Prédiction des constantes cinétiques de complexes protéine-ligand

Sonia Ziada

UNIVERSITÉ D'ORLÉANS

ÉCOLE DOCTORALE SSBCV

Institut de Chimie Organique et Analytique

THÈSE DE DOCTORAT

Présentée par :

SONIA ZIADA

Soutenue le : **23 janvier 2019**

Pour obtenir le grade de : **Docteur de l'Université d'Orléans**

Discipline / Spécialité : Chimie / Modélisation moléculaire

Prédiction des constantes cinétiques de complexes protéine-ligand

MEMBRES DU JURY :

Dr. Marc BAADEN	Directeur de recherche, CNRS Paris	Rapporteur
Pr. Manuel DAUCHEZ	Professeur, Université de Reims	Rapporteur, Président du jury
Dr. Sophie BARBE	Chargée de recherche, INRA Toulouse	Examinatrice
Dr. Pierre DUCROT	Responsable modélisation moléculaire, IdRS Croissy-sur-Seine	Co-encadrant
Dr. Samia ACI-SÈCHE	Chargée de recherche, CNRS Orléans	Co-encadrant
Pr. Pascal BONNET	Professeur, Université d'Orléans	Directeur de Thèse

A Yassin,

A mes parents, Latifa et M’hamed,

« *La science, mon garçon, est faite d'erreurs, mais d'erreurs qu'il est bon de commettre, car elles mènent peu à peu à la vérité*. »

Voyage au centre de la terre - Jules Verne

Remerciements

Je tiens tout d'abord à remercier l'Institut de Recherche Servier (IdRS) pour avoir financé ce projet de thèse.

Il aurait été bien difficile de vivre cette expérience incroyable que représente la thèse, tant sur le plan professionnel qu'humain, sans le concours de plusieurs personnes que je tiens à remercier à travers ces quelques lignes :

Tout d'abord, je remercie les membres de mon jury d'avoir accepté de lire mon manuscrit et d'évaluer mon travail : le Professeur Manuel Dauchez, le Docteur Marc Baaden, le Docteur Sophie Barbe, le Professeur Pascal Bonnet, le Docteur Pierre Ducrot et le Docteur Samia Aci-Sèche.

Je remercie le Professeur Pascal Bonnet de m'avoir accueilli au sein de son équipe et de m'avoir permis de m'épanouir dans ma recherche grâce à la confiance et à liberté qu'il m'a accordé.

Je ne pourrais pas ne pas mentionner dans ces remerciements mon encadrante : le Dr Samia Aci-Sèche, qui été pour moi un mentor, une scientifique inspirante, un manager exemplaire, disponible, à l'écoute, qui a largement contribué au bon déroulement de cette thèse, tant sur le plan scientifique qu'humain. Tu m'as appris énormément de choses scientifiquement durant ces trois ans, volontairement, mais tant aussi involontairement, par tes compétences transversales et comportementales. Sincèrement merci !

Cette thèse n'aurait pas été aussi stimulante et excitante sans les réunions trimestrielles avec mes co-encadrants de l'Institut de Recherche Servier : le Dr Pierre Ducrot et le Dr Eric Raimbaud. Je vous remercie de m'avoir fait confiance, de votre bonne humeur, d'avoir su créer cette ambiance si particulière pendant les réunions, à la fois stimulante, stressante et bienveillante. Cela m'effrayait au début mais j'ai fini par y prendre goût. Quel plaisir !

Je remercie le centre de calcul CaSciModOT Orléans-Tours et le Centre Régional

Informatique et Applications Numériques de Normandie (CRIANN) pour l'accès aux machines Artemis et Myria.

Je remercie les deux ex-dynamiciens de l'équipe SB&C : le Dr Abdennour Braka et le Dr. Julien Diharce, pour vos conseils précieux, pour votre aide, votre plaisir à partager la connaissance scientifique, à transmettre. Abdennour, je te remercie de m'avoir permis d'« éviter les pièges de la cinétique», « spontanément», pour les voyages en train si plaisants, pour les fameux hamburgers de ta mère. Que tu le veuilles ou non tu es un « Milky Man ». Julien, je te remercie pour ta passion du « vivant qui bouge » si communicative, pour tes qualités scientifiques et humaines et surtout pour tes qualités pédagogiques, qui ont naturellement conduit à ton implication dans ce projet de thèse.

Je remercie mon collègue Pascal Krezel avec qui partager le bureau a été un plaisir quotidien malgré des débuts quelque peu difficiles. Je te remercie pour ta liberté de pensée, pour ton courage à remettre en cause des principes établis et supposés inébranlables, pour ton ouverture d'esprit, pour ton aide dans les concepts physiques et mathématiques qui me dépassent. Quelle fraîcheur ! Je n'oublierai jamais cette phrase de toi si pleine de sagesse : « En science, il faut taper dans le tas et regarder ce qui en sort ». A nos bâtons !

Je remercie chaleureusement tout le reste de mes amis et collègues de l'équipe SB&C du laboratoire ICOA pour l'ambiance chaleureuse au quotidien et pour la cordiale ambiance qui règne au sein de l'équipe. Merci au Dr. Stéphane Bourg pour sa compréhension, son écoute et sa sympathie au quotidien ! Merci au Dr. José-Manuel Gally, pour son écoute, pour sa bienveillance et sa gentillesse, quelle positivité ! Merci à Fabrice Carles pour ses qualités humaines, sa simplicité, pour son miel et pour m'avoir, durant ces 3 ans, sensibilisé à de nouvelles technologies informatiques. Ta passion est contagieuse ! Merci au Dr. Jade Fogha pour ses conseils toujours si pleins de sagesse, pour sa bonne humeur si communicative et ses histoires africaines si hilarantes ! Merci aux nouveaux arrivants, Colin Bournez et Gautier Peyrat qui ont apportés leur touche d'humour et leur fraîcheur durant le temps où je les ai côtoyés. Je remercie mon Padawan, Dylan Serillon, dont l'encadrement de stage a été très agréable. Merci au Dr. Baptiste Canault pour m'avoir poussé à l'introspection et à développer mes *soft skills*.

Je remercie tous les membres du laboratoire ICOA, que j'ai pu côtoyer de près ou de loin,

les anciens et les présents, les permanents et non-permanents, en particulier Fatima pour l'accueil souriant du matin ainsi que Nathalie, Marie-Madeleine et les agents de la sécurité.

Plus personnellement,

Je souhaiterai remercier toute ma famille, proche et éloignée, ici et là-bas, particulièrement :

Je remercie profondément mon mari Yassin pour tout. Pour m'avoir poussé à reprendre les études après le bac, pour avoir cru en moi comme personne, pour son soutien constant, inébranlable, pour me pousser à aller de l'avant sans me bousculer, pour m'aimer sans m'étouffer, pour toute cette générosité d'âme…MERCI.

Je souhaiterai remercier également mes adorables parents, pour leur soutien et leur amour, mon père pour m'avoir appris la discipline à la façon d'un maitre-boulanger et ma mère pour m'avoir appris la patience, deux qualités très utiles pour une thèse.

Je souhaiterai remercier mes sœurs Afifa, Karima, Nabila pour leur amour, leur solidarité et leur soutien, chacune à sa façon, et mon petit frère Jihed dont je suis si fière.

Je ne peux finir ces remerciements sans un hommage « par la pensée» à certaines personnalités du corps enseignants qui ont marquées mon parcours sans vraiment le savoir, et que je garde dans mon cœur. Merci.

Table des matières

Table des figures

Table des tableaux

Liste des abréviations

ACP	Analyse en Composantes Principales
AMM	Autorisation de Mise sur le Marché
ATP	Adénosine triphosphate
CBER	*Center for Biologics Evaluation and Research*
CDER	*Center for Drug Evaluation and Research*
CDK8-CycC	Cyclin Dependent Kinase 8 – Cyclin C
clogP	Coefficient de partage octanol/eau calculé
cMD	Dynamique moléculaire conventionnelle
DM	Dynamique moléculaire
FDA	*Food and Drug Administration*
FFT	*Fast Fourier Transform*
GAFF	*General Amber Force Field*
GPCR	*G-Protein Coupled Receptor*
HTS	*High Throughtput Screening*
K_d	Constante de dissociation à l'équilibre
K_i	Constante d'inhibition à l'équilibre
k_{off}	Constante cinétique de dissociation
k_{on}	Constante cinétique d'association
LRT	*Long Residence Time*
MMA	Mécanisme moléculaire d'action
MM-GBSA	*Molecular Mechanics-Generalized Born Surface Area*
MMoA	*Molecular Mechanism of Action*
MRT	*Medium Residence Time*
NME	*New Molecular Entities*
PK	Protéine Kinase
PK/PD	*PharmocoKinetics/pharmacoDynamics*
PME	*Particle Mesh Ewald*
PMF	*Potential of Mean Force*
QSKR	*Quantitative Structure Kinetics Relationship*

RMSD	*Root Mean Square Deviation*
RMSF	*Root Mean Square Fluctuation*
RT	*Residence time*
SKR	*Structure Kinetic Relationships*
SRT	*Short Residence Time*
TMD	*Targeted Molecular Dynamics*

Avant propos

Le coût de plus en plus élevé de l'ensemble du processus de recherche préclinique et de développement clinique d'un médicament pousse la communauté scientifique à limiter au maximum les causes d'échecs. De nombreuses études démontrent qu'une évaluation préclinique des constantes cinétiques d'association et de dissociation du médicament permet de limiter les taux d'échec en phase II des essais cliniques. Ces constantes cinétiques correspondent à la vitesse avec laquelle la molécule active s'associe et se dissocie de la cible. La constante de dissociation est particulièrement intéressante car elle est directement reliée au temps de résidence[1] de la molécule active dans le site de liaison de sa cible thérapeutique. Ce temps de résidence est la durée pendant laquelle la molécule active reste liée à sa cible thérapeutique et donc induit un effet thérapeutique. Un trop faible temps de résidence réduit l'efficacité *in vivo* du candidat-médicament et un trop long temps de résidence peut augmenter sa toxicité. Aujourd'hui, l'importance de l'étude des constantes cinétiques en phase précoce de recherche est largement admise par la communauté scientifique mais souffre d'un manque d'expertise et d'outils adaptés à son étude.

Il existe des méthodes expérimentales pour mesurer ces constantes cinétiques mais en plus d'être très coûteuses, longues et fastidieuses, elles ne permettent pas de comprendre les mécanismes structuraux se produisant entre la cible thérapeutique et la molécule active. Or comprendre la dynamique du processus d'interaction d'une molécule vers sa cible et étudier les évènements structuraux à la base du processus ouvre la voie à l'étude quantitative de la relation structure-cinétique (SKR : *Structure Kinetic Relationships*). Sur la base d'une étude SKR, on pourrait ainsi prédire les modifications à apporter au niveau de la structure d'un candidat-médicament pour améliorer ses constantes cinétiques et donc son effet pharmacologique (toxicité, efficacité).

L'objectif de ce projet de thèse, financé par l'Institut de Recherche Servier à Croissy sur Seine, est de développer un outil basé sur des simulations de dynamique moléculaire pour prédire le temps de résidence de molécules actives. L'outil doit permettre de fournir un classement prédictif d'une série de molécules actives selon leur temps résidence en phase précoce de sélection et d'optimisation des molécules. Pour un usage industriel en routine, l'outil doit être

[1] Le temps de résidence est inversement proportionnel à la constante cinétique de dissociation

simple, rapide et applicable à tous systèmes. Le calcul du temps de résidence par des outils *in silico* constitue un axe majeur de recherche de l'équipe « Bioinformatique Structurale et Chémo-informatique », du Pr. Pascal Bonnet.

Le chapitre I retrace l'histoire des constantes cinétiques de liaison dans le contexte de l'industrie pharmaceutique. Afin de comprendre ce qui a poussé la communauté scientifique à s'intéresser aux constantes cinétiques de liaison, nous situerons l'évolution de l'industrie pharmaceutique dans le contexte du progrès des sciences depuis la naissance du médicament, en passant par les succès remarquables, jusqu'aux difficultés de productivité. Nous expliquerons comment on peut moduler les constantes cinétiques de liaison au niveau moléculaire et leur impact sur la réponse pharmacologique. Nous exposerons les différents moyens d'étude de la cinétique de liaison ainsi que leurs défis actuels en présentant plus en détail les méthodes de simulations numériques dédiées à la prédiction de la cinétique de liaison. Finalement, nous présenterons la famille des protéines kinases à laquelle appartient le système sur lequel nous avons travaillé tout au long de ce projet de thèse, la protéine kinase 8 cycline C dépendante (CDK8-CycC), et qui a été utilisée comme pour la preuve de concept de la méthode développée.

Le chapitre II présente l'étude de la structure et de la dynamique de ce complexe CDK8-CycC et notamment l'interaction entre CDK8 et la CycC. Outre la compréhension structurale et dynamique du complexe CDK8-CycC, cette étude permet de statuer quant à la nécessité ou non de garder la CycC dans nos simulations de DM.

Le chapitre III a pour objectif de présenter la méthode développée de prédiction du temps de résidence basée sur l'utilisation de la dynamique moléculaire dirigée. Il s'agit également d'explorer la relation structure-cinétique du jeu de données ayant servi de preuve de concept à travers l'analyse des interactions protéine-ligand et des différents termes d'énergie d'interaction (van der Waals, électrostatique, solvatation etc.) lors du processus de dissociation. L'outil est ensuite appliqué sur un jeu de données privé appartenant à l'Institut de Recherche Servier. Le but est de vérifier que la méthode développée nous permet d'obtenir un classement correct des composés selon leur temps de résidence sur un autre jeu de données que celui sur lequel elle a été développée.

Le chapitre IV, plus exploratoire, a pour but d'exposer le développement d'une méthode visant à échantillonner plus finement le processus de dissociation de façon à en extraire le profil d'énergie libre et à identifier les états intermédiaires stables et les états de transition. De fait, la méthode développée (dans le chapitre III) est une approche rapide où l'exploration de chaque micro-état tout au long du chemin, autrement dit l'échantillonnage, n'est pas suffisante pour permettre d'analyser plus finement la trajectoire et en extraire des grandeurs ayant un sens thermodynamique. Avec un échantillonnage plus important, on pourrait par exemple analyser le potentiel de force moyenne, qui peut être associé au profil d'énergie libre, à partir duquel l'état de transition et les états intermédiaires pourront être identifiés.

Dans le chapitre V, nous aborderons une description des méthodes et principes utilisés pour la réalisation de ce travail.

I. INTRODUCTION

A. Origine du médicament

1. Les plantes à l'origine des médicaments

La notion de soins aux personnes malades a depuis toujours existé, comme en témoigne la découverte de tablettes d'argile d'écriture cunéiforme sumérienne datant de 2500 à 1000 années avant J.-C. Celles-ci nous apprennent que la pharmacopée de l'époque était composée principalement de végétaux mais aussi de minéraux ou d'organes d'animaux préparés suivant un protocole précis en vue de leur administration : c'est le début de ce que l'on appelle aujourd'hui la galénique. L'origine des médicaments puise finalement sa source dans l'utilisation totalement empirique, basée sur l'expérience, de produits trouvés dans la nature pour soulager des maux, guérir une maladie ou soigner une blessure (Landry, 2018).

2. L'avènement de la chimie et l'extraction des principes actifs des plantes

Au XVIII^{ème} siècle, grâce au développement des sciences physiques et chimiques, on découvre, par la purification, la distillation, et l'extraction chimique, qu'un même végétal peut renfermer à la fois des substances toxiques, des substances thérapeutiques et des substances dont les effets sont opposés. Différentes dilutions des substances extraites montrent qu'une même substance peut se révéler sévèrement toxique ou parfaitement bénéfique selon la dose administrée et font prendre conscience de la notion de fenêtre thérapeutique.

Vers la fin du XVIII^{ème} siècle, les premières écoles de santé sont fondées et le XIX^{ème} siècle marque un tournant dans l'histoire du médicament grâce à l'industrialisation et l'isolement des principes actifs. Parmi les évènements marquants du XIX^{ème} siècle, nous pouvons citer l'isolement de la morphine (en 1803) et celui de l'acide acétylsalicylique (1829) principe actif de l'aspirine (Chast, 2002; Landry, 2018)

B. Les débuts de l'industrie pharmaceutique

1. Progrès à l'origine de l'émergence de l'industrie pharmaceutique

La chimie d'extraction en concomitance avec la chimie organique, qui se met en place à la fin du XIX^{ème} siècle, donneront naissance à une pharmacopée nouvelle, différente de celle des herboristes, prémisse de la naissance de l'industrie pharmaceutique. Cette évolution ainsi que l'émergence du concept «clé-serrure » formulé par Hermann Emil Fischer en 1894, décrivant l'interaction d'un substrat avec une enzyme, et du concept de «récepteur » établit par Paul Ehrlich en 1906, évoquant pour la première fois la notion de « protéine dont l'activité peut être modulée par l'interaction avec une petite molécule », seront à la base de la construction des premiers groupes de l'industrie pharmaceutique des années 1900 à 1980 (Prüll et al., 2009).

2. Le XX^{ème} siècle : entre génie et déclin

a) Succès pharmaceutiques remarquables

De fait, le XX^{ème} siècle est le siècle de l'industrie pharmaceutique. L'intensification de la recherche fondamentale, de la technologie scientifique et de la recherche clinique aboutissent à la mise sur le marché de nouvelles classes thérapeutiques.

Durant la première moitié du XX^{ème} siècle, de nombreux vaccins sont développés, dont les vaccins contre la tuberculose (le BCG, 1921) et la diphtérie (1923) (Plotkin, 2014). D'autres faits marquants du début du XX^{ème} siècle sont l'isolation et la purification de l'insuline à partir de cellules pancréatiques animales en 1922, ainsi que l'isolation de différents alcaloïdes de l'ergot de seigle, parasite du seigle, à partir desquels de nombreux médicaments seront développés et sont d'ailleurs toujours utilisés aujourd'hui. La découverte des sulfamides (1932) et de leurs propriétés antibactériennes ainsi que de la pénicilline par Flemming (1941) ont été un succès majeur dans le contrôle des maladies infectieuses (Landry, 2018).

A partir des années 50, plusieurs anticancéreux sont découverts tels que l'adriamycine dans les années 60. La conception des premiers neuroleptiques dans les années 1960 et 1970 sera suivie par la mise sur le marché de médicaments contrôlant la tension artérielle et le taux de cholestérol et enfin des thérapies contre le sida, représentant un fléau mortel lorsqu'il est apparu dans les années 1980 (Chast, 2002).

b) Émergence des biotechnologies et percées en génétique et biologie moléculaire

Vers 1960, l'arrivée des biotechnologies contemporaines et les avancées majeures en génétique et biologie moléculaire vont largement contribuer à diversifier l'arsenal thérapeutique (thérapie cellulaire, certains vaccins, anticorps monoclonaux, etc.) (Meunier, 2016).

(1) Les anticorps monoclonaux

Vers les années 1980-1985, la production industrielle d'anticorps monoclonaux et de protéines recombinantes est devenue possible et ceux-ci vont faire partie d'une nouvelle catégorie d'outils thérapeutiques : les « biopharmaceutiques ». Malgré leur coût important, les anticorps monoclonaux vont rapidement se trouver au cœur de la thérapie cancéreuse ciblée. Cette médecine personnalisée a été rendue possible par le développement des sciences dites « -omiques » : génomique, protéomique, métabolomique, etc (Drews, 2000). Toutes ces percées en biologie et génétique moléculaire ont grandement favorisé le développement des biopharmaceutiques.

(2) Thérapie génique

Au delà des anticorps, la thérapie génique et la thérapie cellulaire sont d'autres domaines nouveaux des biopharmaceutiques. L'avènement de la thérapie génique permet d'entrevoir la possibilité de mettre en place des traitements pour des maladies génétiques rares. Approuvé par la FDA en 2012, @Kalydeco permet de traiter une forme rare de fibrose pulmonaire due à une mutation ponctuelle dans une protéine régulatrice membranaire (Wood et al., 2013). Le coût du traitement annuel est de 294 000 dollars par an (Meunier, 2016).

(3) Thérapie cellulaire

En ce qui concerne la thérapie cellulaire, Gurdon et Yamanak, prix Nobel de médecine 2012, ont mis au point une technique permettant d'obtenir des cellules souches pluripotentes, c'est à dire des cellules immatures capables de redonner n'importe quelle sorte de cellules de l'organisme, à partir de cellules adultes normales déjà différenciées (Abbott, 2012). Ces cellules sont déjà largement utilisées pour modéliser de nombreuses pathologies et tester l'efficacité de molécules potentiellement thérapeutiques, lors de tests phénotypiques. Elles ouvrent également la voie à la médecine régénérative où on pourrait imaginer remplacer les cellules malades d'un patient par des cellules saines (Abbott, 2012).

c) Les petites molécules toujours au cœur de l'intérêt

Les biomédicaments coûtent généralement plus cher que les petites molécules, comme illustré avec les exemples cités précédemment. Heureusement, les petites molécules ont toujours une place importante dans la pharmacopée moderne comme le montre la **Figure 1** où elles représentent la majorité des médicaments approuvés (Mullard, 2018). Tout comme les biopharmaceutiques, plusieurs progrès technologiques soutiennent le développement des petites molécules.

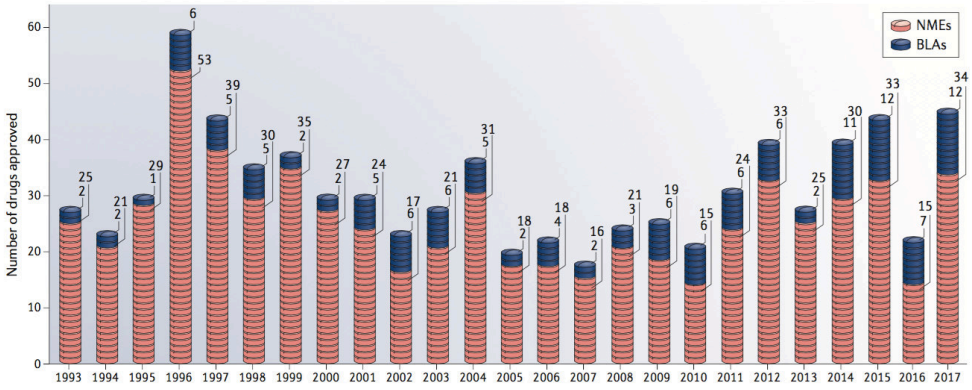


Figure 1 : Nouvelles approbations de la FDA depuis 1993 (Mullard, 2018).
Nouvelles entités moléculaires (NME) et demandes de licences de produits biologiques (BLA) approuvées par le Centre d'évaluation des médicaments et de recherche (CDER en anglais) depuis 1993. Les approbations du Centre d'évaluation des biopharmaceutiques et de recherche (CBER en anglais) ne sont pas incluses dans ce nombre de médicaments.

De fait, depuis 1970-1990, la chimie thérapeutique a fortement bénéficié du développement rapide des nouvelles méthodes d'analyse physique telles que la résonance magnétique nucléaire (RMN), la spectrométrie de masse et la cristallographie. Le développement de ces méthodes physiques associé au début de l'application des méthodes de simulation numérique sur des macromolécules (dynamique moléculaire, chimie quantique), permet de mener des études de relations structure-activité, assurant ainsi une meilleure optimisation des inhibiteurs d'enzymes ou des interactions protéine-protéine (études par amarrage moléculaire) (Pina et al., 2010). Cet ensemble contribue largement à assurer un avenir important aux petites molécules comme agents thérapeutiques. Dans les années 1990, la robotisation permet de mettre en place des criblages à haut débit (*high-throughput screening* ou HTS), mais celles-ci n'ont pas conduit à de nombreuses découvertes en raison du manque de diversité structurale des collections criblées (Payne et al., 2007).

Ainsi, ces éclatants progrès technologiques, apparus dès 1960, ont profondément modifié les méthodes de recherche et de développement et ont ouvert de nouvelles perspectives

d'innovation tant en industrie chimique qu'en biopharmaceutique. Paradoxalement, dès les années 1960-1970, le nombre de nouvelles molécules arrivant sur le marché au terme du processus de R & D commence à diminuer et les coûts de mise au point des médicaments augmentent (**Figure 2**), dépassant aujourd'hui l'équivalent de 200 à 300 millions d'euros (DiMasi et al., 2003).

C. Progrès technologiques et déclin : comment expliquer le paradoxe ?

1. Un contexte défavorable

Lorsque vers les années 1970-1980, la découverte des molécules issues de la chimie traditionnelle commence à s'essouffler, l'industrie s'ouvre alors à cette nouvelle source d'innovation que représente les biopharmaceutiques. L'adoption de ce nouveau terrain de recherche va augmenter les besoins en connaissances, savoir-faires et technologies nécessaires au développement de nouvelles molécules (Scannell et al., 2012). Parallèlement, de nouveaux investissements sont faits en recherche chimique, tels que le criblage à haut débit ou HTS. Sans apporter les résultats escomptés (les bio-médicaments prennent du retard), ces nouvelles techniques accroissent fortement la durée et les coûts de développement (**Figure 2**) qui passent, en moyenne, de 318 millions de dollars en 1987 à 802 millions en 2003 (DiMasi et al., 2003).

A cette augmentation du coût de développement du médicament, s'ajoute l'application des législations sur les brevets d'une durée de 20 ans, qui débute dans les années 1970 (Allemagne : 1967 ; Japon : 1976 ; France : 1978), et produit ses premiers effets au début des années 2000 (Wertheimer and Santella, 2007). À partir de cette date, les premiers brevets arrivent à expiration et les baisses de chiffre d'affaires sont particulièrement importantes pour les *blockbusters*, qui sont peu à peu remplacés par des médicaments génériques. Enfin, dans les années 1960, suite aux affaires du Stalinon®, de la Thalidomide® et du Distilbène® qui ont fait scandale, les autorisations de mise sur le marché (AMM) sont créées, obligeant l'industrie à se plier à un processus de développement strictement réglementé et découpé en quatre phases successives (phases cliniques I à IV) (Hauray, 2006). L'une des conséquences de ces réglementations est un accroissement du coût et du délai avant l'accès au marché pour les industries.

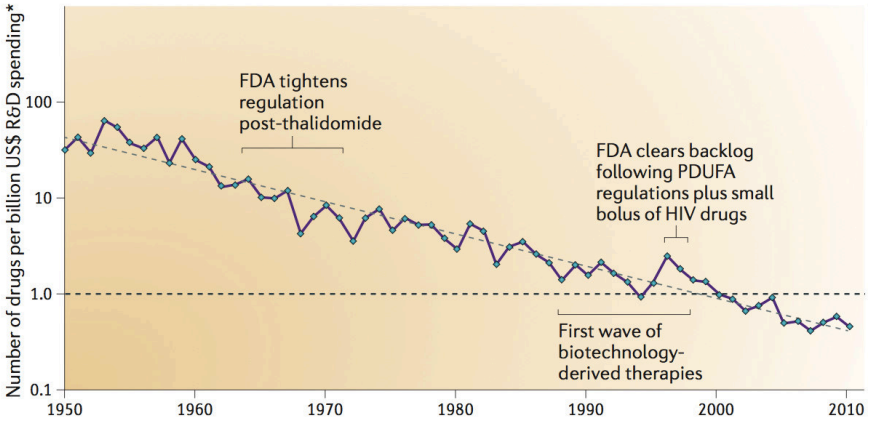


Figure 2 : Tendance générale de l'efficacité de la R&D pharmaceutique ou la Loi d'Eroom (Scannell et al., 2012).
Le nombre de médicaments approuvés par la FDA est divisé par 2 tous les 9 ans, d'où le nom de la loi de Moore inversé : la loi d'Eroom.

Les industries pharmaceutiques ont donc subi plusieurs ondes de choc affectant fortement leurs profits vers la fin des années 2000 : la tombée des brevets, l'essoufflement des découvertes issues de la chimie traditionnelle, les retards dans les innovations issues de la biopharmaceutique, qui augmentent les coûts comme la durée du développement, et l'allongement des phases de développement dû aux nouvelles exigences de l'AMM.

2. Le déclin : quelle stratégie adopter ?

Suite à ces difficultés, de nouvelles stratégies industrielles seront adoptées et vont conduire à une série de restructuration majeure des industries. En parallèle, on commence à chercher les causes d'échec, à identifier les étapes les plus coûteuses pour comprendre les différents facteurs expliquant cette faible productivité.

a) Stratégies d'alliance et de concentration

Parmi les stratégies industrielles mises en place, il y a le processus de fusion-cession, par lequel certains groupes pharmaceutiques s'associent et d'autres, au contraire, se scindent pour se séparer de certaines de leurs activités. En 1993, par exemple, le groupe britannique ICI a scindé ses activités entre la chimie (ICI) et la pharmacie (Zeneca). Cette dernière en fusionnant avec Astra en 1999 va donner le groupe AstraZeneca. Ce processus de fusion-cession a conduit à un recentrage des activités et une spécialisation sur certaines classes thérapeutiques. Ce processus va devenir courant, et on entendra souvent parler de « consolidation de marché » ou de « restructuration » de certains grands groupes. Une autre

stratégie adoptée est l'externalisation de certaines activités notamment en Chine, ce qui permet aux firmes de réduire leurs coûts de R&D. Malgré ces mesures, les coûts de R&D vont suivre une hausse constante (**Figure 2**). Aujourd'hui, les années 1996 et 1997 sont considérées comme l'âge d'or de l'industrie pharmaceutique avec des records en nombre de nouvelles entités découvertes jamais égalés (**Figure 1**) (Abecassis and Coutinet, 2008; Khanna, 2012).

b) Recherche des causes d'échecs

Selon de nombreuses estimations, le coût moyen de la commercialisation d'un médicament est aujourd'hui supérieur à 1 milliard de dollars (0.8-1.8 milliard de dollars) (DiMasi et al., 2016; Munos, 2009).

Plusieurs études ayant analysé les taux de succès en phases cliniques dans les années 1989 à 2012 convergent vers le même résultat : le plus grand taux d'attrition a lieu en phase II, comme le montre le **Tableau 1** (Arrowsmith and Miller, 2013; Bunnage, 2011; Hay et al., 2014). En revanche, les causes d'échecs ont évolué durant ces années grâce à la recherche scientifique. En 1991, la principale cause d'échec en phases cliniques I-III était attribuée à de mauvaises propriétés ADME chez l'homme (**Figure 3**) (Kola and Landis, 2004).

	This study (2013) all indications		This study (2013) lead indications		DiMasi *et al.*[6] lead indications		Kola *et al.*[8] lead indications		Abrantes-Metz *et al.*[9] lead indications	
	Phase success	Phase LOA	Phase success	Phase LOA	Phase success	Phase LOA	Phase success	Phase LOA	Phase success	Phase LOA
Phase 1 to phase 2	64.5%	10.4%	66.5%	15.3%	71%	19%	68%	11%	80.7%	NA
Phase 2 to phase 3	32.4%	16.2%	39.5%	23.1%	45%	27%	38%	16%	57.7%	NA
Phase 3 to NDA/BLA	60.1%	50.0%	67.6%	58.4%	64%	60%	55%	42%	56.7%	NA
NDA/BLA to approval	83.2%	83.2%	86.4%	86.4%	93%	93%	77%	77%	NA	NA
LOA from phase 1[a]		10.4%		15.3%		19%		11%	26.4%[c]	NA
Number of drugs in sample advanced or suspended[b]	5,820		4,736		1,316		NA		2,328	
Dates of source data (duration)	2003–2011 (9 years)				1993–2009 (17 years)		1991–2000 (10 years)		1989–2002 (14 years)	
Number of companies	835				50		10		NA	

[a]Probability of FDA approval for drugs in phase 1 development. [b]Total number of transitions used to calculate the success rate (the *n* value noted in the text). [c]Abrantes-Metz, *et al.*[9] reported 26.4% from phase 1 to phase 3. If we were to conservatively apply the 83.2% NDA/BLA success rate found in this study, Abrantes-Metz would yield the highest LOA from phase 1 (21%). NA, data not available.

Tableau 1 : Taux d'échec en phases cliniques I-III reportés à partir de plusieurs études (Hay et al., 2014).
Phase success : basé sur le nombre de médicaments qui passent d'une phase à l'autre et le *LOA* (*Likelihood Of Approval*) indique la probabilité de parvenir à l'approbation de la FDA à partir de la phase en cours et est également exprimée en pourcentage.

La recherche sur les propriétés ADME et la mise en place de modèles de criblages précliniques (*preclinical screens* tels que les modèles cellulaires) ont permis de mieux caractériser les profils ADME des candidat-médicaments comme en atteste la **Figure 3** (Khanna, 2012).

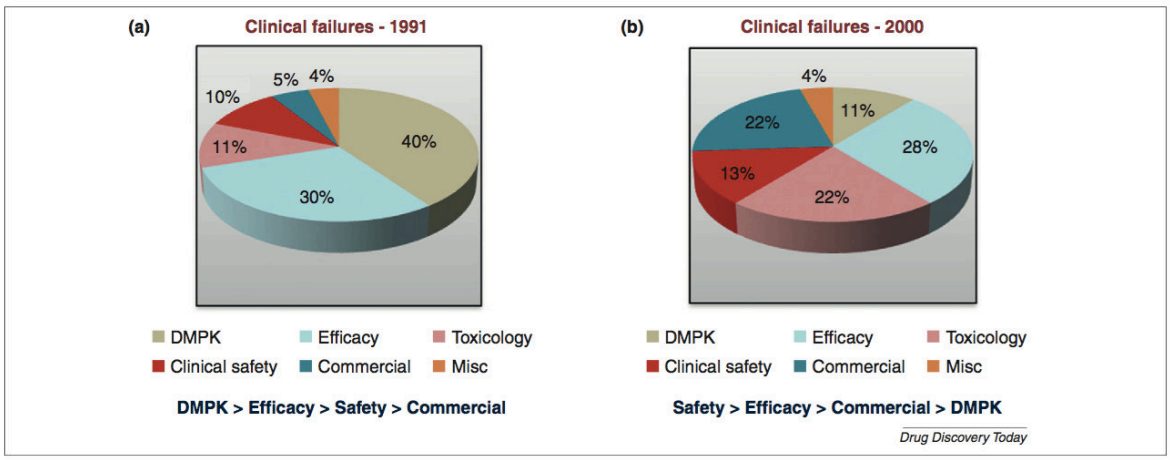


Figure 3 : Causes d'échec des candidat-médicaments en phases cliniques (I-III) dans les années 1991 (a) et 2000 (b) (Khanna, 2012).
DMPK *: Drug Metabolism and PharmacoKinetics*. Misc : *Miscellaneous*.

Les dernières études sur la question (Arrowsmith, 2011a, 2011b) montrent que l'efficacité est maintenant la principale cause d'échec suivie de près par la toxicité, comme illustré sur la **Figure 4** (Khanna, 2012).

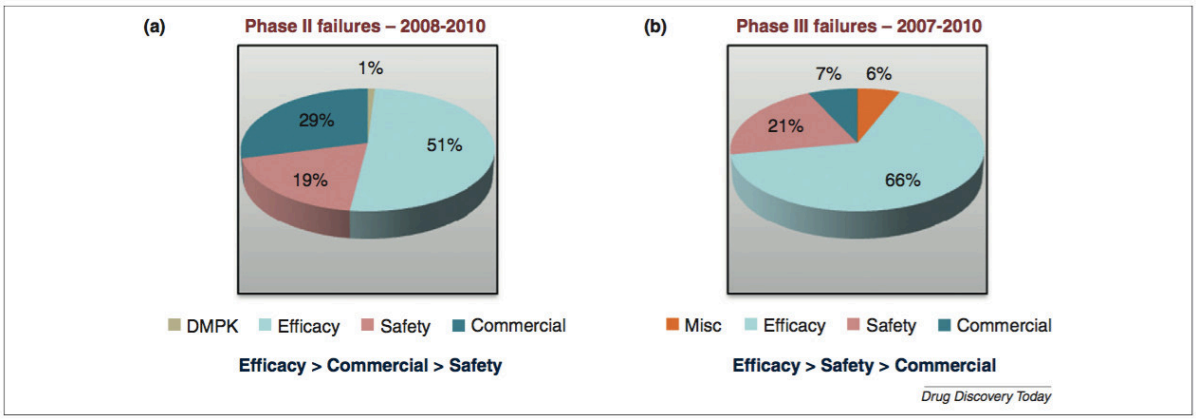


Figure 4 : Causes d'échec des candidat-médicaments (a) en phases II clinique entre 2008 et 2010 et (b) en phase III entre 2007 et 2010 (Khanna, 2012).
DMPK : Drug Metabolism and PharmacoKinetics.

Il convient de souligner que certains facteurs «non-techniques » ont également été avancés comme étant responsables du déclin de l'innovation. L'augmentation de la taille des firmes et l'externalisation de nombreuses activités amènent le chercheur à faire de plus en plus de bureautique (multiplication des rapports d'activité : *reporting*) l'éloignant de ses activités de recherche. Cela favorise une activité de routine plus qu'une activité innovatrice, où le chercheur devient un passeur de dossier, de rapports plus qu'une personne au fait de la science en marche. Avec une faible bureaucratie et une plus grande flexibilité, les petites entreprises de biotechnologie sont reconnues comme des centres axés sur l'innovation et font

généralement preuve d'un engagement plus fort envers la mission de recherche. Ils ont cependant tendance à être privés d'argent et sont influencés par le rapport coût-avantage avant d'entreprendre une étude (LaMattina, 2011).

D. Naissance de l'intérêt pour les constantes cinétiques de liaison

Nous allons voir dans cette partie comment dès les années 2000, différentes études ont été menées afin d'analyser des médicaments approuvés et d'identifier des critères de succès (Copeland, 2011; Copeland et al., 2006; Swinney, 2004, 2006; Swinney and Anthony, 2011). Il s'agit de comprendre s'il existe un (des) paramètre(s) commun(s) caractérisant les médicaments approuvés et pouvant être corrélé(s) à leur efficacité. Ces paramètres pourront être pris en compte lors des étapes précoces du processus de recherche et de développement permettant ainsi de réduire le taux d'échec en phases cliniques. Ces études (puis d'autres études évoquées au paragraphe F) vont amener peu à peu la communauté scientifique à considérer les constantes cinétiques de liaison comme des critères cruciaux de sélection des candidat-médicaments en phase précoce du processus de R&D. Plus particulièrement, le temps de résidence, qui est égal à l'inverse de la constante cinétique de dissociation, va prendre une importance considérable.

1. Analyse des médicaments approuvés

L'étude pionnière est celle menée par Swinney *et al.* en 2004 (Swinney, 2004) que nous détaillerons pour 1) comprendre comment est né l'intérêt pour les constantes cinétiques de liaison et 2) expliquer plusieurs concepts clés en pharmacologie, essentiels pour la compréhension du contexte dans lequel s'inscrit ce projet de thèse. Suite à cette étude, Copeland publie en 2006 (Copeland et al., 2006) une revue défendant les mêmes lignes que Swinney, dans laquelle le terme temps de résidence est défini pour la première fois.

a) Efficacité biochimique

L'efficacité biochimique (noté BE pour *Biochemical Efficiency*) de 50 médicaments commercialisés agissant sur 12 cibles, dont les canaux ioniques, les enzymes, les récepteurs couplés aux protéines G (RCPG ou GPCR en anglais) et les récepteurs nucléaires, a été analysée (Swinney, 2004). L'efficacité biochimique est le terme qui décrit l'efficacité avec laquelle la liaison de l'antagoniste ou de l'inhibiteur à sa cible se traduit en une réponse

fonctionnelle. Mathématiquement, l'efficacité biochimique est le rapport de l'affinité du médicament pour sa cible (K_i) sur la concentration requise pour inhiber de moitié la réponse fonctionnelle (IC_{50}) :

$$BE = \frac{K_i}{IC_{50}} \qquad \text{(Equation 1)}$$

Plus l'efficacité biochimique est grande, plus la liaison de l'inhibiteur à sa cible c'est à dire son K_i, est couplé à la réponse fonctionnelle mesurée durant les tests fonctionnels (IC_{50}), et donc plus il y a de chance que cette réponse s'observe également en phase clinique (EC_{50}) (**Figure 5**). Cela implique également que la quantité d'inhibiteur nécessaire pour obtenir une réponse est plus faible, ce qui abaisse les chances de voir apparaître une toxicité due à l'interaction de l'inhibiteur avec des cibles secondaires dites *off-targets* en anglais. Par conséquent, une élévation de l'efficacité biochimique permet d'augmenter l'index thérapeutique[2].

L'étude montre que sur les 50 médicaments analysés, 38 (soit 76%) présentent une efficacité biochimique supérieur à 0,4. L'analyse plus fine des résultats montre qu'une classe de médicaments, les statines, apparaissent comme des points aberrants de l'étude, avec des efficacités biochimiques inférieures à 0,03. En réalité, ce résultat est pertinent, car ce que nous n'avons pas encore précisé, c'est qu'une grande valeur d'efficacité biochimique, et donc un grand index thérapeutique, n'est pas souhaitable dans le cas d'inhibiteurs ayant un mécanisme moléculaire d'action impliquant une toxicité (*toxicity-based mechanism*). Or c'est le cas des statines qui bloquent la synthèse du cholestérol en inhibant la HMG-CoA reductase. De fait, une inhibition totale de la synthèse de cholestérol n'est pas voulue car elle serait cytotoxique. Lorsque les statines sont retirées de l'analyse, le pourcentage de médicaments commercialisés avec une bonne efficacité biochimique s'élève à 88% ; ceci est considérablement plus grand que le pourcentage de candidat-médicaments qui atteignent le marché.

[2] L'**index thérapeutique** d'un médicament est le rapport de la dose létale 50 (DL 50), soit la quantité d'une substance créant la mort chez 50 % des individus, sur la dose efficace 50 (DE 50), c'est-à-dire la dose nécessaire pour produire les effets désirés chez 50 % des individus.

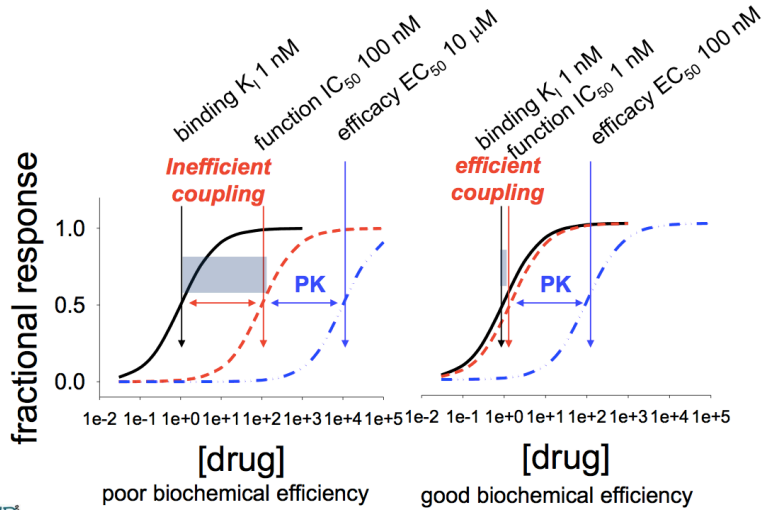


Figure 5 : L'efficacité biochimique d'un médicament (Swinney, 2016).
PK: Pharmacokinetics.

Cette étude va amener l'industrie pharmaceutique à considérer l'efficacité biochimique comme un critère important dans la sélection de molécules comme en attestent ces deux études portées par Pfizer et GlaxoSmithKline (Mourey et al., 2010; Portnoy et al., 2013). Il n'est pas surprenant de voir que la majorité des médicaments analysés ont une efficacité biochimique élevée ; ce qui intéresse les auteurs dans cette étude, c'est de comprendre par quel mécanisme moléculaire ces médicaments parviennent à maximiser leur efficacité biochimique et donc leur fenêtre thérapeutique.

b) Mécanisme moléculaire d'action

Dans ce contexte, ils analysent le mécanisme moléculaire d'action (MMA) de ces médicaments, définit comme étant l'ensemble des caractéristiques de l'interaction entre un médicament et sa cible (ou ses cibles) aboutissant à une réponse spécifique efficace et non toxique (Swinney, 2004). Le site d'interaction (allostérique ou orthostérique), les paramètres décrivant la qualité de la liaison (affinité et constante cinétique de liaison), l'impact fonctionnel (agoniste ou antagoniste) et la spécificité du résultat fonctionnel (par exemple, l'activation de voies de signalisation spécifiques) sont tous des caractéristiques qui contribuent à définir un MMA et par conséquent affectent la réponse pharmacologique.

Les auteurs constatent que la majorité des médicaments étudiés ayant une efficacité biochimique élevée présente un MMA dit « non à l'équilibre ». Un MMA « non à l'équilibre » est un MMA qui présente ou crée une transition vers un état hors équilibre afin d'éviter ou d'amoindrir la compétition directe à l'équilibre avec le(s) ligand(s) (ou substrat(s))

endogène(s) régie(s) par la loi d'action de masse. Les médicaments analysés dans l'étude présentent tous un mode d'interaction de type compétitif, c'est à dire qu'ils se lient au site de liaison du ligand ou substrat endogène. Cependant, la majorité d'entre eux présente une seconde étape qui succède à cette étape d'interaction compétitive à l'équilibre décrite par la loi d'action de masse, qui permet au système d'amoindrir ou d'éviter la compétition avec le ligand (ou substrat) endogène. Parmi ces mécanismes, il y a l'inhibition irréversible (modification covalente de la cible), le changement de conformation induit et l'inhibition dite insurmontable (observée lorsque le processus de dissociation de l'inhibiteur est plus lent qu'un processus compétitif tel que la disponibilité du ligand endogène).

Cette étude va générer une hypothèse selon laquelle l'action d'un médicament est efficace si son MMA connaît des transitions hors équilibre lui permettant d'atténuer la compétition avec le(s) ligand(s) et ou substrat(s) endogène(s), dans le cas où le MMA n'induit pas de toxicité. Dans le cas où la toxicité du médicament découle directement de son MMA, l'équilibre entre le médicament et l'(es) entité(s) endogène(s), autrement dit la compétition directe, est souhaitée et ajustée de façon à réduire la toxicité tout en maintenant une certaine efficacité. Si cette hypothèse est vérifiée, l'impact du MMA sur le processus de découverte de médicaments serait très important, car la sélection des candidats-médicaments se ferait principalement sur la base de constantes d'équilibre thermodynamique telles que l'affinité (K_i), qui rend compte de la force de la liaison d'un inhibiteur à sa cible sans prendre en compte « le bruit » apporté par l'interaction avec les composantes des systèmes biologiques. De fait, le K_d est mesuré par des tests *in vitro* où la cible est exposée à une concentration constante de l'inhibiteur, ce qui est loin de mimer l'aspect variable et dynamique des systèmes *in vivo*.

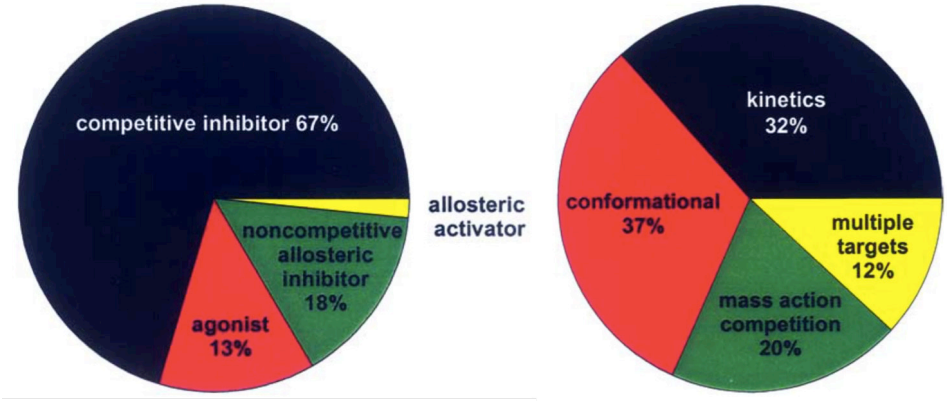


Figure 6 : Répartition des mécanismes moléculaires d'action des NMEs (*New Molecular Entities*) approuvées par la FDA entre 2001 et 2004 (Swinney, 2006).

En 2006, une étude plus générale analysant le MMA des NMEs approuvés entre janvier 2001 et novembre 2004 (85 NMEs au total) est réalisée (Swinney, 2006) et aboutie aux mêmes conclusions. Bien que 80% des NMEs ont un mode d'action compétitif (67% d'inhibiteur + 13 % d'agoniste) seulement 20% agissent par compétition directe à l'équilibre régie par la loi d'action de masse (**Figure 6**). Le MMA des 80 % restant implique soit a) la stabilisation ou l'induction d'une conformation (groupe incluant aussi les inhibitions allostériques et non-compétitives), b) une étape cinétiquement lente ou irréversible (dissociation lente, inhibition irréversible) ou bien c) une interaction avec de multiples cibles ou sites de liaison (**Tableau 2**).

Mechanism	NMEs
Active involving a drug stabilized conformational change R + D ↔ RD ↔ R'D → response	*Agonist*-almotriptan, apomorphine, eletriptan, formoterol, frovatriptan, travoprost *Partial agonists*- tegaserod, aripiprazole *Active antagonist*- eplerenone, fulvestrant, pegvisomant *conformational inhibition*-fondaparinux, gemifloxacin, imatinib, pimecrolimus, epinastine, enfuvirtide *Allosteric or noncompetitive*-rifaximin, cinacalcet, *Uncompetitive-like*-tadalafil, vardenafil, memantine
Active requiring a slow kinetic event contributing to non-equilibrium R + D ↔ RD → R'-D	*Chain termination*-adefovir, emtricitabine, telithromycin, tenofovir *Irreversible*-azacitidine, cefditoren, dutasteride, ertapenem, nitisinone, *Slow dissociation*-bortezomib, rosuvastatin, valdecoxib, aprepitant, desloratadine, olmesartan, tiotropium, duloxetine, palonosetron, oxaliplatin
Mass action equilibrium binding R + D ↔ RD	atazanavir, erlotinib, ibandronate, gefitinib, miglustat, seraconazole, voriconazole, abarelix, alfuzosin, bozentan, solifenacin, atomoxetine
Multiple targets/ Combination therapies	*Multiple targets*-pemetrexed, galantamine, ziprasidone, eletriptan *Combination therapies*- ethinyl estradiol with etonogestrel, norelgestronim, or drospirenone

Tableau 2 : Classification des NMEs approuvées par la FDA entre 2001 et 2004 selon leur mécanisme moléculaire d'action (Swinney, 2006).

2. Mécanisme moléculaire d'action, constantes cinétiques et efficacité

Essayons d'expliquer ces observations par les bases théoriques de l'inhibition enzymatique. L'inhibition compétitive réduit l'affinité apparente du substrat, notée communément Km. L'inconvénient majeur d'un inhibiteur compétitif réside dans le fait que son efficacité peut être diminuée par une compétition avec le substrat, régie par la loi d'action de masse décrite par la **Figure 7**.

La mesure avec laquelle l'efficacité peut être compromise par une interaction compétitive dépend des affinités et des concentrations relatives de l'inhibiteur et du substrat. Lorsque la concentration du substrat est augmentée, des concentrations plus élevées de l'inhibiteur seront nécessaires pour obtenir le même degré d'occupation du site de liaison (**Figure 7**). La quantité d'inhibiteur nécessaire en présence de substrat pour obtenir le même degré d'inhibition qu'en l'absence de substrat est augmentée par le facteur 1 + S / Km, où S est la concentration du substrat. Ce facteur, communément connu sous le nom de correction de Cheng-Prusoff, est utilisé pour calculer le K_i à partir du changement de IC_{50} qui se produit pendant l'inhibition compétitive (Yung-Chi and Prusoff, 1973). Lorsque les concentrations de substrat dépassent Km, des concentrations plus élevées de médicament compétitif devraient être nécessaires pour atteindre le même niveau d'inhibition.

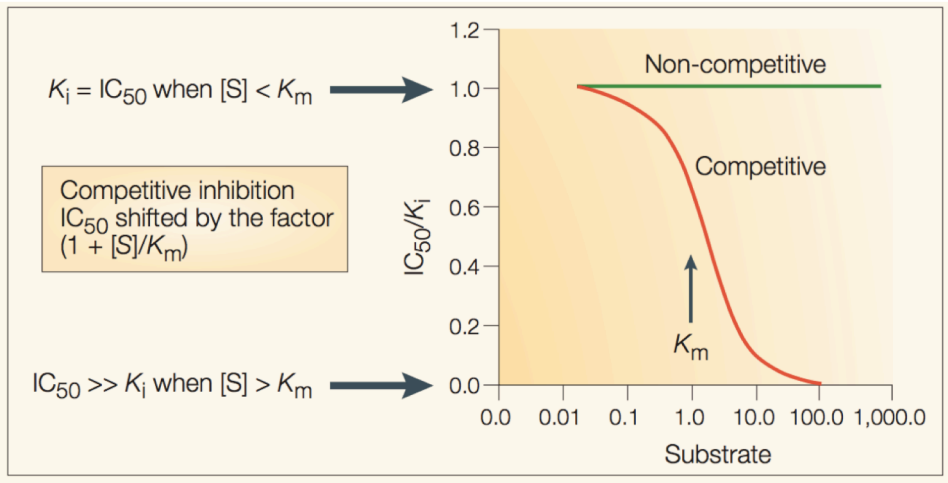


Figure 7 : Effet de la concentration en substrat sur une inhibition compétitive et non-compétitive (Swinney, 2004).

Maintenir une inhibition compétitive dans le contexte cinétique des systèmes biologiques qui sont des systèmes dit «ouverts » (apport constant en substrat et élimination du produit) peut représenter un réel défi (Copeland et al., 2006; Westley and Westley, 1996). De plus,

l'inhibition compétitive augmente la concentration en substrat endogène, parce qu'elle bloque sa liaison au site actif et donc sa catalyse, sauf si un mécanisme d'élimination du substrat existe. Il existe donc un conflit entre le besoin d'avoir un médicament efficace à faible concentration, l'inefficacité potentielle des inhibiteurs compétitifs et le fait que la majorité des médicaments interagissent de façon compétitive avec la cible. Ainsi, la stabilisation ou l'induction d'une conformation, une dissociation lente ou une inhibition irréversible, et l'interaction avec de multiples cibles ou sites de liaison sont des mécanismes efficaces et prédominants parmi les NMEs (**Figure 6** et **Tableau 2**) car ils créent des transitions vers des états dits « non à l'équilibre » c'est à dire non-sensibles à la compétition décrite par la loi d'action de masse, permettant d'éviter ce conflit. De cette façon, lorsque la toxicité n'est pas basée sur le MMA, ces mécanismes permettent d'observer une efficacité à de faibles concentrations en médicament et donc augmentent la fenêtre thérapeutique en minimisant l'interaction avec les cibles secondaires. Déterminer et optimiser le MMA de candidat-médicaments précocement dans le processus de découverte de médicaments permettrait une meilleure transposition de l'activité mesurée lors des tests *in vitro* en réponse efficace.

C'est dans ce contexte qu'est né l'intérêt pour la détermination des constantes cinétiques de liaison (k_{on} et k_{off}) en phase précoce du processus de découverte des médicaments, dont la compréhension et la variation est l'une des façons permettant de moduler et d'optimiser le MMA et donc d'impacter sur l'efficacité et la toxicité. Ces études ont également permis à la communauté scientifique de prendre conscience que l'affinité, autrement dit la force d'interaction entre le médicament et sa cible, mesurée lors de tests *in vitro* (et donc loin de toutes les variations qu'impliquent les système biologiques), ne peut à elle seule constituer un bon paramètre de prédiction de l'efficacité. La détermination des constantes cinétiques ainsi que l'étude de la diversité conformationnelle de la cible sont des moyens de comprendre et d'optimiser le MMA permettant une meilleure transposition de l'activité mesurée lors des tests *in vitro* en réponse pharmacologique efficace.

E. Les constantes cinétiques de liaison

1. Définition

Les constantes cinétiques de liaison sont définies par les constantes cinétiques d'association (k_{on}) et de dissociation (k_{off}) du complexe récepteur-ligand. La notion de temps de résidence est introduite par R. Copeland pour désigner le temps pendant lequel la cible est occupée par

le ligand (Copeland et al., 2006). Le temps de résidence est égale à l'inverse de la constante cinétique de dissociation (*Residence Time* (RT) = $1/k_{off}$). Dans un modèle de liaison simple à deux états, le k_{on} et le k_{off} sont déterminées respectivement par la différence d'énergie entre l'état de transition et les états non lié et lié grâce à l'équation d'Arrhénius (**Figure 8**). Les états de transition sont des structures instables ayant une courte durée de vie et donc impossibles à observer directement. C'est peut-être l'une des raisons pour lesquelles l'optimisation médicamenteuse, guidée par la structure, se concentre sur les affinités de liaison plutôt que sur la cinétique (Schiele et al., 2015a). L'affinité du ligand pour la protéine, mesurée par la constante de dissociation à l'équilibre K_d, est déterminée par la différence d'énergie entre les états lié et non lié, énergétiquement stables, et est égale au rapport k_{off}/k_{on}. La conséquence évidente de ceci pour l'optimisation d'un composé, est qu'il n'est pas possible de manipuler le k_{on} et le k_{off} indépendamment de l'affinité.

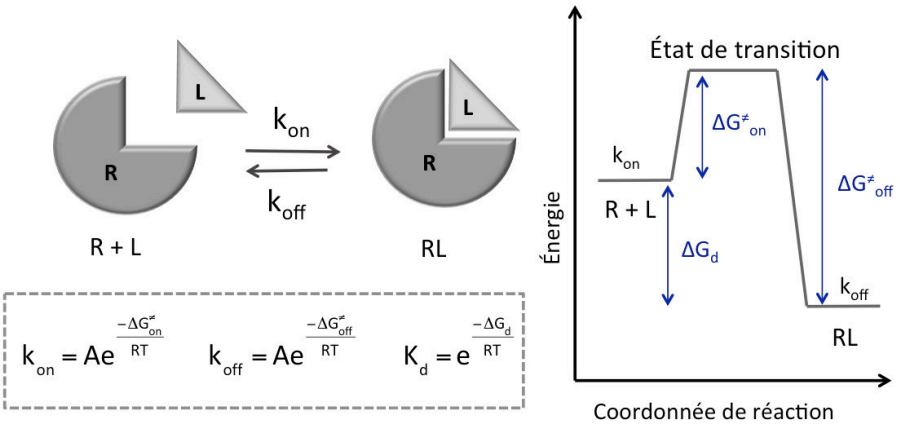


Figure 8 : Profil d'énergie d'un modèle cinétique simple à une étape.
Les relations entre l'énergie et les constantes cinétiques (équation d'Arrhénius) d'une part et l'affinité d'autre part, ont été indiquées.

2. Modèles cinétiques d'interaction

a) Modèle cinétique simple à une étape

À partir du modèle simple à une étape présenté sur la **Figure 8**, il devient évident que la modification des constantes cinétiques peut être réalisée de différentes manières. Par exemple, lorsque l'énergie des états fondamentaux (lié et non lié) est maintenue constante, déstabiliser l'état de transition (augmenter son niveau d'énergie) augmente la hauteur de la barrière énergétique. Par conséquent, l'association et la dissociation du complexe sont toutes deux ralenties, tandis que l'affinité reste constante. Lorsque l'énergie de l'état de transition

est maintenue constante, la stabilisation de l'état lié final (LR) diminue son l'énergie ce qui ralentit la dissociation et augmente l'affinité.

Le modèle de liaison simple à une étape est utile pour illustrer les considérations thermodynamiques générales de la cinétique de liaison, mais est souvent considéré comme trop simpliste. Dans de nombreux cas, la formation du complexe récepteur-ligand s'accompagne de changements conformationnels. Il a été suggéré que deux autres mécanismes sont prépondérants et expliquent le comportement cinétique des complexes récepteur-ligands (Copeland, 2011, 2016)

b) Modèle de sélection conformationnelle

Dans le modèle de sélection conformationnelle, la cible libre existe dans un ensemble de conformations ; seules certaines de ces conformations sont capables de se lier au ligand. Par exemple, en considérant que l'ensemble des conformations est composé de seulement deux états majoritaires, en équilibre l'un avec l'autre : un état qui est incapable de lier le ligand (R) et une autre conformation qui lie le ligand (R*) (Equation 2). Dans ce modèle, l'étape cinétiquement limitante est supposée être l'interconversion entre les deux formes libres R et R* ; une fois formé, R* se lie rapidement au ligand.

$$R \underset{k_2}{\overset{k_1}{\rightleftarrows}} R^* \underset{k_4}{\overset{k_3}{\rightleftarrows}} R^*L \qquad \text{(Equation 2)}$$

k_1, k_2, k_3 et k_4 représentent les constantes de vitesse.

c) Modèle d'ajustement induit

Le modèle d'ajustement induit conduit à la même forme finale du complexe récepteur-ligand, R*L, mais arrive à cet état par une voie cinétique différente (Equation 3). Ici, le récepteur non lié existe dans un seul état conformationnel majoritaire, R, qui est capable de lier le ligand pour former un complexe RL. Les éléments de reconnaissance dans la poche de liaison ne sont pas complémentaires du ligand de manière optimale dans l'état RL. L'acte de liaison du ligand provoque un réajustement conformationnel de la cible pour former une nouvelle conformation (R*L) dans laquelle une complémentarité optimale entre le ligand et la poche de liaison est obtenue. Dans ce modèle, la liaison du ligand à la conformation initiale de la cible R est considérée comme rapide, et l'étape cinétiquement limitante est la transition conformationnelle lente de l'état RL à l'état R*L final.

$$R \underset{k_2}{\overset{k_1}{\rightleftarrows}} RL \underset{k_4}{\overset{k_3}{\rightleftarrows}} R^*L$$

(Equation 3)

Le modèle de liaison simple en une étape, le modèle de sélection conformationnelle et le modèle d'ajustement induit peuvent être distingués expérimentalement les uns des autres ; ceci a été discuté en détail dans (Copeland, 2013; Tummino and Copeland, 2008).

3. Comprendre la cinétique de liaison

a) Mise en place d'un partenariat public-privé : *Kinetics for drug discovery* (*K4DD*)

Comme nous l'avons vu précédemment, une compréhension détaillée de la cinétique de liaison et particulièrement de la cinétique de dissociation (k_{off}) fournit des informations cruciales sur le mécanisme d'action moléculaire d'un composé. Cela pourrait contribuer à améliorer la prise de décision en matière de découverte de médicaments, conduisant ainsi à une meilleure sélection de composés intéressants à développer. Lorsqu'un groupe de chercheurs a commencé à s'intéresser à la cinétique de liaison, il est vite apparu qu'un certain nombre de questions ouvertes devaient être traitées (Schuetz et al., 2017). Celles-ci incluent l'analyse des aspects moléculaires de la cinétique de liaison de molécules dont le but est de mettre en évidence des règles SKR (*Structure Kinetics Relationships*). Ces directives SKR pourront être utilisées pour optimiser des candidat-médicaments. Étant donné que ces questions intéressent de nombreuses sociétés pharmaceutiques, un projet collaboratif de l'Initiative pour les Médicaments Innovants (IMI) a été mis en place (Gottwald et al., 2016) : *Kinetics for Drug Discovery* (K4DD, https://db.k4dd.eu)

Le projet K4DD, d'une durée de cinq ans et doté d'un budget de 21 millions d'euros, a débuté en novembre 2012. Vingt partenaires (neuf instituts universitaires, sept grandes sociétés pharmaceutiques et quatre PMEs) de six pays européens collaborent étroitement sur des objectifs sélectionnés par le consortium.

L'Institut de Recherche Servier a établi une étroite collaboration avec l'ICOA à Orléans en finançant ce projet de thèse. Cette collaboration a pour but la mise en place de protocoles de simulations numériques pour la prédiction des constantes cinétiques. Le projet K4DD se concentre sur trois axes majeurs, appelés WP pour *Work Package* (**Tableau 3**) (https://www.k4dd.eu/k4dd-strategy/).

WP1	Acquérir une compréhension moléculaire de la cinétique de liaison à travers la génération de données cinétiques et thermodynamiques, de structures expérimentales protéine-ligand ainsi que le développement d'outils de modélisation *in silico* .
WP2	Evaluer et développer des tests expérimentaux permettant une évaluation rapide et robuste des constantes cinétiques des composés.
WP3	Développer des modèles pharmacocinétique-pharmacodynamique (PK/PD) à partir de données *in vitro* et *in vivo*

Tableau 3 : Axes de recherche (WP) définis par le consortium K4DD.

b) Modulation des constantes cinétiques au niveau moléculaire

(1) Modulation de la constante d'association (k_{on})

La diffusion moléculaire, à l'origine du processus de collision, joue un rôle décisif dans le processus d'association d'un ligand à un récepteur (Bongrand, 1999). A cet égard, les interactions à longue distance (principalement électrostatiques) jouent un rôle crucial. Déterminer les régions d'attraction au niveau de la cible puis modifier les propriétés structurales du ligand en conséquence pour optimiser les interactions à longue distance sont une façon de moduler la constante d'association (Fedosova et al., 2002; Radić et al., 1997; Spaar et al., 2006). Au voisinage de la cible, il a été rapporté que la diminution de l'entropie conformationnelle du ligand, due à la réduction de son mouvement de rotation, accélère son association à la cible (Alsallaq and Zhou, 2008; Fleck et al., 2012). Enfin, la vitesse d'association du ligand est déterminée en partie par l'accessibilité au site de liaison du récepteur : un accès limité par un passage étroit est intrinsèquement plus lent qu'un accès libre à un site de liaison ouvert, idem pour la sortie. Plus le médicament est gros et (ou) plus le site de liaison est enfoui, plus cette notion simple s'applique(Pan et al., 2013).

Le changement de conformation du récepteur a aussi un impact sur la vitesse d'association du ligand (Frederick et al., 2007). La liaison d'inhibiteurs à InhA (I*nhibin Alpha subunit*), par exemple, l'énoyl-ACP réductase de *Mycobacterium tuberculosis*, implique la conversion d'une boucle désordonnée en une hélice ordonnée sur le site de liaison. Ce grand changement de conformation semble avoir un impact significatif sur la constante d'association de l'inhibiteur qui diminue en raison de la pénalité entropique conformationnelle engendrée (Luckner et al., 2010).

Brady *et al.* et plus récemment Schuetz *et al.*, montrent dans leurs études que l'introduction de groupements hydrophobes dans la partie du ligand qui pointe vers la poche hydrophobe augmente la vitesse d'association en abaissant la barrière énergétique due à la désolvatation du ligand. (**Figure 9**) (Brady et al., 1990; Schuetz et al., 2018). Dror *et al.* ont rapporté que le ligand hydrophobe et la désolvatation du vestibule extracellulaire du récepteur β2-adrénergique sont les principales étapes limitantes de la vitesse d'association ligand-récepteur (Dror et al., 2011). D'autres études mettent en évidence le fort impact de l'énergie de désolvatation du ligand et (ou) du site de liaison sur la cinétique d'association (Deganutti et al., 2017; Gaspari et al., 2016; Mondal et al., 2014; Pearlstein et al., 2013).

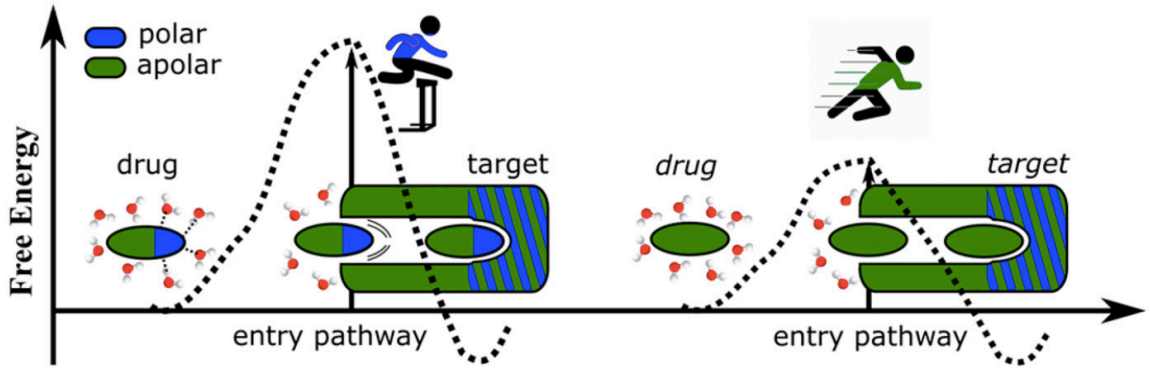


Figure 9 : Orientation du ligand, énergie de désolvatation du ligand et constante cinétique durant le processus d'association (Schuetz et al., 2018).

Au vu de ces observations et avant de passer à la partie traitant de la modulation du k_{off}, il convient de faire un point sur la thermodynamique d'interaction d'un ligand avec son récepteur. De fait, les phénomènes énoncés précédemment tels que l'effet de l'eau, l'effet hydrophobe et la flexibilité conformationnelle peuvent s'expliquer par la thermodynamique qui lie les aspects macroscopiques aux aspects microscopiques. Les aspects thermodynamiques régissant la force de liaison entre un récepteur et son ligand, c'est à dire l'affinité, sont bien connus et largement étudiés, ce qui n'est pas le cas de la cinétique de liaison. Avec la prise de conscience de l'importance de l'aspect dynamique du processus d'association et de dissociation, de plus en plus d'études s'attèlent à mettre en lumière les aspects thermodynamiques contribuant à accélérer (et ou ralentir) le processus d'association et de dissociation d'un ligand vis-à-vis de son récepteur (Freire, 2015; Keserü and Swinney, 2015; Klebe, 2015). L'objet du paragraphe suivant est justement de présenter les connaissances actuelles sur ce sujet afin de comprendre les fondements thermodynamiques à

la base de ces phénomènes (effet hydrophobe, effet de l'eau etc.) et ainsi comprendre leur impact sur la cinétique de liaison.

(2) Aspects thermodynamiques et effet de l'eau

Bien que l'influence de l'eau sur la stabilisation des complexes récepteur-ligands soit bien connue (effet hydrophobe) (Böhm and Klebe, 1996), l'effet de l'eau sur la cinétique de liaison n'a été reconnu que récemment (Klebe, 2015; Setny et al., 2013). Récemment, des chercheurs ont pris conscience de la tendance des nouveaux candidat-médicaments à être excessivement hydrophobes, à présenter une diminution de la solubilité et de la perméabilité et par conséquent un profil *drug-like* médiocre (Leeson and Springthorpe, 2007). Il a été observé que les composés ayant un bon profil *drug-like* sont ceux caractérisés par une forte activité et une faible hydrophobicité simultanément (Ryckmans et al., 2009). D'un point de vue fondamental, une question importante consiste à déterminer si de telles observations ont une base thermodynamique solide et dans quelle mesure celle-ci pourra être utilisée de manière prospective pour ouvrir la voie à une optimisation rationnelle (Keserü and Swinney, 2015).

Dans cette perspective, on a vu émerger de plus en plus de signatures thermodynamiques mesurées par *ITC* (*Isothermal Titration Calorimetry*), fournissant une représentation visuelle de l'ampleur des différentes interactions qui contribuent à la liaison, permettant ainsi de comparer les ligands (**Figure 10**) (Fox et al., 2017; Klebe, 2015; Lafont et al., 2007; Sarver et al., 2007). Rappelons que l'affinité est dictée par l'énergie de liaison de Gibbs ($\Delta G_d = -RT\ln(K_d)$) qui en retour est égale à la somme de la contribution enthalpique de liaison (ΔH) et de la contribution entropique de liaison ($-T\Delta S$).

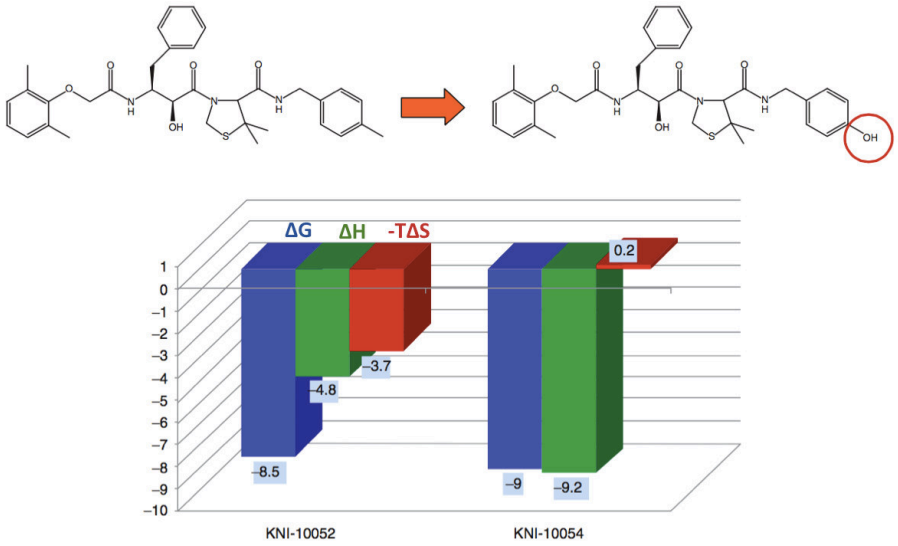


Figure 10 : Signatures thermodynamiques de deux inhibiteurs de protéase (Freire, 2015).

(a) La contribution enthalpique

Les molécules *drug-like* sont composées d'atomes polaires et non polaires qui contribuent très différemment au changement d'enthalpie. Durant le processus d'association, deux phénomènes se produisent simultanément : la désolvatation du composé et la formation d'interactions protéine-ligand.

Il a été montré que les enthalpies de liaison défavorables sont généralement associées à la désolvatation des groupes polaires qui n'établissent pas d'interactions fortes avec la protéine. La pénalité introduite par l'enthalpie de désolvatation et la contribution favorable des liaisons hydrogènes fortes vont diminuer respectivement le k_{on} et le k_{off}, entrainant une compensation sur l'affinité ($K_d = k_{off}/k_{on}$) .

Même lorsque l'enthalpie de liaison parvient à être favorable, malgré l'introduction de groupements polaires, comme illustré dans l'exemple de la **Figure 10**, elle s'accompagne souvent d'une perte entropique, entrainant qu'un léger gain ou bien même une baisse de l'affinité de liaison. Ce résultat illustre bien un fait couramment observé lors de l'optimisation d'un *lead* qui est la compensation enthalpie-entropie : parfois l'introduction de groupes polaires qui établissent des liaisons hydrogènes fortes n'entraîne pas de gains d'affinité de liaison (Freire, 2015; Keserü and Swinney, 2015). Ces résultats indiquent également que pour obtenir des gains d'affinité de liaison avec des groupes polaires, il est nécessaire de surmonter le phénomène omniprésent de la compensation enthalpie-entropie (Deganutti et al., 2017; Lafont et al., 2007).

Contrairement aux groupes polaires, l'introduction de groupes non polaires entraîne des gains d'enthalpie et d'entropie qui conduisent à des améliorations modérées de l'affinité de liaison. L'accumulation de ces améliorations peut entraîner une forte affinité mais également des composés hautement hydrophobes et donc des profils pharmacocinétiques médiocres. Comme l'introduction de fonctionnalités non polaires est dépourvue des grands changements d'enthalpi-entropie compensatoires, cela pousse les scientifiques à en introduire d'avantage au détriment des groupes polaires, ce qui expliquerait la tendance des nouveaux candidat-médicaments à être excessivement hydrophobes (Freire, 2008; Leeson and Springthorpe, 2007).

(b) La contribution entropique

Deux termes sont importants du point de vue de la conception rationnelle : l'entropie de (dé)solvatation et l'entropie conformationnelle. En règle générale, l'entropie de désolvatation des groupes polaires et non polaires est favorable à la liaison. Cependant, elle peut être défavorable si la formation d'une liaison hydrogène force certains groupes apolaires à être plus exposés à l'eau. En outre, la structuration des chaînes latérales, du squelette de la protéine, ou de la molécule, suite à la formation d'une liaison hydrogène par exemple, entraine une pénalité entropique de type conformationnelle importante (Keserü and Swinney, 2015). Ainsi en règle générale, les groupes non polaires présentent une entropie de liaison favorable grâce à l'entropie de désolvatation qui va permettre d'augmenter le k_{on}. Leur faible pénalité enthalpique, quand elle existe, peut être facilement surmontée par des contacts hydrophobes avec la cible ce qui permet de réduire le k_{off} (Freire, 2008).

(c) Conclusion

La principale complication lors de l'optimisation provient de l'introduction (localisation et type) des groupes polaires car elle nécessite de surmonter d'importants effets entropiques défavorables aboutissant à une compensation enthalpie-entropie :

- Les interactions spécifiques, telles que les liaisons hydrogènes, les ponts salins ou les contacts de van der Waals, représentent des gains enthalpiques qui diminuent la valeur du k_{off}, tandis que la désolvatation des groupes polaires entraîne une pénalité enthalpique qui augmente la valeur du k_{on}.
- Les gains entropiques sont généralement associés à la désolvatation du ligand lors de sa liaison avec la protéine cible, ce qui augmente la valeur du k_{on}, tandis que des changements conformationnels structurants au niveau du ligand ou du récepteur

entraînent des pénalités entropiques qui diminuent la valeur du k_{on} tout en augmentant celle du k_{off}.

L'application de certaines règles d'optimisation contribue à limiter la compensation enthalpie-entropie (Freire, 2009) :

- Les interactions spécifiques (liaisons hydrogènes) doivent être dirigées vers des régions structurées de la protéine afin de minimiser les effets structurants et les changements entropiques compensatoires et optimiser à la fois le k_{on} et le k_{off}.
- Bien que, la plupart du temps, l'entropie de désolvatation soit favorable, il arrive qu'elle soit défavorable ; elle peut être surmontée en modifiant la taille ou la géométrie d'un groupement apolaire, afin de remplir une cavité hydrophobe par exemple et optimiser ainsi à la fois le k_{on} et le k_{off}.

(3) Modulation de la constante de dissociation (k_{off})

Une étude sur plus de 2000 médicaments, qui se lient aux GPCR, aux protéines kinases et à d'autres enzymes, provenant de la littérature et de la base de donnée du groupe Pfizer, a montré que les médicaments de poids moléculaire et de lipophilicité (cLogP) élevés avaient tendance à avoir un plus long temps de résidence (Miller et al., 2012). Cependant, cette corrélation du temps de résidence avec le poids moléculaire peut refléter la corrélation connue du poids moléculaire avec l'affinité (à savoir la stabilisation de l'état lié) ou à une corrélation avec la hauteur de la barrière de l'état de transition (déstabilisation de l'état de transition). Dans la même étude, un long temps de résidence a aussi été corrélé à un nombre plus élevé de liaisons rotables. Un ligand plus flexible peut adopter un plus grand nombre de conformations ce qui peut conduire à une probabilité moindre du ligand à adopter la conformation requise pour se dissocier du site de liaison (Waring et al., 2015). Une étude similaire, contenant plus de 1800 composés se liant au récepteur dopaminergique D2 (GPCR), a aboutie aux mêmes conclusions que les composés les plus hydrophobes et de plus grand poids moléculaire ont tendance à avoir un temps de résidence plus long (Tresadern et al., 2011). Plus récemment, trois études illustrent bien cette corrélation del'augmentation du cLogP avec une réduction du k_{off} (Soethoudt et al., 2018; Spagnuolo et al., 2017; Yoshikawa et al., 2018).

