT	NT	NT	NT
312	24	11	-2	-11	NT	NT	NT	NT	NT	NT	NT
313	22	11	13	12	NT	NT	NT	NT	NT	NT	NT
314	24	11	4	0	NT	NT	NT	NT	NT	NT	NT
315	23	11	29	19	NT	NT	NT	NT	NT	NT	NT
316	31	10	-8	-54	NT	NT	NT	NT	NT	NT	NT
317	27	10	21	18	10	5	8	7	21	23	0,4
318	26	10	20	11	NT	NT	NT	NT	NT	NT	NT
319	21	10	26	28	-18	-20	-11	-1	-3	-38	14,9
320	25	10	24	16	12	15	10	14	13	12	0,5
321	38	9	13	-2	NT	NT	NT	NT	NT	NT	NT
322	20	9	22	16	NT	NT	NT	NT	NT	NT	NT
323	21	9	23	12	24	23	4	9	15	-9	11,7
324	23	9	12	4	NT	NT	NT	NT	NT	NT	NT
325	21	8	26	12	NT	NT	NT	NT	NT	NT	NT
326	20	8	6	1	-12	-23	28	21	30	31	12,8
327	22	7	41	28	-25	-32	6	8	5	13	3,0
328	25	7	31	14	NT	NT	NT	NT	NT	NT	NT
329	31	7	13	2	NT	NT	NT	NT	NT	NT	NT
330	25	7	13	5	NT	NT	NT	NT	NT	NT	NT
331	27	7	11	13	NT	NT	NT	NT	NT	NT	NT
332	21	6	22	16	NT	NT	NT	NT	NT	NT	NT
333	24	6	39	48	6	17	23	38	-26	-41	-0,4
334	20	6	21	22	NT	NT	NT	NT	NT	NT	NT
335	21	6	12	13	-5	-22	29	19	15	14	17,2
336	29	6	11	11	NT	NT	NT	NT	NT	NT	NT
337	24	6	20	14	12	10	14	7	37	37	0,7
338	28	5	19	-9	NT	NT	NT	NT	NT	NT	NT
339	24	5	35	31	-23	-14	-8	1	7	-7	8,0
340	30	5	26	-2	NT	NT	NT	NT	NT	NT	NT
341	23	5	30	21	-9	-39	8	-15	21	5	16,6
342	29	5	41	23	NT	NT	NT	NT	NT	NT	NT
343	35	5	54	31	-37	-44	4	-1	6	-2	-3,1
344	26	5	26	14	14	-16	32	1	36	22	6,4
345	25	5	19	18	NT	NT	NT	NT	NT	NT	NT
346	28	4	27	19	NT	NT	NT	NT	NT	NT	NT
347	23	-1	40	15	NT	NT	NT	NT	NT	NT	NT

Mol.	h-IL6 neutra.		h-IL6 neutra. RECONF		h-TNFa neutra.		h-IL1b neutra.		h-IL4 neutra.		Liaison IL-6
	Neutra .	Neutra. COR.	Neutra.	Neutra. COR.	Neutra .	Neutra .COR.	Neutra .	Neutra .COR.	Neutra .	Neutra .COR.	Neutra.
348	26	-2	23	-4	NT	NT	NT	NT	NT	NT	NT
349	22	-3	20	12	NT	NT	NT	NT	NT	NT	NT
350	24	-4	21	-1	NT	NT	NT	NT	NT	NT	NT
351	26	-6	26	14	NT	NT	NT	NT	NT	NT	NT
352	20	-7	25	16	NT	NT	NT	NT	NT	NT	NT
353	24	-8	19	20	NT	NT	NT	NT	NT	NT	NT

Tableau 19. Résultats préliminaires obtenus pour les 353 composés ayant montré un pourcentage de neutralisation de l'activité de l'IL-6 sur la lignée cellulaire HEK-Blue™ (h-IL6 neutra.) supérieur à 20 pourcent, lors de la reconfirmation différentes doses (h-IL neutra. RECONF), lors de l'étude de spécificité des produits pour l'IL-6 à l'encontre du TNFa (h-TNFa neutra.), de l'IL-1 (h-IL1b neutra.) et de l'IL-4 (h-IL4 neutra) et du test de liaison IL-6/IL-6R (liaison IL-6) (avec Neutra : pourcentage de neutralisation ; Neutra. COR. : pourcentage de neutralisation corrigé)

Troisième partie
Conclusion

Les méthodes de criblage virtuel sont communément employées dans les processus de recherche et développement de nouveaux médicaments, en complément de criblages expérimentaux à haut débit.

Cependant, malgré l'intérêt indubitable de ces méthodes, un manque de consensus et de lignes directrices est parfois à déplorer. C'est notamment le cas pour les méthodes de criblage virtuel basées sur la structure de la cible lors de l'étape de sélection de la ou des structures de référence. Nous avons donc décidé de rechercher des critères simples, basés sur les propriétés physico-chimiques du site de liaison, pour guider cette sélection. Le volume du site de liaison est apparu comme étant un critère de sélection des structures tout à fait intéressant, que ce soit pour les approches en structure unique ou en ensemble de structures. Pour les « petites » cibles, la structure dont le volume du site de liaison était le plus faible était la plus adaptée. A l'inverse, pour les plus « grandes » cibles, la structure la plus performante était celle de plus large volume de site de liaison. L'ouverture du site de liaison a aussi été proposée comme critère d'exclusion dans l'approche ensemble de structures, des structures les plus fermées pour les « grandes cibles » et des structures les plus ouvertes pour les « petites » cibles. Ces critères simples et peu coûteux en temps de calculs devraient permettre d'améliorer l'étape de sélection des structures de référence.

Pour pouvoir utiliser ces méthodes de criblage virtuel dans des études prospectives à la recherche de nouveaux médicaments, il est important d'évaluer préalablement leurs performances et ainsi de s'assurer de la fiabilité des résultats obtenus. Pour cela, différentes banques d'évaluation ont été développées au cours du temps, parmi lesquelles la DUD et la DUD_E considérées comme les références. Cependant, la qualité de ces banques peut encore être améliorée, c'est pourquoi nous avons proposé notre propre banque d'évaluation, la NRLiSt BDB. Cette base de données présente l'originalité de n'inclure que des ligands et des structures dont les données pharmacologiques ont été vérifiées manuellement et de prendre en compte le profil pharmacologique des ligands pour la construction des jeux de données. La NRLiSt BDB devrait donc devenir la base de données de référence pour l'évaluation des méthodes basées sur les ligands ou sur les structures des récepteurs nucléaires. Grâce à l'ensemble des données pharmacologiques incluses, la NRLiSt BDB peut aussi être exploitée pour étudier les mécanismes contrôlant la modulation et l'action de ses récepteurs ainsi que pour rechercher de nouveaux ligands des RNs.

Nous avons utilisé la NRLiSt BDB dans une première étude pour évaluer l'impact des partis pris durant la construction de cette banque sur les performances d'une méthode de criblage

virtuel basée sur les structures, Surflex-Dock. Nous avons ainsi pu vérifier que la NRLiSt BDB est une banque d'évaluation fiable et de haute qualité, utilisable pour l'évaluation des méthodes de criblage virtuel. En effet, nous avons confirmé que séparer les jeux de données selon le profil pharmacologique des ligands améliore réellement la qualité de l'évaluation. Le profil pharmacologique du ligand co-cristallisé a aussi été mis en évidence comme paramètre critique. Lorsque cela est possible, une structure avec un ligand co-cristallisé du même profil pharmacologique que les ligands étudiés devrait donc être utilisée pour les criblages. Les modalités de construction de la NRLiSt BDB devraient donc être reproduites pour le développement de nouvelles banques d'évaluation. Nous avons aussi démontré qu'il est nécessaire de remplacer les decoys putatifs actuellement utilisés par des decoys plus adaptés à l'évaluation des méthodes, comme par exemple des ligands decoys expérimentalement validés.

Grâce aux compétences acquises dans l'évaluation des méthodes, nous avons ensuite réalisé un criblage virtuel prospectif à la recherche de nouveaux composés inhibiteurs de l'IL-6, potentiellement utilisable dans le traitement de la polyarthrite rhumatoïde. Pour cela nous avons utilisé le logiciel Surflex-Dock, la chimiothèque ChemBridge préalablement filtrée à l'aide de paramètres ADME-Tox et la structure 1P9M du trimère IL-6/IL-6R/gp130. Après un premier criblage par docking ligand rigide, les 100000 meilleurs composés ont été sélectionnés pour un deuxième criblage, cette fois par docking ligand flexible. Les 1500 composés les plus prometteurs ont ensuite été testés expérimentalement. 208 produits ont donné des résultats intéressants dans un criblage expérimental du potentiel de neutralisation de l'IL-6. La spécificité pour l'IL-6 et envers le TNF α , l'IL-1 et l'IL-4 de ces produits a ensuite été évaluée. Finalement, un test de liaison à l'IL-6 a été mené dans lequel 7 produits ont montré des neutralisations supérieures à 15%. Des résultats préliminaires intéressants ont donc été obtenus. Les résultats *in vitro* devront être confirmés par des tests *in vivo*. D'autre part, nous allons aussi rechercher des analogues des composés les plus prometteurs par des méthodes de criblage virtuel basées sur les ligands.

Bibliographie

1. Sneader, W. *Drug Discovery, A History*, ed. J.W.S. Ltd. 2005. 472.
2. Schmitz, R. *Friedrich Wilhelm Serturner and the discovery of morphine*. *Pharm Hist*, 1985. **27**(2): p. 61-74.
3. Gerhardt, C. *Recherches sur les acides organiques anhydres*. *Ann. de Chim. et de Phys.*, 1853. **37**(3): p. 285-342.
4. Gerhardt, C. *Untersuchungen über die Wassefreien organischen Säuren*. *Ann. d. Chem. u. Pharm.*, 1853. **87**: p. 149-179.
5. Lafont, O. *Du saule à l'aspirine*. *Revue d'histoire de la pharmacie*, 2007. **94**(354): p. 209-216.
6. Nikiéma, J.B.; Djierro, K.; Simporé, J.; et al. *Stratégie d'utilisation des substances naturelles dans la prise en charge des personnes vivant avec le VIH: expérience du Burkina Faso*. *Ethnopharmacologia*, 2009. **43**: p. 47-51.
7. Ankri, J.; Pelicand, J. *Petite histoire du médicament*. *Actualité et dossier en santé publique*, 1999. **27**: p. 22-23.
8. Chanda, S.K.; Caldwell, J.S. *Fulfilling the promise: Drug discovery in the post-genomic era*. *Drug Discovery Today*, 2003. **8**(4): p. 168-174.
9. Paul, S.M.; Mytelka, D.S.; Dunwiddie, C.T.; et al. *How to improve R&D productivity: the pharmaceutical industry's grand challenge*. *Nat Rev Drug Discov*, 2010. **9**(3): p. 203-14.
10. Morgan, S.; Grootendorst, P.; Lexchin, J.; et al. *The cost of drug development: a systematic review*. *Health Policy*, 2011. **100**(1): p. 4-17.
11. Code de la Santé Publique. Available from: <http://www.legifrance.gouv.fr/affichCode.do?cidTexte=LEGITEXT000006072665>. [10/01/2014].
12. Aronson, J.K. *Rare diseases and orphan drugs*. *Br J Clin Pharmacol*, 2006. **61**(3): p. 243-5.
13. Corr, P.W., D. *The pathway from Idea to Regulatory Approval: Examples for Drug Development*, in *Conflict of Interest in Medical Research, Education and Practice*, E.a.P. Bernard Lo and Larykun J Field. Institute of Medicine (US) Committee on Conflict of Interest in Medical Research, Editor. 2009, National Academy Press (US).
14. Owens, J. *Target validation: Determining druggability*. *Nature Reviews Drug Discovery*, 2007. **6**: p. 187.
15. Hopkins, A.L.; Groom, C.R. *The druggable genome*. *Nat Rev Drug Discov*, 2002. **1**(9): p. 727-30.
16. Cheng, A.C.; Coleman, R.G.; Smyth, K.T.; et al. *Structure-based maximal affinity model predicts small-molecule druggability*. *Nat Biotechnol*, 2007. **25**(1): p. 71-5.
17. Hughes, J.P.; Rees, S.; Kalindjian, S.B.; et al. *Principles of early drug discovery*. *Br J Pharmacol*, 2011. **162**(6): p. 1239-49.
18. Yang, Y.; Adelstein, S.J.; Kassis, A.I. *Target discovery from data mining approaches*. *Drug Discov Today*, 2009. **14**(3-4): p. 147-54.
19. Limou, S.; Le Clerc, S.; Coulonges, C.; 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*, 2009. **199**(3): p. 419-26.
20. Le Clerc, S.; Limou, S.; Coulonges, C.; et al. *Genomewide association study of a rapid progression cohort identifies new susceptibility alleles for AIDS (ANRS Genomewide Association Study 03)*. *J Infect Dis*, 2009. **200**(8): p. 1194-201.
21. Zanders, E.D.; Bailey, D.S.; Dean, P.M. *Probes for chemical genomics by design*. *Drug Discov Today*, 2002. **7**(13): p. 711-8.

22. Bajorath, J. *Integration of virtual and high-throughput screening*. Nat Rev Drug Discov, 2002. **1**(11): p. 882-94.
23. Moitessier, N.; Englebienne, P.; Lee, D.; et al. *Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go*. Br J Pharmacol, 2008. **153 Suppl 1**: p. S7-26.
24. Tanrikulu, Y.; Kruger, B.; Proschak, E. *The holistic integration of virtual screening in drug discovery*. Drug Discov Today, 2013. **18**(7-8): p. 358-64.
25. Duffy, B.C.; Zhu, L.; Decornez, H.; et al. *Early phase drug discovery: cheminformatics and computational techniques in identifying lead series*. Bioorg Med Chem, 2012. **20**(18): p. 5324-42.
26. Guha, R.; Van Drie, J.H. *Structure--activity landscape index: identifying and quantifying activity cliffs*. J Chem Inf Model, 2008. **48**(3): p. 646-58.
27. Peltason, L.; Bajorath, J. *Systematic computational analysis of structure-activity relationships: concepts, challenges and recent advances*. Future Med Chem, 2009. **1**(3): p. 451-66.
28. Hubbard, R.E. *3D structure and the drug-discovery process*. Mol. BioSyst., 2005. **1**: p. 391-406.
29. Wang, D.; Bakhai, A. *Clinical Trials: A Practical Guide to Design, Analysis, and Reporting*. Remedica Medical Education and Publishing ed. 2006: Andrew Ward. 480.
30. loi n° 2004-806 du 9 août 2004 relative à la politique de santé publique. Available from: <http://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000000787078>. [15/01/2014].
31. Kola, I.; Landis, J. *Can the pharmaceutical industry reduce attrition rates?* Nat Rev Drug Discov, 2004. **3**(8): p. 711-5.
32. Oprea, T.I.; Matter, H. *Integrating virtual screening in lead discovery*. Curr Opin Chem Biol, 2004. **8**(4): p. 349-58.
33. Lavecchia, A.; Di Giovanni, C. *Virtual screening strategies in drug discovery: a critical review*. Curr Med Chem, 2013. **20**(23): p. 2839-60.
34. Perez-Pineiro, R.; Burgos, A.; Jones, D.C.; et al. *Development of a novel virtual screening cascade protocol to identify potential trypanothione reductase inhibitors*. J Med Chem, 2009. **52**(6): p. 1670-80.
35. Horvath, D. *A virtual screening approach applied to the search for trypanothione reductase inhibitors*. J Med Chem, 1997. **40**(15): p. 2412-23.
36. Bohacek, R.S.; McMartin, C.; Guida, W.C. *The art and practice of structure-based drug design: a molecular modeling perspective*. Med Res Rev, 1996. **16**(1): p. 3-50.
37. Barker, A.; Kettle, J.G.; Nowak, T.; et al. *Expanding medicinal chemistry space*. Drug Discov Today, 2013. **18**(5-6): p. 298-304.
38. Rognan, D. *Le criblage virtuel par docking moléculaire*, in *Chemogénomique, des petites molécules pour explorer le vivant*, E.R. Maréchal, S.; Lafanechère, L., Editor. 2007. p. 258.
39. Hann, M.M.; Oprea, T.I. *Pursuing the leadlikeness concept in pharmaceutical research*. Curr Opin Chem Biol, 2004. **8**(3): p. 255-63.
40. Lagarde, N. NRLiSt BDB. Available from: <http://www.nrlist.drugdesign.fr/>. 2014
41. Bolton, E. *PubChem: Integrated Platform of Small Molecules and Biological Activities*. Annual Reports in Computational Chemistry, 2008.
42. Gaulton, A.; Bellis, L.J.; Bento, A.P.; et al. *ChEMBL: a large-scale bioactivity database for drug discovery*. Nucleic Acids Res, 2012. **40**(Database issue): p. D1100-7.

