



HAL
open science

Accurate prediction of ligand-binding affinities in solution by molecular dynamics simulations

Gilberto Pereira

► **To cite this version:**

Gilberto Pereira. Accurate prediction of ligand-binding affinities in solution by molecular dynamics simulations. Other. Université de Strasbourg, 2021. English. NNT : 2021STRAF044 . tel-03702370

HAL Id: tel-03702370

<https://theses.hal.science/tel-03702370v1>

Submitted on 23 Jun 2022

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

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

ÉCOLE DOCTORALE DES SCIENCES CHIMIQUES (222)

Laboratoire d'Ingénierie des Fonctions Moléculaires, Institut de Chimie, UMR7177

THÈSE

présentée par:

Gilberto PEREIRA

soutenue le: **5 Novembre 2021**

pour obtenir le grade de : **Docteur de l'université de Strasbourg**

Discipline/ Spécialité : Chimie / Chimie théorique et informatique

**Accurate prediction of ligand-binding affinities
in solution by Molecular Dynamics simulations**

THÈSE dirigée par:

M. CECCHINI Marco

Dr, Maitre de Conférence, CNRS - Université de Strasbourg

RAPPORTEURS:

M. CAFLISCH Amedeo

Pr, Professor, Technical University of Zürich

M. SUÁREZ Dimas

Pr, Professor, University of Oviedo

AUTRES MEMBRES DU JURY:

M. STOTE Roland

Dr, Directeur de Recherche, CNRS - Université de Strasbourg

**ACCURATE PREDICTION OF LIGAND-BINDING
AFFINITIES IN SOLUTION BY MOLECULAR
DYNAMICS SIMULATIONS**

GILBERTO PAULO PEREIRA

Institut de Chimie de Strasbourg
École Doctorale des Sciences Chimiques – ED 222
Université de Strasbourg

Thesis Director – Dr. Marco Cecchini

Dedication

As the curtain closes on this stage, I realize that there are not enough words thank you for everything you have done for me. So, please accept this gesture

To my parents,
Sebastião and Alexandra

Acknowledgements

It has been more than three years since I left my home in Lisbon, Portugal, to venture into the unknown. Being raised in a common Portuguese home, with particular hopes and dreams, I never expected to be writing these words at the end of a long journey which made me mature both as a scientist and as a person. Amongst the trials were the adaptation to a new culture, to a new field of study and to living far away from everything I've ever known. Coming to France was challenging, but it gave me many things to which I am grateful. As the page turns and I turn towards new goals, I take this opportunity to be grateful to the people who have shared this path with me.

My first thank you is dedicated to my supervisor, Professor Marco Cecchini. None of this would be possible if, during May of 2018, I had not have the pleasure to discuss with you and arrange a travel to the IFM lab in Strasbourg. I felt incredibly welcomed from the first moment. More than managing my scientific work and path, Marco offered me insight into how a scientist behaves, carries himself and pursues truth and understanding in a strong but humble manner. I am grateful to you for all you have done to help and guide me, which is more than I could ever ask of a supervisor.

It would be a grave mistake if I also did not take this opportunity to thank the professors who pointed me towards Strasbourg and the Cecchini group: Pr. Pedro Lima, Pr. Florbela Pereira, Pr. João Aires de Sousa and Pr. Gilles Marcou. In particular, I would like to thank Pr. Gilles and Pr. João for establishing the Lisbon-Strasbourg link and encouraging me to pursue it by contacting Marco. Professor Gilles was also part of my mid-thesis committee, alongside Pr. Roland Stote, and their advice was helpful going towards my last year of PhD.

I would like to thank Pr. Roland Stote, Pr. Dimas Suárez and Pr. Amedeo Caflisch for accepting my invitation and finding time in their very busy schedule to review the present manuscript.

To our collaborators at the Houdusse group at Institut Curie, thank you for your efforts into the project of Smooth Muscle Myosin II. Our interactions were always positive and enlightening, which contributed to a successful outcome of the project.

I am also grateful to the friends that welcomed me into the lab and that have, in time, left to conquer their own goals: Dr. Joel Montalvo-Acosta, Dr. Diego Gomes and Dr. Florian Blanc. They were always a source of support and inspiration, and continue to be. To Diego, for being a great friend and for your advices. To Florian, for your patience and kindness when discussing various subjects. To Joel, for being the big brother I never had.

As I travelled this road, I would like to also thank the people who have joined the lab over time, my friends Marion Sisqueiras, Katia Galentino, Mariia Avstrikova, Dr. Alessio Bartocci, Dr. Adrien Cerdan, Federica Brando and Alisson Popp. Thank you for always being available to discuss and help in solving the (many) issues we faced in our work, for the moments of joy and fun we had and for being great scientists and friends. You have given me so much and I wish to be able to repay you one day.

To the members of UMR-7177 (Institut de Chimie), I would like to thank you for the moments spent chatting around a cup of coffee even if we work on very different subjects. A kind word or advice can go a long way, and these conversations have certainly helped.

To the friends and family I have made outside of the laboratory, I am also thankful. To Daniel Moreno, Maria de Lourdes, Alex Martin, David Zorilla and Carmen Lázaro-Sánchez, thank you for being there. To these friends, for being a safety net to whom I could always fall back on. To Carmen, for being my home away from home and for your unwavering love and support. Your kindness and gentleness offered me comfort in my times of doubt and need. In more ways than one, you are someone I admire, respect and love deeply.

Finally, it is time to write the last part of my acknowledgements. As it is standard of practice, I reserve this last part for my family. Leaving home to go after my dreams was a very difficult decision, because I knew how much leaving home would hurt. Your strength and elegance in accepting that decision, pushing me to strive for higher and higher accomplishments, has been a source of motivation for me. I am truly proud of being your son and brother to my sister, to whom I would give the world if I could. I am sorry that I do not express these feelings as often as I would like. For your love, support, guidance, and so much more you've given and keep giving me, I wish to thank you and dedicate this dissertation to you. It is as much mine as it is yours.

Accurate prediction of ligand- binding affinities in solution by Molecular Dynamics simulations

Abstract

Computational methodologies are able to accelerate drug discovery campaigns and decrease the associated costs. From the methods available to compute binding affinities, the Molecular Mechanics/Generalized Born Surface Area (MM/GBSA) approach is a compromise between accuracy and efficiency. However, here entropic contributions are often neglected. In this thesis, we developed an automatic method named, Quasi-Harmonic Multi-Basin, to compute the ligand configurational entropy loss upon binding. Apart from achieving quantitative agreement with experimental gas-phase entropies of small molecules, adding this correction to MM/GBSA estimates increased correlation to experiments by 10%. The correction was included into a Virtual Screening (VS) campaign in Smooth Muscle Myosin II (SMM2). From the VS, 26 compounds were experimentally tested and eight showed activity at 190 μM . We thus have established a VS protocol to identify inhibitors of complex allosteric proteins like myosin molecular motors.

Les méthodologies computationnelles sont capables d'accélérer les campagnes de découverte de médicaments et de réduire les coûts associés. Parmi les méthodes disponibles pour calculer les affinités de liaison, l'approche Mécanique Moléculaire/Surface de Born Généralisée (MM/GBSA) est un compromis entre précision et efficacité. Or, les contributions entropiques sont ici souvent négligées. Dans cette thèse, nous avons développé une méthode automatique appelée, Quasi-Harmonic Multi-Basin, pour calculer la perte d'entropie configurationnelle du ligand lors de la liaison. En plus d'obtenir un accord quantitatif avec les entropies expérimentales en phase gazeuse des petites molécules, l'ajout de cette correction aux estimations MM/GBSA a augmenté la corrélation avec les expériences de 10%. La correction a été incluse dans une campagne de criblage virtuel (VS) de la myosine musculaire lisse II (SMM2). À partir du VS, 26 composés ont été testés expérimentalement et huit ont montré une activité à 190 μM . Nous avons donc établi un protocole VS pour identifier des inhibiteurs de protéines allostériques complexes comme les moteurs moléculaires de la myosine.

Introduction

Dans le passé, la découverte de nouveaux médicaments était principalement due à la sérendipité.⁹ Pour maximiser la récupération des résultats et le succès des composés, les approches modernes tendent à s'appuyer de plus en plus sur des stratégies de conception rationnelle de médicaments.⁶ Le temps moyen d'une campagne de découverte de médicaments pour produire un nouveau médicament est compris entre 12 et 15 ans, avec un coût moyen d'environ 2 milliards de dollars.⁵⁹ Les méthodologies computationnelles sont devenues la norme dans les pipelines modernes de découverte de médicaments, fournissant des prédictions de l'affinité de liaison protéine-ligand et accélérant les étapes initiales des campagnes de découverte de médicaments.^{14,28,83,194}

Un autre point attrayant des méthodologies computationnelles est leur potentiel de réduction des coûts associés, à la fois en termes financiers et de temps humain.¹⁴ Ainsi, des efforts ont été consacrés au développement de méthodes capables de calculer de manière fiable les affinités de liaison protéine-ligand.^{6,26,194} Ces méthodes couvrent un spectre, équilibrant l'efficacité et la précision des calculs.⁴ Les méthodes de point final se concentrent sur l'échantillonnage conformationnel des états finaux de la réaction de liaison, représentant un compromis entre la précision et l'efficacité des calculs. Parmi les méthodes populaires de point final, citons la méthode de mécanique moléculaire et de surface de Poisson Boltzmann (MM/PBSA) et ses variantes^{120,193,196,206,265}, où l'énergie potentielle du système est calculée dans le vide par un champ de force de mécanique moléculaire, les effets de solvatation sont traités à l'aide de modèles de solvant implicites et les contributions entropiques sont estimées par l'approximation de l'oscillateur harmonique à rotor rigide (RRHO)^{120,265}.