La flexibilité des récepteurs joue souvent un rôle important dans la modulation de la cinétique de liaison (Pan et al., 2013). Nous avons vu précédemment (cf. I.E.2) deux modèles

d'interaction récepteur-ligand : l'ajustement induit et la sélection conformationnelle. Ces deux modèles offrent des moyens d'augmenter le temps de résidence, en plus d'être des MMA efficaces évitant la compétition avec le ligand endogène (cf. I.D.2). De fait, le temps de résidence dépend de la vitesse du changement de conformation : un changement de conformation lent du récepteur permet d'augmenter le temps de résidence du ligand. Il a été suggéré que la plupart des médicaments présentant un long de résidence agissent via un mécanisme impliquant un changement de conformation du récepteur (Amaral et al., 2017; Copeland, 2011; Lu et al., 2018; Swinney, 2006).

Prenons comme exemple les inhibiteurs de protéine kinase (PK) qui occupent une place importante dans la littérature sur la cinétique de liaison. Dans la majorité des cas, les modifications du temps de résidence ont été associées à des changements de conformation des protéines, en particulier à des mouvements de la "boucle d'activation" impliqués dans l'activation de la PK. Deux inhibiteurs de CDK2 de type I mais possédant des pharmacophores différents, R547 et le Roniciclib, présentent un long temps de résidence comparés aux analogues de leurs séries respectives. Dans ces deux cas, l'observation des structures cristallographiques a permis d'associer l'augmentation du temps de résidence à l'adaptation conformationnelle de la boucle d'activation (Ayaz et al., 2016). Dans le cas de l'inhibiteur de type II de la PK p38α, BIRB796 (inhibiteur de type II dérivés d'aryl-pyrazole), sa liaison induit un changement de conformation de la boucle d'activation qui entraine le changement de conformation de la protéine d'une conformation active (*DFG-in*) à une conformation inactive (*DFG-out*). Ce changement conformationnel est tenu pour responsable du long temps de résidence observé pour BIRB796 (Pargellis et al., 2002). Pour CDK8, les auteurs ont pensé au début que le même mécanisme justifiait les longs temps de résidence obtenus pour d'autres inhibiteurs de type II dérivés d'aryl-pyrazole (Schneider et al., 2013). Cependant, des inhibiteurs de cette même série ayant un court temps de résidence induisent également ce changement conformationnel de *DFG-in* à *DFG-out*. Il a alors été proposé, à partir de l'analyse de structures cristallographiques, que la formation d'interactions hydrogène avec une région très stable et conservée de la famille des PK (région charnière ou *hinge*) et la formation d'interactions hydrophobe contribuent à l'augmentation du temps de résidence (Schneider et al., 2013). Notre analyse de ce jeu de données a révélé d'autres facteurs influençant le temps de résidence (cf. III.A.1 et III.B). Autre exemple, une série d'inhibiteurs de p38α MAP de type $I^{1/2}$, ayant de longs temps de résidence, a été conçue à partir d'un inhibiteur de type I de court temps de résidence. Il a été suggéré que la dissociation lente des inhibiteurs de type $I^{1/2}$ est

associée aux interactions qu'ils forment avec un réseau hydrophobe nommé *R-spine* qui stabilise l'état lié du ligand (**Figure 11**) (Lu et al., 2018; Walter et al., 2017; Wentsch et al., 2017). Ce motif *R-spine,* responsable des mouvements respiratoires de la protéine, est conservé et bien décrit dans la famille des PK. A côté de ces inhibiteurs réversibles, une autre stratégie pour augmenter le temps de résidence des composés est la conception de molécules se liant de façon covalente aux PK. Parmi les plus connus d'entre eux, on trouve l'afatinib et l'ibrutinib (Lu et al., 2018).

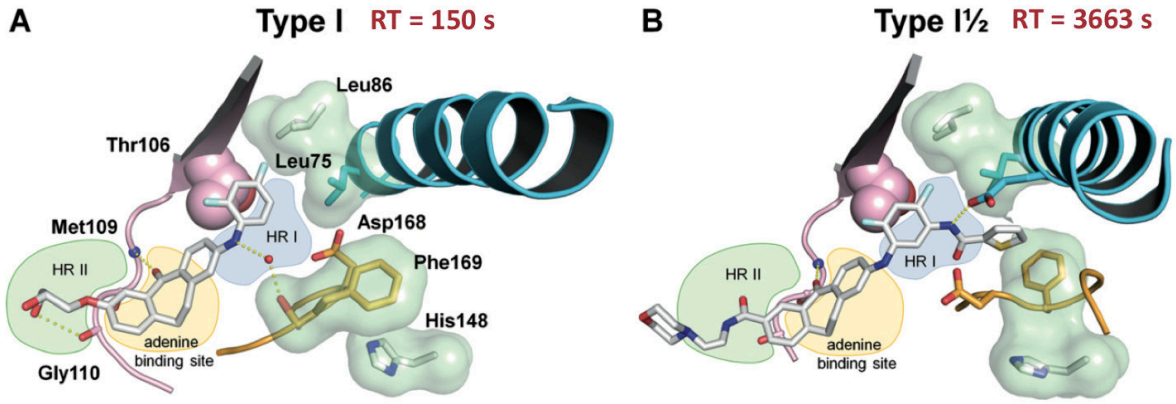


Figure 11 : Comparaison du mode d'interaction de deux inhibiteurs de p38α MAP kinase (Wentsch et al., 2017).
A) Skepinone-L complexé à p38 α MAP kinase (code PDB : 3QUE) : le réseau hydrophobe nommé *R-spine* (représenté en surface vert clair) n'est pas assemblé. **B)** inhibiteur 3 complexé à p38 α MAP kinase (code PDB : 5TBE) : un mode de liaison similaire est observé mais la *R-spine* est assemblée et stabilisée par l'inhibiteur. RT : *Residence Time*.

Comme nous l'avons vu dans la partie I.E.3.b)(2), l'eau influence la thermodynamique de liaison d'un complexe protéine-ligand. L'effet hydrophobe n'est pas une notion nouvelle et son impact sur la stabilisation du complexe protéine-ligand est bien connu (Böhm and Klebe, 1996). La solvatation des zones hydrophobes du site de liaison est défavorable, tout comme celle des groupements apolaires du ligand durant le processus de dissociation. Plusieurs études ont montré que la modulation de la barrière énergétique de solvatation permet de moduler le k_{off} (Liu et al., 2010; Schuetz et al., 2018). Par des approches expérimentales et numériques, Schmidtke et ses collègues ont montré que la formation d'interactions hydrogène protégées de l'eau (*water-shielded hydrogen bonds*) résulte en un complexe plus stable et est donc une façon d'augmenter le temps de résidence (**Figure 12**) (Schmidtke et al., 2011). Ils mettent en lumière le fait que la stabilité cinétique d'un complexe impliquant des liaisons hydrogènes dépend de leurs degrés d'exposition (Schmidtke et al., 2011). Gao *et al.*

quantifient cette effet et révèlent que l'énergie de la liaison hydrogène peut être augmentée de 1.2 kcal.mol^{-1} dans un environnement hydrophobe (Gao et al., 2009).

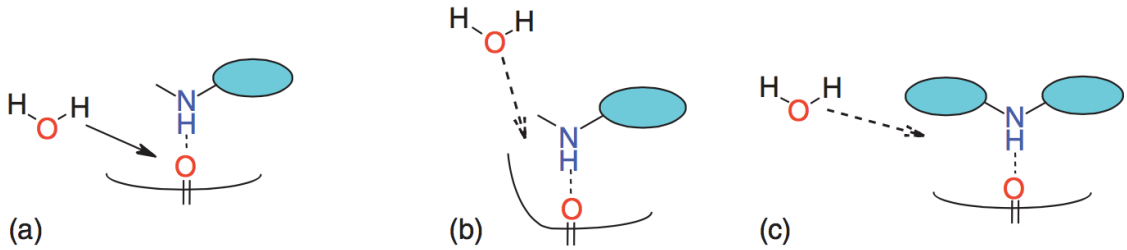


Figure 12 : Impact du degré d'exposition d'une liaison hydrogène au solvant sur la vitesse de dissociation (Waring et al., 2015).
(a) L'eau a facilement accès au site d'interaction et peut former des contacts solvant-soluté lorsque le complexe se dissocie, stabilisant l'état de transition et entraînant une dissociation rapide; **(b)** l'accès à l'eau est entravé par des contraintes stériques de la poche protéique, ce qui signifie que le complexe doit se dissocier davantage avant que des contacts solvant-soluté puissent se former, entraînant une dissociation plus lente; et **(c)** analogue à la situation en (b) sauf qu'ici des groupements chimiques du ligand plutôt que la poche protéine entravent l'accès à l'eau.

(4) Bilan sur la modulation du k_{on} et du k_{off} au niveau moléculaire

Nous venons de voir les facteurs structuraux et les fondements thermodynamiques agissant à l'échelle de l'interaction protéine-ligand, modulant ainsi la cinétique de liaison (k_{on} et k_{off}) et impactant l'occupation du site de liaison au cours du temps. En voici un résumé :

- L'accessibilité au site de liaison (Pan et al., 2013)
- Les interactions protéine-ligand : pour le k_{on} les interactions électrostatiques à longue portée jouent un rôle crucial (Fedosova et al., 2002; Radić et al., 1997; Spaar et al., 2006) tandis que pour le k_{off} ce sont plutôt les interactions de type liaison hydrogène, liaison ionique et interactions hydrophobe (Guo and IJzerman, 2018; Lu et al., 2018; Pan et al., 2013; Strasser et al., 2017).
- L'énergie de solvatation et désolvation ainsi que l'effet hydrophobe, tous deux étroitement liés, impactent fortement le k_{on} et le k_{off} (Deganutti et al., 2017; Mondal et al., 2014; Pearlstein et al., 2013; Schuetz et al., 2018; Setny et al., 2013). Le nombre de groupements polaires et apolaires du ligand ainsi que leur positionnement par rapport aux régions polaires et apolaires du site de liaison impactent aussi bien la cinétique d'association (k_{on}) que celle de dissociation (k_{off}) (Freire, 2015; Klebe, 2015).

- Les liaisons hydrogènes protégées de l'eau sont plus stabilisantes que celles exposées au solvant et augmentent donc le temps de résidence (Gao et al., 2009; Schmidtke et al., 2011).
- La flexibilité conformationnelle : le nombre de liaisons rotables du ligand est positivement corrélé au k_{off} et il a été observé que la rigidification du ligand à l'entrée du site de liaison contribue à augmenter le k_{on} (Alsallaq and Zhou, 2008; Fleck et al., 2012). Concernant la flexibilité de la cible, nous avons vu que les mécanismes d'action impliquant un changement conformationnel de la cible augmentent le temps de résidence (Copeland, 2011; Swinney, 2006; Wentsch et al., 2017). Lorsque l'association du ligand s'accompagne du repliment d'une région désordonnée de la cible en région ordonnée, la perte entropique ralentit le processus d'association, mais ralentit également le processus de dissociation (Luckner et al., 2010).

c) Intérêt particulier pour la constante de dissociation (k_{off})

Cependant, toutes ces considérations sont faites à l'échelle moléculaire du complexe protéine-ligand, loin de la complexité et de la variabilité des systèmes biologiques. De fait, *in vivo*, la cible est exposée à une concentration de médicaments variable qui dépend de facteurs pharmacocinétiques tels que le taux d'absorption et de distribution. Or, les paramètres de liaison (k_{on}, k_{off}, K_d) sont mesurés *in vitro,* dans des conditions expérimentales ne représentant pas les conditions *in vivo* car les concentrations en cible et en médicament restent inchangées au cours de l'expérience. Cependant, bien que le k_{on} et le k_{off} contribuent tous deux à augmenter le taux d'occupation de la cible, le k_{off} (s^{-1}) contrairement au k_{on} ($M^{-1}.s^{-1}$) ne dépend pas de la concentration en médicament, ainsi que le traduisent leurs unités respectives (**Figure 13**). Par conséquent, une faible valeur de k_{on} peut être compensée en augmentant simplement la concentration de médicament (c'est-à-dire la dose). Cela signifie aussi que de bonnes valeurs de k_{on} et K_d mesurées *in vitro* ne sont pas garantes d'une bonne efficacité de la molécule.

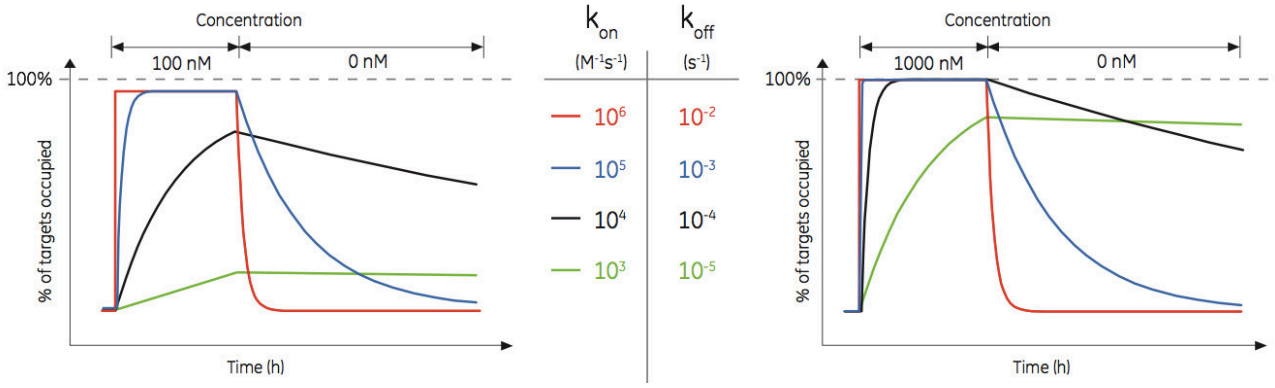


Figure 13 : Effet de la variation du k_{off} et du k_{on} sur le taux d'occupation de la cible pour quatre composés d'affinité identique ($K_d = k_{off} / k_{on}$ = 10 nM) (Hämäläinen, 2014).
Une concentration de 100nM (gauche) et 1000nM (droite) est utilisée. La ligne horizontale représente 100% des cibles bloquées.

Ainsi, comme le k_{off} est le seul paramètre parmi les 3 (k_{on}, k_{off}, K_d) qui ne dépend pas de la concentration locale en médicament libre, il a été suggéré qu'il était le plus prédictif de l'efficacité *in vivo*. C'est la raison pour laquelle la communauté scientifique s'est prioritairement intéressée au k_{off}. Le k_{off} ne dépend que des interactions entre le médicament et sa poche de liaison ; l'optimiser revient à optimiser sa structure chimique. En outre, le k_{on} est limité par le taux de diffusion des deux partenaires de liaison dans une solution physiologique. Or, il n'existe pas de mécanisme permettant de surmonter la limite de diffusion. La diffusion, la désolvatation et l'orientation du ligand à l'approche de la cible, facteurs affectant le k_{on}, sont difficiles à maitriser de manière systématique (Copeland et al., 2006; Keserü, 2015).

F. Impact de la cinétique de liaison sur la réponse pharmacologique

1. Constante d'association et sélectivité

Plusieurs études ont pointé l'importance de la constante cinétique d'association dans la sélectivité du médicament (Folmer, 2017; Freissmuth et al., 2015; de Witte et al., 2016b). De fait, si l'association est trop lente la molécule peut diffuser hors du tissu cible avant d'entrer en contact avec son site de liaison. Hasenhuetl et ses co-auteurs révèlent dans leur étude que la sélectivité du methylphénidate et du desipramine, respectivement pour le transporteur de la serotonine (SERT) et de la dopamine (DAT), s'explique par leurs vitesses d'association (k_{on}) et non leurs temps de résidence (Hasenhuetl et al., 2015).

2. Constante de dissociation et durée d'action : efficacité, dose et toxicité

Comme nous l'avons vu dans la partie I.D.1.b), la majorité des médicaments présentent un MMA « non à l'équilibre ». Une dissociation lente, autrement dit un long temps de résidence, constitue l'un de ces MMA efficaces. L'impact positif de l'augmentation du temps de résidence sur l'efficacité *in vivo* a été démontré sur les GPCR (Hothersall et al., 2016) tels que le récepteur A2 de l'adénosine A2A (Guo et al., 2012), le récepteur à C-C chimiokine de type 5 (CCR5) (Watson et al., 2005), le récepteur adrénergique β2 (Tee et al., 2007) et plus récemment, sur le récepteur métabotropique du glutamate 2 (mGluR2) (Doornbos et al., 2017) et le récepteur de l'histamine H1 (H1R) (Bosma et al., 2017).

Dans la famille des kinases, la lente dissociation de l'inhibiteur lapatinib est corrélée à une baisse prolongée de la phosphorylation de la tyrosine kinase dans les cellules tumorales (Wood et al., 2004). Le composé 584, un analogue de l'imatinib à dissociation lente, inhibe la kinase Abl avec des effets durables (Puttini et al., 2008). Récemment, l'optimisation guidée par le temps de résidence de molécules ciblant la kinase TTK a conduit à la conception d'inhibiteurs ayant une activité antiprolifératives puissante (Uitdehaag et al., 2017). Des études sur d'autres cibles thérapeutiques montrent aussi cet effet positif de l'allongement du temps de résidence sur l'efficacité *in vivo* (Costa et al., 2016; Lee et al., 2014; Ramos et al., 2018).

De nombreux médicaments commercialisés ont évolué pour fournir une inhibition prolongée (**Tableau 4**) (Swinney, 2004) tel que le Candesartan plus efficace que le Losartan alors qu'ils se lient à la même poche de fixation et ont des profils pharmacocinétiques similaires (Bakris et al., 2001; Fuchs et al., 2000; Lacourcière, 1999). Le Candesartan a un temps de demi dissociation ($t_{1/2} = \ln(2)/k_{off}$) de 112 min alors que le Losartan de 2.5 min (Fuchs et al., 2000). Cependant, la supériorité de Candesartan par rapport au Losartan a également été attribuée au phénomène de liaison répétée (*ligand rebinding*) favorisé par le fait que le Candesartan s'accumule dans la membrane à proximité de sa cible, qui est transmembranaire (le récepteur de l'angiotensine II) (Vauquelin, 2015). Dans certains cas, l'allongement de la durée d'action par augmentation du temps de résidence permet de réduire la dose du médicament. C'est le cas de l'Ipratropium ($t_{1/2}$ = 34.7h) où une prise par jour suffit tandis que le Tiotropium ($t_{1/2}$ = 0.26 h) nécessite 4 prises par jour (Durham, 2004).

Classe	Dissociation rapide (Surmontable)	Dissociation lente (Insurmontable)
ARBs (*Angiotensin II Receptor Blockers*)	Losartan	Candesartan
Antihistaminique	Diphénhydramine	Desloratadine
Antimuscarinique	Ipratropium	Tiotropium

Tableau 4 : Exemples de médicament ayant un MMA n'impliquant pas de toxicité et ayant évolué en des médicaments ayant un long temps de résidence.

Il est important de rappeler que l'allongement du temps de résidence n'est pas toujours souhaitable car il peut induire des effets secondaires (Kapur and Seeman, 2001; Vauquelin et al., 2012). Dans le cas où le MMA du médicament implique une toxicité (*toxicity based mechanism*), les nouveaux médicaments d'une classe donnée ont plutôt tendance à avoir des vitesses de dissociation plus rapide (Swinney and Anthony, 2011). Dans le cas de la classe des inhibiteurs de la cyclo-oxygénase (**Tableau 5**), l'aspirine a un MMA irréversible qui lui procure une fonction particulière, celle d'être un anticoagulant en plus d'être un anti-inflammatoire. Or cet effet anticoagulant n'est pas toujours voulu ; d'où la nécessité de développer d'autres médicaments de la même classe tels que l'Indométacine et l'Ibuprofène ayant des cinétiques différentes (**Tableau 5**).

Inhibiteurs de la cyclo-oxygénase	Cinétique de dissociation
Aspirine	Irréversible
Indométacine	Dissociation réversible lente
Ibuprofène	Dissociation réversible rapide

Tableau 5 : Exemples de médicament ayant un MMA impliquant une toxicité et ayant évolué en des médicaments ayant une dissociation plus rapide.
Ces trois inhibiteurs de la cyclo-oxygénase se lient au même site de liaison.

3. Mauvais traduction de la réponse *in vitro* en réponse *in vivo* : que manque t-il ?

Georges Vauquelin déclare dans sa publication « […] *Pharmacokinetics usually prevails over binding kinetics* » (Vauquelin, 2016) ce qui signifie que le profil pharmacocinétique d'un médicament prévaut généralement sur les constantes cinétiques. Lors de ma participation au

congrès annuel du *K4DD* (https://www.k4dd.eu/k4dd-2017-berlin-meeting/general-information/), plusieurs intervenants ont exprimé cette idée. Comme illustré sur la **Figure 14**, les constantes cinétiques constituent un pont entre la pharmacocinétique et la pharmacodynamique : sans un profil pharmacocinétique correcte, l'optimisation du k_{on} et du k_{off} serait vaine et ne produirait pas de réponse *in vivo*. D'ailleurs, il est communément admis que l'optimisation du temps de résidence d'un composé n'est pertinente et ne peut avoir un impact sur la réponse pharmacologique que si sa valeur est de l'ordre de grandeur ou plus grande que le temps de demi vie du composé (Dahl and Akerud, 2013). Le temps de demi-vie est le temps nécessaire pour que sa concentration sanguine dans l'organisme diminue de moitié. Si cela n'est pas le cas, on observera une faible traduction de l'activité observée *in vitro* et caractérisée par les mesures de K_d, k_{off} et k_{on} en efficacité *in vivo* (Dahl and Akerud, 2013).

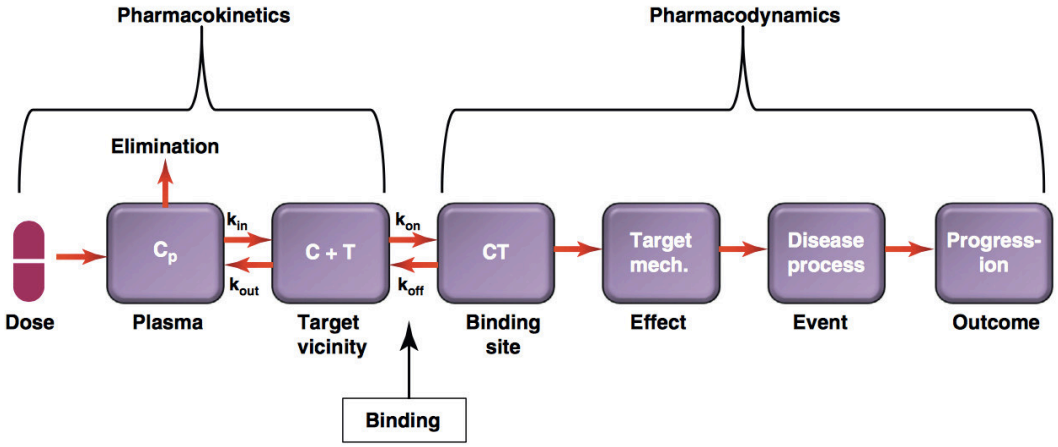


Figure 14 : Le k_{on} et le k_{off} : un lien entre la pharmacocinétique et la pharmacodynamique (Dahl and Akerud, 2013).
Schéma du processus cinétique de l'action du médicament, de l'administration du médicament à la progression de la maladie. Cp : Concentration plasmatique du composé ; C : composé ; T : cible ; Target mech. : mécanisme de la cible.

Cependant, certains médicaments font exception à la règle : leurs propriétés pharmacocinétiques ainsi que leurs temps de résidence ne permettent pas d'expliquer à eux seuls leurs longues durées d'action (Vauquelin, 2010, 2015). Il a été découvert que pour ces médicaments, d'autres phénomènes, dit micro-pharmacocinétiques (c'est à dire au voisinage de la cible), interviennent et augmentent leurs durées d'action. Parmi ces phénomènes, on peut notamment citer i) l'accumulation du médicament à la membrane, ce qui augmente la concentration locale du composé pour les cibles membranaires, et ii) la liaison répétée d'un même médicament à la cible (*ligand rebinding*) (Vauquelin, 2010, 2015; Vauquelin and Charlton, 2010).

Plus généralement, il est important de noter qu'il existe un ensemble de processus schématisés en partie sur la **Figure 14**, qui peuvent influencer la cinétique de liaison protéine-ligand dans un contexte *in vivo* et doivent donc être pris en compte pour une meilleure traduction de la réponse *in vitro* en réponse *in vivo* :

- la compétition entre le médicament et les ligand(s) et ou substrat(s) endogènes (cf. I.D.2).
- la liaison non-spécifique du médicament avec des protéines dans le sang et dans le tissu cible. De fait, un médicament peut se lier aux protéines plasmatiques sanguines. Mais seule la forme libre peut être distribuée vers les tissus (et donc la cible thérapeutique et des cibles secondaires) et être éliminée (Peletier et al., 2010).
- la concentration en médicament dans l'environnement local de la cible peut être modulée par des phénomènes micro-pharmacocinétiques tels que l'accumulation du médicament dans la membrane et le phénomène de liaison répétée (Dahl and Akerud, 2013; Sykes et al., 2014; Vauquelin, 2010, 2015).
- la concentration de la cible dans les tissus (Vauquelin, 2010).
- le renouvellement de la cible (*target turnover*) (Tonge, 2017).

G. Les différents niveaux d'étude de la cinétique

Les différents points sur la cinétique de liaison abordés dans les parties précédentes nous ont implicitement amenés à présenter différents moyens d'étude de la cinétique de liaison. Nous allons consacrer cette partie à donner brièvement une vision d'ensemble des différents niveaux d'étude de la cinétique. Nous porterons une attention particulière sur les challenges actuels dans ces différents domaines d'étude. Nous finirons cette partie en présentant plus en détails les méthodes de simulations numériques dédiées à la prédiction de la cinétique de liaison.

1. Relation quantitative structure-cinétique

La première étude SKR (*Structure Kinetics Relationships*) menée sur un grand nombre de molécules *drug-like* a été réalisée par Miller *et al.* en 2012 (Miller et al., 2012). Ils ont fourni des statistiques sur la distribution des propriétés physicochimiques selon les valeurs de constantes cinétiques k_{on} et k_{off} en examinant les données cinétiques du domaine public et un ensemble de données internes non publiées du groupe Pfizer. Aucun modèle QSKR

(*Quantitative Structure Kinetics Relationship*) n'a été dérivé de cette étude. En ce qui concerne les études QSKR à proprement parler, celles existantes répondent à un besoin très spécifique. Dans ces deux premières études (Andersson and Hämäläinen, 2006; Choulier et al., 2002), les modèles ont été développés à chaque fois sur des petits jeux d'antigènes peptidiques (n<30) ciblant une protéine particulière. Plus récemment, Qu *et al.* ont développé trois modèles pour la prédiction du k_{on}, k_{off} et du K_d respectivement en utilisant un jeu de 37 inhibiteurs de la protéase HIV-1 (Qu et al., 2016). La communauté scientifique manque de modèles QSKR prédictifs, plus globaux, avec un domaine d'applicabilité plus large et utilisables pour l'optimisation des constantes cinétiques de candidat-médicaments. Le consortium K4DD (cf. I.E.3.a)) est conscient de ce désert scientifique (Schuetz et al., 2017) et attribue cela à la pénurie de données cinétiques (en quantité et en qualité) dans les bases de données publiques tel que la ChEMBL (Gaulton et al., 2012) permettant la construction de modèle robuste. Par conséquent, à la fin du consortium (2012-2017) les données de la base de données K4DD ont été publiées dans la ChEMBL (Schuetz, 2018).

2. PK/PD modèles

Les modèles PK/PD (pharmacocinétique/pharmacodynamique) sont des expressions mathématiques ayant pour objectif de modéliser la réponse pharmacologique au cours du temps en fonction de la dose en intégrant les composantes pharmacocinétiques et pharmacodynamiques (Csajka and Verotta, 2006). *In vivo,* plusieurs phénomènes peuvent avoir lieu et influencer la cinétique de liaison (k_{on} et k_{off}), et ont donc un effet sur l'occupation de la cible (cf. I.F.3). Afin de modéliser correctement l'occupation de la cible et donc la réponse pharmacologique, l'ensemble de ces phénomènes doit être pris en compte dans un modèle PK/PD compartimental. Cela rend les modèles PK/PD longs et difficiles à mettre en place : chacun des processus, ainsi que leurs constantes cinétiques respectives, doivent être déterminés et schématisés par des compartiments (exemple : inhibition compétitive, liaison répétée à la cible, cinétique d'association à des cibles secondaires etc.) (**Figure 15**) (de Witte et al., 2016a).

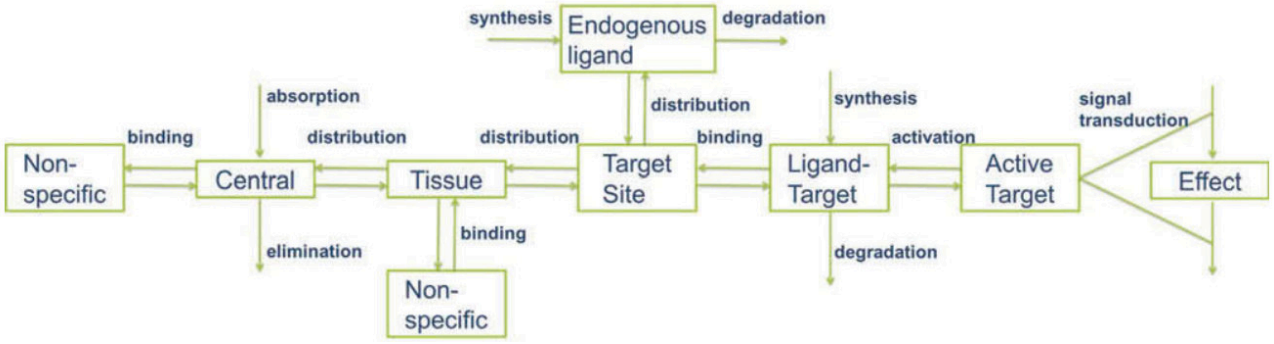


Figure 15 : Représentation schématique en compartiments des processus cinétiques interconnectés qui déterminent la cinétique d'occupation de la cible et de la réponse pharmacologique (de Witte et al., 2016a).
Le compartiment central (*Central*) représente le sang, le compartiment *Target Site* représente l'environnement direct autour de la cible où la liaison se produit.

C'est justement cette compartimentation des processus qui attribuent au modèle PK/PD un intérêt majeur dans le processus de recherche de médicaments (Tuntland et al., 2014; Yu and Wilson, 2010). De fait, l'analyse intégrée de tous les contributeurs à la réponse pharmacologique permet d'optimiser de façon sélective les paramètres les plus pertinents du candidat-médicament (Tuntland et al., 2014; Yu and Wilson, 2010). De plus, avec ces modèles, l'impact de la compétition endogène (de Witte et al., 2016a), de la pharmacocinétique (Taneja et al., 2017; Vauquelin, 2010; de Witte et al., 2017) et de la liaison non-spécifique du médicament (Peletier et al., 2010) sur les constantes cinétiques de liaison et donc sur l'occupation de la cible ont pu être modélisés et mis en lumière. Ces études ainsi que celles sur l'analyse des MMA des médicaments (cf. I.D.1.b)), convergent vers un même constat : en plus de la prise en compte du profil pharmacocinétique du médicament, bien caractériser la cible (renouvellement de la cible, concentration de la cible etc.) et son mécanisme d'action (compétitif, non compétitif, liaison répétée) sont essentiels pour prédire la réponse *in vivo*. Un modèle PK/PD proche du contexte *in vivo* sera un modèle prenant en compte l'ensemble de ces processus (Tuntland et al., 2014).

3. Mesures expérimentales des constantes cinétiques

Malgré l'existence d'un large panel de méthodes expérimentales, celles-ci restent limitées en raison du besoin i) de collectes de données résolues dans le temps, ii) de tests cinétiques à haut débit (très important pour une utilisation industrielle) et, iii) du coût de ces méthodes (Schuetz et al., 2017) (**Figure 16**).

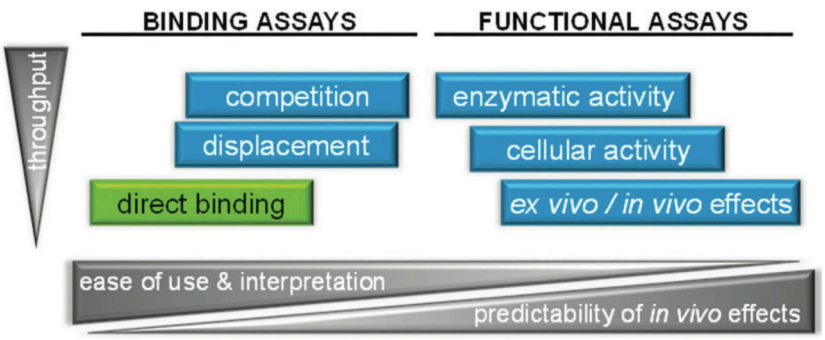


Figure 16 : Comparaison des formats de tests expérimentaux pour l'investigation directe (verte) et indirecte (bleue) de la cinétique de liaison (Georgi et al., 2017).
Les tests de liaison sont les plus appropriés pour la détermination (à haut débit) des constantes cinétiques de liaison. Les tests basés sur la compétition présentent un débit plus élevé, car elles permettent la détermination de k_{on}, k_{off} et K_d dans une même expérience, alors que les tests de déplacement ne déterminent que le k_{off} et nécessitent en plus une étape de pré-incubation. En revanche, les dosages fonctionnels plus complexes et difficiles à interpréter, reflètent mieux la réalité physiologique, mais la quantification de la cinétique de liaison est compliquée ou impossible (en particulier pour les essais cellulaires *in vivo*).

Les tests de liaison étant plus couramment utilisés que les tests fonctionnels, nous nous focaliserons sur ceux-ci dans ce paragraphe. La résonance plasmonique de surface (RPS ; SPR en anglais), qui est une technique sans marqueur, a été appliquée avec succès pour mesurer les constantes cinétiques de liaison des composés associés à des cibles solubles (Zheng et al., 2015). Cependant, la SPR est un test à débit moyen en raison du fait que les ligands sont principalement mesurés en série, et que le test suppose que la protéine soit purifiée et liée de manière stable à une surface (**Figure 17**). Cela rend cette méthode inadaptée aux protéines membranaires qui constituent la majorité des cibles médicamenteuses (GPCR) (Sriram and Insel, 2018).
Les méthodes utilisant un ligand radiomarqué présentent un débit faible à moyen (Georgi et al., 2017; Schuetz et al., 2017). Elles ne nécessitent pas la purification de la protéine de son environnement membranaire natif, ce qui les positionnent comme méthode de choix pour les GPCR. Cependant elles supposent la disponibilité d'un ligand marqué de haute affinité (de Witte et al., 2016a), et surtout elles ne présentent pas un débit élevé.

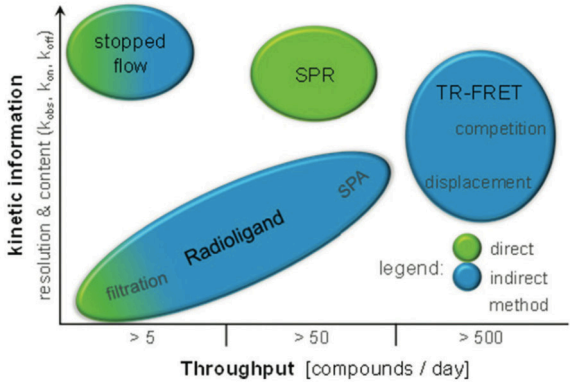


Figure 17 : Comparaison des principaux groupes de tests expérimentaux de liaison en termes de débit et de contenu d'information pour l'étude de la cinétique de liaison (Georgi et al., 2017).

Dans le but d'étudier la cinétique de liaison dans un environnement proche de celui natif et à plus haut débit, de nouvelles technologies ont été développées, soutenues par le consortium K4DD, telles que la technique basée sur le transfert d'énergie (TR-FRET ou BRET) et celle basée sur la microscopie. Ces deux techniques à hauts débits nécessitent que le ligand ou la cible soit marqué par fluorescence (Schuetz et al., 2017) ; beaucoup d'efforts ont été dédiés au développement de ligands fluorescents, surtout pour les GPCR, qui ont été appliqués avec succès (Christiansen et al., 2016; Klein Herenbrink et al., 2016; Nederpelt et al., 2016; Schiele et al., 2015b; Stoddart et al., 2016). Des perspectives futures sont envisagées pour toutes ces méthodes, concernant en priorité, l'augmentation du débit en améliorant notamment le matériel et les logiciels (Schuetz et al., 2017).

4. Méthodes de simulation numérique pour la prédiction des constantes cinétiques

Cette partie vise à présenter les différentes méthodes de simulations numériques développées pour la prédiction des constantes cinétiques de liaison (k_{on}, k_{off}). Le premier objectif est de donner une vision d'ensemble des méthodes existantes avec leur échelle de temps de simulation respectives. Plus l'échelle de temps simulé est grande (microseconde, milliseconde) et plus le temps de calcul sera important. Ce dernier est un point crucial pour une utilisation en industrie de ces méthodes, où des dizaines de composés doivent être analysés. Ensuite, les méthodes ayant été appliquées à la prédiction du temps de résidence (i.e. du k_{off}) seront présentées plus en détails, en précisant notamment leurs limites. Enfin, nous finirons cette partie en expliquant les raisons ayant motivées ce projet de thèse, autrement dit le développement de nouvelles méthodes numériques de prédiction de temps

Méthodes computationnelles pour la prédiction des constantes cinétiques de liaison		
Échelle de temps	**Constante cinétique de dissociation (k_{off})**	**Constante cinétique d'association (k_{on})**
10-100 nanosecondes	**Well tempered-Metadynamics (wt-MTD)** *(Sun et al., 2017)*	
	Scaled-MD (Biki tool) *(Mollica et al., 2015, 2016)*	
	τ-RAMD (τ-Random Accelerated MD) *(Kokh et al., 2018)*	
	Metadynamics (MTD) *(Callegari et al., 2017)*	
	Adiabatic bias MD + wt-MTD *(Bortolato et al., 2015)*	**Brownian Dynamics (BD)** *(Sung et al., 2010)*
	Umbrella sampling (US) *(Sun et al., 2015)*	
	Random accelerated MD (RAMD) + Steered MD (SMD) *(Niu et al., 2016)*	
1µs – 50 µs	**Molecular Dynamics (MD) + Adaptative Markov State Model (AMSM)** *(Doerr and De Fabritiis, 2014)*	**MD + AMSM** *(Doerr and De Fabritiis, 2014)*
	Bias-exchange MTD *(Pietrucci et al., 2009)*	**Bias-exchange MTD** *(Pietrucci et al., 2009)*
	Adaptative Multilevel Splitting (AMS) *(Teo et al. 2016)*	**wt-MTD** *(Tiwary et al., 2015)*
	WExplore tool *(Dickson and Lotz, 2017; Lotz and Dickson, 2018)*	
	wt-MTD *(Tiwary et al., 2015, 2017)*	
	wt-MTD + MSM *(Casasnovas et al., 2017)*	
	MD + BD (SEEKR tool) *(Jagger et al., 2018; Votapka et al., 2017)*	
>50 µs	**MD + Markov State Model (MSM)** *(Buch et al., 2011; Plattner and Noé, 2015)*	**MD + BD** *(Jagger et al., 2018; Votapka et al., 2017; Zeller et al., 2017)*
		Weighted Ensemble MD *(Zwier et al., 2016)*
		MD + MSM *(Buch et al., 2011; Plattner and Noé, 2015)*

Tableau 6 : Méthodes de simulation numérique pour le calcul des constantes cinétiques ligand-récepteur.
Les études où ces méthodes ont été appliquées sont répertoriées ainsi que l'échelle de temps de simulation correspondante. Le temps de simulation dépend des propriétés du système étudié ainsi que des méthodes utilisées.

de résidence.

a) Intérêt des méthodes de simulation numérique dans l'étude de la cinétique de liaison

Malgré les efforts fournis dans le développement de méthodes expérimentales et en dépit de leur nombre, la plupart de ces méthodes ne permettent pas la corrélation des données cinétiques avec les évènements structuraux se produisant à l'échelle atomique durant le processus d'association et de dissociation. De telles informations seraient d'une grande valeur pour aider les chimistes à concevoir et à synthétiser des composés présentant des paramètres cinétiques optimisés. Bien qu'il existe des méthodes expérimentales permettant de capturer des informations structurales sur l'état de transition, elles restent difficiles à appliquer et n'offrent pas autant d'information que les méthodes de dynamique moléculaire (Schiele et al., 2015a). De fait, la simulation de dynamique moléculaire est une technique puissante qui fournit une vue d'ensemble du processus de liaison, notamment les états intermédiaires, et les états de transition avec une résolution temporelle de l'ordre de la femtoseconde.

b) Diversité des méthodes de simulation numérique

Comme on peut le voir dans le **Tableau 6**, on retrouve les mêmes algorithmes et théories utilisés pour la prédiction du k_{on} que pour celle du k_{off}. Du fait de l'intérêt majeur du temps de résidence (cf. I.E.3.c)), le nombre d'études et de méthodes est plus élevé pour la prédiction du k_{off} que pour le k_{on}. Les méthodes nécessitant des plus longs temps de simulation (>1µs) fournissent des valeurs absolues des constantes cinétiques, tandis que celles ayant des temps de simulation de l'ordre de la nanoseconde fournissent des valeurs relatives ou qualitatives (rapide versus lent) en vue d'un classement des composés (excepté l'étude réalisée par (Sun et al., 2017) qui donne des valeurs absolues). Une autre tendance se dégage : la majorité des méthodes ayant un temps de simulation >1µs sont basées sur de la dynamique moléculaire dite « brute force », i.e. non biaisée, tandis que celles de l'ordre de la dizaine ou centaine de nanoseconde sont exclusivement des méthodes de simulation accélérées, dans lesquelles un biais a été ajouté.

c) Prédiction du temps de résidence par simulation numérique

(1) Méthodes basées sur de la dynamique moléculaire brute force

Malgré les progrès de la technologie informatique, qui ont considérablement amélioré la performance de la dynamique moléculaire brute force, celle-ci n'a été utilisée que pour simuler le processus de dissociation de systèmes simples : des petites molécules (fragments) ayant des temps de résidence de l'ordre de 10^{-9} s (variant de 8 ns à 140 ns) et se liant à une cible de petite taille (*FKBP : FK506-binding protein*) (Huang and Caflisch, 2011; Pan et al., 2017). Afin d'accélérer la simulation tout en conservant l'aspect non biaisé, divers théories et algorithmes, utilisant la dynamique moléculaire brute force (en combinaison ou non avec la dynamique brownienne) en tout atome et en solvant explicite, ont été développés pour prédire le k_{off}. Ces méthodes ont été appliquées sur des petits systèmes tels que la trypsine-benzamidine (Buch et al., 2011; Dickson and Lotz, 2017; Plattner and Noé, 2015; Teo et al., 2016; Votapka et al., 2017), qui présente un site de liaison relativement rigide, et sur le système modèle la β-cyclodextrine (Jagger et al., 2018). De plus, les molécules utilisées sont petites et présentent une vitesse de dissociation rapide ($k_{off} > 10^2$ s^{-1}). Ces méthodes incluent : i) l'outil SEEKR (Jagger et al., 2018; Votapka et al., 2017) et l'outil WExplore (Dickson and Lotz, 2017) donnant une valeur prédite de k_{off} à un facteur 10 près, ii) l'AMS (*Adaptative Multilevel Splitting*) donnant une valeur de k_{off} environ 2 fois plus élevée que la valeur expérimentale (Teo et al., 2016) et, iii) les MSM (*Markov State Models*) avec des résultats s'écartant des valeurs de référence expérimentales d'un facteur 10 (Buch et al., 2011) et d'un facteur 20 (Plattner and Noé, 2015). Cependant, les cibles thérapeutiques prédominantes sont des protéines plus grandes et plus souples, telles que les kinases ou les protéines membranaires (GPCR), qui lient souvent des médicaments larges et flexibles. Récemment, l'outil WExplore a été appliqué sur un système cliniquement pertinent, l'époxyde hydrolase liée à l'inhibiteur TPPU, avec une échelle de temps de résidence pertinente sur le plan pharmacologique (11 min) (Lotz and Dickson, 2018).

(2) Méthodes basées sur de la dynamique moléculaire accélérée

Cependant, malgré le gain en puissance de calcul grâce aux progrès informatiques, ces méthodes basées sur la dynamique moléculaire dite « brute force » restent trop coûteuses en temps de calcul, au regard de l'échelle de temps requise (de la milliseconde à l'heure) pour

simuler de nombreux processus de dissociation de ligands pertinents sur le plan clinique. Pour pouvoir être couramment utilisées en industrie, dans la phase d'optimisation des composés où des dizaines d'entre eux doivent être analysés, ces méthodes doivent être plus rapides. Compte tenu de ce besoin croissant, plusieurs protocoles utilisant des méthodes de simulation de dynamique moléculaire biaisées ont été développés et appliqués pour calculer la cinétique de liaison.

Mollica *et al*. ont appliqué la méthode *scaled-MD* (ou *smoothed-MD*), une méthode de classement qui ne nécessite pas la définition d'une coordonnée de réaction, sur plusieurs ligands de HSP90, de Grp78 (*Glucose regulated protein 78*), du récepteur A2a de l'adénosine (Mollica et al., 2015) et de glucokinase (Mollica et al., 2016). Cette méthode a été développée par la spin-off BiKi Technologies qui a établi une collaboration avec l'Institut de Recherche Servier (cf. I.E.3.a)). Les auteurs ont obtenu un bon classement du temps de résidence dans tous les cas (Mollica et al., 2015, 2016). A ce jour, cette méthode est la seule à avoir fourni des résultats probants sur un grand nombre de molécules *drug-likes* se liant à des cibles thérapeutiques d'intérêt. Le temps de résidence est estimé à partir du temps de simulation (t) au bout duquel le composé se dissocie de la protéine et se retrouve complétement solvaté (Mollica et al., 2015, 2016). Plus précisément, le temps de résidence (RT) est estimé par le temps de simulation à la puissance $1/\lambda$ ($RT = t^{1/\lambda}$), où λ est le facteur d'atténuation des barrières énergétiques dans l'algorithme de la méthode *scaled-MD*. L'énergie potentielle totale (U) est atténuée par le facteur λ qui est compris entre 0 et 1 : $\lambda \times U$. Il y a donc un risque accru de distorsion de la protéine. Par conséquent, dans un protocole de simulation par *scaled-MD,* un ensemble de contraintes est systématiquement appliqué sur tous les atomes lourds du squelette de la protéine, à l'exception de ceux du site de liaison, afin de conserver la protéine dans sa conformation native. Ces contraintes peuvent conduire à une description irréaliste du processus de dissociation simulé notamment si celui-ci implique par exemple un changement de conformation.

Deux études (Niu et al., 2016; Sun et al., 2015) abordent la question de la prédiction du k_{off} par un schéma similaire : un potentiel de force moyenne est généré et assimilé au profil d'énergie libre du processus de dissociation (Kumar et al., 1992; Zhu and Hummer, 2012). Dans la première étude, la RAMD (*Random Accelerated Molecular Dynamics*) combinée à la SMD (*Steered Molecular Dynamics*) ont été utilisées pour respectivement explorer les chemins d'accès des ligands et générer les potentiels de force moyenne (Niu et al., 2016) ; dans la deuxième étude, l'US *(Umbrella Sampling)* est utilisée pour réaliser ces deux étapes (Sun et al.,

2015). Pour deux composés donnés, la différence des barrières énergétiques des états de transition calculée à partir des potentiels de force moyenne est qualitativement en accord avec la différence des k_{off} mesurée (Niu et al., 2016; Sun et al., 2015).

Très récemment, une méthode basée sur la RAMD appelée τ-RAMD, a été appliquée sur 70 ligands de protéine HSP90α analogues à des médicaments et a démontré une bonne corrélation ($R^2 = 0.86$) entre le temps de résidence calculé et mesuré pour 78% des composés (les points aberrants ayant été enlevés) (Kokh et al. 2018). Dans la méthode τ-RAMD, une force supplémentaire orientée de manière aléatoire est appliquée sur le ligand pour faciliter sa dissociation. La direction de cette force est réaffectée de manière aléatoire lorsque la distance parcourue par le ligand dans un intervalle de temps défini est en dessous d'une distance seuil spécifiée.

Les méthodes dérivées de la métadynamique (MTD) ont été développées et appliquées en nombre pour la prédiction des constantes cinétiques de liaison (**Tableau 6**) (Bortolato et al., 2015; Callegari et al., 2017; Casasnovas et al., 2017; Pietrucci et al., 2009; Sun et al., 2017; Tiwary et al., 2015, 2017). Au total, on compte trois méthodes : la MTD originale qui est la version originale de la métadynamique (Callegari et al., 2017), la wt-MTD (Bortolato et al., 2015; Casasnovas et al., 2017; Sun et al., 2017; Tiwary et al., 2015, 2017) et la bias-exchange MTD (Pietrucci et al., 2009) (**Tableau 6**). Il est important de noter que pour une même méthode (par exemple la wt-MTD), ce n'est pas forcément la même approche de calcul du temps de résidence qui est envisagée. Parmi les études où l'échelle du temps de simulation est de l'ordre de la nanoseconde, trois approches de prédiction du temps de résidence ont été proposées : l'une utilisant la MTD originale et les deux autres la wt-MTD. L'approche de Bortolato (Bortolato et al., 2015) et l'approche de Tiwary (Salvalaglio et al., 2014; Tiwary and Parrinello, 2013) utilisent la wt-MTD mais estiment le temps de résidence différemment. Dans l'approche de Bortolato, le temps de résidence est estimé par la valeur maximale de l'énergie potentielle biaisée ajoutée pour passer du premier bassin énergétique (état lié) au prochain bassin énergétique (Bortolato et al., 2015). Les auteurs s'appuient sur l'équation d'Eyring et assimile la valeur maximale de l'énergie potentielle biaisée à l'énergie d'activation du processus de dissociation (ΔG_{off}). L'une des limites majeures de cette approche est qu'elle ne prend en compte qu'un seul état de transition et donc suppose que la cinétique du processus est en une étape. L'approche de Tiwary (Salvalaglio et al., 2014; Tiwary and Parrinello, 2013) estime le temps de résidence par le produit $\alpha(t) \times t$ où $\alpha(t)$ est un facteur d'accélération qui donne l'accélération obtenue grâce au biais ajouté lors de la métadynamique ; t est le temps

de simulation qu'il faut pour passer de l'état lié à l'état dissocié. L'application de cette formule n'est possible que si certaines conditions ont été remplies. Ces dernières supposent que le paysage énergétique sous-jacent au processus de dissociation présente peu de barrières hautes et pointues, résultant ainsi en un comportement global d'Arrhenius, c'est à dire une cinétique à une étape (Casasnovas et al., 2017; Salvalaglio et al., 2014; Tiwary and Parrinello, 2013; Tiwary et al., 2017). Malgré certains succès (Bortolato et al., 2015; Casasnovas et al., 2017; Tiwary et al., 2015, 2017), les approches de Bortolato et de Tiwary n'ont pas été capables de prédire correctement le temps de résidence du lapatinib, inhibiteur de la protéine kinase EGFR (Sun et al., 2017). La raison évoquée est la profondeur du site de liaison de EGFR qui rend difficile la définition d'une variable collective (i.e. une coordonnée de réaction) décrivant correctement le processus de dissociation (Sun et al., 2017). De même, en utilisant l'approche de Tiwary, Callegari *et al.* n'ont pas été en mesure de classer un ensemble d'inhibiteurs de la kinase 8 cycline-dépendante (CDK8) par leur k_{off} en accord avec l'expérience (Callegari et al., 2017). En analysant plus finement leurs résultats, ils émettent l'hypothèse que certaines conditions d'application de l'approche de Tiwary ont été violées du fait de la complexité de la surface d'énergie libre du processus de dissociation, présentant probablement plusieurs états de transition avant d'atteindre l'état non lié. Ils ont donc proposé une approche alternative basée sur la MTD originale conduisant à un classement en accord avec l'expérience, dans lequel des valeurs relatives de k_{off} ont été estimées à partir du temps de simulation nécessaire pour conduire les complexes ligand-récepteur à leur état dissocié. Cependant, dans cette étude, la cycline C, complexée à la protéine kinase et disponible dans les structures expérimentales (Schneider et al., 2013), n'a pas été gardée dans les simulations. Nos travaux sur l'importance de la cycline C dans le complexe CDK8-CycC (cf. chapitre II) et une étude très récente (Cholko et al., 2018) ont montré que la cycline C est vitale pour maintenir la structure et la dynamique de CDK8 ainsi que les interactions protéine-ligand.

H. Motivations de ce projet

Notre partenaire industriel l'Institut de Recherche Servier, porteur de ce projet de thèse, souhaite disposer d'une méthode de simulation numérique capable de fournir un classement selon le temps résidence d'une série de molécules *drug-like*. L'outil devra être simple d'utilisation, rapide (≈ 5 molécules par jour) et applicable à tout système pour permettre

l'analyse de touches (*hit*) ou de chef de file (*lead*) en phase d'optimisation des molécules. L'Institut de Recherche Servier n'est pas membre du consortium K4DD (cf I.E.3.a)) mais a, en revanche, établi un partenariat avec l'entreprise Biki Technologies qui est à l'origine du développement de la méthode *scaled-MD* implémentée dans l'outil Biki (Mollica et al., 2015, 2016). Ce partenariat a permis de tester l'outil Biki sur des données internes de l'Institut Servier ; des résultats concluant ont été obtenus et ont été publiés (Mollica et al., 2016). A l'initiation de ce projet de thèse en 2015, la majorité des méthodes présentées dans le **Tableau 6** n'étaient pas développées. De plus, les méthodes disponibles aujourd'hui présentent des limites contraignantes pour une utilisation en industrie (cf. I.G.4.c)). Les limites des méthodes de prédiction du temps de résidence nécessitant des simulations de l'ordre de la nanoseconde, et donc assez rapide pour permettre une utilisation en industrie, sont résumées dans le **Tableau 7**.

Méthodes	Limites
wt-MTD	- Variables collectives difficiles à choisir - Non applicable pour un processus de dissociation complexe (plusieurs états de transition)
Scaled-MD (Biki tool)	- Contrainte imposée sur le squelette de la protéine (sauf le site de liaison) pouvant conduire à une description incorrecte du processus
τ-RAMD	- La direction de déplacement du ligand est imposée et ce de façon aléatoire.
US **RAMD + SMD**	- Pas facile à utiliser en routine (analyse de la qualité de l'échantillonnage, etc.)
MTD	- Pas assez de recul sur cette méthode : une seule étude réalisée dans laquelle une partie de la cible (la cycline C) n'a pas été conservée dans le système simulé.

Tableau 7 : Limites des méthodes de simulation de dynamique moléculaire appliquées pour la prédiction du temps de résidence avec une l'échelle de temps de simulation de l'ordre de la nanoseconde.

Fort de cet état de l'art, ce projet de thèse a pour principal objectif de mettre en place une méthode rapide, simple et applicable à tout système permettant de fournir un classement d'une série de composés *drug-like* selon leur temps de résidence. Il s'agit également

d'explorer la relation structure-cinétique du jeu de données ayant servi de preuve de concept à travers l'analyse des interactions protéine-ligand et des différents termes d'énergie d'interaction (van der Waals, électrostatique, solvatation etc.) lors du processus de dissociation. Le développement de cet outil et les analyses structure-cinétique sont l'objet du chapitre III partie A et partie B, respectivement. L'outil est ensuite appliqué sur un jeu de données privé appartenant à l'Institut Servier, afin de vérifier que la méthode développée peut être appliquée sur un jeu de donnée externe et donner des résultats en accord avec l'expérience (chapitre III.A).

La méthode développée est une approche rapide où l'exploration de chaque micro-état tout au long du chemin, autrement dit l'échantillonnage, n'est pas suffisante pour permettre d'analyser plus finement la trajectoire et en extraire des grandeurs ayant un sens thermodynamique. Avec un échantillonnage plus important, on pourrait par exemple analyser le potentiel de force moyenne, qui peut être associé au profil d'énergie libre de dissociation (Kumar et al., 1992; Zhu and Hummer, 2012) et à partir duquel l'état de transition et les états intermédiaires pourront être identifiés. Par conséquent, le chapitre, plus exploratoire, a pour but de développer une méthode afin d'échantillonner plus finement le processus de dissociation, de façon à extraire le profil d'énergie libre et à identifier les états intermédiaires stables et les états de transition.

La première étape dans ce développement méthodologique est de choisir un système sur lequel la preuve de concept sera faite. Notre choix s'est porté sur la kinase 8 cycline-dépendante (CDK8) pour laquelle des données cinétiques homogènes sont disponibles (Schneider et al., 2013). CDK8 se présente sous forme d'un hétérodimère où elle est liée à la cycline C (CycC). CDK8, en complexe avec CycC, présente certaines particularités structurales et fonctionnelles comparé à ses homologues de la famille CDK (cf. chapitre I.I.5.b)(2)). Le chapitre II a pour rôle d'étudier la structure et la dynamique de ce complexe et notamment l'interaction entre CDK8 et la CycC. Outre la compréhension structurale et dynamique du complexe CDK8-CycC, cette étude permet de statuer quant à la nécessité ou non de garder la CycC dans nos simulations de DM.

I. La famille des protéines kinases

Dans cette partie, nous allons présenter la famille des protéines kinases (PK), dont CDK8, notre cible d'intérêt pour la preuve de concept pour notre méthode, est membre. De plus, la

méthode développée a par la suite été testée sur un jeu de données interne à Servier dont la cible est également une PK.

1. Rôle physiologique des PK

Les PK constituent l'une des principales familles de cible thérapeutique (Ferguson and Gray, 2018; Santos et al., 2017). Elles ont un rôle majeur dans la transduction intracellulaire de signaux physiologiques telles que le contrôle du cycle cellulaire et le métabolisme (**Figure 18**).

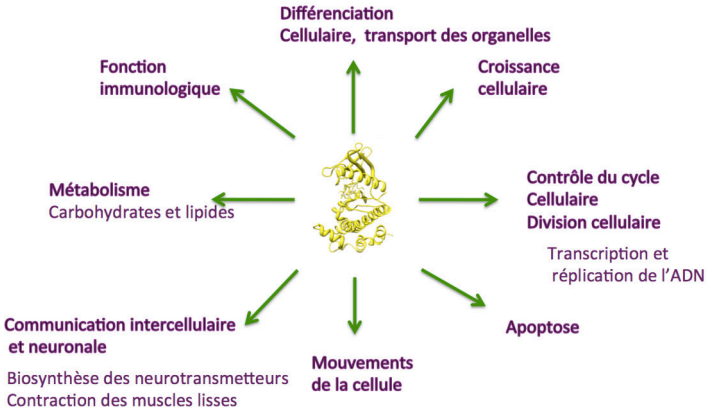


Figure 18 : Rôle physiologique des protéines kinases humaines

La dérégulation de leur fonction cause des désordres pathologiques importants tels que le cancer, les maladies inflammatoires, le diabète et les maladies neurodégénératives (Lahiry et al., 2010). Depuis l'autorisation de mise sur le marché en 2001 d'un premier inhibiteur de kinase, l'imatinib, le nombre d'inhibiteurs approuvés continue de croitre régulièrement pour atteindre 39 en janvier 2018 (Carles et al., 2018) (**Figure 19**).

Appartenant au groupe des transférases, les PK catalysent les réactions de phosphorylation par l'ajout d'un ion phosphate provenant de l'ATP (Adénosine TriPhosphate), sur une molécule cible appelée le substrat. Le substrat peut être une protéine (par exemple une autre PK), un sucre, un lipide ou un acide nucléique (**Figure 20**). Lorsque la phosphorylation se fait sur une sérine ou une thréonine du substrat protéique, on parle de sérine/thréonine kinase ; et sur la tyrosine on parle de tyrosine kinase. La réaction inverse de déphosphorylation est, quant à elle, catalysée par les phosphatases.

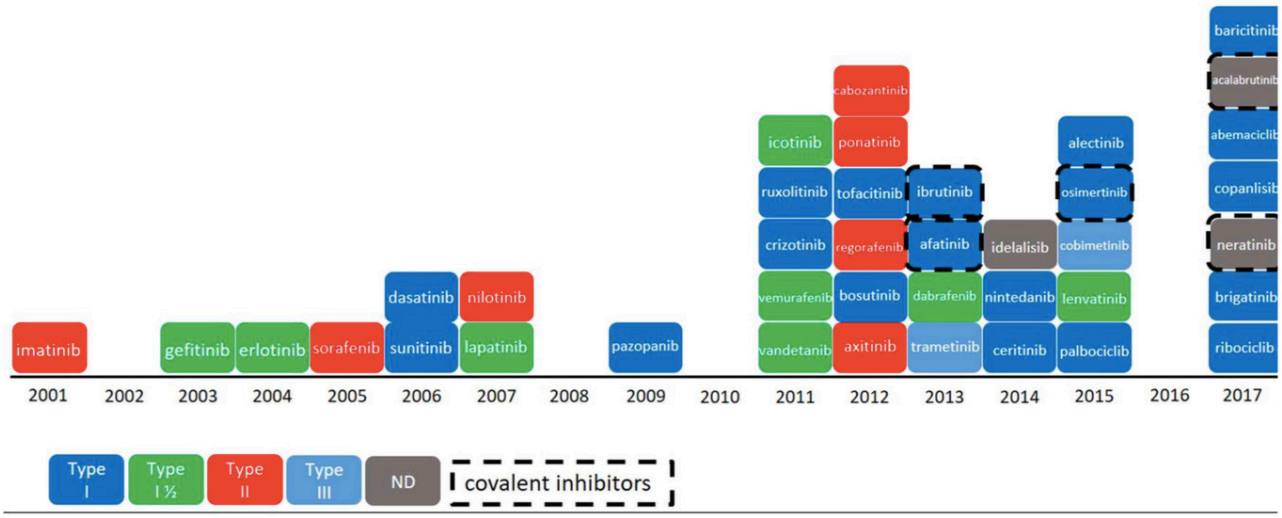


Figure 19 : Inhibiteurs de protéine kinase approuvés entre 2001 et 2017 (Carles et al., 2018).
Tous les inhibiteurs ont été approuvés par la FDA exceptés Icotinib and Baricitinib qui ont été approuvés par la CFDA (*China Food and Drug Administration*) et par l'EMA (*European Medicines Agency*) respectivement.

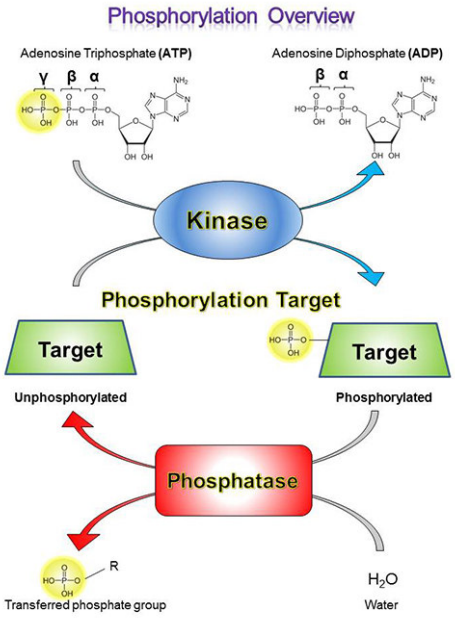


Figure 20 : Schéma de phosphorylation d'une protéine cible par une protéine kinase (Source : https://www.biolegend.com/phospho).
Le phosphate de l'ATP en position γ est transféré sur la protéine cible. La phosphatase catalyse la réaction inverse et libère le phosphate.

2. Structure

De manière générale, les PK eucaryotes présentent toutes le même repliement de leur domaine catalytique, dit domaine kinase, qui est très conservé. Long de 220 à 260 acides

aminés environ, le domaine kinase est formé de deux lobes : le lobe terminal N, le plus petit des deux lobes, et le lobe terminal C (**Figure 21**). Le lobe terminal N est formé de cinq feuillets β antiparallèles (β_1 – β_5) et d'une hélice α dite hélice αC. La boucle P (aussi appelée boucle riche en glycines) relie les 2 premiers brins β (β_1 et β_2) et est cruciale pour l'orientation et la liaison de l'ATP. Le lobe terminal N est relié au lobe terminal C par un coude très peu flexible et très conservé dit région charnière (ou *hinge* en anglais) situé entre le brin β_5 et l'hélice D.

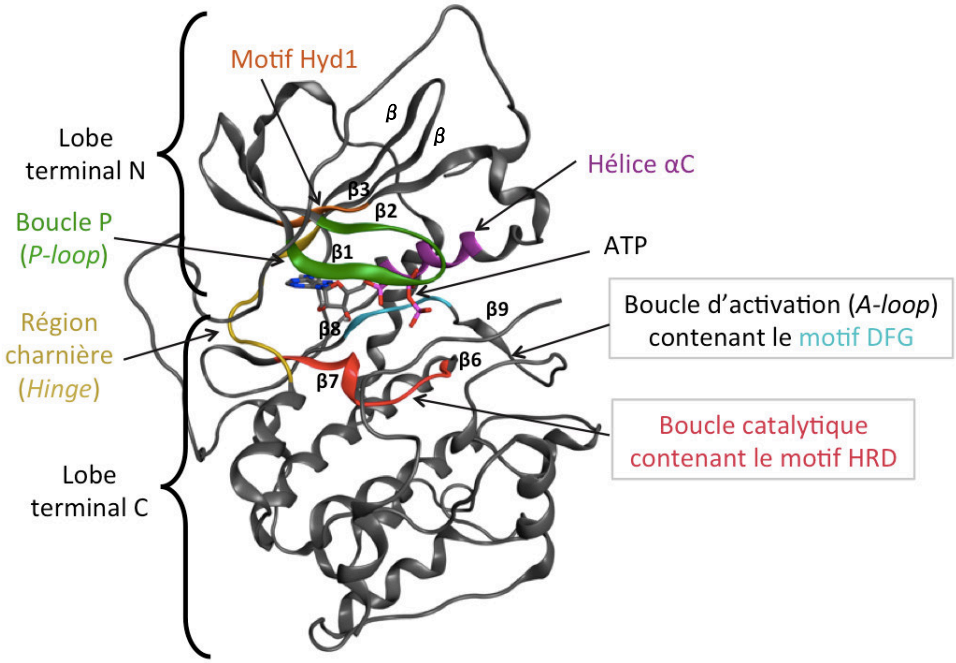


Figure 21 : Caractéristiques de la structure tridimensionnelle du domaine kinase hautement conservées chez les eucaryotes.
La protéine kinase Prkaca est représentée en ruban sous sa forme active complexée à l'ATP représenté en bâton (code PDB : 1ATP).

C'est entre les 2 lobes, au niveau de la région charnière, que se trouve le site catalytique. Dans une conformation active, le lobe terminal C se compose de quatre brins β (β_6-β_9) et d'un nombre variable d'hélices α (entre six et huit) en fonction de la structure considérée. Les brins β_6 et β_7 contiennent la boucle catalytique avec le motif HRD (His-Arg-Asp) à son extrémité N-terminale. L'aspartate du motif HRD joue le rôle d'accepteur pour le transfert du phosphate γ vers le substrat. Puis, entre les brins β_8 et β_9 se trouve le motif DFG (Asp-Phe-Gly) qui marque le début de la boucle d'activation (*A-loop*) et participe au bon positionnement des phosphates β et γ en vue du transfert de ce dernier. La boucle d'activation, elle, permet de positionner le groupement hydroxyle du substrat (Fabbro et al., 2015; Kornev and Taylor, 2010; Taylor and Kornev, 2011; Wang and Cole, 2014).

3. Régulation des PK

Au vu de leur rôle essentiel dans les processus biologiques, comme nous venons de le voir dans le paragraphe précédent, il n'est pas étonnant que les PK soient à leur tour soumises à de multiples mécanismes de régulation. Ces mécanismes divers et variés sont parfois spécifiques à une famille ou à une sous-famille de PK comme nous allons le voir dans la partie I.I.5.a) avec la famille des kinases cycline-dépendante (CDK) qui possède leur propre mode d'activation (Jeffrey et al., 1995). Bien que beaucoup de mécanismes de régulation soient encore inconnus, il est communément admis que :

- dans le milieu cellulaire, les PK transitent entre une conformation catalytiquement active et des conformations inactives, sous l'effet de mécanismes de régulation (Johnson et al., 1996).
- la majorité des PK nécessitent d'être phosphorylée, soit par elle-même soit par une autre kinase, sur un résidu sérine, thréonine ou tyrosine de la boucle d'activation afin d'être actives (Oppermann et al., 2009).
- deux changements conformationnels inactivent la PK : le passage d'une conformation *αC-helix in* à *αC-helix out* et celui d'une conformation *DFG-in* à *DFG-out* (**Figure 22**) (Taylor et al., 2015).

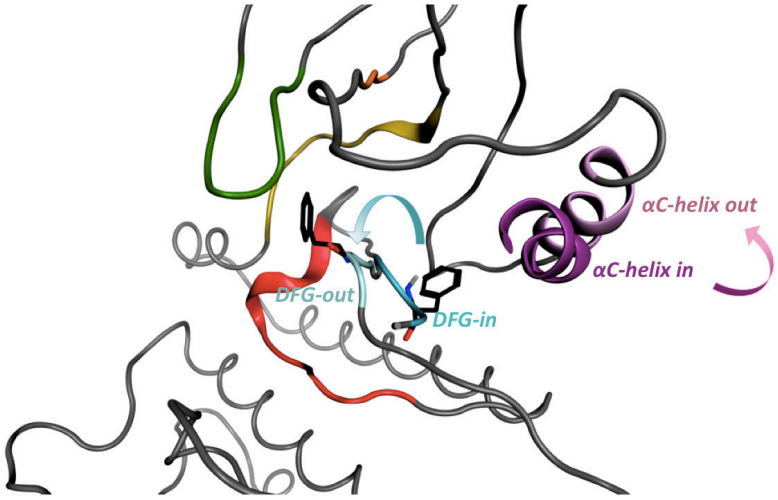


Figure 22 : Changements conformationnels inactivant la protéine kinase.
Représentation en ruban de la protéine kinase EGFR (*DFG-out/ αC-helix out*) (code PDB : 5HG8) et du motif DFG (*DFG-in*) et l'hélice αC (*αC-helix in*) de la protéine kinase SRC (code PDB : 1AD5) centrées sur le site catalytique. La phénylalanine du motif DFG est représentée en bâton avec les atomes de carbone colorés en noir.

4. Les inhibiteurs de PK

Les inhibiteurs de PK ont été regroupés en classes suivant les zones qu'ils occupent dans le site de liaison (Roskoski, 2016). Les inhibiteurs de type I, II et III se lient au niveau de la poche du site catalytique entre les deux lobes terminaux N et C (**Figure 23**). Les inhibiteurs de type IV ne se fixent pas à la poche du site catalytique mais plutôt à un site qui se trouve soit sur le lobe C, soit sur le lobe N ; l'occupation de ce site par l'inhibiteur va induire un changement de conformation, par effet allostérique, qui va entrainer l'inactivation de la PK (Lamba and Ghosh, 2012). Les inhibiteurs de type I se lient à la poche de liaison de l'ATP dans la conformation active de la protéine, c'est à dire la conformation *DFG-in/αC-helix in*. Les inhibiteurs de type $I^{1/2}$ se lient à une conformation *DFG-in/αC-helix out*, dans laquelle le changement de conformation de l'hélice αC lui permet de réaliser des interactions supplémentaires avec une poche hydrophobe. Les inhibiteurs de type II occupent, en plus de la poche de liaison de l'ATP, une poche adjacente, appelée poche allostérique, alors accessible par le changement de conformation de *DFG-in* à *DFG-out*. Les inhibiteurs de type III se lient uniquement à la poche allostérique. Les inhibiteurs de type III et IV sont donc de nature allostérique. Ils existent d'autres classes (type V, type VI) et sous classes (type $I^{1/2}A$, type $I^{1/2}B$) que nous ne détaillerons pas (Roskoski, 2016).

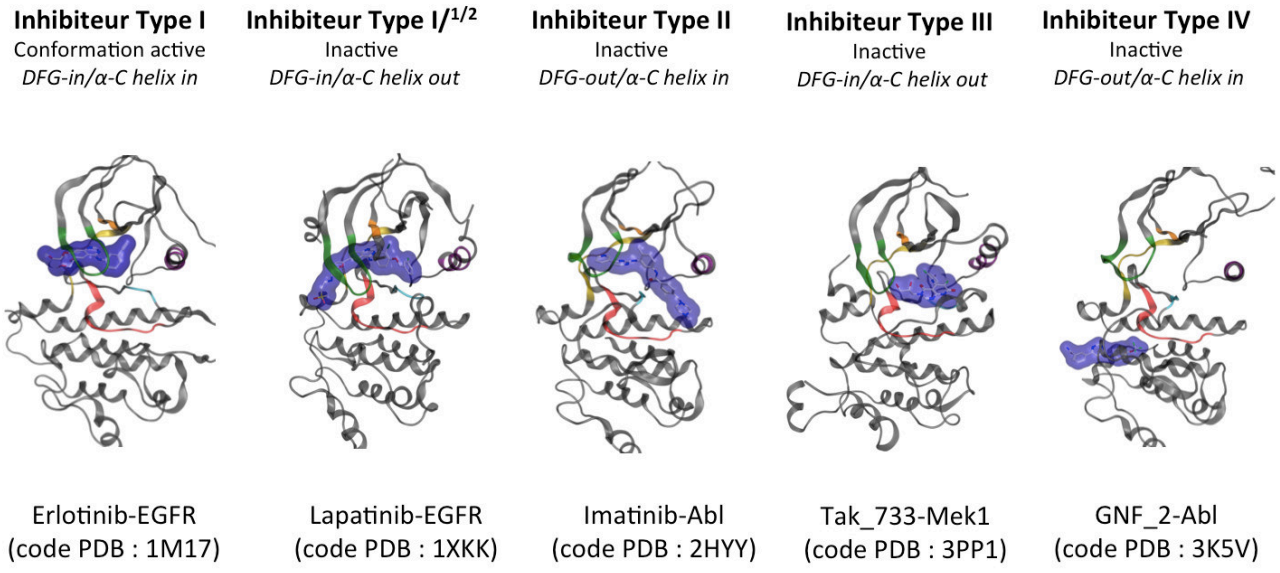


Figure 23 : Mode de liaison des inhibiteurs de type I, $I^{1/2}$, II, III et IV de protéine kinase. La protéine est représentée en ruban et les inhibiteurs en bâton (avec les atomes de carbone colorés en blanc) et avec leur surface.

5. CDK

Les kinases cycline-dépendante (CDK) sont des sérine-thréonine kinases qui, comme leur nom l'indique, nécessitent la liaison avec des protéines régulatrices appelées cycline pour être actives. Les CDK sont les principaux régulateurs du cycle cellulaire et de la transcription des gènes. Le protéome humain contient 20 CDK et 29 cyclines. CDK1 à CDK6 sont impliqués dans la régulation du cycle cellulaire, tandis que CDK7, CDK8, CDK9, CDK11 et CDK20 sont principalement impliqués dans la régulation de la transcription. Plus particulièrement, CDK7, CDK8 et CDK9 contrôlent l'activité de l'ARN polymérase II chez l'homme par la phosphorylation de son domaine C-terminal, qui catalyse la synthèse de tous les précurseurs d'ARNm (Malumbres, 2014). L'inhibition de l'activité des CDK par de petites molécules pour le traitement du cancer a été largement étudiée. Plusieurs inhibiteurs de CDK ont été soumis à des essais cliniques et le premier à être approuvé par la FDA est le palbociclib en février 2015. Le palbociclib est un inhibiteur double des CDK4/6 utilisé pour le traitement du cancer du sein métastatique (Canavese et al., 2012; Sánchez-Martínez et al., 2015).

a) Mécanisme d'activation des CDK

Les CDK sont généralement activées en deux étapes : 1) la liaison de la cycline et 2) la phosphorylation d'un résidu thréonine dans la boucle d'activation des CDK (T160 dans la CDK2 humaine). La liaison de la cycline à la PK induit un changement de conformation de l'hélice αC (*αC-helix out),* qui adopte une conformation *αC-helix in*. La thréonine phosphorylée sur la boucle d'activation sert de point d'ancrage pour ajuster l'orientation de trois résidus arginines conservés, induisant un changement conformationnel de la boucle d'activation qui passe d'une conformation *DFG-out* à *DFG-in* (**Figure 24**) (Pavletich, 1999).

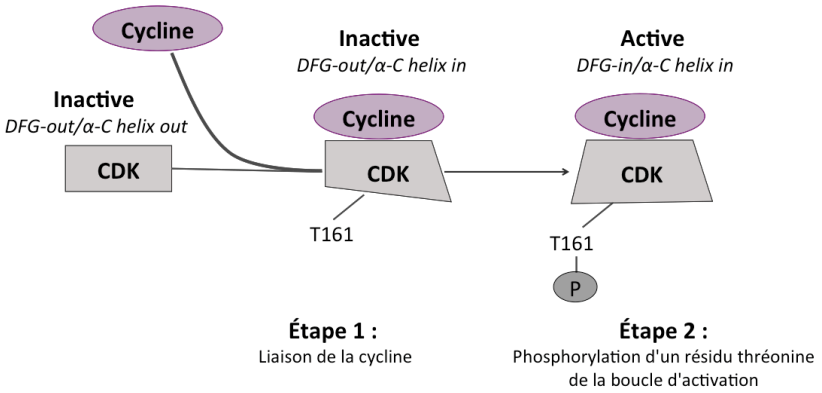


Figure 24 : Mécanisme général d'activation des kinases cycline-dépendante (CDK).

b) Choix d'une cible pour la preuve de concept : CDK8-CycC

Étant donné l'importance des PK en tant que cibles thérapeutiques, et au vu de notre stratégie de recherche au sein de l'équipe SB&C à l'ICOA qui porte principalement sur le développement d'inhibiteurs de PK, nous souhaitions privilégier une cible appartenant à cette famille protéique. Notre choix s'est très vite orienté vers la CDK8-CycC pour laquelle une série congénère d'inhibiteurs avec leurs constantes cinétiques mesurées dans des conditions expérimentales homogènes ont été publiées (Schneider et al., 2013). Pour certains de ces inhibiteurs, nous disposons aussi de la structure expérimentale de CDK8-CycC co-cristallisée avec l'inhibiteur.

(1) CDK8 : une cible d'intérêt majeur

En plus de la disponibilité de données cinétiques, CDK8 est une cible d'intérêt qui a récemment attiré une attention considérable après la publication de nombreuses études génétiques et biochimiques montrant ses nombreux rôles clés dans l'oncogenèse (Philip et al., 2018; Rzymski et al., 2015). Parmi ses diverses fonctions cellulaires, la plus notable est son implication dans la transcription, à travers le complexe médiateur dont elle fait partie (**Figure 25**).