43. Irwin, J.J.; Shoichet, B.K. *ZINC--a free database of commercially available compounds for virtual screening*. *J Chem Inf Model*, 2005. **45**(1): p. 177-82.
44. Bienstock, R.J. *Overview: Fragment-Based Drug Design*, in *Library Design, Search Methods, and Applications of Fragment-Based Drug Design*, R.J. Bienstock, Editor. 2011. p. 204.
45. Zoete, V.; Grosdidier, A.; Michielin, O. *Docking, virtual high throughput screening and in silico fragment-based drug design*. *J Cell Mol Med*, 2009. **13**(2): p. 238-48.
46. Lipinski, C.A.; Lombardo, F.; Dominy, B.W.; et al. *Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings*. *Advanced Drug Delivery Reviews*, 1997. **23**: p. 3-25.
47. Congreve, M.; Carr, R.; Murray, C.; et al. *A 'rule of three' for fragment-based lead discovery?* *Drug Discov Today*, 2003. **8**(19): p. 876-7.
48. Lagunin, A.; Filimonov, D.; Poroikov, V. *Multi-targeted natural products evaluation based on biological activity prediction with PASS*. *Curr Pharm Des*, 2010. **16**(15): p. 1703-17.
49. Butler, M.S. *Natural products to drugs: natural product-derived compounds in clinical trials*. *Nat Prod Rep*, 2008. **25**(3): p. 475-516.
50. Ganesan, A. *The impact of natural products upon modern drug discovery*. *Curr Opin Chem Biol*, 2008. **12**(3): p. 306-17.
51. Harvey, A.L. *Natural products in drug discovery*. *Drug Discov Today*, 2008. **13**(19-20): p. 894-901.
52. Chen, C.Y. *TCM Database@Taiwan: the world's largest traditional Chinese medicine database for drug screening in silico*. *PLoS One*, 2011. **6**(1): p. e15939.
53. Ntie-Kang, F.; Zofou, D.; Babiaka, S.B.; et al. *AfroDb: a select highly potent and diverse natural product library from African medicinal plants*. *PLoS One*, 2013. **8**(10): p. e78085.
54. Gu, J.; Gui, Y.; Chen, L.; et al. *Use of natural products as chemical library for drug discovery and network pharmacology*. *PLoS One*, 2013. **8**(4): p. e62839.
55. Monge, A., *Création et utilisation de chimiothèques optimisées pour la recherche in silico de nouveaux composés bioactifs*, in *Chimie informatique et théorique 2006*, Orléans. p. 185.
56. O'Boyle, N.M.; Banck, M.; James, C.A.; et al. *Open Babel: An open chemical toolbox*. *J Cheminform*, 2011. **3**: p. 33.
57. Hawkins, P.C.; Skillman, A.G.; Warren, G.L.; et al. *Conformer generation with OMEGA: algorithm and validation using high quality structures from the Protein Databank and Cambridge Structural Database*. *J Chem Inf Model*, 2010. **50**(4): p. 572-84.
58. CORINA Molecular Network GmbH, <http://www.molecular-networks.com/products/corina#features>.
59. LigPrep, S.L.N. York, 2.3, <http://www.schrodinger.com>.
60. Weininger, D. *SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules*. *J. Chem. Inf. Comput. Sci.*, 1988. **28**(1): p. 31-36.
61. Weininger, D.; Weininger, A.; Weininger, J.L. *SMILES. 2. Algorithm for Generation of Unique SMILES Notation*. *J. Chem. Inf. Comput. Sci.*, 1989. **29**: p. 97-101.
62. O'Boyle, N.M. *Towards a Universal SMILES representation - A standard method to generate canonical SMILES based on the InChI*. *J Cheminform*, 2012. **4**(1): p. 22.
63. McNaught, A. *The IUPAC International Chemical Identifier: InChI*. *Chemistry International*, 2006. **28**(6): p. 12-15.

64. Heller, S.; McNaught, A.; Stein, S.; et al. *InChI - the worldwide chemical structure identifier standard*. *J Cheminform*, 2013. **5**(1): p. 7.
65. Dalby, A.; Nourse, J.G.; Hounshell, W.D.; et al. *Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited*. *J. Chem. Inf. Comput. Sci.*, 1992. **32**(3): p. 244-255.
66. Bernstein, F.C.; Koetzle, T.F.; Williams, G.J.; et al. *The Protein Data Bank: a computer-based archival file for macromolecular structures*. *J Mol Biol*, 1977. **112**(3): p. 535-42.
67. Tripos Mol2 Format. Available from: http://www.tripos.com/index.php?family=modules,SimplePage,Mol2_File_Format2009.
68. Berman, H.M.; Westbrook, J.; Feng, Z.; et al. *The Protein Data Bank*. *Nucleic Acids Res*, 2000. **28**(1): p. 235-42.
69. Protein Data Bank (PDB). Available from: <http://www.rcsb.org/pdb/>. [04/02/2014].
70. Corina on line, http://www.molecular-networks.com/online_demos/corina_demo.
71. QuacPac, O.S. Software, Editor.
72. Knox, A.J.; Meegan, M.J.; Carta, G.; et al. *Considerations in compound database preparation--"hidden" impact on virtual screening results*. *J Chem Inf Model*, 2005. **45**(6): p. 1908-19.
73. Vogt, A.D.; Di Cera, E. *Conformational selection is a dominant mechanism of ligand binding*. *Biochemistry*, 2013. **52**(34): p. 5723-9.
74. Perola, E.; Charifson, P.S. *Conformational analysis of drug-like molecules bound to proteins: an extensive study of ligand reorganization upon binding*. *J Med Chem*, 2004. **47**(10): p. 2499-510.
75. Bostrom, J. *Reproducing the conformations of protein-bound ligands: a critical evaluation of several popular conformational searching tools*. *J Comput Aided Mol Des*, 2001. **15**(12): p. 1137-52.
76. Veber, D.F.; Johnson, S.R.; Cheng, H.Y.; et al. *Molecular properties that influence the oral bioavailability of drug candidates*. *J Med Chem*, 2002. **45**(12): p. 2615-23.
77. Ridder, L.; Wang, H.; de Vlieg, J.; et al. *Revisiting the rule of five on the basis of pharmacokinetic data from rat*. *ChemMedChem*, 2011. **6**(11): p. 1967-70.
78. Petit, J.; Meurice, N.; Kaiser, C.; et al. *Softening the Rule of Five--where to draw the line?* *Bioorg Med Chem*, 2012. **20**(18): p. 5343-51.
79. Oprea, T.I.; Bologna, C.; Olah, M. *Compound selection for virtual screening*, in *Virtual Screening in Drug Discovery*, J.C.S. Alvarez, B., Editor. 2004, Wiley-VCH.
80. Marchant, C.A.; Briggs, K.A.; Long, A. *In silico tools for sharing data and knowledge on toxicity and metabolism: derek for windows, meteor, and vitic*. *Toxicol Mech Methods*, 2008. **18**(2-3): p. 177-87.
81. Valerio, L.G., Jr. *In silico toxicology for the pharmaceutical sciences*. *Toxicol Appl Pharmacol*, 2009. **241**(3): p. 356-70.
82. Toropov, A.A.; Toropova, A.P.; Raska, I., Jr.; et al. *Comprehension of drug toxicity: Software and databases*. *Comput Biol Med*, 2014. **45**: p. 20-5.
83. Lahl, U.; Gundert-Remy, U. *The Use of (Q)SAR Methods in the Context of REACH*. *Toxicol Mech Methods*, 2008. **18**(2-3): p. 149-58.
84. Benigni, R.; Bossa, C. *Predictivity and reliability of QSAR models: the case of mutagens and carcinogens*. *Toxicol Mech Methods*, 2008. **18**(2-3): p. 137-47.
85. Myshkin, E.; Brennan, R.; Khasanova, T.; et al. *Prediction of organ toxicity endpoints by QSAR modeling based on precise chemical-histopathology annotations*. *Chem Biol Drug Des*, 2012. **80**(3): p. 406-16.

86. Sanguinetti, M.C.; Tristani-Firouzi, M. *hERG potassium channels and cardiac arrhythmia*. *Nature*, 2006. **440**(7083): p. 463-9.
87. Brown, A.M. *Drugs, hERG and sudden death*. *Cell Calcium*, 2004. **35**(6): p. 543-7.
88. Ekins, S.; Balakin, K.V.; Savchuk, N.; et al. *Insights for human ether-a-go-go-related gene potassium channel inhibition using recursive partitioning and Kohonen and Sammon mapping techniques*. *J Med Chem*, 2006. **49**(17): p. 5059-71.
89. Sun, H. *An accurate and interpretable bayesian classification model for prediction of HERG liability*. *ChemMedChem*, 2006. **1**(3): p. 315-22.
90. Gepp, M.M.; Hutter, M.C. *Determination of hERG channel blockers using a decision tree*. *Bioorg Med Chem*, 2006. **14**(15): p. 5325-32.
91. Du, L.P.; Tsai, K.C.; Li, M.Y.; et al. *The pharmacophore hypotheses of I(Kr) potassium channel blockers: novel class III antiarrhythmic agents*. *Bioorg Med Chem Lett*, 2004. **14**(18): p. 4771-7.
92. Cianchetta, G.; Li, Y.; Kang, J.; et al. *Predictive models for hERG potassium channel blockers*. *Bioorg Med Chem Lett*, 2005. **15**(15): p. 3637-42.
93. Yoshida, K.; Niwa, T. *Quantitative structure-activity relationship studies on inhibition of HERG potassium channels*. *J Chem Inf Model*, 2006. **46**(3): p. 1371-8.
94. Du, L.; Li, M.; You, Q.; et al. *A novel structure-based virtual screening model for the hERG channel blockers*. *Biochem Biophys Res Commun*, 2007. **355**(4): p. 889-94.
95. Johnson, M.A.; Maggiora, G.M. *Concepts and Applications of Molecular Similarity*. Wiley ed. 1990.
96. Koeppen, H.; Kriegl, J.; Lessel, U.; et al. *Ligand-Based Virtual Screening*, in *Virtual Screening Principles, Challenges and Practical Guidelines*, C. Sottriffer, Editor. 2011. p. 536.
97. Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors. Methods and Principles in Medicinal Chemistry*, ed. R.K. Mannhold, H.; Timmerman, H. Vol. 11. 2008.
98. Bender, A.; Glen, R.C. *Molecular similarity: a key technique in molecular informatics*. *Org Biomol Chem*, 2004. **2**(22): p. 3204-18.
99. Bielska, E.; Lucas, X.; Czerwoniec, A.; et al. *Virtual screening strategies in drug design - methods and applications*. *BioTechnologia*, 2011. **92**(3): p. 249-264.
100. Barnard, J.M. *Substructure Searching Methods: Old and New*. *J. Chem. Inf. Comput. Sci.*, 1993. **33**: p. 532-538.
101. Cone, M.M.; Venkataraghavan, R.; McLafferty, F.W. *Molecular Structure Comparison Program for the Identification of Maximal Common Substructures*. *Journal of the American Chemical Society*, 1977. **99**(23): p. 7668-7671.
102. *Chemoinformatics A Textbook*. Wiley-VCH ed, ed. J. Gasteiger and T. Engel. 2003.
103. MACCS-II, S.C. MDL Information Systems/symyx, CA,
104. Barnard, J.M.; Downs, G.M. *Chemical Fragment Generation and Clustering Software*. *J. Chem. Inf. Comput. Sci.*, 1997. **37**(1): p. 141-142.
105. Daylight Fingerprint Toolkit, A.V. Daylight Chemical Systems Inc, CA, 4.9,
106. Lengauer, T.; Lemmen, C.; Rarey, M.; et al. *Novel technologies for virtual screening*. *Drug Discov Today*, 2004. **9**(1): p. 27-34.
107. Böhm, H.-J.; Flohr, A.; Stahl, M. *Scaffold hopping*. *Drug Discovery Today: Technologies*, 2004. **1**(3): p. 217-224.
108. Meyer, A.Y.; Richards, W.G. *Similarity of molecular shape*. *Journal of Computer-Aided Molecular Design*, 1991. **5**: p. 427-439.
109. Nicholls, A.; McGaughey, G.B.; Sheridan, R.P.; et al. *Molecular shape and medicinal chemistry: a perspective*. *J Med Chem*, 2010. **53**(10): p. 3862-86.