Les approches par points finaux sont populaires dans la communauté scientifique en raison de leur capacité à produire des classements significatifs tout en restant efficaces.¹²⁰ L'entropie est une propriété thermodynamique essentielle qui régit la plupart des processus biomoléculaires, y compris la liaison des ligands.²⁹² Néanmoins, l'évaluation précise des entropies absolues en solution reste un grand défi.²⁸⁰ Dans le MM/PBSA et sa variante Generalized-Born (MM/GBSA), les termes entropiques sont généralement négligés en raison du coût de calcul et de l'incertitude de la valeur calculée.²⁶⁵ Cependant, négliger les contributions entropiques pourrait biaiser le calcul de l'énergie libre, conduisant à la prédiction de ligands plus grands comme meilleurs liants puisque le coût associé à la retenue du ligand dans le site de liaison est négligé. Les

méthodologies RRHO les plus populaires pour le calcul de l'entropie, l'analyse des modes normaux (NMA)²⁴⁵ et l'analyse quasi-harmonique (QHA)²⁴⁴, souffrent de certaines limitations. Dans la NMA, un potentiel harmonique pur est appliqué et l'anharmonicité au sein de la surface d'énergie potentielle (PES) est négligée, ce qui sous-estime l'entropie absolue du ligand.²⁸⁰ Dans la QHA, les constantes de force sont calculées à partir des fluctuations atomiques au cours d'une simulation MD et la PES est approximée comme un puits d'énergie potentielle harmonique unique et multidimensionnel, ce qui entraîne une surestimation importante de l'entropie absolue.²⁸⁰ Les méthodes de point final sont couramment utilisées dans les pipelines de découverte de médicaments pour aider à identifier les succès biologiques vers une cible pharmacologique donnée.

L'objectif de ce projet était de réaliser des campagnes VS et d'identifier des composés innovants hit vers la myosine II du muscle lisse (SMM2) à partir de bibliothèques chimiques virtuelles. Nous avons d'abord étudié la méthodologie MM/GBSA^{193,262,265} et développé une approche efficace et précise pour calculer la perte d'entropie du ligand lors de la liaison, basée sur la décomposition du PES du ligand en micro-états individuels, suivie de calculs QHA.^{213,244,279,280,292} Après validation de notre méthodologie, nous avons réalisé des campagnes VS sur la structure cristalline SMM2 (PBD **5M05**) en utilisant une approche de docking et de rescoring de l'énergie libre, mise en œuvre par notre logiciel *ChemFlow*, développé en interne, et augmentée d'une pénalité entropique originale calculée en utilisant une méthode développée en interne.³⁷

De nombreuses fonctions cellulaires dépendent de la polymérisation de l'actine et de son interaction avec les molécules de myosine.³³⁵ Les myosines sont une famille de moteurs moléculaires capables d'hydrolyser l'ATP, exploitant l'énergie provenant de sa l'hydrolyse pour effectuer un travail mécanique.³³⁶ Par exemple, si le filament d'actine est une route, la myosine serait la voiture et l'ATP le carburant. Ces protéines fonctionnent de manière cyclique et sont essentielles à un grand nombre de processus cellulaires, allant de la contraction musculaire à la division cellulaire.^{347,368} Le long du cycle moteur, de nombreux états intermédiaires instables ou transitoirement stables peuvent être peuplés lorsque la myosine subit des transitions de conformation entre des états stables (rigor, post-stroke, pré-powerstroke, power-stroke et état de maintien de la force). Certains de ces intermédiaires ont un potentiel de ciblage pharmacologique en vue du développement de nouvelles approches thérapeutiques. En effet, plusieurs études ont rapporté la modulation de l'activité de la myosine par des ligands de petites molécules.^{338,339,343,346,348,349} Le moteur moléculaire de la SMM2 a été choisi comme cible protéique d'intérêt parce que la contractilité des muscles lisses est un élément central de la physiopathologie de plusieurs maladies, comme l'asthme et la bronchopneumopathie chronique obstructive (BPCO).³⁴³ Les relaxants des muscles lisses actuellement disponibles ne sont pas spécifiquement conçus pour se lier à SMM2. Dans une publication

récente, le premier inhibiteur spécifique de SMM2 (CK571)³⁴³, faiblement nanomolaire, a été co-cristallisé en complexe avec SMM2 dans un état intermédiaire de la course de récupération. Cette structure cristalline à haute résolution offre une opportunité sans précédent pour la conception de nouveaux inhibiteurs puissants de SMM2.



Figure 1 - Structure cristallographique du domaine moteur SMM2 avec l'ADP (orange), le CK571 (cyan) et un ion magnésium (vert) liés (code PDB **5M05**). Les parties du domaine moteur SMM2 bordant le site de liaison sont mises en évidence : hélice SH1 (bleu), hélice Relay (rouge foncé), domaine N-terminal (vert) et une partie du domaine convertisseur (violet clair).

Résultats et discussion

Une approche Quasi-harmonique multi-bassin pour le calcul de l'entropie configurationnelle de petites molécules en solution

Nous avons développé une procédure automatique pour calculer avec précision et efficacité les entropies absolues des ligands en solution, basée sur l'approche QHA. Sous l'approximation RRHO et le cadre du "mélange de conformères"^{244,292}, le paysage de l'énergie libre d'un ligand est d'abord décomposé en micro-états individuels par regroupement de simulations MD tout-atome, générant un cluster par micro-état moléculaire. L'entropie des micro-états individuels est évaluée par QHA et l'entropie du paysage est calculée sur la base de l'équation de Shannon-Gibbs, ce qui conduit à une méthode appelée Quasi-Harmonic Multi-Basin (QHMB).²¹³ Pour évaluer les performances de QHMB, des simulations MD dans le vide ont été exécutées pour un ensemble de 22 petites molécules avec des entropies en phase gazeuse disponibles expérimentalement

à partir de la Computational Chemistry Comparison Benchmark DataBase du National Institute of Standards and Technology (NIST).^{291,322} La précision de QHMB a été comparée aux méthodes d'entropie RRHO standard en évaluant l'erreur quadratique moyenne (RMSE) de chaque méthode par rapport aux données expérimentales. Les régressions linéaires pour les résultats de la NMA, de la QHA et de la QHMA sont présentées dans la **Figure 2A**. Les données montrent que les méthodes RRHO standard fournissent des entropies qui sont fortement corrélées avec les données expérimentales. Cependant, l'analyse de la pente pour la NMA et la QHA illustre les lacunes de chaque méthode. La pente de la ligne de régression de la NMA est inférieure à un, indiquant une sous-estimation systématique de l'entropie qui s'aggrave avec la flexibilité intrinsèque des ligands. En revanche, dans le cas du QHA, la pente est plus du double. Pris ensemble, les résultats suggèrent que ni la NMA ni la QHA ne sont suffisamment précises pour reproduire les entropies expérimentales en phase gazeuse et qu'elles représentent plutôt des limites inférieures et supérieures à la véritable entropie, respectivement. En revanche, les résultats des calculs QHMB ont atteint un accord quantitatif avec les données expérimentales (RMSE = 0.36 kcal/mol; coefficient de corrélation de Pearson au carré (R^2) = 0.99), avec une pente de la ligne de régression de 1.02. Pour permettre l'application de routine de QHMB, une mise en œuvre automatique employant l'algorithme de regroupement hiérarchique à liaison moyenne a été construite. L'erreur quadratique moyenne des calculs automatiques de QHMB est passée de 0.36 à 0.65 kcal/mol, ce qui reste bien en dessous de la limite de précision chimique.

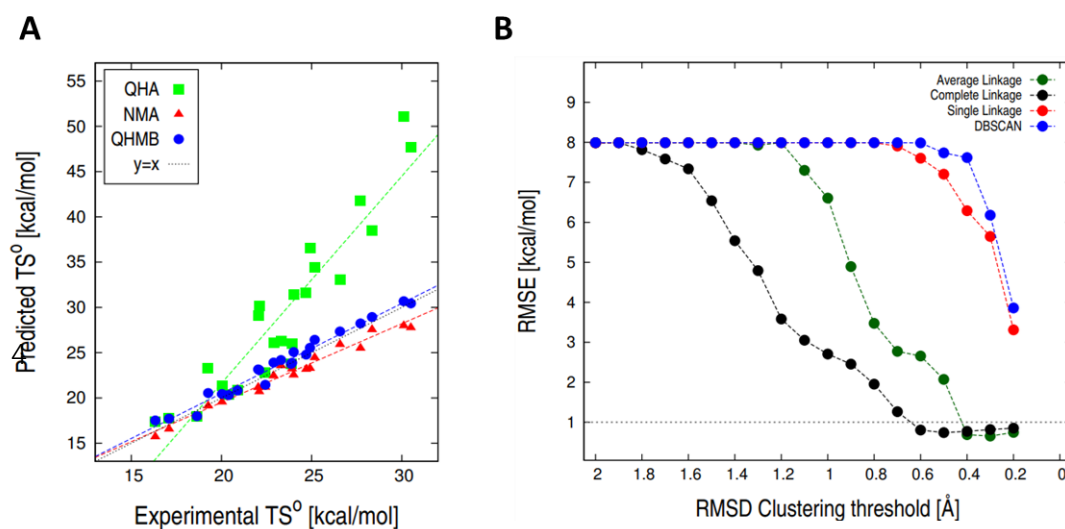


Figure 2 - Résultats obtenus pour les calculs d'entropie RRHO pour les petites molécules en phase gazeuse. Les entropies expérimentales à 298K proviennent du NIST. **A**) Les données montrent les performances de l'approche multi-puits nouvellement introduite (QHMB) par rapport aux approches populaires à puits unique basées sur l'approximation RRHO, c'est-à-dire NMA (rouge) et QHA (vert). **B**) Précision de QHMB en fonction de l'algorithme de regroupement et du seuil de RMSD. La précision est évaluée par l'erreur quadratique moyenne (RMSE) à partir des entropies expérimentales en phase gazeuse pour l'ensemble de

données du NIST ; voir le texte principal. La ligne pointillée à 1 kcal/mol illustre la limite de la précision chimique.