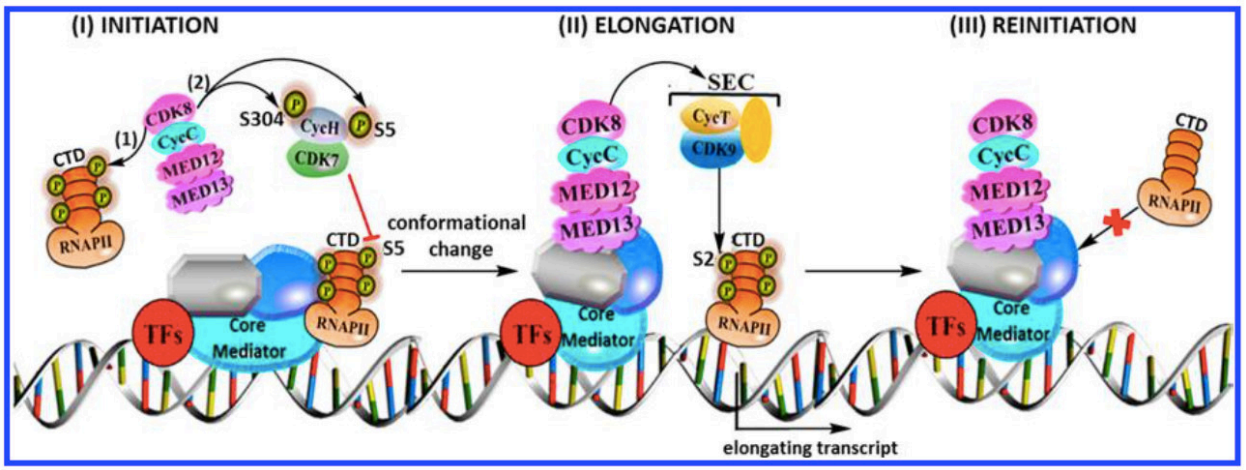


Figure 25 : Rôle de CDK8 dans la régulation de la transcription (Philip et al., 2018).
Rôle régulateur négatif : **(I)** Dans la phase d'initiation, **(1)** le module CDK8 de levure régule négativement la transcription en phosphorylant la CTD RNAPII (domaine carboxy-terminal de l'ARN polymérase II) *in vitro*. **(2)** le module CDK8 humain phosphoryle CycH (cycline H) *in vitro*, empêchant ainsi l'initiation de la transcription via la désactivation de CDK7.
Rôle régulateur positif : **(II)** le module CDK8 peut se lier au noyau du médiateur (*Core Mediator*) après dissociation de la RNAPII, ce qui facilite la libération et l'élongation de certains gènes via le recrutement de la SEC. Cette liaison induit également un changement de la conformation du noyau du médiateur (*Core Mediator*), le rendant incapable de se lier à un autre RNAPII. **(III)** La RNAPII ne pouvant pas se lier au complexe médiateur, ceci met fin à la réinitiation de la transcription.

De fait, CDK8-CycC s'associe à MED12 et MED13 pour former le module CDK8, sous-module du complexe médiateur (Harper and Taatjes, 2017; Wang et al., 2013). Chez l'homme, le module CDK8 inhibe l'initiation de la transcription en phosphorylant la cycline H, entrainant l'inactivation de CDK7. Par conséquent, CDK7 ne peut plus phosphoryler le domaine carboxy-terminal de l'ARN polymérase II, bloquant ainsi l'initiation de la transcription (Akoulitchev et al., 2000). Cependant, ce mode de répression transcriptionnelle n'a pas été observé *in vivo*, et au lieu de cela, un rôle régulateur positif pour CDK8 a été proposé via le recrutement de SEC, ce qui en fait un oncogène (Furumoto et al., 2007; Liu et al., 2004). De fait, l'interaction du complexe médiateur avec SEC facilite l'élongation et la libération de certains gènes (**Figure 25**).
Particulièrement, l'activation médiée par CDK8 de la voie de signalisation de Wnt-β-caténine (Firestein et al., 2008) et de la transcription de gènes inductibles par les œstrogènes (McDermott et al., 2017) contribuent respectivement à l'oncogenèse dans les tumeurs colorectales et mammaires. Les diverses fonctions biologiques de CDK8 et ses rôles apparemment spécifiques dans différents types de cancer ont suscité un grand intérêt et peut-

être une controverse encore plus grande dans le développement d'inhibiteurs de CDK8 en tant qu'agents thérapeutiques potentiels contre le cancer (Philip et al., 2018).

(2) CDK8 : un mécanisme d'activation particulier

Lors de l'initiation d'un projet de *drug design* où la structure de la cible est connue (*structure-based drug design*), la première étape est de se familiariser avec sa structure 3D et ses annotations et la façon dont elles sont impliquées dans sa fonction biologique. En analysant le jeu de données sélectionné, on constate qu'il se compose de 10 inhibiteurs de type II qui se lient non pas à CDK8 seule en conformation *DFG-out/αC-helix out* mais au complexe CDK8-CycC en conformation *DFG-out/αC-helix in* (Schneider et al., 2013). Cette conformation du complexe CDK8-CycC est surprenante car elle correspond à « l'état intermédiaire du mécanisme d'activation » : la liaison de la cycline C à CDK8 a induit le changement de conformation de l'hélice αC de *out* à *in* mais l'étape de phosphorylation n'a pas eu lieu ce qui maintient une conformation *out* pour le motif DFG (**Figure 24**). Cette conformation est très particulière et à ce jour, CDK8 est le seul membre des CDK pour lequel une telle structure est obtenue expérimentalement : co-cristallisée avec un inhibiteur de type II (donc en conformation *DFG-out*) tout en étant associée à la cycline. Une telle structure a été observée la première fois avec l'inhibiteur de type II sorafenib en 2011 (Schneider et al., 2011). Il existe une structure de CDK2 co-cristallisée avec un inhibiteur de type II mais non-complexée à la cycline. Les auteurs constatent que la liaison de l'inhibiteur à CDK2/CycB entraine la dissociation de la cycline B de CDK2 de façon compétitive (Saeed et al., 2012). De plus, toutes les structures de CDK complexées à la cycline sont habituellement en conformation *DFG-in* en accord avec son mécanisme d'activation (**Figure 24**).

Ces observations nous ont poussés à rechercher dans la littérature des particularités du mécanisme d'activation de CDK8, notamment le rôle de la cycline C. En accord avec le mécanisme général d'activation des CDK, l'association de CDK8 à la cycline C (étape 1) entraine bien un changement de conformation de l'hélice αC (*αC-helix out)* qui adopte une conformation *αC-helix in*. Cependant, la seconde étape de phosphorylation n'a pas été prouvée chez CDK8 (Hoeppner et al., 2005; Oppermann et al., 2009). Ainsi un mécanisme autre que la phosphorylation semble induire le changement de conformation de *DFG-out* à *DFG-in* (Hoeppner et al., 2005). L'ensemble de ces observations sont très importantes car elles posent la question du rôle de la cycline C dans le complexe CDK8-CycC : la cycline C influence-t-elle la dynamique de CDK8, plus particulièrement en ce qui concerne les régions proches du site de

liaison et les interactions avec le ligand ? Si l'influence est négligeable, la cycline C peut être ignorée dans les calculs de simulation, ce qui va grandement augmenter la vitesse de calcul. L'objectif du chapitre II est justement de répondre à cette question et également d'essayer de comprendre le mécanisme d'activation de CDK8.

II. ETUDE DU COMPLEXE CDK8-CycC

La protéine kinase 8 cycline C dépendante (CDK8-CycC), est le système sur lequel nous avons travaillé tout au long de ce projet de thèse. Comme présenté dans l'introduction, CDK8-CycC possède des caractéristiques biologiques et structurales particulières, comparé à ses homologues de la famille CDK (cf. I.I.5.b)(2)).

Dans ce chapitre, nous avons étudié la structure et la dynamique de CDK8-CycC et notamment l'interaction entre CDK8 et la CycC à l'aide de simulation de dynamique moléculaire et de calculs d'énergie libre.

Outre la compréhension structurale et dynamique du complexe CDK8-CycC, cette étude permet de statuer quant à la nécessité ou non de garder la CycC dans nos simulations de dynamique moléculaire.

Ce chapitre se présente sous forme d'un article en cours de préparation

A. Article en préparation

Understanding the structural molecular basis of the protein-protein interaction between the cyclin-dependant kinase 8 and the cyclin C

Sonia Ziada[1], Julien Diharce[2], Dylan Serillon[3], Samia Aci-Sèche[1]*, and Pascal Bonnet[1]*

[1]Institut de Chimie Organique et Analytique (ICOA), UMR CNRS-Université d'Orléans 7311, Université d'Orléans BP 6759, 45067 Orléans Cedex 2, France.

[2]Biologie intégrée du Globule Rouge, UMR_S 1134 Inserm - Université Paris 7, Paris Diderot, DSIMB team, Institut National de la Transfusion Sanguine, 6 Rue Alexandre Cabanel, 75015 Paris cedex 15

[3]LSAMM Laboratoire de Synthèse des Assemblages Moléculaires Multifonctionnels, Institut de Chimie de Strasbourg, CNRS/UMR 7177, Université de Strasbourg, 4, rue Blaise Pascal, Strasbourg 67000, France"

*Author to whom all correspondence should be addressed:

Pascal Bonnet: Tel: +33 238 417 254, Fax: +33 238 417 254, E-mail: pascal.bonnet@univ-orleans.fr

Samia Aci-Sèche: Tel: +33 238 419 902, Fax: +33 238 417 254, E-mail: samia.aci@cnrs-orleans.fr

ABSTRACT

A deregulation of the Cyclin-dependent kinase 8 (CDK8) activity has been associated to many diseases including the colorectal and breast cancer. The activity of CDK8 is controlled by a regulatory protein called cyclin C (CycC) that binds to CDK8, which results in its activation. While human CDK family members are generally activated in two steps that is, the binding of the cyclin to CDK and the phosphorylation of a residue in the CDK activation loop, CDK8 does not require the phosphorylation step to be active. Another peculiarity of CDK8 is its ability to be associated to CycC while adopting an inactive form. All those observations questioned the importance of the CycC in the CDK8 dynamical and biological properties Through MD simulations and binding free energy calculations, the role of CycC and its impact on the CDK8 dynamics in several cases (active and inactive form, apo or holo form, mutations) are studied here. Interestingly, our results highlight that CycC stabilizes the CDK8 structure in both active and inactive conformation. Important residues at the interaction surface are identified, and first molecular insights show that the interface presents specific interaction points, in addition of the common interface of CDK-Cyc family. Those results will be very helpful for the

design of peptidic inhibitors targeting specifically the CDK8-CycC interface, in the context of inhibition of CDK8 as an oncogene. Finally, TMD calculation has been used to simulate the transition from the inactive conformation to the active one. Analysis demonstrate that the CycC undergoes a slight rotation during the transition, in order to stabilize the active form of CDK8, via one critical residue, replacing the missing phosphorylate residue present on the other CDKs. Our study shed light on the crucial role of CycC, and the importance to consider it in each computational study concerning CDK8 system.

Keywords: Kinases · Cyclin C · Protein-protein interaction · Drug design · Molecular dynamics simulation · Free energy calculation.

INTRODUCTION

Cyclin-dependent kinases (CDKs) are serine-threonine kinases that require binding with regulatory proteins called cyclin to be active. CDKs are the main regulators of the cell cycle and gene transcription. The human proteome contains 20 CDKs and 29 cyclins. CDK1 to CDK6 are involved in cell cycle regulation, while CDK7, CDK8, CDK9, CDK11, and CDK20 are primarily involved in transcriptional regulation. More particularly, CDK7, CDK8 and CDK9 control the activity of RNA polymerase II in humans by the phosphorylation of its C-terminus domain, which catalyzes the synthesis of all mRNA precursors (Malumbres, 2014). The inhibition of CDKs activity by small molecules for the treatment of cancer has been extensively studied. Several CDKs inhibitors have undergone clinical trials and in February 2015, palbociclib, a CDK4/6 inhibitor, was the first approved by the FDA (Canavese et al., 2012).

CDK8 is a target of interest that has recently attracted considerable attention after the publication of numerous genetic and biochemical studies highlighting its many key roles in oncogenesis (Philip et al., 2018; Rzymski et al., 2015). Among its various cellular functions, the most notable is its involvement in regulating transcription trough diverse mechanisms. CDK8 is a part of the mediator complex, which is a large multisubunit proteins complex that is central to the regulation of transcription in eukaryotes (Allen and Taatjes, 2015). The main function of the mediator complex is to transmit regulatory signals from DNA-bound transcription factors to the RNA polymerase II (RNAPII). The complex CDK8-Cyclin C (CDK8-CycC) associates with MED12 and MED13 to form the CDK8 module, a sub-module of the mediator complex (Harper and Taatjes, 2017; Wang et al., 2013). In humans, it has been demonstrated *in vitro* that the CDK8 module inhibits the initiation of transcription by deactivating CDK7, which can no longer phosphorylate the carboxy-terminal domain of RNAPII, thereby blocking the transcription (Akoulitchev et al., 2000). On the other hand, contrary to this transcriptional repression role, it has been observed *in vivo*, a positive regulatory role for CDK8 via the recruitment of SEC (Super Elongation Complex), making CDK8 an oncogene of interest. In fact, the interaction of the mediator complex with SEC facilitates the elongation and release of certain genes (Furumoto et al., 2007; Liu et al., 2004). In particular, CDK8-mediated activation of the Wnt-β-catenin signaling pathway (Firestein et al., 2008) and of the transcription of estrogen-inducible genes (McDermott et al., 2017) contribute respectively, to oncogenesis in colorectal and mammary tumors.

Since Schneider *et al*. published in 2011 the first crystallographic structure of human CDK8-CycC complexed with sorafenib (PDB id.: 3RGF) (Schneider et al., 2011), a total of 25 experimental structures are currently available. All of these crystal structures present 10 to 20 missing residues within a region that lies outside the active-site cleft called the activation loop. This motif has a central role in regulating the activity of protein kinase by generally adopting a *DFG-in* conformation in the active form and a *DFG-out* conformation in the inactive form (Taylor and Kornev, 2011). In that connection, the first computational study on human CDK8 (with PDB id.: 4RGF) aims at providing insights into two point mutations within the activation loop through 50 ns of all-atom conventional molecular dynamics (cMD) simulation in implicit solvent (Xu et al., 2014). Moreover, the theoretical binding free energy between CDK8 and CycC was also determined using MM-PBSA and MM-GBSA on the basis of 2 ns of all-atom cMD simulation in explicit solvent. However, *in silico* structural studies, a particular attention should be paid to the building of a relevant model of the protein, especially in this study (Xu et al., 2014) where the object of the investigation, the activation loop, is missing and has to be reconstructed. Surprisingly, the authors used a template where the activation loop is in *DFG-in* conformation to model the activation loop of 3RGF (PDB id.), which is in *DFG-out* conformation. Cholko *et al.* studied twelve CDK8-CycC systems using 500 ns all-atom cMD simulations in explicit solvent with the aim of elucidating the system motions and the structural determinants that affect protein-ligand interactions (Cholko et al., 2018). They find that the CycC is important in providing proper interactions for ligand binding whereas the highly flexible activation loop has a little effect. Furthermore, they employed MM-PBSA analysis to characterize protein-ligand interactions from an energetically point of view and discuss the major driving force of protein-ligand binding.

In this study, we investigated the effect of CycC on the structure and dynamics of CDK8 on the one hand, and on the other hand the structural molecular basis of the protein-protein interaction between the two partners. Indeed, the role of the presence of CycC in the complex CDK8-CycC appears to be more complex than other members of CDK family. CDKs are generally activated in two steps: 1) the binding of the cyclin (Cyc) to CDK and 2) the phosphorylation of a threonine residue in the CDK activation loop (T160 in human CDK2). The binding of the Cyc to CDK induces a conformational change of the αC-helix, which adopts a *αC-helix in* conformation (shift toward the binding site) from an *αC-helix out* conformation. The phosphorylated threonine on the activation loop serves as an anchor for adjusting the

orientation of three conserved arginine residues, inducing a conformational change in the activation loop that shifts from a *DFG-out* to a *DFG-in* conformation (Pavletich, 1999). In CDK8 the phosphorylation step has not been observed and is not required to its activation (Hoeppner et al., 2005; Oppermann et al., 2009). Moreover, the first published crystallographic structures of human CDK8-CycC (Schneider et al., 2011, 2013) and also a more recently one (Bergeron et al., 2016) display a surprising conformation corresponding somehow to the "intermediate state of the activation mechanism". Indeed, the αC-helix is in *αC-helix in* conformation, which is expected since CycC is bound to CDK8 in agreement with the activation mechanism; however, the phosphorylation step did not take place, which maintains a *DMG-out* conformation (in CDK8, it is a DMG motif instead of DFG) for the activation loop. All of these structures are co-crystallized with several inhibitors which is told to be responsible of the conformational change from *DMG-in* to *DMG-out* conformation. All these co-crystallized inhibitors belong to the type II or type III class of protein kinase inhibitors, meaning that they bind partially (type II) or only (type III) in the allosteric pocket (also called “hydrophobic pocket”). This particular pocket is only accessible by the rearrangement of the DMG motif from the *in* to the *out* conformation. Up to date, CDK8 is the only CDK member for which such a structure is obtained experimentally: a *DFG-out* conformation (*DMG-out* in CDK8) while being associated with CycC. All CDKs structures complexed with Cyc are usually in *DFG-in* conformation, in accordance with its activation mechanism. Alexander *et al.* try to reproduce this particular conformation with the complex CDK2-CycB. They incubate the CDK2-CycB complex with a type II inhibitor and also observe a *DFG-out* conformation. However, they find that binding of a type II inhibitor to CDK2-CycB results in the dissociation of cyclin B from CDK2 in a competitive manner (Alexander et al., 2015). This state of the art raises the question of the role of CycC in the complex CDK8-CycC in the inactive conformation (*DMG-out*). Particularly, it is interesting to investigate whether the CycC has an impact on the structure and dynamics of CDK8 and whether this impact is the same in active (*DMG-in*) and inactive (*DMG-out*) conformation and in the presence and absence of the ligand. In addition, in view of this unique feature of CDK8 to bind the CycC in both conformations, it is interesting to study the interaction between CDK8 and CycC in order to decipher the interaction molecular basis and to highlight possible important CDK8-specific interaction hotspots.

Through MD simulations and binding free energy calculations, we found that CycC has a stabilizing effect on CDK8, and that it is important to maintain a proper conformation of CDK8 in the active and inactive form of CDK8-CycC. The per residue free energy decomposition method enabled to characterize the CDK8-CycC binding surface, to identify the important residues and to obtain their energy contributions. We found that CDK8-CycC presents specific interaction hotspots within its interaction surface compared to other human CDK/Cyc pairs. Targeting these specific interaction hotspots could be a promising approach in terms of specificity, to effectively disrupt the interaction between CDK8 and CycC and thus, to interfere with the function of CDK8 as an oncogene. The simulation of the conformational transition from the inactive to the active form of CDK8-CycC through TMD simulation, suggests another mechanism that could substitute the missing phosphorylation step in the activation mechanism of CDK8. In a more general view, these results point the importance of keeping the CycC in computational studies when studying the human CDK8 protein in both the active and the inactive form.

MATERIAL AND METHODS

Material description

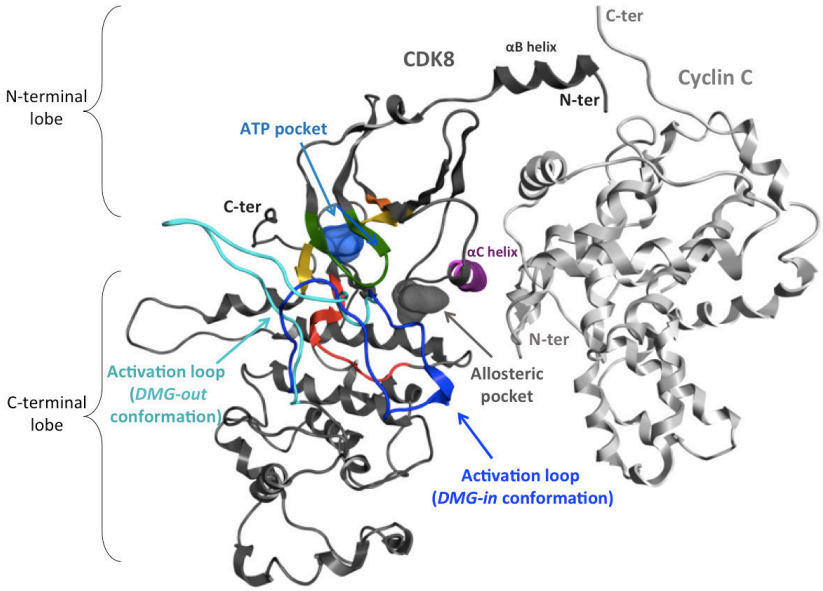


Figure 1: The complex CDK8-CycC.
Representation of the complex in ribbon with CDK8 in dark grey and cyclin C (CycC) in light gray (PDB id.: 4F6U). The regions of the kinase domain containing the conserved motifs are coloured in other colours than grey. Among them, the activation loop in *DMG-out*

conformation (inactive form) was represented in cyan (PDB id.: 4F6U) and that in *DMG-in* conformation (active form) in dark blue (PDB id.: 4F6U).

The catalytic site of CDK8 lies between the N- and C-terminal lobes as in other kinase proteins. Two conformations of CDK8 exist in the PDB that differentiate by the conformation of the activation loop that adopts either a *DMG-in* or a *DMG-out* conformation. CycC interacts mainly with the N-terminal lobe (**Figure 1**). The studied systems are summarized in **Table 1**. The corresponding crystallographic structures are all coming from the paper of Schneider *et al.* (Schneider et al., 2013). The 4F6U structure (PDB id.) presents the best resolution among all *DMG-out* structures resolved up to now. This structure is co-crystallized with a type II inhibitor (system 1a and 1b). To be sure that the results obtained are not ligand-dependent, the apo form of 4F6U (system 3) and another type II inhibitor (PDB id.: 4F7L) (system 6) with a slightly different binding mode (**Figure S1**) were also simulated. Then, to compare our results with experimental mutagenesis results, two residues of the αC-helix were mutated in the structure 4F6U (system 5). Finally, a *DMG-in* conformation of the complex (PDB id.: 4F7S), which is the conformation usually observed in the presence of CycC, has also been simulated in order to compare the behavior of CycC in the complexes *DMG-in* and *DMG-out* conformation (system 8). These systems were also modeled without the CycC in order to investigate the effect of the CycC (except the system 5).

System id.	PDB id.	Ligand name	DMG conformation	Manipulation
1a	4F6U	0SR	DMG-out	(-)
1b (replica)	4F6U	0SR	DMG-out	(-)
2a	4F6U	0SR	DMG-out	Removal of CycC
2b (replica)	4F6U	0SR	DMG-out	Removal of CycC
3	4F6U	(-)	DMG-out	Removal of ligand
4	4F6U	(-)	DMG-out	Removal of ligand and CycC
5	4F6U	0SR	DMG-out	CDK8 mutations: E66A, R65A
6	4F7L	0SO	DMG-out	(-)
7	4F7L	0SO	DMG-out	Removal of CycC
8	4F7S	0SW	DMG-in	(-)

9	4F7S	0SW	DMG-in	Removal of CycC

Table 1: Studied systems and their characteristics.

Model building

The structure of PDB id. 4F6U presents 3 missing loops: the activation loop containing the key DMG-motif (residues 177 to 193) and the loops from residues 116 to 120 and residues 240 to 244. In order to reconstruct these missing residues, we aligned the UniProt canonical sequence of CDK8 on the PDB database to retrieve the most homologous template structures having the missing regions resolved and the activation loop in the *DMG-out* conformation. Two crystallographic structures of the human homologous CDK6 (PDB id.: 1BI8 and 1G3N), were retained and used as template structures. The sequence alignment was performed with Clustal Omega (Sievers et al., 2011) with a particular care on the alignment of domain kinase conserved motifs. CDK6 shares 37 % of identity and 63 % of similarity with CDK8 (**Figure S2 and S3**). Only the missing regions in the target structure were rebuilt in order to keep the coordinates of the resolved parts of the protein unchanged. After adding the sequence of the CycC, the information about the crystallographic water molecules and the ligand (ligand id. 0SR), the alignment file was imported into MODELLER version 9.16 (Sali and Blundell, 1993) to generate the model. We thus obtain a model of CDK8 (residues 1 to 359) complexed to CycC (residues 1 to 264) and to the ligand. The missing C-terminus segments of CDK8 (residues 360 to 464) and of CycC protein (residues 265 to 283) were not reconstructed. The complete model was subjected to structural validation through PROCHECK (Laskowski et al., 1993) and ProSA-web tools (Wiederstein and Sippl, 2007) (**Figure S4**). We did not build another model for the structure of PDB id. 4F7L but rather derive the model by replacing the inhibitor 0SO in that model (chemical replacement). Chemical replacement was considered sufficient because the orientation of the binding site residues is highly conserved in the two structures (PDB id.: 4F6U and 4F7L); and their respective inhibitors (ligand id.: 0SR and 0SO) share a same scaffold bound in the same orientation within the binding site (**Figure S1**). Therefore, the full structure of CDK8-CycC complexed with the ligand 0SO was obtained by, first, aligning the crystallographic structure 4F7L to the model, and then, by placing the ligand and the crystallographic water molecules inside. We manually adjust some residues to be in agreement with protein-ligand interactions observed in the crystallographic structure of PDB id. 4F7L, using Molecular Operating Environment (MOE) version 2016.0802 from the Chemical Computing Group. The same procedure that the one described above was followed

to fill the 3 missing loops of the structure of PDB id. 4F7S (which are the activation loop residues from 187 to 195 and the loops from residues 116 to 121 and from residues 238 to 242). The crystallographic structures of the human homologous CDK1 (PDB id. 1P5E) and CDK2 (PDB id. 1P5E) were retained and used as template structures. CDK1 and CDK2 share respectively 37.8%, 38.3% of identity and 54.5%, 55.9% of similarity with CDK8.

System preparation

In total, 9 systems were prepared (all described in **Table 1**). The AmberTools 14 suite (Case et al., 2015) was employed to protonate, solvate, neutralize and generate topology and coordinate files of the systems. Ligands were prepared by using the Antechamber tool and the GAFF force field after adding hydrogen atoms with the reduce utility (Wang et al., 2004, 2006). The three inhibitors were modeled in their neutral state. Further analysis was carried out for the protonation state of the inhibitor 0SR (ligand id.) (**Figure S5**), since the pKa of alkylmorpholines is about 7.4 (Hall, 1956). The morpholine of the inhibitor 0SR was finally modeled in its unprotonated state since the interaction with Ala100 is not observed any more with the protonated morpholine (**Figure S5**). Partial charges on the ligands were generated with the AM1/BCC method (Jakalian et al., 2002). PROPKA version 3.0 (Olsson et al., 2011) was used to check the protonation state of ionizable residue side-chains at pH = 7. The protein force field ff14SB parameters were assigned (Maier et al., 2015). Then, the system was solvated in a rectangular TIP3P water box, the edge of the box being at least 10 Å away from any solute atom. Finally, Cl^- ions were added to neutralize the positively charged system for a total atoms number of around 110 000 atoms.

Conventional MD simulation (cMD)

A four-cycle minimization was performed with 2000 steps each cycle, minimizing first the solvent, second the residue side-chains, then the solute and finally the entire system. The SHAKE algorithm was applied to constrain bonds involving hydrogen atoms, allowing a time increment of 2 fs. Temperature regulation at 300 K was ensured through Langevin dynamics with a collision frequency of 2 ps^{-1}. The long-range electrostatic interactions were computed by the Particle Mesh Ewald (PME) method beyond 10 Å distance. The system was slowly heated in NVT ensemble from 0 to 300 K over a period of 50 ps, where a harmonic restraint on the solute (20 $kcal.mol^{-1}.Å^{-1}$ force-field constant) prevents the system from structural

distortion. The system was then equilibrated during 10 ns MD simulation in the NPT ensemble at 300 K and 1 atm, through which the harmonic restraint is gradually decreased from 20 kcal.mol^{-1}.Å^{-1} to 3 kcal.mol^{-1}.Å^{-1} in 1.3 ns and then, totally relaxed in 8.7 ns. The pressure relaxation time was set to 1 ps. cMD calculations were performed using the PMEMD.cuda module of the AMBER14 program (Case et al., 2015). We performed 1 µs of cMD production on each system presented in **Table 1** and save the coordinate every 10 ps.

Targeted molecular dynamics (TMD)

The TMD is a simulation technique for determining the pathway of a conformational transition between two states: (un)bound, (un)folded, open-close conformation etc. (Schlitter et al., 1994). It consists in constraining the root mean square deviation (RMSD) between the current structure (which is the starting structure at the beginning of the simulation) and a reference structure ($RMSD_{current}$) to a user-defined value, namely the $RMSD_{target}$. This value of $RMSD_{target}$ is slowly varied from an initial value to a targeted final value ($RMSD_{target\_final}$), which results in the simulation of the process leading to the final desired state. In AMBER14 program, a harmonic restraining potential ($V_{restraint}$) is added to the force field, to help the $RMSD_{current}$ reaching the successive values of $RMSD_{target}$ until the final value ($RMSD_{target\_final}$).

$$V_{restraint} = \frac{1}{2} \times f \times N_{atoms} \times \left(RMSD_{current} - RMSD_{target}\right)^2$$

Equation 1

Where f is the harmonic force constant, N_{atoms} is the number of restrained atoms, that is, the number of atoms on which the RMSD is calculated. Note that the atomic coordinates are mass weighted in the calculation of RMSD. It exists two approaches of TMD: direct TMD and reverse TMD (TMD^{-1}). We apply direct TMD. In direct TMD, the reference structure corresponds to the final targeted structure, so that the value of $RMSD_{target}$ is decreased from the RMSD between the initial and target structure to a value close to 0. In this study, the initial structure is the complex CDK8-CycC in *DMG-out* conformation and the target structure is that in *DMG-in* conformation. The RMSD is calculated on the residues 171 to 182 of the activation loop, after aligning the current and the target structure on the backbone of the less flexible residues of the active site (90 residues in total: residue 26 to 105 and 148 to 158). The spring constant f was set to 2 kcal.mol^{-1}. The $RMSD_{target}$ is changed by step of 0.12 Å every 50 ps from the value of 12.3 Å to 0.01 Å during a total simulation time of 5 ns. TMD runs were performed with the

parallelized version of the SANDER module from the AMBER14 program. The TMD simulation was then continued by 50 ns of cMD simulation following the same parameters as described above.

RMSD, RMSF

The trajectories were aligned on the corresponding crystallographic structures using as mask the heavy atoms of CDK8. The root mean square deviation (RMSD) and the root mean square fluctuation (RMSF) were calculated using the same mask.

PCA

When applying MD simulations on biological system, some questions are often raised: i) do the sampled conformations in a trajectory A "resemble" those of a trajectory B? ii) does the conformational sampling vary over time within a same trajectory? iii) what are the protein regions whose movements contribute the most to explain the conformational diversity? To answer such questions, the principal component analysis method (PCA) is a suitable method. PCA is a linear dimensionality reduction technique that linearly combines a set of variables (here the coordinates of CDK8 backbone residues) into a reduced number of uncorrelated variables called principal components (PCs). The PCs correspond to the directions of largest variance that is the largest-amplitude fluctuations. To obtain the PCs, we first extract the CDK8 backbone of the last 500 ns of a trajectory by selecting one snapshot every 2.5 ns (200 snapshots in total). It is important to align the trajectories to be analyzed on a same referential. Then, a covariance matrix is calculated from the atomic coordinate matrix of the trajectory. The eigenvectors of the covariance matrix are the PCs. The PCs are ordered with PC1 the direction of largest variance, PC2 the direction of second largest variance etc. To visualize the largest amplitude motions, a PDB format trajectory has been produced that interpolates between the most dissimilar structures in the distribution along PC1. PCA analysis were performed with bio3d package (Grant et al., 2006).

MM-GBSA

The molecular mechanics generalized Born surface area continuum solvation (MM-GBSA) method supplied with AMBER was used to calculate protein-protein interaction free energy

(Kollman et al., 2000). 9500 snapshots were extracted from the trajectories in the range of 50ns-1µs (i.e. one snapshot every 100 ps). The binding free energy is calculated as follows:

$$\Delta G_{bind} = \langle G_{complex} \rangle - \langle G_{receptor} \rangle - \langle G_{ligand} \rangle$$

Equation 2

where $\langle G_x \rangle$ corresponds to the average of the total free energy of the component x over snapshots taken from the MD trajectory. The total free energy of each molecule is computed from the following equation:

$$G = E_{MM} + G_{sol} - TS$$

Equation 3

where E_{MM} is the molecular mechanics energy, G_{sol} is the solvation free energy and the term TS is the entropic contribution. The solvation free energy is the sum of polar and non-polar contributions. The non-polar contribution is attributed to cavity formation in the solvent and to van der Waals interactions between the solute and the solvent, which are typically calculated from the solvent-accessible surface area. The polar contribution of G_{sol} is obtained following the generalized Born model (Onufriev et al., 2004) available in AMBER.

Others analysis tools

The VMD program (Humphrey et al., 1996) and the CPPTRAJ module from AMBER14 program (Case et al., 2015) were also used to manipulate and analyze trajectories. The analysis of the protein-protein interactions was performed with the Structure Interaction Diagram module of the maestro suite (Bowers et al., 2006).

RESULTS AND DISCUSSION

Effect of CycC exclusion on structure and dynamics of CDK8

In order to evaluate the effect of CycC on the structure and dynamics of CDK8, the trajectories were analyzed in pairs (with and without Cyc) as shown in the **Table 2**. For the systems in *DMG-out* conformation, the average RMSD is higher in the absence of the CycC, which means that CDK8 structure has deviated from its crystallographic structure in the absence of the CycC. For the system in *DMG-in* conformation, the average RMSD is comparable.

System id. With Cyclin C	Average RMSD (Å) (± Standard deviation)	System id. Without Cyclin C	Average RMSD (Å) (± Standard deviation)
1a	3.6 ± 0.2	2a	5.5 ± 1
1b	3.9 ± 0.3	2b	5.5 ± 0.3
3	3.8 ± 0.3	4	5.4 ± 0.9
6	3.2 ± 0.2	7	5 ± 0.6
8	3.4 ± 0.3	9	3.7 ± 0.4

Table 8: Comparison of the average RMSD calculated on the 1 µs simulation for the systems with and without CycC using as mask the heavy atoms of CDK8.

The RMSF plots (**Figure 2 and Figure S7**) indicates that the absence of CycC increases the motions of one or more of these regions of CDK8: 1) the αC-helix in all cases, which is in direct interaction with the CycC, 2) the αB-helix in all cases except in the system 1b; the αB-helix is also in direct interaction with the CycC and, 3) the activation loop in all cases, except in the system in *DMG-in* conformation (system 9).

In the absence of CycC, the αC-helix has a larger degree of motion and can move toward the region normally occupy by the CycC to adopt an *αC-helix out-like* conformation (**Figure S6**). The αB-helix tends to bend toward CDK8 and to interact with it (**Figure S6**); in system 1b, this leads to its stabilization.

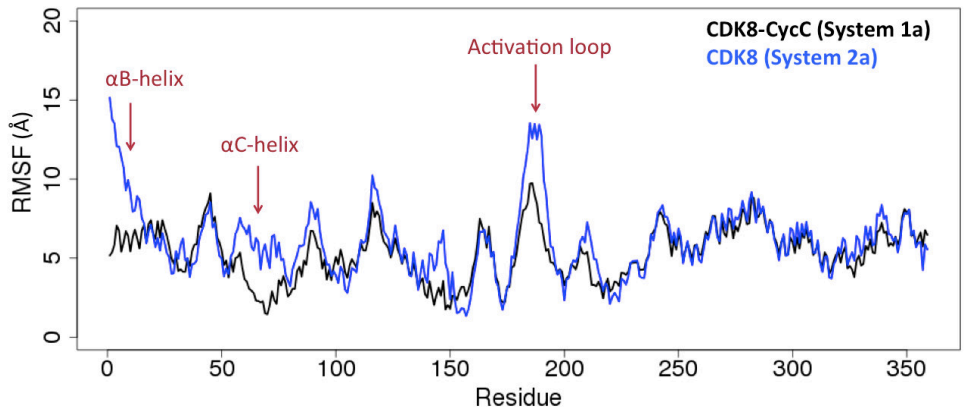


Figure 2: Regions of CDK8 whose atomic fluctuations increase in the absence of CycC. RMSF of system 1a (CDK8-CycC) and system 2a (CDK8) calculated on the 1 µs trajectories.

In order to provide a global view of the effect of the presence of CycC on the structure of CDK8, the trajectories with and without CycC were combined in a single trajectory (two by two: 1a with 2a, 1b with 2b, 3 with 4, 6 with 7, and 8 with 9) by extracting the backbone

coordinates of CDK8 from the two trajectories. Then a PCA was applied (using the conditions described in Materiel and Methods section) on each of the combined trajectories to see if the conformations coming from the simulation with CycC differ from those coming from the simulation without CycC.

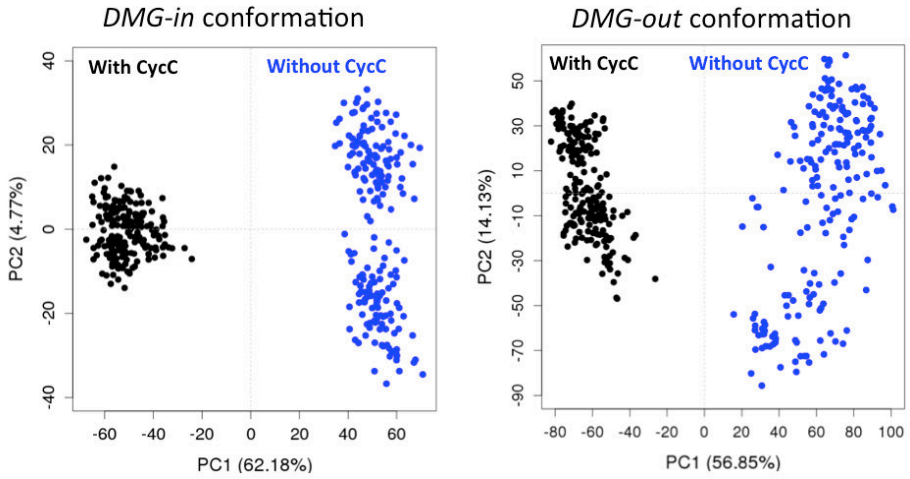


Figure 3: Individual map of the PCA analysis to compare the MD structures of CDK8 coming from a system with and without CycC.
Left: From each of the trajectories of systems 8 and 9 (in *DMG-in* conformation), 200 snapshots were extracted over the last 500 ns. The backbone coordinates of CDK8 were extracted from the total of 400 snapshots and combined in a single trajectory. A PCA analysis was performed on this resulted trajectory: one snapshot of this trajectory is represented by a dot in the individual map of PC1 against PC2. The structures coming from the system with CycC are colored in black and those coming from the system without CycC are colored in blue. **Right**: Same procedure with the trajectories of system 1a and 2a is done, which are in *DMG-out* conformation.

In all cases (5 combined trajectories), two groups are formed along PC1, which correspond respectively to the structures of CDK8 coming from the simulation performed with and without CycC (**Figure 3**). Thus, the PCA analysis is able to separate CDK8 structures coming from the simulation performed with and without CycC along PC1. We also notice that the group of dots corresponding to CDK8 structures coming from the simulation without the CycC is more scattered than the group of dots coming from the one with CycC. It means that the conformational sampling of CDK8 is increased in the absence of CycC. Moreover, in all cases, the two first PCs capture more than 60% of the variance, and PC1 alone traduce more than 50% of the variance. Considering these results together, it appears that PC1 has captured the dynamics of CDK8 regions mostly affected by the presence (or absence) of CycC. It is therefore interesting to analyze the contribution of each residue of CDK8 to PC1, commonly called the loading plot, to identify these regions (**Figure 4**).

First of all, we remark that the PCA loading curves of *DMG-out* conformation systems display a good alignment. Second, a common point emerges from all PCA loading curves: the αB-helix and the activation loop contribute greatly in both cases (*DMG-in* and *DMG-out* conformation systems) to separate the structures coming from the simulations performed with and without CycC. It means that the activation loop and the αB-helix adopt different conformations depending on whether CycC has been kept or not. Note that this observation does not mean that the activation loop and the αB-helix adopt each of them a unique conformation in the absence of CycC. As we saw it previously through the analysis of the RMSF, the activation loop and the αB-helix are more flexible in the absence of CycC. The PCA analysis shows that, in average, the conformations adopted by the activation loop and the αB-helix in the absence of CycC are significantly different from that one adopted in the presence of CycC. Then, other regions contribute at varying level in *DMG-in* and *DMG-out* conformation systems to separate the two groups (with & without CycC) such as: the αF-αG loop that contributes greatly in the *DMG-in* conformation system and not in the *DMG-out* conformation systems, and the contrary for αD-αE loop (**Figure 4**). It is also interesting to note that the αC-helix, which is more flexible in the absence of CycC (**Figure 2** and **Figure S7**), does not show a significant difference in conformations when the CycC is missing for the *DMG-out* conformation systems (**Figure 4**).

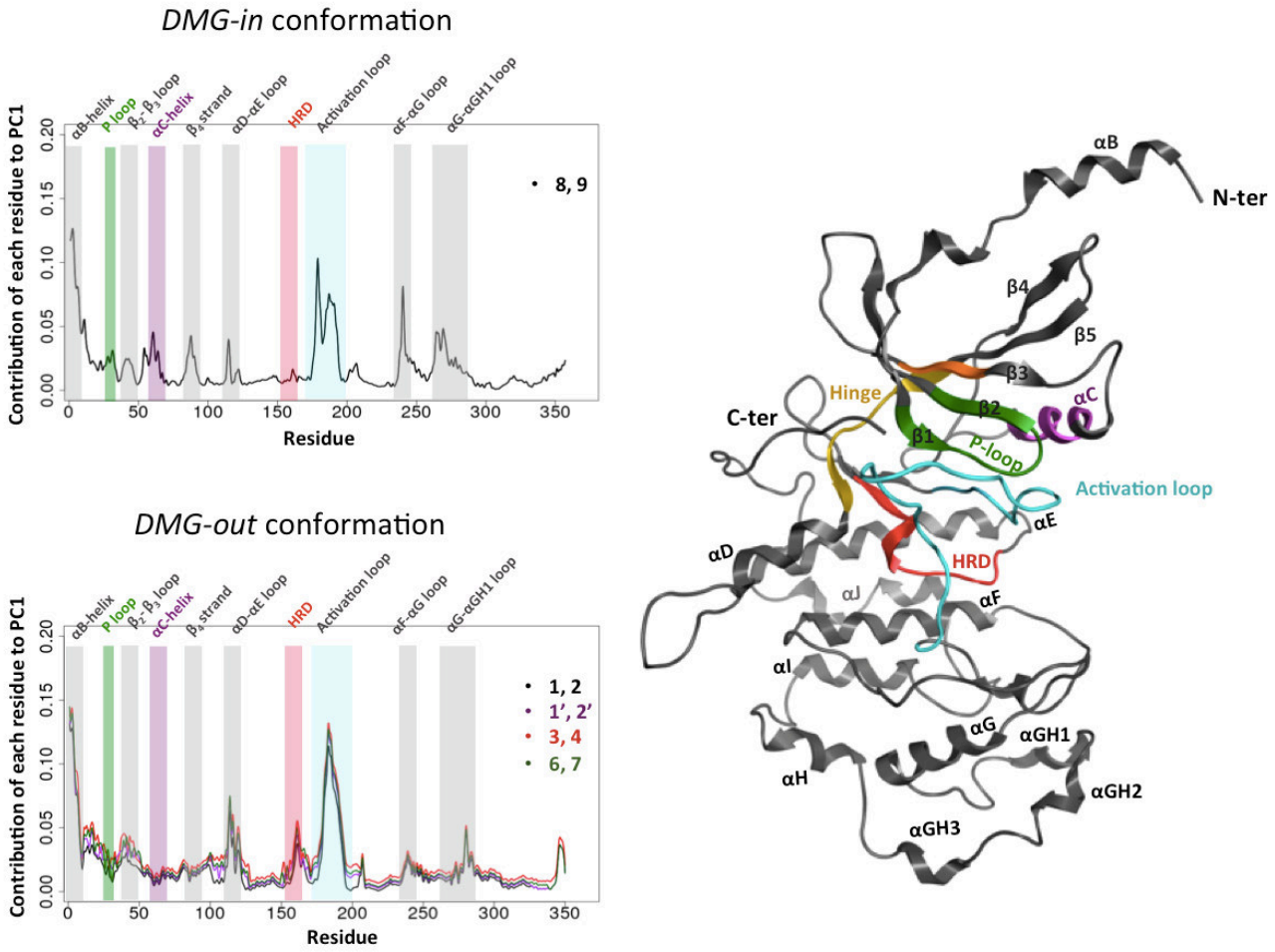


Figure 4: Impact of the CycC on the structure of CDK8: contribution of each residue of CDK8 to PC1.
Loading plot of the PCA done on the combined trajectory (with and without CycC) in *DMG-in* conformation system (**top left**) and *DMG-out* conformation systems (**bottom left**). On **right:** representation of CDK8 in dark gray ribbon except the regions of the kinase domain containing the conserved motifs.

In conclusion of this part, RMSF plots show that CycC stabilizes the αC-helix in both *DMG-in* and *DMG-out* conformation systems and the activation loop of CDK8 in *DMG-out* conformation system. It also reduces the fluctuations of the αB-helix, but in some cases no difference was observed between systems with and without CycC because the αB-helix bend toward CDK8 and stabilizes itself (**Figure S6**). The PCA analysis was able to separate CDK8 structures coming from the simulation performed with and without CycC, which highlights an effect of the CycC on the conformation of CDK8. Particularly, the CycC greatly affects the conformation adopted by the αB-helix and the activation loop. CycC also impacts the dynamics of CDK8, i.e. its breathing motions, since the greatest amplitude motions within CDK8 are not the same depending on whether CycC is present or not (**Figure S8 and S9**). In the literature, Cholko *et al.* also pointed out the importance of the CycC for maintaining a proper structure and

dynamics of CDK8. Through MD simulation on 12 CDK8-CycC systems (6 of *DMG-in* conformation and 6 of *DMG-out* conformation), they observed that CycC stabilizes CDK8 by reducing the fluctuations of the αB-helix, αC-helix and the activation loop. They also mentioned that αC-helix adopt a *αC-helix out* conformation in the absence of the CycC and pointed the importance of the CycC for maintaining proper protein-ligand interaction. Concerning this last point, we also find that the CycC stabilizes the ligand in the binding site (**Figure S10**). In a more general view, these results highlight the importance of keeping the CycC in computational studies and are in agreement with the observations of Cholko *et al.*

Understanding the molecular basis of the interaction between CDK8 and CycC

CDK8-CycC binding free energy

To compute the binding free energy of CycC to CDK8 and gain insights into the binding interaction surface, the MM-GBSA approach was applied on the 9500 snapshots extracted from the trajectories in the range of 50ns-1μs (i.e. one snapshot every 100 ps). We want to know whether CycC has a stabilizing effect in term of binding free energy in: 1) the active form of CDK8-CycC complex (with CDK8 in *DMG-in* conformation), 2) the inactive form of the complex (with CDK8 in *DMG-out* conformation) and, 3) the mutated form of the complex CDK8$^{R65A-E66A}$-CycC. In the presence of Cyc, the active form of CDK-Cyc complex is the form commonly observed in the crystallographic structures of human CDK family members, in agreement with the general activation mechanism of CDKs. In contrast, the inactive form of the CDK8-CycC complex is the first experimental structure exhibiting such conformation. The mutant CDK8$^{R65A-E66A}$-CycC was designed based on experimental mutagenesis data published on CDK8-CycC complex and CDK4-CycD1 complex. A R55A-E56A double point mutation in the αC-helix of CDK4 decreased its binding activity towards cyclin D1 by 85 % (Coleman et al., 1997). On the basis of these results, Barette *et al.* introduce the R65A-E66A double point mutation in CDK8 (corresponding to the R55A-E56A in CDK4) and find that similarly to CDK4, this double point mutation greatly affects the capacity of CDK8 to bind to CycC. However, for the formed complex, they find that CDK8$^{R65A-E66A}$ is still able to stabilize the complex CDK8$^{R65A-E66A}$-CycC (Barette et al., 2001). We therefore calculated the binding energies for the different CDK8-CycC complexes (system 1a, 1b, 3, 5, 6, 8) and summarized the results in **Table 3**.

Systems PDB id. *Conformation*	**System 1a** (4F6U) *DMG-out*	**System 1b** (4F6U-replica) *DMG-out*	**System 3** (4F6U-apo) *DMG-out*	**System 5** (4F6U^{E66A_R65A}) *DMG-out*	**System 6** (4F7L) *DMG-out*	**System 8** (4F7S) *DMG-in*
ΔE_{VDW}	-163.0 ± 0.1	-160.7 ± 0.1	-160.0 ± 0.2	-148.7 ± 0.1	-156.3 ± 0.1	-176.5 ± 0.1
ΔE_{eel}	-508.4 ± 1.0	-436.5 ± 1.0	-490.0 ± 1.5	-573.9 ± 1	-500.6 ± 0.9	-588.5 ± 0.9
ΔE_{GB}	554.3 ± 0.9	491.1 ± 0.9	543.2 ± 1.4	613.9 ± 0.9	541.0 ± 0.8	613.9 ± 0.8
ΔE_{np}	-23.9 ± 0.0	-23.1 ± 0.0	-23.2 ± 0.3	-22.7 ± 0.0	-22.9 ± 0.0	-25.4 ± 0.0
ΔG_{total} (Without entropy)	-141.0± 0.2	-129.2 ± 0.2	-130.0 ± 0.3	-131.5 ± 0.2	-138.8 ± 0.2	-124.6 ± 0.2

Table 3: The binding free energy of the CDK8-CycC complexes and their energy components calculated using the MM-GBSA method.
Average energies in kcal.mol^{-1}, with their corresponding standard errors. ΔE_{eel} and ΔE_{VDW} are respectively electrostatic and van der Waals contributions in gas phase. ΔE_{GB} and ΔE_{np} are respectively electrostatic and non-polar contributions in solvation phase. ΔG_{total} is the total binding free energy without the entropic term.

While the molecular mechanics energy term can be easily obtained from the results of a molecular dynamics simulation, the entropic term is often difficult to achieve. It can be approximated by a quasi-harmonic approximation or calculated through a normal mode analysis. However, the calculation is time-consuming and can be affected by large errors. In addition, the relative contribution of the entropic term to the ΔΔG is often considered to be negligible when comparing two similar systems for example in mutational studies or when comparing similar ligands that bind to the same binding site (as it is the case here), since both contributions are supposed to cancel each other (Massova and Kollman, 1999). Therefore, in this study, ΔG corresponds to the binding free energy without considering the entropic term. In agreement with our structural and dynamical observations, the binding free energy values range from -141.0 ± 0.2 to -124.6 ± 0.2 kcal.mol^{-1}, which confirms the stabilizing effect of CycC. In particular, the result for the mutated system CDK8$^{R65A\text{-}E66A}$ is consistent with the experimental observations, which report that the double point mutation does not affect the stabilization of the complex. The non-polar part of the free energy, composed of the Van der Waals term in gas-phase (ΔE_{VDW}) and the non-polar part of the solvation energy term (ΔE_{np}), is the major favorable component of the CycC binding. Its value is comprised between -171.5 kcal.mol^{-1} and -200.7 kcal.mol^{-1} depending on the system. The highly favorable non-polar part of the free energy might come from the desolvation of the non-polar groups at the binding interface between CDK8 and CycC, as well as the hydrophobic interactions formed between

the two partners. Such phenomenon has been seen in several protein-protein interactions where the main interactions that are responsible for the binding of proteins are hydrophobic in nature (Lo Conte et al., 1999; Tsai et al., 1997). On the other hand, the very favorable electrostatic term in gas phase (ΔE_{eel}) is completely compensated by the unfavorable contribution of the polar part of the solvation free energy (ΔE_{GB}), resulting in a unfavorable total electrostatics interaction comprised between 40.0 kcal.mol^{-1} and 69.5 kcal.mol^{-1}, depending on the system. This compensation phenomenon due to the desolvation penalty of polar groups upon complex formation was discussed in several studies of protein-protein interactions (Kundrotas and Alexov, 2006).

CDK8-CycC binding free energy: decomposition per residue

The method of per-residue binding free energy decomposition can reveal the energy contribution of key residues involved in the protein-protein interaction interface. The total of 9500 snapshots extracted from the trajectories in the range of 50 ns - 1 µs (i.e. one snapshot every 100 ps) were decomposed by the MM-GBSA method. We first identify the common list of residues that significantly contribute to the CDK8-CycC binding in all the studied complexes (system 1a, 1b, 3, 5, 6, 8).

Common hotspots to all studied CDK8-CycC complexes

For each of these systems, the important CDK8-CycC binding residues were extracted using the following condition as cut-off: the absolute value of ΔG_{total} of the residue has to be superior to 1 kcal.mol^{-1}. In the supporting information, the list of the extracted important residues of each system is represented as barplot (**Figure S11**). To extract the common list of important residues shared by all the studied complexes, we took the intersection of these different lists. The heat map presented in **Figure 5** contains the common list of important residues (26 in total) and their binding free energy contributions.

The first obvious result is that no great difference in free energy values is seen between the different studied systems. Second, all the residues present a favorable contribution to CDK8-CycC binding. Moreover, the 26 residues are uniformly distributed on the interaction surface. These first observations suggest that the studied complexes share a large and similar surface of interaction.

Common hotspots to CDK family

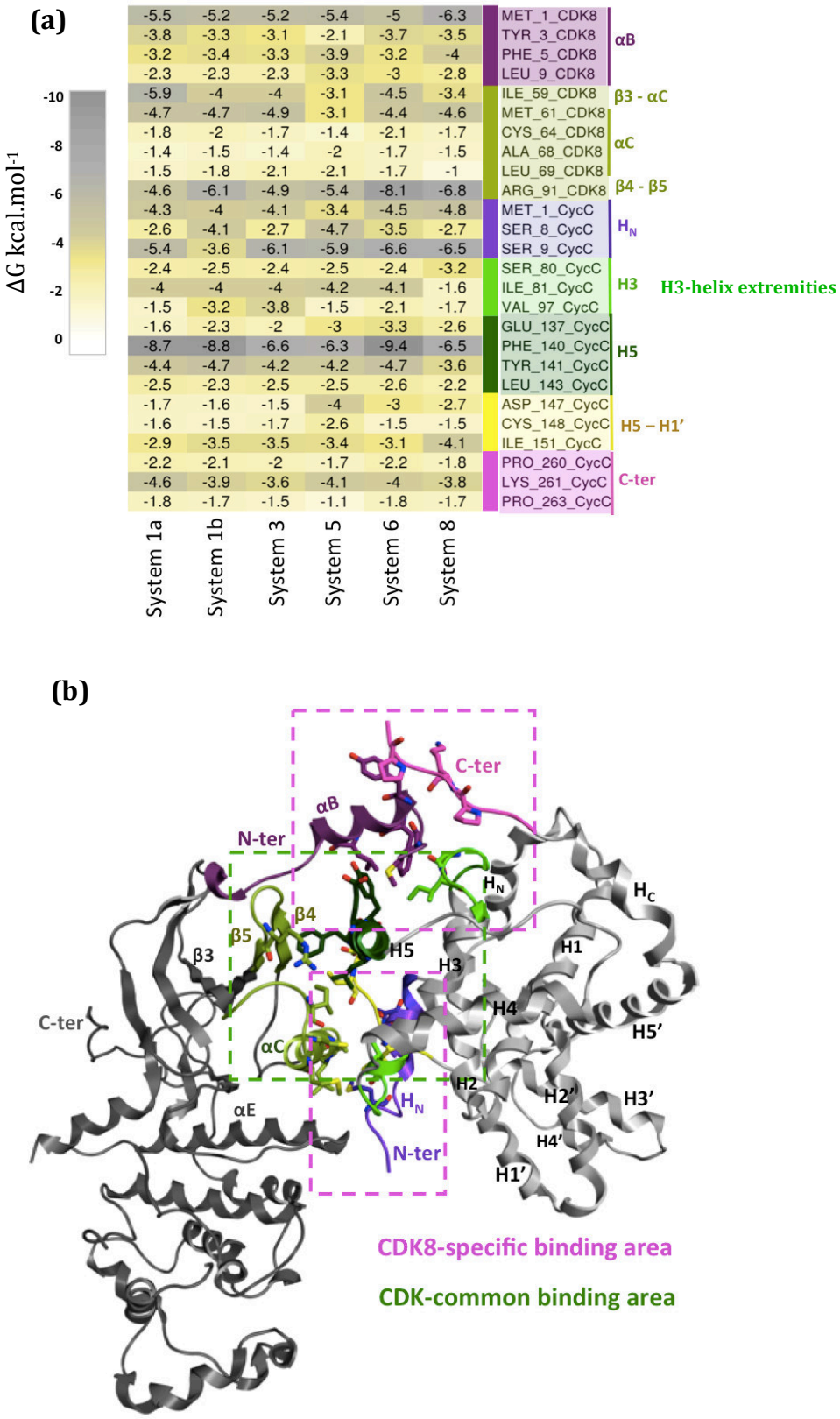


Figure 5: CDK8-CycC binding interface.

(a) Matrix of the per-residue energy contribution (ΔG without entropy). The residues are those that highly contribute to CDK8-CycC binding (absolute (ΔG) > 1 kcal.mol^{-1}) in all studied CDK8-CycC complexes. Residues are tagged according to the secondary structure they belong to. The residues coloured in pink-purple tones are those belonging to CDK8 specific binding

sites. While those coloured in green-yellow tones are those belonging to human CDK common binding sites.
(b) CDK8/CycC structure and the binding site at the CDK8-CycC interface. CDK8 is in ribbon in dark gray except the secondary structures having residues that highly contribute to CDK8-CycC binding in all studied CDK8-CycC complexes. Idem for CycC, which is coloured in light gray. Besides the common binding area (represented by a dashed green box), CDK8/CycC complex forms additional contacts mediated by the CDK8 N-terminus αB-helix and the CycC N-terminus, including the H_N helix.

The members of human CDK family share a conserved common interaction surface with their Cyc partner. This common interaction surface includes the β3–αC region, the αC-helix, and the post-αC region (β4–β5) of the CDK protein in contact with the H5-helix, H5-H1' loop and the residues on both sides of the H3-helix of the Cyc (Echalier et al., 2010; Lolli, 2010). 73.1 % of the identified common important residues of the CDK8-CycC interaction belong to this conserved core as we can see on the heat map (**Figure 5**).

This conserved core is located at the center of the interaction surface and is mainly composed of hydrophobic residues. Among them, the central $Phe140^{CycC}$ situated on the CycC H5-helix seems to have a crucial role in CDK8-CycC binding. Indeed, in a parallel stacking, $Phe140^{CycC}$ establishes a cation-π interaction with $Arg91^{CDK8}$ of the β4–β5 loop, both characterized by a high ΔG absolute value (**Figure 5**). This interaction has an average occupancy of about 78.8% ± 8.3 along all the simulations. A planar cation-π stacking between an arginine and an aromatic side chain has already been described in location critical for the function of a protein, particularly to allow the arginine to form other hydrogen bonds (Flocco and Mowbray, 1994). This is precisely the case here, since $Arg91^{CDK8}$ also establishes a hydrogen bond with $Glu137^{CycC}$ in the H5-helix with occupancy of 75.6 % ± 12.2. Another residue of the CycC H5-helix, the $Leu143^{CycC}$ does hydrophobic contact with $Cys64^{CDK8}$ of the αC-helix with occupancy of 55.7% ± 13.0. Concerning the H5-H1' loop, a hydrogen bond is formed between $Cys148^{CycC}$ and $Arg71^{CDK8}$ with occupancy of 87% ± 7.8 and a water bridge between $Ile151^{CycC}$ and $Glu72^{CDK8}$ of the αC-helix with occupancy of 74 % ± 13.3%. Finally, in the C-terminus of the H3-helix, $Lys96^{CycC}$ interacts with $Ile59^{CDK8}$ localized in β3–αC loop through a hydrogen bond with occupancy of 92.4% ± 7.6%.

In summary, the studied complexes (system 1a, 1b, 3, 5, 6, 8) display a large common binding surface composed of 26 residues distributed uniformly along the interaction surface. This common binding surface is also very similar since the free energy values present few variations from a system to another. All of the 26 residues contributed favorably to CDK8-

CycC binding with free energy values ranging from -9.4 kcal.mol-1 to -1.0 kcal.mol-1. 73.1 % of those residues (19/26 residues) belong to the conserved common interaction interface in the human CDK/Cyc family. Interestingly, we found that the remaining 9 residues belong to regions that are specific to CDK8.

Hotspots specific to CDK8-CycC

Hotspot involving the N-terminus segment of CycC

Although the cyclins are less similar in sequence among themselves compared with the CDKs, they share a common fold constituted of two cyclin boxes comprising five helices each (H1-H5 and H1'-H5'), which are generally associated with two additional helices at the N-terminus and at the C-terminus segments noted H_N and H_C, respectively (**Figure 5**). Unlike cell cycle cyclins (cyclin A/B/D/E), in transcriptional cyclins (cyclin C/T/K/H) (Lolli, 2010), the H_N is located on the side opposite to the CDK binding surface and is not involved in kinase recognition. However, the N-terminus of CycT is still able to maintain some contacts with CDK9. CDK8-CycC appears as an exception since the CycC N-terminus segment is part of the interaction surface positioned below the αC-helix and between CDK8 αE-helix and CycC H5-H1' loop (**Figure 5**). A strong hydrogen bond interaction is observed between the $Glu72^{CDK8}$ and the $Ser9^{CycC}$ with occupancy of 87 ± 9.3% along all the simulations.

Hotspots involving the CDK8-specific N-terminus helix (αB-helix)

CDK8 exhibits an additional N-terminus αB-helix (residues 1-12) preceding the αC-helix, which is unique within human CDK family members (Schneider et al., 2011). Other CDKs display a shorter N-terminus segment by 5-10 residues, except CDK9 where the segment is of equal length, but unstructured (random coil). Among the identified common important binding residues (**Figure 5**), many of them interact with the αB-helix. Particularly, we observed interactions between the proline rich C-terminus segment of CycC and the αB-helix. The $Pro260^{CycC}$ and the $Ser80^{CycC}$ establish both a hydrogen bond with $Asp2^{CDK8}$ with an average occupancy of 82.1% ± 10.3 and 79.1% ± 10.4, respectively. The $Lys261^{CycC}$ interacts with $Tyr3^{CDK8}$ and $Asp4^{CDK8}$ with an average occupancy of 83.4% ± 9.9 and 73.3% ± 11.1, respectively. The CDK8 αB-helix also forms a hydrophobic interaction particularly the $Leu9^{CDK8}$ with $Phe140^{CycC}$ with occupancy of 88.2% ± 7.5.

Taking these results together, it appears that strong and favorable interactions are formed between the proline rich C-terminus segment, which shows a dramatic divergence in length

and in orientation among Cyc partners, and the CDK8 specific αB-helix. Together with the contacts involving the N-terminus segment of CycC, these strong interactions are specific to CDK8-CycC complex and could be one of the mechanisms explaining the selectivity of CDK8 against CycC. Indeed, unlike CDK2 which can bind different Cyc partners (Cyc A/B/E) (Wood and Endicott, 2018), CDK8 is specific to CycC. Moreover, experimental mutational studies converge with our observations since the mutant CDK8-CycC complex missing the αB-helix (first 22 residues in the N-terminus segment of CDK8) has an affinity of 300.71 nM against 7.05 nM for the native complex (Schneider et al., 2011). Thus, in addition to mediating a specific interaction between the CDK8 and CycC, the αB-helix contributes to the tight binding between CDK8 and CycC. For comparison, the affinities of native CDK9-CycT1, CDK2-CycA and CDK7-CycH are weaker by at least 1 order of magnitude: 300 nM (Baumli et al., 2008), 52 nM and 57 nM respectively (Heitz et al., 1997). It is generally assumed that a high affinity to a partner compared to other homologous partners, leads to a highly specific binding to the considered partner. It may be achieved by small structural variations, as seen here with the CDK8-CycC recognition αB-helix**.** Targeting the highlighted specific interaction hotspots between CDK8 and CycC could be a promising approach to design a peptide inhibiting specifically the CDK8-CycC activity by preventing the binding of CDK8 to CycC. Two peptides targeting the CDK2-CycA interface were reported but none of them has yet made it to the clinic. The first one binds at the core of the common binding surface, at the αC-helix/H5-helix interface (Gondeau et al., 2005). The second one targets a surface pocket in CycA, which is a structurally conserved domain comprising H3, H4 and H5 helix of cyclin A (Canela et al., 2006).

Difference in binding surface between the different complexes

After deciphering the common molecular features of the CDK8-CycC interaction surface, we want to assess now if a significant difference exists between the binding surfaces of the studied complexes. In order to highlight possible differences in energy contribution of the residues, we extract the list of the residues that form at least one significant interaction (using the same cutoff as above: absolute (ΔG) > 1 kcal.mol^{-1}) in one of the studied complexes. In other words, instead of taking the intersection of the lists of important residues of each system, as we did previously to get the common molecular features, we took the union of these lists. The resulted matrix has been attached in the supporting information (**Figure S12**). To compare the contributions of the residues of each system with each other in a convenient

way, we calculated a correlation matrix from the contribution matrix and present the results as a scatterplot matrix (**Figure 6**).

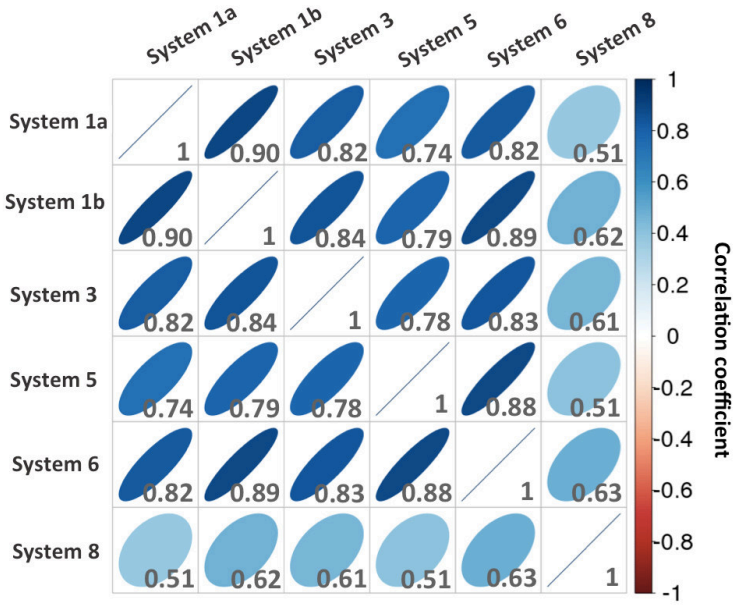


Figure 6: Comparison of the energy contributions of the residues to CDK8-CycC binding of each system with each other through the calculation of a scatterplot correlation matrix.
The compared residues are those presenting at least one significant energy contribution (absolute (ΔG) > 1 kcal.mol^{-1}) in one of the studied CDK8-CycC complexes.

DMG-out CDK8-CycC complexes

The residues of *DMG-out* conformation complexes (1a, 1b, 3, 5, 6) display very similar energy contribution values since the correlation coefficients are comprised between 0.74 and 0.90. The mutated *DMG-out* conformation complex (system 5) does not exhibit significant difference with the others native *DMG-out* conformation complexes (1a, 1b, 3, 6) in terms of energy contribution of residues. Indeed, it presents a correlation coefficient always superior to 0.74 against them. It is particularly close to system 6 (correlation coefficient = 0.88). Together, these results indicate that the *DMG-out* conformation complexes share a similar binding interaction surface. Moreover, the double-point mutation (CDK8^{E66A_R65A}) does not significantly affect the binding interaction network between CDK8 and CycC. Therefore, the mutant complex presents a similar stability (**Table 3**) associated with a similar binding interaction network compared to native systems.

Difference in binding surface between DMG-in and DMG-out

Although the studied CDK8-CycC complexes share a large common interaction surface, as we detailed previously, the distribution of the energy contribution of the residues of the *DMG-in* complex is the least correlated with that of others complexes, with a correlation coefficient comprised between 0.51 and 0.63. In the *DMG-in* conformation complex, the CycC is slightly shifted towards CDK8 as shown in **Figure 7**. This shift increases the contacts between CycC H3-H4 loop and the CDK8 activation loop, which is folded towards the CycC in *DMG-in* conformation.

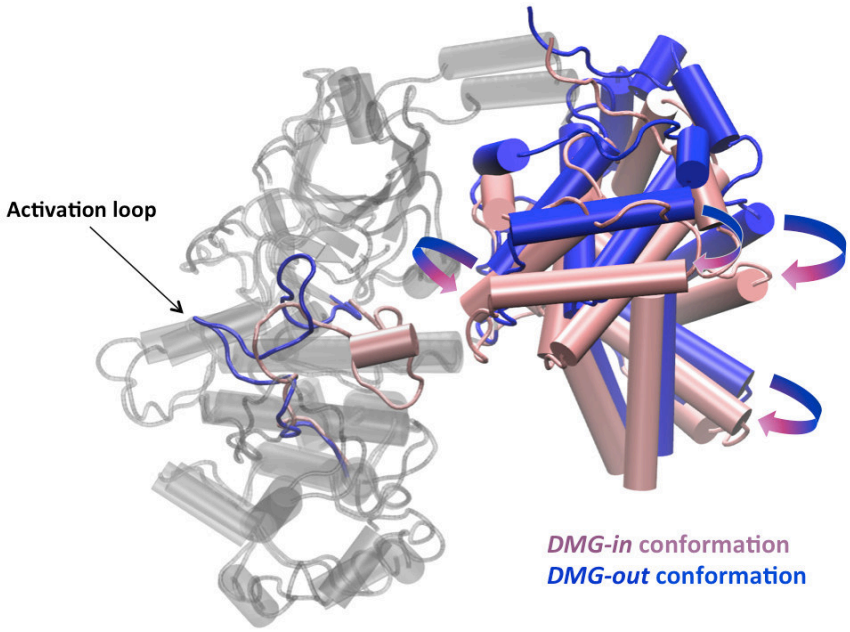


Figure 7: Shift of the Cyclin C toward CDK8 in *DMG-in* conformation.
Representation of the conformations *DMG-in* and *DMG-out* of the CDK8-CycC complex in cartoon. CDK8 is coloured in gray, except the activation loop. The activation loop and the CycC are coloured in pink in the *DMG-in* conformation and in blue in the *DMG-out* conformation.

As a consequence, Arg178^{CDK8} and Pro183^{CDK8} of the activation loop that did not contribute to CDK8-CycC binding in the *DMG-out* conformation complexes are now close to CycC and present a favorable contribution (**Figure S12**). Moreover, the shift of the CycC modifies the interaction network at CDK8-CycC interface which might explain for some residues a change in their energy contribution: Arg13^{CDK8} of the αB-helix, Leu86^{CDK8} on the β4 strand, Asn145^{CDK8}, Trp146^{CDK8} at the C-terminus of the αE-helix, Ala2^{CycC} and Gly3^{CycC} of the N-terminus segment and three residues at the N-terminus of CycC H3-helix (Ile81, Asp82 and leu85) (**Figure S12**). Interestingly, others residues that are far from the interaction surface but part of the binding site also display a difference in their energy contribution in the *DMG-in* conformation system compared to the *DMG-out* one: the Val27^{CDK8} and the Val35^{CDK8} which

are part of the P-loop, Tyr99^{CDK8} and Ala100^{CDK8} in the hinge region and Arg356^{CDK8} of the C-terminus of CDK8.

Activation mechanism of CDK8

In the *DMG-in* conformation complex, we observe that the shift of the CycC toward CDK8 allow the Glu99^{CycC} to be closer to Arg65^{CDK8}. Glu99^{CycC} establishes hydrogen bonds with Arg65^{CDK8}, Arg178^{CDK8} and to a lesser extent with Arg150^{CDK8}. The three arginines also interact with each other through water-mediated hydrogen bonds. This interaction network is maintained over time (**Figure S13**) and could therefore have a role in the stabilization of the activation loop in the *DMG-in* conformation. In this context, we turn to literature to find a possible known role of Arg65^{CDK8}, Arg150^{CDK8} and Arg178^{CDK8} in the activation mechanism of CDK8. In that regard, it was reported that these three arginines are conserved within human CDK members and are involved in the second step of the activation mechanism (Hoeppner et al., 2005). As mentioned in the introduction, the second step of the general activation mechanism of CDKs is the phosphorylation of a residue within the activation loop. The phospho-residue serves as anchor to adjust the orientation of three conserved arginines, thereby inducing a *DFG-in* conformation of the activation loop. In CDK8, these three conserved arginines are Arg65^{CDK8}, Arg150^{CDK8} and Arg178^{CDK8}. However, since in CDK8 there is no phosphorylation, on the basis of crystallographic structure analysis, Glu99^{CycC} was hypothesized to mimic the missing phospho-residue within CDK8 and serves as anchor to adjust the orientation of the three important arginines, Arg65^{CDK8}, Arg150^{CDK8} and Arg178^{CDK8} in CDK8 (Hoeppner et al., 2005). The stable interaction network formed by the three arginines and Glu99^{CycC} observed during the MD simulation supports this hypothesis.

To further investigate this hypothesis and brought a dynamical view of the process, we simulate through targeted molecular dynamic simulation the conformational transition from a *DMG-out* conformation complex to a *DMG-in* one. The restraint was applied only on the activation loop (and not on the whole complex) to verify if a relationship exists between the shift of the CycC and the conformational change of the activation loop (residue 171 to 182). We first check on the stability of the protein structure over time during the TMD simulation, by verifying the RMSF, the RMSD of the protein and the restraint potential over time (**Figure S14**). The *DMG-in* conformation obtained through TMD simulation followed by 50 ns of cMD is in agreement with the one of system 8 (**Figure S15**).

To monitor the shift of the CycC toward CDK8, we measure the distance between Glu99^{CycC} and a stable residue of CDK8, the Lys153^{CDK8} (according to its RMSF, **Figure 2**). As the activation loop gets closer to the CycC, the CycC shifts toward CDK8 as shown by the Lys153^{CDK8} - Glu99^{CycC} distance curve over time (**Figure 8**). At the beginning of the TMD simulation, Arg178^{CDK8} first interacts with the Arg150^{CDK8} and at this stage the CycC already undergoes a small shift (≈ 2.5 Å). This displacement of the CycC enables the Glu99^{CycC} to become closer to Arg65^{CDK8} and to optimize its interaction with it. Then we observe that the interaction of Arg178^{CDK8} with Arg65^{CDK8} and Glu99^{CDK8} occurs at the same time as a second shift of the CycC (≈ 6 Å) toward CDK8. Therefore, the displacement of the CycC might be an important event to adjust the orientation of three conserved arginine residues. During the following 50 ns of cMD production, the Glu99^{CycC}–mediated hydrogen bond interaction network stabilizes and the same interaction network as in the system 8 is formed (**Figure S13**): Arg178^{CDK8} becomes sandwiched between Arg150^{CDK8} and Arg65^{CDK8} and the three arginines forms hydrogen bonds with Glu99^{CycC} (**Figure 8**). The three arginines also interact with each other through water-mediated hydrogen bonds. It may be noted that finding this network is not trivial since only the activation loop (residue 171 to 182), and therefore only Arg178^{CDK8} was submitted to the restraint potential (Arg150^{CDK8}, Arg65^{CDK8} and Glu99^{CycC} were not under restraint).

From these results, it appears that the Glu99^{CycC} and the shift of the CycC are important to orient and stabilize the three conserved arginines known to be involved in the second step of the general activation mechanism of other CDK members. Therefore, our observations support the hypothesis that the Glu99^{CycC} in CDK8 mimics the missing phospho-residue which role is to adjust the orientation of three conserved arginines thereby inducing a *DFG-in* conformation of the activation loop. In addition to that, our results suggest that a shift of the CycC toward the CDK8 is also required to obtain the active form of CDK8-CycC.

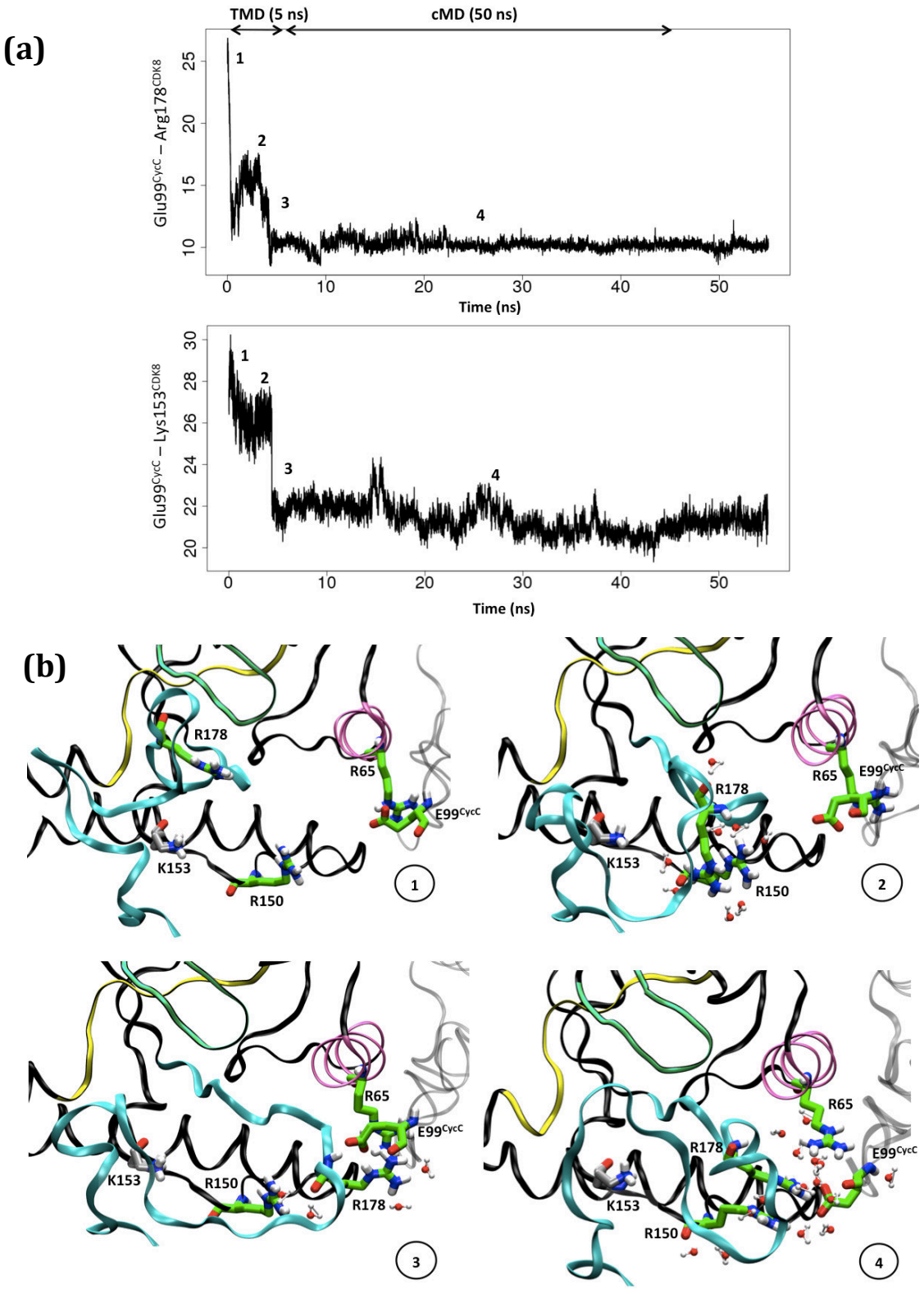


Figure 9: Conformational transition of CDK8 activation loop from a *DMG-out* to a *DMG-in* conformation.