110. Ballester, P.J.; Richards, W.G. *Ultrafast shape recognition to search compound databases for similar molecular shapes*. J Comput Chem, 2007. **28**(10): p. 1711-23.
111. Bemis, G.W.; Kuntz, I.D. *A fast and efficient method for 2D and 3D molecular shape description*. J Comput Aided Mol Des, 1992. **6**(6): p. 607-28.
112. Nilakantan, R.; Bauman, N.; Venkataraghavan, R. *New method for rapid characterization of molecular shapes: applications in drug design*. J Chem Inf Comput Sci, 1993. **33**(1): p. 79-85.
113. Good, A.C.; Ewing, T.J.; Gschwend, D.A.; et al. *New molecular shape descriptors: application in database screening*. J Comput Aided Mol Des, 1995. **9**(1): p. 1-12.
114. Zauhar, R.J.; Moyna, G.; Tian, L.; et al. *Shape signatures: a new approach to computer-aided ligand- and receptor-based drug design*. J Med Chem, 2003. **46**(26): p. 5674-90.
115. Meyer, A.Y.; Richards, W.G. *Similarity of molecular shape*. J Comput Aided Mol Des, 1991. **5**(5): p. 427-39.
116. Hahn, M. *Three-Dimensional Shape-Based Searching of Conformationally Flexible Compounds*. J. Chem. Inf. Comput. Sci., 1997. **37**: p. 80-86.
117. Good, A.C.; Richards, W.G. *Rapid Evaluation of Shape Similarity Using Gaussian Functions*. J. Chem. Inf. Comput. Sci., 1993. **33**(1): p. 112-116.
118. Masek, B.B.; Merchant, A.; Matthew, J.B. *Molecular shape comparison of angiotensin II receptor antagonists*. J Med Chem, 1993. **36**(9): p. 1230-8.
119. Grant, J.A.; Pickup, B.T. *A Gaussian Description of Molecular Shape*. J. Phys. Chem, 1995. **99**: p. 3503-3510.
120. ROCS, O.S. Software, <http://www.eyesopen.com>.
121. Lalonde, J.M.; Elban, M.A.; Courter, J.R.; et al. *Design, synthesis and biological evaluation of small molecule inhibitors of CD4-gp120 binding based on virtual screening*. Bioorg Med Chem, 2011. **19**(1): p. 91-101.
122. Hevener, K.E.; Mehboob, S.; Su, P.C.; et al. *Discovery of a novel and potent class of F. tularensis enoyl-reductase (FabI) inhibitors by molecular shape and electrostatic matching*. J Med Chem, 2012. **55**(1): p. 268-79.
123. Willett, P. *Chemical Similarity Searching*. J. Chem. Inf. Comput. Sci., 1998. **38**: p. 983-996.
124. Willett, P. *Similarity searching using 2D structural fingerprints*. Methods Mol Biol, 2011. **672**: p. 133-58.
125. Dixon, S.L.; Koehler, R.T. *The hidden component of size in two-dimensional fragment descriptors: side effects on sampling in bioactive libraries*. J Med Chem, 1999. **42**(15): p. 2887-900.
126. Hodgkin, E.E.; Richards, W.G. *Molecular similarity*. Chemistry in Britain, 1988. **24**: p. 1141-1144.
127. Carbo, R.; Leyda, L.; Arnau, M. *How similar is a molecule to another? An electron density measure of similarity between two molecular structures*. International Journal of Quantum Chemistry, 1980. **17**: p. 1185-1189.
128. Holliday, J.D.; Salim, N.; Whittle, M.; et al. *Analysis and display of the size dependence of chemical similarity coefficients*. J Chem Inf Comput Sci, 2003. **43**(3): p. 819-28.
129. Ehrlich, P. *Über die Constitution des Diphtheriegiftes*. Deut. Med. Wochschr., 1898. **24**: p. 597-600.
130. Guner, O.F.; Bowen, J.P. *Setting the record straight: the origin of the pharmacophore concept*. J Chem Inf Model, 2014. **54**(5): p. 1269-83.
131. Schueler, F.W. *Chemobiodynamics and Drug Design*. McGraw-Hill ed. 1960.

132. Beckett, A.H.; Harper, N.J.; Clitherow, J.W. *The importance of stereoisomerism in muscarinic activity*. J Pharm Pharmacol, 1963. **15**: p. 362-71.
133. Kier, L.B. *Molecular orbital calculation of preferred conformations of acetylcholine, muscarine, and muscarone*. Mol Pharmacol, 1967. **3**(5): p. 487-94.
134. Wermuth, C.-G.; Gannelin, C.R.; Lindberg, P.; et al. *Glossary of Terms Used in Medicinal Chemistry (IUPAC Recommendations 1997)*, in *Annual Reports in Medicinal Chemistry*, J.A. Bristol, Editor. 1998. p. 385-395.
135. Horvath, D. *Topological Pharmacophores*, in *Chemoinformatics Approaches to Virtual Screening*, A.T. Varnek, A., Editor. 2008. p. 338.
136. Schuffenhauer, A.; Floersheim, P.; Acklin, P.; et al. *Similarity metrics for ligands reflecting the similarity of the target proteins*. J Chem Inf Comput Sci, 2003. **43**(2): p. 391-405.
137. Stiefl, N.; Watson, I.A.; Baumann, K.; et al. *ErG: 2D pharmacophore descriptions for scaffold hopping*. J Chem Inf Model, 2006. **46**(1): p. 208-20.
138. Gillet, V.J.; Willett, P.; Bradshaw, J. *Similarity searching using reduced graphs*. J Chem Inf Comput Sci, 2003. **43**(2): p. 338-45.
139. Schneider, G.; Neidhart, W.; Giller, T.; et al. *"Scaffold-Hopping" by Topological Pharmacophore Search: A Contribution to Virtual Screening*. Angew Chem Int Ed Engl, 1999. **38**(19): p. 2894-2896.
140. Reutlinger, M.; Koch, C.P.; Reker, D.; et al. *Chemically Advanced Template Search (CATS) for Scaffold-Hopping and Prospective Target Prediction for 'Orphan' Molecules*. Mol Inform, 2013. **32**(2): p. 133-138.
141. Hessler, G.; Baringhaus, K.-H. *The scaffold hopping potential of pharmacophores*. Drug Discov Today Technol, 2010. **7**(4): p. e203-70.
142. Rarey, M.; Dixon, J.S. *Feature trees: a new molecular similarity measure based on tree matching*. J Comput Aided Mol Des, 1998. **12**(5): p. 471-90.
143. Dror, O.; Shulman-Peleg, A.; Nussinov, R.; et al. *Predicting molecular interactions in silico: I. A guide to pharmacophore identification and its applications to drug design*. Curr Med Chem, 2004. **11**(1): p. 71-90.
144. Martin, Y.C. *DISCO: What We Did Right and What We Missed*, in *Pharmacophore Perception, Development, and Use in Drug Design*, O.F. Güner, Editor. 2000. p. 537.
145. Dolata, D.P.; Parrill, A.L.; Walters, W.P. *CLEW: The Generation of Pharmacophore Hypotheses Through Machine Learning*. SAR and QSAR in Environmental Research, 1998. **9**: p. 53-81.
146. Li, H.; Sutter, J.; Hoffmann, R. *HypoGen: An Automated System for Generating 3D Predictive Pharmacophore Models*, in *Pharmacophore Perception, Development, and Use in Drug Design*, O.F. Güner, Editor. 2000. p. 537.
147. Martin, Y.C.; Bures, M.G.; Danaher, E.A.; et al. *A fast new approach to pharmacophore mapping and its application to dopaminergic and benzodiazepine agonists*. J Comput Aided Mol Des, 1993. **7**(1): p. 83-102.
148. Finn, P.W.; Kavraki, L.E.; Latombe, J.C.; et al. *RAPID: Randomized pharmacophore identification for drug design*. Computational Geometry Theory and Applications, 1998. **10**: p. 263-272.
149. Chen, X.; Rusinko, A., 3rd; Tropsha, A.; et al. *Automated pharmacophore identification for large chemical data sets*. J Chem Inf Comput Sci, 1999. **39**(5): p. 887-96.
150. Jones, G.; Willett, P.; Glen, R.C. *A genetic algorithm for flexible molecular overlay and pharmacophore elucidation*. J Comput Aided Mol Des, 1995. **9**(6): p. 532-49.

151. Handschuh, S.; Wagener, M.; Gasteiger, J. *Superposition of three-dimensional chemical structures allowing for conformational flexibility by a hybrid method.* J Chem Inf Comput Sci, 1998. **38**(2): p. 220-32.
152. Holliday, J.D.; Willett, P. *Using a genetic algorithm to identify common structural features in sets of ligands.* J Mol Graph Model, 1997. **15**(4): p. 221-32.
153. CATALYST, Accelrys Inc., San Diego, CA, USA, www.accelrys.com.
154. Clement, O.O.; Mehl, A.T. *HipHop: Pharmacophores Based on Multiple Common-Feature Alignments*, in *Pharmacophore Perception, Development, and Use in Drug Design*, O.F. Güner, Editor. 2000. p. 537.
155. MOE, Chemical Computing Group, Montreal, QC, CA, www.chemcomp.com.
156. Dixon, S.L.; Smondyrev, A.M.; Knoll, E.H.; 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, 2006. **20**(10-11): p. 647-71.
157. Chopra, M.; Gupta, R.; Gupta, S.; et al. *Molecular modeling study on chemically diverse series of cyclooxygenase-2 selective inhibitors: generation of predictive pharmacophore model using catalyst.* J Mol Model, 2008. **14**(11): p. 1087-99.
158. Deschênes, A.; Sourial, E. *Ligand Scaffold Replacement using MOE Pharmacophore Tools.* Journal of Chemical Computing Group, 2007.
159. Sen, D.; Chatterjee, T.K. *Pharmacophore modeling and 3D quantitative structure-activity relationship analysis of febrifugine analogues as potent antimalarial agent.* J Adv Pharm Technol Res, 2013. **4**(1): p. 50-60.
160. Pickett, S. *The Biophore Concept*, in *Protein-Ligand Interactions: From Molecular Recognition to Drug Design*, H.-J.S. Böhm, G., Editor. 2006. p. 262.
161. Laggner, C.; Wolber, G.; Kirchmair, J.; et al. *Pharmacophore-based Virtual Screening in Drug Discovery*, in *Chemoinformatics Approaches to Virtual Screening*, A.T. Varnek, A., Editor. 2008. p. 338.
162. Gillet, V.J. *Pharmacophore Models in Drug Design*, in *Physico-Chemical and Computational Approaches to Drug Discovery*, J.B. Luque, X., Editor. 2012. p. 418.
163. Finn, P.W. *Computer-based screening of compound database for the identification of novel leads.* Drug Discovery Today, 1996. **1**(9): p. 363-370.
164. Hoffman, R. *Which conformations in Catalyst Databases? Sharing an Observation.* Accelrys Ressource Center http://accelrys.com/resource-center/case-studies/archive/studies/catalystdb_full.html.
165. Chem-X, <http://www.chem.ac.ru/Chemistry/Soft/CHEM-X.en.html>.
166. Fang, X.W., S. *A Web-Based 3D-Database Pharmacophore Searching Tool for Drug Discovery.* J. Chem. Inf. Comput. Sci., 2002. **42**: p. 192-198.
167. Moock, T.E.; Douglas, R.H.; Ozkabak, A.G.; et al. *Conformational searching in ISIS/3D databases.* J Chem Inf Comput Sci, 1994. **34**(1): p. 184-189.
168. Wang, T.; Zhou, J. *3DFS: A New 3D Flexible Searching System for Use in Drug Design.* J. Chem. Inf. Comput. Sci., 1998. **38**(1): p. 71-77.
169. Seidel, T.; Ibis, G.; Bendix, F.; et al. *Strategies for 3D pharmacophore-based virtual screening.* Drug Discovery Today: Technologies, 2010. **7**(4): p. e221-e228.
170. Jakes, S.; Watts, N.; Willett, P.; et al. *Pharmacophoric pattern matching in files of 3D chemical structures: evaluation of search performance.* Journal of Molecular Graphics, 1987. **5**(1): p. 41-48.
171. Brint, A.T.; Willett, P. *Identifying 3D maximal common substructures using transputer networks.* Journal of Molecular Graphics, 1987. **5**(4): p. 200-207.
172. Patrick, G.L. *Aspects quantitatifs des relations structure-activité (RSA)*, in *Chimie Pharmaceutique*. 2002. p. 648.

173. Damale, M.G.; Harke, S.N.; Kalam Khan, F.A.; et al. *Recent advances in multidimensional QSAR (4D-6D): a critical review*. *Mini Rev Med Chem*, 2014. **14**(1): p. 35-55.
174. Andrade, C.H.; Pasqualoto, K.F.; Ferreira, E.I.; et al. *4D-QSAR: perspectives in drug design*. *Molecules*, 2010. **15**(5): p. 3281-94.
175. Vedani, A.; Dobler, M. *5D-QSAR: the key for simulating induced fit?* *J Med Chem*, 2002. **45**(11): p. 2139-49.
176. Vedani, A.; Descloux, A.V.; Spreafico, M.; et al. *Predicting the toxic potential of drugs and chemicals in silico: a model for the peroxisome proliferator-activated receptor gamma (PPAR gamma)*. *Toxicol Lett*, 2007. **173**(1): p. 17-23.
177. Hansch, C.; Leo, A.; Hoekman, D.H. *Exploring QSAR: Fundamentals and applications in chemistry and biology*. American Chemical Society ed. 1995. 580.
178. Burden, F.R. *A chemically intuitive molecular index based on the eigenvalues of a modified adjacency matrix*. *Quantitative Structure-Activity Relationships*, 1997. **16**: p. 309-314.
179. Garg, R.; Kurup, A.; Hansch, C. *Comparative QSAR: on the toxicology of the phenolic OH moiety*. *Crit Rev Toxicol*, 2001. **31**(2): p. 223-45.
180. Leo, A.; Jow, P.Y.; Silipo, C.; et al. *Calculation of hydrophobic constant (log P) from pi and f constants*. *J Med Chem*, 1975. **18**(9): p. 865-8.
181. Taft, R.W. *Steric Effects in Organic Chemistry*. Wiley ed, ed. M.S. Newman. 1956.
182. Scior, T.; Medina-Franco, J.L.; Do, Q.T.; et al. *How to recognize and workaround pitfalls in QSAR studies: a critical review*. *Curr Med Chem*, 2009. **16**(32): p. 4297-313.
183. Sippl, W. *3D-QSAR - Applications, recent advances, and limitations*, in *Recent Advances in QSAR Studies Methods and Applications*, T.L. Puzyn, K.; Cronin, M.T.D., Editor. 2010. p. 414.
184. Cramer, R.D.; Milne, M., *The lattice model: A general paradigm for shape-related structure/activity correlation*, in *American Chemical Society Meeting, Computer Chemistry Section* 1979.
185. Cramer, R.D.; Patterson, D.E.; Bunce, J.D. *Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins*. *J Am Chem Soc*, 1988. **110**(18): p. 5959-67.
186. Höltje, H.D.; Sippl, W.; Rognan, D.; et al. *Molecular Modeling. Basic Principles and Applications*. Wiley-VCH ed. 2003.
187. Bohm, H.J.; Klebe, G.; Kubinyi, H. *Wirkstoffdesign*. Spektrum-Akademischer Vlg ed. 1996.
188. Klebe, G.; Abraham, U.; Mietzner, T. *Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity*. *J Med Chem*, 1994. **37**(24): p. 4130-46.
189. Kearsley, S.K.; Smith, G.M. *An alternative method for the alignment of molecular structures: Maximizing electrostatic and steric overlap*. *Tetrahedron Computer Methodology*, 1990. **3**(6): p. 615-633.
190. Cichero, E.; Cesarini, S.; Mosti, L.; et al. *CoMFA and CoMSIA analyses on 1,2,3,4-tetrahydropyrrolo[3,4-b]indole and benzimidazole derivatives as selective CB2 receptor agonists*. *J Mol Model*, 2010. **16**(9): p. 1481-98.
191. Reynolds, C.A.; Wade, R.C.; Goodford, P.J. *Identifying targets for bioreductive agents: using GRID to predict selective binding regions of proteins*. *J Mol Graph*, 1989. **7**(2): p. 103-8, 100.