Tableau 1 - Précision de diverses méthodes RRHO pour le calcul des entropies moléculaires absolues.

Experiment	RMSE ^[*]	RMSE ^[+]	R ²	Pente	MUE ^a
NMA ^b	1.12	1.67	0.96	0.86	0.93
NMA-multi ^b	1.17	1.75	0.96	0.86	0.97
NMA-clust ^b	1.13	1.64	0.96	0.87	0.96
QHA ^b	8.09	13.36	0.86	2.24	5.87
QHA-clust ^b	8.42	13.95	0.86	2.30	6.07
QHMB ^c	0.36	0.24	0.99	1.02	0.28

Toutes les valeurs RMSE sont données en kcal/mol. [*] - RMSE pour l'ensemble complet de référence. [+] - RMSE pour le sous-ensemble de ligands avec 3 torsions non redondantes ou plus. [a] - Erreur moyenne non signée (MUE) entre les entropies absolues prédites et expérimentales. [b] - Résultats obtenus en appliquant le numéro de symétrie à un. [c] - Résultats obtenus en utilisant des nombres de symétrie appropriés.

Un calcul QHMB comprend trois étapes: i. la détermination des conformères stables avec leur probabilité d'équilibre à partir d'un MD convergé; ii. le calcul de l'entropie par bassin et par conformère par QHA; et iii. le calcul de l'entropie du paysage. Bien qu'une mise en œuvre manuelle de QHMB basée sur l'analyse des distributions de dièdres et le raffinement visuel puisse être suffisamment précise, cette procédure n'est pas pratique, si l'objectif est de l'appliquer à des centaines ou des milliers de composés. C'est dans ce but qu'une procédure automatique de QHMB a été développée. Suivant Suarez *et al.*,^{280,292} l'implémentation vise à: i. identifier les conformères stables par regroupement RMSD d'une trajectoire MD étendue; ii. extraire une série de sous-trajectoires correspondant à chacun d'eux; et iii. analyser ces sous-trajectoires par QHA automatiquement. À cette fin, plusieurs algorithmes hiérarchiques, dont le Average-Linkage³³⁰, le Single-Linkage³³¹ et le Complete-Linkage³³², ainsi que le DBSCAN³³³ basé sur la densité, ont été envisagés; notons que toutes les méthodes de clustering font partie de la suite logicielle Amber18¹²². En outre, étant donné qu'une décomposition correcte de l'espace configurationnel en puits d'énergie potentielle distincts est essentielle pour

une évaluation correcte de l'entropie absolue, l'analyse QHMB a été répétée en faisant varier le seuil de RMSD pour le regroupement de 2.0Å à 0.1Å par décrets de 0.1Å. Pour valider la procédure, l'ensemble de données du NIST pour lequel des entropies expérimentales sont disponibles (voir ci-dessus) a été utilisé comme référence. Les résultats sont présentés dans la **Figure 2B**. À des seuils élevés, c'est-à-dire RMSD \sim 2Å, toutes les méthodologies sont équivalentes et donnent des résultats d'entropie avec une erreur systématique aussi importante que celle de l'QHA standard. Plus le seuil est bas, plus l'erreur de calcul de l'entropie est faible. Il est intéressant de noter qu'en diminuant le seuil de RMSD en dessous de 0.5Å, les algorithmes de Average-Linkage (vert) et de Complete-Linkage (noir) améliorent tous deux les prédictions de QHMB de manière assez significative et atteignent un plateau avec une RMSE inférieure à 1 kcal/mol. D'autre part, ni l'algorithme de liaison simple ni l'algorithme DBSCAN n'ont atteint une précision satisfaisante dans la gamme de seuils étudiée. Puisque le regroupement basé sur la liaison moyenne a produit la RMSE la plus faible à des seuils inférieurs à 0.5Å, ce protocole a été choisi pour toutes les études ultérieures. Nous notons au passage qu'à des cutoffs trop grands (\geq 2 Å), tous les algorithmes de clustering échouent car ils mélangent des conformations appartenant à des bassins différents et QHMB se réduit à un QHA standard.

Application du QHMB aux calculs de l'énergie libre de liaison au point final MM/GBSA

MM/PBSA et MM/GBSA sont des approches très populaires pour le calcul des affinités de liaison relatives des ligands.^{88,108,120,196,207,208,334} Cependant, il a été remarqué que leurs performances dépendent fortement du système.^{206,209} De plus, de nombreux chercheurs préfèrent ne pas inclure les contributions entropiques dans leurs calculs MM-PB(GB)SA en raison du coût de calcul supplémentaire et des observations précédentes selon lesquelles une inclusion explicite de l'entropie peut détériorer la corrélation des affinités de liaison prédites avec les expériences.^{196,206,209,280} D'autres, pour surmonter les limites des calculs d'entropie dans les réactions de liaison, ont proposé des stratégies alternatives qui ne reposent pas sur l'approximation harmonique, comme la méthode de l'entropie d'interaction.⁴⁹ Motivés par la précision des résultats du QHMB en phase gazeuse (voir ci-dessus), nous avons sélectionné un ensemble de données de 21 complexes protéine-ligand à partir de l'ensemble de données de Greenidge²⁰⁹ et utilisé le QHMB pour quantifier la perte d'entropie lors de la liaison. Plus précisément, QHMB a été utilisé pour évaluer l'entropie configurationnelle du ligand dans ses états lié et non lié à partir de simulations MD indépendantes, afin d'estimer l'entropie de liaison à partir de la différence entre les deux. Ce terme a ensuite été introduit comme une correction

d'entropie dans les calculs MM/GBSA standard. Les résultats de l'énergie libre de liaison de MM/GBSA avec et sans la correction d'entropie du QHMB sont présentés dans la **Figure 3**. À titre de comparaison, la perte d'entropie du ligand a également été accessible par le QHA²⁷⁹ standard et les performances des deux protocoles ont été comparées; voir le **Tableau 2**. Les données montrent que les calculs MM/GBSA standard tels qu'ils sont mis en œuvre dans AmberTools18¹²² font un travail raisonnable avec cet ensemble de données, donnant un $R^2 = 0,67$. L'application d'une correction de l'entropie du ligand basée sur QHA introduit une erreur plus importante et la corrélation diminue de 17% ($R^2 = 0,5$), ce qui est cohérent avec les rapports précédents.^{206,270} En revanche, l'introduction de la correction de l'entropie du ligand par QHMB augmente la corrélation de 11% et donne un R^2 final = 0,78; voir **Tableau 2**. Ces résultats conduisent aux observations suivantes. Premièrement, la correction de l'entropie par QHMB introduit une pénalité dans le ΔG° calculé, qui tient compte de la restriction du volume configurationnel accessible au ligand dans son état lié. Deuxièmement, la taille de la correction dépend fortement du ligand et introduit une pénalité plus importante pour les ligands grands et flexibles; c'est-à-dire que la correction QHMB est > 9 kcal/mol pour quatre ligands de l'ensemble de données, alors qu'elle est de 6 kcal/mol en moyenne. De plus, la correction repose sur l'échantillonnage configurationnel par MD, ce qui permet de sonder directement le changement du volume configurationnel du ligand indépendamment de sa flexibilité intrinsèque. Pris ensemble, ces résultats suggèrent que l'introduction d'une correction d'entropie dépendante du ligand basée sur le QHMB augmente la précision des calculs d'affinité de liaison relative.

Tableau 2 - Inclusion de la perte d'entropie du ligand dans les calculs du MMGBSA.

Experiment	R	R ²	Pente	$\rho^{[4]}$
MM-GBSA ^[1]	0.82	0.67	2.15	0.79
MM-GBSA+QHA ^[2]	0.71	0.50	1.33	0.66
MM-GBSA+QHMB ^[3]	0.88	0.78	1.87	0.88

[1] Résultats du MMGBSA à trajectoire unique. [2] Résultats MMGBSA avec correction d'entropie par QHA. [3] Résultats MMGBSA avec correction d'entropie par QHMB. [4] Coefficient de corrélation de Spearman. Toutes les simulations ont été effectuées à 298,15 K et 1M.

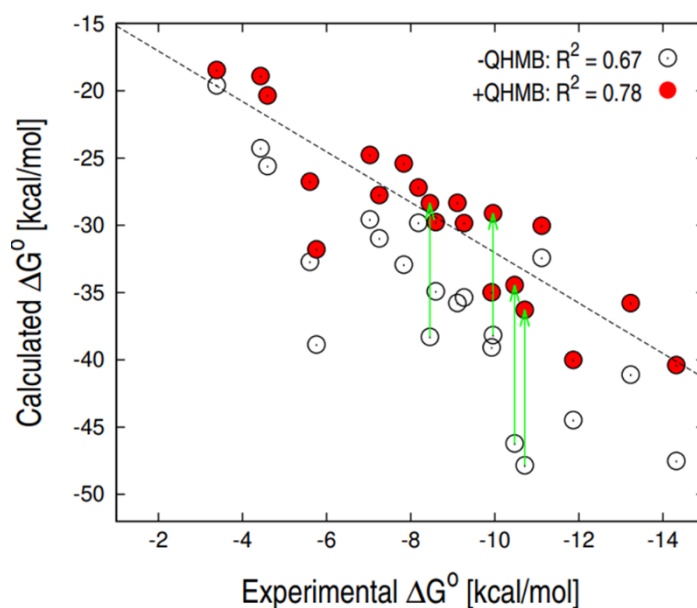


Figure 3 - Corrélation entre les affinités de liaison expérimentales et prédites par MMGBSA avec (rouge) et sans (points vides) correction d'entropie par QHMB. Les données montrent que l'introduction de la correction entropique augmente la corrélation avec les expériences de 11%. Pour certains ligands grands et flexibles, la correction est aussi grande que 10 kcal/mol (flèches vertes).