(a) The distance between $Glu99^{CycC}$ and $Arg178^{CDK8}$ enables to monitor the transition of the activation loop over the simulation time course since $Arg178^{CDK8}$ is part of the activation loop and restraint in the TMD simulation (activation loop: residue 171 to 182). $Lys153^{CDK8}$ belongs

to a few flexible region (cf **Figure 2**). The distance between $Glu99^{CycC}$ and $Lys153^{CDK8}$ enables to monitor the displacement of the CycC toward CDK8. **(b)** Orientation of the three conserved arginines $Arg65^{CDK8}$, $Arg150^{CDK8}$ and $Arg178^{CDK8}$ and, the $Glu99^{CycC}$ over the simulation time course. CDK8 is represented in dark gray ribbon except the regions of the kinase domain containing the conserved motifs, particularly the activation loop is in cyan. CycC is in light gray. $Arg65^{CDK8}$, $Arg150^{CDK8}$, $Arg178^{CDK8}$, $Glu99^{CycC}$ and $Lys153^{CDK8}$ are represented in sticks. The carbon atom of the three arginines and the glutamate are in light green and that of $Lys153^{CDK8}$ is in gray.

CONCLUSION

Theoretical studies were conducted on the human CDK8-CycC complex in order to provide more structural information about the binding of CycC to CDK8 that is an important target in cancer therapy. We first investigated the role of CycC on the structure and dynamics of CDK8. We found that the CycC brings stability on the CDK8 structure and impacts its dynamics in both the active (*DMG-in*) and inactive form (*DMG-out*) of the complex. Unlike CDK2, where the binding of a type II inhibitor to CDK2-CycB results in the dissociation of CycB from CDK2 in a competitive manner (Alexander et al., 2015), Schneider *et al* have shown that the binding of a type II inhibitor to CDK8-CycC does not dissociate CycC (Schneider et al., 2013). Our findings converge to this result since the presence of a type II inhibitor does not affect the stabilizing effect of the CycC on CDK8. The interaction free energy values of CDK8-CycC binding calculated through the MM-GBSA method confirm these results, and show that the CycC stabilizes both CDK8 forms (active and inactive) to the same extent.

The analysis of the interaction between CDK8 and CycC, through the per-residue binding free energy decomposition, highlighted 26 hotspot residues uniformly distributed on the interaction surface. In all the studied simulations, those residues establish strong and favorable interactions ($\Delta G_{total} < -1 kcal.mol^{-1}$) that contribute to CDK8-CycC binding. 19 of the 26 important residues belong to the conserved common interaction surface in the human CDK family with Cyc partners. On the contrary, the remaining 7 hotspot residues are situated in two binding sites of the interaction surface that are specific to CDK8-CycC complex and involve the proline rich C-terminus segment, the CDK8 αB-helix and the N-terminus segment of CycC. These key amino acids proposed in this work are valuable information to design an inhibitor, that will effectively prevent the binding of the CycC to CDK8, which will block the activation of the CDK8-CycC complex, thereby interfering with the function of CDK8 as an oncogene. The active and the inactive forms display some differences in their CDK8-CycC binding energy contribution values. These differences might be explained by the flip of the

activation loop from a *DMG-out* to a *DMG-in* conformation and the displacement of the CycC toward CDK8 in the active form.

The simulation of the conformational transition from the inactive to the active form through TMD simulation showed that this displacement of the CycC toward CDK8 occurs during the conformational change. This displacement is an important event to adjust the orientation of three conserved arginine residues ($Arg65^{CDK8}$, $Arg178^{CDK8}$ and $Arg150^{CDK8}$), which is meditated by the $Glu99^{CycC}$, thereby inducing a *DMG-in* conformation (active form). The active form is maintained through a hydrogen bond interaction network involving the three arginines and the $Glu99^{CycC}$. In human CDK family, the three conserved arginine residues, together with a phosphorylated residue, are known to have a role in the conformational change of CDK and in the stabilization of the active form. Our TMD simulation suggests that $Glu99^{CycC}$ assumes the role of the missing phosphorylated residue in CDK8, confirming the hypothesis found in the litterature.

Our study provides interesting molecular insights, describing the interaction between CDK8 and CycC in terms of structure and energy. Since this interaction is essential to the activity of CDK8, the particular characteristics of this interaction and of its mechanism of activation highlighted in this study, are valuable information to design specific compounds targeting the CDK8/CycC interface. In a more general view, these results demonstrate the importance of keeping the CycC in computational studies when studying the human CDK8 protein in both the inactive and active form.

SUPPORTING INFORMATION

Section S1: Model building and system preparation.

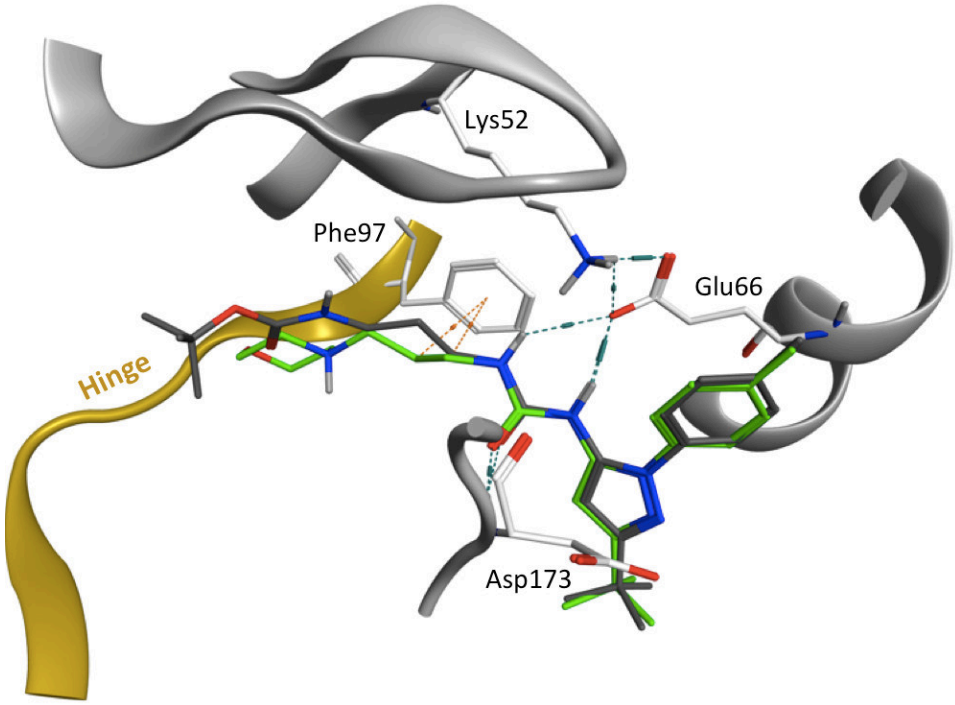


Figure S1: Binding site of the crystal structure of human CDK8 in complex with inhibitors 0SR and 0SO from respectively PDB id. 4F6U and 4F7L.
Inhibitors 0SR to 0SO are colored respectively in green and black. Interactions between the urea and the common scaffold of the inhibitors and the residues Glu66 and Asp173 are shown in blue dashed lines. The protein is represented in white cartoon except the hinge region colored in yellow cartoon.

```
Uniprot_Cyclin-C      ---MAGNFWQSSHYLQWILDKQDLLKERQKDLKFLSEEEYWKLQIFFTNVIQALGEHLKL
4F6U.B                DKAMAGNFWQSSHYLQWILDKQDLLKERQKDLKFLSEEEYWKLQIFFTNVIQALGEHLKL
                         ************************************************************

Uniprot_Cyclin-C      RQQVIATATVYFKRFYARYSLKSIDPVLMAPTCVFLASKVEEFGVVSNTRLIAAATSVLK
4F6U.B                RQQVIATATVYFKRFYARYSLKSIDPVLMAPTCVFLASKVEEFGVVSNTRLIAAATSVLK
                      ************************************************************

Uniprot_Cyclin-C      TRFSYAFPKEFPYRMNHILECEFYLLELMDCCLIVYHPYRPLLQYVQDMGQEDMLLPLAW
4F6U.B                TRFSYAFPKEFPYRMNHILECEFYLLELMDCCLIVYHPYRPLLQYVQDMGQEDMLLPLAW
                      ************************************************************

Uniprot_Cyclin-C      RIVNDTYRTDLCLLYPPFMIALACLHVACVVQQKDARQWFAELSVDMEKILEIIRVILKL
4F6U.B                RIVNDTYRTDLCLLYPPFMIALACLHVACVVQQKDARQWFAELSVDMEKILEIIRVILKL
                      ************************************************************

Uniprot_Cyclin-C      YEQWKNFDERKEMATILSKMPKPKPPPNSEGEQGPNGSQNSSYSQS
4F6U.B                YEQWKNFDERKEMATILSKMPKPKPPP------------------
                      ***************************
```

Figure S2: Alignment of the UNIPROT sequence of the human Cyclin C with the Cyclin C sequence of 4F6U (PDB id.).

```
Uniprot_CDK8      --MDYDFKVKLSSERERVEDLFEYEGCKVGRGTYGHVYKAKR-KDGKDDKDYALKQIE--
4F6U.A            DKMDYDFKVKLSSERERVEDLFEYEGCKVGRGTYGHVYKAKR-KDGKDDKDYALKQIE--
1BI8.A            --MEKDGLCRADQQYECVAE--------IGEGAYGKVFKARDLKNG--GRFVALKRVRVQ
1G3N.A            --MEKDGLCRADQQYECVAE--------IGEGAYGKVFKARDLKNG--GRFVALKRVRVQ
                    *: *   : ..: * * :         :*.*:**:*:**:  *:*  .:  ***::.

Uniprot_CDK8      -GT-GISMSACREIALLRELK---HPNVISLQKVF-LSHADR--KVWLLFDYAEHDLWHI
4F6U.A            -GT-GISMSACREIALLRELK---HPNVISLQKVF-LSHADR--KVWLLFDYAEHDLWHI
1BI8.A            TGEEGMPLSTIREVAVLRHLETFEHPNVVRLFDVCTVSRTDRETKLTLVFEHVDQDLTTY
1G3N.A            TGEEGMPLSTIREVAVLRHLETFEHPNVVRLFDVCTVSRTDRETKLTLVFEHVDQDLTTY
                   *  *:.:*: **:*:**.*:   ****: * .*  :*::**  *: *:*::.::**

Uniprot_CDK8      IKFHRASKANKKPVQLPRGMVKSLLYQILDGIHYLHANWVLHRDLKPANILVMGEGPERG
4F6U.A            IKFHRASK-----VQLPRGMVKSLLYQILDGIHYLHANWVLHRDLKPANILVMGEGPERG
1BI8.A            L-----DKVPEPGV--PTETIKDMMFQLLRGLDFLHSHRVVHRDLKPQNILVTSSG----
1G3N.A            L-----DKVPEPGV--PTETIKDMMFQLLRGLDFLHSHRVVHRDLKPQNILVTSSG----
                  :     .*     *  *   :*.:::*:* *:.:**:: *:****** **** ..*

Uniprot_CDK8      RVKIADMGFARLFNSPLKPLADLDPVVVTFWYRAPELLLGARHYTKAIDIWAIGCIFAEL
4F6U.A            RVKIADMGF-----------------VVTFWYRAPELLLGARHYTKAIDIWAIGCIFAEL
1BI8.A            QIKLADFGLARIYSFQMA----LTSVVVTLWYRAPEVLLQSS-YATPVDLWSVGCIFAEM
1G3N.A            QIKLADFGLARIYSFQMA----LTSVVVTLWYRAPEVLLQSS-YATPVDLWSVGCIFAEM
                  ::*:**:*:                ***:******:** :  *:..:*:*::******:

Uniprot_CDK8      LTSEPIFHCRQEDIKTSNPYHHDQLDRIFNVMGFPADKDWEDIKKMPEHSTLMKDFRRNT
4F6U.A            LTSEPIFHCRQE-----NPYHHDQLDRIFNVMGFPADKDWEDIKKMPEHSTLMKDFRRNT
1BI8.A            FRRKPLFR-GSSDV--------DQLGKILDVIGLPGEEDWPRDVALPRQAFHSKSAQP--
1G3N.A            FRRKPLFR-GSSDV--------DQLGKILDVIGLPGEEDWPRDVALPRQAFHSKSAQP--
                  :  :*:*:  ..          ***.:*::*:*:*.::**     :*.::   *. :

Uniprot_CDK8      YTNCSLIKYMEKHKVKPDSKAFHLLQKLLTMDPIKRITSEQAMQDPYFLEDPLPTSDVFA
4F6U.A            YTNCSLIKYMEKHKVKPDSKAFHLLQKLLTMDPIKRITSEQAMQDPYFLEDPLPTSDVFA
1BI8.A            ---------IEKFVTDIDELGKDLLLKCLTFNPAKRISAYSALSHPYFQDLERCKENLDS
1G3N.A            ---------IEKFVTDIDELGKDLLLKCLTFNPAKRISAYSALSHPYFQDLERCKENLDS
                          :**. .. *. . .** * **::* ***:: .*:..*** :    ..:: :

Uniprot_CDK8      GCQIPYPKREFLTEEEPDDKGDKKNQQQQQGNNHTNGTGHPGNQDSSHTQGPPLKKVRVV
4F6U.A            GCQIPYPKREFLTEEEPDDKGDKKNQQQQQGNNHTNGTGHPGNQDSSHTQGPPLKK----
1BI8.A            ------------------------------------HLPPSQNTSELNTA--------
1G3N.A            ------------------------------------HLPPSQNTSELNTA--------
                                                       * .*::*. : .
```

Figure S3: Alignment of the human sequences of CDK8 (canonical UNIPROT sequence and 4F6U (PDB id.) sequence) with the template sequences of human CDK6 (1BI8 and 1G3N (PDB id.)).

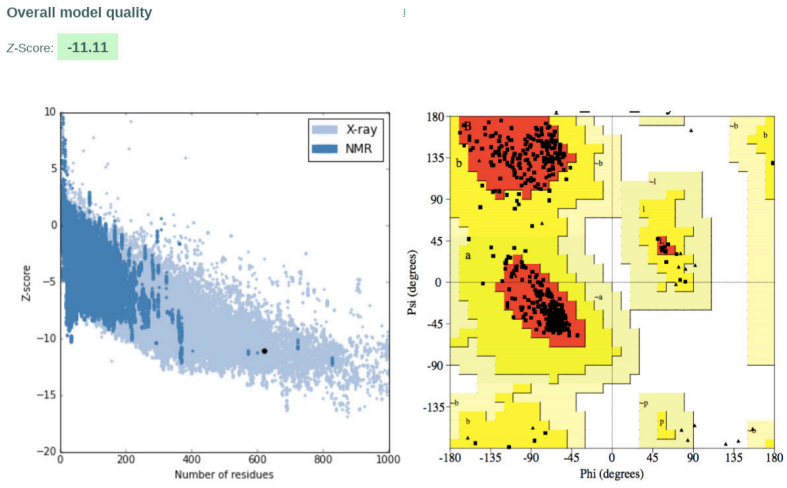


Figure S4: Model validation.
(**Left**) Plot of the Z-score. The *Z*-score indicates overall model quality. The plot contains the *z*-scores of all experimentally determined protein chains in current PDB. In this plot, groups of structures from different sources (X-ray, NMR) are distinguished by different colors. The z-score of the reconstructed model of 4F6U (*PDB id.*) is represented by a black point. Its value is within the range of scores typically found for native proteins of similar size. **(Right)** Ramachandran plot showing that most of the model residues display backbone dihedral angle phi and psi values in the energetically allowed regions.

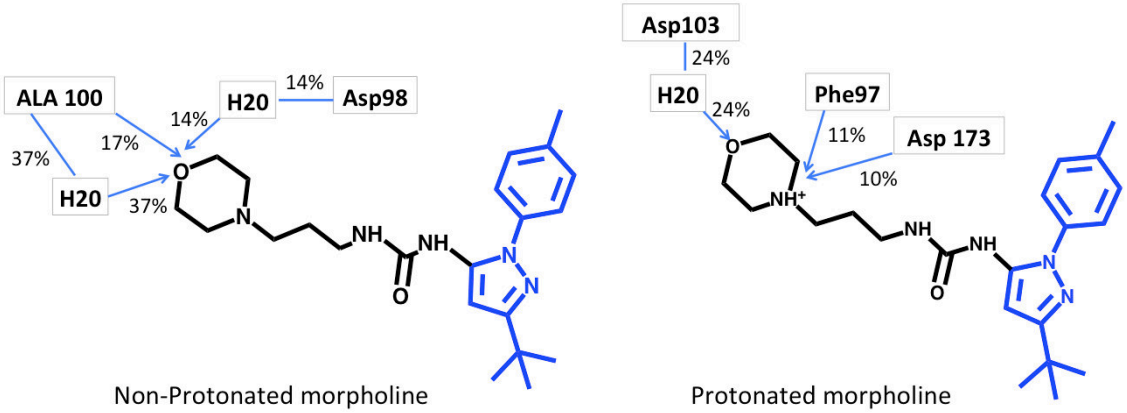


Figure S5: Difference in protein-ligand interaction between the form non-protonated and protonated of the morpholine of inhibitor 0SR.
The interactions involving the scaffold (drawn in blue) and the urea where not represented because no difference has been noted. These interactions have been calculated along a brute force MD simulation of 1 microseconde. The percentage represents the time of the simulation the interaction is maintained. With the protonated morpholine the interaction with Ala100 is not observed any more.

Section S2: Effect of the exclusion of CycC on structure and dynamics of CDK8.

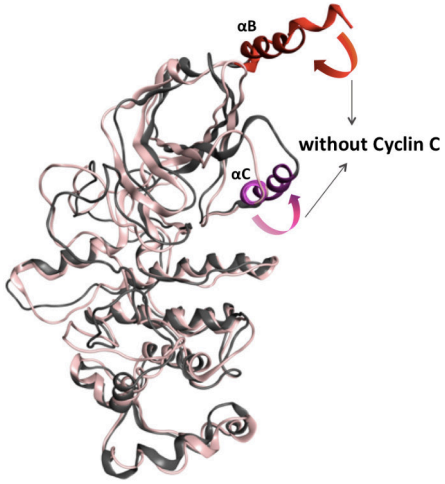


Figure S6: Conformational changes observed in the absence of CycC.
The protein is represented in gray ribbon in the system 2a (without CycC) and light pink in the system 1a (with CycC), except αB and αC helix.

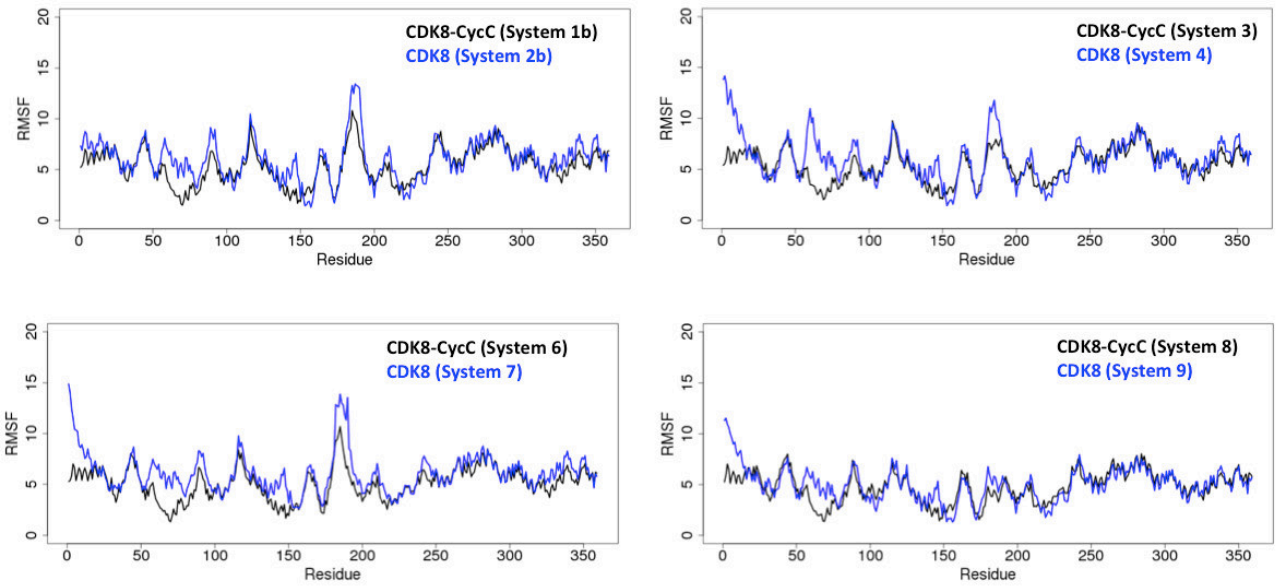


Figure S7: Comparison of the root mean square atomic fluctuations of CDK8 in the presence and absence of CycC.

Besides the application of PCA on combined trajectory (with and without CycC), PCA was also applied on each individual trajectory. The goal is to capture the major CDK8 motions observed in the different systems and assess if the presence of the CycC impacts the dynamics of CDK8. To visualize the largest amplitude motions, a PDB format trajectory has been produced that interpolates between the most dissimilar structures in the distribution along PC1.

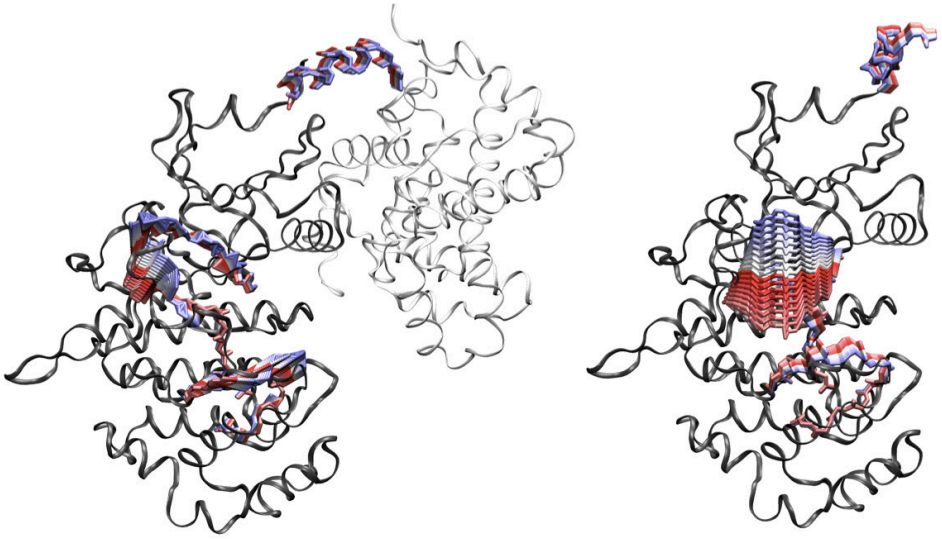


Figure S8: Largest fluctuation motions (PC1) in the *DMG-out* conformation complex in the presence (left) and absence (right) of CycC.
The regions of largest moves are represented in a color gradient from blue to red to capture the dynamics.

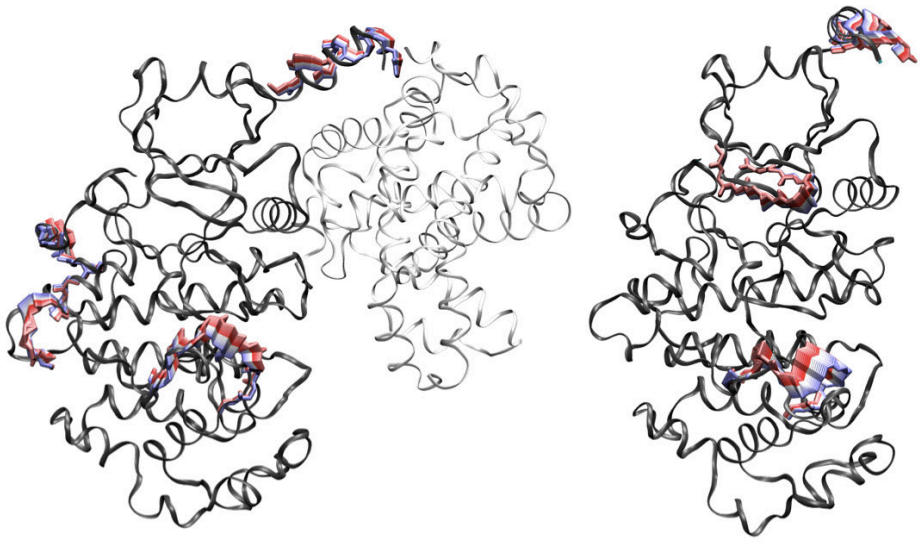


Figure S9: Largest fluctuation motions (PC1) in the *DMG-in* conformation complex in the presence (left) and absence (right) of CycC.
The regions of largest moves are represented in a color gradient from blue to red to capture the dynamics.

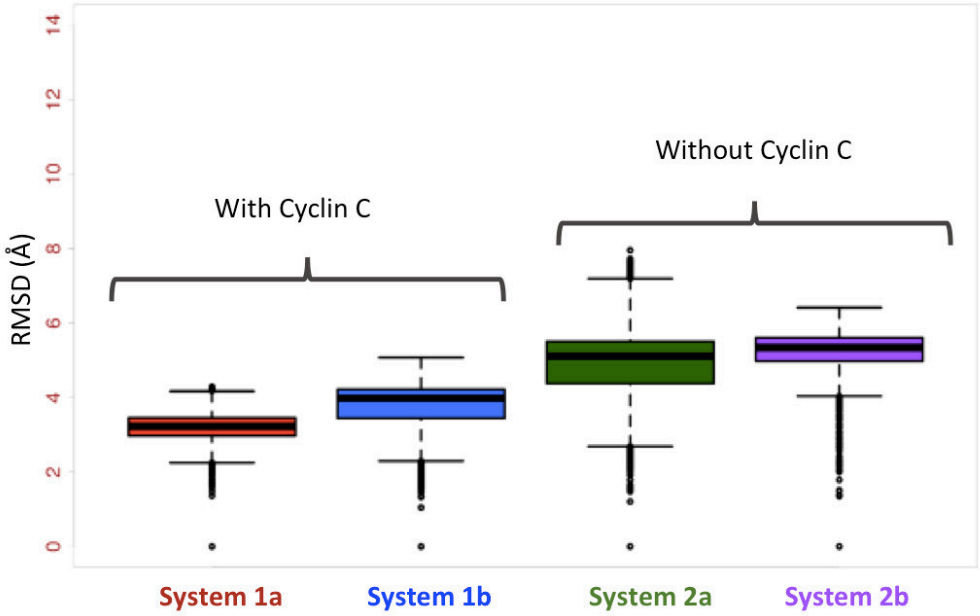


Figure S10: Distribution of the ligand RMSD in the presence and absence of CycC.

Section S3: Characterization of the protein-protein interactions between CDK8 and CycC.

(a)
(b)
(c)
(d)
Binding free energy (kcal.mol-1)
2
0
-2
-4
-6
-8
-10
MET_1 ASP_2 TYR_3 PHE_5 LEU_9 ARG_13 ILE_59 MET_61 CYS_64 ARG_65 ILE_67 ALA_68 LEU_69 GLU_72 LEU_73 LEU_86 ARG_91 TYR_141 ASN_145 TRP_146 MET_1 ASN_4 GLN_7 SER_8 SER_9 LYS_79 SER_80 ILE_81 ASP_82 LEU_85 VAL_97 ASN_133 GLU_137 PHE_140 TYR_141 LEU_143 LEU_145 ASP_147 CYS_148 CYS_149 LEU_150 ILE_151 PRO_260 LYS_261 PRO_263 PRO_264
MET_1 ASP_2 TYR_3 PHE_5 LEU_9 ARG_13 GLU_14 ILE_59 MET_61 CYS_64 ARG_65 ILE_67 ALA_68 LEU_69 ARG_71 GLU_72 LEU_86 ARG_91 TYR_141 ASN_145 TRP_146 MET_1 ALA_2 ASN_4 SER_8 SER_9 LYS_79 SER_80 ILE_81 ASP_82 VAL_97 GLU_137 PHE_140 TYR_141 LEU_143 LEU_145 ASP_147 CYS_148 LEU_150 ILE_151 TYR_153 PRO_260 LYS_261 PRO_263 PRO_264
MET_1 ASP_2 TYR_3 PHE_5 LYS_6 LEU_9 ARG_13 ILE_59 MET_61 CYS_64 ARG_65 ILE_67 ALA_68 LEU_69 VAL_84 LEU_86 HIE_88 ARG_91 ASN_145 TRP_146 ARG_150 MET_1 ALA_2 SER_8 SER_9 LYS_79 SER_80 ILE_81 ASP_82 LEU_85 VAL_97 GLU_99 GLU_137 PHE_140 TYR_141 LEU_143 GLU_144 LEU_145 ASP_147 CYS_148 LEU_150 ILE_151 PRO_260 LYS_261 PRO_263 PRO_264
MET_1 ASP_2 TYR_3 PHE_5 LYS_6 LEU_9 ARG_13 ILE_59 MET_61 CYS_64 ALA_65 ALA_68 LEU_69 ARG_71 GLU_72 VAL_84 LEU_86 HIE_88 ARG_91 ASN_145 TRP_146 MET_1 ALA_2 SER_8 SER_9 TYR_76 SER_80 ILE_81 ASP_82 LEU_85 LYS_96 VAL_97 GLU_137 PHE_140 TYR_141 LEU_143 GLU_144 ASP_147 CYS_148 CYS_149 ILE_151 TYR_153 PRO_260 LYS_261 PRO_263 PRO_264

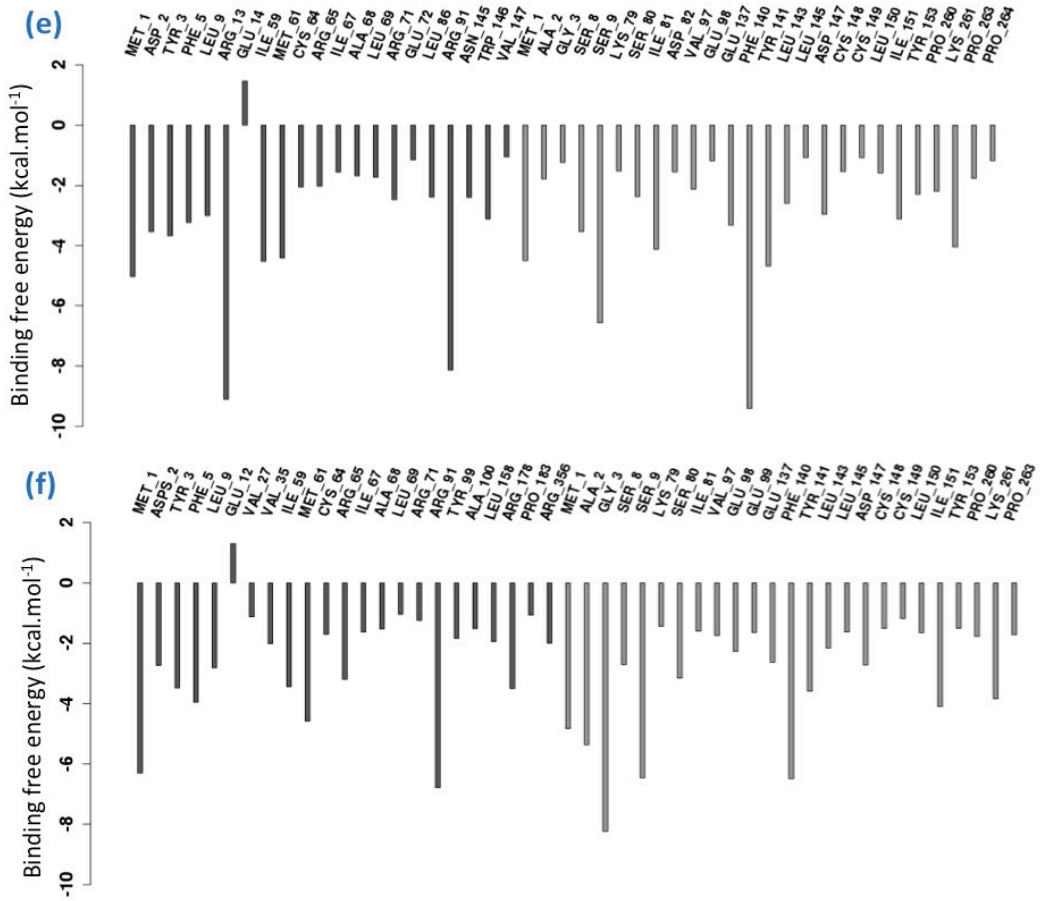


Figure S11: Energy contribution of important residues of each studied CDK8-CycC complex.

An important residue is defined as a residue having its absolute value of ΔG_{total} superior to 1kcal.mol^{-1}. (a): system 1a, (b): system 1b, (c): system 3, (d): system 5, (e): system 6, (f): system 8. Residues whose bars are coloured in dark gray belong to CDK8 and those in light gray to CycC.

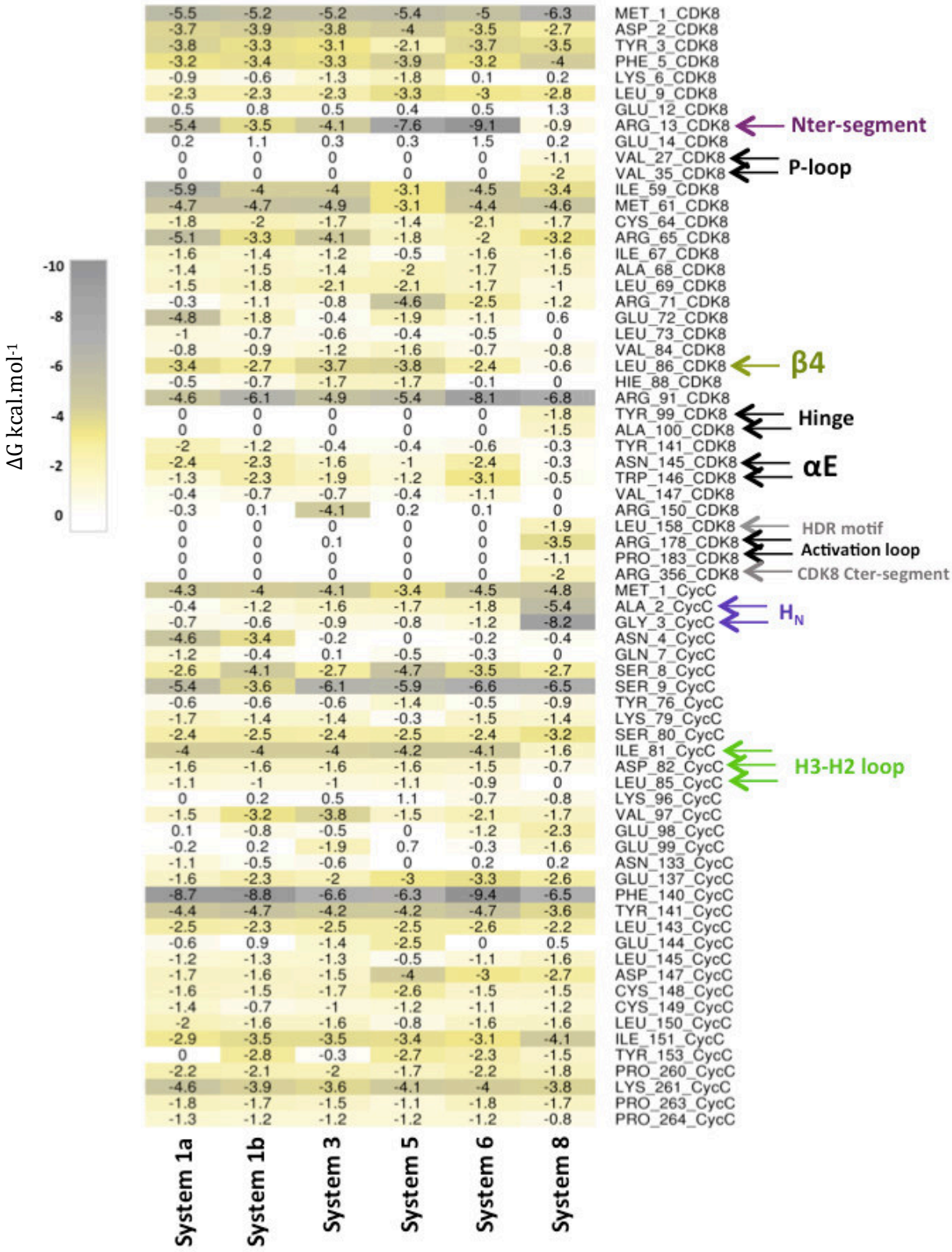


Figure S12: Matrix of the per-residue energy contribution (ΔG without entropy in kcal.mol^{-1}) of the residues that present at least one significant energy contribution (absolute(ΔG) > 1 kcal.mol^{-1}) in one of the studied CDK8-CycC complexes.
Arrows point residues that contribute differently in DMG-in and DMG-out conformation. The secondary structure they belong to is indicated also and coloured following the same colour code as Figure 5 (gray color was used to distinct two narrow indications).

Section S4: Activation mechanism of CDK8

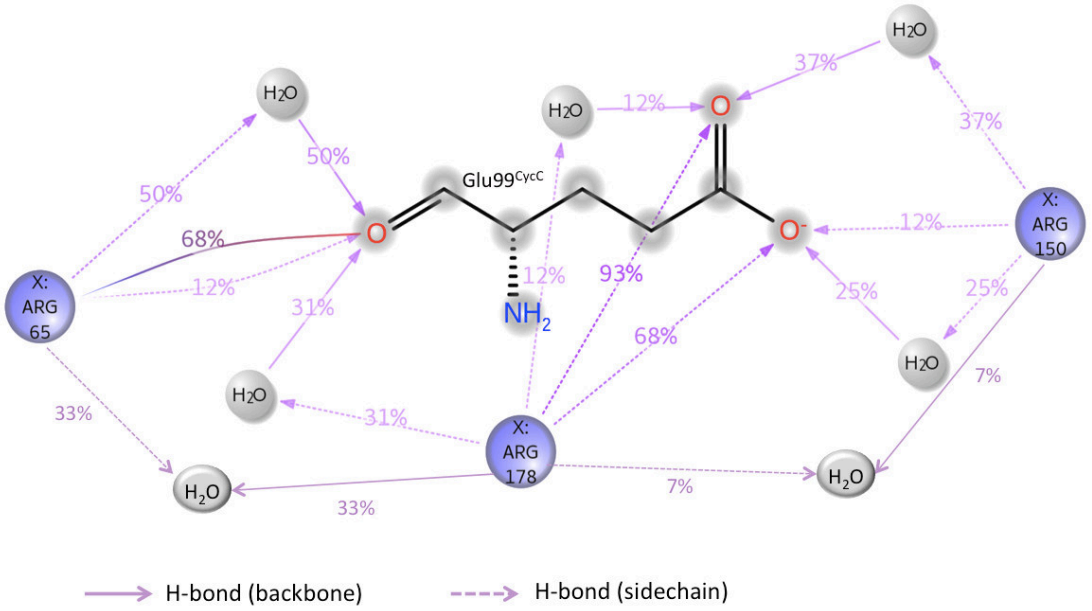


Figure S13: Interaction network between the three conserved arginines of CDK8 (Arg65^{CDK8}, Arg150^{CDK8} and Arg178^{CDK8}) and the glutamate 99 of the CycC in the *DMG-in* conformation system.
The interactions were calculated on 1μs simulation of the system 8.

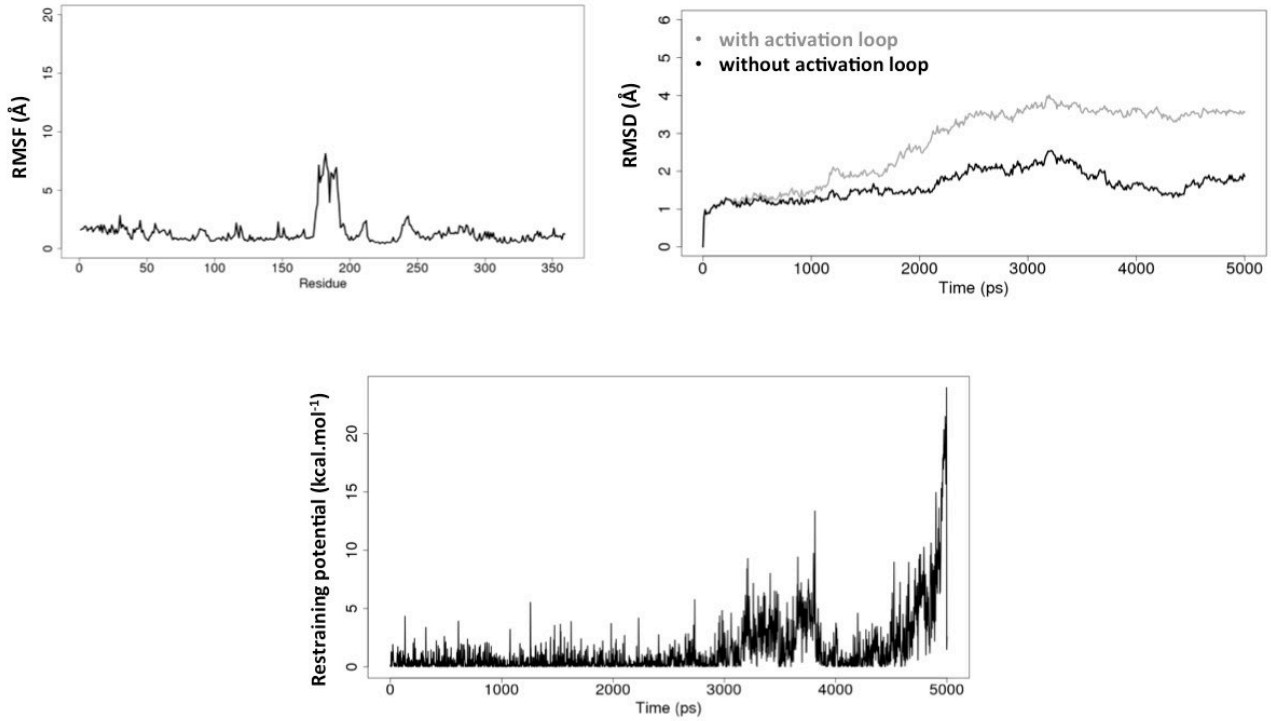


Figure S14: Analysis of the stability of the protein structure over time during the TMD simulation through the investigation of the RMSF, RMSD and the restraining potential ($V_{restraint}$).

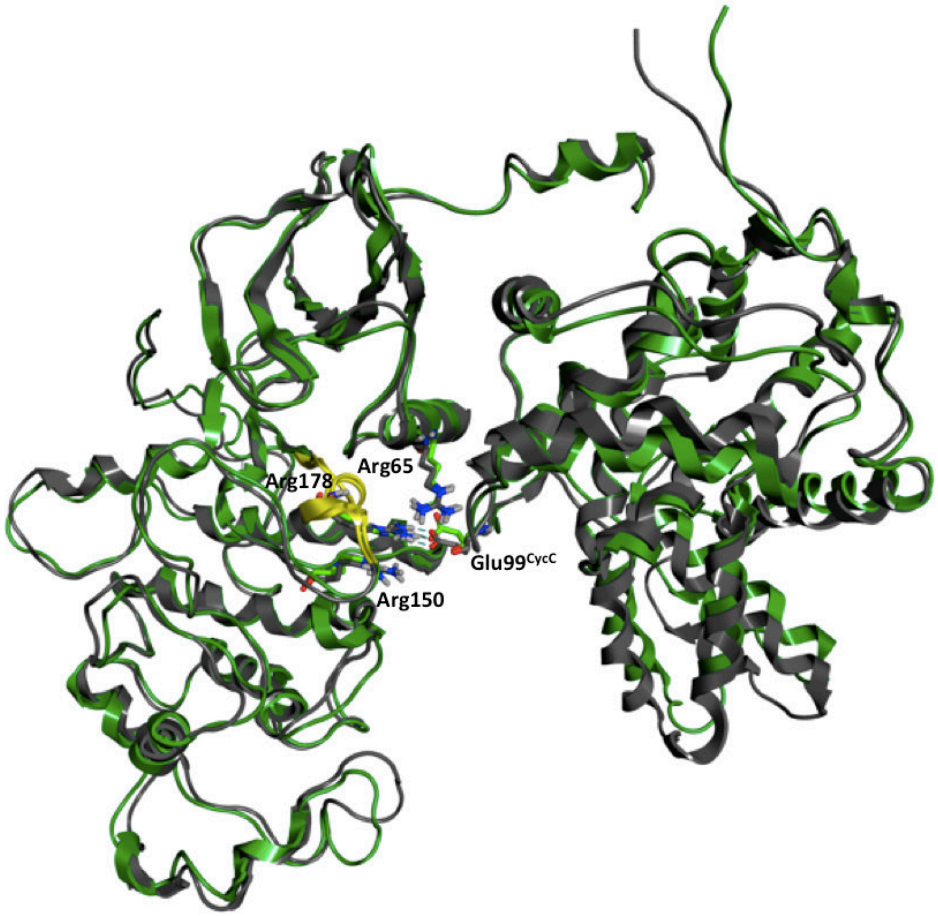


Figure S15: Comparison of the structures of CDK8-CycC in *DMG-in* conformation obtained from TMD and cMD simulation.
The structure obtained from 5 ns of TMD simulation followed by 50 ns of cMD (in green ribbon) was compared to the structure obtained from the 1μs of cMD simulation (in dark gray ribbon). Residues 171 to 182 of CDK8 protein were coloured in yellow. These are the residues constrained during the TMD simulation. The three conserved arginines (Arg65^{CDK8}, Arg150 ^{CDK8} and Arg 178 ^{CDK8}) were represented as sticks.

B. Conclusion

Par le biais de simulations de dynamique moléculaire et de calculs d'énergie libre de liaison, le rôle de la CycC et son impact sur la structure et la dynamique de CDK8 dans plusieurs cas (forme active et inactive, forme apo ou holo, mutations) a été étudié. Nous avons mis en évidence que la CycC stabilise la structure de CDK8 dans ses formes active et inactive. Elle est indispensable pour maintenir la structure native de CDK8 observées dans le complexe CDK8-CycC, et a un impact déterminant dans le comportement dynamique de cette protéine kinase. Les résidus contribuant fortement à l'interaction CDK8-CycC ont été mis en évidence. Faits intéressants, parmi les résidus importants identifiés, certains appartiennent à des sites d'interaction spécifiques à CDK8 tandis que les autres résidus se trouvent sur des sites d'interaction communs à la famille des CDKs humaines. Compte tenu de l'émergence récente de CDK8 en tant que cible d'intérêt pour le cancer colorectal et mammaire, ces résultats sont très utiles pour la conception d'inhibiteurs peptidiques ciblant spécifiquement l'interface CDK8-CycC. Enfin, nous avons simulé le changement de conformation de la forme inactive à la forme active du complexe CDK8-CycC au moyen de simulations de dynamique moléculaire dirigée. Les analyses montrent que le changement de conformation s'accompagne d'une légère rotation de la CycC. Ce léger déplacement semble indispensable pour stabiliser la forme active de CDK8, via une interaction critique avec un résidu de la CycC. Le rôle de ce résidu dans la stabilisation du complexe actif est normalement assuré par un résidu phosphorylé de CDK, absent dans le cas de CDK8.

Outre les connaissances théoriques apportées sur le complexe CDK8-CycC, notre étude a mis en lumière le rôle crucial de la CycC dans le complexe CDK8-CycC et la nécessité de la prendre en compte dans les études de modélisation de CDK8. En outre, cette étude exemplifie l'importance de bien caractériser le mécanisme moléculaire d'action (MMA) d'une cible avant le commencement d'un projet de *drug design* (I.D.1.b), page28) ; investiguer le rôle de la cyclin C, de même qu'étudier la cinétique d'interaction d'un complexe protéine-ligand en font parties. La partie suivante s'attèle justement à présenter une méthode computationnelle pour estimer la constante de dissociation d'un ligand de sa cible. Nous verrons d'ailleurs que la cycline C, bien que ne contenant pas le site de liaison, peut impacter la cinétique de dissociation de l'inhibiteur.

III. PREDICTION QUALITATIVE DU TEMPS DE RESIDENCE ET SKR

A. Méthode de prédiction du temps de résidence

Dans cette première partie, nous présenterons la méthode développée de prédiction du temps résidence basée sur de la dynamique moléculaire dirigée. Notre partenaire industriel l'Institut de Recherche Servier, porteur de ce projet de thèse, souhaite disposer d'une méthode de simulation numérique capable de fournir un classement selon le temps résidence d'une série de molécules *drug-like*. L'outil devra être simple d'utilisation, rapide (≈ 5 molécules par jour) et applicable à tout système pour permettre l'analyse de touches (*hit*) ou de chef de file (*lead*) en phase d'optimisation des molécules. Cette méthode a été développée et testée sur un ensemble de 10 inhibiteurs du complexe CDK8-CycC pour lesquels nous disposons de mesures de temps de résidence homogènes ainsi que de structures cristallographiques pour certains des inhibiteurs (**Figure 26**, **Figure 27**).

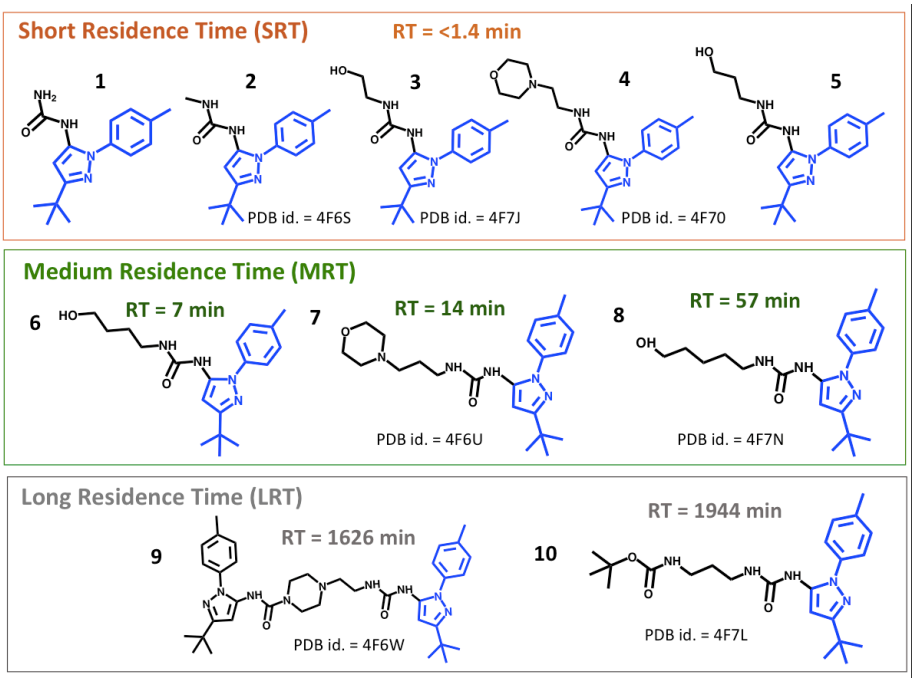


Figure 26: Les structures chimiques des dix inhibiteurs de CDK8.
Les molécules ont été numérotées de 1 à 10. Le temps de résidence expérimental est noté RT pour *Residence Time*. Les inhibiteurs sont classés en trois groupes les courts (SRT), les moyens (MRT) et les longs temps de résidence (LRT). Le *scaffold* commun le 1- (3-tert-butyl-1-p-tolyl-1H-pyrazol-5-yl) urée est coloré en bleu.

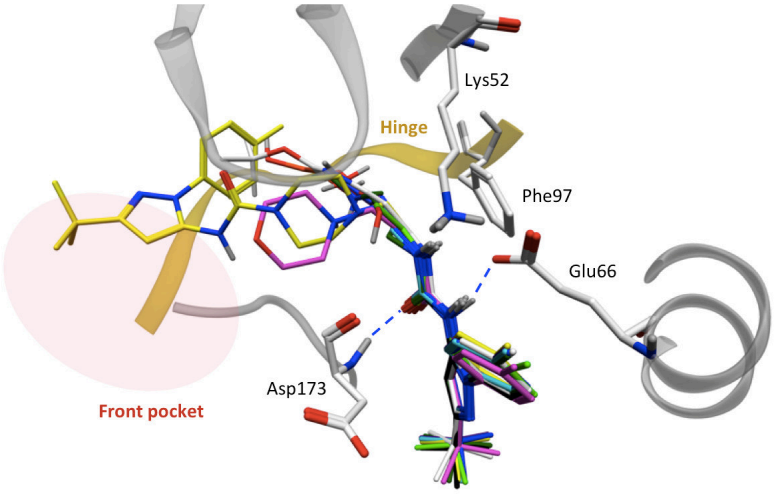


Figure 27 : Site de liaison de la structure cristallographique de CDK8 humaine liée avec les inhibiteurs 1-10.
Les interactions entre l'urée du *scaffold* commun des inhibiteurs et les résidus Glu66 et Asp173 sont représentées par des lignes en pointillées bleues. La protéine est représentée en ruban blanc sauf la région charnière (*hinge motif*) colorée en ruban jaune. Un cercle rouge représente l'emplacement de la *front pocket*. Les inhibiteurs 1 à 10 sont colorés en gris, noir, cyan, vert clair, rose, violet, vert foncé, bleu, jaune et blanc.

Id. inhibiteur	Kd, µM	k_{on}, s^{-1}. μM^{-1}	k_{off}, s^{-1}	cLogP
1	3.24	(-)	(-)	3.03
2	1.57	(-)	(-)	3.26
3	5.82	(-)	(-)	2.75
4	1.82	(-)	(-)	2.86
5	3.53	(-)	(-)	2.85
6	1.30	1.85×10^{-3}	2.41×10^{-3}	3.31
7	0.70	1.68×10^{-3}	1.18×10^{-3}	2.97
8	0.08	3.65×10^{-3}	2.90×10^{-4}	3.72
9	0.03	2.99×10^{-4}	1.02×10^{-5}	6.28
10	0.01	5.73×10^{-4}	8.57×10^{-6}	3.86

Tableau 8 : Mesures expérimentales du K_d, k_{on}, k_{off}, et valeurs calculées du cLogP des 10 inhibiteurs de CDK8.
Le cLogP est une mesure de la lipophilicité du composé. Il a été calculé avec ChemDraw.

A partir des simulations de dynamique moléculaire décrivant la dissociation des inhibiteurs, une étude structure-cinétique (*SKR : Structure Kinetics Relationship*) est menée sur la base des

interactions protéine-ligands calculées tout au long du processus de dissociation. L'ensemble de ces résultats est présenté sous forme d'un article en cours de préparation.

1. Article en préparation

Estimation of drug-target residence time by targeted molecular dynamics simulations

Sonia Ziada[1], Julien Diharce[2], Eric Rimbaud[3], Pierre Ducrot[3], Samia Aci-Sèche[1]*, and Pascal Bonnet[1]*

[1]Institut de Chimie Organique et Analytique (ICOA), UMR CNRS-Université d'Orléans 7311, Université d'Orléans BP 6759, 45067 Orléans Cedex 2, France.

[2]Biologie intégrée du Globule Rouge, UMR_S 1134 Inserm - Université Paris 7, Paris Diderot, DSIMB team, Institut National de la Transfusion Sanguine, 6 Rue Alexandre Cabanel, 75015 Paris cedex 15.

[3]Institut de Recherches Servier, 125 Chemin de Ronde, 78290 Croissy-sur-Seine, France.

*Author to whom all correspondence should be addressed:

Pascal Bonnet: Tel: +33 238 417 254, Fax: +33 238 417 254, E-mail: pascal.bonnet@univ-orleans.fr

Samia Aci-Sèche: Tel: +33 238 419 902, Fax: +33 238 417 254, E-mail: samia.aci@cnrs-orleans.fr

ABSTRACT

Since researchers have uncovered that the binding duration of a drug molecule to its protein target could significantly impact its clinical efficacy, drug-target residence time has emerged as a key selection factor in drug discovery. The challenge in studying the residence time, in early drug discovery stages lies in how to cost-effectively determine the residence time at sufficient throughput for systematic assessment of compounds. Today it remains a lack of computational protocols to estimate this parameter, particularly for large and flexible protein target and drugs. Here, we report an efficient computational protocol, based on the targeted molecular dynamics method, for the ranking of drug candidates by their residence time and obtaining insights into ligand-target dissociation mechanisms. The method was assessed on a data set of 10 arylpyrazole inhibitors of CDK8 (Schneider et al., 2013), a large, flexible and clinically important target (Philip et al., 2018), for which experimental residence time of the inhibitors ranges from minutes to hours. The compounds were correctly ranked according to the computed residence time score in agreement with their experimental residence time. The

analysis of protein-ligand interactions along the dissociation trajectories highlighted the favorable contribution of hydrophobic contacts to residence time and revealed key residues that strongly affect compound residence time.

Keywords: Binding kinetics · Residence time · Molecular dynamics simulation · Drug design · Structure-kinetics relationship · Cyclin-dependent kinase 8.

INTRODUCTION

Despite technological and methodological progress, attrition rates remain high in drug discovery and development program. A large proportion of drug candidates fails in the late phase of clinical trials due to a lack of efficacy (Arrowsmith and Miller, 2013), while those compounds appeared promising in the early stages of the drug discovery process. The efficacy is the maximum response that a drug can produce. In order for a drug to have an effect, it needs to bind to its target. Therefore, increasing the target occupancy will increase the efficacy of the drug. Understanding and considering the parameters that can influence the target occupancy can help improving the efficacy and so reducing the attrition rate. In the complexity of dynamic biological system, target occupancy is influenced by many processes including the time course of the drug concentration (pharmacokinetics), the affinity, the rates of synthesis and degradation of the target molecule (target turnover), the concentrations of endogenous ligands competing for the same target and the drug-target binding kinetics (de Witte et al., 2016b). This last ten years, drug-target binding kinetics is increasingly considered as an important selection criterion in drug discovery, in addition to the traditional focus on drug target-binding affinity (Bernetti et al., 2017). The binding kinetics of a drug on its protein target is characterized by the association rate constant (k_{on}), which is the rate the drug binds, and the dissociation rate constant (k_{off}), which is the rate of unbinding. High k_{on} can compensate a suboptimal pharmacokinetics context, for example a poor bioavailability, and can be the driving force to get the drug into the binding site. A low k_{off}, that is a high target residence time (residence time (RT) = $1/k_{off}$), increases the duration of action of the drug. More globally, optimizing the pharmacological response through binding kinetics opens the way towards a control of several drug properties, among them the *in vivo* efficacy, the duration of action, the selectivity, and the safety (Copeland, 2010).

Since Swinney *et al.* have noted that among FDA-approved drug candidates there is an enrichment of “non-equilibrium” drug (Swinney, 2004), the residence time has emerged as an important criteria to evaluate the *in vivo* efficacy in the early phases of drug discovery (Schuetz et al., 2017). They define a “non-equilibrium” drug as a drug with a molecular mechanism of action (MMoA) that prevents the competition between the drug and the endogenous ligand from reaching equilibrium. Indeed, such equilibrium competition reduces the pharmacological response, and so increases the amount of drug required to obtain the same degree of activity in the presence of the endogenous ligand. There are numerous ways

to prevent an equilibrium competition and an intuitive way is the irreversible inhibition involving covalent modification of the drug target, which is a type of a "non-equilibrium" MMoA. Another "non-equilibrium" MMoA is the insurmountable reversible inhibition favored by a slow dissociation i.e., an increased target residence time. The positive impact of increasing the residence time on the *in vivo* efficacy has been demonstrated on G-protein-coupled receptors (GPCRs) such as the A2A adenosine receptor (Guo et al., 2012) and recently, the metabotropic glutamate receptor 2 (mGluR2) (Doornbos et al., 2017) and the Histamine H1 Receptor (H_1R) (Bosma et al., 2017). In the kinase family, the slow dissociation of the dual tyrosine kinase inhibitor lapatinib correlates with a prolonged down-regulation of receptor tyrosine phosphorylation in tumor cells (Wood et al., 2004). Compound 584, an analogue of imatinib with a slow dissociation, inhibits Abl kinase with lasting effects (Puttini et al., 2008). Recently, target residence time-guided optimization on TTK kinase results in inhibitors with potent anti-proliferative activity (Uitdehaag et al., 2017). Studies on other therapeutic target proteins show a correlation of the residence time with *in vivo* efficacy (Costa et al., 2016; Lee et al., 2014; Ramos et al., 2018). Many commercialized drugs have evolved to provide insurmountable inhibition (Swinney, 2004) as Candesartan (RT=112 min) that is more effective than Losartan (RT = 2.5 min) while they bind to the same binding pocket and have similar pharmacokinetics profiles (Fuchs et al., 2000).

Despite the great improvements in experimental methods and the large available panel assay, measurements of binding kinetics are operatively limited due to the need of time-resolved data collections, high throughput binding kinetics assays and the cost of these methods (de Witte et al., 2016a). Moreover, these experimental methods do not allow the correlation of kinetics data with structural interactions, that is the description of the full (un)binding process at the atomic level including the high-energy transition states and the stability of ground (or metastable) states. Though, such information would be of great importance to help chemists in the design and synthesis of compounds with optimized kinetics parameters.

In this context, molecular dynamics (MD) simulation is an interesting method to study binding kinetics owing to the supplied comprehensive structural view at the atomic level of the (un)binding process. Despite recent advances in computer technology that have increased the speed of brute-force MD, the simulation of ligand unbinding with solely brute-force MD has only been achieved on small system (FK506 binding protein, FKBP) with millimolar drug fragments (weak binders) and residence times in the 10^{-9} s timescale (ranged from 8ns to

140ns) (Huang and Caflisch, 2011; Pan et al., 2017). Various theories and algorithms using brute-force all-atom MD simulations (in combination or not with Brownian dynamics) in explicit solvent on microsecond scale have been developed to predict the k_{off}. These methods have been applied on a small target (trypsin-benzamidine) with a relatively rigid binding site and small molecules with fast association (k_{on} >10^7 $M^{-1}s^{-1}$) and fast dissociation (k_{off} >10^2 s^{-1}) rates. These methods include the software SEEKR (Votapka et al., 2017) and the WExplore tool (Dickson and Lotz, 2017) giving a predicted value of the k_{off} within a factor of 10, the adaptive multistate splitting (AMS) yielding to a k_{off} value about 2-fold higher than experiment (Teo et al., 2016) and the markov state models (MSMs) with results deviating from the experimental reference values by a factor of 10 (Buch et al., 2011), and a factor of 20 (Plattner and Noé, 2015). However, predominant therapeutic targets in drug discovery are larger and more flexible proteins, such as kinases or membrane proteins (GPCR), and often bind large and flexible drugs. Recently, WExplore tool has been applied on a clinically relevant system, the epoxide hydrolase bound to the inhibitor TPPU, with a pharmacologically relevant residence time scale (11min) (Lotz and Dickson, 2018). However, despite progress in computational power, those brute force MD based methods remain too computationally expensive regarding the timescale from millisecond to hours needed to simulate many clinically relevant ligand unbinding processes and hence, unsuitable for a routine industrial use where series of compounds must be analyzed during the hit-to-lead and the lead optimization stages. With this rising need in mind, several protocols using biased sampling methods have been developed and applied to compute binding kinetics.

Mollica *et al.* applied scaled-MD, a simple ranking method that does not require the definition of a reaction coordinate, on several ligands of HSP90, of Glucose-Regulated Protein (Grp78), of adenosine A2A receptor (A2A) (Mollica et al., 2015) and of glucokinase (Mollica et al., 2016), and obtained a correct k_{off} ranking in all cases. However, a set of restraints is applied on all the backbone heavy atoms with the exception of those in the binding site, to keep the protein in its native conformation. These restraints can lead to an unrealistic description of the simulated unbinding process if the dissociation of the ligand from the binding site involves for example a conformational change. In another study, random acceleration molecular dynamics (RAMD) in combination with steered molecular dynamics (SMD) were used to explore ligand (un)binding pathways and to generate potentials of mean force respectively (Niu et al., 2016). The difference in transition state barrier calculated from the potentials of mean force is in qualitative agreement with the measured difference in binding

kinetics. Very recently, an RAMD based method called τRAMD, was applied on 70 diverse drug-like ligands of HSP90α protein and demonstrated a good correlation (R^2 = 0.86) between computed and measured residence time for 78% of the compounds (Kokh et al., 2018). Several approaches derived from metadynamics (MTD) have been developed for the prediction of binding kinetics constants (Bortolato et al., 2015; Callegari et al., 2017; Casasnovas et al., 2017; Pietrucci et al., 2009; Sun et al., 2017; Tiwary et al., 2015, 2017). Among them, through microseconde time scale simulations, results in good agreement with experiment were obtained for the k_{off} calculation of an urea-based allosteric inhibitor of p38 MAP kinase (Casasnovas et al., 2017) and the dasatinib of c-Src kinase (Tiwary et al., 2017). For this MTD approach to work, the underlying energy landscape of the unbinding process is assumed to display few high and sharp barriers (Tiwary et al., 2017). Using this approach, Callegari *et al*. were not able to accurately rank a set of cyclin-dependent kinase 8 (CDK8) inhibitors by their k_{off} in agreement with experiment (Callegari et al., 2017). Therefore, they proposed an alternative MTD approach leading to a ranking in agreement with experiment, in which the relative k_{off} values were estimated from the simulation time required for driving the ligands to the point of dissociation using simulations of dozen nanoseconds. However, in that study, the cyclin C, which is complexed to the kinase and available in the experimental structures (Schneider et al., 2013), was not kept in the simulations. Our work on the importance of the cyclin C in the complex CDK8-CycC (data not shown)[3] and a very recently published study (Cholko et al., 2018), showed that cyclin C is vital for maintaining the structure of CDK8 and providing proper interactions for ligand binding.

In this study, we developed a protocol involving an ensemble of targeted molecular dynamics simulations to allow the ranking of a series of congeneric compounds by their residence time. With a relatively low computational cost, this method is suitable for an industrial use, to analyze chemical series during hit-to-lead or lead optimization stages. The protocol does not require *a priori* knowledge or hypothesis on the exit pathway and applies sufficiently slow variation to provide a realistic description of the unbinding process. The method was validated on a set of arylpyrazole inhibitors of CDK8 (Schneider et al., 2013), a clinically relevant target (Philip et al., 2018), for which experimental residence time of the inhibitors ranges from minutes to hours. We obtained a good ranking of the compounds according to

[3] Cette étude est l'objet du chapitre II.

their computed residence time score in agreement with their experimental residence time. Moreover, we carried out a Structure-Kinetics Relationship (SKR) study to depict the molecular determinants of protein-ligand interaction responsible of a slow dissociation. We hope that it will help optimizing kinetics profile of CDK8-CycC inhibitors.

MATERIEL AND METHODS

Data

We used the data published by Schneider *et al.* (Schneider et al., 2013), containing a set of 10 arylpyrazole inhibitors of CDK8 complexed with the cyclin C (CDK8-CycC). Experimental residence times of these inhibitors are provided and for some of them, the crystallographic structure of CDK8-CycC complexed with the inhibitor is also available. We classify the inhibitors into three groups according to their residence time: Short Residence Time (SRT), Medium Residence Time (MRT) and Long Residence Time (LRT) (**Figure 1**). These compounds display a common pyrazol-5-yl urea scaffold. The SRT inhibitors 1-5 (RT < 1.4 min) comprise an inhibitor without a pendant chain (1), an inhibitor with a methyl derivative (2) and 3 other inhibitors with hydrophilic substituents (hydroxyethyl for 3, morpholinoethyl for 4 and hydroxypropyl for 5). The MRT group contains 3 inhibitors 6, 7, 8 which carry an hydroxybutyl, a morpholinopropyl and a hydroxypentyl chain with residence times of 7, 14 and 57 min, respectively. The third group includes the LRT inhibitors 9 and 10, which have a 1- (2-(4-(3-tert-butyl-1-p-tolyl-1H-pyrazol-5-ylcarbamoyl)- piperazin-1-yl)ethyl) or tert-butoxycarbonylaminopropyl chain and show extremely long residence times of 1626 and 1944 min, respectively.

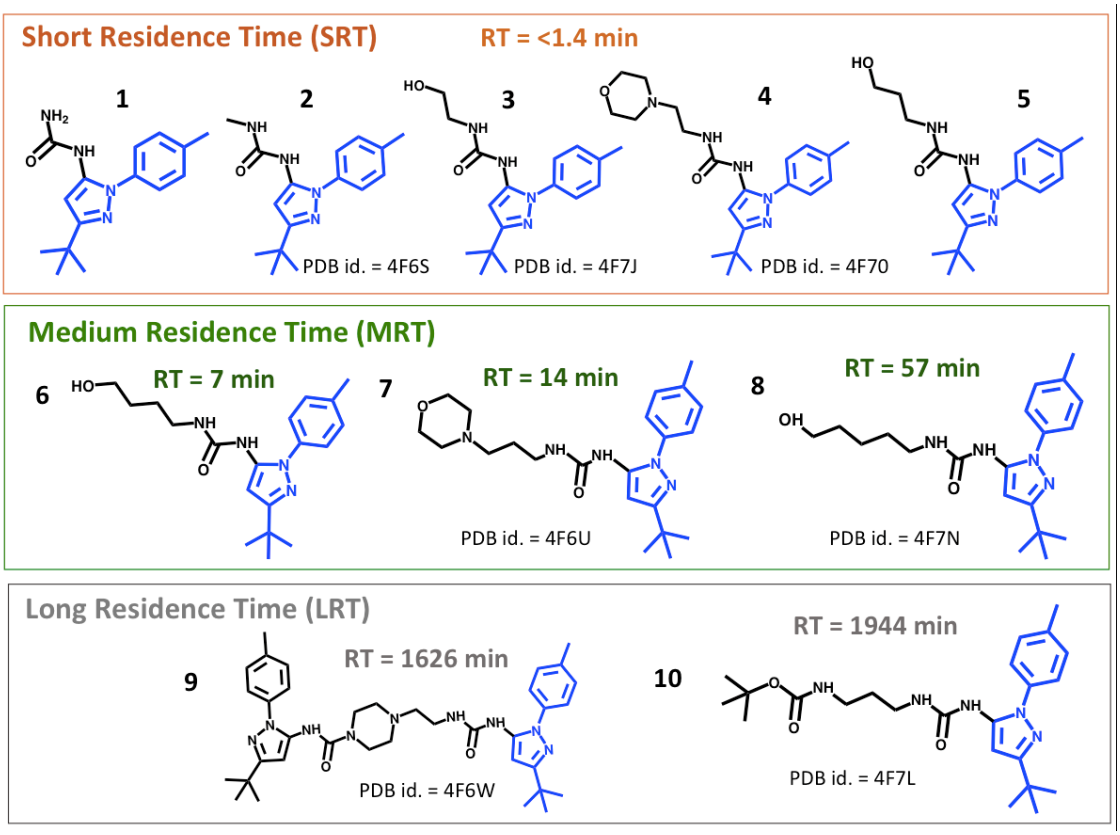


Figure 1: Chemical structures of the ten CDK8 inhibitors, numbered from 1 to 10 (inhibitor id.) and their experimental residence times classified in three groups: SRT MRT and LRT inhibitors.
The common 1-(3- tert-butyl-1-p-tolyl-1H-pyrazol-5-yl)urea scaffold is highlighted in blue.

The common 1-(3-tert-butyl-1-p-tolyl-1H-pyrazol-5-yl)urea scaffold of these compounds is anchored in the kinase allosteric pocket (also called hydrophobic pocket) and interacts with the conserved DMG motif through a hydrogen bond (HB) interaction with the backbone nitrogen atom of Asp173 and with the carboxylate group of Glu66 through two HB interactions. The scaffold extends with variable functional groups toward the hinge region or the front pocket (**Figure 2**).

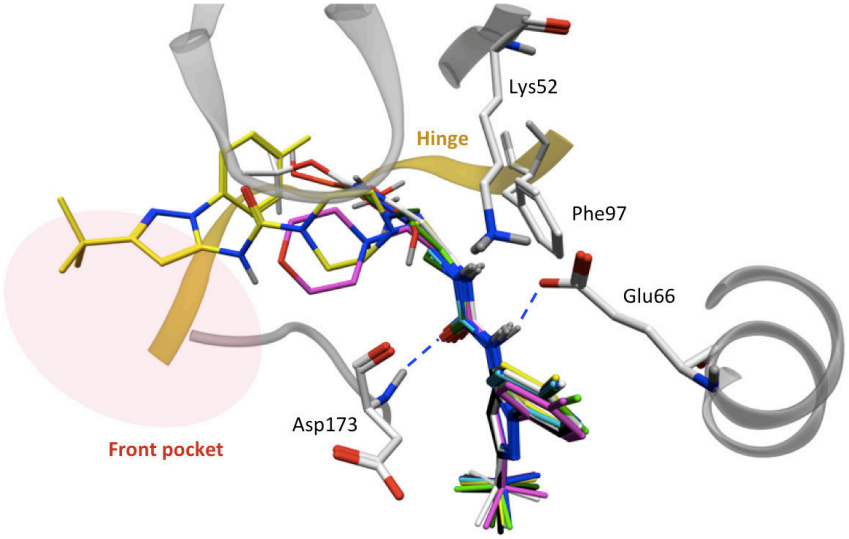


Figure 2: Binding site of the crystal structure of human CDK8 in complex with inhibitors 1-10.
Interactions between the urea of the common scaffold of the inhibitors and the residues Glu66 and Asp173 are shown in blue dashed lines. The protein is represented in white cartoon except the hinge colored in yellow cartoon. A red circle represents the location of the front pocket. Inhibitors 1 to 10 are colored respectively in gray, black, cyan, light green, pink, purple, dark green, blue, yellow and white.

Model Building

As in Mollica *et al.* (Mollica et al., 2016), we make the choice to construct a unique model of CDK8-CycC protein–ligand complex by homology modeling. The other protein-ligand systems were obtained by replacing the inhibitors in that model (chemical replacement), in such a way to keep the same protocol even with inhibitors with no crystallographic structure. Chemical replacement was considered sufficient because the inhibitors of this congeneric series that have a crystallographic structure consistently display a conserved orientation within the binding site (**Figure 2**). So the first step consists in choosing the crystallographic structure from which the homology model will be constructed. All the available experimental structures (Schneider et al., 2013) present 3 missing loops: the activation loop following the key DMG-motif (residues 177 to 193 in the structure of PDB id. 4F6U) which is in *DMG-out* conformation in all structures and the loops from residues 116 to 120 and residues 240 to 244. After checking the absence of mutations in its sequence, the structure of PDB id. 4F6U was chosen as target structure to construct the model since it presents the best resolution (R = 2.1 Å) among the available ones. We have then aligned the UniProt canonical sequence of CDK8 on the PDB database to retrieve the most homologous template structures having the

missing regions resolved and the activation loop in the *out* conformation. Two crystallographic structures of the human homologous CDK6, (PDB id 1BI8 and 1G3N) were retained and used as template structures. The sequence alignment was performed with Clustal Omega (Sievers et al., 2011) with a particular care on the alignment of domain kinase conserved motifs. CDK6 shares 37 % of identity and 63 % of similarity with CDK8 (**Figure S1** & **S2**). Only the missing regions in the target structure were rebuilt in order to keep the coordinates of the resolved parts of the protein unchanged. After adding the sequence of the cyclin C, the crystallographic molecules of water, and the ligand (compound 7), the alignment file was imported into MODELLER version 9.16 (Sali and Blundell, 1993) to generate the model. We thus obtain a model of CDK8 (residues 1 to 359) complexed to cyclin C (residues 1 to 264) and to compound 7. The missing C-terminal segments of CDK8 (residues 360 to 464) and of cyclin C protein (residues 265 to 283) were not reconstructed. The complete model was subjected to structural validation through PROCHECK (Laskowski et al., 1993) and ProSA-web tools (Wiederstein and Sippl, 2007) (**Figure S3**). The complexes CDK8CycC-inhibitor for compounds 2, 3, 4, 7-10 (for which a crystallographic structure is available) were obtained by, first, aligning the crystallographic structure to the model, and then, by placing the ligand and the crystallographic molecules of water inside. We manually rotated some residues, when necessary, to be in agreement with protein-ligand interactions observed in the respective crystallographic structures. The complexes CDK8CycC-inhibitor of compounds 1, 5, 6, were generated from previous reconstructed models, by manually modifying the chemical structure of the most similar analogues. These manipulations were done with the Molecular Operating Environment (MOE) software version 2016.0802 from the Chemical Computing Group. We ensure that there is no steric hindrance.

System Preparation

In total, 10 systems (**Figure 1**) were prepared. The AmberTools 15 suite (Case et al., 2015) was employed to protonate, solvate, neutralize and generate the topology and coordinate files of the systems. Ligands were prepared by using the Antechamber tool and the GAFF force field after adding hydrogen atoms with the reduce utility (Wang et al., 2004, 2006). All compounds were modeled in their neutral state. Further analysis was carried out for the protonation state of compound 7 (**Figure S4**), since the pKa of alkylmorpholines is about 7.4 (Hall, 1956). The morpholine of compound 7 was finally modeled in its unprotonated state since the interaction with Ala100 is not observed any more with the protonated morpholine

(**Figure S4**). Partial charges on the ligands were generated with the AM1/BCC method (Jakalian et al., 2002). PROPKA version 3.0 (Olsson et al., 2011) was used to check the protonation state of ionizable residue side-chains at pH = 7. The protein force field ff14SB parameters were assigned (Maier et al., 2015). After being solvated with a rectangular TIP3P water box, the edge of the box is at least 10 Å away from any solute atom. Finally, Cl^- ions were added to neutralize the positively charged systems for a total number of atoms around 110 000 atoms.

Simulation protocols

For each of the 10 systems, 11 replicas were launched. A same ligand can form slightly different interactions within the protein binding site. Accordingly, in order to better reflect this reality, we have prepared each replica with a new cycle of minimization and equilibration, and indeed, slightly different protein-ligand interaction network can be observed among replicas. A four-cycle minimization was performed with 2000 steps each cycle, minimizing first the solvent, second the residue side-chains, then the solute and finally the entire system. The SHAKE algorithm was applied to constrain bonds involving hydrogen atoms by using a time increment of 2 fs. Temperature regulation at 300K was ensured through Langevin dynamics with a collision frequency of 2 ps^{-1}. The long-range electrostatic interactions were computed by the particle mesh Ewald method beyond 10Å distance. The system was slowly heated in NVT ensemble from 0 to 300 K over a period of 50 ps, where a harmonic restraint on the solute (20 $kcal.mol^{-1}.Å^{-2}$ force-field constant) prevents the system from structural distortion. The system was then equilibrated during 10 ns MD simulation in the NPT ensemble at 300K and 1 atm, through which the harmonic restraint is gradually decreased from 20 $kcal.mol^{-1}.Å^{-2}$ to 3 $kcal.mol^{-1}.Å^{-2}$ in 1.3 ns and then, totally relaxed in 8.7 ns. The pressure relaxation time was set to 1 ps. Brute force MD calculations were performed using the PMEMD.cuda module of the AMBER14 program (Case et al., 2015).