192. Nilsson, J.; Wikstrom, H.; Smilde, A.; et al. *GRID/GOLPE 3D quantitative structure-activity relationship study on a set of benzamides and naphthamides, with affinity for the dopamine D3 receptor subtype*. J Med Chem, 1997. **40**(6): p. 833-40.
193. Ragno, R.; Simeoni, S.; Valente, S.; et al. *3-D QSAR studies on histone deacetylase inhibitors. A GOLPE/GRID approach on different series of compounds*. J Chem Inf Model, 2006. **46**(3): p. 1420-30.
194. Naerum, L.; Norskov-Lauritsen, L.; Olesen, P.H. *Scaffold hopping and optimization towards libraries of glycogen synthase kinase-3 inhibitors*. Bioorg Med Chem Lett, 2002. **12**(11): p. 1525-8.
195. Singh, J.; Van Vlijmen, H.; Liao, Y.; et al. *Identification of potent and novel alpha4beta1 antagonists using in silico screening*. J Med Chem, 2002. **45**(14): p. 2988-93.
196. Flohr, S.; Kurz, M.; Kostenis, E.; 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, 2002. **45**(9): p. 1799-805.
197. Mustata, G.; Follis, A.V.; Hammoudeh, D.I.; et al. *Discovery of novel Myc-Max heterodimer disruptors with a three-dimensional pharmacophore model*. J Med Chem, 2009. **52**(5): p. 1247-50.
198. Rush, T.S., 3rd; Grant, J.A.; Mosyak, L.; et al. *A shape-based 3-D scaffold hopping method and its application to a bacterial protein-protein interaction*. J Med Chem, 2005. **48**(5): p. 1489-95.
199. Bostrom, J.; Berggren, K.; Elebring, T.; et al. *Scaffold hopping, synthesis and structure-activity relationships of 5,6-diaryl-pyrazine-2-amide derivatives: a novel series of CBI receptor antagonists*. Bioorg Med Chem, 2007. **15**(12): p. 4077-84.
200. Bologna, C.G.; Revankar, C.M.; Young, S.M.; et al. *Virtual and biomolecular screening converge on a selective agonist for GPR30*. Nat Chem Biol, 2006. **2**(4): p. 207-12.
201. 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, 2008. **16**(2): p. 838-53.
202. Al-Sha'er, M.A.; Taha, M.O. *Elaborate ligand-based modeling reveals new nanomolar heat shock protein 90alpha inhibitors*. J Chem Inf Model, 2010. **50**(9): p. 1706-23.
203. Ward, W.H.; Cook, P.N.; Slater, A.M.; et al. *Epidermal growth factor receptor tyrosine kinase. Investigation of catalytic mechanism, structure-based searching and discovery of a potent inhibitor*. Biochem Pharmacol, 1994. **48**(4): p. 659-66.
204. Herbst, R.S.; Fukuoka, M.; Baselga, J. *Gefitinib--a novel targeted approach to treating cancer*. Nat Rev Cancer, 2004. **4**(12): p. 956-65.
205. Laurie, A.T.; Jackson, R.M. *Methods for the prediction of protein-ligand binding sites for structure-based drug design and virtual ligand screening*. Curr Protein Pept Sci, 2006. **7**(5): p. 395-406.
206. Kuntz, I.D.; Blaney, J.M.; Oatley, S.J.; et al. *A geometric approach to macromolecule-ligand interactions*. J Mol Biol, 1982. **161**(2): p. 269-88.
207. DesJarlais, R.L.; Sheridan, R.P.; Seibel, G.L.; et al. *Using shape complementarity as an initial screen in designing ligands for a receptor binding site of known three-dimensional structure*. J Med Chem, 1988. **31**(4): p. 722-9.
208. Laskowski, R.A.; Luscombe, N.M.; Swindells, M.B.; et al. *Protein clefts in molecular recognition and function*. Protein Sci, 1996. **5**(12): p. 2438-52.
209. MOLCAD, Molcad GmbH. Available at <http://www.molcad.de/index.html.en>,

210. Levitt, D.G.; Banaszak, L.J. *POCKET: a computer graphics method for identifying and displaying protein cavities and their surrounding amino acids*. *J Mol Graph*, 1992. **10**(4): p. 229-34.
211. Hendlich, M.; Rippmann, F.; Barnickel, G. *LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins*. *J Mol Graph Model*, 1997. **15**(6): p. 359-63, 389.
212. Laurie, A.T.; Jackson, R.M. *Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites*. *Bioinformatics*, 2005. **21**(9): p. 1908-16.
213. Desaphy, J.; Azdimousa, K.; Kellenberger, E.; et al. *Comparison and druggability prediction of protein-ligand binding sites from pharmacophore-annotated cavity shapes*. *J Chem Inf Model*, 2012. **52**(8): p. 2287-99.
214. Connolly, M.L. *Solvent-accessible surfaces of proteins and nucleic acids*. *Science*, 1983. **221**(4612): p. 709-13.
215. Peters, K.P.; Fauck, J.; Frommel, C. *The automatic search for ligand binding sites in proteins of known three-dimensional structure using only geometric criteria*. *J Mol Biol*, 1996. **256**(1): p. 201-13.
216. Dundas, J.; Ouyang, Z.; Tseng, J.; et al. *CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues*. *Nucleic Acids Res*, 2006. **34**(Web Server issue): p. W116-8.
217. Laskowski, R.A. *SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions*. *J Mol Graph*, 1995. **13**(5): p. 323-30, 307-8.
218. CASTp. Available at: <http://sts-fw.bioengr.uic.edu/castp/background.php>,
219. Goodford, P.J. *A computational procedure for determining energetically favorable binding sites on biologically important macromolecules*. *J Med Chem*, 1985. **28**(7): p. 849-57.
220. Ruppert, J.; Welch, W.; Jain, A.N. *Automatic identification and representation of protein binding sites for molecular docking*. *Protein Sci*, 1997. **6**(3): p. 524-33.
221. Pupko, T.; Bell, R.E.; Mayrose, I.; et al. *Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues*. *Bioinformatics*, 2002. **18 Suppl 1**: p. S71-7.
222. Glaser, F.; Pupko, T.; Paz, I.; et al. *ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information*. *Bioinformatics*, 2003. **19**(1): p. 163-4.
223. Dai, T.; Liu, Q.; Gao, J.; et al. *A new protein-ligand binding sites prediction method based on the integration of protein sequence conservation information*. *BMC Bioinformatics*, 2011. **12 Suppl 14**: p. S9.
224. Hendlich, M. *Databases for protein-ligand complexes*. *Acta Crystallogr D Biol Crystallogr*, 1998. **54**(Pt 6 Pt 1): p. 1178-82.
225. Stark, A.; Sunyaev, S.; Russell, R.B. *A model for statistical significance of local similarities in structure*. *J Mol Biol*, 2003. **326**(5): p. 1307-16.
226. Shulman-Peleg, A.; Nussinov, R.; Wolfson, H.J. *Recognition of functional sites in protein structures*. *J Mol Biol*, 2004. **339**(3): p. 607-33.
227. Kinoshita, K.; Furui, J.; Nakamura, H. *Identification of protein functions from a molecular surface database, eF-site*. *J Struct Funct Genomics*, 2002. **2**(1): p. 9-22.
228. Laskowski, R.A.; Watson, J.D.; Thornton, J.M. *ProFunc: a server for predicting protein function from 3D structure*. *Nucleic Acids Res*, 2005. **33**(Web Server issue): p. W89-93.
229. Gold, N.D.; Jackson, R.M. *SitesBase: a database for structure-based protein-ligand binding site comparisons*. *Nucleic Acids Res*, 2006. **34**(Database issue): p. D231-4.

230. Liang, M.P.; Banatao, D.R.; Klein, T.E.; et al. *WebFEATURE: An interactive web tool for identifying and visualizing functional sites on macromolecular structures*. *Nucleic Acids Res*, 2003. **31**(13): p. 3324-7.
231. *WebFEATURE*. Available at: <http://feature.stanford.edu/index.php>.
232. Bohm, H.J. *The computer program LUDI: a new method for the de novo design of enzyme inhibitors*. *J Comput Aided Mol Des*, 1992. **6**(1): p. 61-78.
233. Bohm, H.J. *LUDI: rule-based automatic design of new substituents for enzyme inhibitor leads*. *J Comput Aided Mol Des*, 1992. **6**(6): p. 593-606.
234. Carlson, H.A.; Masukawa, K.M.; Rubins, K.; et al. *Developing a dynamic pharmacophore model for HIV-1 integrase*. *J Med Chem*, 2000. **43**(11): p. 2100-14.
235. Masukawa, K.M.; Carlson, H.A.; McCammon, J.A. *Technique for developing a pharmacophore model that accommodates inherent protein flexibility: an application to HIV-1 integrase*, in *Pharmacophore Perception, Development, and Use in Drug Design*, O.F. Güner, Editor. 2000. p. 537.
236. Koes, D.R.; Camacho, C.J. *ZINCPharmer: pharmacophore search of the ZINC database*. *Nucleic Acids Res*, 2012. **40**(Web Server issue): p. W409-14.
237. Wolber, G.; Langer, T. *LigandScout: 3-D pharmacophores derived from protein-bound ligands and their use as virtual screening filters*. *J Chem Inf Model*, 2005. **45**(1): p. 160-9.
238. Chen, Z.; Li, H.L.; Zhang, Q.J.; et al. *Pharmacophore-based virtual screening versus docking-based virtual screening: a benchmark comparison against eight targets*. *Acta Pharmacol Sin*, 2009. **30**(12): p. 1694-708.
239. Ortiz, A.R.; Pisabarro, M.T.; Gago, F.; et al. *Prediction of drug binding affinities by comparative binding energy analysis*. *J Med Chem*, 1995. **38**(14): p. 2681-91.
240. Head, R.D.; Smythe, M.L.; Oprea, T.I.; et al. *VALIDATE: A new method for the receptor-based prediction of binding affinities of novel ligands*. *J Am Chem Soc* 1996. **118**: p. 3959-3969.
241. Di Santo, R.; Costi, R.; Artico, M.; et al. *Design, synthesis and QSAR studies on N-aryl heteroarylisopropanolamines, a new class of non-peptidic HIV-1 protease inhibitors*. *Bioorg Med Chem*, 2002. **10**(8): p. 2511-26.
242. Gil-Redondo, R.; Klett, J.; Gago, F.; et al. *gCOMBINE: A graphical user interface to perform structure-based comparative binding energy (COMBINE) analysis on a set of ligand-receptor complexes*. *Proteins*, 2010. **78**(1): p. 162-72.
243. Liu, S.; Fu, R.; Cheng, X.; et al. *Exploring the binding of BACE-1 inhibitors using comparative binding energy analysis (COMBINE)*. *BMC Struct Biol*, 2012. **12**: p. 21.
244. Wermuth, C.G.; Gannelin, C.R.; Lindberg, P.; et al. *Glossary of terms used in medicinal chemistry*. *Pure and Applied Chemistry*, 1998. **70**: p. 1129-1143.
245. Lewis, R.A.; Dean, P.M. *Automated site-directed drug design: the formation of molecular templates in primary structure generation*. *Proc R Soc Lond B Biol Sci*, 1989. **236**(1283): p. 141-62.
246. Gillett, V.A.; Johnson, A.P.; Mata, P.; et al. *Automated structure design in 3D*. *Tetrahedron Computer Methodology*, 1990. **3**: p. 681-696.
247. Lewis, R.A. *Automated site-directed drug design: approaches to the formation of 3D molecular graphs*. *J Comput Aided Mol Des*, 1990. **4**(2): p. 205-10.
248. Lewis, R.A.; Roe, D.C.; Huang, C.; et al. *Automated site-directed drug design using molecular lattices*. *J Mol Graph*, 1992. **10**(2): p. 66-78, 106.
249. Nishibata, Y.; Itai, A. *Automatic creation of drug candidate structures based on receptor structure. Starting point for artificial lead generation*. *Tetrahedron*, 1991. **47**: p. 8985-8990.

250. Tschinke, V.; Cohen, N.C. *The NEWLEAD program: a new method for the design of candidate structures from pharmacophoric hypotheses*. *J Med Chem*, 1993. **36**(24): p. 3863-70.
251. Ho, C.M.W.; Marshall, G.R. *SPLICE: a program to assemble partial query solutions from three-dimensional database searches into novel ligands*. *Journal of Computer-Aided Molecular Design*, 1993. **7**: p. 623-647.
252. Rotstein, S.H.; Murcko, M.A. *GenStar: a method for de novo drug design*. *J Comput Aided Mol Des*, 1993. **7**(1): p. 23-43.
253. Rotstein, S.H.; Murcko, M.A. *GroupBuild: a fragment-based method for de novo drug design*. *J Med Chem*, 1993. **36**(12): p. 1700-10.
254. Pearlman, D.A.; Murcko, M.A. *CONCEPTS: new dynamic algorithm for de novo design suggestion*. *Journal of Computational Chemistry*, 1993. **14**: p. 1184-1193.
255. Gillett, V.J.; Myatt, G.; Zsodos, Z.; et al. *SPROUT, HIPPO and CAESA: tools for de novo structure generation and estimation of synthetic accessibility*. *Perspectives in Drug Discovery and Design*, 1995. **3**: p. 34-50.
256. Eisen, M.B.; Wiley, D.C.; Karplus, M.; et al. *HOOK: a program for finding novel molecular architectures that satisfy the chemical and steric requirements of a macromolecule binding site*. *Proteins*, 1994. **19**(3): p. 199-221.
257. Miranker, A.; Karplus, M. *Functionality maps of binding sites: a multiple copy simultaneous search method*. *Proteins*, 1991. **11**(1): p. 29-34.
258. 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*. *Journal of American Chemical Society*, 1994. **116**: p. 5560-5571.
259. Gehlhaar, D.K.; Moerder, K.E.; Zichi, D.; et al. *De novo design of enzyme inhibitors by Monte Carlo ligand generation*. *J Med Chem*, 1995. **38**(3): p. 466-72.
260. Glen, R.C.; Payne, A.W. *A genetic algorithm for the automated generation of molecules within constraints*. *J Comput Aided Mol Des*, 1995. **9**(2): p. 181-202.
261. Miranker, A.; Karplus, M. *An automated method for dynamic ligand design*. *Proteins*, 1995. **23**(4): p. 472-90.
262. Frenkel, D.; Clark, D.E.; Li, J.; et al. *PRO_LIGAND: an approach to de novo molecular design. 4. Application to the design of peptides*. *J Comput Aided Mol Des*, 1995. **9**(3): p. 213-25.
263. DeWitte, R.S.; Shakhnovich, E.I. *SMoG de novo design method based on simple, fast, and accurate free energy estimates. 1. Methodologies and supporting evidence*. *Journal of American Chemical Society*, 1996. **118**: p. 11733-11744.
264. Ishchenko, A.V.; Shakhnovich, E.I. *SMall Molecule Growth 2001 (SMoG2001): an improved knowledge-based scoring function for protein-ligand interactions*. *J Med Chem*, 2002. **45**(13): p. 2770-80.
265. Roe, D.C.; Kuntz, I.D. *BUILDER v.2: improving the chemistry of a de novo design strategy*. *J Comput Aided Mol Des*, 1995. **9**(3): p. 269-82.
266. Pearlman, D.A.; Murcko, M.A. *CONCERTS: dynamic connection of fragments as an approach to de novo ligand design*. *J Med Chem*, 1996. **39**(8): p. 1651-63.
267. Luo, Z.; Wang, R.; Lai, L. *RASSE: a new method for structure-based drug design*. *J Chem Inf Comput Sci*, 1996. **36**(6): p. 1187-94.
268. Murray, C.W.; Clark, D.E.; Auton, T.R.; et al. *PRO_SELECT: combining structure-based drug design and combinatorial chemistry for rapid lead discovery. 1. Technology*. *J Comput Aided Mol Des*, 1997. **11**(2): p. 193-207.
269. Todorov, N.P.; Dean, P.M. *Evaluation of a method for controlling molecular scaffold diversity in de novo ligand design*. *J Comput Aided Mol Des*, 1997. **11**(2): p. 175-92.