Identification de composés "hit" dans la quête de modulateurs allostériques du moteur moléculaire de la myosine

Résultats de la précédente campagne VS réalisée sur la Myosine Muscle Lisse II

L'identification d'inhibiteurs de la SMM2 est une tâche difficile.³⁴⁷ Le site de liaison de la CK571 est une poche allostérique, s'ouvrant transitoirement pendant la phase de récupération du cycle du moteur de la myosine.³⁴³ Notre approche VS emploie une stratégie originale basée sur la préparation d'une bibliothèque de ligands, le docking moléculaire des ligands à la structure cristallographique SMM2-CK571 (PDB **5M05**) et le rescoring de l'énergie libre des composés les plus prometteurs. Lors d'une précédente campagne VS, un sous-ensemble de composés de la Chimiotèque Nationale du CNRS (CN)356 (> 60k composés) a été priorisé à l'aide de calculs MM/GBSA et testé expérimentalement (**Figure 4**). L'expérience consistait à évaluer si les ligands pouvaient ralentir la vitesse d'hydrolyse de l'ATP de SMM2 en évaluant la diminution de l'absorbance à 340 nm (A340) due à la consommation de NADH lorsque l'ATP est hydrolysé en ADP. Comme on peut l'observer sur la **Figure 5**, aucun des composés n'a été actif car les valeurs des pentes obtenues par l'ajustement d'une fonction à la décroissance de l'A340 sont similaires à celles des expériences sans aucun composé et avec le seul DMSO.

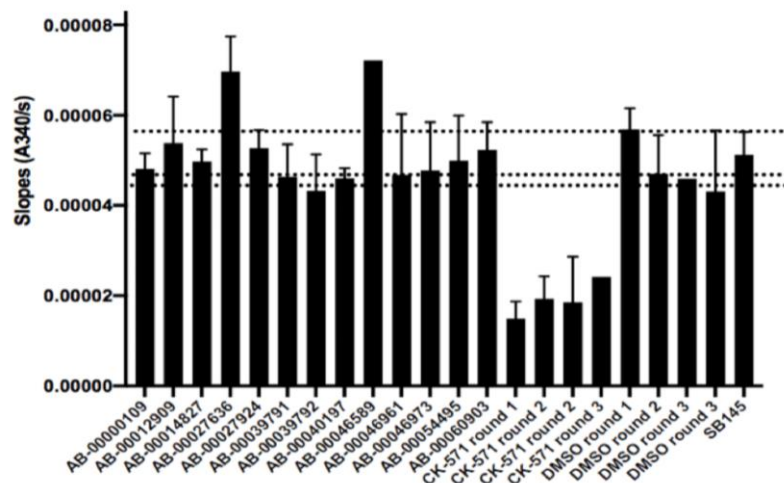


Figure 4 - Mesures expérimentales de l'activité ATPase. Les essais ont été réalisés sur le CK-571 (contrôle positif), le DMSO (contrôle négatif) et les composés prioritaires du CN. Le test surveille les changements d'absorbance à 340 nanomètres (A340), qui sont couplés à l'oxydation du NADH par une série de réactions enzymatiques couplées. Les lignes en pointillés sont données par les mesures de l'A340 sur l'expérience de contrôle négatif, qui est réalisée en utilisant du DMSO et aucun inhibiteur.

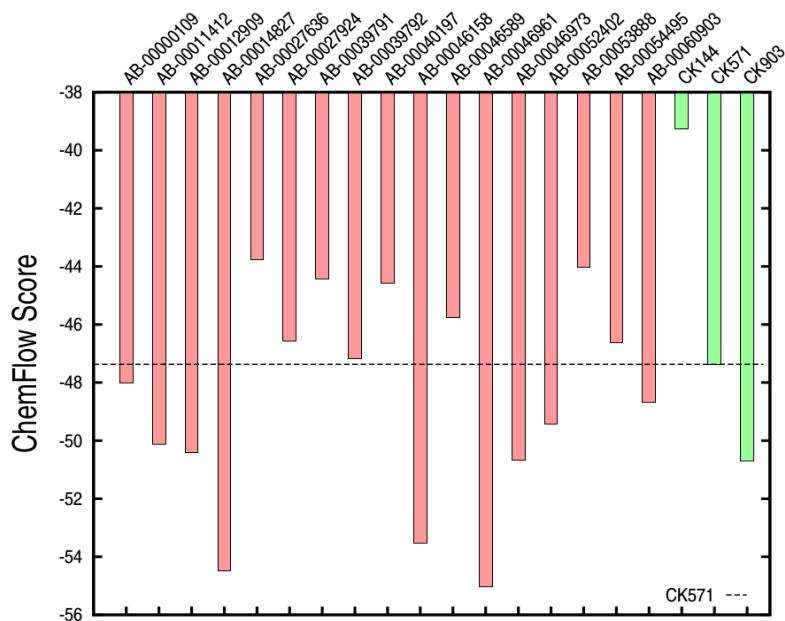


Figure 5 - Energies libres de liaison prédites pour le sous-ensemble de composés prioritaires de la précédente campagne VS ciblant SMM2.

Bien que ces résultats soient décevants, car ils montrent que l'approche VS n'a pas été capable de trouver des composés actifs, ils ont permis de réfléchir sur la configuration MM/GBSA employée. En inspectant les paramètres utilisés dans les calculs, trois termes possibles qui pourraient être optimisés dans une future campagne ont été trouvés : la constante diélectrique interne du soluté (ϵ), le modèle GB et l'inclusion de termes entropiques dans les calculs MM/GBSA. Le fait de retenir des ligands grands et flexibles

dans un site de liaison entraîne un coût entropique qui, lorsqu'il n'est pas payé, conduit à des calculs MM/GBSA biaisés où les ligands plus grands sont prédits comme de meilleurs liants.

Calibration des calculs MM/GBSA pour une deuxième campagne VS sur SMM2

Avant d'entreprendre une nouvelle campagne VS, le flux de travail VS a dû être ajusté pour essayer d'éviter une deuxième série d'expériences infructueuses. Les paramètres pour les calculs MM/GBSA ont été calibrés sur la base de données de référence, où le jeu de données était composé des trois inhibiteurs connus et du sous-ensemble de composés prioritaires trouvés inactifs dans la campagne VS précédente. Les énergies libres de liaison MM/GBSA obtenues à partir de simulations MD à solvant explicite pour cet ensemble de données ont été utilisées comme données de référence et corrélées avec les résultats MM/GBSA obtenus à partir d'ensembles configurationnels échantillonnés par MD à solvant implicite. L'effet de ϵ et du modèle GB sur la corrélation entre les prédictions des ensembles MD à solvant implicite et explicite a été interrogé simultanément. La configuration MM/GBSA sélectionnée pour la nouvelle campagne VS était celle qui présentait la plus forte corrélation entre ces calculs (**Tableau 3**). La corrélation a été évaluée en évaluant le coefficient de corrélation de Pearson au carré (R^2) et la corrélation de Spearman des rangs (ρ).

Tableau 3 - Calculs de référence effectués sur le jeu de données des composés prioritaires du criblage précédent. En rouge, la configuration utilisée lors du premier criblage est mise en évidence, tandis qu'en vert, la configuration pour les études futures est mise en évidence.

GB Model	R^2 ($\epsilon = 1$)	ρ ($\epsilon = 1$)	R^2 ($\epsilon = 2$)	ρ ($\epsilon = 2$)	R^2 ($\epsilon = 4$)	ρ ($\epsilon = 4$)	R^2 ($\epsilon = 10$)	ρ ($\epsilon = 10$)
GB1	0.54	0.62	0.63	0.75	0.64	0.75	0.64	0.74
GB2	0.37	0.45	0.59	0.69	0.64	0.74	0.64	0.76
GB5	0.29	0.43	0.55	0.65	0.62	0.76	0.64	0.75
GB7	0.35	0.47	0.59	0.68	0.64	0.76	0.64	0.76
GB8	0.07	0.40	0.48	0.65	0.60	0.76	0.63	0.76

Le benchmark montre que la configuration utilisée dans le criblage précédent (GB = 2, $\epsilon = 1$) présentait une faible corrélation entre les calculs effectués sur les données de simulation MD à solvant implicite et explicite ($R^2 = 0,37$, $\rho = 0,45$). Les calculs MM/GBSA

à solvant implicite configurent une étape de filtrage dans le workflow VS dans le but de sélectionner un sous-ensemble de composés à étudier avec des simulations MD à solvant explicite plus longues et donc plus coûteuses. Ainsi, le fait que la corrélation soit faible diminue notre confiance dans l'étape de filtrage et soulève la question de savoir si le sous-ensemble de composés prioritaires est significatif. L'objectif final de ce benchmark était de trouver une configuration MM/GBSA avec le plus grand accord possible entre les calculs effectués sur les deux types de simulations, car cela signifierait que les étapes de rescore de l'énergie libre seraient cohérentes entre elles. Un degré de corrélation plus élevé a toujours été obtenu lorsque ϵ était fixé à 4 ou 10, comme le montre le **Tableau 3**.