Targeted molecular dynamics (TMD)

The TMD is a simulation technique for that aims to determine the pathway of a conformational transition between two states: (un)bound, (un)folded, open-close conformation etc. (Schlitter et al., 1994). It consists in constraining the root mean square deviation (RMSD) between the current structure (which is the starting structure at the

beginning of the simulation) and a reference structure ($RMSD_{current}$) to a user-defined value, namely the $RMSD_{target}$. This value of $RMSD_{target}$ is slowly varied from an initial value to a targeted final value ($RMSD_{target\_final}$), which results in the simulation of the process leading to the final desired state. In AMBER14 program, a harmonic restraining potential ($V_{restraint}$) is added to the force field, to help the $RMSD_{current}$ reaching the successive values of $RMSD_{target}$ until the final value ($RMSD_{target\_final}$).

$$V_{restraint} = \frac{1}{2} \times f \times N_{atoms} \times \left(RMSD_{current} - RMSD_{target}\right)^2$$

Equation 1

Where f is the harmonic force constant, N_{atoms} is the number of restrained atoms, that is, the number of atoms on which the RMSD is calculated. Note that the atomic coordinates are mass weighted in the calculation of RMSD. It exists two approaches of TMD: direct TMD and reverse TMD (TMD^{-1}). In direct TMD, the reference structure corresponds to the final targeted structure, so that the value of $RMSD_{target}$ is decreased from the RMSD between the initial and target structure to a value close to 0. In TMD^{-1}, the reference structure corresponds to the initial structure, so that the RMSD value is increased from 0 to a predefined value. Therefore, TMD^{-1} is less constraining than direct TMD, since no information on the final desired state, and so no indication of the search direction is given.

In this study, we apply TMD^{-1} using the equilibrated CDK8CycC-inhibitor complex as reference structure. Therefore, as the $RMSD_{target}$ increases, the ligand moves away from the binding site, aided by the addition of the restraining potential ($V_{restraint}$). The more difficult the sampling at the considered $RMSD_{target}$ is, the higher the energy ($V_{restraint}$) added. The RMSD is calculated on the heavy atoms of the ligand, after aligning the reference and the current structure on the backbone of a set of 22 residues of the binding site. These 22 residues are at a distance of 4 Å from the center of mass of the binding site (**Section S2**). Since the ligands of our series do not have the same number of heavy atoms, N_{atoms} varies and so the spring constant $f \times N_{atoms}$ varies also. To ensure comparability between the results of the diverse ligands, the spring constant $f \times N_{atoms}$ was set constant by adapting the value of f, and after several tests, the value of $f \times N_{atoms}$ was fixed to 80 $kcal.mol^{-1}$. The $RMSD_{target}$ is changed by step of 0.01 Å every 0.2 ps from the value of 0.001 to 75.001 Å during a total simulation time of 1500 ps. Hence, at the beginning of the simulation, the ligand is in the binding site and

is asked to sample at a $RMSD_{target}$ value of 0.001 Å, namely the bound state. After 0.2 ps of sampling at this value, $RMSD_{target}$ increases to a value of 0.011 Å and the ligand is again asked to sample at this value during 0.2 ps etc., until it exits from the protein (the unbound state is defined in the Results and Discussion section). A snapshot was saved every 0.2 ps. TMD runs were performed with the parallelized version of the SANDER module from the AMBER14 program.

Data analysis

The simulations were analyzed using VDM (Humphrey et al., 1996), the CPPTRAJ module from AMBER14 program (Case et al., 2015) and the Structure Interaction Diagram (SID) module of the maestro suite (Maestro, Schrödinger, LLC, New York, NY, 2016). The SID module was used to calculate the protein-ligand interactions after converting the trajectories to the maestro format (see **Section S6**). For each simulation, we analyzed the protein-ligand interactions of a total of 150 conformations (one snapshot every 10 ps). All the data extracted from the simulations were processed and analyzed using the R package version 3.4.1 (R Core Team, 2017).

RESULTS AND DISCUSSION

Residence time (RT) and RT_{score}

The residence time is directly related to the height of the energy barrier(s) between thermodynamically stable states crossed by a drug, to move from the bound to the unbound state, as described by the Arrhenius law (**section S4**). Ligands displaying long residence time are assumed to cross higher or multiple energy barriers and so need more energy to be expelled from their binding sites. On this basis, the restraining potential added during a TMD^{-1} simulation to cross energy barriers and reach the unbound state could be considered as a relevant estimator of the residence time of a drug. However, in view of the nature of the bias (harmonic restraining potential) which forces the system to sample every micro-state along the pathway, the increase of the restraining potential is not only associated to the action of "pushing" the ligand to help it crossing high energy barriers, but also to the action of "retaining" the ligand, when the system is on an energetic descent approaching a metastable state. In that connection, we derived a new function $V_{restraint\_push}$ from $V_{restraint}$ than only

encompasses the restraining potential added to cross energy barriers and so, traduces the difficulty encountered by the ligand to escape the binding site. $V_{restraint\_push}$ is defined as:

$$V_{restraint\_push} = V_{restraint} \times f(RMSD_{current}, RMSD_{target})$$

Equation 2

- Where $f(RMSD_{current}, RMSD_{target}) = 1$, when $RMSD_{current} - RMSD_{target} < 0$
- And $f(RMSD_{current}, RMSD_{target}) = 0$, when $RMSD_{current} - RMSD_{target} > 0$

Accordingly, when the ligand has to overcome an energetic peak, the value of $RMSD_{current}$ becomes inferior to the value of the wanted $RMSD_{target}$ due to the difficulty to move forward and in this case $V_{restraint\_push} = V_{restraint}$. On the contrary, when the ligand is about to reach an energetic minimum, the ligand advances faster, so the value of $RMSD_{current}$ becomes significantly superior to the value of the wanted $RMSD_{target}$. In this case, the increase in $V_{restraint}$ is not considered, so $V_{restraint\_push} = 0$. Integrating $V_{restraint\_push}$ over time leads to a quantity (noted RT_{score}) comparable to an action in physics expressed in [energy].[time] which reflects here the overall "action" supplied to pull the ligand out of the binding site.

$$RT_{score} = \int_{t0}^{texit\_time} V_{restraint\_push} = \frac{1}{2} \times f \times N_{atoms} \int_{t0}^{texit\_time} (RMSD_{current} - RMSD_{target})^2$$

Equation 3

The integral is calculated from the beginning of the simulation (t = 0 ns) to the exit time (t_{exit_time}) defined as the time at which the ligand is unbound. A geometrical criteria was used to identify the unbound state that is, any atoms of the ligand are at least at 6 Å of any atoms of the protein (here the complex CDK8-CycC). We assume that ligands having to cross higher or multiple energy barriers need more energy ($V_{restraint\_push}$) from the TMD^{-1} protocol to be expelled from the binding site, and so a higher value of RT_{score}, is expected. Therefore, RT_{score} is assumed to be positively correlated to the RT (**Section S4**). As a ligand could exit through different unbinding paths during TMD^{-1} simulations, the values of RT_{score} were collected and averaged from replicated simulations. In **Table 1**, the average RT_{score} values for compounds 1 to 10 and their standard errors are reported. The **Figure 3** shows that the method is able to correctly rank the SRT, MRT and LRT inhibitors according to their residence time on the basis of RT_{score}. The method clearly distinguishes LRT inhibitors from SRT and MRT inhibitors while MRT inhibitors are less well separated from SRT inhibitors in light of the error bars.

Inhibitor id.	RT (minute)	RT_{score} (kcal.mol^{-1}.ps)
1	<1.4	1048 ± 66
2	<1.4	1562 ± 243
3	<1.4	1560 ± 206
4	<1.4	1610 ± 217
5	<1.4	1580 ± 205
6	7	1991 ± 298
7	14	1933 ± 245
8	57	2014 ± 428
9	1626	4494 ± 419
10	1944	3990 ± 289

Table 9: Experimental Residence Times (RT) and values of the estimator of the residence time (RT_{score}) calculated from the simulations for CDK8 inhibitors 1-10.
RT_{score} values correspond to the means calculated on the 11 replicas, with their standard errors.

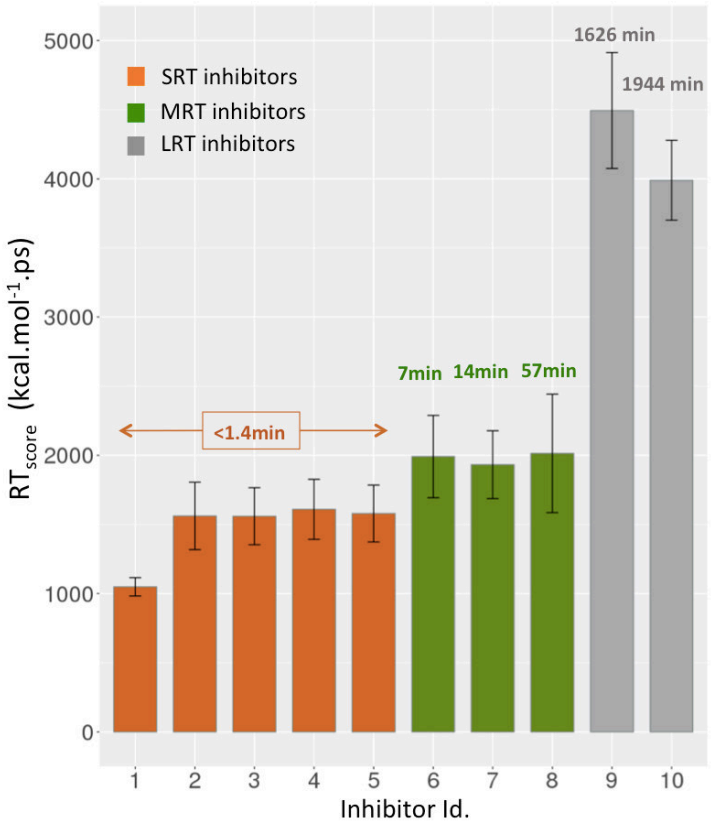


Figure 3: Estimator of the residence time (RT_{score}) calculated from the simulations for CDK8 inhibitors 1-10.
The bars are colored according to experimental residence time groups (SRT, MRT and LRT). The error bars are based on the standard error.

The difficulty in predicting residence time is to set up a method that is applicable to both a simple 1-step kinetics process and a complex process involving several kinetics steps. Despite being commonly found in textbooks, the static lock-and-key model in which the binding occurs in a 1-step kinetics is seldom adequate to describe the association and dissociation of a drug with its target. Indeed, both the association of a drug to its target, and the subsequent dissociation of the drug–target complex, are often controlled by conformational changes, especially involving structural changes in the immediate vicinity of the drug-binding pocket (Copeland, 2011). However, the transition state theory that relates the activation energy to the kinetics constant (k_{off}, k_{on}) is based on the Arrhenius equation that describes on a 1-step kinetics process. In a complex multi-steps kinetics process, given the fact that energy barriers can be multiple, more or less important and depend on the degree of resolution of the energy profile (free energy, restraint energy or other) calculated by numerical methods, determining the k_{off} is complex and challenging. Bortolato and co-workers proposed an approach derived from MTD simulation that only takes into account the first barrier of the bias potential energy. They calculate a RT score defined as the maximum bias potential energy required to move the ligand from the starting energy basin (bound state) to the next (Bortolato et al., 2015). Another approach (Tiwary's approach), also derived from MTD simulation, assumes that the energetic landscape underlying the dissociation process presents few high and sharp barriers resulting thus, in a global Arrhenius behavior, ie a 1-step kinetics process (Tiwary and Parrinello, 2013; Tiwary et al., 2017). Sun *et al.* tested Bortolato's and Tiwary's approach and were not able to accurately predict the k_{off} of a complex dissociation process involving EGFR kinase protein that presents a deep binding site (Sun et al., 2017). Callegari *et al.* applied Tiwary's approach and failed also in predicting the k_{off}. They concluded that the free energy landscape of the CDK8– arylpyrazole inhibitor unbinding process is too complex (Callegari et al., 2017). The advantage of RT_{score} is that it encompasses all the energy barriers encountered during the dissociation process, which makes the method applicable on complex multistep kinetics processes (Equation 3).

Analysis of the Structure-Kinetics Relationship (SKR)

Pathways

Visual inspection of the unbinding trajectories revealed three pathways taken by the 10 inhibitors: the allosteric channel when the ligand exits through the allosteric pocket (also

called hydrophobic pocket), the front channel (also called the ATP channel) when it passes through the front pocket and, the one we call the hinge channel, when it passes under the hinge (**Figure 4**).

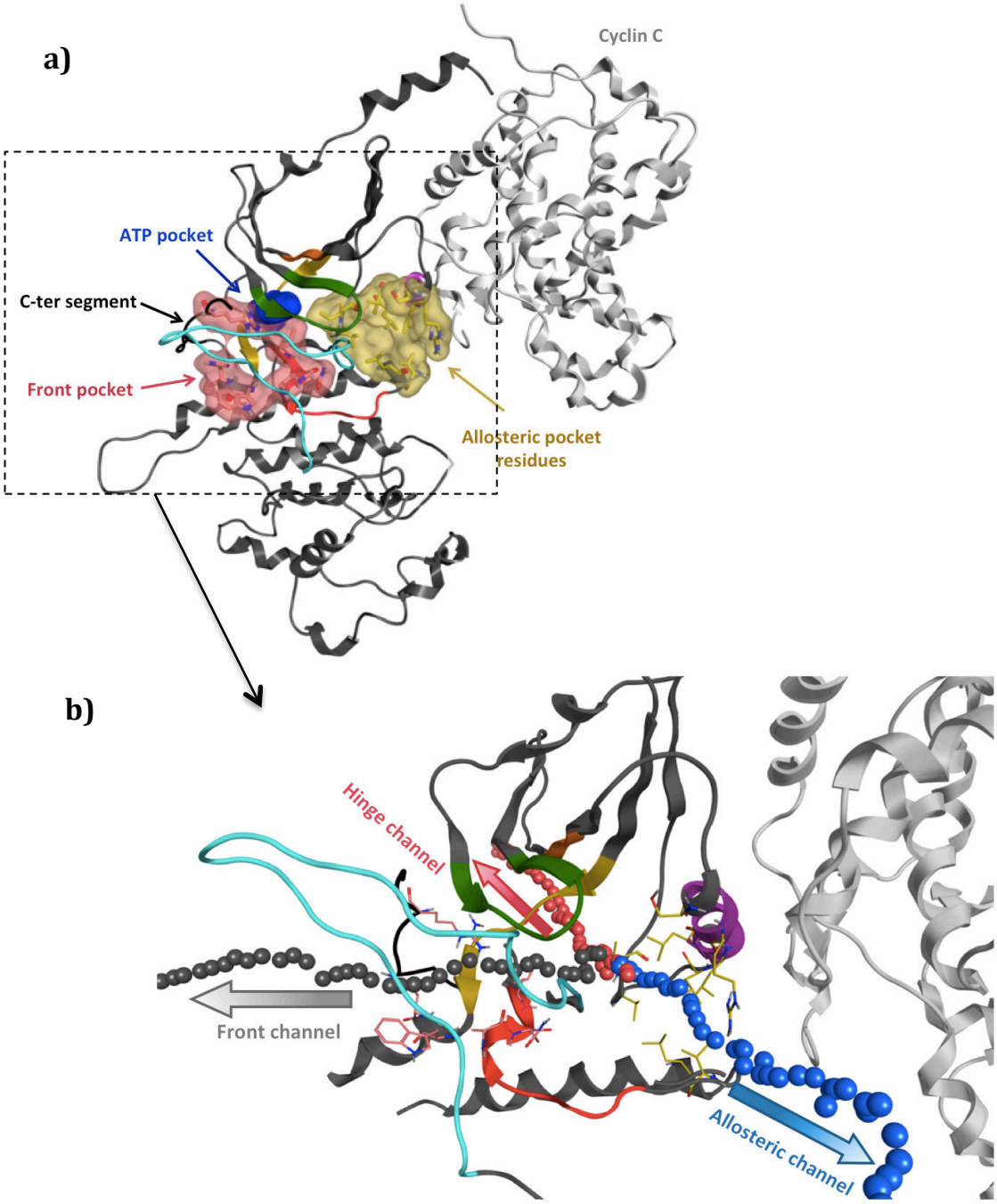


Figure 4: Exit pathways taken by the inhibitors.
(a) Representation of CDK8-CycC in cartoons. The cyclin C is in light gray. The kinase domain is in dark gray except the kinase conserved motifs, the C-ter segment (black) and the activation loop (cyan). The conserved motifs of the kinase domain are colored as follows: α-C helix (purple), hinge (yellow), P-loop (dark green), Hyd1 (orange), HRD (red). The activation loop (containing the DMG motif) is coloured in cyan. The allosteric pocket residues, the ATP pocket and the front pocket residues are displayed in surface coloured in yellow, dark blue and pink, respectively. **(b)** Zoom of picture (a) where an exemple of each pathway was represented by balls corresponding to the centers of mass of the ligand along the path. The allosteric, front and hinge channels were represented in blue, black and red, respectively. The residues of the allosteric pocket and the front procket are represented in sticks with their carbons colored in yellow and pink, respectively.

LRT inhibitors take mainly the front channel whereas MRT and SRT inhibitors follow mostly the allosteric channel (**Figure 5**). The allosteric and the front channel have already been observed in several computational studies as possible routes in kinase family (Niu et al., 2016; Patel et al., 2014; Sun et al., 2015). A type II inhibitor of p38 MAP kinase called BIRB796, having the same scaffold (1-(3-tert-butyl-1-p-tolyl-1H-pyrazol-5-yl)urea) as our series and the same binding mode as MRT inhibitors, has also been shown to exit preferentially through the allosteric channel (Sun et al., 2015), which is in agreement with our results. Inhibitors exiting through the allosteric channel go either toward the cyclin C partner, interacting with its surface in some cases, or go towards the P-loop or the C-lobe of the kinase. For the front channel, the routes also diversify when approaching the solvent, which can be attributed to the movements of the long flexible C-terminal segment and of the activation loop that directly impact the ease of access to the solvent (**Figure 4**).

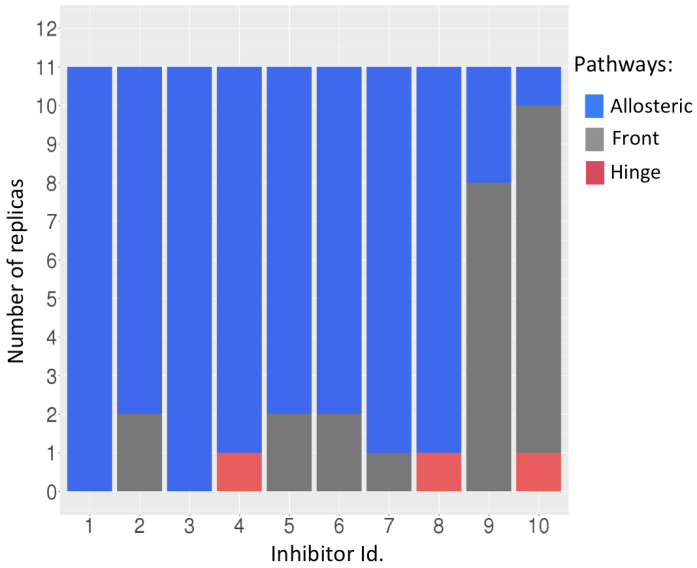


Figure 5: Barplot counting the number of each pathway type taken by each inhibitor.

We then analyzed protein-ligand interactions along the unbinding process to depict the structure-kinetics relationship. Schneider *et al.* have described the differences in protein-ligand interactions between the inhibitors on the basis of their static crystallographic structures and related them to their residence times. They suggest that hydrogen bonds with the hinge region are indispensable to provide a detectable residence time to compounds leading to MRT inhibitors, whereas large hydrophobic complementarities within the front pocket (**Figure 6**) significantly optimize the compound residence time leading to LRT inhibitors (Schneider et al., 2013).

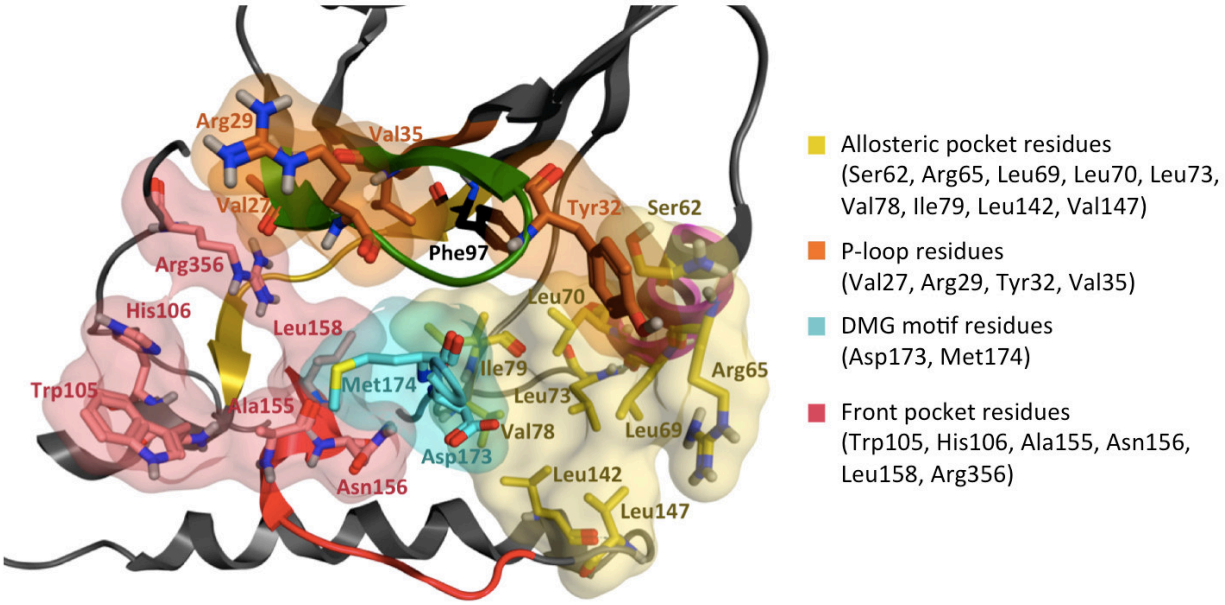


Figure 6: Representation of the regions of the binding site containing some residues that influence the residence time. (Nomenclature adapted from (Schneider et al., 2013)). The kinase domain is represented in ribbon coloured in dark gray except the kinase conserved motifs and the C-ter segment (black).The conserved motifs of the kinase domain are coloured as follows: α-C helix (purple), hinge (yellow), P-loop (dark green), Hyd1 (orange), HRD (red), and DMG motif (cyan). The allosteric pocket residues, the P-loop residues, the DMG motif residues and the front pocket residues are displayed in surface respectively in yellow, orange, cyan and pink. They are also represented in stick where the carbon atoms follow the same color code. The residue Phe97, called gatekeeper residue, is represented in stick, with the carbon atom coloured in black.

The MD brings the dynamical view of the phenomenon and along with several replicas, provides a sampling of the unbinding event leading to a relevant observation. Therefore, in the following section, we analyzed protein-ligand interactions in the bound state and also, along the dissociation path by comparing replicas of a same inhibitor and the inhibitors with each other. The goal is to discuss and hypothesize on possible impact of some protein-ligand interactions on the residence time.

SRT inhibitors

For the SRT inhibitors (1 - 5), replicas that establish hydrophobic interactions between the trimethyl group of the scaffold and the residues of the allosteric pocket (**Figure 6**), in particular Leu70, Leu73, Val78, Ile79, Leu142 and Val147, present higher RT_{score} values. In addition to these interactions, inhibitors 2 to 5 differ from the inhibitor 1 by their ability to establish hydrophobic interactions between the Phe97, known as the "gatekeeper" residue, and the hydrophobic part of their variable fragment (the methyl group for inhibitor 2, and the alkyl chain for the inhibitors 3 to 5). The variable fragment, which is the most flexible part of

the inhibitors, is stabilized by the hydrophobic interactions with the gatekeeper Phe97. As a result, the inhibitors are maintained close to the hinge region, deep in the binding site. In this location, hydrophobic interactions between the allosteric pocket residues on one hand, and the trimethyl group of the scaffold as well as the hydrophobic segment of the variable fragment on the other hand, are promoted. Replicas of SRT inhibitors 2 to 5 displaying the highest RT_{score} values always form such hydrophobic interactions involving the gatekeeper Phe97 and the allosteric pocket residues. Consequently, it results in a lower number of hydrophobic contacts for the inhibitor 1, that does not have a variable fragment, compared to other inhibitors (**Figure 7, Figure S7**). This could be a reason explaining the gap in average RT_{score} values between the inhibitor 1 and the inhibitors 2 to 5 (**Figure 3**). Callegari *et al.* also observed comparable results, that is a lower predicted residence time for inhibitor 1 compared to inhibitors 2 – 5 (Callegari et al., 2017). In the supporting information, we show the positive correlation between the number of hydrophobic contacts involving the allosteric pocket residues and the Phe97, and the value of RT_{score} (**Figure S7**). Inhibitors 2, 3 and 5 do not interact with the hinge: inhibitor 2 is too short for such interaction and inhibitors 3 and 5 orient instead their hydroxyl group toward Met174 or Asp173 or Glu66. In addition to the previously described hydrophobic interactions, we observe, for the replicas with the highest RT_{score} of the inhibitor 4, a water bridge interaction between the oxygen of the morpholine and the Ala100 of the hinge region. However, this interaction is unstable because the inhibitor is not long enough to interact with the hinge region and to maintain, at the same time, the strong HB interactions with Glu66 and Asp173 involving the urea of the scaffold.

MRT inhibitors

From this analysis, we can easily imagine that a longer alkyl chain will stabilize the interaction with the hinge region and lead to an increased residence time. Such inhibitors correspond to the MRT inhibitors where stronger interactions with the hinge region residues are observed without compromising the HB interactions involving the scaffold. Indeed, the hydroxyl group of inhibitor 6 establishes stable water bridge interactions with Ala100 and Asp98, (inhibitor 6 with a $RT_{experimental}$ of 7min) and the hydroxyl group of inhibitor 8 ($RT_{experimental}$ of 57 min) forms a HB interaction with Asp98 and a water bridge with Ala100. The morpholine of inhibitor 7 interacts with the hinge through a HB interaction with Ala100 and a water bridge with Asp98. These interactions with the hinge region lead to higher average RT_{score} values for the MRT, compared to SRT as show in **Figure 3.** In agreement with the results of Schneider *et*

al., the HB interactions with the hinge region are indispensable to detect residence time ($RT_{experimental}$ > 1.4 min). Among the replicas of the MRT inhibitors, when those HB interactions are formed along with hydrophobic contacts involving the allosteric pocket residues and Phe97, higher values of RT_{score} are observed (**Figure S8 and S9**).

Then, we investigated whether there is a relationship between the formed protein-ligand interactions and the taken pathway for SRT and MRT inhibitors, as some of them take the front or the hinge channel instead of the main allosteric channel (**Figure 5**). We found that all replicas taken the front channel display optimized hydrophobic interactions involving the gatekeeper Phe97 and the allosteric pocket residues. It results in higher RT_{score} for those replicas (**Figure S8 and S9**). However these interactions are not exclusive to the front channel since they are also observed in some replicas exiting through the allosteric channel.

LRT inhibitors

Finally, we analyze the interactions made by LRT inhibitors. The binding mode of LRT inhibitors differs from SRT and MRT inhibitors since the variable fragment is oriented toward the front pocket (**Figure 2**) and not the hinge. Schneider *et al.*, stated that large hydrophobic complementarities within the front pocket significantly optimize the residence time leading to LRT inhibitors. These interactions are indeed observed, but they are not the ones with the greatest impact on the RT_{score}. Our analysis revealed that the replicas with the highest RT_{score} tend to establish more interactions with the DMG motif (Asp173, Met174) through HB interactions, and with the P-loop residues (Val27, Arg29, Val35) through hydrophobic and HB interactions at the beginning of the simulation. Those replicas exit through the front channel. On the contrary, the replicas with the lowest RT_{score} tend to present tighter HB interactions with Glu66 (involving the urea of the scaffold) and less contacts with the DMG motif and the P-loop residues at the beginning of the simulation. Those replicas exit through the allosteric channel (**Figure S10**). We did not establish any relationship between the exit through the hinge pathway and the formed protein-ligand interactions.

For the replicas that exit through the front channel, the increased interactions with the P-loop and the DMG motif is accompanied with an increasing number hydrophobic contacts with the front pocket residues. Then, the HB interactions between the urea of the scaffold and Glu66 are broken, which leads to increase hydrophobic contacts with the allosteric pocket residues. In this intermediate state, the two extremities of the inhibitor form hydrophobic contacts with the front pocket and the allosteric pocket, respectively. This intermediate state is not

observed with SRT and MRT inhibitors that exit through the front channel because the compound is not long enough to interact with the front pocket and the allosteric pocket residues at the same time.

In summary, our SKR analysis suggests that hydrophobic contacts with the allosteric pocket residues (Leu70, Leu73, Val78, Ile79, Leu142 and Val147) and the gatekeeper Phe97, in addition to the HB interactions with the hinge residues (Ala100 and Asp98) are indispensable for an inhibitor to show a medium residence time (1.4 min < $RT_{experimental}$ < 57 min), besides the conserved HB interactions involving the urea of scaffold and residues Glu66 and Asp173. For a LRT inhibitor (> 57min), the main positive contributions to residence time are brought by HB and hydrophobic interactions with P-loop residues (Val27, Arg29, Val35) and HB interactions with DMG motif residues (Asp173, Met174) and to a less extent the hydrophobic interactions with the front pocket residues.

Importance of the hydrophobic contacts

From our SKR results, it appears obvious that the hydrophobic interactions strongly contribute to slow the dissociation process. In line with these results, when analyzing the average total number of each contact type (HB, water bridge, ionic, cation-Pi and Pi-Pi stacking) for each inhibitor during the unbinding process (**Figure 7, Figure S11**), it appears that hydrophobic contacts are significantly increased for LRT inhibitors compared to SRT and MRT inhibitors. The strong impact of the hydrophobic contacts has already been discussed in the literature (Liu et al., 2010; Schuetz et al., 2018). In that connection, Schmidtke *et al.*, demonstrated in their study that the formation of water-shielded hydrogen bonds between a ligand and its receptor protein is a viable strategy to increase the residence time. They showed that the kinetics stability provided by hydrogen bonds depends on their degree of solvent exposure (Schmidtke et al., 2011). Gao *et al.*, quantified this effect and revealed that hydrogen bonds can be up to 1.2 $kcal.mol^{-1}$ stronger in hydrophobic environments (Gao et al., 2009). Therefore we can hypothesize that the hydrophobic contacts involving the allosteric pocket residues in one side, and those mediated by the P-loop and the front pocket residues in the other side, increase the residence time by reducing the solvent exposure of the inhibitor respectively in the allosteric channel side and front channel side. In this context, the desolvation energy barrier required for the ligand to exit the binding site and get solvated, is higher, contributing thus to increase the residence time.

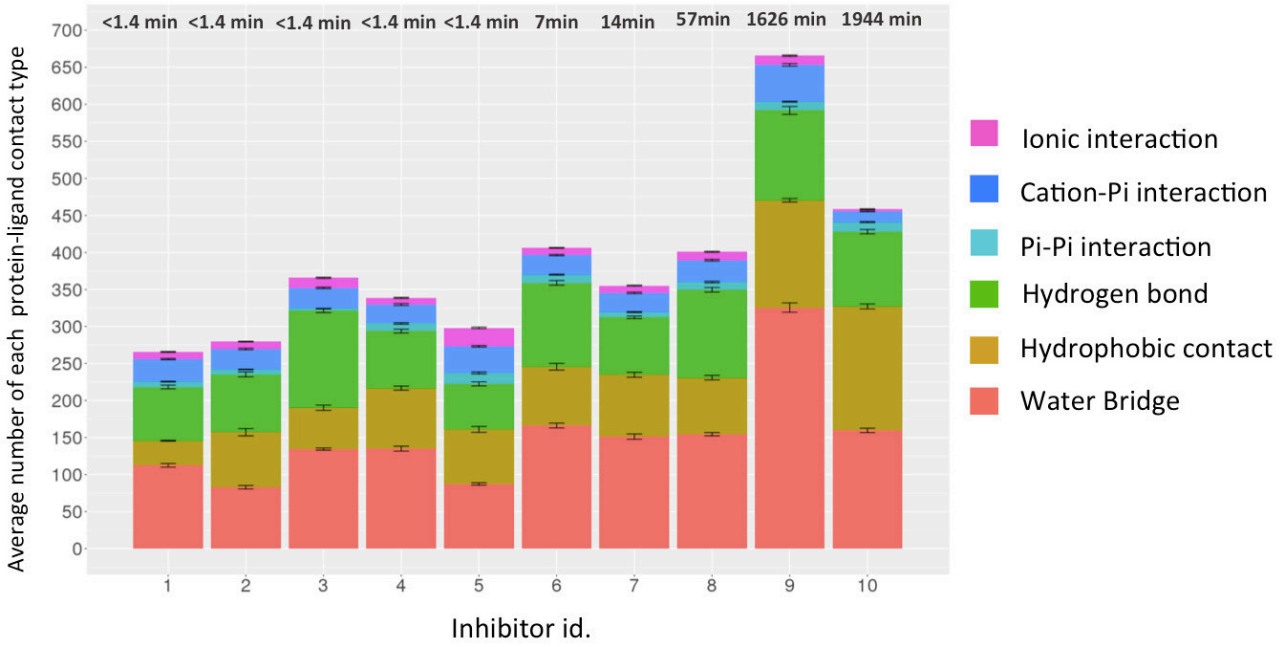


Figure 7: Stacked barplot of the average number of each protein-ligand contact type (pi-pi, HB, etc.) established during the simulation for each CDK8 inhibitors.
For each inhibitor and each replica, the total number of each protein-ligand contact type has been calculated along the simulation. This number is then averaged on the 11 replicas of each inhibitor and we calculated the error bars (based on the standard deviation).

Besides hydrophobic interactions, inhibitors also form ionic and cation-pi interactions (**Figure 7**) mostly with Arg65 and Arg150, in addition to Lys52 and Arg356, through their interaction with the scaffold rings (imidazole and phenyl). Arg65 and Arg150 are two of the three conserved arginines of CDKs family (the third one is Arg178) and are located on both sides of the access gate of the allosteric channel belonging to the α-C helix and to the HRD motif, respectively (**Figure 8**). Arg150 and Arg65, when interacting with the ligand, tend to decrease the RT_{score} when the exit occurs through the allosteric channel. This effect is clearer with Arg150 than Arg65. Indeed, we systematically observe a low RT_{score} when this interaction is formed (**Figure S12**). For the interaction with Arg65, it is a bit more complex, because this residue can adopt two conformations: toward the allosteric pocket or toward the Glu99 of the CylinC ($Glu99^{CycC}$) (**Figure 8**). When Arg65 is oriented toward the allosteric pocket the interaction between Arg65 and the inhibitor is relatively stable despite the restraining potential. However, as the inhibitor moves forward, toward the solvent in the allosteric channel, the Arg65 forms HB interaction with $Glu99^{CycC}$. In this position, Arg65 interacts alternatively with $Glu99^{CycC}$ and with the inhibitor, which drives the ligand toward the cyclin C, so toward the solvent and facilitates the dissociation process. In the literature, $Glu99^{CycC}$ was hypothesized to have an important role in the activation mechanism of CDK8. It

has been suggested that Glu99^{CycC} mimics the absent phosphoresidue within CDK8 and interacts with the three conserved arginine residues (Arg65, Arg150, and Arg178) to adjust their orientation and induce an open conformation of the activation loop (Hoeppner et al., 2005).

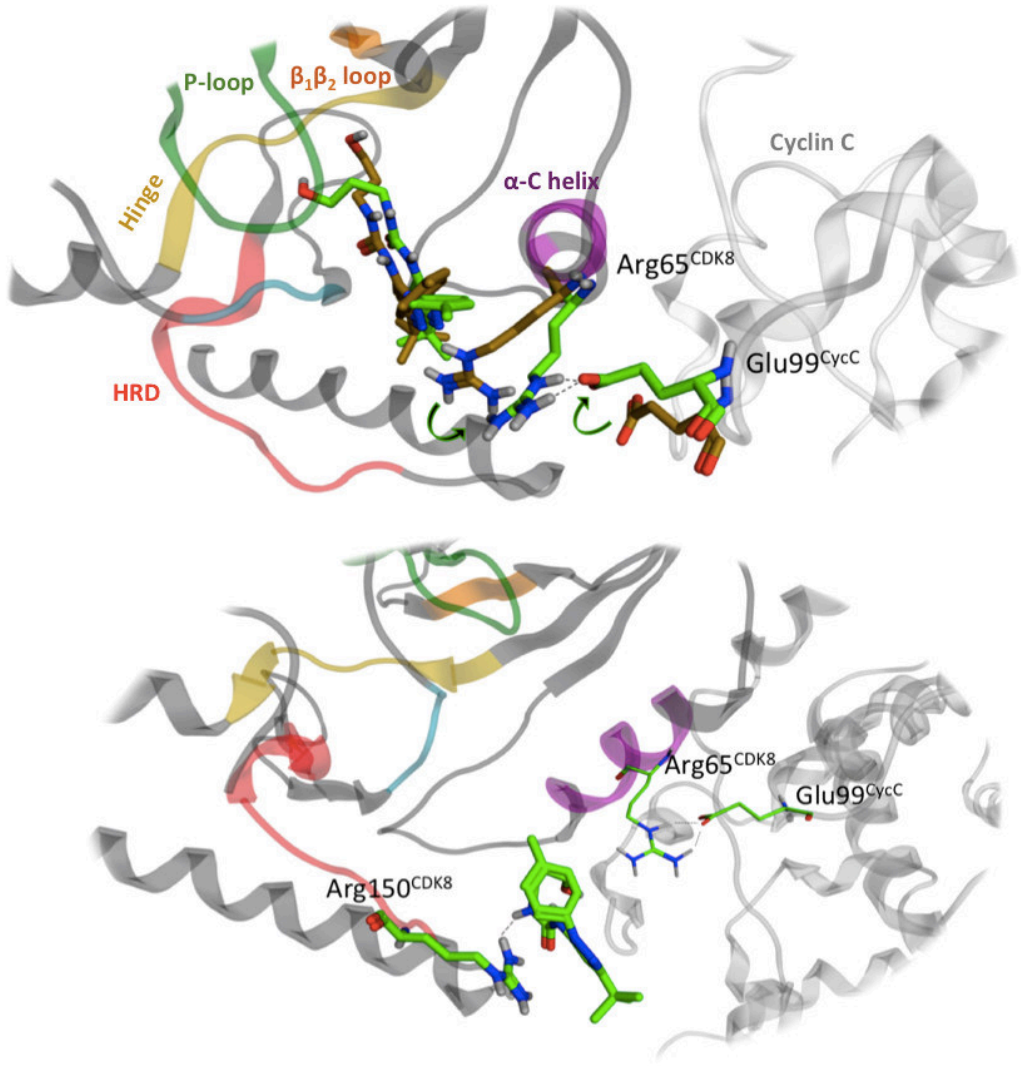


Figure 8: Arg65 and Arg 150 two CDK8 conserved residues that accelerate the dissociation process.
(Top) The Arg65^{CDK8} can adopt two conformations: toward the allosteric pocket where it interacts with the ligand through cation-pi or ionic interactions mostly (represented in stick with carbon coloured in brown) or toward the cyclin C where it interacts with the Glu99^{CycC} through a HB interaction (represented in stick with carbon coloured in green stick). **(Bottom)** The ligand interacts with Arg150^{CDK8} through mostly ionic and cation-pi interaction.

In summary, these results suggest that the interaction of the ligand with Arg65 and Arg150, two residues positioned on both sides of the entrance of the allosteric channel, have a negative contribution to its residence time. In a more general view, this overall protein-ligand

interaction study indirectly shows the importance of keeping cyclin C in the system to produce meaningful simulations. Indeed, we showed that the $Glu99^{CycC}$ through its interaction with the Arg65 could have an impact on the residence time. Moreover, besides the role of cyclin C in the biological function of CDK8, a recent study, in good agreement with our results[4], demonstrates the crucial role of cyclin C in the dynamics and structure of CDK8, particularly in its fundamental role in providing proper interaction for ligand binding through its impact on α-C helix conformation (Cholko et al., 2018). Indeed, Cholko *et al.*, observed that in the absence of cyclin C, the αC helix of CDK8 adopts an αC-out conformation, whereby Glu66 moves away from the DMG motif. By losing the H-bond from Glu66, the allosteric binding site collapses, thereby disabling the binding of type-II ligands. Surprisingly, Callegari *et al.* still obtained a correct ranking of these SRT, MRT and LRT inhibitors according to their experimental residence time using a MTD approach while they do not keep the cyclin C in their simulations (Callegari et al., 2017). Note that the experimental data were measured in the presence of cyclin C (Schneider et al., 2013). Moreover, we observed that some replicas, when they leave the binding site through allosteric channel, are not directly solvated but interact on the surface of the cyclin C. This phenomenon lengthens the exit time of these replicas (**Section S6**), exit time that is used to estimate the residence time in Callegari *et al.* (Callegari et al., 2017).

CONCLUSION

Drug-target residence time has emerged as an appealing parameter in modern drug discovery programs to select and optimize lead candidates with improved *in vivo* efficacy, in addition to the traditional focus on drug target-binding affinity (Bernetti et al., 2017). Despite the great improvements in experimental and computational methods combined with the availability of powerful computer resources, the determination of the residence time is still a challenging task. We presented here a computational method using an ensemble of targeted molecular dynamics simulations to estimate the unbinding kinetics constant of protein-ligand complexes. The method was able to properly rank a set of arylpyrazole inhibitors of cyclin-dependent kinase 8 (CDK8-CycC) according to their experimental residence time ranging from <1.4min to 1944min. One of our major concerns was to develop a method with a relatively low computational cost to be suitable for an industrial use, where dozens of

[4]Cette étude est l'objet du chapitre II.

compounds must be prioritized in the hit-to-lead and the lead optimization phases. For a kinase system prepared as mentioned in the Material and Methods section our method has a throughput of 5 ligands per day (11 replicas per ligand) using 11 computers of 8 cpu (Intel(R) Xeon(R) Gold 5115 CPU @ 2.40GHz). Moreover, another advantage lays in its simplicity and in the fact it does not require any specific *a priori* knowledge on the exit pathway. The used reaction coordinate (RMSD) has the advantage to induce soft changes since the ligand can increase its RMSD just by changing its conformation without moving ahead its mass center, which provides realistic description of the unbinding process. We subsequently focused on establishing structure–kinetics relationships, to identify the chemical features impacting the residence time. The results highlight the importance of hydrophobic interactions with the allosteric pocket, the P-loop and the front pocket residues, and the HB interactions with the hinge and the DMG motif residues. This SKR study could be of valuable help in designing CDK8-CycC inhibitors with an optimized kinetics profile and thus an improved *in vivo* profile.

SUPPORTING INFORMATION

Section S1: Model building and system preparation.

```
Uniprot_Cyclin-C      ---MAGNFWQSSHYLQWILDKQDLLKERQKDLKFLSEEEYWKLQIFFTNVIQALGEHLKL
4F6U.B                DKAMAGNFWQSSHYLQWILDKQDLLKERQKDLKFLSEEEYWKLQIFFTNVIQALGEHLKL
                         ************************************************************

Uniprot_Cyclin-C      RQQVIATATVYFKRFYARYSLKSIDPVLMAPTCVFLASKVEEFGVVSNTRLIAAATSVLK
4F6U.B                RQQVIATATVYFKRFYARYSLKSIDPVLMAPTCVFLASKVEEFGVVSNTRLIAAATSVLK
                      ************************************************************

Uniprot_Cyclin-C      TRFSYAFPKEFPYRMNHILECEFYLLELMDCCLIVYHPYRPLLQYVQDMGQEDMLLPLAW
4F6U.B                TRFSYAFPKEFPYRMNHILECEFYLLELMDCCLIVYHPYRPLLQYVQDMGQEDMLLPLAW
                      ************************************************************

Uniprot_Cyclin-C      RIVNDTYRTDLCLLYPPFMIALACLHVACVVQQKDARQWFAELSVDMEKILEIIRVILKL
4F6U.B                RIVNDTYRTDLCLLYPPFMIALACLHVACVVQQKDARQWFAELSVDMEKILEIIRVILKL
                      ************************************************************

Uniprot_Cyclin-C      YEQWKNFDERKEMATILSKMPKPKPPPNSEGEQGPNGSQNSSYSQS
4F6U.B                YEQWKNFDERKEMATILSKMPKPKPPP------------------
                      ***************************
```

Figure S1: Alignment of the UNIPROT sequence of the human Cyclin C with the Cyclin C sequence in the PDB file 4F6U.

```
Uniprot_CDK8      --MDYDFKVKLSSERERVEDLFEYEGCKVGRGTYGHVYKAKR-KDGKDDKDYALKQIE--
4F6U.A            DKMDYDFKVKLSSERERVEDLFEYEGCKVGRGTYGHVYKAKR-KDGKDDKDYALKQIE--
1BI8.A            --MEKDGLCRADQQYECVAE--------IGEGAYGKVFKARDLKNG--GRFVALKRVRVQ
1G3N.A            --MEKDGLCRADQQYECVAE--------IGEGAYGKVFKARDLKNG--GRFVALKRVRVQ
                    *: *   : ..: * * :        :*.*:**:*:**:  *:*  .:  ***::.

Uniprot_CDK8      -GT-GISMSACREIALLRELK---HPNVISLQKVF-LSHADR--KVWLLFDYAEHDLWHI
4F6U.A            -GT-GISMSACREIALLRELK---HPNVISLQKVF-LSHADR--KVWLLFDYAEHDLWHI
1BI8.A            TGEEGMPLSTIREVAVLRHLETFEHPNVVRLFDVCTVSRTDRETKLTLVFEHVDQDLTTY
1G3N.A            TGEEGMPLSTIREVAVLRHLETFEHPNVVRLFDVCTVSRTDRETKLTLVFEHVDQDLTTY
                   *  *:.:*: **:*:**.*:   ****: * .*  :*::**  *: *:*::.::**

Uniprot_CDK8      IKFHRASKANKKPVQLPRGMVKSLLYQILDGIHYLHANWVLHRDLKPANILVMGEGPERG
4F6U.A            IKFHRASK-----VQLPRGMVKSLLYQILDGIHYLHANWVLHRDLKPANILVMGEGPERG
1BI8.A            L-----DKVPEPGV--PTETIKDMMFQLLRGLDFLHSHRVVHRDLKPQNILVTSSG----
1G3N.A            L-----DKVPEPGV--PTETIKDMMFQLLRGLDFLHSHRVVHRDLKPQNILVTSSG----
                  :     .*     *  *   :*.:::*:* *:.:**:: *:****** **** ..*

Uniprot_CDK8      RVKIADMGFARLFNSPLKPLADLDPVVVTFWYRAPELLLGARHYTKAIDIWAIGCIFAEL
4F6U.A            RVKIADMGF----------------VVTFWYRAPELLLGARHYTKAIDIWAIGCIFAEL
1BI8.A            QIKLADFGLARIYSFQMA----LTSVVVTLWYRAPEVLLQSS-YATPVDLWSVGCIFAEM
1G3N.A            QIKLADFGLARIYSFQMA----LTSVVVTLWYRAPEVLLQSS-YATPVDLWSVGCIFAEM
                  ::*:**:*:                ***:******:** :  *:..:*:*::******:

Uniprot_CDK8      LTSEPIFHCRQEDIKTSNPYHHDQLDRIFNVMGFPADKDWEDIKKMPEHSTLMKDFRRNT
4F6U.A            LTSEPIFHCRQE-----NPYHHDQLDRIFNVMGFPADKDWEDIKKMPEHSTLMKDFRRNT
1BI8.A            FRRKPLFR-GSSDV--------DQLGKILDVIGLPGEEDWPRDVALPRQAFHSKSAQP--
1G3N.A            FRRKPLFR-GSSDV--------DQLGKILDVIGLPGEEDWPRDVALPRQAFHSKSAQP--
                  :  :*:*:  ..          ***.:*::*:*:*.::**     :*.::   *. :

Uniprot_CDK8      YTNCSLIKYMEKHKVKPDSKAFHLLQKLLTMDPIKRITSEQAMQDPYFLEDPLPTSDVFA
4F6U.A            YTNCSLIKYMEKHKVKPDSKAFHLLQKLLTMDPIKRITSEQAMQDPYFLEDPLPTSDVFA
1BI8.A            ---------IEKFVTDIDELGKDLLLKCLTFNPAKRISAYSALSHPYFQDLERCKENLDS
1G3N.A            ---------IEKFVTDIDELGKDLLLKCLTFNPAKRISAYSALSHPYFQDLERCKENLDS
                          :**. .. *. . .** * **::* ***:: .*:..*** :    ..:: :

Uniprot_CDK8      GCQIPYPKREFLTEEEPDDKGDKKNQQQQQGNNHTNGTGHPGNQDSSHTQGPPLKKVRVV
4F6U.A            GCQIPYPKREFLTEEEPDDKGDKKNQQQQQGNNHTNGTGHPGNQDSSHTQGPPLKK----
1BI8.A            -------------------------------------HLPPSQNTSELNTA--------
1G3N.A            -------------------------------------HLPPSQNTSELNTA--------
                                                       * .*::*. : .
```

Figure S2: Alignment of the human sequences of CDK8 (the canonical UNIPROT sequence and the sequence in the PDB file 4F6U) with the template sequences of human CDK6 (sequences in the PDB files 1BI8 and 1G3N).

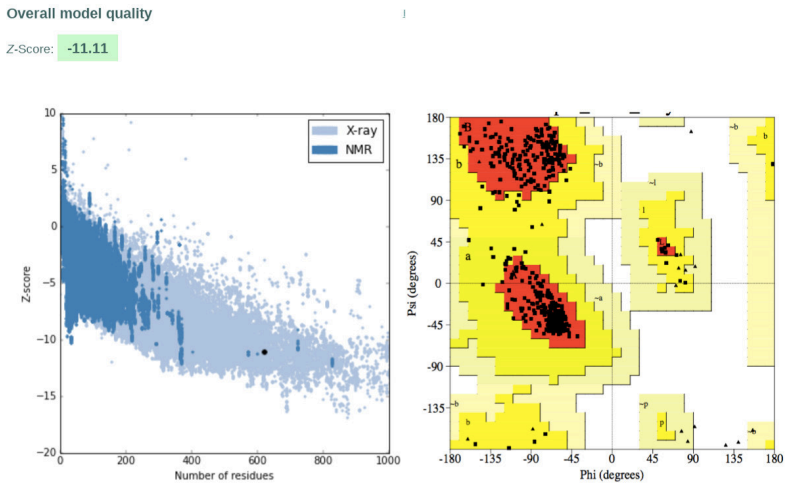


Figure S3: Model validation.
(**Left**) Plot of the Z-score. The *Z*-score indicates overall model quality. The plot contains the *z*-scores of all experimentally determined protein chains in current PDB. In this plot, groups of structures from different sources (X-ray, NMR) are distinguished by different colors. The z-score of the reconstructed model of CDK8-CycC complex is represented by a black point. Its value is within the range of scores typically found for native proteins of similar size. **(Right)** Ramachandran plot showing that most of the model residues present values of the backbone dihedral angle phi and psi in the energetically allowed regions.

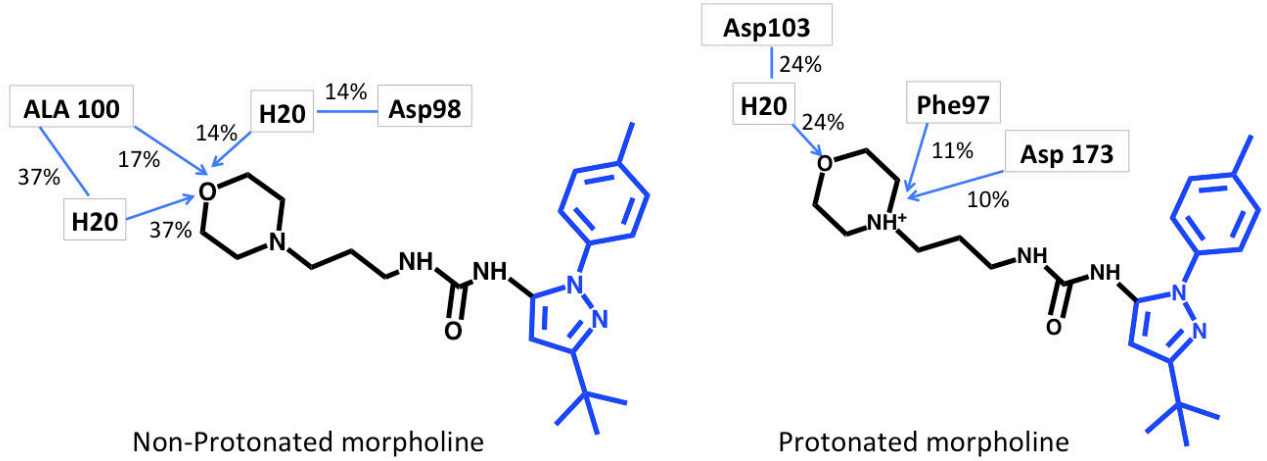


Figure S4: Difference in protein-ligand interaction between the form non-protonated and protonated of the morpholine of inhibitor 0SR.
The interactions involving the scaffold (drawn in blue) and the urea are not represented because no difference has been noted. Considering the other interactions, they have been measured along a brute force MD simulation of 1 microseconde. The percentage represents the time of the simulation during which the interaction is maintained. The interaction with Ala100 is not observed when the morpholine is protonated.

Section S2: TMD Setup

Calculation of the reaction coordinate: RMSD

The RMSD is calculated on the heavy atoms of the ligand, after aligning the reference and the current structures on the backbone of a set of 22 residues belonging to the binding site. These 22 residues are apart of 4 Å from the center of mass of the binding site. This center of mass is calculated on the conserved motifs of kinase protein: the P-loop (residues 27-35), Hyd1 motif (residues 50-52), αC-helix (residues 63-70), the hinge (residues 97-103), the HRD motif (149-158) and the DMG motif (residues 172-175).

Section S3: Impact of the harmonic restraining potential on the stability of the system

To investigate the impact of adding a harmonic restraining potential ($V_{restraint}$) on the overall stability of the system, the values of RMSF and RMSD were computed and compared to the ones of brute force molecular dynamics (cMD) (**Figure S5**). Protein structure and dynamics are comparable in TMD and cMD simulation. The backbone RMSD values are always below or equal to 2Å at the beginning of the simulation and may display an abrupt increase up to 3-4 Å

during the transition toward the unbound state. Once the ligand is solvated, the RMSD values go back to the original values.

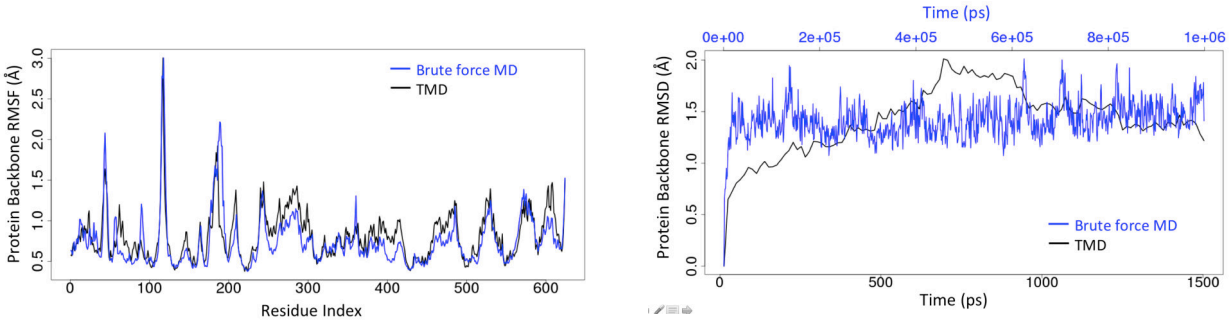


Figure S5: Stability of the system.
The Root Mean Square Fluctuation (RMSF) (left) and the Root Mean Square Deviation (RMSD) of TMD simulation are compared to those of brute force MD (1 microseconde).

We then examine the distribution of $V_{restraint}$ of each trajectory, with a particular attention on the maximum values. The distributions of $V_{restraint}$ are centered on a value <1 kcal.mol^{-1}. The values of $V_{restraint}$ never exceed 22.5 kcal.mol^{-1} and values between 2.5 kcal.mol^{-1} and 22.5 kcal.mol^{-1} are extreme and rare values (**Figure S6**). Moreover, the maximum values of $V_{restraint}$ are always observed at the beginning of the unbinding process when the ligand leaves the binding site. In light of these observations, $V_{restraint}$ is sufficiently low and imposes sufficiently slow variation during the unbinding process to consider the simulation as a realistic description of the unbinding process.

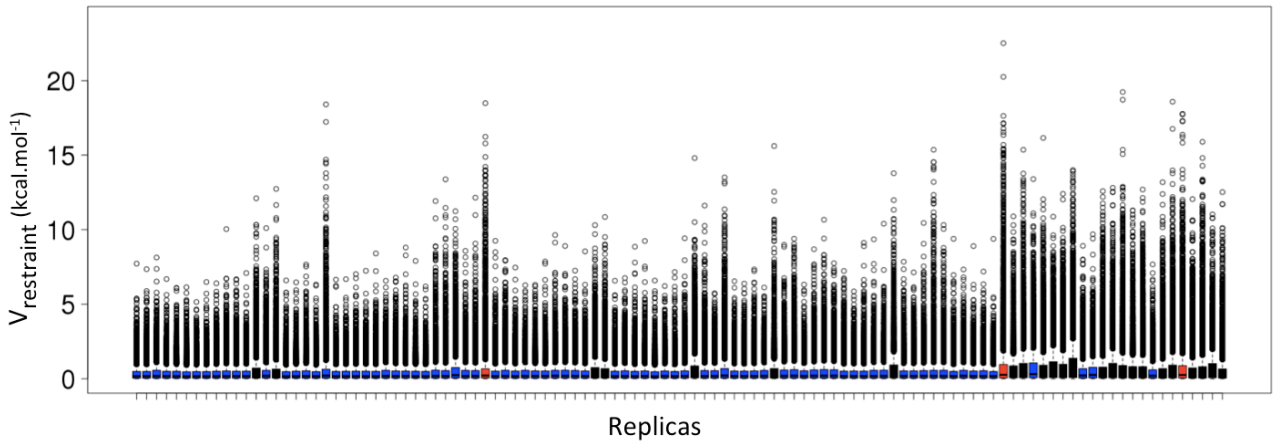


Figure S6: Boxplots of the harmonic restraining potential ($V_{restraint}$) of each replica (11 replica×10 inhibitors = 110 replicas).
The boxplots are colored according to the taken pathway: black for ATP channel, blue for the allosteric channel and red for the hinge channel.

Section S4: Relationship between RT_{score} and $RT_{experimental}$

The quantity noted RT_{score} represents the overall action required to dissociate the ligand from the binding site and get it solvated. We hypothesize a linear relationship between RT_{score} and the activation energy (ΔG_{off}) defined in the Arrhenius equation:

$$k_{off} = A.e\frac{-\Delta G_{off}}{B}$$

Equation d'Arrhenius

With B = R × T, R is the ideal gas constant and T the temperature. A is the frequency factor. By replacing ΔG_{off} with RT_{score}, we get:

$$k_{off} = A.e\frac{-RT_{score}}{B}$$

Taking the natural logarithm linearizes the equation; we then obtain:

$$\ln(k_{off}) = \ln\left(A.e\frac{-RT_{score}}{B}\right)$$

$$\ln(k_{off}) = \ln(A) - \frac{RT_{score}}{B}$$

Considering that:

$$\ln(k_{off}) = \ln\left(\frac{1}{RT_{experimental}}\right) = -\ln\left(RT_{experimental}\right)$$

We obtain:

$$\mathbf{\ln\left(RT_{experimental}\right) = -\ln(A) + \frac{RT_{score}}{B}}$$

Therefore, we make the assumption that the RT_{score} is linearly related to the natural logarithm of $RT_{experimental}$.

Section S6: Analysis of protein-ligand interactions

Details on the calculation of the protein-ligand interactions

For each simulation, the protein-ligand interactions were calculated on a total of 150 conformations (one snapshot every 10 ps) using the SID module from the maestro suite (Maestro, Schrödinger, LLC, New York, NY, 2016) after converting the trajectories to the maestro format. The different calculated interactions are defined as follows:

- Considering a hydrogen bond between a donor atom (D) and an acceptor atom (A) noted D-H...A-X, a hydrogen bond is considered when the D-A distance is less than 2.5Å and the D-H-A angle greater than 120° and the H-A-X angle greater than 90°.
- A ionic interaction is defined by an interaction between oppositely charged atoms on the ligand and the protein that are within 3.7Å.
- A water bridge interaction involves a hydrogen bond via a water bridge molecule. The geometric criteria are a D-A distance less than 2.7Å, a D-H-A angle greater than 110° and a H-A-X angle greater than 80°.
- A pi-pi interaction is defined as an interaction between two aromatic rings in which either (a) the angle between the ring planes is less than 30° and the distance between the ring centroids is less than 4.4 Å (face-to-face), or (b) the angle between the ring planes is between 60° and 120° and the distance between the ring centroids is less than 5.5 Å (edge-to-face).
- A cation-pi interaction is considered when the maximum distance between the cation center and the ring center is 6.6 Å and the angle between the ring plane and the line between the cation center and the ring center does not deviate from the perpendicular by more than 30°.
- A hydrophobic contact is considered when a hydrophobic side chain is within 3.6Å of ligand aromatic or aliphatic carbon.

Structure kinetic relationship

SRT inhibitors

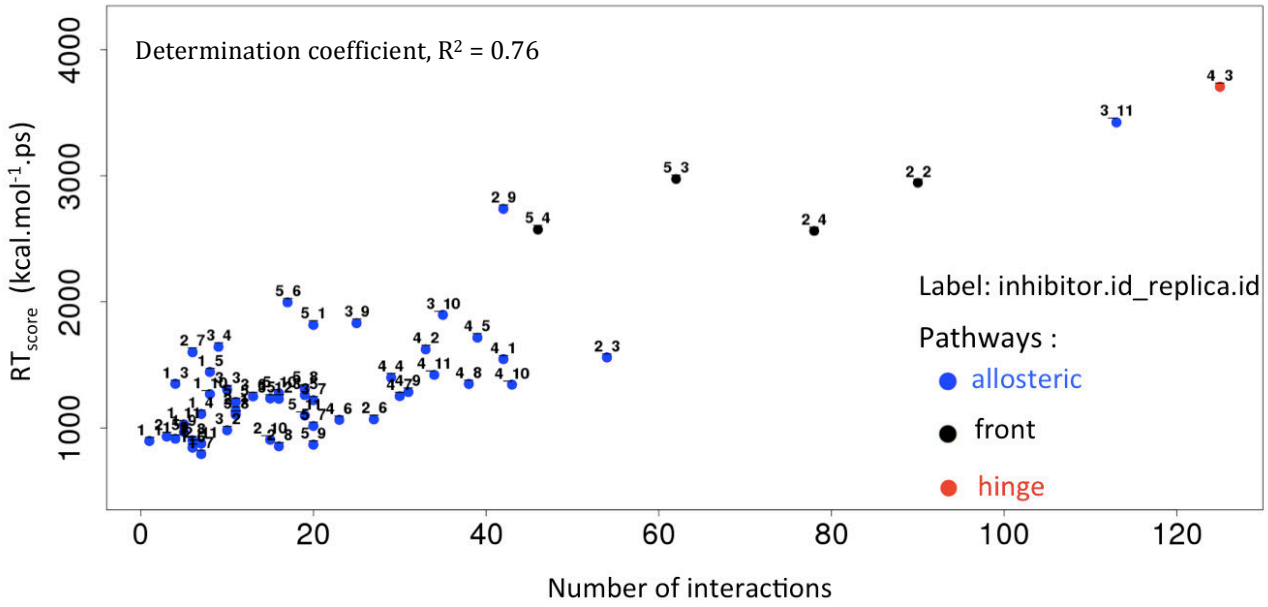


Figure S7: RT_{score} of SRT inhibitors as a function of the total number of hydrophobic contacts established along the simulation involving the allosteric pocket residues and the gatekeeper Phe97.

Only the replicas of the SRT inhibitors are plotted here. For each replica, we sum the total number of hydrophobic contacts involving Leu70, Leu73, Val78, Ile79, Leu142 and Val147 (allosteric pocket residues) and Phe97 established during the simulation.

MRT inhibitors

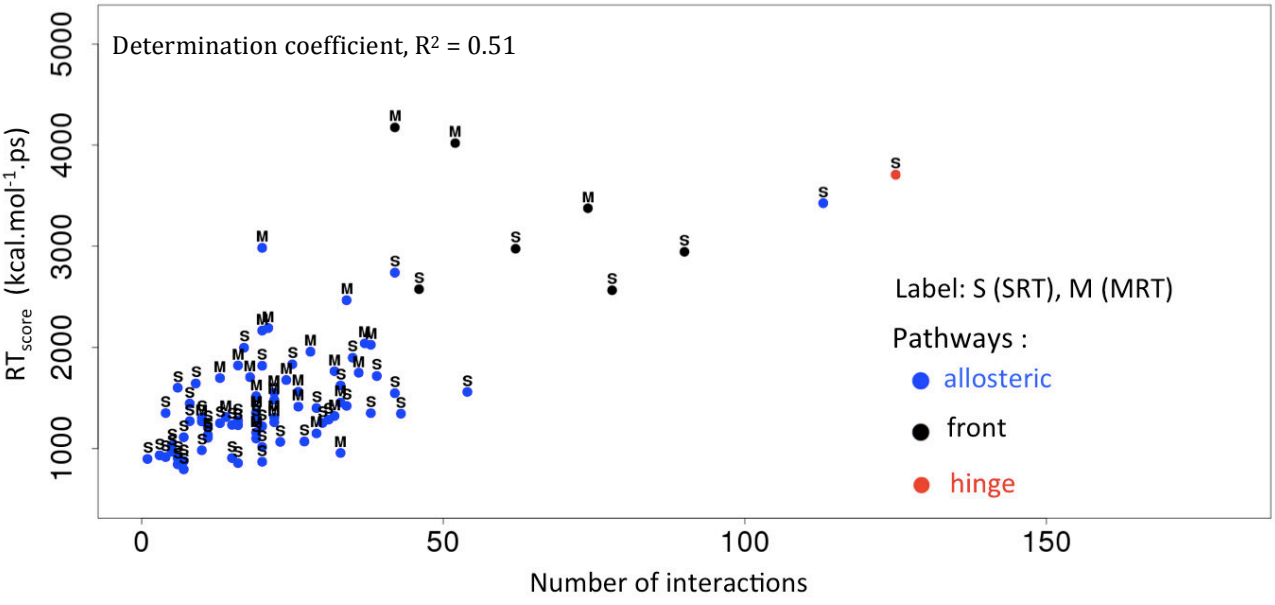


Figure S8: RT_{score} of SRT and MRT inhibitors as a function of the total number of hydrophobic contacts established along the simulation involving the allosteric pocket residues and the gatekeeper Phe97.

All the replicas of SRT and MRT inhibitors are plotted here. For each replica, we sum the total number of hydrophobic contacts involving Leu70, Leu73, Val78, Ile79, Leu142 and Val147 (allosteric pocket residues) and Phe97 established during the simulation.

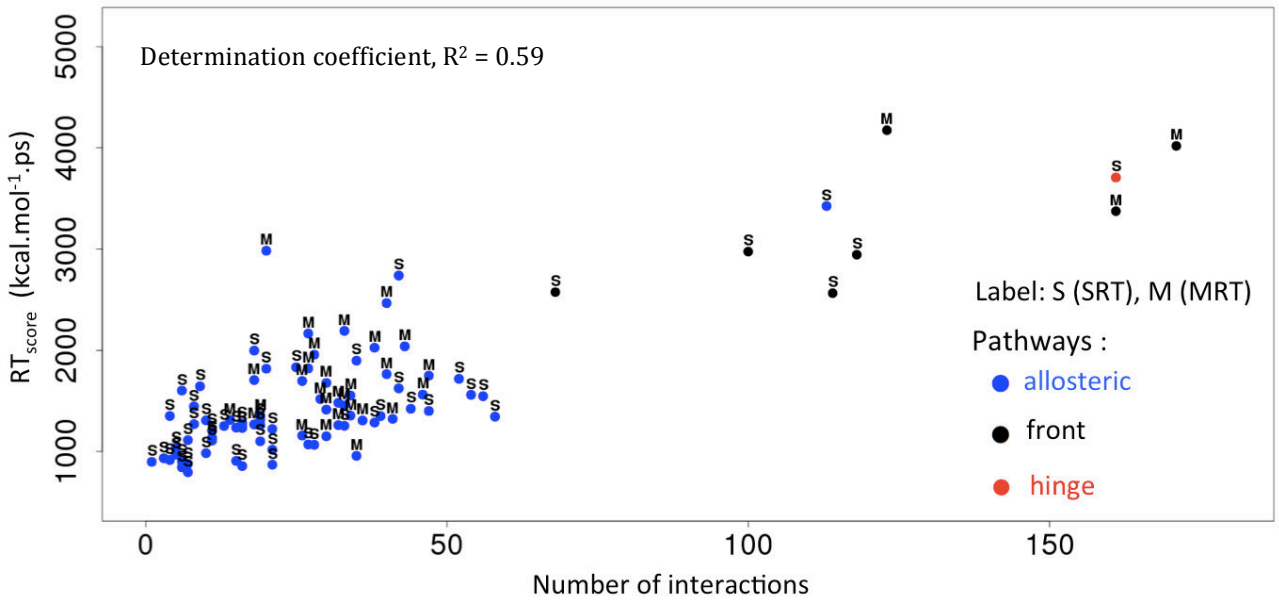


Figure S9: RT_{score} of SRT and MRT inhibitors as a function of the total number of hydrophobic contacts involving the allosteric pocket residues and the gatekeeper Phe97, and the HB interactions involving the hinge region.

All the replicas of SRT and MRT inhibitors are plotted here. For each replica, we sum the total number of hydrophobic contacts involving Leu70, Leu73, Val78, Ile79, Leu142 and Val147 (allosteric pocket residues) and Phe97 and, the HB interactions involving Asp98, Tyr99, Ala100, established during the simulation.

LRT inhibitors

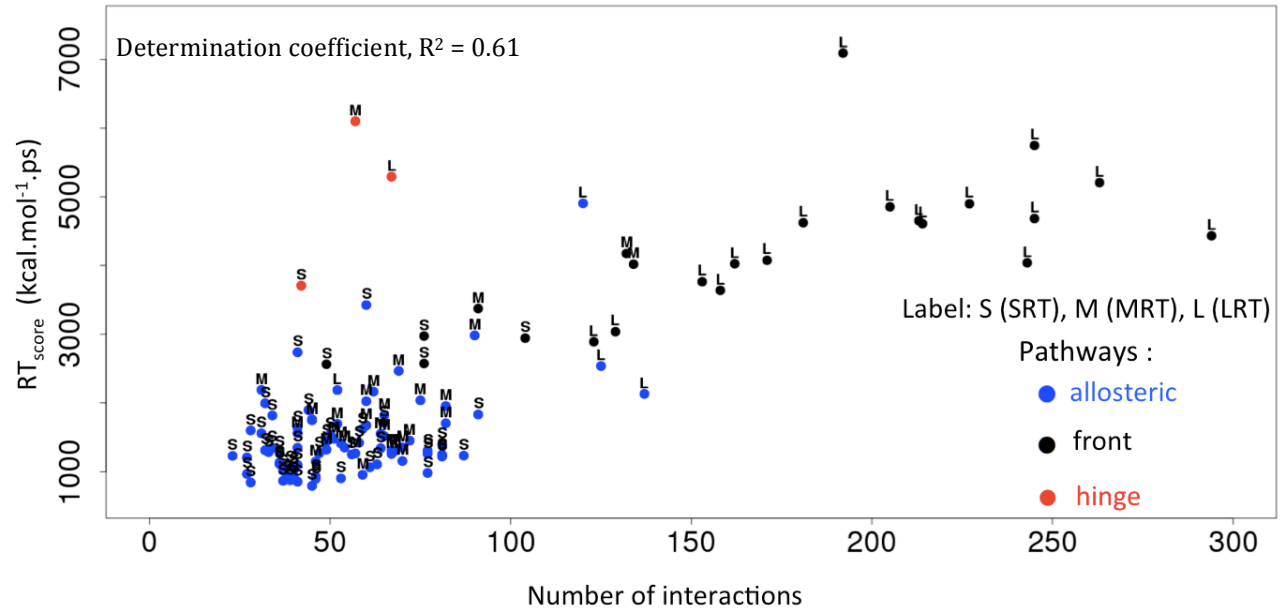


Figure S10: RT_{score} of SRT, MRT and LRT inhibitors as a function of the total number of hydrophobic contacts and HB interactions involving the residues of the P-loop (Val27, Arg29 and Val35) and the DMG motif (Asp173, Met174).

All the replicas of SRT, MRT and LRT inhibitors are plotted here. For each replica we sum the total number of hydrophobic contacts and HB interactions involving the P-loop (Val27, Arg29 and Val35) and the DMG motif (Asp173, Met174), established during the simulation.

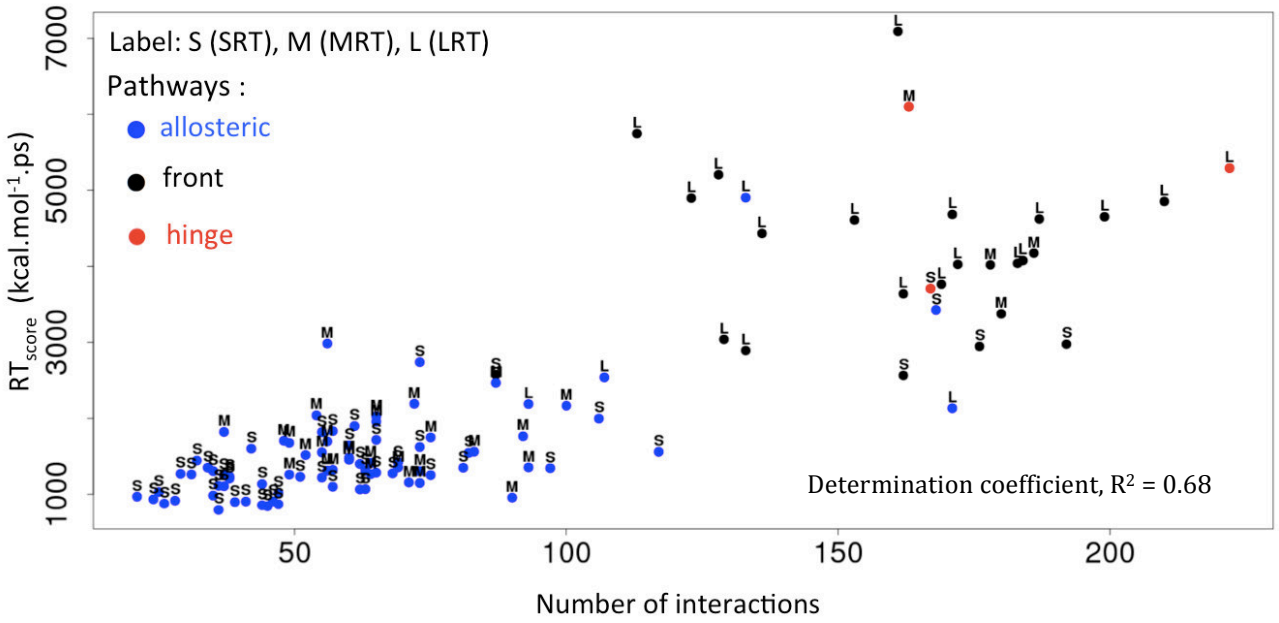
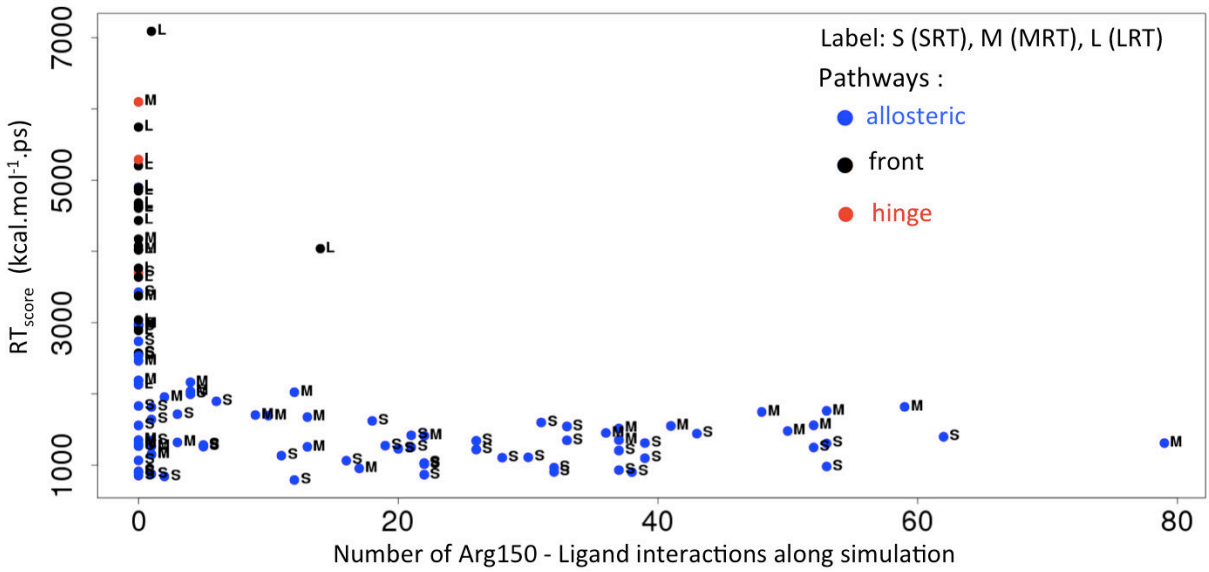


Figure S11: RT_{score} of SRT, MRT and LRT inhibitors as a function of the total number of hydrophobic contacts established during the simulation (with all residues).
All the replicas of SRT, MRT and LRT inhibitors are plotted here. For each replica, we sum the total number of hydrophobic contacts with all residues established during the simulation.



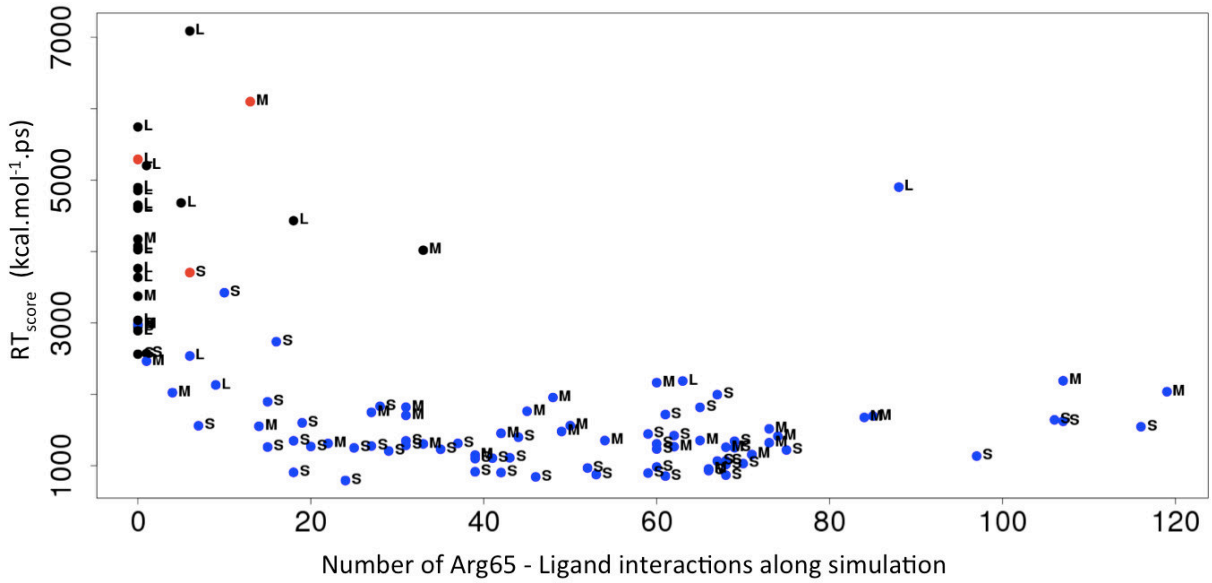


Figure S12: RT$_{score}$ of SRT, MRT and LRT inhibitors as a function of the total number contacts established during the simulation with Arg150 (top) or Arg65 (bottom).
All the replicas of SRT, MRT and LRT inhibitors are plotted here. For each replica, we sum the total number of contacts involving Arg150 **(top)** and Arg65 (**bottom**) established during the simulation.