270. Todorov, N.P.; Dean, P.M. *A branch-and-bound method for optimal atom-type assignment in de novo ligand design*. *J Comput Aided Mol Des*, 1998. **12**(4): p. 335-49.
271. Nachbar, R.B. *Molecular evolution: automated manipulation of hierarchical chemical topology and its application to average molecular structures*. *Genetic Programming and Evolvable Machines*, 2000. **1**: p. 57-94.
272. Globus, A.; Lawton, J.; Wipke, W.T. *Automatic Molecular design using evolutionary algorithms*. *Nanotechnology*, 1999. **10**: p. 290-299.
273. Liu, H.; Duan, Z.; Luo, Q.; et al. *Structure-based ligand design by dynamically assembling molecular building blocks at binding site*. *Proteins*, 1999. **36**(4): p. 462-70.
274. 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*, 2000. **14**(5): p. 449-66.
275. Wang, R.; Gao, Y.; Lai, L. *LigBuilder: a multi-purpose program for structure-based drug design*. *J. Mol. Model.*, 2000. **6**: p. 498-516.
276. Schneider, G.; Lee, M.L.; Stahl, M.; et al. *De novo design of molecular architectures by evolutionary assembly of drug-derived building blocks*. *J Comput Aided Mol Des*, 2000. **14**(5): p. 487-94.
277. Zhu, J.; Fan, H.; Liu, H.; et al. *Structure-based ligand design for flexible proteins: application of new F-DycoBlock*. *J Comput Aided Mol Des*, 2001. **15**(11): p. 979-96.
278. Pegg, S.C.; Haresco, J.J.; Kuntz, I.D. *A genetic algorithm for structure-based de novo design*. *J Comput Aided Mol Des*, 2001. **15**(10): p. 911-33.
279. Pelligrini, E.; Field, M.J. *Development and testing of a de novo drug-design algorithm*. *Journal of Computer-Aided Molecular Design*, 2003. **17**: p. 621-641.
280. Vinkers, H.M.; de Jonge, M.R.; Daeyaert, F.F.; et al. *SYNOPSIS: SYNthesize and OPTimize System in Silico*. *J Med Chem*, 2003. **46**(13): p. 2765-73.
281. Brown, N.; McKay, B.; Gilardoni, F.; et al. *A graph-based genetic algorithm and its application to the multiobjective evolution of median molecules*. *J Chem Inf Comput Sci*, 2004. **44**(3): p. 1079-87.
282. Pierce, A.C.; Rao, G.; Bemis, G.W. *BREED: Generating novel inhibitors through hybridization of known ligands. Application to CDK2, p38, and HIV protease*. *J Med Chem*, 2004. **47**(11): p. 2768-75.
283. Douguet, D.; Munier-Lehmann, H.; Labesse, G.; et al. *LEA3D: a computer-aided ligand design for structure-based drug design*. *J Med Chem*, 2005. **48**(7): p. 2457-68.
284. Nikitin, S.; Zaitseva, N.; Demina, O.; et al. *A very large diversity space of synthetically accessible compounds for use with drug design programs*. *J Comput Aided Mol Des*, 2005. **19**(1): p. 47-63.
285. Moore, W.R., Jr. *Maximizing discovery efficiency with a computationally driven fragment approach*. *Curr Opin Drug Discov Devel*, 2005. **8**(3): p. 355-64.
286. Degen, J.; Rarey, M. *FlexNovo: structure-based searching in large fragment spaces*. *ChemMedChem*, 2006. **1**(8): p. 854-68.
287. Jorgensen, W.L.; Ruiz-Caro, J.; Tirado-Rives, J.; et al. *Computer-aided design of non-nucleoside inhibitors of HIV-1 reverse transcriptase*. *Bioorg Med Chem Lett*, 2006. **16**(3): p. 663-7.
288. Dey, F.; Caflisch, A. *Fragment-based de novo ligand design by multiobjective evolutionary optimization*. *J Chem Inf Model*, 2008. **48**(3): p. 679-90.
289. Moriaud, F.; Doppelt-Azeroual, O.; Martin, L.; et al. *Computational fragment-based approach at PDB scale by protein local similarity*. *J Chem Inf Model*, 2009. **49**(2): p. 280-94.

290. Pearce, B.C.; Langley, D.R.; Kang, J.; et al. *E-novo: an automated workflow for efficient structure-based lead optimization*. J Chem Inf Model, 2009. **49**(7): p. 1797-809.
291. Nisius, B.; Rester, U. *Fragment shuffling: an automated workflow for three-dimensional fragment-based ligand design*. J Chem Inf Model, 2009. **49**(5): p. 1211-22.
292. Hecht, D.; Fogel, G.B. *A novel in silico approach to drug discovery via computational intelligence*. J Chem Inf Model, 2009. **49**(4): p. 1105-21.
293. Durrant, J.D.; Amaro, R.E.; McCammon, J.A. *AutoGrow: a novel algorithm for protein inhibitor design*. Chem Biol Drug Des, 2009. **73**(2): p. 168-78.
294. Nicolaou, C.A.; Apostolakis, J.; Pattichis, C.S. *De novo drug design using multiobjective evolutionary graphs*. J Chem Inf Model, 2009. **49**(2): p. 295-307.
295. Huang, Q.; Li, L.L.; Yang, S.Y. *PhDD: a new pharmacophore-based de novo design method of drug-like molecules combined with assessment of synthetic accessibility*. J Mol Graph Model, 2010. **28**(8): p. 775-87.
296. Ishchenko, A.; Liu, Z.; Lindblom, P.; et al. *Structure-based design technology contour and its application to the design of renin inhibitors*. J Chem Inf Model, 2012. **52**(8): p. 2089-97.
297. Schneider, G.; Fechner, U. *Computer-based de novo design of drug-like molecules*. Nat Rev Drug Discov, 2005. **4**(8): p. 649-63.
298. Reddy, A.S.C., L.; Zhang, S. *Structure-Based De Novo Drug Design*, in *De novo molecular design*, G. Schneider, Editor. 2013. p. 480.
299. 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 Lond B Biol Sci, 1989. **236**(1283): p. 101-13.
300. Kitchen, D.B.; Decornez, H.; Furr, J.R.; et al. *Docking and scoring in virtual screening for drug discovery: methods and applications*. Nat Rev Drug Discov, 2004. **3**(11): p. 935-49.
301. Barril, X.; Soliva, R. *Molecular modelling*. Mol Biosyst, 2006. **2**(12): p. 660-81.
302. Fischer, E. *Einfluss der Configuration auf die Wirkung der Enzyme*. Ber. Dtsch. Chem. Ges., 1894. **27**: p. 2985-2993.
303. McGann, M. *FRED pose prediction and virtual screening accuracy*. J Chem Inf Model, 2011. **51**(3): p. 578-96.
304. Miteva, M.A.; Lee, W.H.; Montes, M.O.; et al. *Fast structure-based virtual ligand screening combining FRED, DOCK, and Surflex*. J Med Chem, 2005. **48**(19): p. 6012-22.
305. Sousa, S.F.; Fernandes, P.A.; Ramos, M.J. *Protein-ligand docking: current status and future challenges*. Proteins, 2006. **65**(1): p. 15-26.
306. Yuriev, E.; Agostino, M.; Ramsland, P.A. *Challenges and advances in computational docking: 2009 in review*. J Mol Recognit, 2011. **24**(2): p. 149-64.
307. Brooijmans, N.; Kuntz, I.D. *Molecular recognition and docking algorithms*. Annu Rev Biophys Biomol Struct, 2003. **32**: p. 335-73.
308. Friesner, R.A.; Banks, J.L.; Murphy, R.B.; et al. *Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy*. J Med Chem, 2004. **47**(7): p. 1739-49.
309. Huang, S.Y.; Zou, X. *Advances and challenges in protein-ligand docking*. Int J Mol Sci, 2010. **11**(8): p. 3016-34.
310. Jain, A.N. *Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine*. J Med Chem, 2003. **46**(4): p. 499-511.

311. Kuntz, I.D.; Leach, A.R. *Conformational analysis of flexible ligands in macromolecular receptor sites*. Journal of Computational Chemistry, 1992. **13**: p. 730-748.
312. Rarey, M.; Kramer, B.; Lengauer, T.; et al. *A fast flexible docking method using an incremental construction algorithm*. J Mol Biol, 1996. **261**(3): p. 470-89.
313. Lang, P.T.; Brozell, S.R.; Mukherjee, S.; et al. *DOCK 6: combining techniques to model RNA-small molecule complexes*. RNA, 2009. **15**(6): p. 1219-30.
314. Welch, W.; Ruppert, J.; Jain, A.N. *Hammerhead: fast, fully automated docking of flexible ligands to protein binding sites*. Chem Biol, 1996. **3**(6): p. 449-62.
315. Jain, A.N. *Morphological similarity: a 3D molecular similarity method correlated with protein-ligand recognition*. J Comput Aided Mol Des, 2000. **14**(2): p. 199-213.
316. Metropolis, N.; Ulam, S. *The Monte Carlo method*. J Am Stat Assoc, 1949. **44**(247): p. 335-41.
317. Abagyan, R.; Totrov, M.; Kuznetov, D. *ICM - A new method for protein modeling and design: Applications to docking and structure prediction from the distorted native conformation*. Journal of Computational Chemistry, 1994. **15**: p. 488-506.
318. McMartin, C.; Bohacek, R.S. *QXP: powerful, rapid computer algorithms for structure-based drug design*. J Comput Aided Mol Des, 1997. **11**(4): p. 333-44.
319. Trosset, J.Y.; Scheraga, H.A. *Prodock: Software package for protein modeling and docking*. Journal of Computational Chemistry, 1999. **20**: p. 412-427.
320. Liu, M.; Wang, S. *MCDOCK: a Monte Carlo simulation approach to the molecular docking problem*. J Comput Aided Mol Des, 1999. **13**(5): p. 435-51.
321. Darwin, C. *On the Origin of Species*. Harvard Univ. Press, Cambridge, Massachusetts ed. 1859.
322. Jones, G.; Willett, P.; Glen, R.C.; et al. *Development and validation of a genetic algorithm for flexible docking*. J Mol Biol, 1997. **267**(3): p. 727-48.
323. Morris, G.M.; Goodsell, D.S.; Halliday, R.S.; et al. *Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function*. Journal of Computational Chemistry, 1998. **19**: p. 1639-1662.
324. Baxter, C.A.; Murray, C.W.; Clark, D.E.; et al. *Flexible docking using Tabu search and an empirical estimate of binding affinity*. Proteins, 1998. **33**(3): p. 367-82.
325. Pei, J.; Wang, Q.; Liu, Z.; et al. *PSI-DOCK: towards highly efficient and accurate flexible ligand docking*. Proteins, 2006. **62**(4): p. 934-46.
326. Chen, H.M.; Liu, B.F.; Huang, H.L.; et al. *SODOCK: swarm optimization for highly flexible protein-ligand docking*. J Comput Chem, 2007. **28**(2): p. 612-23.
327. Chen, K.; Li, T.; Cao, T. *Tribe-PSO: A novel global optimization algorithm and its application in molecular docking*. Chemom. Intell. Lab. Systems, 2006. **82**: p. 248-259.
328. Namasivayam, V.; Gunther, R. *pso@autodock: a fast flexible molecular docking program based on Swarm intelligence*. Chem Biol Drug Des, 2007. **70**(6): p. 475-84.
329. Liu, Y.; Zhao, L.; Li, W.; et al. *FIPSDock: a new molecular docking technique driven by fully informed swarm optimization algorithm*. J Comput Chem, 2013. **34**(1): p. 67-75.
330. Di Nola, A.; Roccatano, D.; Berendsen, H.J. *Molecular dynamics simulation of the docking of substrates to proteins*. Proteins, 1994. **19**(3): p. 174-82.
331. Nakajima, N.N., H.; Kidera, A. *Multicanonical Ensemble Generated by Molecular Dynamics Simulation for Enhanced Conformational Sampling of Peptides*. Journal of Physical Chemistry, 1997. **101**: p. 817-824.
332. Dias, R.; de Azevedo, W.F., Jr. *Molecular docking algorithms*. Curr Drug Targets, 2008. **9**(12): p. 1040-7.

333. Pearlman, D.A.; Charifson, P.S. *Are free energy calculations useful in practice? A comparison with rapid scoring functions for the p38 MAP kinase protein system.* J Med Chem, 2001. **44**(21): p. 3417-23.
334. Pauling, L.; Delbruck, M. *The Nature of the Intermolecular Forces Operative in Biological Processes.* Science, 1940. **92**(2378): p. 77-9.
335. Kauzmann, W. *Some factors in the interpretation of protein denaturation.* Adv Protein Chem, 1959. **14**: p. 1-63.
336. Halperin, I.; Ma, B.; Wolfson, H.; et al. *Principles of docking: An overview of search algorithms and a guide to scoring functions.* Proteins, 2002. **47**(4): p. 409-43.
337. Weiner, P.K.; Kollman, P.A. *AMBER: Assisted model building with energy refinement. A general program for modeling molecules and their interactions.* Journal of Computational Chemistry, 1981. **2**: p. 287-303.
338. Weiner, S.J.; Kollman, P.A.; Case, D.A.; et al. *A new force field for molecular mechanical simulation of nucleic acids and proteins.* Journal of American Chemical Society, 1984. **106**: p. 765-784.
339. Verdonk, M.L.; Cole, J.C.; Hartshorn, M.J.; et al. *Improved protein-ligand docking using GOLD.* Proteins, 2003. **52**(4): p. 609-23.
340. Nemethy, G.; Gibson, K.D.; Palmer, K.A.; 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.* Journal of Physical Chemistry, 1992. **96**: p. 6472-6484.
341. Rocchia, W.; Sridharan, S.; Nicholls, A.; et al. *Rapid grid-based construction of the molecular surface and the use of induced surface charge to calculate reaction field energies: applications to the molecular systems and geometric objects.* J Comput Chem, 2002. **23**(1): p. 128-37.
342. Liu, H.; Kuntz, I.D.; Zou, X. *Pairwise GB/SA Scoring Function for Structure-based Drug Design.* The Journal of Physical Chemistry B, 2004. **108**: p. 5453-5462.
343. Huang, S.Y.; Grinter, S.Z.; Zou, X. *Scoring functions and their evaluation methods for protein-ligand docking: recent advances and future directions.* Phys Chem Chem Phys, 2010. **12**(40): p. 12899-908.
344. Eldridge, M.D.; Murray, C.W.; Auton, T.R.; et al. *Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes.* J Comput Aided Mol Des, 1997. **11**(5): p. 425-45.
345. Rognan, D.; Lauemoller, S.L.; Holm, A.; et al. *Predicting binding affinities of protein ligands from three-dimensional models: application to peptide binding to class I major histocompatibility proteins.* J Med Chem, 1999. **42**(22): p. 4650-8.
346. Jain, A.N. *Surflex-Dock 2.1: robust performance from ligand energetic modeling, ring flexibility, and knowledge-based search.* J Comput Aided Mol Des, 2007. **21**(5): p. 281-306.
347. Pham, T.A.; Jain, A.N. *Customizing scoring functions for docking.* J Comput Aided Mol Des, 2008. **22**(5): p. 269-86.
348. Muegge, I. *PMF scoring revisited.* J Med Chem, 2006. **49**(20): p. 5895-902.
349. Yang, C.Y.; Wang, R.; Wang, S. *M-score: a knowledge-based potential scoring function accounting for protein atom mobility.* J Med Chem, 2006. **49**(20): p. 5903-11.
350. Leach, A.R.; Shoichet, B.K.; Peishoff, C.E. *Prediction of protein-ligand interactions. Docking and scoring: successes and gaps.* J Med Chem, 2006. **49**(20): p. 5851-5.
351. Muegge, I.; Martin, Y.C. *A general and fast scoring function for protein-ligand interactions: a simplified potential approach.* J Med Chem, 1999. **42**(5): p. 791-804.
352. Gohlke, H.; Hendlich, M.; Klebe, G. *Knowledge-based scoring function to predict protein-ligand interactions.* J Mol Biol, 2000. **295**(2): p. 337-56.