Dans les calculs où ϵ est fixé à 4 ou 10, de petites variations dans la prévisibilité des calculs MM/GBSA ont été observées lors de l'utilisation de différents modèles de GB. En considérant le coefficient de Spearman et le coefficient de Pearson au carré, on a constaté que GB = 7 et ϵ = 4 ou 10 étaient les combinaisons les plus prédictives. En ajustant simultanément le modèle GB et ϵ , nous avons obtenu une corrélation modérée entre les calculs, de $R^2 = 0,37$ à $0,64$ et de $\rho = 0,45$ à $0,76$. Ainsi, $\epsilon = 4$ a été choisi car il s'agit d'une valeur habituellement utilisée selon la littérature pour les systèmes protéine-ligand.^{196,251} Les données indiquent que lors du premier criblage, la configuration utilisée pour le rescore de l'énergie libre MM/GBSA était sous-optimale et produisait donc de nombreux faux positifs. L'introduction d'une constante diélectrique plus élevée et l'optimisation du modèle GB utilisé pour calculer la contribution polaire à l'énergie libre de solvation semblent résoudre ce problème. La configuration optimisée permet de différencier les composés actifs connus des inactifs, ce qui n'était pas possible auparavant. Pour résoudre le problème du biais dû à la négligence des contributions entropiques, un terme de correction tenant compte de la perte d'entropie configurationnelle du ligand lors de la liaison en utilisant le QHMB a été ajouté dans l'étape finale de recalibrage de la dernière campagne VS. Ainsi, le biais en faveur des ligands plus flexibles trouvé dans la campagne VS initiale devrait également être pris en compte.

Flux de travail de la nouvelle campagne VS

La Chimiotèque Nationale du CNRS (CN)³⁵⁶ (> 60k composés) a été sélectionnée pour la nouvelle campagne VS, normalisée et préparée par PrepFlow^{165,369}, un pipeline automatisé développé en interne pour la préparation de bibliothèques de ligands (**Figure 6**), et filtrée par DataWarrior.¹ Un schéma décrivant la préparation de la bibliothèque CN est présenté à la **Figure 6**. Après la préparation de la bibliothèque et le filtrage R-o-5, 39 mille composés ont été retenus. Ceux-ci ont ensuite été filtrés en fonction de la présence de

sous-structures PAINS spécifiques.¹⁶⁰ L'ensemble de données filtré par PAINS contenait 22 000 composés qui ont ensuite été élagués sur la base de propriétés physicochimiques telles que la surface polaire (PSA), la solubilité calculée (clogS), le nombre de liaisons rotatives et la présence de certains groupes d'ogives toxiques dans leurs structures moléculaires (RTECS).⁵⁴ Le but de cette étape était de concentrer la bibliothèque sur les composés de type médicament ayant des caractéristiques physicochimiques similaires à celles du CK-571. La taille de l'ensemble de données a considérablement diminué à ce stade, pour un total de 8K composés. Enfin, un descripteur de flexophore a été calculé pour chaque ligand.³⁵⁸ La bibliothèque a été regroupée à l'aide d'un descripteur de flexophore 3D, ce qui a permis de réduire l'ensemble de données à environ 2300 structures chimiquement diverses, chaque structure représentant un groupe de composés, regroupés sur la base de la similarité 3D.

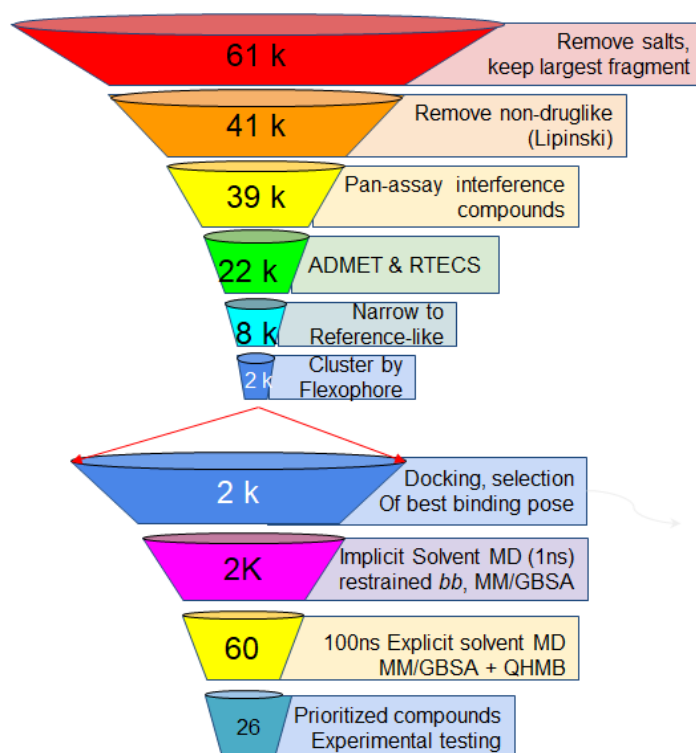


Figure 6 - Représentation schématique du déroulement de la campagne VS menée sur le CN en vue de la découverte de nouveaux modulateurs SMM2.

Deuxième campagne de criblage virtuelle sur SMM2

La méthode MM/GBSA^{189,196,206} a été utilisée pour calculer les énergies libres de liaison des complexes protéine-ligand simulés en solvant implicite en utilisant le même modèle GB et la même constante diélectrique interne du soluté que ceux utilisés pour exécuter la simulation.

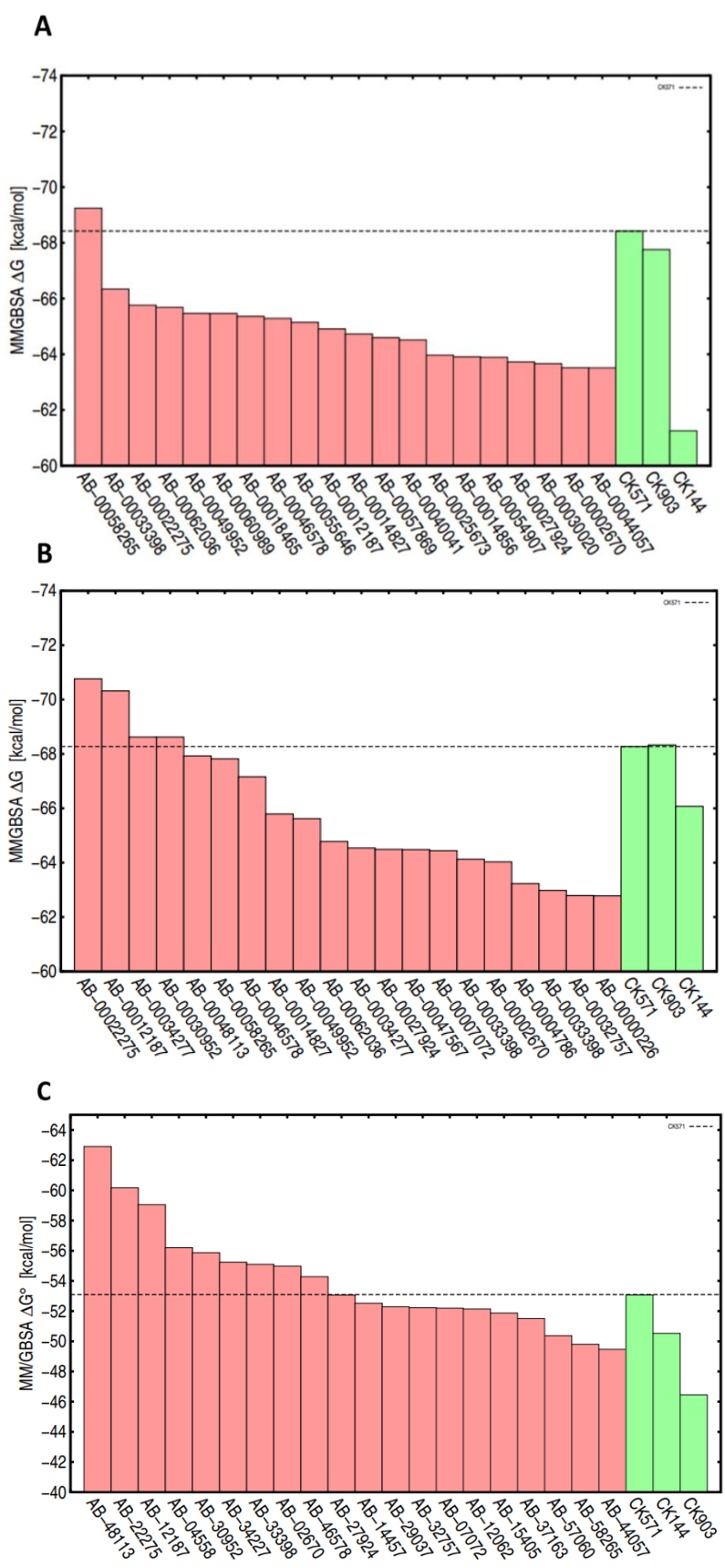


Figure 7 - Energies libres de liaison obtenues pour les 20 premiers composés de la CN à différentes étapes

de la campagne de criblage sur la Chimiotèque Nationale. Les barres rouges font référence aux énergies libres de liaison prédites pour les composés de la Chimiotèque et les barres vertes correspondent aux énergies libres de liaison prédites des composés CK de référence. La ligne pointillée met en évidence l'énergie libre de liaison prédite du CK-571, que nous utilisons comme référence. A) Les 20 premiers composés obtenus suite aux calculs MM/GBSA utilisant des trajectoires de solvant implicites. B) Les 20 premiers composés obtenus après les calculs MM/GBSA utilisant des trajectoires de solvant explicites. C) Les 20 premiers composés obtenus suite aux calculs MM/GBSA utilisant des trajectoires de solvant explicites et incluant la correction entropique QHMB.