Section S6: Exit time

We also compute the average exit time of each inhibitor defined as the time at which all atoms of the ligand are at least at 6 Å of any atoms of the protein (the protein being the CDK8-CycC complex). The goal was to compare our results to those of Callegari *et al.* (Callegari et al., 2017) that uses this criterion to predict the residence time and do not kept the cyclin C in their simulations. As we can see on the **Figure S13**, the exit time calculated, following the definition above, do not allow to separate SRT, MRT and LRT inhibitors. A more detailed analysis of the results reveals that for some replicas the exit time is lengthened because the ligand continues to migrate on the protein surface, either on the cyclin C or on the C-lobe of the kinase, before eventually being dragged away under the imposed TMD forces. The **Figure S13** is consistent with this observation where we see a better discrimination between LRT inhibitors and the SRT and MRT inhibitors when considering only CDK8 or only the residues that are part of the exit pathway (binding site residues from 20 to 200) in the calculation of the exit time.

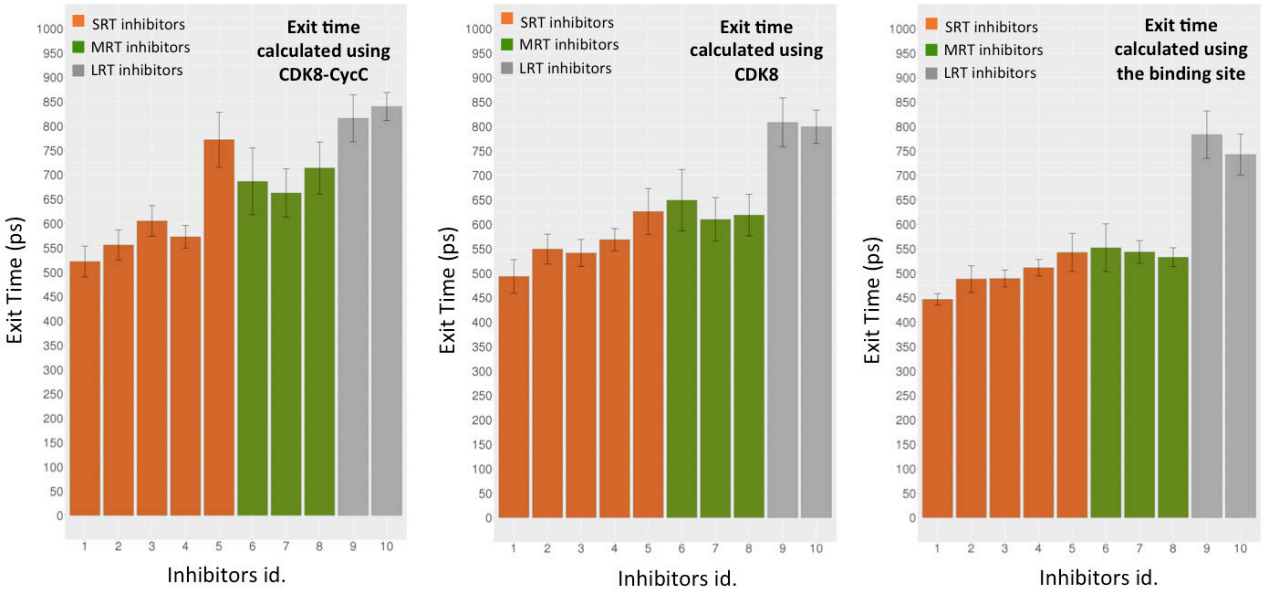


Figure S13: Exit time of CDK8 inhibitors calculated from the simulations.
The exit time is the time at which any atoms of the ligand are at least at 6 Å of any atoms of the receptor. From left to right, we took as receptor: the complex CDK8-CycC, CDK8 and the binding site (residues 20-200, that is residues that are part of the exit pathways).

B. Etude QSKR à partir de profils énergétiques d'interaction

Dans cette partie**,** nous présenterons une approche originale d'analyse automatique de données issues de dynamique moléculaire. L'approche a été appliquée ici, pour développer un modèle QSKR (*Quantitative Structure Kinetics Relationship*) à partir des profils énergétiques d'interaction calculés à partir des simulations de dynamique moléculaire (obtenues dans la partie A).

1. Préambule

La simulation de dynamique moléculaire (DM) est une méthode de plus en plus utilisée dans le processus de recherche et de développement de nouveaux médicaments. Avec l'augmentation de la puissance de calculs et de la capacité de stockage des données, nous sauvegardons de plus en plus de trajectoires issues de DM sur les disques durs qui constituent une mine d'information complexes à exploiter. Une trajectoire de DM représente l'évolution des coordonnées du système étudié au cours du temps. Cette information brute (les coordonnées du système) constitue un ensemble de séries temporelles. Des analyses réalisées sur cette information brute peuvent également se présenter sous forme de séries temporelles telles que : le nombre de liaisons hydrogène au cours du temps, l'évolution d'une énergie au cours du temps, le volume du site de liaison etc. Lorsqu'il s'agit de comparer ce type de données entre plusieurs trajectoires, les chercheurs ont souvent recours à des fonctions d'agrégation (moyenne, variance, etc.) ou à des méthodes de partitionnement de données (*clustering*) pour résumer l'information et faciliter l'analyse. Outre le fait que l'on perd de l'information, ce type d'analyse n'est pas adapté lorsque le processus étudié est lui-même dynamique (par exemple la dissociation d'un ligand de sa protéine) et que l'on cherche justement à capturer et comparer l'évolution d'un évènement entre 2 trajectoires. Il peut s'agir par exemple de comparer l'évolution, au cours du processus de dissociation, des contacts hydrophobes de plusieurs ligands ayant des temps de résidence différents. L'objectif serait alors de vérifier s'il existe un lien entre le temps de résidence et l'évolution des contacts hydrophobes. De fait, le nombre moyen de contacts hydrophobes au cours de la dynamique peut être similaire entre plusieurs ligands de temps de résidence différents, donc non corrélé au temps de résidence. Cependant, l'évolution de ces contacts au cours du temps peut être, elle, corrélée au temps de résidence des ligands. Pour répondre à ce type de problématique, les méthodes d'apprentissage automatique et la construction d'un modèle statistique sont des

solutions plus adaptées. Dans cette perspective, la variable à expliquer (Y) serait le temps de résidence et les variables explicatives X, communément appelées descripteurs, proviendraient des séries temporelles extraites des trajectoires de DM. Il s'agirait de quantifier la relation entre le temps de résidence Y et les variables X issues des séries temporelles extraites des trajectoires de DM. Cependant, les valeurs constituant une série temporelle ne sont pas des variables indépendantes utilisables en tant que descripteurs dans un algorithme d'apprentissage automatique. De plus, lorsqu'on souhaite comparer plusieurs courbes, ce ne sont pas les valeurs brutes de la série temporelle qui nous intéressent mais plutôt les descripteurs qui en découlent tels que la pente de la courbe, la présence de motifs particuliers, le nombre de minima, leur profondeur, etc. C'est ce type de donnée qu'il faut extraire des séries temporelles et analyser par les méthodes d'apprentissage automatique afin d'évaluer un lien potentiel avec une propriété Y telle que le temps de résidence. La littérature sur les séries temporelles est très riche lorsqu'il s'agit de comprendre comment faire des prévisions (*forecasting*) sur une série temporelle au temps *t+1* sachant la valeur au temps *t*, ou encore lorsqu'on cherche à traiter un signal comme le débruitage par exemple. En revanche, elle est pauvre concernant la façon dont on peut appliquer des méthodes d'apprentissage automatique sur un ensemble de séries temporelles en vue de les clustériser ou de construire des modèles statistiques de régression ou classification. C'est pourquoi, nous proposons dans cette partie de montrer comment on peut extraire efficacement des informations contenues dans un ensemble de séries temporelles tel que la pente de la courbe, la présence de motifs temporels particuliers, la concavité de la courbe etc., et les exploiter afin de comprendre le phénomène sous-jacent simulé.

2. Objectif

Nous nous proposons d'investiguer, d'un point de vue énergétique, la relation structure-cinétique du même jeu de données que décrit précédemment, composé de 10 inhibiteurs de CDK8-CycC (**Figure 26** and **Figure 27**)**.** Il s'agit de comprendre quelle(s) composante(s) énergétique(s) de l'interaction protéine-ligand pourrai(en)t contribuer le plus à expliquer le temps de résidence. Pour cela, les termes de l'énergie non-liée du champ de forces ainsi que ceux de l'énergie de solvatation sont calculés par la méthode MM-GBSA à partir des simulations de DM biaisées décrivant la dissociation des inhibiteurs (partie A). Ces profils énergétiques, qui sont des séries temporelles, sont ensuite traités et soumis à une méthode d'apprentissage par arbres de décision (*random forest*) pour en dériver un modèle QSKR.

Quantifier la relation entre les différentes composantes énergétiques de l'interaction protéine-ligand et le temps de résidence, peut suggérer des modifications structurales à introduire sur ligand pour augmenter ou réduire son temps de résidence. Au delà de la compréhension, on pourra envisager par la suite d'utiliser un tel modèle pour prédire le temps de résidence d'un composé à partir de ses profils énergétiques calculés par la méthode MM-GBSA. Enfin, la méthodologie développée est une procédure générale qui peut s'appliquer à d'autres études où l'on souhaite faire de l'apprentissage automatique sur des séries temporelles afin de comprendre et ou de prédire une propriété Y donnée.

3. Méthodologie

Le processus utilisé pour identifier les termes énergétiques pertinents issus des simulations de DM est décrit ci-dessous.

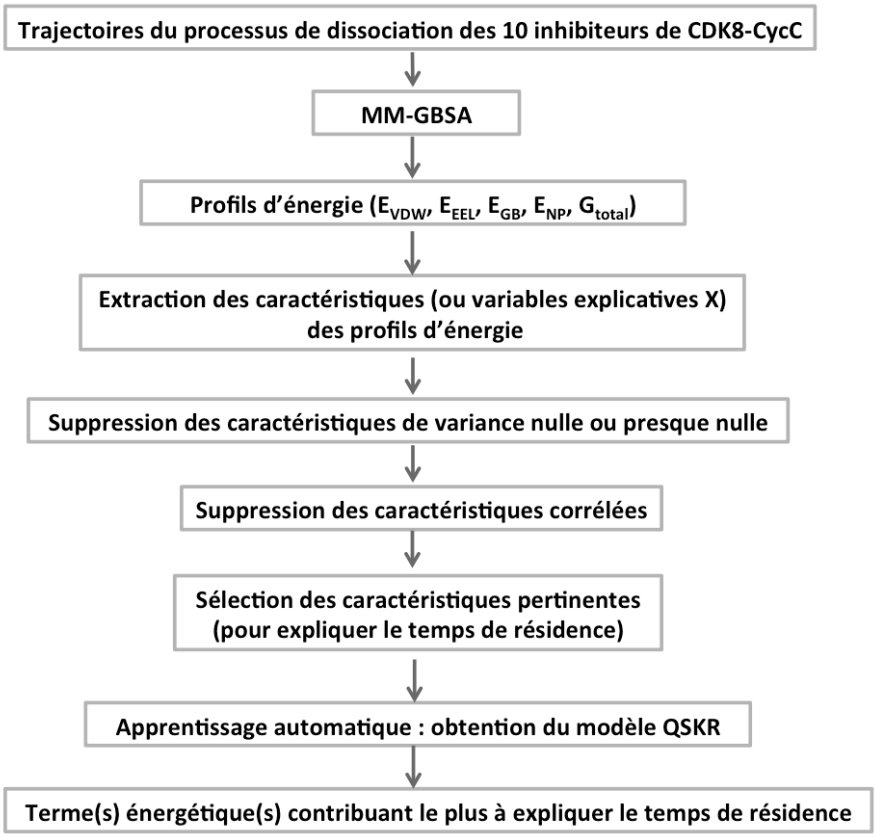


Figure 28 : Schéma général du protocole utilisé

a) Préparation des données : extraction des descripteurs

Comme la dissociation de chaque inhibiteur a été simulée 11 fois, nous disposons d'un total de 110 simulations de DM biaisées. La méthode MM-GBSA (chapitre V.J.1) a été appliquée sur

chacune de ces trajectoires en considérant l'ensemble des conformations enregistrées dans chaque trajectoire, soit 7501 au total pour un temps de simulation de 1500 ps. Le modèle II de Born généralisé (*igb=5*) disponible dans AMBER, a été utilisé pour calculer la contribution polaire de l'énergie de solvatation (cf. V.J.1.b)) (Onufriev et al., 2004). Pour chaque trajectoire, l'ensemble des termes énergétiques calculés pour chaque conformation a été sauvegardé : termes d'énergie liée et non-liée du champ de forces ainsi que les composantes de l'énergie de solvatation. Dans le cadre de l'analyse des interactions protéine-ligand, les termes qui nous intéressent sont l'énergie de van der Waals (E_{VDW}), l'énergie électrostatique (E_{EEL}), les contributions polaires (E_{GB}) et non-polaires (E_{NP}) de l'énergie de solvatation et aussi l'énergie libre totale (G_{total}) (sans prendre en compte l'entropie). Les profils énergétiques de ces termes sont extraits des simulations. Ainsi, 5 profils d'énergie sont associés à chaque simulation. Le **Tableau 9** donne une vue d'ensemble de la structure des données brutes.

	Y		X				
Id. de l'inhibiteur	Temps de résidence (min)	Réplica	Valeurs énergétiques (kcal.mol^{-1})				
1	< 1.4	1	$Profil_{VDW}$ $[X_1 \ldots X_{Nvaleur = 7501}]$	$Profil_{EEL}$	$Profil_{GB}$	$Profil_{NP}$	$Profil_{Gtot}$
		. . .	. . .	. . .	. . .	. . .	. . .
		$N_{replica} = 11$	$Profil_{VDW}$ $[X_1 \ldots X_{Nvaleur = 7501}]$	$Profil_{EEL}$	$Profil_{GB}$	$Profil_{NP}$	$Profil_{Gtot}$
. . .	. . .	. . .	. . .				
10	1944	1	$Profil_{VDW}$ $[X_1 \ldots X_{Nvaleur = 7501}]$	$Profil_{EEL}$	$Profil_{GB}$	$Profil_{NP}$	$Profil_{Gtot}$
		. . .	. . .	. . .	. . .	. . .	. . .
		$N_{replica} = 11$	$Profil_{VDW}$ $[X_1 \ldots X_{Nvaleur = 7501}]$	$Profil_{EEL}$	$Profil_{GB}$	$Profil_{NP}$	$Profil_{Gtot}$

Tableau 9 : Tableau récapitulant la structure des données brutes.
Un individu est caractérisé par une variable Y (le temps de résidence). A chaque individu est associé 5 profils d'énergies calculés par MM-GBSA. Le chiffre 7501 correspond au nombre de structures (*snapshots*) pour chaque simulation.

Dans le **Tableau 9**, Y représente la variable à expliquer par le modèle QSKR, soit le temps de résidence. X représente les variables explicatives, c'est à dire les variables décrivant chacun des 110 simulations (individus), communément appelées des descripteurs. Cependant, en l'état, nos variables X ne peuvent être considérées comme des descripteurs car elles sont le résultat de la concaténation des valeurs des 5 profils d'énergie. Or, un profil d'énergie est une

suite de valeurs représentant l'évolution de l'énergie au cours du temps. En statistique, le profil d'énergie est une série temporelle. On ne peut pas appliquer un algorithme d'apprentissage automatique sur une série de valeurs ordonnées dans le temps que l'on considère comme des variables explicatives. De plus, au vue du nombre de valeurs par profil (7501), le nombre total de descripteurs serait trop grand (7501×5) par rapport au nombre d'individu (110 simulations). D'autre part, le nombre de descripteurs corrélés et donc apportant une information redondante serait très important. De fait, les valeurs d'un terme énergétique au temps *t*, ont de fortes chances d'être corrélées à celles du temps *t + 1*. Nous pourrions envisager de supprimer les descripteurs corrélés. Cependant, cela reviendrait à ne garder que certains points du profil, et donc potentiellement de perdre des informations importantes telles que la pente de la courbe, la présence de motif particuliers, le nombre de minima, leur profondeur, etc. Or, lorsqu'on analyse un profil énergétique ce qui nous intéresse ce sont justement ces caractéristiques qui décrivent l'allure de la courbe sous tous ses angles et de façon pertinente. Nous avons donc cherché à calculer un ensemble de caractéristiques[5] à partir des profils d'énergie afin de les décrire le plus pertinemment possible et pouvoir par la suite appliquer des méthodes d'apprentissage sur ces variables descriptives X. Pour extraire un ensemble de caractéristiques d'une série temporelle automatiquement, nous avons utilisé le paquet *tsfresh* dans python disponible depuis 2016 (Christ et al., 2018). Il est important de souligner que jusqu'en 2016, il n'existait aucun paquet permettant d'extraire automatiquement des caractéristiques d'une série temporelle (Fulcher, 2018). Le programme Matlab, qui est un outil de référence pour le traitement du signal dans le domaine de l'électronique et des automates, n'a vu son premier outil d'extraction de caractéristiques *hctsa* n'être publié qu'en 2016 (Fulcher and Jones, 2017). 794 caractéristiques décrivant une série temporelle sont implémentées dans *tsfresh*. Une définition de ces caractéristiques est donnée dans le lien suivant : https://tsfresh.readthedocs.io/en/latest/text/list_of_features.html et on peut voir que l'ensemble de ces caractéristiques permettent de décrire la série temporelle de façon très pertinente et complète. Nous avons calculé l'ensemble de ces 794 variables sur chacun des 5 profils d'énergie de chacun des 110 individus. On obtient alors une matrice de données de dimension 110×794×5 soit 436700. Les 110 lignes correspondent à chacune de nos trajectoires (ou individus) labellisées par une réponse Y qui est ici le temps de résidence.

[5] On parle de *features* pour une série temporelle, soit caractéristiques en français plutôt que de descripteurs comme il est d'usage dans les modèles statistiques.

L'apprentissage sera supervisé par cette réponse Y. Chaque individu est donc décrit par 3970 colonnes (794×5) correspondant aux caractéristiques (variable X) extraites des 5 profils d'énergie. Nos données ont maintenant la forme appropriée pour appliquer les procédures classiques d'apprentissage automatique. La procédure suivie est décrite dans la **Figure 28**.

b) Sélection des caractéristiques

(1) Suppression des caractéristiques non-pertinentes

Le traitement et l'analyse statistique des données ont été réalisés avec le paquet caret (Kuhn, 2008) de R. La fonction *nearZeroVar* est appliquée à l'ensemble des 3970 caractéristiques afin de supprimer les caractéristiques ayant une variance nulle ou presque nulle. En effet, ces variables ne permettent pas de différencier les individus en fonction de leur temps de résidence. Une variable est définie comme présentant une variance presque nulle lorsqu'elle remplit deux conditions : 1) elle contient peu de valeurs uniques par rapport au nombre de valeurs totales ; le seuil utilisé est celui par défaut, soit moins de 10% de valeur unique comparé au nombre de valeur total et 2) le rapport de l'effectif de la valeur la plus commune par l'effectif de la seconde valeur la plus commune est supérieur à 19 (95/5). 468 variables ont été supprimées.

Les 3502 variables restantes sont ensuite soumises à la fonction *findCorrelation* qui identifie les variables corrélées avec un coefficient de corrélation (en valeur absolue) supérieur à 0.8. Lorsque deux variables corrélées sont identifiées, il faut décider laquelle supprimer. Pour cela, chacune des deux variables est comparée par rapport à toutes les autres variables X de façon à supprimer la plus corrélée aux autres variables X. La variable qui présente la corrélation moyenne (en valeur absolue) la plus élevée avec les autres variables est supprimée. A l'issue de cette étape il reste 1264 variables X, ce qui représentent 2238 variables supprimées. Les 1264 caractéristiques ont été centré-réduites.

(2) Sélection des caractéristiques pertinentes

L'une des étapes cruciales dans la construction d'un modèle statistique est la sélection de variables explicatives X pertinentes. Les méthodes pour la sélection de variables explicatives X pertinentes sont nombreuses et peuvent être classées en deux catégories, les méthodes de filtre (*filter methods*) et les méthodes *wrapper*. Dans une méthode de filtre, les variables X sont sélectionnées sur la base de tests statistiques évaluant la corrélation de la variable X avec la réponse Y. Dans les méthodes *wrapper*, plusieurs modèles statistiques sont construits à

partir de plusieurs jeux de variables X différents. En fonction de la performance du modèle, on décide de garder ou non certaines variables X. C'est ce type de méthode *wrapper* qui a été appliquée ici, plus particulièrement le *recursive feature elimination* (rfe, *rfe function* dans caret), et la performance du modèle a été définie comme la capacité des variables X sélectionnées à expliquer Y. La procédure est la suivante. Un modèle est construit avec toutes les variables X. La performance du modèle ainsi que l'importance de chacune des variables X sont calculées. Les variables X sont triées par ordre décroissant de leur importance. Les *n* variables X les plus importantes (*n* choisi par l'utilisateur) sont sélectionnés. Pour chaque sous-ensemble de *n* variables X les plus importantes (*1, 2, [...], n-1, n*), un modèle est construit. La performance de chacun de ces modèles est calculée et gardée en mémoire. L'ensemble du processus est répété *k* fois. A l'issue de la procédure, les performances de tous les modèles générés ainsi que leurs jeux de variables X respectives sont analysés. L'algorithme fournit finalement la liste optimale de variables X, i.e. de caractéristiques, à conserver. La variable *n* a été fixée à 100, ce qui est suffisant car la précision converge vers un plateau à partir de *n = 40* (**Figure 29**). La précision est définie comme le ratio de l'effectif des individus dont la classe (Y) a été bien prédite sur celui dont la classe (Y) a été mal prédite. De même, une valeur de 5 pour le paramètre *k* a été jugée suffisante car les résultats convergent (**Figure 29**).

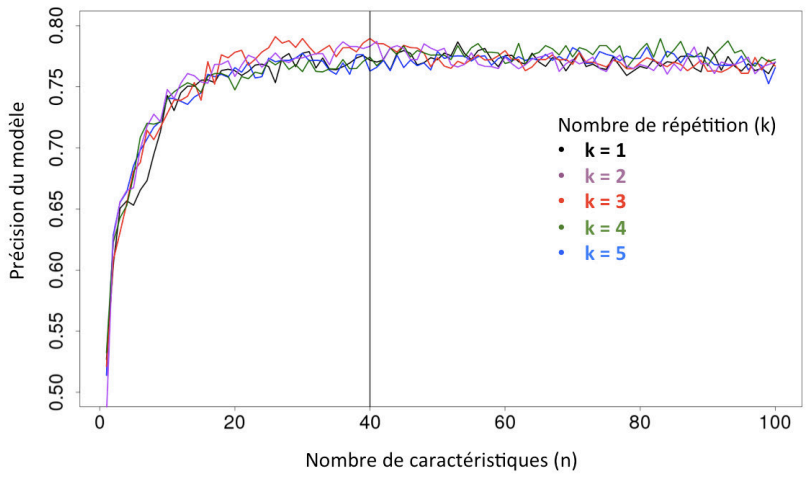


Figure 29 : Modèles QSKR de forêts aléatoires créés durant l'algorithme rfe de sélection des caractéristiques pertinentes.

Enfin, il reste à définir quel algorithme d'apprentissage à choisir pour cette procédure de sélection des caractéristiques pertinentes. Son choix dépend i) de la nature des variables explicatives X et de la réponse Y (qualitative, quantitative continue ou discrète, etc.) et ii) de la nature de la relation entre les variables X et la variable Y (linéaire, logistique, etc.). Le temps

de résidence (Y) est une variable qualitative à 6 classes (<1.4 min, 7 min, 14 min, 57 min, 1624min et 1944 min). La relation entre le temps de résidence et les différentes caractéristiques énergétiques extraites des profils d'énergie n'est, *a priori,* pas simple et supposée complexe. Une méthode dite non-linéaire et complexe est alors nécessaire. Notre choix s'est porté sur les Forêts Aléatoires (RF pour *Random Forest*). Le détail de la méthode se trouve dans le chapitre V partie V.K**.** Les RF présentent de nombreux avantages. Comparées à d'autres méthodes, la préparation des données est simple, l'algorithme gère aussi bien des données binaires, catégoriques, numériques sans aucun prérequis sur les données tel que la normalité, l'absence de multi-colinéarité, etc. La méthode est rapide et s'applique à toutes situations, aussi bien pour un modèle de régression qu'un modèle de classification. De plus comme chaque arbre est un modèle statistique, et que le résultat de la prédiction est la combinaison de l'ensemble des arbres, soit de la forêt, le RF réduit le phénomène de sur-apprentissage. L'inconvénient souvent évoqué du RF est son manque d'interprétabilité contrairement aux méthodes dans lesquelles le modèle peut s'écrire sous forme d'une équation de Y en fonction de X et son manque de reproductibilité. Pour l'objet de notre étude, cela ne représente pas une limite car le but est de comprendre quels sont les termes énergétiques contribuant le plus à expliquer le temps de résidence pour notre jeu de donnée. Lors de l'application de la méthode RF, il est important de ré-étirer le processus de construction du modèle plusieurs fois (comme nous l'avons fait avec le paramètre $k = 5$) car il n'y a aucune garantie de construire exactement le même modèle à chaque itération. De fait, les RF utilisent un algorithme d'échantillonnage *bootstrap* qui permet de générer diffèrent arbres sur un même jeu de donnée (cf. Forêts aléatoires).

L'algorithme *rfe* nous suggère de garder un ensemble de 26 caractéristiques. Vue le nombre d'individu (110), le nombre de caractéristiques est correcte car il est généralement admis que le nombre d'individu par variables descriptives X doit être d'environ 5 (Topliss and Costello, 1972). Un trop grand nombre de variables descriptives X par rapport au nombre d'individus nous expose au risque du sur-apprentissage du modèle. Celui-ci aura alors une faible capacité prédictive car il ne décrira pas seulement les règles générales de la relation entre le temps de résidence et les caractéristiques X des profils, mais aussi les particularités de chaque individu. Le ratio du nombre d'individus sur le nombre de caractéristiques de notre modèle est de $110/26 = 4.2$. Dans le cadre de cette étude, l'objectif n'est pas de construire un modèle prédictif mais de comprendre la relation entre Y (le temps de résidence) et les caractéristiques X de profils pour le jeu de donnée considéré. Le facteur 4.2 est donc tout à fait

correct. Parmi les 26 caractéristiques sélectionnées, 15 ont été extraites des profils de la contribution non-polaire de l'énergie de solvatation (E_{NP}), 7 des profils de l'énergie totale (G_{total}) et 4 des profils de l'énergie de van der Waals (E_{VDW}).

4. Construction du modèle.

Le modèle est construit à partir des 26 caractéristiques sélectionnées par l'algorithme ***rfe***. Afin d'assurer la robustesse du modèle, c'est à dire sa capacité à ne pas être impactée par une petite modification dans les données, une méthode d'échantillonnage, dite validation croisée de type *k-fold,* a été utilisée pour la construction du modèle. Notre jeu de données est divisé en 3 groupes (*k = 3*), et itérativement le modèle est construit sur *k-1* groupes et prédit sur le $k^{ème}$ groupe restant. Ainsi, 3 modèles sont produits à l'issue de la procédure de validation croisée. Cette dernière est répétée 5 fois, ce qui fait un total de 15 modèles. A l'issue de ce processus, le modèle qui présente la meilleure capacité explicative et prédictive est sélectionné. Ce modèle est alors entrainé sur l'ensemble du jeu de donnée et on obtient le modèle final.

Cependant, dans tous les algorithmes d'apprentissage automatique, il y a des hyper-paramètres à ajuster afin d'obtenir la meilleure performance du modèle. Tandis que les paramètres d'un modèle sont ceux calculés durant l'apprentissage (tels que la pente et l'ordonnée à l'origine pour une régression linéaire), les hyper-paramètres doivent être déterminés par l'utilisateur avant l'apprentissage. Dans le cas du RF, les hyper-paramètres généralement ajustés sont le nombre d'arbres à considérer (*ntree*) et le nombre de variables (*mtry*) à tester à chaque nœud d'un arbre, tandis que les paramètres du RF sont les variables à considérer à chaque nœud ainsi que la valeur seuil utilisée à chaque nœud pour diviser les données (cf. V.K). Nous avons testé les valeurs de 500 (valeur par défaut), 1000, 1500 et 2000 pour *ntree* et les 22 valeurs de 5 à 26 pour *mtry*. Ainsi, un total de 88 modèles finaux (4*(26-5+1)) ont été produits. Le modèle final retenu est celui ayant la meilleure précision (0.82), obtenue avec un nombre de 1000 arbres (*ntree*) et un nombre de variables testées à chaque nœud de 25 (*mtry*) (**Figure 30**).

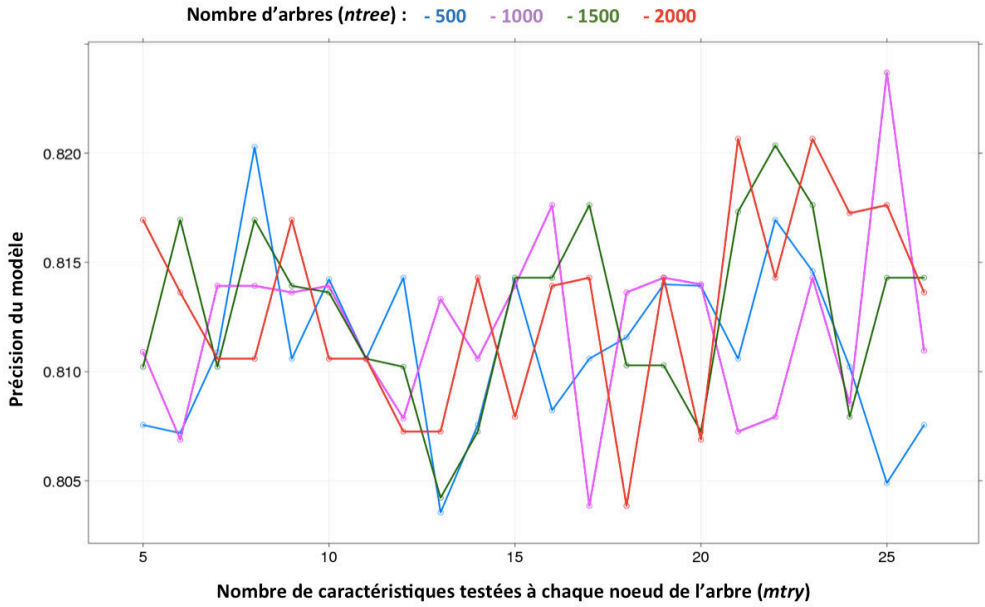


Figure 30 : Précision des modèles QSKR générés lors de la procédure de validation croisée répétée.

5. Résultats et discussion

a) Description du modèle

Nous avons ensuite analysé l'importance de ces 26 variables X dans le modèle (**Figure 31**). Rappelons que ces variables X correspondent à des caractéristiques qui ont été calculées chacune sur l'un des profils énergétiques et sélectionnées par l'algorithme de sélection des variables pertinentes. Plus une caractéristique contribue à séparer les individus en fonction de leur temps de résidence, plus elle est considérée importante. Donnons d'abord une explication qualitative de la signification de ces 26 caractéristiques. Ils appartiennent à 5 catégories différentes :

- 20 caractéristiques, dont la nomenclature générique est *Eterme*_dft_*nombre,* sont les coefficients issus de la transformée de Fourier discrète appliquée aux profils (dft pour *Discrete Fourier Transform*). Une explication détaillée de la DFT et de ses coefficients est donnée dans la partie V.L. Les coefficients traduisent la contribution d'une onde sinusoïdale d'une fréquence donnée au signal, c'est à dire au profil d'énergie. La fréquence est indiquée par la partie *nombre* de la nomenclature générique. Les coefficients de basse fréquence décrivent les fluctuations de large amplitude tandis que ceux de haute fréquence décrivent celles de basse amplitude. Ainsi, lorsque la valeur d'un coefficient de basse fréquence est élevée cela signifie que de larges fluctuations

(de l'ordre de l'onde sinusoïdale de la fréquence considérée) sont présentes dans le profil d'énergie.

- 2 caractéristiques, dont la nomenclature générique est *Eterme*_allsr_*nombre_paramètre*, correspondent à la valeur d'un des *paramètres* d'un modèle de régression linéaire des moindres carrés réalisée sur le profil d'énergie. L'une de ces caractéristiques correspond à la pente de la droite (*slope*), le second au coefficient de corrélation (*rvalue*). Cette régression n'est pas réalisée sur les données brutes du profil. Les données ont été préalablement agrégées en utilisant comme fonction d'agrégation la variance (allsr pour *Aggregated Linear Least-Squares Regression*). Les valeurs du profil sont découpées par tranche de *n* valeurs (*n* = *nombre*) et la variance est calculée sur chacune de ces tranches. La régression est réalisée sur ce profil de variances.
- 1 caractéristique, dont la nomenclature générique est *Eterme*_lsbm, correspond à la longueur de la plus longue suite de valeurs consécutives du profil, supérieures à la moyenne de l'ensemble des valeurs.
- 2 caractéristiques, dont la nomenclature générique est *Eterme*_rb_*nombre*sigma, correspondent à la fraction des valeurs du profil qui sont à plus de *n* écart-type de la moyenne (*n*=*nombre*).
- 1 caractéristique, dont la nomenclature générique est *Eterme*_varautocc, correspond à la variance calculée sur les coefficients d'autocorrélation du profil. Un coefficient d'autocorrélation est le coefficient de corrélation calculé entre deux groupes de points d'une série temporelle espacé d'un intervalle de temps Δt. L'autocorrélation permet de détecter des motifs qui se répètent au cours du temps par exemple. Plusieurs coefficients sont calculés en faisant varier Δt. La caractéristique correspond à la variance de l'ensemble de ces coefficients d'autocorrélation.

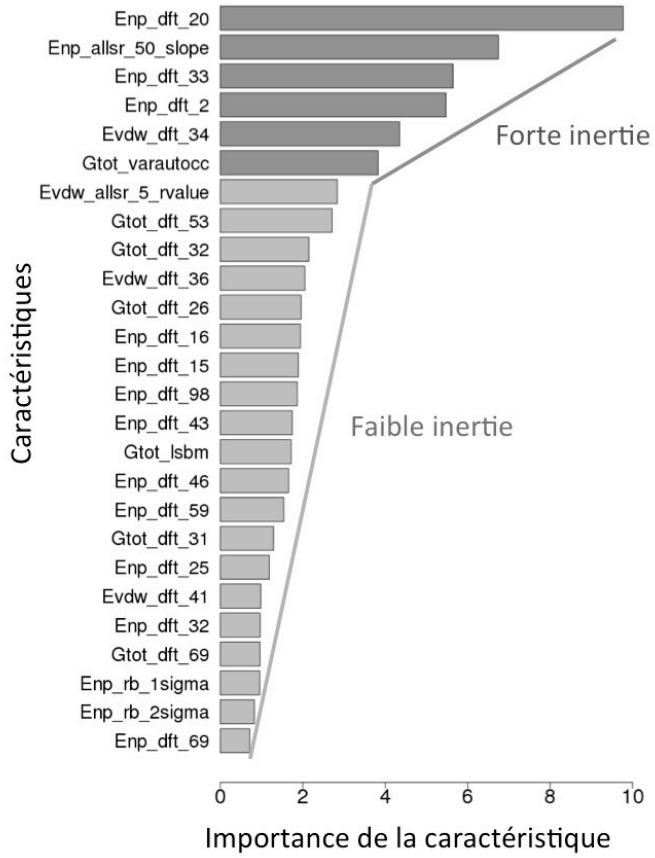


Figure 31 : Contribution des caractéristiques au modèle QSKR final.
Les noms des caractéristiques suivent la nomenclature suivante : *TermeEnergetique_TypeDeCaractéristique.* On peut ainsi voir sur quel profil énergétique a été calculé chacune des caractéristiques. Le sens de chaque caractéristique est expliqué ci-dessus. Les caractéristiques qui contribuent le plus au modèle sont en gris foncé et ceux qui contribuent le moins sont en gris clair. L'importance de la caractéristique est calculée suivant le critère de Gini (*Gini importance* ou *mean decrease impurity*) qui est une mesure de la capacité de la variable à diminuer l'impureté du nœud.

b) Relation structure – énergie - cinétique

Parmi les 6 variables contribuant le plus à séparer les molécules en fonction de leur temps de résidence, les 4 premières sont des caractéristiques extraites du profil de la contribution non-polaire de l'énergie de solvatation (**Figure 31**). L'énergie de van der Waals (E_{VDW}) et l'enthalpie (G_{total}) sont en 5ème et 6ème positions. Parmi les 26 caractéristiques pertinentes sélectionnées, 15 sont des caractéristiques du profil de la contribution non-polaire de l'énergie de solvatation, correspondant pour la majorité (12/15) à des coefficients de la transformée de Fourier discrète (DFT). Si l'on considère, par exemple, la caractéristique qui contribue le plus à moduler le temps de résidence, elle correspond à un coefficient de la DFT de fréquence 20 calculée sur le profil d'énergie de solvatation non-polaire (*Enp_dft_20*). Cela

signifie que l'amplitude (dans l'espace des fréquences) de la composante sinusoïdale de fréquence 20, issue de la décomposition du profil de la contribution non-polaire de l'énergie de solvatation par DFT, est corrélée au temps de résidence. Autrement dit, l'amplitude de la composante sinusoïdale de fréquence 20 du profil de la contribution non-polaire de l'énergie de solvatation calculée par DFT est significativement différente suivant que le profil ait été calculé sur une molécule ayant un court ou un long temps de résidence. De même, la seconde caractéristique, dont la définition a été donnée ci-dessus, indique que la pente de la droite modélisant la dispersion des valeurs du profil de la contribution non-polaire de l'énergie de solvatation au cours du temps, est corrélée au temps de résidence.

Ces résultats indiquent que le temps de résidence de la série d'inhibiteurs étudiée est modulé majoritairement par la contribution non-polaire de leur énergie de solvatation. En accord avec ce résultat, sur la base de l'analyse des interactions protéine-ligand au cours du processus de dissociation, nous avions mis en évidence que les contacts hydrophobes jouent un rôle crucial dans l'augmentation du temps de résidence (partie A). De fait, nous avions montré que les contacts hydrophobes avec les résidus de la poche allostéristique et le *gatekeeper* (Phe97) sont indispensables pour qu'un inhibiteur présente un temps de résidence moyen (1.4 < temps de résidence $\leq$ 57min). Pour les composés de long temps de résidence (> 57min), les interactions hydrophobes avec les résidus de la boucle P (*P-loop*) et de la *front pocket* jouent un rôle primordial. Toujours sur la base de l'analyse des interactions protéine-ligand, nous avions également suggéré que ces interactions hydrophobes limitent l'exposition du ligand au solvant ce qui a pour effet de stabiliser le complexe protéine-ligand, et donc d'augmenter le temps de résidence. Les résultats obtenus dans cette étude nous confortent dans cette hypothèse puisque d'une part, les caractéristiques extraites du profil de la contribution non-polaire de l'énergie de solvatation sont les plus corrélées au temps de résidence (**Figure 31**) et d'autre part, le temps de résidence de ces molécules n'est pas corrélé à leur lipophilicité estimé par le cLogP (**Tableau 8**).

Ainsi les évènements moléculaires identifiés comme étant déterminants à la cinétique de dissociation se répercutent sur les profils de la contribution non-polaire de l'énergie de solvatation qui présentent un signal différent suivant le temps de résidence du composé. Lorsqu'on compare visuellement les profils de la contribution non-polaire de l'énergie de solvatation de composés de différents temps de résidence, la différence de signal n'est pas triviale. C'est l'extraction de caractéristiques particulières du profil, majoritairement les coefficients de la DFT (**Figure 31**), qui nous a permis de mettre en évidence la relation entre

la contribution non-polaire de l'énergie de solvatation et le temps de résidence. L'ensemble de ces résultats est également en accord avec la littérature, détaillée dans l'introduction, relatant l'importance de l'effet du solvant sur la modulation du k_{off} et donc du temps de résidence.
Plusieurs études ont montré que moduler la barrière énergétique de solvatation qu'un ligand doit franchir au cours de sa dissociation, permet de moduler son temps de résidence (Schuetz et al., 2018)(Liu et al., 2010). Plus la barrière est élevée, plus le temps de résidence augmente. Cette barrière énergétique augmente avec l'hydrophobicité du ligand. Une étude statistique réalisée sur plus de 2000 médicaments provenant de la littérature et de la base de données du groupe Pfizer, a montré que les médicaments de lipophilicité (cLogP) élevée avaient tendance à avoir un plus long temps de résidence (Miller et al., 2012). Contrairement à notre étude, ce type d'analyse, tout comme les modèles QSAR et QSKR classiquement construits qui se basent sur la description d'un conformation des molécules et ou de la cible, sont restrictifs. De fait, il n'y a pas de garantie que deux molécules ayant des caractéristiques physico-chimiques similaires aient des profils énergétiques similaires (E_{VDW}, E_{EEL}, E_{GB}, E_{NP}) au cours de la dissociation et donc un même temps de résidence. Cela dépend aussi du chemin emprunté, de l'orientation et de la flexibilité du ligand au cours de sa dissociation, des contacts protéine-ligand établis au cours de sa sortie etc. La cinétique de liaison est avant tout un processus dynamique et hors équilibre. Dans ce contexte, les profils d'énergie d'interaction protéine-ligand au cours de la dissociation du ligand constituent des données plus pertinentes pour comprendre et prédire le temps de résidence. Cependant, il est important de souligner que les valeurs des profils d'énergie ne doivent pas être considérées en valeur absolue mais uniquement de façon relative en vue de comparer les profils entre eux. En effet, ces valeurs énergétiques sont loin de correspondre à des valeurs calculées à l'équilibre (cf. V.G.1) car : i) chaque valeur du profil correspond à l'énergie d'une seule conformation (1 *snapshot*) de la DM et ii) la DM n'est pas à l'équilibre car il s'agit d'une DM biaisée. Néanmoins, dans la littérature, le profil d'énergie (G_{total}) calculé par la méthode MM-GBSA sur une dynamique moléculaire biaisée (*umbrella sampling*) a déjà été utilisé pour approximer le potentiel de force moyen (PMF) (Sun et al., 2015)(Bai et al., 2013) calculé par la méthode WHAM ((Kumar et al., 1992).

6. Conclusion et perspectives

Dans cette étude, nous avons présenté un protocole général permettant d'extraire des informations pertinentes de séries temporelles que l'on peut analyser par la suite grâce à des

algorithmes d'apprentissage automatique afin de décrire et ou de prédire une propriété Y donnée. Ce protocole à été appliqué pour réaliser une étude QSKR afin de rechercher et mettre en évidence un potentiel lien entre le temps de résidence de 10 inhibiteurs de CDK8 et leurs profils énergétiques des différents termes d'interaction (E_{VDW}, E_{EEL}, E_{GB}, E_{NP}, G_{total}) calculés par la méthode MM-GBSA.

Chacun de ces 5 profils énergétiques a été décrit par 794 caractéristiques qui représentent soit des descripteurs simples du profil tels que la médiane des valeurs du profil ou leur moyenne, ou bien des paramètres plus complexes qui décrivent des motifs temporels tels que les coefficients de la transformée de Fourier. A l'issue de la procédure de sélection des caractéristiques, 26 caractéristiques ont été retenues et considérées comme étant pertinentes pour expliquer le temps résidence. Un modèle QSKR a été développé sur ces 26 caractéristiques en utilisant un algorithme de forêt aléatoire et une précision de 0.82 en apprentissage a été obtenue.

Parmi les 26 caractéristiques composant le modèle, 15 sont des caractéristiques décrivant le profil de la contribution non-polaire de l'énergie de solvatation, dont 4 sont les plus importantes du modèle (**Figure 31**). Ainsi, la contribution non-polaire de l'énergie de solvatation est le terme énergétique qui contribue le plus à expliquer, c'est à dire à moduler, le temps de résidence. Notre modèle QSKR suggère donc que le temps de résidence de cette série congénère d'inhibiteurs de CDK8 est modulé majoritairement par la barrière énergétique de la contribution non-polaire de l'énergie de solvatation. L'augmentation de cette barrière augmente le temps de résidence. Ces résultats sont en accord avec les résultats issus de l'analyse des interactions protéine-ligand au cours du processus de dissociation, où nous avions mis en évidence le rôle clé des contacts hydrophobes dans l'augmentation du temps de résidence (cf. III.A.1). L'ensemble de ces résultats est également en accord avec la littérature, détaillée dans l'introduction, relatant l'importance de l'effet du solvant et donc, la solvatation du ligand, dans la modulation du temps de résidence.

Par la suite, on pourra envisager d'utiliser un tel modèle QSKR pour prédire le temps de résidence d'un inhibiteur de CDK8 à partir de ses profils énergétiques calculés par la méthode MM-GBSA à partir de sa trajectoire du processus de dissociation. Il serait intéressant de développer un tel modèle QSKR sur un jeu de données plus grand et plus divers incluant des inhibiteurs de plusieurs protéines différentes pour lesquelles nous disposons de la mesure du temps de résidence. Un tel modèle nous permettrait 1) de comprendre la relation structure-

énergie-cinétique et d'en extraire des règles générales et 2) de prédire le temps de résidence d'une molécule à partir de ses profils énergétiques issus de la méthode MM-GBSA.

Autre point intéressant, les coefficients de la transformée de Fourier discrète (DFT) constituent 77% (20/26) des 26 caractéristiques composant le modèle QSKR développé. Parmi les 20 coefficients de la DFT, 12 d'entre eux décrivent le profil de la contribution non-polaire de l'énergie de solvatation. La transformée de Fourier discrète, ainsi que la transformée en ondelette sont des méthodes de réduction de données très utilisées en informatique pour compresser des données. Elles permettent de réduire les données initiales en un ensemble de coefficients (fichier compressé) à partir desquels les données initiales peuvent être recouvertes. Calculés sur une série temporelle, ces coefficients sont un résumé interprétable de faible dimension capable de capturer des schémas temporels compliqués qui se déroulent à différentes échelles de temps (Fulcher, 2018). Dans le domaine biomédical, plusieurs modèles statistiques utilisant la transformée de Fourier ou en ondelette ont été développés pour détecter des anomalies sur des signaux d'électromyogramme ou d'électroencéphalogramme et ainsi prédire respectivement une insuffisance cardiaque (Adam et al., 2018) et une crise d'épilepsie (Samiee et al., 2015). Il n'est donc pas étonnant de voir que les coefficients de la transformée de Fourier sont les caractéristiques de plus fortes inerties dans notre modèle.

Dans notre approche, la mesure de la (di)similarité entre les séries temporelles est basée sur les caractéristiques calculées à partir des séries temporelles (*Feature-based dissimilarity*). Il existe une autre approche qui consiste à mesurer une distance entre les valeurs de deux séries temporelles (*Time-domain dissimilarity*). A cet effet, plusieurs algorithmes ont été développés dont le populaire DTW (*Dynamic Time Warping*) qui fournit la distance entre deux séries temporelles après avoir trouver le meilleur alignement entre celles-ci en prenant en compte un décalage possible des valeurs dans le temps (Besse et al., 2015). Cependant, le calcul de distance entre des séries temporelles est coûteux en temps de calcul et ne permet pas de traduire la (di)similarité entre deux profils en terme de présence (ou non) de motifs particuliers.

Ainsi, de façon générale nous pensons que ce protocole d'analyse de séries temporelles est une approche prometteuse qui sera de plus en plus utilisée pour extraire de la connaissance à partir d'un ensemble de séries temporelles (notamment des trajectoires de DM) supervisées (régression, classification) ou non (*clustering*) par une propriété Y. La DM est une méthode de plus en plus utilisée dans les projets de recherche et de développement de nouveaux

médicaments. Les forts progrès réalisés ces dernières années dans les technologies informatiques et algorithmiques, notamment les cartes GPU y ont fortement contribué. La DM est utilisée afin de décrire l'interaction entre une protéine et un ligand, au sein de son site de liaison, mais également lors du processus d'association et de dissociation. Elle prend en compte la flexibilité des deux partenaires, l'effet du solvant et décrit l'ensemble du processus de liaison. Ces informations sont d'une grande valeur et ne sont pas prises en compte (ou très peu), dans les modèles prédictifs *in silico* tels que les modèles pharmacophoriques et les modèles QSAR utilisés pour filtrer et sélectionner les molécules potentiellement actives. De fait, ces modèles ont été développés sur la base de structures statiques (ou peu flexible) de la protéine et du ligand. Ainsi, nous avons d'une part une accumulation exponentielle de données issue des trajectoires de DM grâce aux progrès technologiques et d'autre part, des modèles de criblage (QSAR, pharmacophore) pas assez précis. Parallèlement, nous assistons à un développement croissant des méthodes d'apprentissage automatique dont le but est justement de traiter, d'analyser et de créer de la valeur à partir d'une grande quantité de données. Il est donc intéressant de combiner les méthodes d'apprentissage automatique aux méthodes de DM. C'est en 2008 que le cabinet d'études Gartner introduit pour la première fois l'expression *Big Data* (Richard, 2012) et c'est cette même année qu'une première étude combine la DM à une méthode d'apprentissage automatique afin d'annoter les protéines par leur fonction (Glazer et al., 2008). Puis plusieurs études ont combiné les deux approches pour répondre à diverses problématiques telles que l'identification de transitions de phase en science des matériaux (Li et al., 2018). Plus en rapport avec la conception de médicament, l'équipe de Denis Fourches, calcule des descripteurs en prenant la moyenne sur un ensemble de conformations issues de trajectoires de DM au lieu de les calculer à partir d'une structure unique comme il est d'usage de le faire dans les modèles QSARs. Il montre que les caractéristiques moyennées issues de la DM sont faiblement corrélées aux caractéristiques non moyennées habituellement calculés. De plus, ils permettent une meilleur discrimination entre les molécules actives et les molécules non actives (Ash and Fourches, 2017). Enfin les constantes cinétiques sont aujourd'hui considérées comme des paramètres importants à prendre en compte pour la sélection de candidat-médicaments. Leur étude par DM suppose de prendre en compte l'évolution dans le temps de l'ensemble des évènements se produisant durant le processus de liaison. Le protocole développé dans cette étude permet justement d'analyser un grand nombre de séries temporelles et d'en extraire des informations pertinentes.

C. Application de la méthode de prédiction du temps de résidence à un jeu de données privé

1. Préambule et objectif

J'ai eu l'occasion de me rendre à l'Institut de Rechercher Servier (IdRS), situé à Croissy-sur-Seine, afin de tester notre méthode de prédiction du temps de résidence sur un jeu de données privé appartenant à l'IdRS. Ce jeu de données est associé à l'un des projets de recherche au sein de cet institut. La cible thérapeutique de ce jeu de données est une protéine kinase impliquée dans les désordres cardiovasculaires et le cancer. Pour des raisons de confidentialité, nous ne présenterons ni la cible, ni les structures du jeu de données. L'objectif est de vérifier que la méthode développée nous permet d'obtenir un classement correct des composés selon leur temps de résidence sur un autre jeu de données que celui sur lequel elle a été développée. Cela a permis également de tester la faisabilité de l'installation et de l'utilisation de l'outil dans un cadre extérieur, ici l'IdRS, et d'évaluer les performances de l'outil sur les moyens de calcul locaux. Particulièrement, nous souhaitions avoir une estimation du temps de calcul des simulations de dynamique moléculaire (DM) ainsi que de leurs analyses. *In fine*, même si la cible est toujours une protéine kinase, l'objectif est de savoir si la précision de la méthode ainsi que sa performance permettraient de l'intégrer comme un outil utilisable en routine dans un cadre industriel pour fournir un classement prédictif des chefs de file en phase d'optimisation.

2. Matériels et méthodes

Pour cette étude, nous disposons de 21 composés pour lesquels le temps de résidence expérimental a été mesuré par la méthode de fluorescence *stopped-flow.* Le temps de résidence de ces composés varie de 30 s à 89 min**.** Pour 11 d'entre eux, nous disposons également des paramètres thermodynamiques (l'enthalpie et l'entropie) mesurés par titrage calorimétrique isotherme. Parmi les 21 composés, cinq ont été co-cristallisés avec la protéine. L'une de ces structures cristallographiques a été utilisée comme modèle de la protéine car elle est entièrement résolue à une résolution de 3**.**05 Å. Avec le logiciel MOE[6], les 20 composés ont été replacés dans cette structure en alignant leur *scaffold* commun. Un soin particulier a été

[6] Molecular Operating Environment (MOE) software version 2016.0802; Chemical Computing Group Inc. http://www.chemcomp.com.

apporté à la vérification des interactions protéine-ligand dans chacune de ces 20 structures. De fait, nos précédents essais sur CDK8 ont montré que la position initiale du ligand dans le site de liaison conditionne fortement la pertinence du résultat de la simulation.

Les 21 systèmes ont été préparés, minimisés et équilibrés en suivant la même procédure et en utilisant les mêmes outils que ceux présentés dans la partie III.A.1. Brièvement, les ligands ont été préparés avec l'outil Antechamber et le champ de forces GAFF sous leur forme non-chargée. A l'aide du module tleap, le système est protoné, puis solvaté en utilisant un modèle d'eau TIP3P et placé dans une boite d'eau cubique dont le bord est à au moins 10 Å de tout atome du soluté. Des ions Cl^- et Na^+ ont été ajoutés afin de neutraliser le système. L'énergie du système est minimisée en quatre étapes en minimisant d'abord le solvant, puis les chaines latérales de la protéine, puis le soluté et enfin l'ensemble du système. Le système est ensuite thermalisé dans le milieu NVT avec une augmentation graduelle de la température de 0 à 300K en imposant une contrainte harmonique sur le soluté pour éviter toute distorsion de la protéine. Enfin, le système est équilibré durant 10 ns dans le milieu NPT avec une diminution graduelle de la contrainte harmonique jusqu'à son annulation.

Le processus de dissociation a été simulé en utilisant le même protocole que dans la partie III.A.1. Pour le calcul du RMSD, l'alignement se fait sur le même ensemble de résidus que ceux sélectionnés sur CDK8 (III.A.1). Nous avons déterminé cet ensemble en alignant la séquence de CDK8 sur celle de la protéine kinase de cette étude de façon à identifier les résidus homologues. Le RMSD, calculé sur les atomes lourds du ligand, augmente de 0,001 à 38,001 Å avec un pas de 0,01 Å tous les 0,2 ps. Le RMSD final est de 38,001 Å et non de 75,001 Å comme présenté dans la partie III.A.1 car la cible est plus petite ici. Il s'agit d'une protéine kinase seule (un monomère) et non d'un hétérodimère comme dans le cas de CDK8 (complexée à la cycline C). Ainsi, un RMSD de 38,001 Å est suffisant pour que le ligand se situe à une distance de plus 6 Å de tout atome de la protéine. Enfin la valeur de la constante de force du potentiel harmonique utilisée est la même, soit 80 $kcal.mol^{-1}$. Chaque composé a été simulé 11 fois (11 réplicas). En ayant à disposition 160 cpu, nous avons pu simuler 11 réplicas de 5 composés par jour. Ce débit convient à l'usage industriel. Chaque simulation est lancée sur 8 CPUs et dure 8,6 heures. Comme dans la partie 1, le temps de résidence prédit (RT_{score}) est déterminé en prenant la moyenne sur les 11 réplicas.

3. Résultats

Le score (RT_{score}) est utilisé pour estimer le temps de résidence des 21 composés en vue de fournir un classement prédictif de ces composés. Il est également calculé sur la moyenne des 11 réplicas. Bien que le calcul de RT_{score} ait été détaillé dans la partie 1, nous allons ici repréciser certains points qui nous semblent importants. RT_{score} est calculé en prenant l'aire sous la courbe du potentiel harmonique ajouté pour franchir les barrières énergétiques ($V_{restraint\_push}$). Intégrer $V_{restraint\_push}$ sur le temps conduit à une quantité (notée RT_{score}) comparable à une action en physique exprimée en [Energie].[Temps] qui reflète ici «l'action» globale fournie pour extraire le ligand du site de liaison (partie 1, équation 3). La difficulté dans la prédiction du temps de résidence est de mettre en place une méthode applicable aussi bien à un processus cinétique simple en 1 étape qu'à un processus complexe comprenant plusieurs étapes cinétiques. Or, la théorie des états de transition est fondée sur l'équation d'Arrhenius qui relie l'énergie d'activation à la constante cinétique (k_{off}) pour un processus cinétique en une étape. Pour un processus plus complexe, il n'existe aucune théorie ou règle sur laquelle se baser pour dériver le k_{off} à partir de plusieurs barrières énergétiques. De plus, les barrières énergétiques peuvent être multiples, plus ou moins importantes et dépendentes du degré de résolution du profil d'énergie calculé par des méthodes numériques. L'avantage de RT_{score} est que sa valeur englobe toutes les barrières d'énergie rencontrées tout au long du processus de dissociation ce qui rend la méthode applicable aux processus cinétiques complexes à plusieurs étapes.

Dans la partie III.A.1, $RT_{expérimental}$ est une variable qualitative à 6 classes {<1.4, 7, 14, 57, 1626, 1944 min} car nous ne disposions pas de valeur exacte du temps de résidence pour les inhibiteurs de CDK8 de faible temps de résidence. Cependant, pour le jeu de données privé, $RT_{expérimental}$ est bien une variable quantitative. Il convient alors de poser clairement la relation entre $RT_{expérimental}$ et RT_{score}. RT_{score} est l'énergie globale nécessaire pour dissocier le ligand de son site de liaison (III.A.1). Nous faisons l'hypothèse qu'il y a une relation linéaire entre RT_{score} et l'énergie d'activation (ΔG_{off}) définie dans l'équation d'Arrhenius en se ramenant donc à un processus à une étape, comme suit :

$$k_{off} = A.\, e\frac{-\Delta G_{off}}{B}$$

Equation d'Arrhenius

Avec B = R×T, R est la constante des gaz parfaits et T la température. Et A est le facteur de fréquence. En remplaçant ΔG_{off} par RT_{score}, on obtient :

$$k_{off} = A.e\frac{-RT_{score}}{B}$$

L'équation est linéarisée en prenant le logarithme népérien ; on obtient alors :

$$\ln(k_{off}) = \ln\left(A.e\frac{-RT_{score}}{B}\right)$$

$$\ln(k_{off}) = \ln(A) - \frac{RT_{score}}{B}$$

Or on a :

$$\ln(k_{off}) = \ln\left(\frac{1}{RT_{experimental}}\right) = -\ln(RT_{experimental})$$

On obtient alors :

$$\mathbf{\ln\left(RT_{experimental}\right) = -\ln(A) + \frac{RT_{score}}{B}}$$

Ainsi, pour tester la corrélation entre $RT_{expérimental}$ et RT_{score} nous prendrons le logarithme népérien de $RT_{experimental}$.

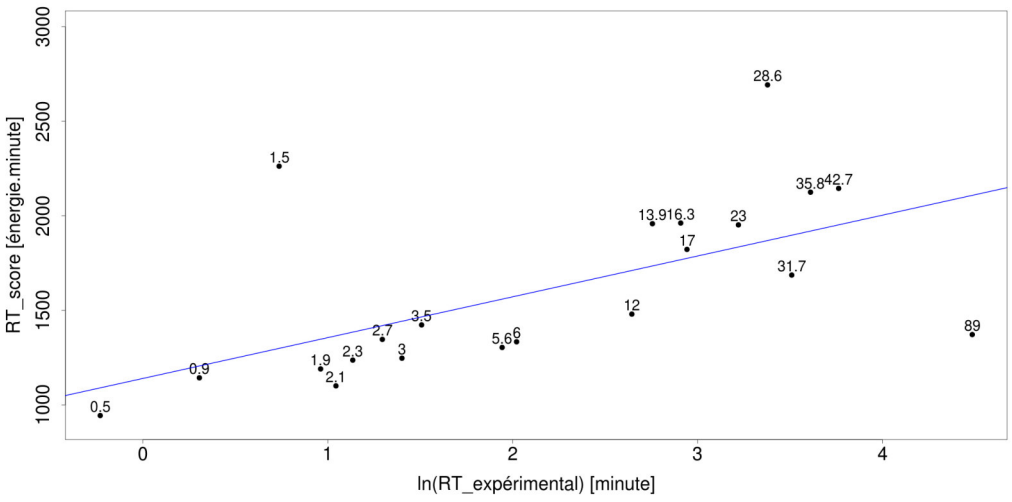
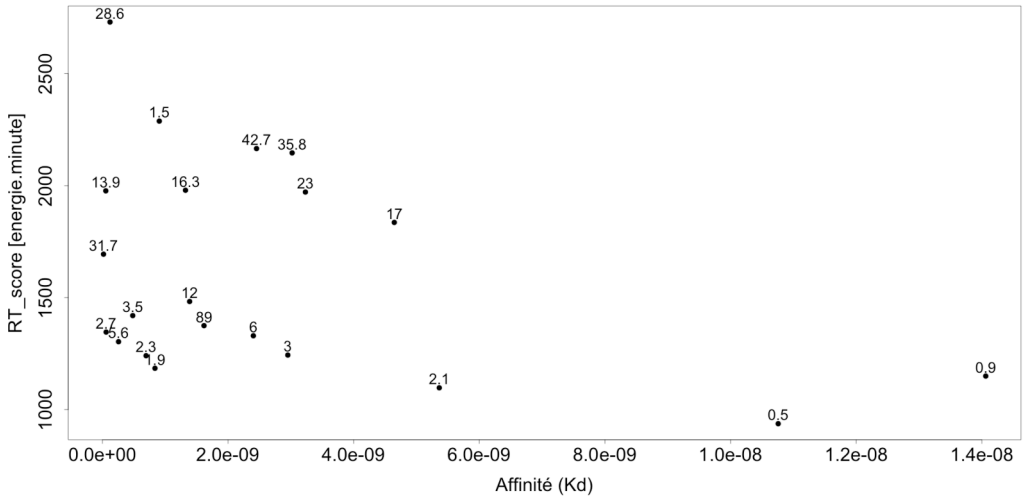


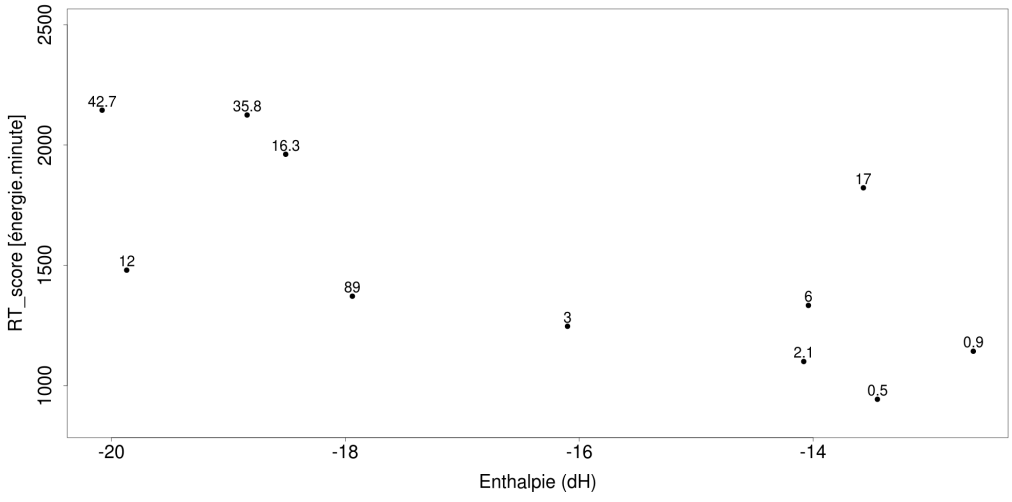
Figure 32 : RT_{score} en fonction du temps de résidence expérimental ($RT_{expérimental}$) obtenu pour les 21 composés.
Chaque point du graphique correspond à un composé et porte comme label le temps de résidence expérimental exprimé en minute.

Nous observons qu'il existe une corrélation positive entre le RT_{score} et le temps de résidence expérimental des 21 composés (**Figure** 32). Trois composés de temps de résidence expérimental 1.5, 28.6 et 89 minutes apparaissent comme des points aberrants. Le coefficient de détermination est de 0.40 en considérant l'ensemble des points et de 0.85 sans les trois points aberrants. Cela signifie que 85% de la variance observée au sein des 18 composés est expliquée par RT_{score}. Cela montre que RT_{score} est un bon prédicteur du temps de résidence. Comme pour 11 de ces 21 molécules, nous disposons également des paramètres thermodynamiques (enthalpie, entropie et affinité), nous décidons de vérifier s'il existe une corrélation entre chacun de ces paramètres thermodynamiques et RT_{score}.

(a)



(b)



(c)

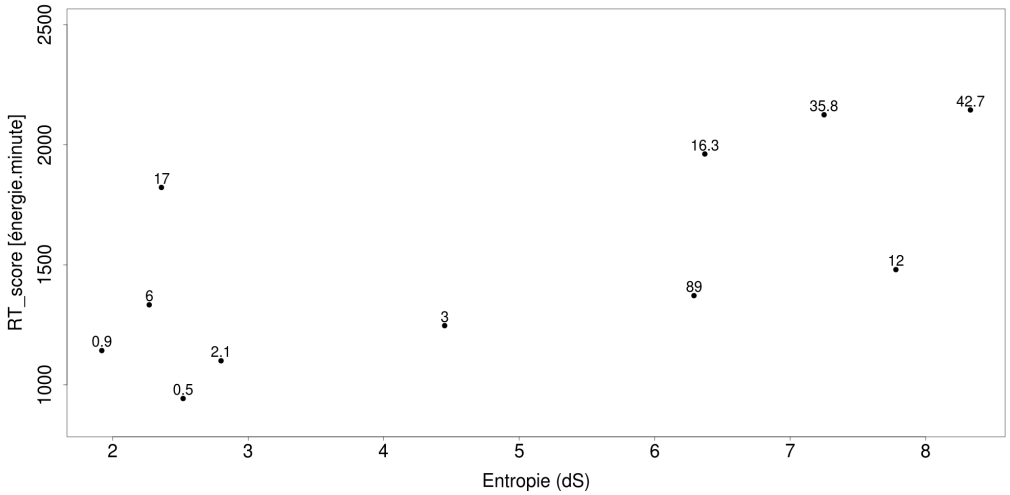


Figure 33 **: RT_{score} en fonction de l'affinité (a), de l'enthalpie (b) et de l'entropie (c).**
Chaque point du graphique correspond à un composé et porte comme label le temps de résidence expérimental exprimé en minute.

La **Figure** 33 montre que le RT_{score} n'est pas corrélé à l'affinité, ni à l'enthalpie et ni à l'entropie, et que ce qu'il représente le mieux est bien le temps de résidence.

4. Conclusion

Ainsi, notre méthode de prédiction du temps de résidence a été appliquée avec succès sur un jeu de données privé constitué de 21 composés et appartenant à l'Institut de Recherche Servier. Bien que notre méthode n'ait pas été développée sur ce jeu de données, elle nous a permis de fournir un classement correct de 18 molécules en fonction de leur temps de résidence avec un coefficient de détermination de 85%. Durant les deux semaines passées chez Servier, nous n'avons pas eu le temps d'analyser plus en détails les trajectoires, notamment celles correspondantes aux trois points aberrants afin d'identifier et de comprendre les possibles sources d'erreur. Il serait également intéressant d'analyser les interactions protéine-ligand afin de mettre en lumière des règles SKR. L'Institut de Recherche Servier dispose d'un autre outil de prédiction du temps de résidence basé sur la méthode *scaled-MD* implémenté dans l'outil *Biki* **(Mollica et al., 2015, 2016)**, développé par la *spin-off* BiKi Technologies située à Gêne **(Decherchi et al., 2018)**. Notre méthode a une performance équivalente à la méthode *scaled-MD,* en terme de temps de calculs et d'analyse. Dans notre méthode, contrairement à la *scaled-MD*, un terme additionnel, qui correspond au potentiel harmonique ($V_{restraint}$) appliqué sur le ligand, est ajouté à l'énergie potentielle totale (U) du champ de forces : $U + V_{restraint}$. Dans la *scaled-MD*, c'est l'énergie potentielle totale qui est atténuée par un facteur λ compris entre 0 et 1 : $\lambda \times U$. Il y a donc un risque accru de distorsion

de la protéine dans la méthode *scaled-MD*. Par conséquent, dans un protocole de simulation par *scaled-MD,* un ensemble de contraintes est systématiquement appliqué sur tous les atomes lourds du squelette de la protéine, à l'exception de ceux du site de liaison, afin de conserver la protéine dans sa conformation native. Ces contraintes peuvent conduire à une description irréaliste du processus de dissociation simulé, notamment si celui-ci implique par exemple un changement de conformation. Ainsi, la méthode proposée offre une description plus réaliste du processus de dissociation et a montré des résultats forts encourageants sur deux jeux de données distincts. De plus, elle est adaptée à un usage industriel et permet de fournir un classement prédictif du temps de résidence de molécules d'intérêt grâce au RT_{score}.

IV. DU PROFIL D'ENERGIE AU TEMPS DE RESIDENCE : UNE ETUDE QUANTITATIVE

A. Préambule

Nous avons vu dans le chapitre III comment la dynamique moléculaire (DM) peut être utilisée pour estimer le temps de résidence et fournir un classement prédictif d'une série de composés tout en limitant le temps de calcul. Avec cette approche rapide, l'exploration de chaque micro-état tout au long du chemin, autrement dit l'échantillonnage, n'est pas suffisante pour permettre d'analyser plus finement la trajectoire et d'en extraire des grandeurs thermodynamiques et cinétiques. Nous allons voir dans cette partie comment la DM peut être utilisée pour caractériser plus finement le chemin de dissociation, et extraire le profil d'énergie libre. Cette étude n'est pas encore aboutie et se situe encore au stade exploratoire. Nous présenterons le protocole utilisé, les premiers résultats obtenus et les interrogations soulevées par nos premières analyses.

Les grandeurs thermodynamiques (telles que l'affinité, autrement dit l'énergie libre de fixation) ne dépendent que de l'état initial (état lié du complexe récepteur-ligand) et de l'état final du chemin (état dissocié), qui sont des états stables. Contrairement aux constantes cinétiques, qui dépendent, elles, de l'ensemble des caractéristiques du chemin, notamment de(s) état(s) de transition. Comme nous l'avons expliqué dans l'introduction, le k_{on} et le k_{off} sont déterminés respectivement par la différence d'énergie entre l'état de transition et les états non liés et liés par l'intermédiaire de l'équation d'Arrhenius, dans un modèle cinétique à une étape. Par conséquent, disposer d'informations structurales et énergétiques (i.e. l'énergie libre) sur les états de transition et les états intermédiaires constitue une aide précieuse pour moduler et optimiser la cinétique d'un composé et donc également son affinité. Cependant, les états de transition sont des structures instables ayant une courte durée de vie et donc difficiles à observer aussi bien expérimentalement que théoriquement.

Ainsi, pour identifier les états intermédiaires stables et les états de transition d'un processus de dissociation et déterminer le profil d'énergie libre associé, il nous faut : i) utiliser une méthode de DM biaisée pour surmonter les barrières énergétiques associées aux états de transition, comme nous l'avons fait dans notre méthode d'estimation du temps de résidence et

ii) échantillonner suffisamment chaque micro-état tout au long du chemin afin de pouvoir en extraire des grandeurs macroscopiques (cf. V.G.1). Il serait théoriquement envisageable d'estimer le temps de résidence à partir des barrières énergétiques identifiées sur le profil. L'objectif de cette étude est de caractériser le mécanisme moléculaire associé au processus de dissociation d'un complexe protéine-ligand et d'extraire les constantes cinétiques des profils d'énergie libre liés à ces phénomènes. Il s'agit d'en identifier les états stables tels que les états intermédiaires, les états de transition, ainsi que les étapes cinétiquement déterminantes. Dans le cadre d'un projet de découverte de médicaments, cette connaissance est d'une grande importance pour optimiser la cinétique et l'affinité d'un *lead*.

B. Matériels et méthodes

Nous avons utilisé le jeu de données des 10 inhibiteurs de CDK8-CycC (**Figure 26** et **Figure 27**)

1. De la simulation au profil d'énergie libre : un protocole en 2 étapes

Nous avons mis au point un protocole en 2 étapes :

- La première étape est la simulation du processus de dissociation du complexe protéine-ligand. Pour cela, nous avons utilisé le même protocole que celui de notre méthode d'estimation du temps de résidence (cf. III.A.1, page 117). La seule différence réside dans le temps total de simulation qui est réduit d'un facteur 10 (150 ps au lieu de 1500 ps). Bien entendu, ici il ne s'agit pas d'estimer le temps de résidence, mais seulement d'obtenir la trajectoire de dissociation. Pour chaque inhibiteur, nous avons simulé 11 fois le processus de dissociation.
- La deuxième étape consiste à ré-échantillonner le chemin de dissociation obtenu lors de l'étape 1 et à appliquer la méthode d'analyse par histogramme pondéré (WHAM) sur les échantillons collectés afin de calculer le profil d'énergie libre associé au processus de dissociation. Cette étape est expliquée en détails dans la partie V.J.2 du chapitre V. Cette étape étant coûteuse en temps de calcul, nous n'avons pas ré-échantillonné les 11 répliques de chacun des 10 inhibiteurs. Nous avons sélectionné et ré-échantillonné 3 chemins de dissociation par inhibiteur, détaillés dans la partie

« Analyse et sélection des chemins ». Nous avons donc calculé un total de 30 profils d'énergie libre.

2. Choix des paramètres pour la 2ème étape du protocole

Le chemin obtenu à l'issue de l'étape 1 doit être ré-échantillonné à des positions données notées $RMSD_i^{référence}$ (ou $\xi_i^{référence}$), situées à des intervalles réguliers de RMSD. Pour cela, il faut sélectionner un ensemble de N conformations successives sur le chemin, espacées par des intervalles réguliers de RMSD, à partir desquelles nous avons lancé N simulations de TMD. Ces N simulations sont communément appelées des fenêtres d'échantillonnage. La valeur initiale du RMSD est fixée à 0.001 et correspond à l'état lié du système ($RMSD_1^{0,001}$). Cette valeur initiale n'a pas été fixée exactement égale à zéro afin d'éviter d'imposer une contrainte initiale trop forte sur le système. La valeur finale correspond à l'état où le complexe protéine-ligand est dissocié avec le ligand complétement solvaté ($RMSD_N^{RMSD\_final}$). La définition de l'état dissocié est la même que celle utilisée dans la partie III.A.1 (page 117). Pour les chemins analysés dans cette étude, le RMSD_final varie globalement de 20 Å à 50 Å. Après plusieurs tests, l'écart entre deux coordonnées de réaction successives ($RMSD_{i+1}^{référence} - RMSD_i^{référence}$), la valeur de la constante de force K du potentiel harmonique $w_i(RMSD)$, ainsi que le temps de simulation dans chacune des fenêtres i ont été fixés à 0.25 Å, 80 $kcal.mol^{-1}.Å^{-2}$ et 2 ns, respectivement. Ainsi, en fonction du RMSD_final, le nombre de fenêtres par chemin varie de 80 (20/0.25) à 200 (50/0.25) et le temps simulation total varie de 160 ns (80×2) à 400 (200x2) ns. Pour l'ensemble des inhibiteurs, nous avons réalisé un total de 8,5 µs de simulation. Enfin, à partir de chaque fenêtre d'échantillonnage i, nous avons sauvegardé 1000 valeurs de RMSD, dont les distributions sont par la suite analysées avec la méthode WHAM.

C. Résultats et discussion

1. Analyse et sélection des chemins

A l'issue de l'étape 1 du protocole, nous avons analysé les chemins de dissociation obtenus et nous avons constaté que nous obtenons une répartition similaire à celle obtenue dans la partie III.A.1 (page 117). Les inhibiteurs de court (SRT) et moyen (MRT) temps de résidence empruntent majoritairement la voie *allosteric channel* et les inhibiteurs de long temps de

résidence (LRT), sauf l'inhibiteur 10 où le chemin *allosteric* est majoritaire, empruntent majoritairement la voie *front channel*. (**Figure 34)**.

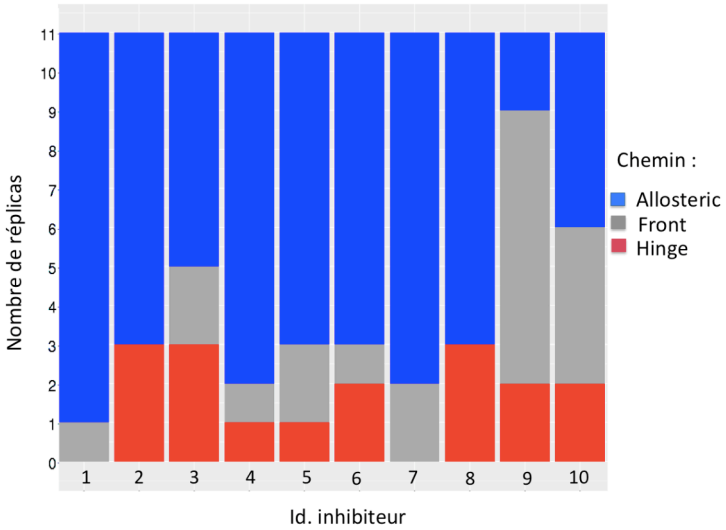


Figure 34 : Barplot comptant le nombre de chemins de chaque type emprunté par chaque inhibiteur.

Nous souhaitions dans un premier temps vérifier si des chemins similaires présentent des profils d'énergie libre similaires pour un inhibiteur donné. Etant donné que la voie *allosteric* est majoritairement empruntée par les inhibiteurs SRT et MRT, nous avons sélectionné pour chacun d'entre eux 3 chemins représentatifs de cette voie. Pour les 2 inhibiteurs LRT, nous avons choisi 2 chemins qui passent par la voie majoritaire *front channel* et 1 chemin qui passe par la voie *allosteric channel*.

2. Analyse des distributions de RMSD

Avant de calculer les profils d'énergie libre, il faut analyser les N distributions de RMSD afin de s'assurer qu'elles répondent bien à certains critères (cf. V.J.2). C'est une étape cruciale car l'énergie libre est estimée directement à partir de ces distributions. Cette analyse a été réalisée sur les 30 ensembles de distributions. Nous présentons un exemple représentatif de ré-échantillonnage du processus de dissociation, celui de l'inhibiteur 9 sortant par un chemin *front channel* (**Figure 35**).