353. Bar-Haim, S.; Aharon, A.; Ben-Moshe, T.; et al. *SeleX-CS: a new consensus scoring algorithm for hit discovery and lead optimization*. J Chem Inf Model, 2009. **49**(3): p. 623-33.
354. Wang, R.; Wang, S. *How does consensus scoring work for virtual library screening? An idealized computer experiment*. J Chem Inf Comput Sci, 2001. **41**(5): p. 1422-6.
355. Bissantz, C.; Folkers, G.; Rognan, D. *Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations*. J Med Chem, 2000. **43**(25): p. 4759-67.
356. Charifson, P.S.; Corkery, J.J.; Murcko, M.A.; et al. *Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins*. J Med Chem, 1999. **42**(25): p. 5100-9.
357. Clark, R.D.; Strizhev, A.; Leonard, J.M.; et al. *Consensus scoring for ligand/protein interactions*. J Mol Graph Model, 2002. **20**(4): p. 281-95.
358. Stahl, M.; Rarey, M. *Detailed analysis of scoring functions for virtual screening*. J Med Chem, 2001. **44**(7): p. 1035-42.
359. Rognan, D. *Le criblage virtuel par docking moléculaire*, in *Des petites molécules pour explorer le vivant*, E.R. Maréchal, S.; Lafanechère, L., Editor. 2007. p. 258.
360. Jain, A.N. *Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine*. J Med Chem, 2003. **46**(4): p. 499-511.
361. Yuriev, E.; Ramsland, P.A. *Latest developments in molecular docking: 2010-2011 in review*. J Mol Recognit, 2013. **26**(5): p. 215-39.
362. PDB: Protein Data Bank. Available from: www.rcsb.org.
363. Thomas, M.P.; McInnes, C.; Fischer, P.M. *Protein structures in virtual screening: a case study with CDK2*. J Med Chem, 2006. **49**(1): p. 92-104.
364. Hawkins, P.C.; Warren, G.L.; Skillman, A.G.; et al. *How to do an evaluation: pitfalls and traps*. J Comput Aided Mol Des, 2008. **22**(3-4): p. 179-90.
365. McGovern, S.L.; Shoichet, B.K. *Information decay in molecular docking screens against holo, apo, and modeled conformations of enzymes*. J Med Chem, 2003. **46**(14): p. 2895-907.
366. Kirchmair, J.; Distinto, S.; Schuster, D.; et al. *Enhancing drug discovery through in silico screening: strategies to increase true positives retrieval rates*. Curr Med Chem, 2008. **15**(20): p. 2040-53.
367. Giganti, D.; Guillemain, H.; Spadoni, J.L.; et al. *Comparative evaluation of 3D virtual ligand screening methods: impact of the molecular alignment on enrichment*. J Chem Inf Model, 2010. **50**(6): p. 992-1004.
368. Ben Nasr, N.; Guillemain, H.; Lagarde, N.; et al. *Multiple structures for virtual ligand screening: defining binding site properties-based criteria to optimize the selection of the query*. J Chem Inf Model, 2013. **53**(2): p. 293-311.
369. Liebeschuetz, J.W. *Evaluating docking programs: keeping the playing field level*. J Comput Aided Mol Des, 2008. **22**(3-4): p. 229-38.
370. Tuccinardi, T. *Docking-based virtual screening: recent developments*. Comb Chem High Throughput Screen, 2009. **12**(3): p. 303-14.
371. Poornima, C.S.; Dean, P.M. *Hydration in drug design. 1. Multiple hydrogen-bonding features of water molecules in mediating protein-ligand interactions*. J Comput Aided Mol Des, 1995. **9**(6): p. 500-12.
372. Levinson, N.M.; Boxer, S.G. *A conserved water-mediated hydrogen bond network defines bosutinib's kinase selectivity*. Nat Chem Biol, 2014. **10**(2): p. 127-32.
373. Sondergaard, C.R.; Garrett, A.E.; Carstensen, T.; et al. *Structural artifacts in protein-ligand X-ray structures: implications for the development of docking scoring functions*. J Med Chem, 2009. **52**(18): p. 5673-84.

374. 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, 1996. **3**(12): p. 973-80.
375. Schnecke, V.; Kuhn, L. *Virtual screening with solvation and ligand-induced complementary.* Perspectives in Drug Discovery and Design, 2000. **20**: p. 171-190.
376. Rarey, M.; Kramer, B.; Lengauer, T. *The particle concept: placing discrete water molecules during protein-ligand docking predictions.* Proteins, 1999. **34**(1): p. 17-28.
377. Carlson, H.A. *Protein flexibility and drug design: how to hit a moving target.* Curr Opin Chem Biol, 2002. **6**(4): p. 447-52.
378. Durrant, J.D.; McCammon, J.A. *Computer-aided drug-discovery techniques that account for receptor flexibility.* Curr Opin Pharmacol, 2010. **10**(6): p. 770-4.
379. Cavasotto, C.N.; Abagyan, R.A. *Protein flexibility in ligand docking and virtual screening to protein kinases.* J Mol Biol, 2004. **337**(1): p. 209-25.
380. Claussen, H.; Buning, C.; Rarey, M.; et al. *FlexE: efficient molecular docking considering protein structure variations.* J Mol Biol, 2001. **308**(2): p. 377-95.
381. Leach, A.R. *Ligand docking to proteins with discrete side-chain flexibility.* J Mol Biol, 1994. **235**(1): p. 345-56.
382. Jiang, F.; Kim, S.H. *"Soft docking": matching of molecular surface cubes.* J Mol Biol, 1991. **219**(1): p. 79-102.
383. Alex, A.A.; Millan, D.S. *Contribution of Structure-Based Drug Design to the Discovery of Marketed Drugs,* in *Drug Design Strategies*, D.D. Livingstone, A.M., Editor. 2009. p. 498p.
384. Liebeschuetz, J.W.; Jones, S.D.; Morgan, P.J.; et al. *PRO_SELECT: combining structure-based drug design and array-based chemistry for rapid lead discovery. 2. The development of a series of highly potent and selective factor Xa inhibitors.* J Med Chem, 2002. **45**(6): p. 1221-32.
385. Baldwin, J.J.; Ponticello, G.S.; Anderson, P.S.; et al. *Thienothiopyran-2-sulfonamides: novel topically active carbonic anhydrase inhibitors for the treatment of glaucoma.* J Med Chem, 1989. **32**(12): p. 2510-3.
386. Greer, J.; Erickson, J.W.; Baldwin, J.J.; et al. *Application of the three-dimensional structures of protein target molecules in structure-based drug design.* J Med Chem, 1994. **37**(8): p. 1035-54.
387. Graves, B.J.; Hatada, M.H.; Miller, J.K.; et al. *The three-dimensional x-ray crystal structure of HIV-1 protease complexed with a hydroxyethylene inhibitor.* Adv Exp Med Biol, 1991. **306**: p. 455-60.
388. Krohn, A.; Redshaw, S.; Ritchie, J.C.; et al. *Novel binding mode of highly potent HIV-proteinase inhibitors incorporating the (R)-hydroxyethylamine isostere.* J Med Chem, 1991. **34**(11): p. 3340-2.
389. Kempf, D.J.; Sham, H.L.; Marsh, K.C.; et al. *Discovery of ritonavir, a potent inhibitor of HIV protease with high oral bioavailability and clinical efficacy.* J Med Chem, 1998. **41**(4): p. 602-17.
390. Lin, J.H. *Role of pharmacokinetics in the discovery and development of indinavir.* Adv Drug Deliv Rev, 1999. **39**(1-3): p. 33-49.
391. Gershell, L.J.; Atkins, J.H. *A brief history of novel drug discovery technologies.* Nat Rev Drug Discov, 2003. **2**(4): p. 321-7.
392. von Itzstein, M.; Wu, W.Y.; Kok, G.B.; et al. *Rational design of potent sialidase-based inhibitors of influenza virus replication.* Nature, 1993. **363**(6428): p. 418-23.
393. Bossart-Whitaker, P.; Carson, M.; Babu, Y.S.; et al. *Three-dimensional structure of influenza A N9 neuraminidase and its complex with the inhibitor 2-deoxy 2,3-dehydro-N-acetyl neuraminic acid.* J Mol Biol, 1993. **232**(4): p. 1069-83.

394. Woods, J.M.; Bethell, R.C.; Coates, J.A.; et al. *4-Guanidino-2,4-dideoxy-2,3-dehydro-N-acetylneuraminic acid is a highly effective inhibitor both of the sialidase (neuraminidase) and of growth of a wide range of influenza A and B viruses in vitro*. *Antimicrob Agents Chemother*, 1993. **37**(7): p. 1473-9.
395. Kim, C.U.; Lew, W.; Williams, M.A.; et al. *Influenza neuraminidase inhibitors possessing a novel hydrophobic interaction in the enzyme active site: design, synthesis, and structural analysis of carbocyclic sialic acid analogues with potent anti-influenza activity*. *J Am Chem Soc*, 1997. **119**(4): p. 681-90.
396. Raja, A.; Lebbos, J.; Kirkpatrick, P. *Atazanavir sulphate*. *Nat Rev Drug Discov*, 2003. **2**(11): p. 857-8.
397. Wire, M.B.; Shelton, M.J.; Studenberg, S. *Fosamprenavir : clinical pharmacokinetics and drug interactions of the amprenavir prodrug*. *Clin Pharmacokinet*, 2006. **45**(2): p. 137-68.
398. Gustafsson, D.; Bylund, R.; Antonsson, T.; et al. *A new oral anticoagulant: the 50-year challenge*. *Nat Rev Drug Discov*, 2004. **3**(8): p. 649-59.
399. Thaisrivongs, S.; Watenpaugh, K.D.; Howe, W.J.; et al. *Structure-based design of novel HIV protease inhibitors: carboxamide-containing 4-hydroxycoumarins and 4-hydroxy-2-pyrones as potent nonpeptidic inhibitors*. *J Med Chem*, 1995. **38**(18): p. 3624-37.
400. Turner, S.R.; Strohbach, J.W.; Tommasi, R.A.; 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*, 1998. **41**(18): p. 3467-76.
401. Sun, C.L.; Christensen, J.G.; McMahon, G. *Discovery and Development of Sunitinib (SU11248): A Multitarget Tyrosine Kinase Inhibitor of Tumor, Growth, Survival and Angiogenesis*, in *Kinase Inhibitor Drugs*, R.S. Li, J.A., Editor. 2009. p. 510.
402. de Béthude, M.P.; Sekar, V.; Spinosa-Guzman, S.; et al. *Darunavir (Prezista, TMC114): From Bench to Clinic, Improving Treatment Options for HIV-Infected Patients*, in *Antiviral Drugs: From Basic Discovery through Clinical Trials*, W.M. Kazmierski, Editor. 2011. p. 438.
403. Blay, J.Y.; von Mehren, M. *Nilotinib: a novel, selective tyrosine kinase inhibitor*. *Semin Oncol*, 2011. **38 Suppl 1**: p. S3-9.
404. Goschke, R.; Cohen, N.C.; Wood, J.M.; et al. *Design and synthesis of novel 2,7-dialkyl substituted 5(S)-amino-4(S)-hydroxy- 8-phenyl-octanecarboxamides as in vitro potent peptidomimetic inhibitors of human renin*. *Bioorg. Med. Chem. Lett.*, 1997. **7**: p. 2735-2740.
405. van Ryn, J.; Goss, A.; Huel, N.; et al. *The discovery of dabigatran etexilate*. *Front Pharmacol*, 2013. **4**: p. 12.
406. Harris, P.A.; Stafford, J.A. *Discovery of Pazopanib: A Pan Vascular Endothelial Growth Factor Kinase Inhibitor*, in *Kinase Inhibitor Drugs*, R.S. Li, J.A., Editor. 2009. p. 510.
407. Njoroge, F.G.; Chen, K.X.; Shih, N.Y.; et al. *Challenges in modern drug discovery: a case study of boceprevir, an HCV protease inhibitor for the treatment of hepatitis C virus infection*. *Acc Chem Res*, 2008. **41**(1): p. 50-9.
408. Irwin, J.J. *Community benchmarks for virtual screening*. *J Comput Aided Mol Des*, 2008. **22**(3-4): p. 193-9.
409. Kirchmair, J.; Markt, P.; Distinto, S.; et al. *Evaluation of the performance of 3D virtual screening protocols: RMSD comparisons, enrichment assessments, and decoy selection--what can we learn from earlier mistakes?* *J Comput Aided Mol Des*, 2008. **22**(3-4): p. 213-28.

410. Kroemer, R.T.; Vulpetti, A.; McDonald, J.J.; et al. *Assessment of docking poses: interactions-based accuracy classification (IBAC) versus crystal structure deviations*. J Chem Inf Comput Sci, 2004. **44**(3): p. 871-81.
411. Abagyan, R.A.; Totrov, M.M. *Contact area difference (CAD): a robust measure to evaluate accuracy of protein models*. J Mol Biol, 1997. **268**(3): p. 678-85.
412. Yusuf, D.; Davis, A.M.; Kleywegt, G.J.; et al. *An alternative method for the evaluation of docking performance: RSR vs RMSD*. J Chem Inf Model, 2008. **48**(7): p. 1411-22.
413. Pham, T.A.; Jain, A.N. *Parameter estimation for scoring protein-ligand interactions using negative training data*. J Med Chem, 2006. **49**(20): p. 5856-68.
414. Perkins, E.; Sun, D.; Nguyen, A.; et al. *Novel inhibitors of poly(ADP-ribose) polymerase/PARP1 and PARP2 identified using a cell-based screen in yeast*. Cancer Res, 2001. **61**(10): p. 4175-83.
415. Doman, T.N.; McGovern, S.L.; Witherbee, B.J.; et al. *Molecular docking and high-throughput screening for novel inhibitors of protein tyrosine phosphatase-1B*. J Med Chem, 2002. **45**(11): p. 2213-21.
416. Wang, R.; Fang, X.; Lu, Y.; et al. *The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures*. J Med Chem, 2004. **47**(12): p. 2977-80.
417. Perola, E.; Walters, W.P.; Charifson, P.S. *A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance*. Proteins, 2004. **56**(2): p. 235-49.
418. Montes, M.; Miteva, M.A.; Villoutreix, B.O. *Structure-based virtual ligand screening with LigandFit: pose prediction and enrichment of compound collections*. Proteins, 2007. **68**(3): p. 712-25.
419. Diller, D.J.; Li, R. *Kinases, homology models, and high throughput docking*. J Med Chem, 2003. **46**(22): p. 4638-47.
420. Lorber, D.M.; Shoichet, B.K. *Hierarchical docking of databases of multiple ligand conformations*. Curr Top Med Chem, 2005. **5**(8): p. 739-49.
421. Irwin, J.J.; Raushel, F.M.; Shoichet, B.K. *Virtual screening against metalloenzymes for inhibitors and substrates*. Biochemistry, 2005. **44**(37): p. 12316-28.
422. Tiikkainen, P.; Bellis, L.; Light, Y.; et al. *Estimating error rates in bioactivity databases*. J Chem Inf Model, 2013. **53**(10): p. 2499-505.
423. Huang, N.; Shoichet, B.K.; Irwin, J.J. *Benchmarking sets for molecular docking*. J Med Chem, 2006. **49**(23): p. 6789-801.
424. Good, A.C.; Oprea, T.I. *Optimization of CAMD techniques 3. Virtual screening enrichment studies: a help or hindrance in tool selection?* J Comput Aided Mol Des, 2008. **22**(3-4): p. 169-78.
425. Mysinger, M.M.; Shoichet, B.K. *Rapid context-dependent ligand desolvation in molecular docking*. J Chem Inf Model, 2010. **50**(9): p. 1561-73.
426. Wallach, I.; Lilien, R. *Virtual decoy sets for molecular docking benchmarks*. J Chem Inf Model, 2011. **51**(2): p. 196-202.
427. Mysinger, M.M.; Carchia, M.; Irwin, J.J.; et al. *Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking*. J Med Chem, 2012. **55**(14): p. 6582-94.
428. Truchon, J.F.; Bayly, C.I. *Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem*. J Chem Inf Model, 2007. **47**(2): p. 488-508.
429. Triballeau, N.; Acher, F.; Brabet, I.; et al. *Virtual screening workflow development guided by the "receiver operating characteristic" curve approach. Application to high-*