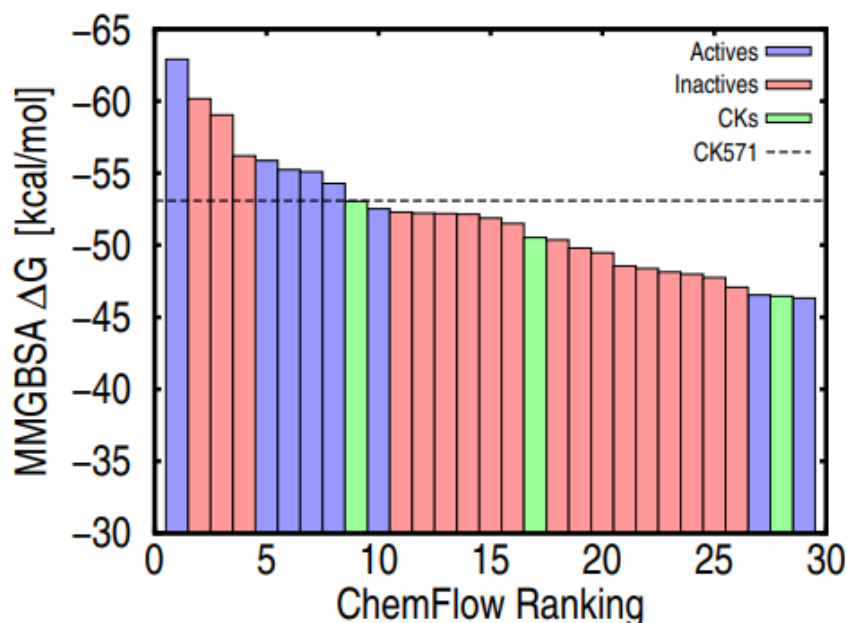
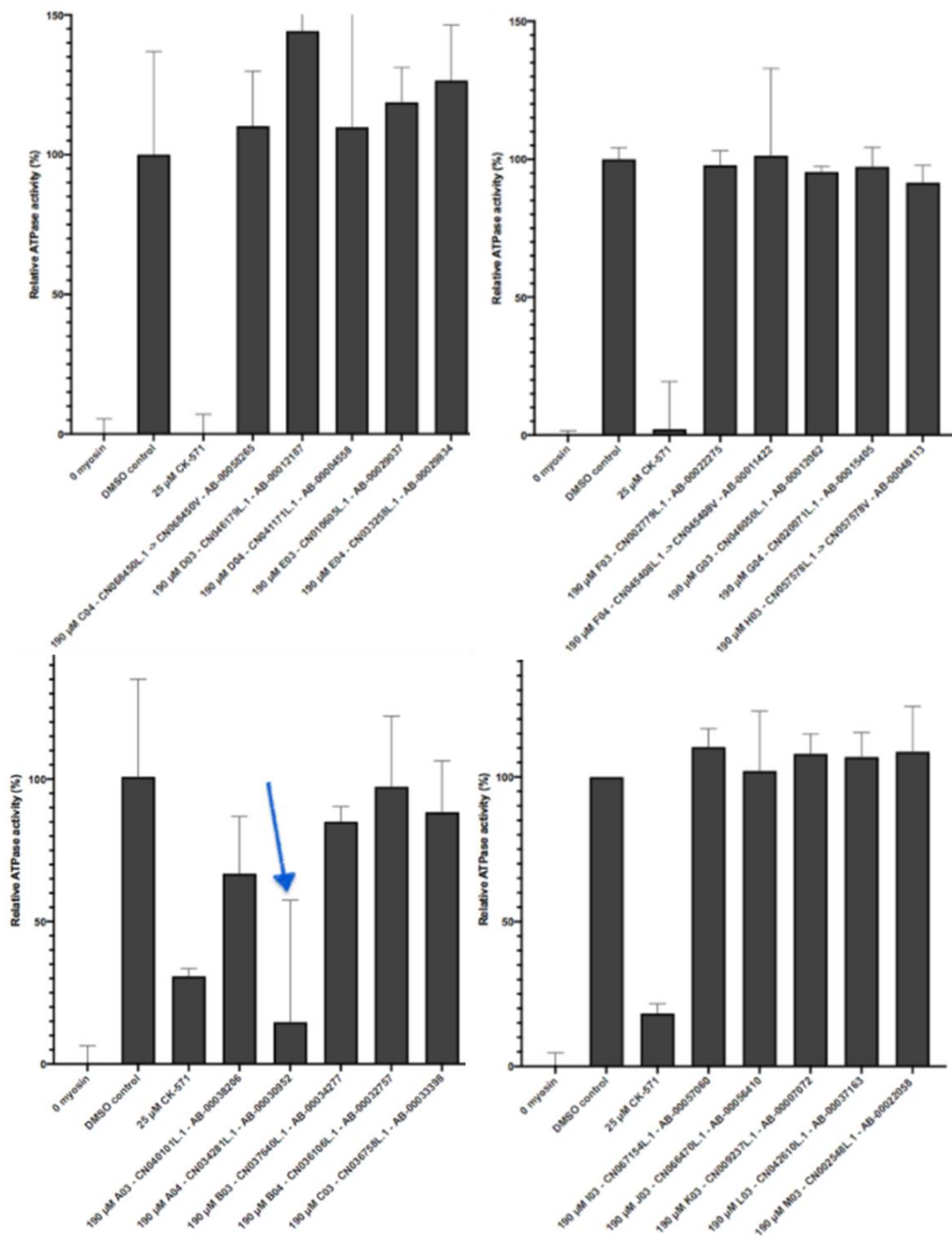


Figure 8 - Résultats du criblage virtuel de la campagne contre la myosine II du muscle lisse (SMM2). L'affinité prédite des composés prioritaires de la Chimiotèque Nationale est indiquée en rouge. L'affinité prédite de trois inhibiteurs connus de Cytokinetics est indiquée en vert. Les composés ayant une activité détectable dans les tests in-vitro (voir ci-dessous) ou les hits sont indiqués en bleu.

Un sous-ensemble contenant les 60 composés les mieux classés à partir des calculs effectués à l'étape du solvant implicite a été sélectionné pour des études plus poussées en utilisant le MD en solvant explicite. À l'étape du solvant implicite, un seul ligand, AB-00058265, a été prédit avec une meilleure affinité de liaison que le CK-571. Après 100ns de simulations MD avec solvant explicite, l'énergie libre de liaison des composés a été calculée en utilisant MM/GBSA et les composés ont été classés à nouveau. Un certain reclassement a été observé, mais en général, les résultats des solvants implicites et explicites étaient bien corrélés (**Figure 7A** et **7B**). Cependant, on a remarqué que les composés les mieux classés étaient dans la plupart des cas des ligands grands et flexibles, ce qui est un artefact connu des calculs MM/GBSA lorsque les termes entropiques sont négligés, comme c'était le cas dans nos calculs. Ainsi, ces calculs ont été complétés en utilisant la correction QHMB pour tenir compte du coût entropique de la retenue du ligand dans le site de liaison, calculé en prenant la différence entre l'entropie QHMB du

ligand dans l'état lié, en extrayant les coordonnées du ligand de la simulation du complexe, et dans l'état non lié en solution. Ainsi, le protocole émergent est le suivant : (1) utiliser le docking pour produire les coordonnées initiales du complexe ; (2) utiliser le MD à solvant implicite pour un classement et un filtrage rapides ; (3) compléter par le MD à solvant explicite et le QHMB pour le classement final. Après la correction QHMB, un reclassement significatif des composés a été observé. Nous avons trouvé plusieurs composés avec des énergies libres de liaison prédites similaires à celles du CK-571 (Figure 7B et 7C).



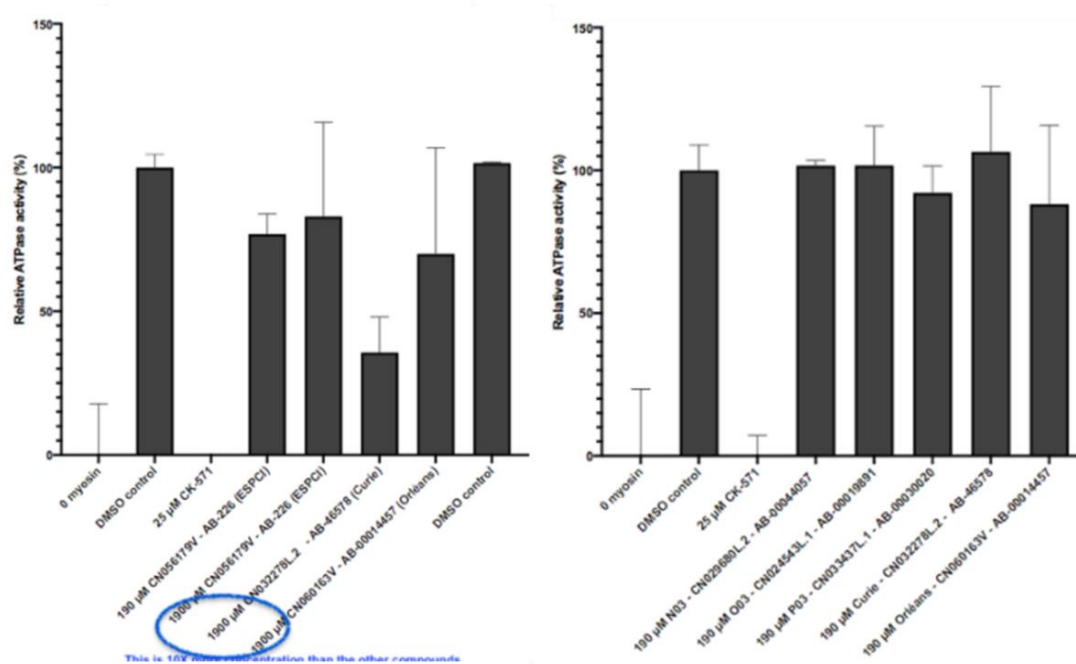


Figure 9 - Résultats du test d'inhibition de l'ATPase SMM2 réalisé à l'Institut Curie. Sur l'axe des y est reportée l'activité ATPase relative de SMM2 en présence de DMSO, de CK-571 ou de chaque ligand du CN acquis à 190 μ M. La première colonne correspond à l'activité ATPase en absence de myosine. La flèche bleue met en évidence le composé AB-00030952, qui est le composé le plus actif. Cependant, ce composé a une barre d'erreur très large. Trois composés sont représentés avec un cercle bleu autour, correspondant aux composés dont l'activité a été évaluée à très haute concentration (1,9 mM).