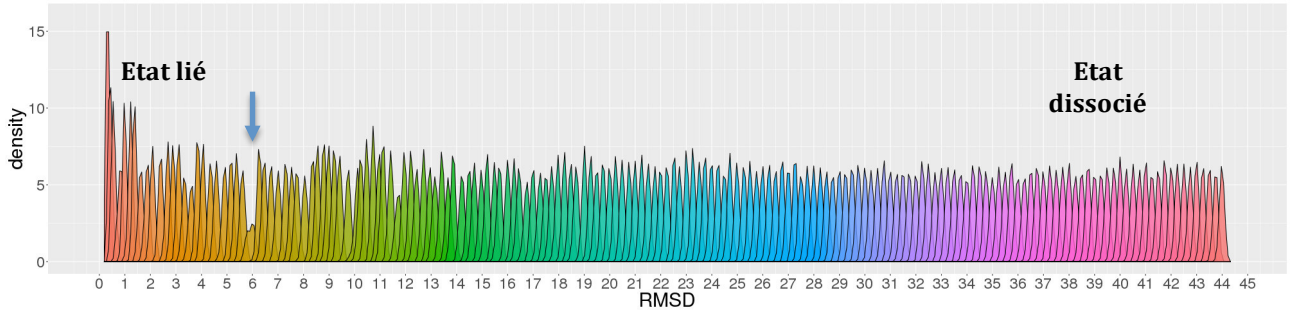


Figure 35 : Distributions de RMSD extraites de chaque fenêtre i de simulation à l'issue de l'étape de ré-échantillonnage.
Sur le graphique, nous avons indiqué l'état lié et dissocié et marqué par une flèche une distribution atypique. 177 fenêtres ont été simulées ici. 86% des distributions suivent une loi normale et présentent donc une forme gaussienne.

Les distributions successives semblent toutes se chevaucher. Elles sont de forme gaussienne, ce qui signifie que l'échantillonnage est confiné autour d'une région. Le test de Shapiro montre que 86% des distributions suivent une loi normale, avec un risque d'erreur de 5 %. On remarquera tout de même la distribution atypique autour de 6 Å. Le fait que cette distribution soit singulière et de forme aplatie suggère que la position $RMSD_i^{référence}$ associée à cette fenêtre correspond à une région de haute énergie. A ce stade, bien que la grande majorité des distributions présente un échantillonnage confiné, on ne sait pas si cette exploration se fait bien autour du RMSD demandé ($RMSD_i^{référence}$). Cependant, il est tout de même peu probable que toutes ces distributions soient déviées alors qu'elles présentent une forme gaussienne. Pour le vérifier, il faut analyser la déviation moyenne de chaque distribution par rapport à sa valeur de $RMSD_i^{référence}$. La **Figure 36** confirme que toutes les distributions se recouvrent. Ce critère est très important car le calcul du potentiel de force moyenne ne peut converger sans cela. Puis, la courbe de déviation montre que plus le processus de dissociation progresse, et plus la déviation tend vers 0, pour finir par s'annuler une fois le ligand solvaté. Or, on sait qu'une fois que le ligand est complétement solvaté et qu'il diffuse dans l'eau, l'énergie libre est sensée être stable et ne pas subir de variation. Cela suggère que la déviation et donc l'échantillonnage au niveau microscopique semble traduire l'observable macroscopique. L'ensemble de ces analyses montre donc que le choix des paramètres effectués (constante de force et écart de RMSD entre les fenêtres) conduit à un échantillonnage satisfaisant pour valider les critères d'application de la méthode WHAM, mais pas forcément suffisant pour observer tous les phénomènes du processus de dissociation.

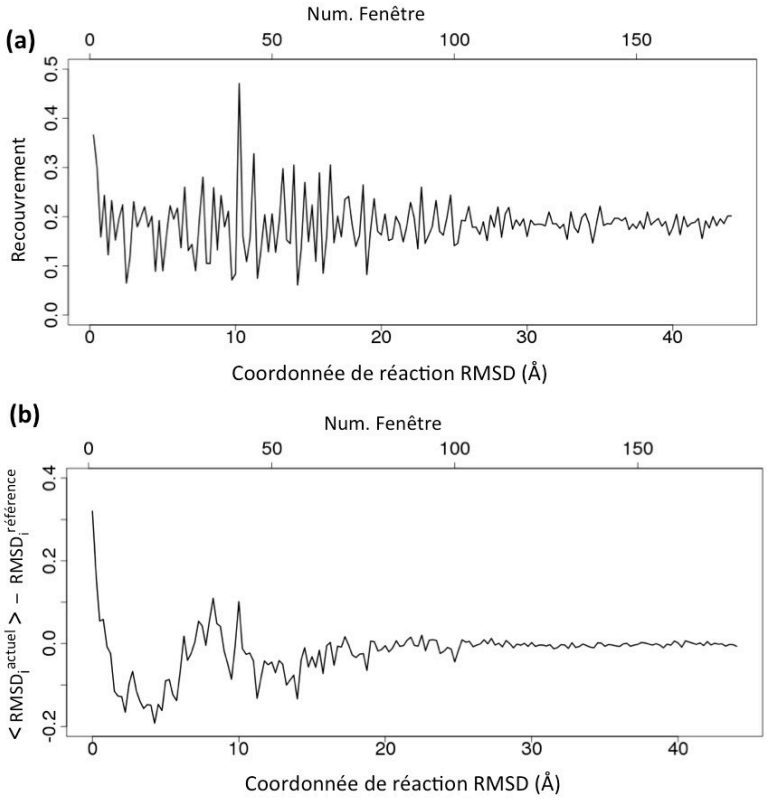


Figure 36 : Analyse du recouvrement et de la déviation des distributions.
(a) Proportion de recouvrement entre deux fenêtres successives. **(b)** Ecart entre la moyenne de RMSD échantillonné dans chaque fenêtre et le RMSD voulu.

Le profil d'énergie obtenu est présenté sur la **Figure 37**. Lorsque la courbe de déviation décroit ou augmente fortement, cela se traduit respectivement par une augmentation ou une baisse d'énergie libre, comme expliqué dans le chapitre V.J.2.

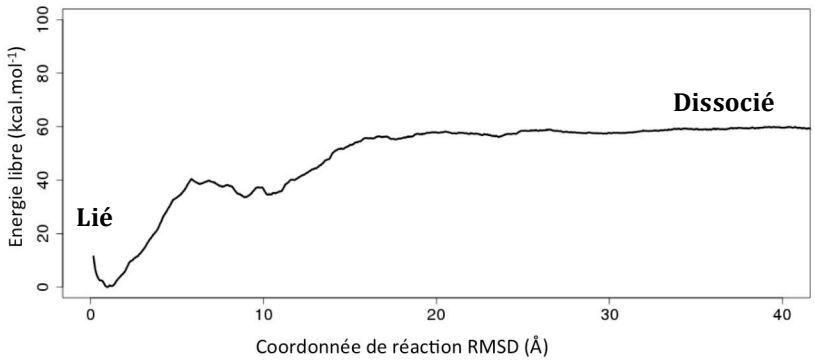


Figure 37 : Profil d'énergie libre associé au processus de dissociation de l'inhibiteur 9 par le chemin front.

L'analyse des 30 ensembles de distributions a montré que les recommandations d'application de la méthode WHAM sont remplies (cf. V.J.2).

3. Analyse des profils d'énergie

a) Le RMSD : un lien complexe entre l'énergie libre et la structure

Nous ne présenterons pas l'analyse des 30 profils, seulement un sous-ensemble illustrant nos résultats. L'analyse des profils révèle que pour un même inhibiteur empruntant des chemins similaires, comme l'inhibiteur n°7 ici, les profils d'énergie peuvent être différents ainsi qu'illustré sur la **Figure 38** avec les répliques 1 et 2. Au contraire, les répliques 1 et 3, qui sont plus dissimilaires en termes de chemins empruntés, présentent des courbes énergétiques plus similaires.

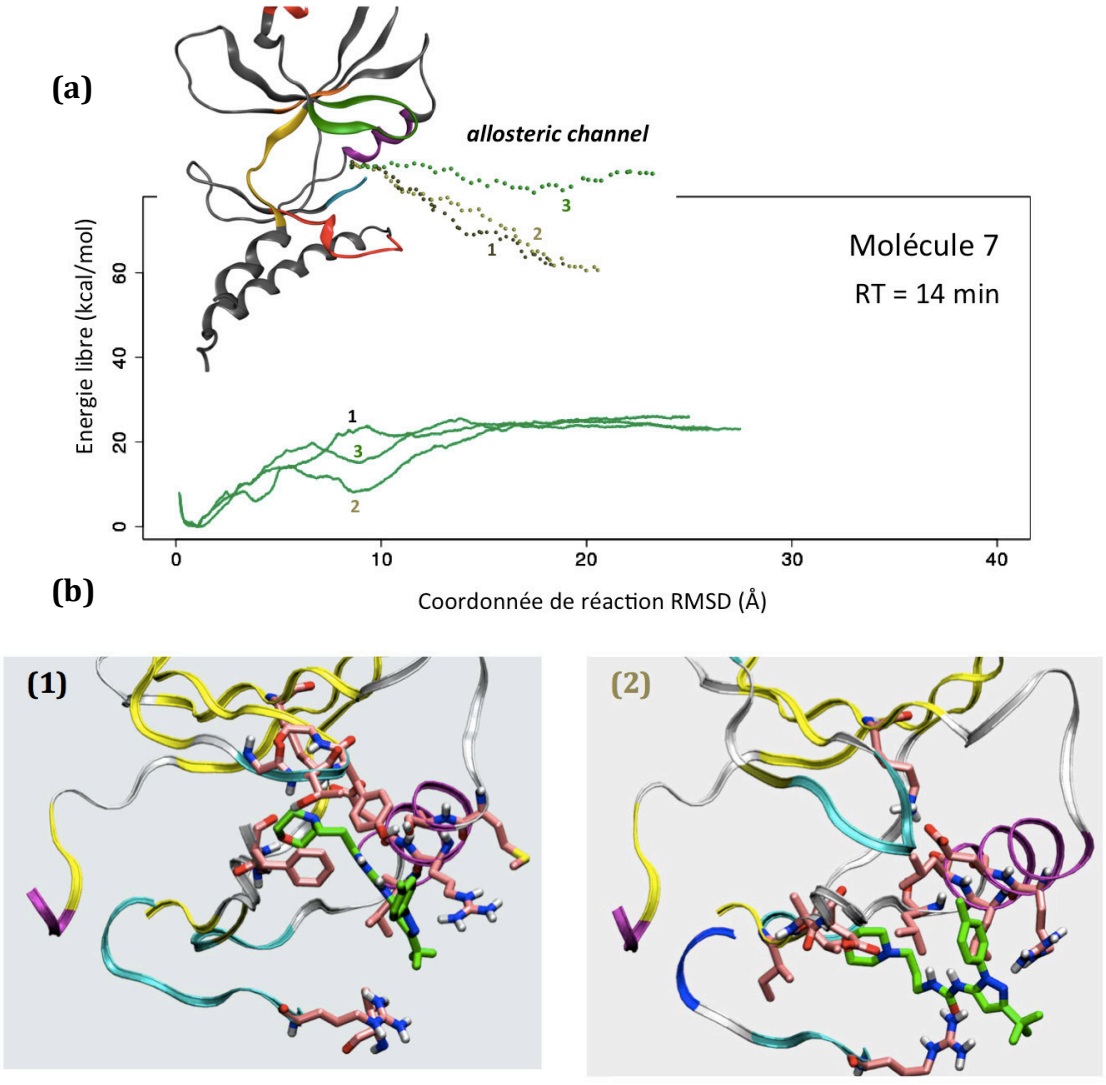


Figure 38 : Profils d'énergie libre associés au processus de dissociation de l'inhibiteur 7.
(a) Le site de liaison de CDK8-CycC est représenté en ruban gris sauf les motifs conservés des protéines kinases. Les centres de masse de la molécule au cours de sa dissociation sont

représentés par des sphères. Les profils des trois répliques sont tracés. **(b)** La conformation 1 est issue de la fenêtre à 8.75 Å de RMSD de la réplique 1 ; idem pour la conformation 2 qui provient de la réplique 2.

En analysant la fenêtre qui correspond au point de plus grande divergence entre les profils des répliques 1 et 2 (à 8.75 Å de RMSD), on constate que l'inhibiteur ne se situe pas dans la même position dans le site de liaison (**Figure 38**). Cela illustre bien le fait que, l'inhibiteur peut adopter une conformation différente et ou se trouver à une distance complétement différente pour une même valeur de RMSD.

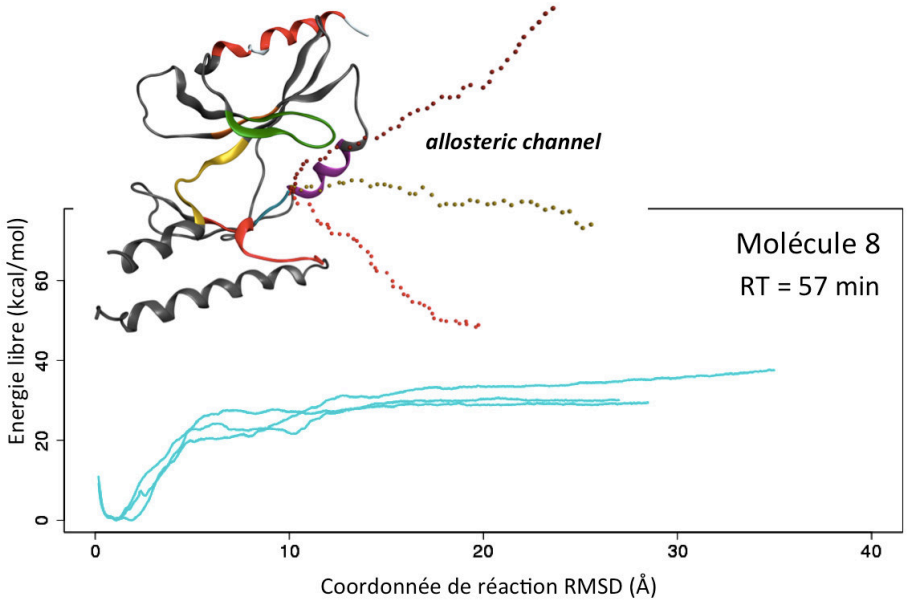


Figure 39 : Profils d'énergie libre associés au processus de dissociation de la molécule 8.
Le site de liaison de CDK8-CycC est représenté en ruban gris sauf les motifs conservés des protéines kinases. Les centres de masse de la molécule au cours de sa dissociation sont représentés par des sphères. Les profils des trois répliques sont tracés.

A l'inverse, nous avons également observé que des trajectoires de dissociation différentes peuvent conduire des profils énergétiques de même allure. Les 3 répliques de l'inhibiteur 8 présentent des trajectoires très diverses bien qu'appartenant toutes à la voie *allosteric channel*. Étonnamment, les profils d'énergie sont relativement similaires. Ils ont une forme légèrement aplatie, ne permettant d'identifier clairement un état de transition particulier. Cela peut être dû à un échantillonnage insuffisant, et cela constitue une des pistes futures à explorer pour améliorer notre approche.

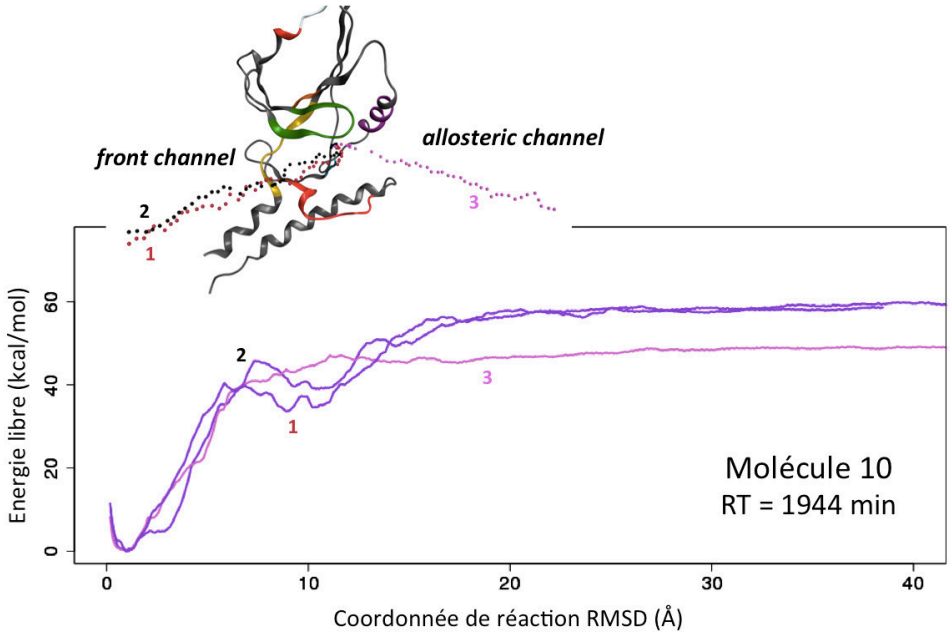


Figure 40 : Profils d'énergie libre associés au processus de dissociation de la molécule 10.
Le site de liaison de CDK8-CycC est représenté en ruban gris sauf les motifs conservés des protéines kinases. Les centres de masse de la molécule au cours de sa dissociation sont représentés par des sphères. Les profils des trois répliques sont tracés.

Pour la molécule 10, les 2 chemins *front channel* sont très similaires tout comme leurs profils respectifs qui présentent une bonne superposition. Sur ces profils, il semble y avoir au moins un état de transition. Le profil énergétique du chemin *allosteric channel* se distingue nettement de celui du chemin *front channel.*

Ainsi, bien que l'étape de ré-échantillonnage ait été lancée avec le même jeu de paramètres (écart entre deux fenêtres, constante de force, temps de simulation) pour tous les inhibiteurs, il est difficile de dégager un profil d'énergie libre caractéristique d'un chemin de sortie donné. Deux trajectoires de dissociation différentes d'un même inhibiteur peuvent mener à deux profils très similaires. A l'inverse, deux répliques d'un même inhibiteur, empruntant des chemins similaires, peuvent présenter des profils d'énergie différents. Cependant, on peut se demander si les profils d'énergie sont comparables. En effet, comme illustré avec la molécule 7, pour une même valeur de RMSD, l'inhibiteur peut adopter une conformation différente et ou se trouver à une distance complétement différente.

Cette observation illustre bien la difficulté d'interprétation qu'implique le RMSD. Tous les points vérifiant une valeur donnée de la distance euclidienne peuvent être représentés sur la surface d'une sphère (donc en 2D), tandis que ceux vérifiant une valeur donnée du RMSD représentent un volume 3D plus complexe. L'espace des possibilités est plus grand avec le RMSD. Il est donc plus difficile de relier le profil énergétique, qui est exprimé en fonction du

RMSD, aux évènements structuraux. Cela nous amène à nous questionner quant au sens que revêt la comparaison des profils. Si l'on considère la distance euclidienne entre le centre de masse du site de liaison et celui de l'inhibiteur à la place du RMSD, ce problème est beaucoup moins prononcé car, dans l'espace confiné que représente un site de liaison, il est difficile de déplacer le centre de masse du ligand sur un cercle (dont le centre est le centre de masse du site de liaison). Rappelons tout de même que la complexité géométrique du RMSD constitue également son principal avantage, ce qui nous a orienté vers son utilisation. En effet, le fait qu'à une valeur donnée de RMSD correspond un large spectre de possibilités permet d'imposer une contrainte « plus douce » sur le système, et donc un échantillonnage plus proche des conditions non-biaisées. De plus, le processus que l'on essaie de simuler est de nature complexe, et donc l'état de transition associé peut être variable d'un représentant du système à l'autre. Ainsi, malgré l'apparente difficulté de caractérisation structurale de ces états, cette coordonnée de réaction laisse une plus grande liberté au système pour explorer divers états de transition et donc plusieurs barrières énergétiques possibles, comme cela peut se produire dans une mesure expérimentale. Ainsi, la détermination des constantes cinétiques est sans doutes possible, et même peut-être plus pertinente, qu'avec l'utilisation d'une coordonnée de réaction plus simpliste. L'idée pourrait être, par exemple, de multiplier les répliques afin d'engranger un échantillonnage plus important sur ces barrières, les moyenner et ainsi obtenir une valeur quantitative du $\Delta G^{\#}_{off}$, directement liée au k_{off}.

b) Du profil d'énergie libre au temps de résidence

Une autre mesure intéressante pouvant être extraite de nos profils d'énergie libre est la différence entre l'état initial et l'état final du PMF (appelé ici ΔG_{exit}). Cette mesure se rapproche du calcul du ΔG_D, couramment associé à l'affinité et à l'énergie libre de liaison. Plusieurs études ont montré que la différence entre les valeurs de ΔG_{exit} de deux inhibiteurs, calculée à partir des potentiels de force moyenne, est qualitativement en accord avec la différence de leurs k_{off} mesurés (Niu et al., 2016; Sun et al., 2015). La **Figure 41** représente l'ensemble des PMFs obtenus dans notre travail. Nous pouvons voir, dans un premier temps, que le ΔG_{exit} est en effet corrélé au temps de résidence. Cependant, nous nous apercevons également que l'affinité (K_D) évolue dans le même sens que le temps de résidence et est donc potentiellement corrélé au ΔG_{exit}. (**Tableau 8**). Ainsi, notre jeu de donnée actuel ne nous permet pas de conclure définitivement sur une corrélation entre nos profils d'énergie libre et le temps de résidence des composés. Néanmoins, l'approche développée ici est un premier

signe encourageant vers une possible prédiction quantitative des temps de résidence d'inhibiteurs.

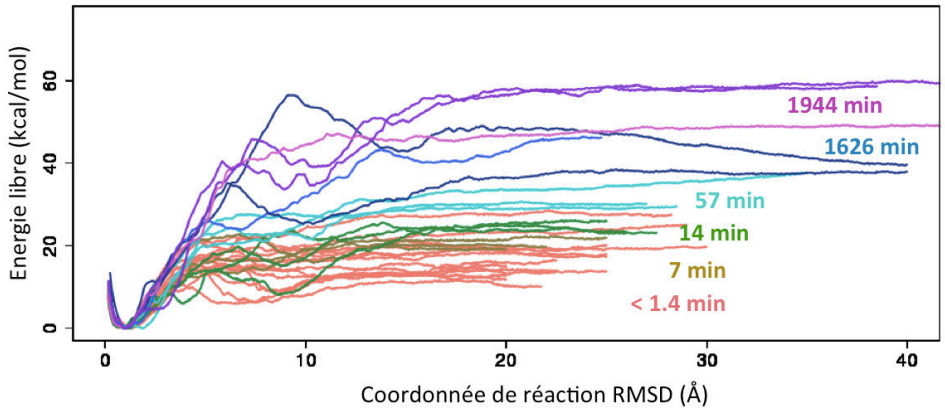


Figure 41 : Profils d'énergie libre associés au processus de dissociation des 10 inhibiteurs simulés chacun 3 fois.
Les courbes ont été colorées selon le temps de résidence expérimental. Il y a 3 réplicas par molécule, ce qui fait un total de 30 réplicas.

D. Conclusion et perspectives

Cette partie exploratoire a pour but la mise au point d'un protocole pour obtenir quantitativement une estimation du temps de résidence, et identifier les états de transitions associés à la dissociation d'un complexe protéine-ligand, ceci afin d'optimiser la structure des composés. On souhaite identifier les états intermédiaires stables, les états de transition et déterminer le profil d'énergie libre associé. Nous avons pour cela mis au point un protocole en 2 étapes qui consiste à ré-échantillonner un chemin, obtenu à l'issue d'une 1ère étape, et appliquer la méthode WHAM, pour calculer un potentiel de force moyenne à partir des distributions de la coordonnée de réaction extraites du ré-échantillonnage. Ce protocole est basé sur la dynamique moléculaire dirigée, qui utilise un potentiel harmonique pour contraindre le RMSD, notre coordonnée de réaction.

Le paramétrage de la méthode a nécessité plusieurs tests avant l'obtention de distributions pertinentes, répondant aux recommandations de l'utilisation de la méthode WHAM. L'analyse des profils d'énergie associés aux processus de dissociation simulés, a permis de prendre conscience de la difficulté de comparer, d'analyser et d'interpréter des profils d'énergie exprimés en fonction du RMSD. Le grand nombre de possibilités géométriques que vérifie une valeur de RMSD rend difficile l'établissement du lien entre le profil énergétique et les évènements structuraux. Cependant, considérant la nature variable de la structure possible de

l'état de transition sur un phénomène complexe comme un processus de dissociation, il n'est pas surprenant d'observer cette difficulté de caractérisation structurale des états de transitions. Nous pouvons également tirer profit de nos profils pour mesurer des barrières d'énergie et donc prédire le temps de résidence. Cependant, considérant notre jeu de données particulier, dans lequel l'affinité évolue dans le même sens que le temps de résidence, nous ne pouvons pas conclure sur une corrélation directe entre nos barrières d'énergies et le temps de résidence.

En perspective de cette étude, il serait intéressant de regarder, dans un premier temps, si le temps d'échantillonnage employé est suffisant pour faire converger nos fenêtres. Plusieurs études signalent cependant qu'augmenter le nombre de fenêtres, pour assurer un recouvrement important des distributions, est plus important pour la précision du profil qu'augmenter le temps de simulation par fenêtre (Kästner, 2011; Zhu and Hummer, 2012). Le nombre de fenêtres est également un critère qu'il faudra explorer et tester.

Dans un second temps, notre approche pourrait être appliquée sur un autre jeu de données, dans lequel le temps de résidence et l'affinité n'évolueraient pas dans le même sens. Ceci permettrait de vérifier si il existe une réelle corrélation entre nos barrières d'énergie et le temps de résidence.

Enfin, notre approche pourrait être couplée à un modèle de Markov, dans lequel nos simulations de ré-échantillonnage permettraient d'enrichir le modèle mathématique markovien. Il identifie clairement les états cinétiquement déterminants et calcule les probabilités de transition entre les différents états. Cela pourrait être comparé à nos profils afin d'en vérifier la pertinence.

V. PRINCIPES ET METHODES

Dans ce chapitre, nous aborderons une description des méthodes utilisées pour la réalisation de ce travail.

A. Modélisation moléculaire

La modélisation moléculaire est une discipline regroupant un ensemble de méthodes utilisées pour modéliser ou simuler le comportement de molécules. Il existe deux types de modèles moléculaires : quantique ou classique. Dans le modèle quantique les électrons sont représentés explicitement. Les méthodes *ab initio* de chimie quantique sont précises mais restent très coûteuses en temps de calcul malgré les progrès en technologie informatique. Elles ne sont donc pas adaptées aux systèmes de grande taille comme les macromolécules biologiques. Basé sur les principes de la mécanique moléculaire dite classique de Newton, le modèle classique décrit le noyau et ses électrons comme une même entité : une sphère indéformable de masse, de charge et de rayon donnés. On dispose ainsi d'une représentation du système à l'échelle atomique dont on peut déterminer l'énergie potentielle. L'énergie potentielle du système est calculée à l'aide d'un champ de forces qui modélise les interactions s'exerçant entre les différents atomes. Un champ de forces est l'ensemble des paramètres associés aux termes d'énergie d'interaction permettant de modéliser l'énergie potentielle d'un système de particules. Ces paramètres ont été ajustés et optimisés à partir de données expérimentales (cristallographie aux rayons X, RMN et spectroscopie infrarouge) et de calculs de chimie quantique. Les termes énergétiques du champ de forces peuvent être groupés en deux catégories : les termes d'énergie liés et non-liés, comme présentés dans le paragraphe suivant.

B. Champ de forces

Les champs de forces couramment utilisés sont AMBER (Case et al., 2015), CHARMM (Brooks et al., 2009), GROMOS (Oostenbrink et al., 2004) et OPLS (Jorgensen et al., 1996). Les champs de forces doivent être choisis avec soin selon la nature du système à simuler. Dans cette étude, nous avons utilisé le champ de forces AMBER, car ce champ de forces convient aux études sur les interactions protéine-protéine et protéine-ligand.

Dans le champ de forces AMBER, l'énergie potentielle d'un système est donnée par la relation suivante :

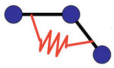
$$V_{potentielle} = \underbrace{E_{liaison} + E_{angle} + E_{dièdre}}_{\text{Termes liés}} + \underbrace{E_{électrostatique} + E_{VdW}}_{\text{Termes non liés}}$$

Equation 4

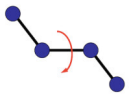
Où chaque terme est défini de la façon suivante :

$$V_{potentielle} = \sum_{\text{Liaison}} k_b(b - b_0)^2$$

$$+ \sum_{\text{Angle}} k_\theta(\theta - \theta_0)^2$$



$$+ \sum_{\text{Dièdre}} k_\varphi[\cos(n\varphi + \delta) + 1]$$



$$+ \sum_{\text{atome i}} \sum_{\text{VdW}} 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{d_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{d_{ij}}\right)^{6}\right]$$

$$+ \sum_{\text{i}} \sum_{\text{Electrostatique}} \frac{q_i q_j}{4\pi\varepsilon d_{ij}}$$

Equation 5

Les trois premiers termes sont les termes d'énergie liés, respectivement les termes d'énergie de la liaison entre 2 atomes, de l'angle formé entre 3 atomes et de l'angle de torsion formé par 4 atomes successivement liés. Les énergies de liaison et d'angle correspondent aux énergies de déformation de la liaison et de l'angle considéré. Elles ont la forme d'un potentiel harmonique où b_0 et θ_0 sont respectivement la distance à l'équilibre et l'angle à l'équilibre des atomes considérés. L'énergie de torsion, associée à la rotation autour d'une liaison, est représentée par un potentiel périodique de type sinusoïdal où φ représente la valeur de l'angle dièdre tandis que δ représente la phase de l'angle.

Les deux derniers termes énergétiques, les termes d'énergie non-liés, correspondent aux interactions de van der Waals (décrites par un potentiel de type Lennard-Jones) ainsi qu'aux interactions électrostatiques entre les atomes i et j séparés par au moins trois liaisons. De fait, la distance entre 2 atomes séparés par deux liaisons ou trois liaisons étant courte, les énergies d'interactions "non-liées" entre ces atomes voisins seront très élevées. Par conséquence, pour

les interactions "non-liées", seules les interactions séparées d'au moins 3 liaisons sont prises en compte. Pour les interactions de van der Waals, le premier terme en puissance 12 correspond à la répulsion entre deux atomes i et j due à l'exclusion de Pauli et le second terme en puissance 6 représente la dispersion attractive de London. σ_{ij} représente la distance à l'équilibre entre les atomes i et j. Ce terme attractif englobe les interactions faibles de type dipôles-dipôles permanents (London), dipôles-dipôles induits (Debye) et dipôles induits-dipôles induits (Keesom). Enfin, l'énergie électrostatique dérive de la loi de Coulomb et décrit la force de l'interaction entre deux atomes i et j de charge respective q_i et q_j, séparés par une distance d_{ij}. Elle dépend de la constance diélectrique du milieu ε.

C. Modèle d'eau

Afin de reproduire autant que possible les conditions biologiques et mimer le milieu aqueux, la macromolécule biologique est placée dans une boite d'eau. Nous avons utilisé une représentation explicite du solvant avec un modèle à 3 points appelé TIP3P (Jorgensen et al., 1983). Dans ce modèle, afin de simplifier les calculs sur le grand nombre de molécules d'eau du système, la molécule est maintenue rigide par 3 liaisons, deux entre O et H et une pseudo-liaison H⋯H.

D. Distance de troncature et PME

Le nombre de paires d'atomes, et donc le nombre d'interactions non-liées à calculer, augmente avec la taille du système. Afin de limiter le temps de calcul des interactions non-liées, une distance de troncature (d_{cut}) est utilisée au-delà de laquelle les interactions de van der Waals sont considérées comme négligeables et ne sont plus calculées ; le calcul des interactions électrostatiques est, quant à lui, simplifié au moyen de la méthode du maillage particulaire d'Ewald (*Particle Mesh Ewald*, PME). Cette méthode consiste à calculer les interactions électrostatiques à longue distance ($d > d_{cut}$) en appliquant une transformée de Fourier du potentiel électrostatique. A courte distance ($d < d_{cut}$), le calcul de l'interaction électrostatique est réalisé de façon classique comme décrit précédemment en utilisant un potentiel de Coulomb (Equation 5). Dans leurs espaces respectifs, le calcul des interactions électrostatiques à courte distance (espace des réels) et celui des interactions à longue distance (espace de Fourier) convergent rapidement ce qui permet de réduire significativement le temps de calcul. La distance de troncature est généralement fixée entre 8

et 10 Angströms. La méthode du maillage particulaire d'Ewald suppose implicitement que le système est infiniment périodique d'où le terme « maille » qui fait référence à l'unité répétée. Le système est représenté de façon périodique grâce aux conditions périodiques aux limites.

E. Conditions périodiques aux limites

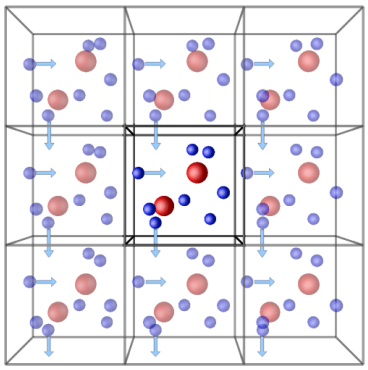


Figure 42 : Schéma représentant le principe des conditions périodiques aux limites (Tesson, 2016)

Les conditions périodiques aux limites consistent en la réplication à l'infini du système initial, dit système unitaire, dans les trois directions de l'espace (**Figure 42**). Cela permet de s'affranchir des effets de bords et d'éviter la dispersion des molécules de solvant dans le cas, par exemple, d'un système biologique placé dans une boite de solvant. Lorsqu'une molécule sort de la boite par un côté, elle y est aussitôt réintroduite par le côté opposé. Le nombre de particules dans la boite unitaire reste donc constant au cours du temps. Le calcul des interactions entre la particule i et la particule j s'effectue toujours avec l'image de j la plus proche de i (qui ne se trouve pas nécessairement dans la même boite que i) : c'est le principe de la convention d'image minimale. Ainsi les conditions périodiques aux limites permettent, d'une part, de modéliser le système dans les conditions les plus réalistes possibles, et d'autre part, de calculer des grandeurs macroscopiques à partir d'une seule unité du système.

F. Minimisation de l'énergie

En modélisation moléculaire l'étape de minimisation est une étape cruciale pour finaliser la préparation du système. L'objectif est d'en optimiser la structure géométrique de façon à obtenir une structure présentant l'énergie potentielle la plus basse possible. Elle permet également d'adapter le système au champ de forces. La minimisation consiste à se déplacer sur l'hypersurface énergie-coordonnée de façon à diminuer l'énergie potentielle du système et

à atteindre un puits de potentiel le plus bas possible (**Figure 43**). Il existe différents algorithmes de minimisation dont les plus communément utilisés sont la méthode de la plus grande pente (*Steepest-descent*) appelée aussi méthode du gradient simple, et la méthode du gradient conjugué. Ces deux approches sont complémentaires et sont souvent appliquées successivement. La méthode de la plus grande pente permet une optimisation importante de la structure, c'est à dire qu'elle accepte des géométries initiales se situant loin du minimum, tandis que la méthode du gradient conjugué permet d'affiner la structure de façon à converger vers un minimum énergétique. Ces deux algorithmes cherchent un minimum énergétique à partir des informations de la dérivée première de l'énergie potentielle.

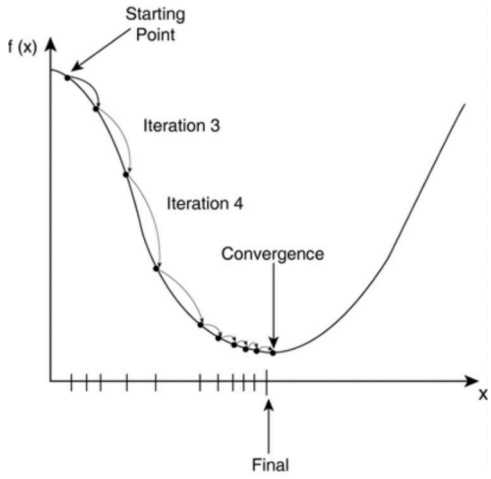


Figure 43 : Schéma du principe de la minimisation

1. Méthode de la plus grande pente (Steepest-descent)

Cette approche consiste à rechercher la direction de plus grande pente c'est à dire la direction selon laquelle l'énergie potentielle $V(\vec{r}_n)$ décroît la plus fortement. La direction de recherche $\vec{D}_n$ est donnée par l'opposé du gradient de l'énergie potentielle (Equation 6). Les coordonnées du système à l'itération n+1 ($\vec{r}_{n+1}$) sont calculées à partir des coordonnées à l'itération n ($\vec{r}_n$), de α_n et de la direction de recherche $\vec{D}_n$ (Equation 7). Le coefficient α_n est ajusté à chaque itération n en fonction de l'énergie du système à l'itération n+1 : si celle-ci baisse, α_n est augmenté et si elle augmente, α_n est réduit.

$$\vec{D}_n = -\vec{\nabla}V(\vec{r}_n)$$

Equation 6

$$\vec{r}_{n+1} = \vec{r}_n + \alpha_n \vec{D}_n$$

Equation 7

Rapide et peu coûteux en mémoire, cet algorithme est appliqué en première approche pour éliminer les mauvais contacts entre atomes. Cependant à l'approche d'un minimum énergétique, l'algorithme présente des problèmes de convergence, où des oscillations de géométrie autour du minimum d'énergie sont observées.

2. Méthode du gradient conjugué

Comme dans la méthode de la plus grande pente, les nouvelles coordonnées du système sont obtenues suivant l'Equation 7. Cependant, les directions de recherche ($\vec{D}_n$) sont déterminées de sorte qu'elles soient conjuguées aux directions de recherche précédentes ($\vec{D}_{n-1}$) (Equation 8).

$$\vec{D}_n = -\vec{\nabla}V(\vec{r}_n) + \frac{\vec{\nabla}V(\vec{r}_n)^T.\vec{\nabla}V(\vec{r}_n)}{\vec{\nabla}V(\vec{r}_{n-1})^T.\vec{\nabla}V(\vec{r}_{n-1})}\vec{D}_{n-1}$$

Equation 8

Plusieurs implémentations de cet algorithme existent et c'est celui de Fletcher-Reeves qui est utilisé dans AMBER (Fletcher and Reeves, 1964). L'avantage de cet algorithme est qu'il permet de converger vers un minimum énergétique lorsque le système s'en rapproche, évitant ainsi les problèmes d'oscillation rencontrés avec la méthode de la plus grande pente. En revanche, cette méthode est peu efficace si la structure initiale possède beaucoup de défauts et est trop distante du minimum d'énergie.

G. Dynamique moléculaire

La simulation de dynamique moléculaire (DM) connaît un succès fort pour l'étude des systèmes biologiques tels que les protéines et les acides nucléiques. Elle permet de simuler le comportement dynamique d'un système moléculaire au cours du temps. L'objectif est d'appréhender des processus temporels ainsi que de comprendre, prédire et calculer les propriétés du système étudié. La capacité d'une protéine à adapter sa conformation en réponse à son environnement est une étape clé des processus biologiques tels que les interactions protéine-ligand, la catalyse enzymatique, l'interaction protéine-protéine, le transport de protéines et la transduction du signal. Au niveau moléculaire, l'interaction entre les deux partenaires (protéine-ligand, protéine-protéine etc.) peut impliquer un mouvement

subtil de chaînes latérales de quelques acides aminés comme elle peut entraîner un changement de conformation spectaculaire. Une compréhension approfondie de la fonction de la protéine dans un processus biologique implique nécessairement une description structurale et une quantification énergétique des événements dynamiques associés. L'intégration de ces connaissances dans un projet de conception de candidat-médicaments facilite la découverte, la conception et le développement de nouveaux médicaments et constitue une avancée fondamentale dans la connaissance des processus biologiques clés.

1. De la DM aux propriétés macroscopiques

La DM permet de calculer les propriétés macroscopiques ou thermodynamiques (par exemple, pression, énergie, enthalpie) du système à partir des données microscopiques obtenues lors de la simulation. Le lien entre les données microscopiques et les propriétés macroscopiques est établi par la mécanique statistique. La mécanique statistique définit une grandeur (ou un observable) expérimentale comme une valeur obtenue à partir d'une moyenne d'ensemble (*ensemble average*) représentant les états possibles du système obtenus de manière instantanée (différents représentants du système). La DM, quant à elle, offre la possibilité de calculer la valeur d'une grandeur en considérant la moyenne temporelle calculée à partir d'un échantillonnage suffisamment long d'un seul représentant du système. Il existe un axiome important en mécanique statistique, connu sous le nom d'hypothèse d'Ergodicité, qui permet de relier la moyenne d'ensemble à la moyenne temporelle. L'hypothèse d'Ergodicité stipule que la moyenne temporelle d'une grandeur calculée à partir des données d'une simulation de DM suffisamment longue approche la moyenne d'ensemble calculée sur un grand nombre d'individus obtenus de manière instantanée. Ainsi, si l'échantillonnage est suffisamment important au cours de la simulation, l'hypothèse d'Ergodicité est vérifiée et il devient possible de calculer des grandeurs thermodynamiques, telle que l'enthalpie d'interaction d'un complexe récepteur-ligand par exemple, et de les comparer directement à l'expérience.

2. Mécanique classique de Newton

Au cours d'une simulation de DM les positions et les vitesses des particules i du système à l'instant t+1 sont calculées à partir de leurs positions et de leurs vitesses au temps t. Pour cela, les forces agissant sur l'ensemble des particules i du système sont dérivées du gradient de l'énergie potentielle suivant l'équation :

$$F_i = -\frac{dV}{dr_i}$$

Equation 9

Avec :

- F_i la force exercée sur la particule i.
- r_i la position de la particule i.
- V la fonction d'énergie potentielle.

Connaissant les forces agissant sur l'ensemble des particules du système ainsi que leurs masses, on peut alors calculer leurs accélérations en appliquant la seconde loi du mouvement de Newton **:**

$$F_i = m_i a_i = m_i \ddot{r}_i$$

Equation 10

Avec, a_i l'accélération de la particule i.

En combinant les deux équations précédentes, on obtient une relation entre l'énergie potentielle et les positions des particules au cours du temps :

$$-\frac{dV}{dr_i} = m_i \ddot{r}_i$$

Equation 11

L'intégration de cette équation par rapport au temps permet à partir des accélérations déterminées grâce au gradient de l'énergie potentielle, aux positions initiales et aux vitesses initiales, de calculer les positions et les vitesses au temps t+1 et d'obtenir ainsi une trajectoire. Les positions initiales des particules sont les coordonnées cartésiennes issues de structures expérimentales résolues par cristallographie aux rayons X ou par spectroscopie RMN (ou encore issues de structures reconstruites par modélisation par homologie). Les vitesses initiales sont déterminées pour une température donnée T. Elles sont sélectionnées aléatoirement suivant une distribution de Maxwell-Boltzmann qui donne la probabilité qu'une particule i a une vitesse v_{iX} dans la direction X à la température T :

$$pr(v_{iX}) = \left(\frac{m_i}{2\pi k_B T}\right)^{\frac{1}{2}} \exp\left[-\frac{1}{2}\frac{m_i v_{iX}^2}{k_B T}\right]$$

Equation 12

Avec :

- v_{ix} la vitesse de la particule i dans la direction X.
- k_B est la constante de Boltzmann.

Les vitesses sont corrigées de façon à ce que la somme des quantités de mouvement de toutes les particules du système (P) soit nulle (Equation 13). Cela empêche que le système se déplace dans l'espace.

$$P = \sum_{i=1}^{N} m_i v_i = 0$$

Equation 13

3. Intégration numérique de l'équation

L'énergie potentielle est une fonction complexe qui s'exprime en fonction des coordonnées de toutes les particules du système. L'intégration de l'Equation 11 nécessite d'utiliser des méthodes numériques d'intégration. Plusieurs algorithmes ont été développés à cet effet, tels que l'algorithme de Verlet, l'algorithme de Verlet vitesse et l'algorithme de Leap-Frog. Le principe de base de ces algorithmes consiste à développer en série de Taylor les expressions des nouvelles positions, vitesses et accélérations au temps $t + \delta t$ en fonction de celles au temps t.

$$r(t + \delta t) = r(t) + v(t)\delta t + a(t)\frac{\delta t^2}{2} + \cdots$$

$$v(t + \delta t) = v(t) + a(t)\delta t + b(t)\frac{\delta t^2}{2} + \cdots$$

$$a(t + \delta t) = a(t) + b(t)\delta t + c(t)\frac{\delta t^2}{2} + \cdots$$

Equation 14

δ t est le pas d'intégration. La valeur du pas d'intégration est choisie de façon à capturer le mouvement de tous les degrés de liberté du système (liaisons, angles etc.), notamment les mouvements moléculaires les plus rapides. La valeur du pas d'intégration δ t doit être inférieure à la fréquence de Nyqvist du système. Sachant que la fréquence de vibration de liaison la plus élevée est celle de la liaison H-X (avec X étant un hétéroatome) et vaut ≈ 1 fs, il faudrait prendre un pas d'intégration de 0.5 fs afin de capturer le mouvement de toutes les liaisons du système. L'utilisation de l'algorithme SHAKE (Ryckaert et al., 1977) permet d'augmenter le pas d'intégration et de diminuer ainsi le temps de calcul d'une simulation.

L'algorithme SHAKE consiste à figer les liaisons X-H dont les vibrations sont considérées comme ayant un faible impact sur la dynamique globale du système. Il est ainsi possible de considérer un pas d'intégration de 2 fs ce qui permet de simuler l'évolution d'un système sur une durée biologiquement pertinente, en un temps de calcul raisonnable sans perdre en précision.

Algorithme Leap-Frog

Toutes nos simulations de dynamique moléculaire ont été réalisées avec AMBER qui utilise l'algorithme Leap-Frog pour intégrer les équations de mouvement. Contrairement à l'algorithme de Verlet qui n'utilise pas les vitesses pour calculer les nouvelles positions, l'algorithme Leap-Frog utilise les vitesses au temps $t + \frac{1}{2}\delta t$ pour calculer les nouvelles positions au temps $t + \delta t$, ce qui lui procure une plus grande précision. L'algorithme Leap-Frog calcule d'abord la vitesse au temps $t + \frac{1}{2}\delta t$ comme suit :

$$v\left(t + \frac{1}{2}\delta t\right) = v\left(t - \frac{1}{2}\delta t\right) + a(t)\delta t$$

Equation 15

La vitesse au temps $t + \frac{1}{2}\delta t$ est ensuite utilisée pour calculer la position au temps $t + \delta t$:

$$r(t + \delta t) = r(t) + v\left(t + \frac{1}{2}\delta t\right)\delta t$$

Equation 16

L'accélération est ensuite déterminée à partir de l'Equation 11 et un nouveau cycle recommence. Ainsi, les vitesses progressent de $t - \frac{1}{2}\delta t$ à $t + \frac{1}{2}\delta t$ alors que les positions sont déterminées de t à $t + \delta t$. Par conséquent, à chaque cycle de l'algorithme, les positions et les vitesses sont décalées dans le temps comme illustré dans la **Figure 44** et se « dépassent » constamment, d'où le nom de Leap-Frog donné à l'algorithme.

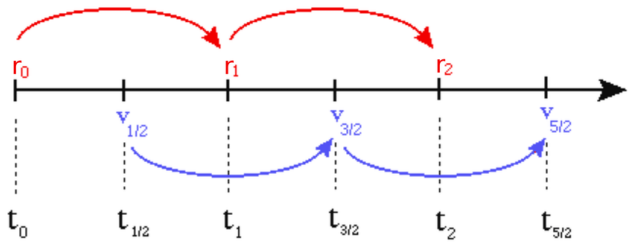


Figure 44 : Schéma de la méthode d'intégration numérique Leap-Frog

L'expression pour le calcul de la vitesse au temps t est la suivante :

$$v(t) = \frac{1}{2}\left[v\left(t - \frac{1}{2}\delta t\right) + v\left(t + \frac{1}{2}\delta t\right)\right]$$

Equation 17

H. Ensemble thermodynamique

Par définition, une simulation de DM est réalisée dans l'ensemble micro-canonique (NVE) où l'énergie totale E du système est constante, à moins de modifier l'Hamiltonien du système pour simuler d'autres ensembles thermodynamiques. La simulation peut être réalisée dans l'ensemble NVT (ou ensemble canonique) où le nombre de particules N, le volume V et la température T sont conservés au cours de la simulation. Dans l'ensemble isotherme-isobare (NPT), ce sont le nombre de particules N, la pression P et la température T qui demeurent constantes. Pour maintenir la température et ou la pression constante(s), il est nécessaire de coupler le système moléculaire à un thermostat et ou à un barostat respectivement.

1. Contrôle de la température : thermostat de Langevin

Il existe principalement trois façons de contrôler la température dans une simulation de dynamique moléculaire :

- En corrigeant les vitesses, comme dans le thermostat de Berendsen
- En ajoutant des forces et ou vitesses stochastiques, comme dans les thermostats d'Andersen ou de Langevin.
- En utilisant un formalisme Lagrangien étendu comme dans le thermostat de Nosé-Hoover.

Pour nos simulations de DM, nous avons utilisé le thermostat de Langevin. Le thermostat de Langevin est le thermostat le plus couramment utilisé pour la simulation des systèmes biologiques. A chaque pas d'intégration δt, une force stochastique $Fs_i(t)$ et une force de

friction $-m_i\gamma_i\dot{r}_i$ sont appliquées à toutes les particules du système, avec m_i la masse de la particule i et γ_i son coefficient de friction. L'équation du mouvement de la particule i est modifiée et l'évolution du système est alors décrite par l'équation de Langevin :

$$F_i = m_i\ddot{r}_i = -\frac{dV}{dr_i} - m_i\gamma_i\dot{r}_i + Fs_i(t)$$

Equation 18

Avec $Fs_i(t)$ qui dépend du coefficient de friction γ_i et de la température T :

$$|Fs_i(t)| = \sqrt{\frac{2m_i\gamma_i k_B T}{\delta t}} \cdot \sigma(t)$$

Equation 19

$\sigma(t)$ représente un nombre aléatoire tiré d'une distribution gaussienne à chaque pas. La force de friction $-m_i\gamma_i\dot{r}_i$ dissipe l'énergie cinétique du soluté en s'opposant au mouvement de la particule. Cette dissipation d'énergie est directement liée aux fluctuations de la vitesse de la particule induites par la force aléatoire $Fs_i(t)$ qui dépend elle même de γ_i et T. Les atomes sont ainsi thermalisés par l'intermédiaire de leurs vitesses, elles-mêmes couplées aux termes de friction γ_i. Un avantage du thermostat de Langevin est qu'il permet l'utilisation d'un pas d'intégration supérieur à celui utilisé lors de simulation NVE sans affecter la stabilité de la simulation. Cela est un atout non négligeable notamment pour la simulation de grands systèmes.

En règle générale, après la minimisation du système, la phase d'équilibration débute par une thermalisation du système dans l'ensemble NVT où la température est ajustée graduellement à la température finale. Il est fortement recommandé de thermaliser le système dans l'ensemble NVT, avant d'ajuster la densité du système dans l'ensemble NPT.

2. Contrôle de la pression : barostat de Monte Carlo

L'ensemble des simulations de DM de cette étude ont été effectuées dans l'ensemble isobare-isotherme (NPT). L'ensemble NPT est celui qui se rapproche le plus des conditions expérimentales en biologie. En réalisant les simulations dans l'ensemble NPT, il est possible de comparer les grandeurs thermodynamiques calculées, avec des données expérimentales mesurées à pression constante. La pression est maintenue constante grâce à un barostat qui agit sur le volume du système en ajustant les positions des particules.

Nous avons utilisé le barostat de Monte Carlo pour maintenir la pression à 1 bar (1 atmosphère). Le barostat de Monte Carlo utilise le critère de Metropolis pour comparer l'énergie de la configuration initiale du système (*old*) à celle générée par l'algorithme de Monte Carlo (*new*) :

$$\chi_{acc} = \min\left[1,\ \left(\frac{V_{new}}{V_{old}}\right)^{N} \exp\left(-\frac{1}{kT}(U_{new} - U_{old} + P(V_{new} - V_{old}))\right)\right]$$

Equation 20

χ_{acc} est la probabilité d'accepter la nouvelle configuration de volume V_{new}. χ_{acc} est déterminée par le produit de deux facteurs. Le premier facteur est le ratio du volume de la nouvelle configuration V_{new} par l'ancienne V_{old} à la puissance N, N étant le nombre de particules du système. Le second facteur est le facteur de Boltzmann basé sur la somme de l'énergie potentielle du système U et de la pression P liés au changement de volume. k est la constante de Boltzmann et T est la température du système imposée (300 K) régulée par un thermostat (celui de Langevin dans notre cas). Comparé aux autres barostats implémentés dans AMBER14, le barostat de Monte Carlo est simple et moins nécessiteux en temps de calcul. De fait, le déplacement de la position des particules nécessite seulement le calcul de l'énergie potentielle du système ainsi que celui de son volume.

I. Dynamique moléculaire dirigée

En dépit des progrès en technologie informatique et malgré notamment l'avènement des cartes GPU, la simulation des processus et l'échantillonnage de l'espace conformationnel restent limités pour les systèmes biologiques. C'est dans ce contexte que nous avons utilisé la dynamique moléculaire dirigée pour simuler, entre autres, le processus de dissociation d'un ligand de sa cible.

La *Targeted molecular dynamics* (TMD) est une technique de simulation permettant de simuler la transition conformationnelle entre deux états : ligand lié-dissocié, conformation ouverte-fermée, etc. (Schlitter et al., 1994). Cette méthode consiste à contraindre la structure simulée à adopter une certaine conformation en utilisant une structure de référence. Pour cela, un terme supplémentaire est ajouté à la fonction d'énergie du champ de forces de la forme d'un potentiel harmonique :

$$V_{potentielle} + V^{contrainte}$$

$$V^{contrainte} = \frac{1}{2} \times k \times N_{atomes} \times \left(RMSD^{actuel} - RMSD^{référence}\right)^2$$

Equation 21

Où :

- k est la constante de force du potentiel harmonique.
- N_{atomes} est le nombre d'atomes sur lequel le RMSD est calculé.
- $RMSD^{actuel}$ est la déviation moyenne quadratique (RMSD) de la structure à l'instant t par rapport à la structure de référence.
- $RMSD^{référence}$ est le RMSD demandé par rapport à la structure de référence.

Le RMSD permet de mesurer la distance entre deux structures tridimensionnelles S_1 et S_2 d'une molécule. Dans la TMD implémentée dans AMBER, les coordonnées atomiques sont pondérées par la masse lors du calcul du RMSD :

$$RMSD = \sqrt{\frac{\sum_{i=0}^{N_{atomes}} [m_i \times (\vec{r}_{1i} - \vec{r}_{2i})^2]}{\sum_{i=0}^{N_{atomes}} m_i}}$$

Equation 22

Où :

- m_i est la masse de l'atome i.
- $\vec{r}_{1i}$ est la position de l'atome i dans la structure S_1.
- $\vec{r}_{2i}$ est la position de l'atome i dans la structure S_2.

La valeur du $RMSD^{actuel}$ est contrainte à une valeur définie par l'utilisateur à savoir le $RMSD^{référence}$. Le $RMSD^{référence}$ varie progressivement d'une valeur initiale à une valeur finale. Ainsi, $V^{contrainte}$ permet de contraindre la structure simulée, ayant un $RMSD^{actuel}$ donné, à atteindre les valeurs successives de $RMSD^{référence}$.

On peut utiliser la dynamique dirigée de deux façons (**Figure 45**) :

a) TMD directe

Dans la TMD directe, la structure de référence correspond à la structure vers laquelle on souhaite faire converger notre structure initiale. Ainsi, la valeur du $RMSD^{référence}$ décroit progressivement au cours de la simulation : sa valeur initiale correspond au RMSD entre la structure initiale et la structure de référence et sa valeur finale est proche de 0. C'est ce

protocole qui a été utilisé pour simuler le changement de conformation du complexe CDK8-CycC pour passer d'un état à l'autre dans le chapitre II.

b) TMD inverse (TMD^{-1})

Dans la TMD^{-1}, la même structure joue à la fois le rôle de structure initiale et de structure de référence. La valeur du $RMSD^{référence}$ augmente alors progressivement au cours de la simulation : sa valeur initiale est de 0 et sa valeur finale est choisie par l'utilisateur. La TMD^{-1} n'a pas pour but de faire converger une structure vers une autre, mais de contraindre une structure à s'éloigner de sa conformation initiale sans lui imposer une direction particulière. Ainsi, dans le cas de la simulation du processus de dissociation d'un ligand de son site de liaison, utiliser la TMD^{-1} au lieu de la TMD directe permet de laisser le ligand choisir librement son chemin de sortie. C'est ce protocole qui a été utilisé pour la simulation du processus de dissociation.

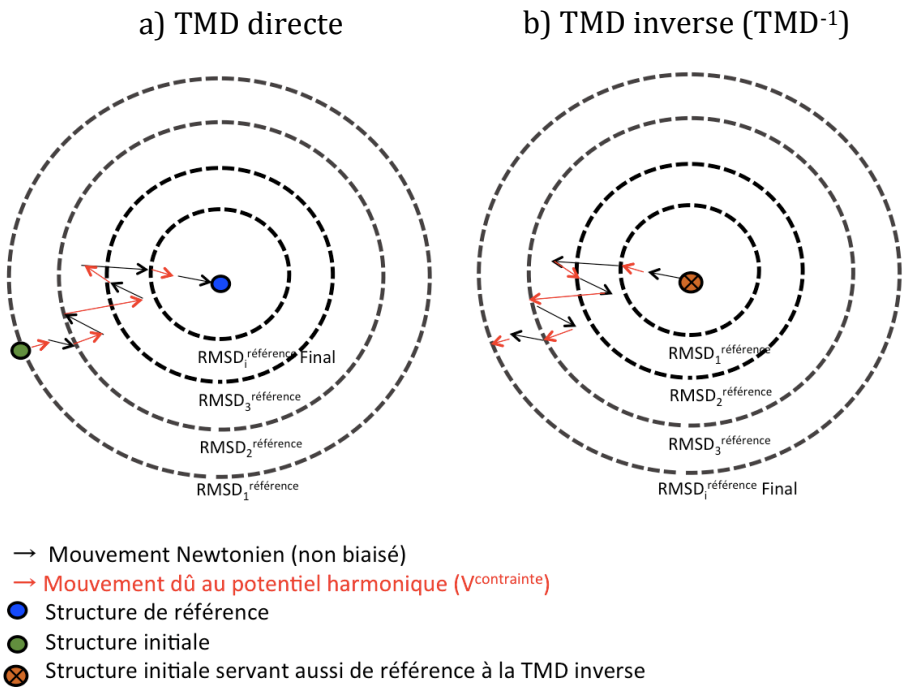


Figure 45 : Représentation schématique du fonctionnement de la TMD.
A chaque cercle i représenté en pointillé, est associé une valeur de $RMSD_i^{référence}$. $V^{contrainte}$ a pour but de maintenir la valeur de $RMSD_i^{actuel}$ proche de $RMSD_i^{référence}$.

Dans nos protocoles de TMD, le $RMSD^{référence}$ ne varie pas à chaque pas de temps, mais toutes les m étapes. Cela permet de laisser le système s'équilibrer, voir même échantillonner (si m est assez grand) autour de chaque valeur de $RMSD^{référence}$. Une trajectoire de TMD peut donc

être vue comme un ensemble de n trajectoires successives numérotées par i (avec i = 1, 2 … n) où le potentiel harmonique $V_i^{contrainte}$ a pour but de maintenir la valeur de $RMSD_i^{actuel}$ proche du $RMSD_i^{référence}$. Le protocole de TMD s'apparente à une simulation de type *umbrella sampling*, sauf qu'au lieu d'utiliser une distance euclidienne en tant que coordonnée de réaction, on utilise un RMSD. Le choix du RMSD permet d'imposer une contrainte « plus douce » sur le système qu'une distance. En effet, augmenter la distance euclidienne entre le centre de masse du site de liaison et le centre de masse d'un ligand implique nécessairement que le ligand s'éloigne du site de liaison par déplacement de son centre de masse. Or, cela n'est pas le cas avec le RMSD. Si on considère le RMSD calculé sur les atomes du ligand, l'augmentation du RMSD peut se faire uniquement par changement de conformation du ligand sans nécessairement déplacer son centre de masse.

J. Energie libre de liaison

Déterminer l'énergie libre de liaison (énergie de Gibbs ou enthalpie libre) est crucial lorsque l'on souhaite comprendre la relation entre la structure, la fonction et la stabilité de biomolécules. Plusieurs approches numériques ont été développées pour le calcul de l'énergie libre de liaison, chacune d'entre elles représente un compromis entre coût de calcul et précision.

Les méthodes rapides de prédiction de l'énergie libre utilisent des fonctions énergétiques simples, empiriques ou basées sur la connaissance (*knowledge-based approach*) (Böhm, 1994; Jain, 1996). L'absence d'échantillonnage conformationnel rend ces méthodes plus rapides mais cela au détriment de la précision. Les méthodes les plus coûteuses en temps de calcul et les plus précises sont celles basées sur des champs de forces moléculaires et utilisent des simulations de DM ou de Monte Carlo pour générer des ensembles conformationnels. Dans cette catégorie, on peut citer les méthodes alchimiques telles que la perturbation d'énergie libre (*free energy perturbation*) (Kollman, 1993) et l'intégration thermodynamique (*thermodynamic integration*) (Gouda et al., 2003). Elles donnent des estimations précises de l'énergie libre mais sont coûteuses en temps de calcul, ce qui les rend difficilement applicable à de grands complexes protéiques. Dans ce contexte, des méthodes *end-point* (c'est à dire qu'elles ne dépendent que de l'évaluation des états initiaux et finaux, et non du chemin suivi), moins coûteuses en temps de calcul mais de précision correcte, ont été développées telles que l'approche d'énergie d'interaction linéaire (LIE pour *linear interaction energy*) (Aqvist et al.,

1994) et les approches MM-PBSA (*Molecular Mechanics–Poisson Boltzmann Surface Area*) (Fogolari et al., 2002) et MM-GBSA (*Molecular Mechanics–Generalized-Born Surface Area*) (Kollman et al., 2000). Les méthodes MM-GBSA et MM-PBSA combinent les énergies mécaniques moléculaires en phase gazeuse, l'approche de Poisson–Boltzmann ou de Born Généralisé pour évaluer la contribution polaire de l'énergie de solvatation, et une fonction empirique pour prendre en compte l'énergie liée à la surface accessible au solvant (la contribution non-polaire de l'énergie de solvatation). Elles consistent à analyser un ensemble de conformations issues de simulations de DM. Il est important de noter que même si les simulations de DM ont été effectuées avec un modèle de solvant explicite (comme c'est le cas dans nos études), celui-ci n'est pas utilisé et est remplacé par l'usage d'un modèle de solvant implicite lors du calcul de l'énergie libre par MM-PBSA ou MM-GBSA. La représentation implicite du solvant réduit considérablement les coûts de calcul qui deviennent souvent prohibitifs, en particulier pour les grands systèmes ou les simulations de longue durée. Elle permet également de diminuer les erreurs statistiques résultant d'un échantillonnage incomplet des conformations du solvant.

1. MM-GBSA : principe

La méthode MM-GBSA, plus rapide que la méthode MM-PBSA, estime l'enthalpie libre de liaison entre deux états en solution (lié et non-lié par exemple). Elle a été largement utilisée pour étudier les complexes protéine-ligand, protéine-protéine, et protéine-peptide. Cette approche calcule l'énergie libre de liaison ($\Delta G_{liaison}$), à partir d'une somme de composantes énergétiques en utilisant un cycle thermodynamique :

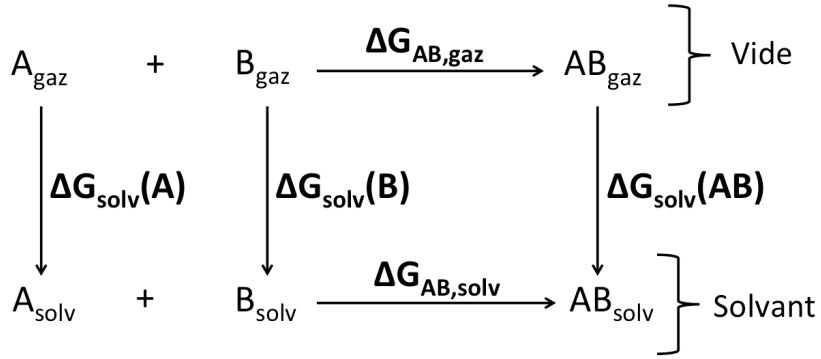


Figure 46 : Cycle thermodynamique pour calculer l'énergie libre de liaison du complexe AB en solution $\Delta G_{solv}(AB)$
$\Delta G_{solv}(AB)$ est décomposée en énergie libre de liaison en phase gazeuse (ou dans le vide, noté $\Delta G_{gaz}(AB)$) et en énergies de solvatation des espèces en réaction.

D'une part, l'énergie libre de solvatation d'une espèce X s'exprime de la façon suivante :

$$X_{gaz} \xrightarrow{\Delta G_{solv}(X)} X_{solv}$$

$$\Delta G_{solv}(X) = G(X)_{solv} - G(X)_{gaz}$$

Equation 23

Où $G(X)_{solv}$ est l'énergie libre de X dans le solvant. $G(X)_{gaz}$ est l'énergie libre de X dans le vide.

A partir de l'Equation 23, on peut exprimer l'énergie libre de X dans le solvant :

$$G(X)_{solv} = \Delta G_{solv}(X) + G(X)_{gaz}$$

Equation 24

D'autre part, l'énergie libre de liaison du complexe AB s'écrit :

$$\Delta G_{AB,solv} = G(AB)_{solv} - G(A)_{solv} - G(B)_{solv}$$

Equation 25

En exprimant $G(AB)_{solv}$, $G(A)_{solv}$ et $G(B)_{solv}$ en fonction de l'Equation 24, l'Equation 25 peut s'écrire :

$$\Delta G_{AB,solv} = \left[\Delta G_{solv}(AB) + G(AB)_{gaz}\right] - \left[\Delta G_{solv}(A) + G(A)_{gaz}\right] - \left[\Delta G_{solv}(B) + G(B)_{gaz}\right]$$

$$\Delta G_{AB,solv} = \Delta G_{AB,gaz} + \underbrace{\Delta G_{solv}(AB) - \Delta G_{solv}(A) - \Delta G_{solv}(B)}_{\Delta\Delta G_{AB,solv}}$$

Equation 26

Ainsi dans la méthode MM-GBSA ou MM-PBSA, l'énergie libre de liaison $\Delta G_{AB,solv}$ est calculée à partir de l'énergie libre de liaison dans le vide $\Delta G_{AB,gaz}$ et un terme de solvatation $\Delta\Delta G_{AB,solv}$ sur la base du cycle thermodynamique.

Sachant que le calcul est effectué à partir d'un ensemble de conformations issues d'une trajectoire de DM, l'équation s'écrit alors :

$$\Delta G_{AB,solv} = \langle \Delta E_{MM} \rangle + \Delta\Delta G_{solv} - T\Delta S$$

Equation 27

Où $\langle \Delta E_{MM} \rangle$ est l'énergie potentielle moyenne issue de la mécanique moléculaire (décrite à l'Equation 5), ΔS est la contribution entropique et T la température absolue exprimée en Kelvin.

a) Contribution entropique

Le terme entropique ΔS est la somme de la variation entropique translationnelle (ΔS_{tr}),

rotationnelle (ΔS_{rot}) et vibrationnelle (ΔS_{vib}) :

$$\Delta S = \Delta S_{tr} + \Delta S_{rot} + \Delta S_{vib}$$

Equation 28

L'estimation de l'entropie peut se faire par une approximation quasi-harmonique ou par une analyse des modes normaux à partir des trajectoires de DM. Le calcul de l'entropie est coûteux en temps de calcul et peut être affecté par une grande incertitude statistique. Par conséquent, il est d'usage de négliger le terme entropique lorsqu'il s'agit de comparer des systèmes similaires (par exemple dans des études de mutation ou lors de la comparaison de ligands interagissant avec la même cible protéique).

b) Contribution de la solvatation

Dans la méthode de MM-GBSA, l'énergie de solvatation ($\Delta\Delta G_{solv}$) est définie comme la somme de la contribution non-polaire de l'énergie de solvatation (ΔG_{SA}) et la contribution polaire de l'énergie de solvatation (ΔG_{GB}) :

$$\Delta\Delta G_{solv} = \Delta G_{SA} + \Delta G_{GB}$$

Equation 29

- ΔG_{SA} est attribuée à la formation de cavité dans le solvant et aux interactions de van der Waals entre le soluté et le solvant. ΔG_{SA} se définit par l'expression suivante :

$$\Delta G_{SA} = \gamma A + b$$

Equation 30

Où A est la surface totale accessible au solvant, et γ et b sont des constantes empiriques.

- ΔG_{GB} est la composante électrostatique de l'énergie de solvatation. Elle est liée à la création d'une charge q_i dans une cavité sphérique de rayon a_i dans un solvant de constante diélectrique ε. Pour N charges situées à une distance r_{ij}, l'équation utilisée pour la définition de ΔG_{GB} est la suivante :

$$\Delta G_{GB} = -\frac{1}{2}\left(1 - \frac{1}{\varepsilon}\right) \sum_{i,j\ (i \neq j)}^{N} \frac{q_i q_j}{f(r_{ij}, a_{ij})}$$

Equation 31

Avec :

- $f(r_{ij}, a_{ij}) = \sqrt{r_{ij}^2 + a_{ij}^2 e^{-D}}$
- $a_{ij} = \sqrt{a_i a_j}$
- $D = \frac{r_{ij}^2}{(2a_{ij})^2}$

2. Méthode d'analyse par histogramme pondéré (WHAM)

La méthode d'analyse par histogramme pondéré ou WHAM (W*eighted Histogram Analysis Method)*(Kumar et al. 1992; Roux 1995) est une technique utilisée pour estimer un profil d'énergie libre associé à un processus (dissociation d'un ligand de sa protéine, transition d'une conformation ouverte à fermée etc.) ayant été simulé par DM biaisée. L'énergie libre est approximée par un potentiel de force moyenne qui est calculé en fonction d'une ou plusieurs coordonnées de réaction ξ. La variation de la coordonnée de réaction ξ doit permettre de décrire le processus à simuler. La méthode consiste à :

1) réaliser une série de n simulations sous une contrainte de type potentiel harmonique ($V^{contrainte}$ noté ici w, cf. Equation 32). Dans notre cas, nous avons réalisé des simulations de TMD inverse ; ξ est donc le RMSD. Pour chacune de ces simulations i, le potentiel de contrainte $w_i(\xi)$ sert à confiner les variations de ξ autour d'une valeur prédéfinie $\xi_i^{référence}$. Cela permet d'échantillonner efficacement autour de la région $\xi_i^{référence}$. On parle de fenêtre pour désigner chacune de ces simulations i.

$$w_i(\xi) = \frac{1}{2} \times K \times \left(\xi - \xi_i^{référence}\right)^2$$

Equation 32

Où K est la constante de force.

2) calculer la distribution de probabilité biaisée $P_i^b(\xi)$ de chaque simulation i. $P_i^b(\xi_j)$ est la probabilité (biaisée) que le système atteigne la valeur ξ_j dans la fenêtre i. Comme on connait la contrainte $w_i(\xi)$ appliquée à chaque fenêtre i, il est possible de «débiaiser » chacune des distributions obtenues et de déterminer ainsi les distributions de probabilité non-biaisées $P_i^u(\xi)$ de chaque fenêtre i :

$$P_i^u(\xi) = P_i^b(\xi).\exp[-\beta w_i(\xi)].\langle \exp[\beta w_i(\xi)] \rangle$$

Equation 33

Avec $\beta = 1/k_bT$, k_b étant la constante de Boltzmann et T la température.

3) calculer la distribution globale $P^u(\xi)$ grâce aux équations de WHAM qui permettent de combiner les histogrammes $P_i^u(\xi)$ de chaque fenêtre i, pondérés par un facteur $\rho_i(\xi)$:

$$P^u(\xi) = \sum_i^n \rho_i(\xi).\, P_i^u(\xi)$$

Equation 34

Avec :

- n le nombre total de fenêtre.
- $\rho_i = \frac{a_i}{\sum_i^n a_i}$.
- $a_i(\xi) = N_i \exp[-\beta w_i(\xi) + \beta F_i]$.
- N_i est le nombre total d'échantillons générés à la fenêtre i.

F_i est calculé comme suit :

$$\exp(-\beta F_i) = \int P^u(\xi) \exp[-\beta w_i(\xi)] d\xi$$

Equation 35

4) déduire l'énergie libre exprimée en fonction de la coordonnée de réaction ξ. Une fois les deux équations précédentes itérées sur toutes les fenêtres, le potentiel de force moyenne non biaisé, correspondant à l'énergie libre $A(\xi)$, est calculé à partir de $P^u(\xi)$ comme suit :

$$A(\xi) = -k_b T \ln(P^u(\xi))$$

Equation 36

Plusieurs paramètres peuvent avoir une influence sur la précision de la valeur de $A(\xi)$ et donc la pertinence du profil :

- La valeur de la constante de force K du potentiel harmonique $w_i(\xi)$ doit être assez grande pour permettre au système d'échantillonner correctement autour de $\xi_i^{référence}$, surtout lorsque $\xi_i^{référence}$ correspond à un état de haute énergie, comme par exemple à un état de transition. Cependant une valeur trop grande de K peut conduire à un échantillonnage très étroit autour de $\xi_i^{référence}$. Cela aura pour conséquence de limiter le recouvrement entre deux distributions de ξ consécutives i et i+1 ce qui nécessitera

de multiplier les fenêtres pour obtenir un échantillonnage continu tout au long de ξ. Ainsi la constante de force K et le nombre de fenêtres n sont des paramètres importants et interdépendants, à choisir avec précaution.

- Le temps de simulation doit être le même pour toutes les fenêtres et doit être assez long pour permettre une convergence de l'échantillonnage.

Pour s'assurer du bon paramétrage des simulations, il faut vérifier :

- Le recouvrement entre deux distributions de ξ consécutives i et i+1. Plus le recouvrement est important meilleur sera l'estimation du profil d'énergie libre (Kästner, 2011).
- La normalité des distributions. De fait, une distribution normale garantit une distribution unimodale et symétrique autour d'une valeur moyenne ξ. Si la distribution est bimodale, cela signifie qu'il faut augmenter le nombre de fenêtres ou la constante de force K pour confiner l'échantillonnage autour d'une valeur moyenne.
- L'échantillonnage. Il doit se faire autour de $\xi_i^{\text{référence}}$ dans les régions de basses énergies. Pour l'évaluer, il faut calculer l'écart entre la valeur moyenne de ξ à la fenêtre i (ξ_i^{moyenne}) et la valeur de référence ($\xi_i^{\text{référence}}$). Celui-ci doit être quasi-nulle. Si ce n'est pas le cas, il faut augmenter la valeur de la constante de force K. De fait, si le système ne parvient pas à échantillonner correctement autour de $\xi_i^{\text{référence}}$ lorsque celle-ci correspond à une région de basse énergie, alors il ne parviendra pas à explorer les états de transition qui sont des états instables de hautes énergies.

K. Forêts aléatoires

Les forêts aléatoires ou forêts d'arbres décisionnels (*Random forest*, RF) sont un algorithme d'apprentissage automatique supervisé applicable aussi bien pour construire un modèle de régression qu'un modèle de classification. Le RF consiste à construire plusieurs modèles statistiques, dit arbres décisionnels. Un arbre est composé de plusieurs nœuds correspondant chacun à une variable. Chacun des arbres se construit comme suit. A chaque nœud de l'arbre, un sous-ensemble de variables est tiré aléatoirement, avec remise (*bootstrap*), et chacune de ces variables est testée. La variable qui fournit le meilleur découpage des données suivant la réponse étudiée Y est sélectionnée au nœud considéré. La qualité du découpage est mesurée par une fonction de coût que l'on cherche à minimiser dans les deux sous-ensembles issus du découpage (nœuds adjacents). Cette fonction de coût est la variance dans le cas d'une

régression et le coefficient d'impureté de Gini dans le cas d'une classification. On construit ainsi plusieurs arbres indépendants car basés sur des sous-ensembles de variables tirés aléatoirement. Le nombre d'arbres à considérer (*ntree*) ainsi que le nombre de variables (*mtry*) à tester à chaque nœud sont les hyper-paramètres du modèle et doivent donc être choisis *a priori.* Les observations à prédire sont ensuite calculées pour chaque arbre. Le résultat final de la prédiction correspondant à l'agrégation (*aggregating*) des résultats obtenus pour chaque arbre : c'est un vote pour une classification et une moyenne des valeurs pour une régression. L'algorithme du RF est connu sous le nom de *bagging (bagging = bootstrap + aggregating)* car il contient un échantillonnage de type *bootstrap* et une étape d'agrégation (*aggregating*).

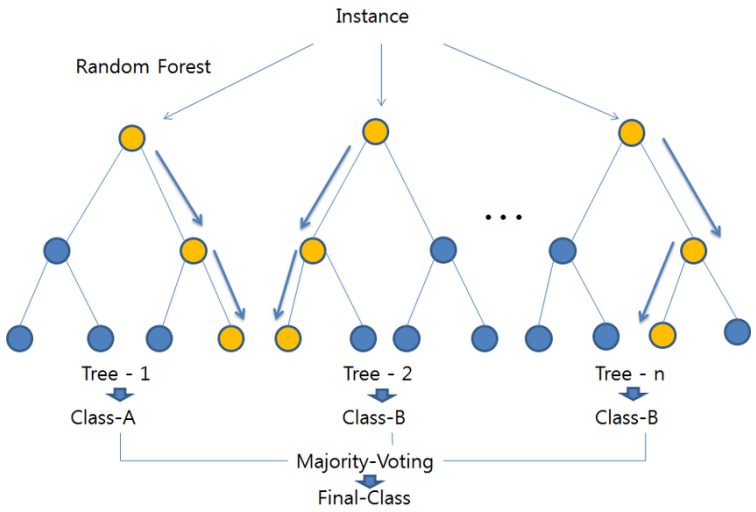


Figure 47 : Représentation schématique de la méthode de forêt d'arbres décisionnels (Dimitriadis et al., 2018).

L. Transformée de fourier discrète

La transformée de Fourier discrète ou DFT (*Discret Fourier Transform*) est une méthode qui permet de décomposer un signal discret en ondes sinusoïdales de différentes fréquences. Dans notre étude, la DFT est appliquée sur une série de valeurs temporelles (énergie libre calculée par la méthode MM-GBSA à intervalle de temps régulier). Il s'agit donc d'un signal discret. La DFT permet de passer d'une représentation temporelle du signal, c'est à dire son évolution en fonction du temps, à une représentions fréquentielle qui décrit l'amplitude des différentes composantes sinusoïdales en fonction de leurs fréquences (**Figure 48**).