- throughput docking on metabotropic glutamate receptor subtype 4.* J Med Chem, 2005. **48**(7): p. 2534-47.
430. Sheridan, R.P.; Singh, S.B.; Fluder, E.M.; et al. *Protocols for bridging the peptide to nonpeptide gap in topological similarity searches.* J Chem Inf Comput Sci, 2001. **41**(5): p. 1395-406.
431. Barril, X.; Morley, S.D. *Unveiling the full potential of flexible receptor docking using multiple crystallographic structures.* J Med Chem, 2005. **48**(13): p. 4432-43.
432. Craig, I.R.; Essex, J.W.; Spiegel, K. *Ensemble docking into multiple crystallographically derived protein structures: an evaluation based on the statistical analysis of enrichments.* J Chem Inf Model, 2010. **50**(4): p. 511-24.
433. Huang, S.Y.; Zou, X. *Ensemble docking of multiple protein structures: considering protein structural variations in molecular docking.* Proteins, 2007. **66**(2): p. 399-421.
434. Bolstad, E.S.; Anderson, A.C. *In pursuit of virtual lead optimization: pruning ensembles of receptor structures for increased efficiency and accuracy during docking.* Proteins, 2009. **75**(1): p. 62-74.
435. Broughton, H.B. *A method for including protein flexibility in protein-ligand docking: improving tools for database mining and virtual screening.* J Mol Graph Model, 2000. **18**(3): p. 247-57, 302-4.
436. Cavasotto, C.N.; Kovacs, J.A.; Abagyan, R.A. *Representing receptor flexibility in ligand docking through relevant normal modes.* J Am Chem Soc, 2005. **127**(26): p. 9632-40.
437. Frimurer, T.M.; Peters, G.H.; Iversen, L.F.; et al. *Ligand-induced conformational changes: improved predictions of ligand binding conformations and affinities.* Biophys J, 2003. **84**(4): p. 2273-81.
438. Sperandio, O.; Mouawad, L.; Pinto, E.; et al. *How to choose relevant multiple receptor conformations for virtual screening: a test case of Cdk2 and normal mode analysis.* Eur Biophys J, 2010. **39**(9): p. 1365-72.
439. Yoon, S.; Welsh, W.J. *Identification of a minimal subset of receptor conformations for improved multiple conformation docking and two-step scoring.* J Chem Inf Comput Sci, 2004. **44**(1): p. 88-96.
440. Rueda, M.; Bottegoni, G.; Abagyan, R. *Recipes for the selection of experimental protein conformations for virtual screening.* J Chem Inf Model, 2010. **50**(1): p. 186-93.
441. Rao, S.; Sanschagrin, P.C.; Greenwood, J.R.; et al. *Improving database enrichment through ensemble docking.* J Comput Aided Mol Des, 2008. **22**(9): p. 621-7.
442. Bottegoni, G.; Rocchia, W.; Rueda, M.; et al. *Systematic exploitation of multiple receptor conformations for virtual ligand screening.* PLoS One, 2011. **6**(5): p. e18845.
443. Rueda, M.; Totrov, M.; Abagyan, R. *ALiBERO: evolving a team of complementary pocket conformations rather than a single leader.* J Chem Inf Model, 2012. **52**(10): p. 2705-14.
444. Birch, L.; Murray, C.W.; Hartshorn, M.J.; et al. *Sensitivity of molecular docking to induced fit effects in influenza virus neuraminidase.* J Comput Aided Mol Des, 2002. **16**(12): p. 855-69.
445. Zhang, J.; Aizawa, M.; Amari, S.; et al. *Development of KiBank, a database supporting structure-based drug design.* Comput Biol Chem, 2004. **28**(5-6): p. 401-7.
446. Wang, R.; Fang, X.; Lu, Y.; et al. *The PDBbind database: methodologies and updates.* J Med Chem, 2005. **48**(12): p. 4111-9.
447. Fang, H.; Tong, W.; Shi, L.M.; et al. *Structure-activity relationships for a large diverse set of natural, synthetic, and environmental estrogens.* Chem Res Toxicol, 2001. **14**(3): p. 280-94.

448. Jorissen, R.N.; Gilson, M.K. *Virtual screening of molecular databases using a support vector machine*. J Chem Inf Model, 2005. **45**(3): p. 549-61.
449. Wright, L.; Barril, X.; Dymock, B.; et al. *Structure-activity relationships in purine-based inhibitor binding to HSP90 isoforms*. Chem Biol, 2004. **11**(6): p. 775-85.
450. Dymock, B.W.; Barril, X.; Brough, P.A.; et al. *Novel, potent small-molecule inhibitors of the molecular chaperone Hsp90 discovered through structure-based design*. J Med Chem, 2005. **48**(13): p. 4212-5.
451. Hennequin, L.F.; Thomas, A.P.; Johnstone, C.; et al. *Design and structure-activity relationship of a new class of potent VEGF receptor tyrosine kinase inhibitors*. J Med Chem, 1999. **42**(26): p. 5369-89.
452. Hennequin, L.F.; Stokes, E.S.; Thomas, A.P.; et al. *Novel 4-anilinoquinazolines with C-7 basic side chains: design and structure activity relationship of a series of potent, orally active, VEGF receptor tyrosine kinase inhibitors*. J Med Chem, 2002. **45**(6): p. 1300-12.
453. Sun, L.; Tran, N.; Liang, C.; et al. *Design, synthesis, and evaluations of substituted 3-[(3- or 4-carboxyethylpyrrol-2-yl)methylidene]indolin-2-ones as inhibitors of VEGF, FGF, and PDGF receptor tyrosine kinases*. J Med Chem, 1999. **42**(25): p. 5120-30.
454. Jacobsson, M.; Liden, P.; Stjernschantz, E.; et al. *Improving structure-based virtual screening by multivariate analysis of scoring data*. J Med Chem, 2003. **46**(26): p. 5781-9.
455. Bohm, M.; St rzebecher, J.; Klebe, G. *Three-dimensional quantitative structure-activity relationship analyses using comparative molecular field analysis and comparative molecular similarity indices analysis to elucidate selectivity differences of inhibitors binding to trypsin, thrombin, and factor Xa*. J Med Chem, 1999. **42**(3): p. 458-77.
456. Sutherland, J.J.; O'Brien, L.A.; Weaver, D.F. *A comparison of methods for modeling quantitative structure-activity relationships*. J Med Chem, 2004. **47**(22): p. 5541-54.
457. Varney, M.D.; Palmer, C.L.; Romines, W.H., 3rd; et al. *Protein structure-based design, synthesis, and biological evaluation of 5-thia-2,6-diamino-4(3H)-oxypyrimidines: potent inhibitors of glycinamide ribonucleotide transformylase with potent cell growth inhibition*. J Med Chem, 1997. **40**(16): p. 2502-24.
458. Van Zandt, M.C.; Jones, M.L.; Gunn, D.E.; et al. *Discovery of 3-[(4,5,7-trifluorobenzothiazol-2-yl)methyl]indole-N-acetic acid (lidorestat) and congeners as highly potent and selective inhibitors of aldose reductase for treatment of chronic diabetic complications*. J Med Chem, 2005. **48**(9): p. 3141-52.
459. Powers, R.A.; Morandi, F.; Shoichet, B.K. *Structure-based discovery of a novel, noncovalent inhibitor of AmpC beta-lactamase*. Structure, 2002. **10**(7): p. 1013-23.
460. Graves, A.P.; Brenk, R.; Shoichet, B.K. *Decoys for docking*. J Med Chem, 2005. **48**(11): p. 3714-28.
461. Tondi, D.; Morandi, F.; Bonnet, R.; et al. *Structure-based optimization of a non-beta-lactam lead results in inhibitors that do not up-regulate beta-lactamase expression in cell culture*. J Am Chem Soc, 2005. **127**(13): p. 4632-9.
462. Wang, J.; Kang, X.; Kuntz, I.D.; et al. *Hierarchical database screenings for HIV-1 reverse transcriptase using a pharmacophore model, rigid docking, solvation docking, and MM-PB/SA*. J Med Chem, 2005. **48**(7): p. 2432-44.
463. Tikhe, J.G.; Webber, S.E.; Hostomsky, Z.; et al. *Design, synthesis, and evaluation of 3,4-dihydro-2H-[1,4]diazepino[6,7,1-hi]indol-1-ones as inhibitors of poly(ADP-ribose) polymerase*. J Med Chem, 2004. **47**(22): p. 5467-81.

464. Ealick, S.E.; Babu, Y.S.; Bugg, C.E.; et al. *Application of crystallographic and modeling methods in the design of purine nucleoside phosphorylase inhibitors*. Proc Natl Acad Sci U S A, 1991. **88**(24): p. 11540-4.
465. Oikonomakos, N.G.; Tsitsanou, K.E.; Zographos, S.E.; et al. *Allosteric inhibition of glycogen phosphorylase a by the potential antidiabetic drug 3-isopropyl 4-(2-chlorophenyl)-1,4-dihydro-1-ethyl-2-methyl-pyridine-3,5,6-tricarboxylate*. Protein Sci, 1999. **8**(10): p. 1930-45.
466. Oikonomakos, N.G.; Zographos, S.E.; Skamnaki, V.T.; et al. *The 1.76 Å resolution crystal structure of glycogen phosphorylase B complexed with glucose, and CP320626, a potential antidiabetic drug*. Bioorg Med Chem, 2002. **10**(5): p. 1313-9.
467. Hollenberg, S.M.; Weinberger, C.; Ong, E.S.; et al. *Primary structure and expression of a functional human glucocorticoid receptor cDNA*. Nature, 1985. **318**(6047): p. 635-41.
468. Green, S.; Walter, P.; Kumar, V.; et al. *Human oestrogen receptor cDNA: sequence, expression and homology to v-erb-A*. Nature, 1986. **320**(6058): p. 134-9.
469. Greene, G.L.; Gilna, P.; Waterfield, M.; et al. *Sequence and expression of human estrogen receptor complementary DNA*. Science, 1986. **231**(4742): p. 1150-4.
470. Imai, Y.; Youn, M.Y.; Inoue, K.; et al. *Nuclear receptors in bone physiology and diseases*. Physiol Rev, 2013. **93**(2): p. 481-523.
471. Park, S.-J.; Kufareva, I.; Abagyan, R. *Improved docking, screening and selectivity prediction for small molecule nuclear receptor modulators using conformational ensembles*. J Comput Aided Mol Des, 2010. **24**(5): p. 459-71.
472. Via, M. *Nuclear Receptors: The Pipeline Outlook*. Cambridge Healthtech Institute ed. 2010.
473. Sladek, F.M. *Nuclear receptors as drug targets: new developments in coregulators, orphan receptors and major therapeutic areas*. Expert Opin Ther Targets, 2003. **7**(5): p. 679-84.
474. Margolis, R.N.; Evans, R.M.; O'Malley, B.W.; et al. *The Nuclear Receptor Signaling Atlas: development of a functional atlas of nuclear receptors*. Mol Endocrinol, 2005. **19**(10): p. 2433-6.
475. McKenna, N.J.; Cooney, A.J.; DeMayo, F.J.; et al. *Minireview: Evolution of NURSA, the Nuclear Receptor Signaling Atlas*. Mol Endocrinol, 2009. **23**(6): p. 740-6.
476. Sharman, J.L.; Mpamhanga, C.P. *IUPHAR-DB: an open-access, expert-curated resource for receptor and ion channel research*. ACS Chem Neurosci, 2011. **2**(5): p. 232-5.
477. Vroling, B.; Thorne, D.; McDermott, P.; et al. *NuclearRDB: information system for nuclear receptors*. Nucleic Acids Res, 2012. **40**(Database issue): p. D377-80.
478. Duarte, J.; Perriere, G.; Laudet, V.; et al. *NUREBASE: database of nuclear hormone receptors*. Nucleic Acids Res, 2002. **30**(1): p. 364-8.
479. Van Durme, J.J.; Bettler, E.; Folkertsma, S.; et al. *NRMD: Nuclear Receptor Mutation Database*. Nucleic Acids Res, 2003. **31**(1): p. 331-3.
480. Bourguet, W.; Germain, P.; Gronemeyer, H. *Nuclear receptor ligand-binding domains: three-dimensional structures, molecular interactions and pharmacological implications*. Trends Pharmacol Sci, 2000. **21** (10): p. 381-8.
481. Gatica, E.A.; Cavasotto, C.N. *Ligand and decoy sets for docking to G protein-coupled receptors*. J Chem Inf Model, 2012. **52**(1): p. 1-6.
482. Talete, s., *Dragon (Software for Molecular Descriptor Calculation)Version 6.0 - 2013* %W <http://www.talete.mi.it>.