En particulier, les ligands très flexibles ont été pénalisés plus fortement que les ligands plus rigides par le QHMB. Par exemple, le composé AB-00048113 était classé 7ème avant la correction QHMB. Après l'application de la correction, il est devenu le composé le mieux classé alors que le composé qui était classé 8ème (AB-00058265) avant la correction QHMB est maintenant classé 21ème. Les résultats de la figure 7 illustrent l'effet du passage de simulations MD à solvant implicite à des simulations MD à solvant explicite, ainsi que l'effet de l'inclusion de la correction QHMB dans les prédictions d'énergie libre de liaison MM/GBSA. En outre, la CK-571 et la CK-903 ont été prédites avec des énergies libres de liaison similaires avant la correction, tandis que la CK-144 est moins affine.

Après la correction du QHMB, le CK-903 a été fortement pénalisé et est devenu le pire des trois inhibiteurs. Parmi les composés les mieux classés, trois ont une affinité de liaison supérieure à celle de CK571, et environ 20 composés présentent des affinités comparables à celles des inhibiteurs connus de SMM2 de Cytokinetics. Sur un total de 60 composés, 26 ont été acquis et envoyés pour des tests expérimentaux. Les résultats expérimentaux ont été recueillis par nos collaborateurs de l'Institut Curie à une concentration de 190 μ M en présence de 2 μ M de SMM2 et de 25 μ M ou 40 μ M d'actine

en utilisant un test d'inhibition de l'ATPase (**Figure 9**). En fixant comme zéro de l'inhibition le signal recueilli en présence de DMSO, correspondant à l'expérience de contrôle, les données montrent 8 composés avec une inhibition détectable. La plupart de ces composés ont une activité légère mais détectable ($IC_{50} > 100 \mu M$) et sont des entités chimiques uniques. Fait intéressant, et malgré une barre d'erreur importante, AB-00030952 présente une inhibition de 85 % de SMM2 à $190 \mu M$. En considérant ces composés comme des succès, le taux de succès du protocole de criblage est d'environ 30 %. Ces composés sont des inhibiteurs légers et on pourrait donc dire que l'activité capturée est si faible qu'ils ne devraient pas être considérés comme des succès. Cependant, comme l'indique la revue de Hevener *et al.*⁴⁷, le seuil utilisé pour définir les composés actifs ou inactifs dans les campagnes HTS varie dans la littérature. En particulier, Hevener *et al.* détaillaient à l'époque que 56 études utilisaient un seuil d'activité compris entre 100 et $500 \mu M$ et 25 études utilisaient un critère supérieur à $500 \mu M$. Ils justifient le fait de définir comme hits des composés aussi peu actifs par la volonté d'enrichir la bibliothèque de hits en termes de diversité structurale.⁴⁷ Parmi les composés présentant une activité détectable *in vitro*, 5 sur 8 figurent dans les 10 meilleures prédictions ; voir les barres bleues de la **Figure 8**.

Analyse structurale des composés prédits

Suite à la campagne de criblage *in silico*, huit nouveaux inhibiteurs de SMM2 ont été trouvés. Bien que ceux-ci présentent une activité légère, ils peuvent être utiles pour comprendre comment mieux explorer la poche de liaison et à quel point cette poche est plastique et adaptable.

La mise à disposition d'une structure cristallographique du complexe SMM2/CK-571 constitue une avancée majeure pour le développement de nouvelles voies thérapeutiques ciblant SMM2. En particulier, cette structure montre une poche allostérique inconnue jusqu'alors où la liaison du ligand inhibe l'activité de SMM2 en stabilisant un état intermédiaire avec une faible affinité pour l'actine et en piégeant le domaine moteur de la myosine. On a découvert que l'inhibiteur cocrystallin avait un IC_{50} de 12 nM, ce qui en fait un inhibiteur SMM2 très puissant. Plus récemment, les structures cristallines de deux autres inhibiteurs de la cytokinétique ont été résolues (données non publiées) en complexe avec SMM2. De façon intéressante, ces composés ciblent la même poche allostérique dans le même état, et ont des modes de liaison très similaires à ceux du CK571. Pour ces ligands, connus sous le nom de CK-144 et CK-903, aucune valeur IC_{50} n'est disponible mais ils sont des inhibiteurs de SMM2. L'inhibiteur CK-571 se lie à SMM2 en insérant un fragment hydrophobe de part et d'autre de l'hélice SH1, la partie

carbamate d'isoquinoléine (Poche 1; P1) étant maintenue fixe entre l'hélice relais et l'hélice SH1, une partie hydrophobe de l'autre côté (Poche 2; P2) et une queue s'étendant vers l'extrémité N-terminale (Poche 3; P3). (**Figure 10A et 10B**). Comme l'ont noté Sirigu et al.³⁴³, le CK-571 n'établit aucune liaison hydrogène directe avec les résidus protéiques et ses interactions sont principalement hydrophobes, s'appuyant sur un réseau de contacts de van der Waals pour interagir avec SMM2 (**Figure 10A et 10B**). En se liant à la poche allostérique de SMM2 de cette manière, la CK-571 empêche la réprimande du bras de levier et arrête le cycle dans cet état intermédiaire entre l'état de rigueur et l'état de pré-puissance (**Figure 10B**).

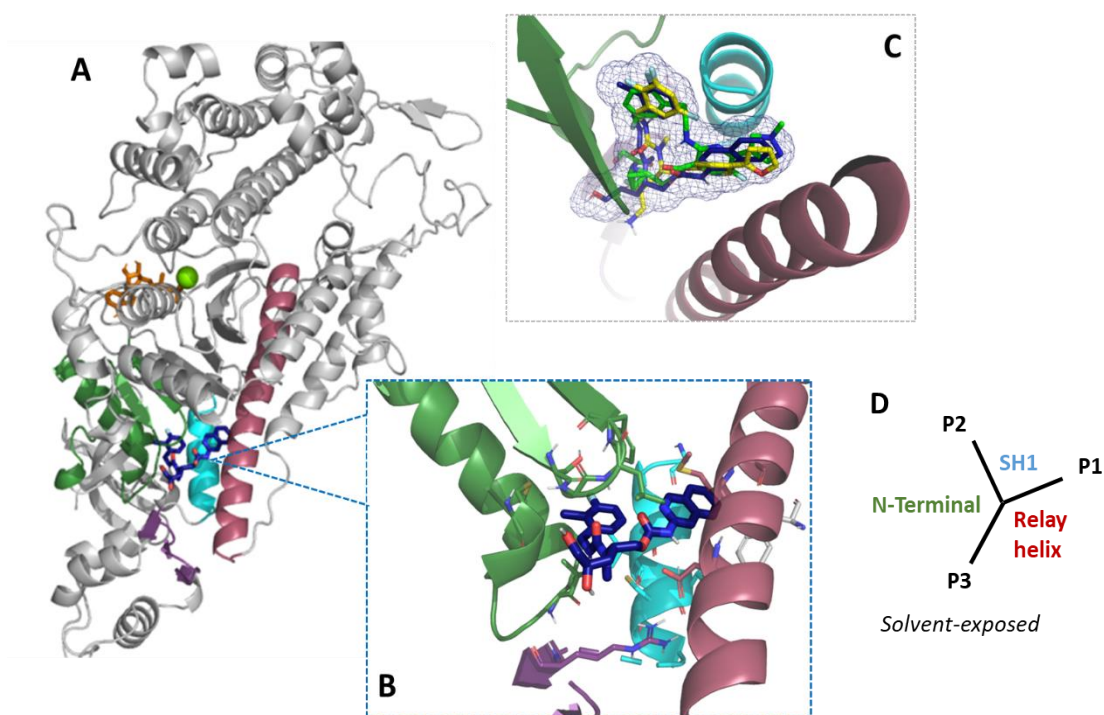


Figure 10 - Vue tridimensionnelle des systèmes SMM2/inhibiteurs. Le domaine N-terminal est représenté en vert foncé, l'hélice Relay est représentée en rouge, l'hélice SH1 est surlignée en cyan et les trois feuillets β du domaine convertisseur sont représentés en violet. **A)** Structure du complexe biomoléculaire SMM2/CK-571. La CK-571 est représentée en bleu foncé, l'ADP en orange et l'ion magnésium sous la forme d'une sphère verte. **B)** Zoom sur le site de liaison de la CK-571, montrant les résidus bordant le site de liaison sous forme de bâtonnets et colorés selon le type d'atome. La vue de dessus permet de voir le groupement carbamate de la CK-571 inséré entre le Relais et l'hélice SH1 et le groupement chloro-fluoro-phényle de la CK-571 inséré de l'autre côté de l'hélice SH1. Elle met également en évidence la présence d'une queue polaire s'étendant vers le domaine N-terminal. **C)** Superposition des trois inhibiteurs de la cytokinétique SMM2 dans le site de liaison. Une représentation maillée du volume de la CK-571 est représentée en bleu, la structure moléculaire de la CK-144 est représentée en jaune et la structure moléculaire de la CK-903 est représentée en vert clair. **D)** Schéma illustratif du mode de liaison des inhibiteurs de la CK dans la poche allostérique du SMM2.

Le mode de liaison de CK-571 est partagé par les deux autres inhibiteurs de la cytokinétique, où l'hélice SH1 est entourée de chaque côté et une queue polaire s'étend

vers l'extérieur dans le domaine N-terminal. Cependant, la CK-903 enveloppe l'hélice SH1 sans explorer complètement le volume disponible sur le côté gauche de l'hélice SH1 au lieu d'insérer ses groupes chimiques profondément dans P2 (**Figure 10C**). Pour la CK-144, le mode de liaison est similaire à celui de la CK-571, en insérant un groupement dans P2 et un autre gros groupement hydrophobe dans P1, entre l'hélice SH1 et l'hélice Relay. Un autre point important à soulever est que les interactions entre la CK-144 et SMM2 sont principalement non polaires et, comme pour la CK-571, il semble que la CK-144 n'établisse aucune liaison hydrogène directe avec SMM2. La **Figure 10A** montre que le site de liaison allostérique de la CK-571 est éloigné du site de liaison des nucléotides et du domaine de liaison de l'ATP.