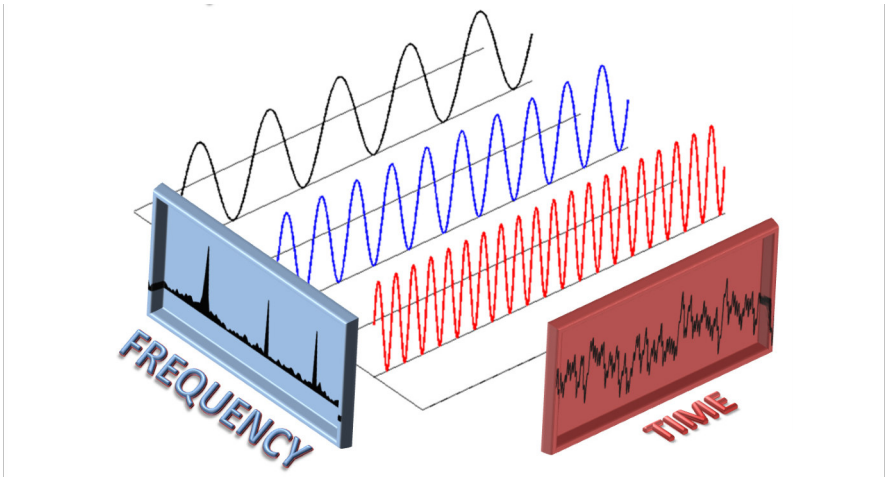


Figure 48 : Schéma du principe de la transformée de Fourier discrète
Le signal dépendant du temps est décomposé en ses composantes sinusoïdales dans l'espace des fréquences.

Ce sont ces amplitudes sinusoïdales à des fréquences données que l'on appelle coefficients de la DFT. Il y a autant de coefficients que de valeurs dans le signal. Plus l'amplitude à une fréquence donnée d'une composante sinusoïdale (⇔la valeur du coefficient) est importante plus la composante contribue à décrire le signal. Les coefficients de basse fréquence traduisent les fluctuations de fortes amplitudes telles que la tendance du signal et ceux de hautes fréquences traduisent les fluctuations de faible amplitude tel que le bruit. La DFT est considérée comme une méthode de réduction de dimension au même titre que l'analyse en composante principale. La DFT est aussi couramment utilisée pour débruiter un signal en reconstruisant le signal par DFT inverse à partir uniquement des composantes sinusoïdales d'intérêt.

Sachant un signal s de N valeurs, les coefficients de la DFT notés S(k) se calculent de la façon suivante :

$$S(k) = \sum_{n=0}^{N-1} s(n) \exp\left(-2i\pi k \frac{n}{N}\right) \qquad \text{avec} \quad 0 \leq k < N$$

Equation 37

A partir de l'ensemble des coefficients d'un signal il est possible de reconstruire le signal d'origine. La DFT permet ainsi de fournir une description résumée et complète du signal sous forme de variables discrètes (les coefficients) qui décrivent des motifs temporels complexes. A partir de ces variables discrètes descriptives de chacun des signaux il devient possible d'appliquer des méthodes d'apprentissage automatique en vue de classifier les signaux par

exemple.

VI. CONCLUSION

De nombreux candidat-médicaments présentant des résultats *in vitro* prometteurs échouent lors d'expériences *in vivo* ou sont retirés des essais cliniques en raison d'un manque d'efficacité ou d'innocuité. Ce constat a poussé la communauté scientifique à remettre en cause les critères de sélection utilisés en phase précoce du processus de découverte du médicament. Un ensemble d'études a mis en évidence qu'une évaluation préclinique des constantes cinétiques d'association et de dissociation d'un complexe récepteur-ligand permet de limiter les taux d'échec en phase II d'essais cliniques. Aujourd'hui, la cinétique de liaison est considérée comme un critère essentiel de sélection et de priorisation de molécules en phase précoce du processus de recherche. La constante de dissociation d'une molécule de sa cible a fait l'objet d'un attrait particulier depuis que plusieurs études ont montré qu'elle est significativement corrélée à son efficacité clinique. En effet la constante de dissociation d'une molécule de sa cible contrôle directement son temps de résidence qui est la durée pendant laquelle la molécule active reste liée à sa cible thérapeutique et donc induit un effet thérapeutique.

L'objectif de ce projet de thèse est de répondre à un besoin formulé par l'Institut de Recherche Servier (IdRS), porteur de ce projet, qui souhaite disposer d'un outil basé sur des simulations de dynamique moléculaire pour prédire le temps de résidence de molécules actives. L'outil doit permettre de fournir un classement prédictif d'une série de molécules actives selon leur temps de résidence en phase précoce de sélection et d'optimisation des molécules. Pour un usage industriel en routine, l'outil doit être simple, rapide et applicable à tout système.
En parallèle, une des thématiques de l'équipe SB&C porte sur la prédiction quantitative des constantes cinétiques basée sur les simulations de dynamique moléculaire. Un deuxième volet de ce travail de thèse porte sur la mise au point d'un protocole permettant d'accéder à de telles grandeurs.

CDK8, une protéine kinase cliniquement pertinente car impliquée dans différents cancers, a été choisie pour établir la preuve de concept de l'outil. Nous avons tout d'abord été confronté à une particularité structurale, CDK8 qui se complexe à la CycC sous sa forme inactive. Cette particularité est une singularité au sein de la famille des CDKs humaines. De fait, la cycline se lie à CDK pour l'activer et former ainsi le complexe CDK-Cyc actif. Nous nous sommes alors

questionnés quant au rôle de la cycline, dans le complexe CDK8-CycC inactif et sur la nécessité de la conserver ou non dans le système pour nos simulations. Par le biais de simulations de dynamique moléculaire et de calculs d'énergie libre de liaison, le rôle de la CycC et son impact sur la structure et la dynamique de CDK8 dans plusieurs cas (forme active et inactive, forme apo ou holo, mutations) a été étudié. Nous avons mis en évidence que la CycC stabilise la structure de CDK8 dans ses formes active et inactive. Elle est indispensable pour maintenir la structure native de CDK8 observées dans le complexe CDK8-CycC, et a un impact déterminant dans le comportement dynamique de cette protéine kinase. Les résidus contribuant fortement à l'interaction CDK8-CycC ont été mis en évidence. Faits intéressants, parmi les résidus importants identifiés, certains appartiennent à des sites d'interaction spécifiques à CDK8 tandis que les autres résidus se trouvent sur des sites d'interaction communs à la famille des CDKs humaines. Compte tenu de l'émergence récente de CDK8 en tant que cible d'intérêt pour le cancer colorectal et mammaire, ces résultats sont très utiles pour la conception d'inhibiteurs peptidiques ciblant spécifiquement l'interface CDK8-CycC. Enfin, nous avons simulé le changement de conformation de la forme inactive à la forme active du complexe CDK8-CycC au moyen de simulations de dynamique moléculaire dirigée. Les analyses montrent que le changement de conformation s'accompagne d'une légère rotation de la CycC. Ce léger déplacement semble indispensable pour stabiliser la forme active de CDK8, via une interaction critique avec un résidu de la CycC. Le rôle de ce résidu dans la stabilisation du complexe actif est normalement assuré par un résidu phosphorylé de CDK, absent dans le cas de CDK8. Notre étude a mis en lumière le rôle crucial de la CycC dans le complexe CDK8-CycC et la nécessité de la prendre en compte dans les études de modélisation de CDK8.

La méthode d'estimation du temps de résidence développée est basée sur la dynamique moléculaire dirigée. L'outil a été développé et appliqué sur un ensemble de 10 inhibiteurs d'arylpyrazole du complexe CDK8-CycC dont les temps de résidence expérimentaux varient de <1,4 à 1944 min. A partir des simulations de dissociations de ces complexes, nous avons établi un estimateur du temps de résidence, basé sur l'énergie de contrainte appliquée au système le long du processus. La méthode a été capable de classer correctement les inhibiteurs par rapport aux données expérimentales. L'outil a ensuite été testé sur un jeu de donnée privé appartenant à l'Institut de Recherche Servier et associé à un projet de recherche en cours.

Nous souhaitions vérifier que la méthode développée permet d'obtenir un classement correct des composés selon leur temps de résidence sur un autre jeu de données que celui sur lequel elle a été développée. Il s'agissait également de voir si la précision de la méthode ainsi que sa performance permettaient de l'intégrer comme outil utilisable en routine dans un cadre industriel pour fournir un classement prédictif des chefs de file en phase d'optimisation. Grâce à notre méthode, nous avons fourni un classement correct de 18 molécules sur 21 en fonction de leurs temps de résidence avec un coefficient de détermination de 85%. L'une des exigences du cahier des charges concerne le coût en temps de calcul qui devait être suffisamment faible pour convenir à un usage industriel, où des dizaines de composés doivent être priorisés lors des phases d'optimisation de *lead*. En ayant eu à disposition 160 CPUs, nous avons pu simuler 11 répliques de 5 composés par jour, ce qui est acceptable pour notre partenaire industriel. De plus, un autre avantage de notre méthode réside dans sa simplicité et dans le fait qu'elle ne nécessite aucune connaissance spécifique *a priori* du chemin emprunté par le ligand lors de sa dissociation. La coordonnée de réaction utilisée (le RMSD) présente l'avantage d'induire des modifications douces, offrant ainsi une description plus réaliste du processus de dissociation. Une limite possible de cette méthode est que le choix de la constante de force est directement dépendant du système que l'on veut étudier. Ainsi, la méthode proposée a montré des résultats forts encourageants sur deux jeux de données distincts. De plus, elle est adaptée à un usage industriel et permet de fournir un classement prédictif du temps de résidence de molécules d'intérêt grâce au RT_{score}. Il serait donc intéressant de vérifier encore plus la robustesse de notre méthode en l'appliquant sur des jeux de données plus importants, impliquant des protéines kinases ou d'autres familles de protéines.

Nous nous sommes ensuite intéressés à l'étude des relations structure-cinétiques du jeu de données des 10 inhibiteurs d'arylpyrazole du complexe CDK8-CycC. Cette étude a été conduite par deux approches différentes et complémentaires.

La première approche consiste en l'analyse des interactions protéine-ligand des trajectoires de dissociation de ces inhibiteurs. Cette analyse a permis d'identifier des interactions affectant fortement le temps de résidence. Nos résultats soulignent l'importance des interactions hydrophobes ainsi que les interactions hydrogènes avec deux motifs conservés des protéines kinases : la boucle P et le motif DMG. Cette étude SKR pourrait être d'une aide

précieuse pour la conception d'inhibiteurs de CDK8-CycC avec un profil cinétique optimisé et donc un profil *in vivo* amélioré.

Dans la seconde approche, nous avons calculé les différents termes d'interaction protéine-ligand (E_{VDW}, E_{EEL}, E_{GB}, E_{NP}, G_{total}) par la méthode MM-GBSA à partir des simulations de DM biaisées décrivant la dissociation des inhibiteurs. A partir de ces profils énergétiques, qui sont des séries temporelles, un grand nombre de caractéristiques est calculé de façon à décrire finement les profils et à capturer des schémas temporels complexes. Cet ensemble de caractéristiques est ensuite soumis à une méthode d'apprentissage automatique pour en dériver un modèle QSKR. Il s'agit de comprendre quelle(s) composante(s) énergétique(s) de l'interaction protéine-ligand pourrai(en)t contribuer le plus à expliquer le temps de résidence. Cette étude QSKR a montré que le temps de résidence de la série d'inhibiteurs étudiée est modulé majoritairement par la contribution non polaire de l'énergie de solvatation. Ces résultats sont en accord avec les résultats issus de l'analyse des interactions protéine-ligand au cours du processus de dissociation, dans laquelle nous avons mis en évidence le rôle clé des contacts hydrophobes dans l'augmentation du temps de résidence. L'ensemble de ces résultats est également en accord avec la littérature, relatant l'importance de l'effet du solvant et donc, la solvatation du ligand, dans la modulation du temps de résidence. Par la suite, on pourra envisager d'utiliser un tel modèle pour prédire le temps de résidence d'un inhibiteur de CDK8 à partir de ses profils énergétiques calculés par la méthode MM-GBSA, issus de la trajectoire du processus de dissociation. D'une façon générale, nous pensons que ce protocole d'analyse de séries temporelles est une approche prometteuse qui pourrait être employée pour extraire des informations différentes que celles obtenues par des analyses classiques.

La méthode d'estimation du temps de résidence est une approche rapide qui ne permet pas un échantillonnage fin du chemin de dissociation. Ainsi, dans cette partie plus exploratoire, nous nous sommes proposés de caractériser plus finement le chemin de dissociation. L'objectif est d'obtenir quantitativement une estimation du temps de résidence, et d'identifier les états stables tels que les états intermédiaires, les états de transition, ainsi que les étapes cinétiquement déterminantes, en se basant sur les profils d'énergie libre. Dans le cadre d'un projet de découverte de médicaments, cette connaissance est d'une grande importance pour optimiser la cinétique et l'affinité d'un *lead*. Nous avons mis au point un protocole en deux étapes qui consiste à ré-échantillonner un chemin, obtenu à l'issue d'une 1ère étape, et à appliquer la méthode WHAM, pour calculer un potentiel de force moyenne à partir des

distributions de la coordonnée de réaction extraites du ré-échantillonnage. Ce protocole est basé sur la dynamique moléculaire dirigée. Le paramétrage de la méthode a nécessité plusieurs tests avant l'obtention de distributions pertinentes, aboutissant à des profils d'énergie libre corrects et exploitables, desquels nous pouvons extraire des grandeurs thermodynamiques. De fait, la différence entre l'état lié et dissocié des profils d'énergie libre est corrélée au temps de résidence. Cependant, considérant notre jeu de donnée particulier, dans lequel l'affinité évolue dans le même sens que le temps de résidence, nous ne pouvons pas conclure sur une corrélation directe entre nos barrières d'énergies et le temps de résidence. Il serait intéressant d'appliquer le protocole sur un autre jeu de données, dans lequel le temps de résidence et l'affinité n'évolueraient pas dans le même sens. Ceci permettrait de vérifier s'il y a une corrélation réelle entre nos barrières d'énergie et le temps de résidence. D'autre part, sur certains profils l'état de transition n'est pas visible. D'une façon générale, il faudrait donc optimiser le paramétrage de notre approche, par exemple en cherchant à augmenter l'échantillonnage du chemin et en faisant des analyses supplémentaires, telles que l'estimation de la convergence, pour affiner les profils, les optimiser et en extraire des valeurs prédites de constantes cinétiques. Comme ce protocole génère beaucoup de trajectoires, il serait intéressant de le coupler à un modèle de Markov, dans lequel nos simulations de ré-échantillonnage permettraient d'enrichir le modèle mathématique. Cette méthode permet de clairement identifier les états cinétiquement déterminants et calcule les probabilités de transition entre les différents états. Cela pourrait être comparé à nos profils afin d'en vérifier la pertinence.

En conclusion, nous avons réussi à réaliser le premier objectif de ce travail, c'est-à-dire la mise en place d'une méthode de prédiction qualitative du temps de résidence. De par sa simplicité et sa rapidité d'exécution, elle est tout à fait adaptée à un usage routinier en industrie. Ainsi, nous avons développé un outil innovant et efficace, permettant d'estimer qualitativement la constante cinétique de dissociation d'un complexe ligand-récepteur. Ce nouvel outil ouvre des perspectives intéressantes pour l'optimisation des chefs de file et la conception de médicaments plus efficaces *in vivo,* et vise à réduire les coûts de développement de nouveaux candidats-médicaments en diminuant les taux d'échecs en phase 2 des essais cliniques.

VII. BIBLIOGRAPHIE

Abbott, A. (2012). Cell rewind wins medicine Nobel. Nature *490*, 151–152.

Abecassis, P., and Coutinet, N. (2008). Caractéristiques du marché des médicaments et stratégies des firmes pharmaceutiques. Horiz. Strat. 111–139.

Adam, M., Oh, S.L., Sudarshan, V.K., Koh, J.E., Hagiwara, Y., Tan, J.H., Tan, R.S., and Acharya, U.R. (2018). Automated characterization of cardiovascular diseases using relative wavelet nonlinear features extracted from ECG signals. Comput. Methods Programs Biomed. *161*, 133–143.

Akoulitchev, S., Chuikov, S., and Reinberg, D. (2000). TFIIH is negatively regulated by cdk8-containing mediator complexes. Nature *407*, 102–106.

Alexander, L.T., Möbitz, H., Drueckes, P., Savitsky, P., Fedorov, O., Elkins, J.M., Deane, C.M., Cowan-Jacob, S.W., and Knapp, S. (2015). Type II Inhibitors Targeting CDK2. ACS Chem. Biol. *10*, 2116–2125.

Allen, B.L., and Taatjes, D.J. (2015). The Mediator complex: a central integrator of transcription. Nat. Rev. Mol. Cell Biol. *16*, 155–166.

Alsallaq, R., and Zhou, H.-X. (2008). Electrostatic Rate Enhancement and Transient Complex of Protein-Protein Association. Proteins *71*, 320–335.

Amaral, M., Kokh, D.B., Bomke, J., Wegener, A., Buchstaller, H.P., Eggenweiler, H.M., Matias, P., Sirrenberg, C., Wade, R.C., and Frech, M. (2017). Protein conformational flexibility modulates kinetics and thermodynamics of drug binding. Nat. Commun. *8*.

Andersson, K., and Hämäläinen, M.D. (2006). Replacing affinity with binding kinetics in QSAR studies resolves otherwise confounded effects. J. Chemom. *20*, 370–375.

Aqvist, J., Medina, C., and Samuelsson, J.E. (1994). A new method for predicting binding affinity in computer-aided drug design. Protein Eng. *7*, 385–391.

Arrowsmith, J. (2011a). Trial watch: Phase II failures: 2008–2010.

Arrowsmith, J. (2011b). Trial watch: phase III and submission failures: 2007-2010. Nat. Rev. Drug Discov. *10*, 87.

Arrowsmith, J., and Miller, P. (2013). Trial Watch: Phase II and Phase III attrition rates 2011–2012.

Ash, J., and Fourches, D. (2017). Characterizing the Chemical Space of ERK2 Kinase Inhibitors Using Descriptors Computed from Molecular Dynamics Trajectories. J. Chem. Inf. Model. *57*, 1286–1299.

Ayaz, P., Andres, D., Kwiatkowski, D.A., Kolbe, C.-C., Lienau, P., Siemeister, G., Lücking, U., and Stegmann, C.M. (2016). Conformational Adaption May Explain the Slow Dissociation Kinetics

of Roniciclib (BAY 1000394), a Type I CDK Inhibitor with Kinetic Selectivity for CDK2 and CDK9. ACS Chem. Biol. *11*, 1710–1719.

Bai, F., Xu, Y., Chen, J., Liu, Q., Gu, J., Wang, X., Ma, J., Li, H., Onuchic, J.N., and Jiang, H. (2013). Free energy landscape for the binding process of Huperzine A to acetylcholinesterase. Proc. Natl. Acad. Sci. U. S. A. *110*, 4273–4278.

Bakris, G., Gradman, A., Reif, M., Wofford, M., Munger, M., Harris, S., Vendetti, J., Michelson, E.L., and Wang, R. (2001). Antihypertensive Efficacy of Candesartan in Comparison to Losartan: The CLAIM Study. J. Clin. Hypertens. *3*, 16–21.

Barette, C., Jariel-Encontre, I., Piechaczyk, M., and Piette, J. (2001). Human cyclin C protein is stabilized by its associated kinase cdk8, independently of its catalytic activity. Oncogene *20*, 551–562.

Baumli, S., Lolli, G., Lowe, E.D., Troiani, S., Rusconi, L., Bullock, A.N., Debreczeni, J.É., Knapp, S., and Johnson, L.N. (2008). The structure of P-TEFb (CDK9/cyclin T1), its complex with flavopiridol and regulation by phosphorylation. EMBO J. *27*, 1907–1918.

Bergeron, P., Koehler, M.F.T., Blackwood, E.M., Bowman, K., Clark, K., Firestein, R., Kiefer, J.R., Maskos, K., McCleland, M.L., Orren, L., et al. (2016). Design and Development of a Series of Potent and Selective Type II Inhibitors of CDK8. ACS Med. Chem. Lett. *7*, 595–600.

Bernetti, M., Cavalli, A., and Mollica, L. (2017). Protein–ligand (un)binding kinetics as a new paradigm for drug discovery at the crossroad between experiments and modelling. MedChemComm *8*, 534–550.

Besse, P., Guillouet, B., Loubes, J.-M., and François, R. (2015). Review & Perspective for Distance Based Trajectory Clustering.

Böhm, H.J. (1994). The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. J. Comput. Aided Mol. Des. *8*, 243–256.

Böhm, H.-J., and Klebe, G. (1996). What Can We Learn from Molecular Recognition in Protein–Ligand Complexes for the Design of New Drugs? Angew. Chem. Int. Ed. Engl. *35*, 2588–2614.

Bongrand, P. (1999). LIGAND-RECEPTOR INTERACTIONS. Rep. Prog. Phys. *62*, 921–968.

Bortolato, A., Deflorian, F., Weiss, D.R., and Mason, J.S. (2015). Decoding the Role of Water Dynamics in Ligand–Protein Unbinding: CRF1R as a Test Case. J. Chem. Inf. Model. *55*, 1857–1866.

Bosma, R., Witt, G., Vaas, L.A.I., Josimovic, I., Gribbon, P., Vischer, H.F., Gul, S., and Leurs, R. (2017). The Target Residence Time of Antihistamines Determines Their Antagonism of the G Protein-Coupled Histamine H1 Receptor. Front. Pharmacol. *8*, 667.

Bowers, K., Chow, E., Xu, H., Dror, R., Eastwood, M., Gregersen, B., Klepeis, J., Kolossvary, I., Moraes, M., Sacerdoti, F., et al. (2006). Scalable Algorithms for Molecular Dynamics Simulations on Commodity Clusters. In ACM/IEEE SC 2006 Conference (SC'06), (Tampa, FL, USA: IEEE), pp. 43–43.

Brady, K., Wei, A.Z., Ringe, D., and Abeles, R.H. (1990). Structure of chymotrypsin-trifluoromethyl ketone inhibitor complexes: comparison of slowly and rapidly equilibrating inhibitors. Biochemistry *29*, 7600–7607.

Brooks, B.R., Brooks, C.L., Mackerell, A.D., Nilsson, L., Petrella, R.J., Roux, B., Won, Y., Archontis, G., Bartels, C., Boresch, S., et al. (2009). CHARMM: The biomolecular simulation program. J. Comput. Chem. *30*, 1545–1614.

Buch, I., Giorgino, T., and De Fabritiis, G. (2011). Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations. Proc. Natl. Acad. Sci. *108*, 10184–10189.

Bunnage, M.E. (2011). Getting pharmaceutical R&D back on target. Nat. Chem. Biol. *7*, 335–339.

Callegari, D., Lodola, A., Pala, D., Rivara, S., Mor, M., Rizzi, A., and Capelli, A.M. (2017). Metadynamics Simulations Distinguish Short- and Long-Residence-Time Inhibitors of Cyclin-Dependent Kinase 8. J. Chem. Inf. Model. *57*, 159–169.

Canavese, M., Santo, L., and Raje, N. (2012). Cyclin dependent kinases in cancer: Potential for therapeutic intervention. Cancer Biol. Ther. *13*, 451–457.

Canela, N., Orzáez, M., Fucho, R., Mateo, F., Gutierrez, R., Pineda-Lucena, A., Bachs, O., and Pérez-Payá, E. (2006). Identification of an hexapeptide that binds to a surface pocket in cyclin A and inhibits the catalytic activity of the complex cyclin-dependent kinase 2-cyclin A. J. Biol. Chem. *281*, 35942–35953.

Carles, F., Bourg, S., Meyer, C., Bonnet, P., Carles, F., Bourg, S., Meyer, C., and Bonnet, P. (2018). PKIDB: A Curated, Annotated and Updated Database of Protein Kinase Inhibitors in Clinical Trials. Molecules *23*, 908.

Casasnovas, R., Limongelli, V., Tiwary, P., Carloni, P., and Parrinello, M. (2017). Unbinding Kinetics of a p38 MAP Kinase Type II Inhibitor from Metadynamics Simulations. J. Am. Chem. Soc. *139*, 4780–4788.

Case, D., Babin, V., Berryman, J., Betz, R., Cai, Q., Cerutti, D., Cheatham, T., Darden, T., Duke, R., Gohlke, H., et al. (2015). {Amber 15} (University of California, San Francisco).

Chast, F. (2002). Histoire contemporaine des médicaments (Paris: La Découverte).

Cholko, T., Chen, W., Tang, Z., and Chang, C.A. (2018). A molecular dynamics investigation of CDK8/CycC and ligand binding: conformational flexibility and implication in drug discovery. J. Comput. Aided Mol. Des. *32*, 671–685.

Choulier, L., Andersson, K., Hämäläinen, M.D., Regenmortel, V., H.v, M., Malmqvist, M., and Altschuh, D. (2002). QSAR studies applied to the prediction of antigen–antibody interaction kinetics as measured by BIACORE. Protein Eng. Des. Sel. *15*, 373–382.

Christ, M., Braun, N., Neuffer, J., and Kempa-Liehr, A.W. (2018). Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests (tsfresh – A Python package). Neurocomputing *307*, 72–77.

Christiansen, E., Hudson, B.D., Hansen, A.H., Milligan, G., and Ulven, T. (2016). Development and Characterization of a Potent Free Fatty Acid Receptor 1 (FFA1) Fluorescent Tracer. J. Med. Chem. *59*, 4849–4858.

Coleman, K.G., Wautlet, B.S., Morrissey, D., Mulheron, J., Sedman, S.A., Brinkley, P., Price, S., and Webster, K.R. (1997). Identification of CDK4 sequences involved in cyclin D1 and p16 binding. J. Biol. Chem. *272*, 18869–18874.

Copeland, R.A. (2010). The dynamics of drug-target interactions: drug-target residence time and its impact on efficacy and safety. Expert Opin. Drug Discov. *5*, 305–310.

Copeland, R.A. (2011). Conformational adaptation in drug–target interactions and residence time. Future Med. Chem. *3*, 1491–1501.

Copeland, R.A. (2013). Evaluation of Enzyme Inhibitors in Drug Discovery: A Guide for Medicinal Chemists and Pharmacologists (Hoboken, NJ, USA: John Wiley & Sons, Inc.).

Copeland, R.A. (2016). The drug–target residence time model: a 10-year retrospective. Nat. Rev. Drug Discov. *15*, 87–95.

Copeland, R.A., Pompliano, D.L., and Meek, T.D. (2006). Drug–target residence time and its implications for lead optimization. Nat. Rev. Drug Discov. *5*, 730.

Costa, B., Da Pozzo, E., Giacomelli, C., Barresi, E., Taliani, S., Da Settimo, F., and Martini, C. (2016). TSPO ligand residence time: a new parameter to predict compound neurosteroidogenic efficacy. Sci. Rep. *6*, 18164.

Csajka, C., and Verotta, D. (2006). Pharmacokinetic-pharmacodynamic modelling: history and perspectives. J. Pharmacokinet. Pharmacodyn. *33*, 227–279.

Dahl, G., and Akerud, T. (2013). Pharmacokinetics and the drug–target residence time concept. Drug Discov. Today *18*, 697–707.

Decherchi, S., Bottegoni, G., Spitaleri, A., Rocchia, W., and Cavalli, A. (2018). BiKi Life Sciences: A New Suite for Molecular Dynamics and Related Methods in Drug Discovery. J. Chem. Inf. Model.

Deganutti, G., Zhukov, A., Deflorian, F., Federico, S., Spalluto, G., Cooke, R.M., Moro, S., Mason, J.S., and Bortolato, A. (2017). Impact of protein–ligand solvation and desolvation on transition state thermodynamic properties of adenosine A2A ligand binding kinetics. Silico Pharmacol. *5*.

Dickson, A., and Lotz, S.D. (2017). Multiple Ligand Unbinding Pathways and Ligand-Induced Destabilization Revealed by WExplore. Biophys. J. *112*, 620–629.

DiMasi, J.A., Hansen, R.W., and Grabowski, H.G. (2003). The price of innovation: new estimates of drug development costs. J. Health Econ. *22*, 151–185.

DiMasi, J.A., Grabowski, H.G., and Hansen, R.W. (2016). Innovation in the pharmaceutical industry: New estimates of R&D costs. J. Health Econ. *47*, 20–33.

Dimitriadis, S.I., Liparas, D., and Alzheimer's Disease Neuroimaging Initiative (2018). How random is the random forest? Random forest algorithm on the service of structural imaging biomarkers for Alzheimer's disease: from Alzheimer's disease neuroimaging initiative (ADNI) database. Neural Regen. Res. *13*, 962–970.

Doerr, S., and De Fabritiis, G. (2014). On-the-Fly Learning and Sampling of Ligand Binding by High-Throughput Molecular Simulations. J. Chem. Theory Comput. *10*, 2064–2069.

Doornbos, M.L.J., Cid, J.M., Haubrich, J., Nunes, A., van de Sande, J.W., Vermond, S.C., Mulder-Krieger, T., Trabanco, A.A., Ahnaou, A., Drinkenburg, W.H., et al. (2017). Discovery and Kinetic Profiling of 7-Aryl-1,2,4-triazolo[4,3- *a*]pyridines: Positive Allosteric Modulators of the Metabotropic Glutamate Receptor 2. J. Med. Chem. *60*, 6704–6720.

Drews, J. (2000). Drug discovery: a historical perspective. Science *287*, 1960–1964.

Dror, R.O., Pan, A.C., Arlow, D.H., Borhani, D.W., Maragakis, P., Shan, Y., Xu, H., and Shaw, D.E. (2011). Pathway and mechanism of drug binding to G-protein-coupled receptors. Proc. Natl. Acad. Sci. U. S. A. *108*, 13118–13123.

Durham, M.C. (2004). Tiotropium (Spiriva): a once-daily inhaled anticholinergic medication for chronic obstructive pulmonary disease. Proc. Bayl. Univ. Med. Cent. *17*, 366–373.

Echalier, A., Endicott, J.A., and Noble, M.E.M. (2010). Recent developments in cyclin-dependent kinase biochemical and structural studies. Biochim. Biophys. Acta *1804*, 511–519.

Fabbro, D., Cowan-Jacob, S.W., and Moebitz, H. (2015). Ten things you should know about protein kinases: IUPHAR Review 14: Ten things you should know about protein kinases. Br. J. Pharmacol. *172*, 2675–2700.

Fedosova, N.U., Champeil, P., and Esmann, M. (2002). Nucleotide Binding to Na,K-ATPase: The Role of Electrostatic Interactions. Biochemistry *41*, 1267–1273.

Ferguson, F.M., and Gray, N.S. (2018). Kinase inhibitors: the road ahead. Nat. Rev. Drug Discov. *17*, 353–377.

Firestein, R., Bass, A.J., Kim, S.Y., Dunn, I.F., Silver, S.J., Guney, I., Freed, E., Ligon, A.H., Vena, N., Ogino, S., et al. (2008). CDK8 is a colorectal cancer oncogene that regulates beta-catenin activity. Nature *455*, 547–551.

Fleck, B.A., Hoare, S.R.J., Pick, R.R., Bradbury, M.J., and Grigoriadis, D.E. (2012). Binding kinetics redefine the antagonist pharmacology of the corticotropin-releasing factor type 1 receptor. J. Pharmacol. Exp. Ther. *341*, 518–531.

Fletcher, R., and Reeves, C.M. (1964). Function minimization by conjugate gradients. Comput. J. *7*, 149–154.

Flocco, M.M., and Mowbray, S.L. (1994). Planar stacking interactions of arginine and aromatic side-chains in proteins. J. Mol. Biol. *235*, 709–717.

Fogolari, F., Brigo, A., and Molinari, H. (2002). The Poisson–Boltzmann equation for biomolecular electrostatics: a tool for structural biology. J. Mol. Recognit. *15*, 377–392.

Folmer, R.H.A. (2017). Drug target residence time: a misleading concept. Drug Discov. Today.

Fox, J.M., Kang, K., Sastry, M., Sherman, W., Sankaran, B., Zwart, P.H., and Whitesides, G.M. (2017). Water‐Restructuring Mutations Can Reverse the Thermodynamic Signature of Ligand Binding to Human Carbonic Anhydrase. Angew Chem Int Ed 6.

Frederick, K.K., Marlow, M.S., Valentine, K.G., and Wand, A.J. (2007). Conformational entropy in molecular recognition by proteins. Nature *448*, 325–329.

Freire, E. (2008). Do Enthalpy and Entropy Distinguish First in Class From Best in Class? Drug Discov. Today *13*, 869–874.

Freire, E. (2009). A thermodynamic approach to the affinity optimization of drug candidates. Chem. Biol. Drug Des. *74*, 468–472.

Freire, E. (2015). The Binding Thermodynamics of Drug Candidates. In Methods and Principles in Medicinal Chemistry, G.M. Keserü, and D.C. Swinney, eds. (Weinheim, Germany: Wiley-VCH Verlag GmbH & Co. KGaA), pp. 1–13.

Freissmuth, M., Hasenhuetl, P., Sucic, S., Sitte, H., and Sandtner, W. (2015). Association Rate Constants as Determinants of Ligand Selectivity:Lessons from The Dopamine And The Serotonin Transporter. FASEB J. *29*, 932.1.

Fuchs, B., Breithaupt-GröGler, K., Belz, G.G., Roll, S., Malerczyk, C., Herrmann, V., Spahn-Langguth, H., and Mutschler, E. (2000). Comparative Pharmacodynamics and Pharmacokinetics of Candesartan and Losartan in Man. J. Pharm. Pharmacol. *52*, 1075–1083.

Fulcher, B.D. (2018). Feature-Based Time-Series Analysis.

Fulcher, B.D., and Jones, N.S. (2017). hctsa: A Computational Framework for Automated Time-Series Phenotyping Using Massive Feature Extraction. Cell Syst. *5*, 527-531.e3.

Furumoto, T., Tanaka, A., Ito, M., Malik, S., Hirose, Y., Hanaoka, F., and Ohkuma, Y. (2007). A kinase subunit of the human mediator complex, CDK8, positively regulates transcriptional activation. Genes Cells Devoted Mol. Cell. Mech. *12*, 119–132.

Gao, J., Bosco, D.A., Powers, E.T., and Kelly, J.W. (2009). Localized Thermodynamic Coupling between Hydrogen Bonding and Microenvironment Polarity Substantially Stabilizes Proteins. Nat. Struct. Mol. Biol. *16*, 684–690.

Gaspari, R., Rechlin, C., Heine, A., Bottegoni, G., Rocchia, W., Schwarz, D., Bomke, J., Gerber, H.-D., Klebe, G., and Cavalli, A. (2016). Kinetic and Structural Insights into the Mechanism of Binding of Sulfonamides to Human Carbonic Anhydrase by Computational and Experimental Studies. J. Med. Chem. *59*, 4245–4256.

Gaulton, A., Bellis, L.J., Bento, A.P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B., et al. (2012). ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res. *40*, D1100-1107.

Georgi, V., Andres, D., Fernandez-Montalvan, A.E., Stegmann, C.M., Becker, A., and Mueller-Fahrnow, A. (2017). Binding kinetics in drug discovery - A current perspective. Front. Biosci. Landmark Ed. *22*, 21–47.

Glazer, D.S., Radmer, R.J., and Altman, R.B. (2008). Combining molecular dynamics and machine learning to improve protein function recognition. Pac. Symp. Biocomput. Pac. Symp. Biocomput. 332–343.

Gondeau, C., Gerbal-Chaloin, S., Bello, P., Aldrian-Herrada, G., Morris, M.C., and Divita, G. (2005). Design of a Novel Class of Peptide Inhibitors of Cyclin-dependent Kinase/Cyclin Activation. J. Biol. Chem. *280*, 13793–13800.

Gottwald, M., Becker, A., Bahr, I., and Mueller-Fahrnow, A. (2016). Public-Private Partnerships in Lead Discovery: Overview and Case Studies. Arch. Pharm. (Weinheim) *349*, 692–697.

Gouda, H., Kuntz, I.D., Case, D.A., and Kollman, P.A. (2003). Free energy calculations for theophylline binding to an RNA aptamer: Comparison of MM-PBSA and thermodynamic integration methods. Biopolymers *68*, 16–34.

Grant, B.J., Rodrigues, A.P.C., ElSawy, K.M., McCammon, J.A., and Caves, L.S.D. (2006). Bio3d: an R package for the comparative analysis of protein structures. Bioinforma. Oxf. Engl. *22*, 2695–2696.

Guo, D., and IJzerman, A.P. (2018). Molecular Basis of Ligand Dissociation from G Protein-Coupled Receptors and Predicting Residence Time. Methods Mol. Biol. Clifton NJ *1705*, 197–206.

Guo, D., Mulder-Krieger, T., IJzerman, A.P., and Heitman, L.H. (2012). Functional efficacy of adenosine A2A receptor agonists is positively correlated to their receptor residence time. Br. J. Pharmacol. *166*, 1846–1859.

Hall, H.K. (1956). Potentiometric Determination of the Base Strength of Amines in Non-protolytic Solvents. J. Phys. Chem. *60*, 63–70.

Hämäläinen, M.D. (2014). The importance of drug-target binding kinetics for drug efficacy – Without on you are off! (International Workshop : new approaches in drug design & discovery, Marburg, Germain).

Harper, T.M., and Taatjes, D.J. (2017). The complex structure and function of Mediator. J. Biol. Chem. jbc.R117.794438.

Hasenhuetl, P.S., Schicker, K., Koenig, X., Li, Y., Sarker, S., Stockner, T., Sucic, S., Sitte, H.H., Freissmuth, M., and Sandtner, W. (2015). Ligand Selectivity among the Dopamine and Serotonin Transporters Specified by the Forward Binding Reaction. Mol. Pharmacol. *88*, 12–18.

Hauray, B. (2006). Chapitre 7. Politique et expertise scientifique. In L'Europe du médicament, (Paris: Presses de Sciences Po (P.F.N.S.P.)), pp. 270–295.

Hay, M., Thomas, D.W., Craighead, J.L., Economides, C., and Rosenthal, J. (2014). Clinical development success rates for investigational drugs. Nat. Biotechnol. *32*, 40.

Heitz, F., Morris, M.C., Fesquet, D., Cavadore, J.-C., Dorée, M., and Divita, G. (1997). Interactions of Cyclins with Cyclin-Dependent Kinases: A Common Interactive Mechanism. Biochemistry *36*, 4995–5003.

Hoeppner, S., Baumli, S., and Cramer, P. (2005). Structure of the Mediator Subunit Cyclin C and its Implications for CDK8 Function. J. Mol. Biol. *350*, 833–842.

Hothersall, J.D., Brown, A.J., Dale, I., and Rawlins, P. (2016). Can residence time offer a useful strategy to target agonist drugs for sustained GPCR responses? Drug Discov. Today *21*, 90–96.

Huang, D., and Caflisch, A. (2011). The free energy landscape of small molecule unbinding. PLoS Comput. Biol. *7*, e1002002.

Humphrey, W., Dalke, A., and Schulten, K. (1996). VMD – Visual Molecular Dynamics. J. Mol. Graph. *14*, 33–38.

Jagger, B.R., Lee, C.T., and Amaro, R.E. (2018). Quantitative Ranking of Ligand Binding Kinetics with a Multiscale Milestoning Simulation Approach. J. Phys. Chem. Lett. 4941–4948.

Jain, A.N. (1996). Scoring noncovalent protein-ligand interactions: a continuous differentiable function tuned to compute binding affinities. J. Comput. Aided Mol. Des. *10*, 427–440.

Jakalian, A., Jack, D.B., and Bayly, C.I. (2002). Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. J. Comput. Chem. *23*, 1623–1641.

Jeffrey, P.D., Russo, A.A., Polyak, K., Gibbs, E., Hurwitz, J., Massagué, J., and Pavletich, N.P. (1995). Mechanism of CDK activation revealed by the structure of a cyclinA-CDK2 complex. Nature *376*, 313–320.

Johnson, L.N., Noble, M.E., and Owen, D.J. (1996). Active and inactive protein kinases: structural basis for regulation. Cell *85*, 149–158.

Jorgensen, W.L., Chandrasekhar, J., Madura, J.D., Impey, R.W., and Klein, M.L. (1983). Comparison of simple potential functions for simulating liquid water. J. Chem. Phys. *79*, 926–935.

Jorgensen, W.L., Maxwell, D.S., and Tirado-Rives, J. (1996). Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. J. Am. Chem. Soc. *118*, 11225–11236.

Kapur, S., and Seeman, P. (2001). Does fast dissociation from the dopamine d(2) receptor explain the action of atypical antipsychotics?: A new hypothesis. Am. J. Psychiatry *158*, 360–369.

Kästner, J. (2011). Umbrella sampling: Umbrella sampling. Wiley Interdiscip. Rev. Comput. Mol. Sci. *1*, 932–942.

Keserü, G.M. (2015). Drug-Target Residence Time. Thermodyn. Kinet. Drug Bind. 12.

Keserü, G.M., and Swinney, D.C. (2015). Thermodynamics and Binding Kinetics in Drug Discovery. In Methods and Principles in Medicinal Chemistry, G.M. Keserü, and D.C. Swinney, eds. (Weinheim, Germany: Wiley-VCH Verlag GmbH & Co. KGaA), pp. 313–329.

Khanna, I. (2012). Drug discovery in pharmaceutical industry: productivity challenges and trends. Drug Discov. Today *17*, 1088–1102.

Klebe, G. (2015). The Use of Thermodynamic and Kinetic Data in Drug Discovery: Decisive Insight or Increasing the Puzzlement? ChemMedChem *10*, 229–231.

Klein Herenbrink, C., Sykes, D.A., Donthamsetti, P., Canals, M., Coudrat, T., Shonberg, J., Scammells, P.J., Capuano, B., Sexton, P.M., Charlton, S.J., et al. (2016). The role of kinetic context in apparent biased agonism at GPCRs. Nat. Commun. *7*, 10842.

Kokh, D.B., Amaral, M., Bomke, J., Grädler, U., Musil, D., Buchstaller, H.-P., Dreyer, M.K., Frech, M., Lowinski, M., Vallee, F., et al. (2018). Estimation of Drug-Target Residence Times by τ-Random Acceleration Molecular Dynamics Simulations. J. Chem. Theory Comput. *14*, 3859–3869.

Kola, I., and Landis, J. (2004). Can the pharmaceutical industry reduce attrition rates? Nat. Rev. Drug Discov. *3*, 711–715.

Kollman, Peter. (1993). Free energy calculations: Applications to chemical and biochemical phenomena. Chem. Rev. *93*, 2395–2417.

Kollman, P.A., Massova, I., Reyes, C., Kuhn, B., Huo, S., Chong, L., Lee, M., Lee, T., Duan, Y., Wang, W., et al. (2000). Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. Acc. Chem. Res. *33*, 889–897.

Kornev, A.P., and Taylor, S.S. (2010). Defining the conserved internal architecture of a protein kinase. Biochim. Biophys. Acta BBA - Proteins Proteomics *1804*, 440–444.

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. J. Stat. Softw. *28*, 1–26.

Kumar, S., Rosenberg, J.M., Bouzida, D., Swendsen, R.H., and Kollman, P.A. (1992). The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. J. Comput. Chem. *13*, 1011–1021.

Kundrotas, P.J., and Alexov, E. (2006). Electrostatic Properties of Protein-Protein Complexes. Biophys. J. *91*, 1724–1736.

Lacourcière, Y. (1999). A comparison of the efficacy and duration of action of candesartan cilexetil and losartan as assessed by clinic and ambulatory blood pressure after a missed dose, in truly hypertensive patients A placebo-controlled, forced titration study. Am. J. Hypertens. *12*, 1181–1187.

Lafont, V., Armstrong, A.A., Ohtaka, H., Kiso, Y., Mario Amzel, L., and Freire, E. (2007). Compensating enthalpic and entropic changes hinder binding affinity optimization. Chem. Biol. Drug Des. *69*, 413–422.

Lahiry, P., Torkamani, A., Schork, N.J., and Hegele, R.A. (2010). Kinase mutations in human disease: interpreting genotype-phenotype relationships. Nat. Rev. Genet. *11*, 60–74.

LaMattina, J.L. (2011). The impact of mergers on pharmaceutical R&D. Nat. Rev. Drug Discov. *10*, 559–560.

Lamba, V., and Ghosh, I. (2012). New directions in targeting protein kinases: focusing upon true allosteric and bivalent inhibitors. Curr. Pharm. Des. *18*, 2936–2945.

Landry, Y. (2018). Petite histoire des médicaments (Paris).

Laskowski, R.A., MacArthur, M.W., Moss, D.S., and Thornton, J.M. (1993). PROCHECK: a program to check the stereochemical quality of protein structures. J. Appl. Crystallogr. *26*, 283–291.

Lee, K.S.S., Liu, J.-Y., Wagner, K.M., Pakhomova, S., Dong, H., Morisseau, C., Fu, S.H., Yang, J., Wang, P., Ulu, A., et al. (2014). Optimized inhibitors of soluble epoxide hydrolase improve in vitro target residence time and in vivo efficacy. J. Med. Chem. *57*, 7016–7030.

Leeson, P.D., and Springthorpe, B. (2007). The influence of drug-like concepts on decision-making in medicinal chemistry. Nat. Rev. Drug Discov. *6*, 881–890.

Li, L., Yang, Y., Zhang, D., Ye, Z.-G., Jesse, S., Kalinin, S.V., and Vasudevan, R.K. (2018). Machine learning–enabled identification of material phase transitions based on experimental data: Exploring collective dynamics in ferroelectric relaxors. Sci. Adv. *4*, eaap8672.

Liu, L., Michelsen, K., Kitova, E.N., Schnier, P.D., and Klassen, J.S. (2010). Evidence that Water Can Reduce the Kinetic Stability of Protein−Hydrophobic Ligand Interactions. J. Am. Chem. Soc. *132*, 17658–17660.

Liu, Y., Kung, C., Fishburn, J., Ansari, A.Z., Shokat, K.M., and Hahn, S. (2004). Two cyclin-dependent kinases promote RNA polymerase II transcription and formation of the scaffold complex. Mol. Cell. Biol. *24*, 1721–1735.

Lo Conte, L., Chothia, C., and Janin, J. (1999). The atomic structure of protein-protein recognition sites. J. Mol. Biol. *285*, 2177–2198.

Lolli, G. (2010). Structural dissection of cyclin dependent kinases regulation and protein recognition properties. Cell Cycle *9*, 1551–1561.

Lotz, S.D., and Dickson, A. (2018). Unbiased Molecular Dynamics of 11 min Timescale Drug Unbinding Reveals Transition State Stabilizing Interactions. J. Am. Chem. Soc. *140*, 618–628.

Lu, H., Iuliano, J.N., and Tonge, P.J. (2018). Structure–kinetic relationships that control the residence time of drug–target complexes: insights from molecular structure and dynamics. Curr. Opin. Chem. Biol. *44*, 101–109.

Luckner, S.R., Liu, N., am Ende, C.W., Tonge, P.J., and Kisker, C. (2010). A slow, tight binding inhibitor of InhA, the enoyl-acyl carrier protein reductase from Mycobacterium tuberculosis. J. Biol. Chem. *285*, 14330–14337.

Maier, J.A., Martinez, C., Kasavajhala, K., Wickstrom, L., Hauser, K.E., and Simmerling, C. (2015). ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. J. Chem. Theory Comput. *11*, 3696–3713.

Malumbres, M. (2014). Cyclin-dependent kinases. Genome Biol. *15*, 122.

Massova, I., and Kollman, P.A. (1999). Computational Alanine Scanning To Probe Protein−Protein Interactions: A Novel Approach To Evaluate Binding Free Energies. J. Am. Chem. Soc. *121*, 8133–8143.

McDermott, M.S.J., Chumanevich, A.A., Lim, C.-U., Liang, J., Chen, M., Altilia, S., Oliver, D., Rae, J.M., Shtutman, M., Kiaris, H., et al. (2017). Inhibition of CDK8 mediator kinase suppresses estrogen dependent transcription and the growth of estrogen receptor positive breast cancer. Oncotarget *8*, 12558–12575.

Meunier, B. (2016). L'innovation thérapeutique : évolution et tendances : Leçon inaugurale prononcée le jeudi 6 novembre 2014 (Paris: Collège de France).

Miller, D.C., Lunn, G., Jones, P., Sabnis, Y., Davies, N.L., and Driscoll, P. (2012). Investigation of the effect of molecular properties on the binding kinetics of a ligand to its biological target. MedChemComm *3*, 449.

Mollica, L., Decherchi, S., Zia, S.R., Gaspari, R., Cavalli, A., and Rocchia, W. (2015). Kinetics of protein-ligand unbinding via smoothed potential molecular dynamics simulations. Sci. Rep. *5*.

Mollica, L., Theret, I., Antoine, M., Perron-Sierra, F., Charton, Y., Fourquez, J.-M., Wierzbicki, M., Boutin, J.A., Ferry, G., Decherchi, S., et al. (2016). Molecular Dynamics Simulations and Kinetic Measurements to Estimate and Predict Protein–Ligand Residence Times. J. Med. Chem. *59*, 7167–7176.

Mondal, J., Friesner, R.A., and Berne, B.J. (2014). Role of Desolvation in Thermodynamics and Kinetics of Ligand Binding to a Kinase. J. Chem. Theory Comput. *10*, 5696–5705.

Mourey, R.J., Burnette, B.L., Brustkern, S.J., Daniels, J.S., Hirsch, J.L., Hood, W.F., Meyers, M.J., Mnich, S.J., Pierce, B.S., Saabye, M.J., et al. (2010). A benzothiophene inhibitor of mitogen-activated protein kinase-activated protein kinase 2 inhibits tumor necrosis factor alpha production and has oral anti-inflammatory efficacy in acute and chronic models of inflammation. J. Pharmacol. Exp. Ther. *333*, 797–807.

Mullard, A. (2018). 2017 FDA drug approvals. Nat. Rev. Drug Discov. *17*, 81–85.

Munos, B. (2009). Lessons from 60 years of pharmaceutical innovation. Nat. Rev. Drug Discov. *8*, 959–968.

Nederpelt, I., Georgi, V., Schiele, F., Nowak-Reppel, K., Fernández-Montalván, A.E., IJzerman, A.P., and Heitman, L.H. (2016). Characterization of 12 GnRH peptide agonists - a kinetic perspective. Br. J. Pharmacol. *173*, 128–141.

Niu, Y., Li, S., Pan, D., Liu, H., and Yao, X. (2016). Computational study on the unbinding pathways of B-RAF inhibitors and its implication for the difference of residence time: insight

from random acceleration and steered molecular dynamics simulations. Phys. Chem. Chem. Phys. *18*, 5622–5629.

Olsson, M.H.M., Søndergaard, C.R., Rostkowski, M., and Jensen, J.H. (2011). PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa Predictions. J. Chem. Theory Comput. *7*, 525–537.

Onufriev, A., Bashford, D., and Case, D.A. (2004). Exploring protein native states and large-scale conformational changes with a modified generalized born model. Proteins *55*, 383–394.

Oostenbrink, C., Villa, A., Mark, A.E., and van Gunsteren, W.F. (2004). A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force-field parameter sets 53A5 and 53A6. J. Comput. Chem. *25*, 1656–1676.

Oppermann, F.S., Gnad, F., Olsen, J.V., Hornberger, R., Greff, Z., Kéri, G., Mann, M., and Daub, H. (2009). Large-scale Proteomics Analysis of the Human Kinome. Mol. Cell. Proteomics MCP *8*, 1751–1764.

Pan, A.C., Borhani, D.W., Dror, R.O., and Shaw, D.E. (2013). Molecular determinants of drug–receptor binding kinetics. Drug Discov. Today *18*, 667–673.

Pan, A.C., Xu, H., Palpant, T., and Shaw, D.E. (2017). Quantitative Characterization of the Binding and Unbinding of Millimolar Drug Fragments with Molecular Dynamics Simulations. J. Chem. Theory Comput. *13*, 3372–3377.

Pargellis, C., Tong, L., Churchill, L., Cirillo, P.F., Gilmore, T., Graham, A.G., Grob, P.M., Hickey, E.R., Moss, N., Pav, S., et al. (2002). Inhibition of p38 MAP kinase by utilizing a novel allosteric binding site. Nat. Struct. Biol. *9*, 268–272.

Patel, J.S., Berteotti, A., Ronsisvalle, S., Rocchia, W., and Cavalli, A. (2014). Steered Molecular Dynamics Simulations for Studying Protein–Ligand Interaction in Cyclin-Dependent Kinase 5. J. Chem. Inf. Model. *54*, 470–480.

Pavletich, N.P. (1999). Mechanisms of cyclin-dependent kinase regulation: structures of Cdks, their cyclin activators, and Cip and INK4 inhibitors. J. Mol. Biol. *287*, 821–828.

Payne, D.J., Gwynn, M.N., Holmes, D.J., and Pompliano, D.L. (2007). Drugs for bad bugs: confronting the challenges of antibacterial discovery. Nat. Rev. Drug Discov. *6*, 29–40.

Pearlstein, R.A., Sherman, W., and Abel, R. (2013). Contributions of water transfer energy to protein-ligand association and dissociation barriers: Watermap analysis of a series of p38 α MAP kinase inhibitors: Water Transfer in Structure-Kinetic Relationships. Proteins Struct. Funct. Bioinforma. *81*, 1509–1526.

Peletier, L.A., Benson, N., and van der Graaf, P.H. (2010). Impact of protein binding on receptor occupancy: a two-compartment model. J. Theor. Biol. *265*, 657–671.

Philip, S., Kumarasiri, M., Teo, T., Yu, M., and Wang, S. (2018). Cyclin-Dependent Kinase 8: A New Hope in Targeted Cancer Therapy? J. Med. Chem. *61*, 5073–5092.

Pietrucci, F., Marinelli, F., Carloni, P., and Laio, A. (2009). Substrate Binding Mechanism of HIV-1 Protease from Explicit-Solvent Atomistic Simulations. J. Am. Chem. Soc. *131*, 11811–11818.

Pina, A.S., Hussain, A., and Roque, A.C.A. (2010). An Historical Overview of Drug Discovery. In Ligand-Macromolecular Interactions in Drug Discovery, A.C.A. Roque, ed. (Totowa, NJ: Humana Press), pp. 3–12.

Plattner, N., and Noé, F. (2015). Protein conformational plasticity and complex ligand-binding kinetics explored by atomistic simulations and Markov models. Nat. Commun. *6*.

Plotkin, S. (2014). History of vaccination. Proc. Natl. Acad. Sci. U. S. A. *111*, 12283–12287.

Portnoy, A., Kumar, S., Behm, D.J., Mahar, K.M., Noble, R.B., Throup, J.P., and Russ, S.F. (2013). Effects of Urotensin II Receptor Antagonist, GSK1440115, in Asthma. Front. Pharmacol. *4*.

Prüll, C., Maehle, A., and Halliwell, R. (2009). A Short History of the Drug Receptor Concept (Springer).

Puttini, M., Redaelli, S., Moretti, L., Brussolo, S., Gunby, R.H., Mologni, L., Marchesi, E., Cleris, L., Donella-Deana, A., Drueckes, P., et al. (2008). Characterization of compound 584, an Abl kinase inhibitor with lasting effects. Haematologica *93*, 653–661.

Qu, S., Huang, S., Pan, X., Yang, L., and Mei, H. (2016). Constructing Interconsistent, Reasonable, and Predictive Models for Both the Kinetic and Thermodynamic Properties of HIV-1 Protease Inhibitors. J. Chem. Inf. Model. *56*, 2061–2068.

R Core Team (2017). R: A Language and Environment for Statistical Computing (Vienna, Austria: R Foundation for Statistical Computing).

Radić, Z., Kirchhoff, P.D., Quinn, D.M., McCammon, J.A., and Taylor, P. (1997). Electrostatic Influence on the Kinetics of Ligand Binding to Acetylcholinesterase: DISTINCTIONS BETWEEN ACTIVE CENTER LIGANDS AND FASCICULIN. J. Biol. Chem. *272*, 23265–23277.

Ramos, I., Aparici, M., Letosa, M., Puig, C., Gavaldà, A., Huerta, J.M., Espinosa, S., Vilella, D., and Miralpeix, M. (2018). Abediterol (LAS100977), an inhaled long-acting β 2-adrenoceptor agonist, has a fast association rate and long residence time at receptor. Eur. J. Pharmacol. *819*, 89–97.

Richard, P. (2012). Dossier Big Data : l'analyse des données intéresse de plus en plus les entreprises.

Roskoski, R. (2016). Classification of small molecule protein kinase inhibitors based upon the structures of their drug-enzyme complexes. Pharmacol. Res. *103*, 26–48.

Ryckaert, J., Ciccotti, G., and Berendsen, H.J.C. (1977). Numerical integration of the Cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. J Comput Phys 327–341.

Ryckmans, T., Edwards, M.P., Horne, V.A., Correia, A.M., Owen, D.R., Thompson, L.R., Tran, I., Tutt, M.F., and Young, T. (2009). Rapid assessment of a novel series of selective CB2 agonists

using parallel synthesis protocols: A Lipophilic Efficiency (LipE) analysis. Bioorg. Med. Chem. Lett. *19*, 4406–4409.

Rzymski, T., Mikula, M., Wiklik, K., and Brzózka, K. (2015). CDK8 kinase—An emerging target in targeted cancer therapy. Biochim. Biophys. Acta BBA - Proteins Proteomics *1854*, 1617–1629.

Saeed, U., Jalal, N., and Ashraf, M. (2012). Roles of Cyclin Dependent Kinase and Cdk-Activating Kinase in Cell Cycle Regulation: Contemplation of Intracellular Interactions and Functional Characterization. 7.

Sali, A., and Blundell, T.L. (1993). Comparative protein modelling by satisfaction of spatial restraints. J. Mol. Biol. *234*, 779–815.

Salvalaglio, M., Tiwary, P., and Parrinello, M. (2014). Assessing the Reliability of the Dynamics Reconstructed from Metadynamics. J. Chem. Theory Comput. *10*, 1420–1425.

Samiee, K., Kovács, P., and Gabbouj, M. (2015). Epileptic seizure classification of EEG time-series using rational discrete short-time fourier transform. IEEE Trans. Biomed. Eng. *62*, 541–552.

Sánchez-Martínez, C., Gelbert, L.M., Lallena, M.J., and de Dios, A. (2015). Cyclin dependent kinase (CDK) inhibitors as anticancer drugs. Bioorg. Med. Chem. Lett. *25*, 3420–3435.

Santos, R., Ursu, O., Gaulton, A., Bento, A.P., Donadi, R.S., Bologa, C.G., Karlsson, A., Al-Lazikani, B., Hersey, A., Oprea, T.I., et al. (2017). A comprehensive map of molecular drug targets. Nat. Rev. Drug Discov. *16*, 19–34.

Sarver, R.W., Peevers, J., Cody, W.L., Ciske, F.L., Dyer, J., Emerson, S.D., Hagadorn, J.C., Holsworth, D.D., Jalaie, M., Kaufman, M., et al. (2007). Binding thermodynamics of substituted diaminopyrimidine renin inhibitors. Anal. Biochem. *360*, 30–40.

Scannell, J.W., Blanckley, A., Boldon, H., and Warrington, B. (2012). Diagnosing the decline in pharmaceutical R&D efficiency. Nat. Rev. Drug Discov. *11*, 191–200.

Schiele, F., Ayaz, P., and Müller-Fahrnow, A. (2015a). The Use of Structural Information to Understand Binding Kinetics. In Methods and Principles in Medicinal Chemistry, G.M. Keserü, and D.C. Swinney, eds. (Weinheim, Germany: Wiley-VCH Verlag GmbH & Co. KGaA), pp. 237–256.

Schiele, F., Ayaz, P., and Fernández-Montalván, A. (2015b). A universal homogeneous assay for high-throughput determination of binding kinetics. Anal. Biochem. *468*, 42–49.

Schlitter, J., Engels, M., and Krüger, P. (1994). Targeted molecular dynamics: a new approach for searching pathways of conformational transitions. J. Mol. Graph. *12*, 84–89.

Schmidtke, P., Luque, F.J., Murray, J.B., and Barril, X. (2011). Shielded Hydrogen Bonds as Structural Determinants of Binding Kinetics: Application in Drug Design. J. Am. Chem. Soc. *133*, 18903–18910.

Schneider, E.V., Böttcher, J., Blaesse, M., Neumann, L., Huber, R., and Maskos, K. (2011). The Structure of CDK8/CycC Implicates Specificity in the CDK/Cyclin Family and Reveals Interaction with a Deep Pocket Binder. J. Mol. Biol. *412*, 251–266.

Schneider, E.V., Böttcher, J., Huber, R., Maskos, K., and Neumann, L. (2013). Structure–kinetic relationship study of CDK8/CycC specific compounds. Proc. Natl. Acad. Sci. *110*, 8081–8086.

Schuetz, D.A. (2018). K4DD drug target binding kinetics data (EMBL-EBI).

Schuetz, D.A., de Witte, W.E.A., Wong, Y.C., Knasmueller, B., Richter, L., Kokh, D.B., Sadiq, S.K., Bosma, R., Nederpelt, I., and Segala, E. (2017). Kinetics for Drug Discovery: an industry-driven effort to target drug residence time. Drug Discov. Today.

Schuetz, D.A., Richter, L., Amaral, M., Grandits, M., Grädler, U., Musil, D., Buchstaller, H.-P., Eggenweiler, H.-M., Frech, M., and Ecker, G.F. (2018). Ligand Desolvation Steers On-Rate and Impacts Drug Residence Time of Heat Shock Protein 90 (Hsp90) Inhibitors. J. Med. Chem. *61*, 4397–4411.

Setny, P., Baron, R., Kekenes-Huskey, P.M., McCammon, J.A., and Dzubiella, J. (2013). Solvent fluctuations in hydrophobic cavity–ligand binding kinetics. Proc. Natl. Acad. Sci. *110*, 1197–1202.

Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., et al. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. *7*, 539.

Soethoudt, M., Hoorens, M.W.H., Doelman, W., Martella, A., van der Stelt, M., and Heitman, L.H. (2018). Structure-kinetic relationship studies of cannabinoid CB2 receptor agonists reveal substituent-specific lipophilic effects on residence time. Biochem. Pharmacol. *152*, 129–142.

Spaar, A., Dammer, C., Gabdoulline, R.R., Wade, R.C., and Helms, V. (2006). Diffusional Encounter of Barnase and Barstar. Biophys. J. *90*, 1913–1924.

Spagnuolo, L.A., Eltschkner, S., Yu, W., Daryaee, F., Davoodi, S., Knudson, S.E., Allen, E.K.H., Merino, J., Pschibul, A., Moree, B., et al. (2017). Evaluating the Contribution of Transition-State Destabilization to Changes in the Residence Time of Triazole-Based InhA Inhibitors. J. Am. Chem. Soc. *139*, 3417–3429.

Sriram, K., and Insel, P.A. (2018). G Protein-Coupled Receptors as Targets for Approved Drugs: How Many Targets and How Many Drugs? Mol. Pharmacol. *93*, 251–258.

Stoddart, L.A., White, C.W., Nguyen, K., Hill, S.J., and Pfleger, K.D.G. (2016). Fluorescence- and bioluminescence-based approaches to study GPCR ligand binding. Br. J. Pharmacol. *173*, 3028–3037.

Strasser, A., Wittmann, H.-J., and Seifert, R. (2017). Binding Kinetics and Pathways of Ligands to GPCRs. Trends Pharmacol. Sci. *38*, 717–732.

Sun, H., Tian, S., Zhou, S., Li, Y., Li, D., Xu, L., Shen, M., Pan, P., and Hou, T. (2015). Revealing the favorable dissociation pathway of type II kinase inhibitors via enhanced sampling simulations and two-end-state calculations. Sci. Rep. *5*.

Sun, H., Li, Y., Shen, M., Li, D., Kang, Y., and Hou, T. (2017). Characterizing Drug–Target Residence Time with Metadynamics: How To Achieve Dissociation Rate Efficiently without Losing Accuracy against Time-Consuming Approaches. J. Chem. Inf. Model. *57*, 1895–1906.

Sung, J.C., Wynsberghe, A.W.V., Amaro, R.E., Li, W.W., and McCammon, J.A. (2010). Role of Secondary Sialic Acid Binding Sites in Influenza N1 Neuraminidase. J. Am. Chem. Soc. *132*, 2883–2885.

Swinney, D.C. (2004). Biochemical mechanisms of drug action: what does it take for success? Nat. Rev. Drug Discov. *3*, 801–808.

Swinney, D.C. (2006). Biochemical mechanisms of New Molecular Entities (NMEs) approved by United States FDA during 2001-2004: mechanisms leading to optimal efficacy and safety. Curr. Top. Med. Chem. *6*, 461–478.

Swinney, D.C. (2016). The role of Binding Kinetics in Drug Action (Mastering Medicinal Chemistry Congress, Boston, U.S.).

Swinney, D.C., and Anthony, J. (2011). How were new medicines discovered? Nat. Rev. Drug Discov. *10*, 507–519.

Sykes, D.A., Parry, C., Reilly, J., Wright, P., Fairhurst, R.A., and Charlton, S.J. (2014). Observed Drug-Receptor Association Rates Are Governed by Membrane Affinity: The Importance of Establishing "Micro-Pharmacokinetic/Pharmacodynamic Relationships" at the 2-Adrenoceptor. Mol. Pharmacol. *85*, 608–617.

Taneja, A., Vermeulen, A., Huntjens, D.R.H., Danhof, M., De Lange, E.C.M., and Proost, J.H. (2017). Modeling of prolactin response following dopamine D_2 receptor antagonists in rats: can it be translated to clinical dosing? Pharmacol. Res. Perspect. *5*, e00364.

Taylor, S.S., and Kornev, A.P. (2011). Protein Kinases: Evolution of Dynamic Regulatory Proteins. Trends Biochem. Sci. *36*, 65–77.

Taylor, S.S., Shaw, A.S., Kannan, N., and Kornev, A.P. (2015). Integration of signaling in the kinome: Architecture and regulation of the α C Helix. Biochim. Biophys. Acta BBA - Proteins Proteomics *1854*, 1567–1574.

Tee, A.K.H., Koh, M.S., Gibson, P.G., Lasserson, T.J., Wilson, A.J., and Irving, L.B. (2007). Long-acting beta2-agonists versus theophylline for maintenance treatment of asthma. Cochrane Database Syst. Rev. CD001281.

Teo, I., Mayne, C.G., Schulten, K., and Lelièvre, T. (2016). Adaptive Multilevel Splitting Method for Molecular Dynamics Calculation of Benzamidine-Trypsin Dissociation Time. J. Chem. Theory Comput. *12*, 2983–2989.

Tesson, S. (2016). Un champ de forces polarisable pour l'étude des argiles à l'échelle moléculaire.

Tiwary, P., and Parrinello, M. (2013). From Metadynamics to Dynamics. Phys. Rev. Lett. *111*, 230602.

Tiwary, P., Limongelli, V., Salvalaglio, M., and Parrinello, M. (2015). Kinetics of protein–ligand unbinding: Predicting pathways, rates, and rate-limiting steps. Proc. Natl. Acad. Sci. *112*, E386–E391.

Tiwary, P., Mondal, J., and Berne, B.J. (2017). How and when does an anticancer drug leave its binding site? Sci. Adv. *3*.

Tonge, P.J. (2017). Drug–Target Kinetics in Drug Discovery. ACS Chem. Neurosci. *9*, 29–39.

Topliss, J.G., and Costello, R.J. (1972). Change correlations in structure-activity studies using multiple regression analysis. J. Med. Chem. *15*, 1066–1068.

Tresadern, G., Bartolome, J.M., Macdonald, G.J., and Langlois, X. (2011). Molecular properties affecting fast dissociation from the D2 receptor. Bioorg. Med. Chem. *19*, 2231–2241.

Tsai, C.J., Lin, S.L., Wolfson, H.J., and Nussinov, R. (1997). Studies of protein-protein interfaces: a statistical analysis of the hydrophobic effect. Protein Sci. Publ. Protein Soc. *6*, 53–64.

Tummino, P.J., and Copeland, R.A. (2008). Residence time of receptor-ligand complexes and its effect on biological function. Biochemistry *47*, 5481–5492.

Tuntland, T., Ethell, B., Kosaka, T., Blasco, F., Zang, R.X., Jain, M., Gould, T., and Hoffmaster, K. (2014). Implementation of pharmacokinetic and pharmacodynamic strategies in early research phases of drug discovery and development at Novartis Institute of Biomedical Research. Front. Pharmacol. *5*.

Uitdehaag, J.C.M., de Man, J., Willemsen-Seegers, N., Prinsen, M.B.W., Libouban, M.A.A., Sterrenburg, J.G., de Wit, J.J.P., de Vetter, J.R.F., de Roos, J.A.D.M., Buijsman, R.C., et al. (2017). Target Residence Time-Guided Optimization on TTK Kinase Results in Inhibitors with Potent Anti-Proliferative Activity. J. Mol. Biol. *429*, 2211–2230.

Vauquelin, G. (2010). Rebinding: or why drugs may act longer in vivo than expected from their in vitro target residence time. Expert Opin. Drug Discov. *5*, 927–941.

Vauquelin, G. (2015). On the 'micro'-pharmacodynamic and pharmacokinetic mechanisms that contribute to long-lasting drug action. Expert Opin. Drug Discov. *10*, 1085–1098.

Vauquelin, G. (2016). Effects of target binding kinetics on in vivo drug efficacy: koff, kon and rebinding. Br. J. Pharmacol. *173*, 2319–2334.

Vauquelin, G., and Charlton, S.J. (2010). Long-lasting target binding and rebinding as mechanisms to prolong in vivo drug action. Br. J. Pharmacol. *161*, 488–508.

Vauquelin, G., Bostoen, S., Vanderheyden, P., and Seeman, P. (2012). Clozapine, atypical antipsychotics, and the benefits of fast-off D2 dopamine receptor antagonism. Naunyn. Schmiedebergs Arch. Pharmacol. *385*, 337–372.

Votapka, L.W., Jagger, B.R., Heyneman, A.L., and Amaro, R.E. (2017). SEEKR: Simulation Enabled Estimation of Kinetic Rates, A Computational Tool to Estimate Molecular Kinetics and Its Application to Trypsin–Benzamidine Binding. J. Phys. Chem. B *121*, 3597–3606.

Walter, N.M., Wentsch, H.K., Bührmann, M., Bauer, S.M., Döring, E., Mayer-Wrangowski, S., Sievers-Engler, A., Willemsen-Seegers, N., Zaman, G., Buijsman, R., et al. (2017). Design, Synthesis, and Biological Evaluation of Novel Type I $^{1}/_{2}$ p38 α MAP Kinase Inhibitors with Excellent Selectivity, High Potency, and Prolonged Target Residence Time by Interfering with the R-Spine. J. Med. Chem. *60*, 8027–8054.

Wang, Z., and Cole, P.A. (2014). Catalytic Mechanisms and Regulation of Protein Kinases. Methods Enzymol. *548*, 1–21.

Wang, J., Wolf, R.M., Caldwell, J.W., Kollman, P.A., and Case, D.A. (2004). Development and testing of a general amber force field. J. Comput. Chem. *25*, 1157–1174.

Wang, J., Wang, W., Kollman, P.A., and Case, D.A. (2006). Automatic atom type and bond type perception in molecular mechanical calculations. J. Mol. Graph. Model. *25*, 247–260.

Wang, X., Wang, J., Ding, Z., Ji, J., Sun, Q., and Cai, G. (2013). Structural flexibility and functional interaction of mediator Cdk8 module. Protein Cell *4*, 911–920.

Waring, M.J., Leach, A.G., and Miller, D.C. (2015). Challenges in the Medicinal Chemical Optimization of Binding Kinetics. In Methods and Principles in Medicinal Chemistry, G.M. Keserü, and D.C. Swinney, eds. (Weinheim, Germany: Wiley-VCH Verlag GmbH & Co. KGaA), pp. 191–210.

Watson, C., Jenkinson, S., Kazmierski, W., and Kenakin, T. (2005). The CCR5 receptor-based mechanism of action of 873140, a potent allosteric noncompetitive HIV entry inhibitor. Mol. Pharmacol. *67*, 1268–1282.

Wentsch, H.K., Walter, N.M., Bührmann, M., Mayer-Wrangowski, S., Rauh, D., Zaman, G.J.R., Willemsen-Seegers, N., Buijsman, R.C., Henning, M., Dauch, D., et al. (2017). Optimized Target Residence Time: Type I1/2 Inhibitors for p38 α MAP Kinase with Improved Binding Kinetics through Direct Interaction with the R-Spine. Angew. Chem. Int. Ed. *56*, 5363–5367.

Wertheimer, A., and Santella, T. (2007). The history and economics of pharmaceutical patents. In The Value of Innovation: Impact on Health, Life Quality, Safety, and Regulatory Research, (Emerald Group Publishing Limited), pp. 101–119.

Westley, A.M., and Westley, J. (1996). Enzyme Inhibition in Open Systems SUPERIORITY OF UNCOMPETITIVE AGENTS. J. Biol. Chem. *271*, 5347–5352.

Wiederstein, M., and Sippl, M.J. (2007). ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. Nucleic Acids Res. *35*, W407–W410.

de Witte, W.E., Wong, Y.C., Nederpelt, I., Heitman, L.H., Danhof, M., van der Graaf, P.H., Gilissen, R.A., and de Lange, E.C.M. (2016a). Mechanistic models enable the rational use of *in vitro* drug-target binding kinetics for better drug effects in patients. Expert Opin. Drug Discov. *11*, 45–63.

de Witte, W.E.A., Danhof, M., van der Graaf, P.H., and de Lange, E.C.M. (2016b). In vivo Target Residence Time and Kinetic Selectivity: The Association Rate Constant as Determinant. Trends Pharmacol. Sci. *37*, 831–842.

de Witte, W.E.A., Vauquelin, G., van der Graaf, P.H., and de Lange, E.C.M. (2017). The influence of drug distribution and drug-target binding on target occupancy: The rate-limiting step approximation. Eur. J. Pharm. Sci. Off. J. Eur. Fed. Pharm. Sci. *109S*, S83–S89.

Wood, D.J., and Endicott, J.A. (2018). Structural insights into the functional diversity of the CDK–cyclin family. Open Biol. *8*, 180112.

Wood, E.R., Truesdale, A.T., McDonald, O.B., Yuan, D., Hassell, A., Dickerson, S.H., Ellis, B., Pennisi, C., Horne, E., Lackey, K., et al. (2004). A unique structure for epidermal growth factor receptor bound to GW572016 (Lapatinib): relationships among protein conformation, inhibitor off-rate, and receptor activity in tumor cells. Cancer Res. *64*, 6652–6659.

Wood, M.E., Smith, D.J., Reid, D.W., Masel, P.J., France, M.W., and Bell, S.C. (2013). Ivacaftor in severe cystic fibrosis lung disease and a G551D mutation. Respirol. Case Rep. *1*, 52–54.

Xu, W., Amire-Brahimi, B., Xie, X.-J., Huang, L., and Ji, J.-Y. (2014). All-atomic molecular dynamic studies of human CDK8: Insight into the A-loop, point mutations and binding with its partner CycC. Comput. Biol. Chem. *51*, 1–11.

Yoshikawa, M., Saitoh, M., Katoh, T., Seki, T., Bigi, S.V., Shimizu, Y., Ishii, T., Okai, T., Kuno, M., Hattori, H., et al. (2018). Discovery of 7-Oxo-2,4,5,7-tetrahydro-6 H-pyrazolo[3,4- c]pyridine Derivatives as Potent, Orally Available, and Brain-Penetrating Receptor Interacting Protein 1 (RIP1) Kinase Inhibitors: Analysis of Structure-Kinetic Relationships. J. Med. Chem. *61*, 2384–2409.

Yu, X.-Q., and Wilson, A.G. (2010). The role of pharmacokinetic and pharmacokinetic/pharmacodynamic modeling in drug discovery and development. Future Med. Chem. *2*, 923–928.

Yung-Chi, C., and Prusoff, W.H. (1973). Relationship between the inhibition constant (KI) and the concentration of inhibitor which causes 50 per cent inhibition (I50) of an enzymatic reaction. Biochem. Pharmacol. *22*, 3099–3108.

Zeller, F., Luitz, M.P., Bomblies, R., and Zacharias, M. (2017). Multiscale Simulation of Receptor–Drug Association Kinetics: Application to Neuraminidase Inhibitors. J. Chem. Theory Comput. *13*, 5097–5105.

Zheng, X., Bi, C., Li, Z., Podariu, M., and Hage, D.S. (2015). Analytical methods for kinetic studies of biological interactions: A review. J. Pharm. Biomed. Anal. *113*, 163–180.

Zhou, G., Pantelopulos, G.A., Mukherjee, S., and Voelz, V.A. (2017). Bridging Microscopic and Macroscopic Mechanisms of p53-MDM2 Binding with Kinetic Network Models. Biophys. J. *113*, 785–793.

Zhu, F., and Hummer, G. (2012). Convergence and error estimation in free energy calculations using the weighted histogram analysis method. J. Comput. Chem. *33*, 453–465.

Zwier, M.C., Pratt, A.J., Adelman, J.L., Kaus, J.W., Zuckerman, D.M., and Chong, L.T. (2016). Efficient Atomistic Simulation of Pathways and Calculation of Rate Constants for a Protein–Peptide Binding Process: Application to the MDM2 Protein and an Intrinsically Disordered p53 Peptide. J. Phys. Chem. Lett. *7*, 3440–3445.

Sonia ZIADA

Prédiction des constantes cinétiques de complexes protéine-ligand

Résumé :

Le coût de plus en plus élevé de l'ensemble du processus de recherche préclinique et de développement clinique d'un médicament pousse la communauté scientifique à limiter au maximum les causes d'échecs. De nombreuses études démontrent qu'une évaluation préclinique des constantes cinétiques de liaison permet de limiter les taux d'échec en phase II des essais cliniques. L'objectif de ce projet de thèse, financé par l'Institut de Recherche Servier (IdRS) à Croissy-sur-seine, est de développer un outil informatique visant à prédire ces constantes cinétiques en un temps de calcul acceptable, afin de permettre son utilisation en routine en phase précoce de sélection et d'optimisation des molécules actives.

Une première partie de cette thèse a été dévolue à l'étude du jeu de données utilisé pour réaliser ces développements techniques. Il est constitué d'inhibiteurs de la kinase dépendante de la cycline 8 (CDK8), une cible thérapeutique émergente de la famille des protéines kinases impliquée dans le cancer colorectal.

La deuxième partie de ce travail a porté sur le développement d'un outil visant à classer les composés en fonction de leurs temps de résidence. Cet outil a été ensuite validé sur un jeu de données interne de l'IdRS.

Enfin, dans une troisième partie, un protocole a été initié pour prédire les constantes cinétiques de manière quantitative et non qualitative et pouvoir également identifier les déterminants structuraux responsables des propriétés cinétiques des composés.

Mots clés : dynamique moléculaire, constantes cinétiques, temps de résidence, processus de dissociation

Kinetics constants prediction of protein-ligand complexes

Abstract:

Over the past decade, there is an increasing interest for pharmaceutical companies in measuring and understanding drugs' binding kinetics, in order to limit the most possible failure causes. Indeed, numerous studies demonstrate that preclinical evaluation of binding kinetics constants allow a limitation of failure rates in phase II clinical trials. The main objective of this thesis, founded by the Institut de Recherche Servier (IdRS) in Croissy-sur-seine, is to develop an informatics tool with the aim to predict those kinetics constants at low computational cost, to allow its use in an industrial context.

The first part of this work focuses on the study of the dataset used to develop the methodology. This dataset is constituted by Cyclin 8 Dependent Kinase (CDK8) inhibitors, a new therapeutic target of protein kinase family, involved in colorectal cancer.

The second part of this thesis aims to develop a tool able to classify compounds according to their experimental residence time. This *in silico* tool has been validated on another private dataset, which belongs to the IdRS.

Finally, in a last part, a protocol is currently under development to determine in a quantitative way the kinetics constant. It will also be able to identify the structural insights responsible for the kinetics properties of inhibitors.

Keywords: molecular dynamics, kinetics constant, residence time, dissociation process.

Institut de Chimie Organique et Analytique
UMR CNRS-Université d'Orléans 7311
Rue de Chartres, 45067 Orléans cedex 2