483. Togashi, M.; Borngraeber, S.; Sandler, B.; et al. *Conformational adaptation of nuclear receptor ligand binding domains to agonists: potential for novel approaches to ligand design*. *J Steroid Biochem Mol Biol*, 2005. **93**(2-5): p. 127-37.
484. Schapira, M.; Raaka, B.M.; Samuels, H.H.; et al. *Rational discovery of novel nuclear hormone receptor antagonists*. *Proc Natl Acad Sci U S A*, 2000. **97**(3): p. 1008-13.
485. Pettersen, E.F.; Goddard, T.D.; Huang, C.C.; et al. *UCSF Chimera--a visualization system for exploratory research and analysis*. *J Comput Chem*, 2004. **25**(13): p. 1605-12.
486. Solt, L.A.; Wang, Y.; Banerjee, S.; et al. *Regulation of circadian behaviour and metabolism by synthetic REV-ERB agonists*. *Nature*, 2012. **485**(7396): p. 62-8.
487. Shi, Y.; Zhang, J.; Shi, M.; et al. *Cyanoguanidine-based lactam derivatives as a novel class of orally bioavailable factor Xa inhibitors*. *Bioorg Med Chem Lett*, 2009. **19**(15): p. 4034-41.
488. Desvergne, B.; Wahli, W. *Peroxisome proliferator-activated receptors: nuclear control of metabolism*. *Endocr Rev*, 1999. **20**(5): p. 649-88.
489. Vazquez, M.; Silvestre, J.S.; Prous, J.R. *Experimental approaches to study PPAR gamma agonists as antidiabetic drugs*. *Methods Find Exp Clin Pharmacol*, 2002. **24**(8): p. 515-23.
490. Krey, G.; Braissant, O.; L'Horsset, F.; et al. *Fatty acids, eicosanoids, and hypolipidemic agents identified as ligands of peroxisome proliferator-activated receptors by coactivator-dependent receptor ligand assay*. *Mol Endocrinol*, 1997. **11**(6): p. 779-91.
491. Straus, D.S.; Glass, C.K. *Anti-inflammatory actions of PPAR ligands: new insights on cellular and molecular mechanisms*. *Trends Immunol*, 2007. **28**(12): p. 551-8.
492. Bissantz, C.; Folkers, G.; Rognan, D. *Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations*. *J Med Chem*, 2000. **43**(25): p. 4759-67.
493. Pham, T.A.; Jain, A.N. *Parameter estimation for scoring protein-ligand interactions using negative training data*. *J Med Chem*, 2006. **49**(20): p. 5856-68.
494. Lagarde, N.; Ben Nasr, N.; Jeremie, A.; et al. *NRLiSt BDB, the manually curated nuclear receptors ligands and structures benchmarking database*. *J Med Chem*, 2014. **57**(7): p. 3117-25.
495. Ben Nasr, N.; Guillemain, H.; Lagarde, N.; et al. *Multiple structures for virtual ligand screening: defining binding site properties-based criteria to optimize the selection of the query*. *J Chem Inf Model*, 2013. **53**(2): p. 293-311.
496. McGovern, S.L.; Shoichet, B.K. *Information decay in molecular docking screens against holo, apo, and modeled conformations of enzymes*. *J Med Chem*, 2003. **46**(14): p. 2895-907.
497. Hawkins, P.C.D.; Warren, G.L.; Skillman, A.G.; et al. *How to do an evaluation: pitfalls and traps*. *J Comput Aided Mol Des*, 2008. **22**(3-4): p. 179-90.
498. Thomas, M.P.; McInnes, C.; Fischer, P.M. *Protein structures in virtual screening: a case study with CDK2*. *J Med Chem*, 2006. **49**(1): p. 92-104.
499. Smolen, J.S.; Breedveld, F.C.; Eberl, G.; et al. *Validity and reliability of the twenty-eight-joint count for the assessment of rheumatoid arthritis activity*. *Arthritis Rheum*, 1995. **38**(1): p. 38-43.
500. Gabriel, S.E. *The epidemiology of rheumatoid arthritis*. *Rheum Dis Clin North Am*, 2001. **27**(2): p. 269-81.
501. Smolen, J.S.; Steiner, G. *Therapeutic strategies for rheumatoid arthritis*. *Nat Rev Drug Discov*, 2003. **2**(6): p. 473-88.

502. Meier, F.M.; Frerix, M.; Hermann, W.; et al. *Current immunotherapy in rheumatoid arthritis*. *Immunotherapy*, 2013. **5**(9): p. 955-74.
503. He, M.M.; Smith, A.S.; Oslob, J.D.; et al. *Small-molecule inhibition of TNF-alpha*. *Science*, 2005. **310**(5750): p. 1022-5.
504. Chan, D.S.; Lee, H.M.; Yang, F.; et al. *Structure-based discovery of natural-product-like TNF-alpha inhibitors*. *Angew Chem Int Ed Engl*, 2010. **49**(16): p. 2860-4.
505. Choi, H.; Lee, Y.; Park, H.; et al. *Discovery of the inhibitors of tumor necrosis factor alpha with structure-based virtual screening*. *Bioorg Med Chem Lett*, 2010. **20**(21): p. 6195-8.
506. Leung, C.H.; Chan, D.S.; Kwan, M.H.; et al. *Structure-based repurposing of FDA-approved drugs as TNF-alpha inhibitors*. *ChemMedChem*, 2011. **6**(5): p. 765-8.
507. Kumar, K.S.; Kumar, P.M.; Kumar, K.A.; et al. *A new three-component reaction: green synthesis of novel isoindolo[2,1-a]quinazoline derivatives as potent inhibitors of TNF-alpha*. *Chem Commun (Camb)*, 2011. **47**(17): p. 5010-2.
508. Leung, C.H.; Zhong, H.J.; Yang, H.; et al. *A metal-based inhibitor of tumor necrosis factor-alpha*. *Angew Chem Int Ed Engl*, 2012. **51**(36): p. 9010-4.
509. Montes, M.; Mouhsine, H.; Guillemain, H.; et al., *In vivo orally active small molecule inhibitors of TNFa*. *Submitted*.
510. Hirano, T.; Yasukawa, K.; Harada, H.; et al. *Complementary DNA for a novel human interleukin (BSF-2) that induces B lymphocytes to produce immunoglobulin*. *Nature*, 1986. **324**(6092): p. 73-6.
511. Tanaka, T.; Kishimoto, T. *Immunotherapeutic implication of IL-6 blockade*. *Immunotherapy*, 2012. **4**(1): p. 87-105.
512. Heinrich, P.C.; Castell, J.V.; Andus, T. *Interleukin-6 and the acute phase response*. *Biochem J*, 1990. **265**(3): p. 621-36.
513. Kimura, A.; Kishimoto, T. *IL-6: regulator of Treg/Th17 balance*. *Eur J Immunol*, 2010. **40**(7): p. 1830-5.
514. Okada, M.; Kitahara, M.; Kishimoto, S.; et al. *IL-6/BSF-2 functions as a killer helper factor in the in vitro induction of cytotoxic T cells*. *J Immunol*, 1988. **141**(5): p. 1543-9.
515. Kotake, S.; Sato, K.; Kim, K.J.; et al. *Interleukin-6 and soluble interleukin-6 receptors in the synovial fluids from rheumatoid arthritis patients are responsible for osteoclast-like cell formation*. *J Bone Miner Res*, 1996. **11**(1): p. 88-95.
516. Hashizume, M.; Hayakawa, N.; Mihara, M. *IL-6 trans-signalling directly induces RANKL on fibroblast-like synovial cells and is involved in RANKL induction by TNF-alpha and IL-17*. *Rheumatology (Oxford)*, 2008. **47**(11): p. 1635-40.
517. Nakahara, H.; Song, J.; Sugimoto, M.; et al. *Anti-interleukin-6 receptor antibody therapy reduces vascular endothelial growth factor production in rheumatoid arthritis*. *Arthritis Rheum*, 2003. **48**(6): p. 1521-9.
518. Grossman, R.M.; Krueger, J.; Yourish, D.; et al. *Interleukin 6 is expressed in high levels in psoriatic skin and stimulates proliferation of cultured human keratinocytes*. *Proc Natl Acad Sci U S A*, 1989. **86**(16): p. 6367-71.
519. Duncan, M.R.; Berman, B. *Stimulation of collagen and glycosaminoglycan production in cultured human adult dermal fibroblasts by recombinant human interleukin 6*. *J Invest Dermatol*, 1991. **97**(4): p. 686-92.
520. Yamasaki, K.; Taga, T.; Hirata, Y.; et al. *Cloning and expression of the human interleukin-6 (BSF-2/IFN beta 2) receptor*. *Science*, 1988. **241**(4867): p. 825-8.
521. Kishimoto, T.; Akira, S.; Taga, T. *Interleukin-6 and its receptor: a paradigm for cytokines*. *Science*, 1992. **258**(5082): p. 593-7.

522. Boulanger, M.J.; Chow, D.C.; Brevnova, E.E.; et al. *Hexameric structure and assembly of the interleukin-6/IL-6 alpha-receptor/gp130 complex*. *Science*, 2003. **300**(5628): p. 2101-4.
523. Hirano, T. *Interleukin 6 in autoimmune and inflammatory diseases: a personal memoir*. *Proc Jpn Acad Ser B Phys Biol Sci*, 2010. **86**(7): p. 717-30.
524. Ishihara, K.; Hirano, T. *IL-6 in autoimmune disease and chronic inflammatory proliferative disease*. *Cytokine Growth Factor Rev*, 2002. **13**(4-5): p. 357-68.
525. Kamimura, D.; Arima, Y.; Hirano, T.; et al. *IL-6 and Inflammatory Diseases*, in *Cytokine Frontiers: Regulation of Immune Responses in Health and Disease*, T.Y. Yoshimoto, T., Editor. 2014. p. 396.
526. Shetty, A.; Hanson, R.; Korsten, P.; et al. *Tocilizumab in the treatment of rheumatoid arthritis and beyond*. *Drug Des Devel Ther*, 2014. **8**: p. 349-64.
527. Genentech; Available from: www.actemrahcp.com/pjia/landing-pjia.html. [24/04/2014].
528. Lagorce, D.; Sperandio, O.; Galons, H.; et al. *FAF-Drugs2: free ADME/tox filtering tool to assist drug discovery and chemical biology projects*. *BMC Bioinformatics*, 2008. **9**: p. 396.
529. InvivoGen Hek Blue IL-6. Available from: <http://www.invivogen.com/hek-blue-il6>. [03/06/2014].
530. InvivoGen Quanti-Blue. Available from: <http://www.invivogen.com/quant-blue>. [03/06/2014].
531. Meshulam, T.; Levitz, S.M.; Christin, L.; et al. *A simplified new assay for assessment of fungal cell damage with the tetrazolium dye, (2,3)-bis-(2-methoxy-4-nitro-5-sulphenyl)-(2H)-tetrazolium-5-carboxanil ide (XTT)*. *J Infect Dis*, 1995. **172**(4): p. 1153-6.
532. InvivoGen Hek Blue IL-1. Available from: http://www.invivogen.com/PDF/HEKBlue_TNF_IL1_TDS.pdf. [03/06/2014].
533. InvivoGen Hek Blue IL-4 IL-13. Available from: <http://www.invivogen.com/hek-blue-il4-il13>. [03/06/2014].

Liste des publications

1. Cebrià-Torrejón G, Assad Kahn S, **Lagarde N**, Castellano F, Leblanc K, Rodrigo J, Molinier-Frenkel V, Rojas de Arias A, Ferreira ME, Thirant C, Fournet A, Figadère B, Chneiweiss H, Poupon E. « **Antiproliferative activity of trans-avicennol from *Zanthoxylum chiloperone* var. *angustifolium* against human cancer stem cells** », *Journal of Natural Products*, 2012, 75 (2), 257-61.
2. Ben Nasr N*, Guillemain H*, **Lagarde N***, Zagury JF, Montes M. « **Multiple structures for Virtual Ligand Screening: defining binding sites properties-based criteria to optimize the selection of the query** », *Journal of Chemical Information and Modeling*, 2013, 53, 293-311. (*) contribution égale des co-auteurs
3. Bordessa A, Keita M, Maréchal X, Formicola L, **Lagarde N**, Rodrigo J, Bernadat G, Bauvais C, Soulier JL, Dufau L, Milcent T, Crousse B, Reboud-Ravaux M, Onger S. « **α - and β -Hydrazino acid-based pseudopeptides inhibit the chymotrypsin-like activity of eukaryotic 20S proteasome** », *European Journal of Medicinal Chemistry*, 2013, 70, 505-524.
4. **Lagarde N**, Ben Nasr N, Jérémie A., Guillemain H, Laville V., Labib T., Zagury JF, Montes, M. « **NRLiSt BDB, the Manually Curated Nuclear Receptors Ligands and Structures Benchmarking Database** », *Journal of Medicinal Chemistry*, 2014, 57(7), 3117-3125.
5. **Lagarde N**, Zagury JF, Montes, M. « **Importance of the pharmacological profile of the bound-ligand in enrichment on Nuclear Receptors: towards using decoy ligands** », *Journal of Chemical Information and Modeling* (Accepté le 24 septembre 2014)

Liste des communications orales

1. « **NRLiSt BDB, the Manually Curated Nuclear Receptors Ligands and Structures Benchmarking Database** », Lagarde N., *Workshop: "From Bioinformatics to Therapeutics"*, Doha, 7-9 avril 2014.

Posters

1. « **Modeling, design, synthesis and evaluation of novel Hsp 90 inhibitors** », Lagarde N., Peyrat J.-F., Messaoudi S., Brion J.-D., Alami M., Rodrigo J. *Journées de l'Ecole Doctorale (ED) 425 « Innovation thérapeutique du fondamental à l'appliqué* », Paris, 25 - 26 mai 2010.

2. « **Design and synthesis of novel non-covalent inhibitors of proteasome** » Keita M., Bordessa A., Maréchal X., Soulier J.-L., Milcent T., Lagarde N., Rodrigo J., Reboud-Ravaux, M., Onger, S. *Journées de l'ED 425 « Innovation thérapeutique du fondamental à l'appliqué* », Paris, 25 - 26 mai 2010.

3. « **NRLiSt BDB, the Manually Curated Nuclear Receptors Ligands and Structures Benchmarking Database** », Lagarde N., Zagury J.F., Montes M., *Strasbourg Summer School in Chemoinformatics, Strasbourg, 23-27 juin 2014*

Méthodes de criblage virtuel *in silico* : importance de l'évaluation et application à la recherche de nouveaux inhibiteurs de l'interleukine 6

Résumé

Le criblage virtuel est largement employé pour la recherche de nouveaux médicaments.

La sélection de structures pour les méthodes de criblage virtuel basées sur la structure reste problématique. Nous avons montré que les propriétés physico-chimiques du site de liaison, critères simples et peu coûteux en temps de calcul, pouvaient être utilisées pour guider celle-ci.

L'évaluation des méthodes de criblage virtuel, critique pour vérifier leur fiabilité, repose sur la qualité de banques d'évaluation. Nous avons construit la NRLiSt BDB, n'incluant que des données vérifiées manuellement et prenant en compte le profil pharmacologique des ligands. Une étude à l'aide du logiciel Surflex-Dock montre qu'elle devrait devenir la base de données de référence, pour l'évaluation des méthodes de criblage virtuel et pour rechercher de nouveaux ligands des récepteurs nucléaires.

L'application d'un protocole hiérarchique de criblage *in silico/in vitro*, a permis d'identifier de nouveaux composés inhibiteurs de l'IL-6, potentiellement utilisables dans le traitement de la polyarthrite rhumatoïde. Les résultats *in vitro* devront être confirmés par des tests *in vivo*.

Mots clés : criblage virtuel, docking, flexibilité, banque d'évaluation, récepteurs nucléaires, interleukine 6, polyarthrite rhumatoïde

Résumé en anglais

Virtual screening is widely used in drug discovery processes.

Structure selection in structure-based virtual screening methods is still problematic. We showed that simple and "low cost" binding site physico-chemical properties could be used to guide structure selection.

The evaluation of virtual screening methods, necessary to ensure their reliability, relies on benchmarking databases quality. We created the NRLiSt BDB, gathering only manually curated data and taking into account ligands pharmacological profiles. A study using Surflex-Dock showed that the NRLiSt BDB should become the reference, both for the evaluation of virtual screening methods and for the identification of new ligands of the nuclear receptors.

The use of a *in silico/in vitro* hierarchical approach screening allowed to identify new IL-6 inhibitors, that could be used in rheumatoid arthritis treatment. *In vitro* results should be confirmed *in vivo*.

Key words: Virtual screening, docking, flexibility, benchmarking database, nuclear receptors, interleukin 6, rheumatoid arthritis