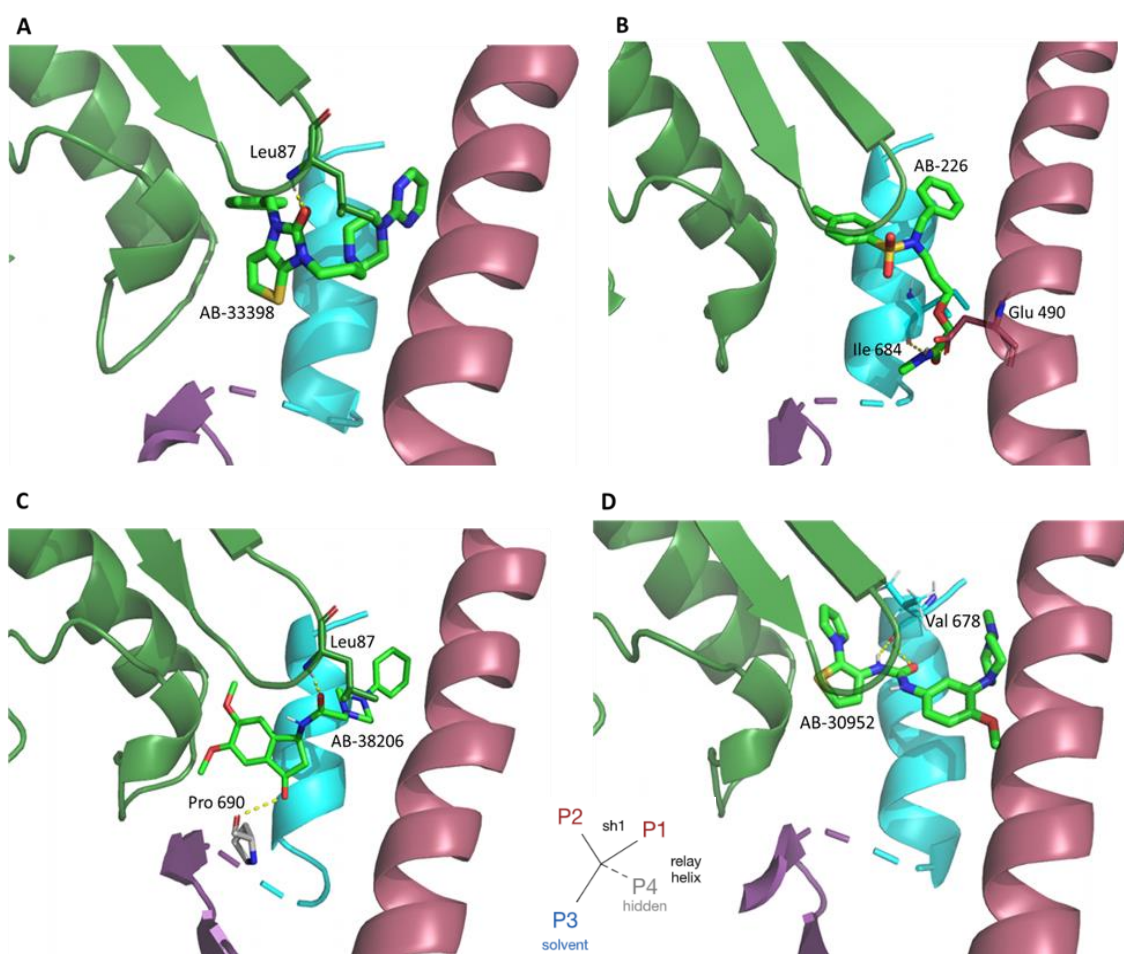


Figure 11 - Mode de liaison de certains des composés trouvés par criblage virtuel de la Chmiotèque National du CNRS. Les liaisons hydrogène établies entre les ligands (vert clair) et SMM2 sont indiquées en jaune à côté du nom et du numéro du résidu. **A)** Mode de liaison de AB-00033398, classé 6ème par notre approche VS. **B)** Mode de liaison de AB-0000226, classé 37ème par notre approche VS. **C)** Mode de liaison de AB-38206, classé 40ème par notre approche VS. **D)** Mode de liaison de AB-00030952, le composé le plus actif, dans les essais expérimentaux, classé 5ème par notre approche VS.

Cependant, en limitant le mouvement de l'hélice Relais, il empêche l'hydrolyse de l'ATP et arrête le cycle au début de la transition conformationnelle vers l'état de pré-course. Il est évident que bien que la poche ne semble pas être entièrement remplie par les ligands CK, les ligands remplissent la grande majorité de la poche, établissant de nombreux contacts de van der Waals, et il n'est donc pas surprenant que les interactions non polaires semblent être la principale force motrice stabilisant la liaison des CK. De plus, le fait que les trois ligands se lient dans la poche avec des modes de liaison très similaires implique qu'une géométrie particulière du ligand est nécessaire pour inhiber l'activité de SMM2 en ciblant cette poche allostérique. Les informations provenant de la structure cristalline soulignent que les liants potentiels doivent avoir deux parties hydrophobes pour envelopper l'hélice SH1 de chaque côté et une queue polaire s'étendant vers l'extérieur vers le domaine N-terminal pour une liaison appropriée dans la poche, en forme de Y comme le montre la **Figure 10D**. La comparaison du mode de liaison des composés CK aux composés hits montre que P1 n'est que partiellement occupé par les CKs et pourrait être mieux rempli (**Figures 10 et 11**). Le mode de liaison de certains des hits, notamment AB-00033398, explore P1 en insérant une terminaison pipérazine+aromatique qui s'enfonce dans la poche entre les hélices SH1 et Relay (**Figure 11A**). Ce ligand établit également une liaison hydrogène directe avec la protéine au niveau de la leucine 87, un résidu qui fait partie d'une boucle flexible. Cependant, comme la boucle est flexible, cette liaison hydrogène n'est pas maintenue tout au long de la simulation et l'ampleur de sa contribution est donc probablement négligeable. D'autres composés, comme AB-00000226 (**Figure 11B**), explorent P1 et P2 de manière moins profonde. Cependant, AB-00000226 explore également longitudinalement la crevasse le long des hélices SH1 et Relay, s'insérant entre elles et empêchant éventuellement leurs mouvements. De plus, en explorant cette crevasse s'étendant vers le bas à partir de P1, occupant effectivement une quatrième poche (P4), il est capable de maintenir deux liaisons hydrogène avec l'isoleucine 684 et la glutamine 490. Nous constatons également que P2 pourrait être mieux remplie. En particulier, il semble qu'elle puisse accueillir des groupes plus grands et plus volumineux, comme dans le mode de liaison de AB-00038206 (**Figure 11C**). En outre, AB-00038206 établit deux interactions de liaison hydrogène, avec la leucine 87 et la proline 690. De la même manière que pour AB-00033398 (**Figure 11D**), la liaison hydrogène avec la leucine 87 peut ne pas être très pertinente en raison de la flexibilité intrinsèque de la partie boucle. Cependant, la présence de la liaison hydrogène à la proline 690 est susceptible de contribuer à la stabilisation du mode de liaison.

Conclusion

De nombreuses études font état des limites associées aux calculs MM/PBSA, parmi lesquelles la négligence des contributions entropiques est un aspect essentiel. Pour résoudre ce problème tout en conservant l'efficacité des calculs, nous avons développé un schéma original pour calculer la perte d'entropie du ligand lors de la liaison dans le cadre du "mélange de conformères".^{280,297} En appliquant le QHMB, nous avons pu reproduire quantitativement les entropies absolues expérimentales de 23 petites molécules en phase gazeuse. Puisque dans QHA et QHMB l'effet des molécules de solvant est implicitement capturé à partir des fluctuations des degrés de liberté du soluté, cela ouvre au calcul des entropies des ligands en solution avec une application directe dans les calculs de l'énergie libre de liaison au point final. L'application de QHMB pour calculer la perte d'entropie du ligand lors de la liaison pour un ensemble de 21 complexes protéine-ligand a augmenté la corrélation avec les énergies libres de liaison expérimentales. De plus, les corrections d'entropie dépendent du ligand, pénalisant fortement les ligands les plus flexibles, ce qui rend compte du coût de la retenue des molécules flexibles et de grande taille à l'intérieur du site de liaison. Après le développement de QHMB, une campagne de VS a été menée à la recherche de nouveaux et puissants inhibiteurs de SMM2. La bibliothèque chimique CN a été criblée en utilisant le protocole implémenté dans *ChemFlow* complété par QHMB pour obtenir le classement final des composés. L'efficacité de la correction QHMB dans la prise en compte du coût de la retenue des ligands dans le site de liaison de la protéine représente une opportunité d'améliorer les résultats de la VS en réduisant le taux de faux positifs dans les calculs MM/GBSA. L'application de la correction QHMB aux calculs MM/GBSA effectués sur les composés du CN a conduit à un reclassement significatif. A partir du CN, 26 composés ont été acquis et testés expérimentalement dans un essai ATPase. Parmi ces 26 composés, huit ont montré une activité significative à une concentration de 190 μM , ce qui a conduit à un taux de réussite de 30 %. De plus, la structure moléculaire de ces composés est diverse et différente de celle du ligand cocrystallin, le CK571. Ainsi, cela ouvre la possibilité d'étudier différentes portions de l'espace chimique à la recherche d'inhibiteurs SMM2 innovants.

List of Publications

Adrien H. Cerdan, Marion Sisqueillas, Gilberto Pereira, Diego E. Gomes, Jean-Pierre Changeux, & Marco Cecchini. The Glycine Receptor Allosteric Ligand Library (GRALL). *Bioinformatics*. **2020**, 36(11), 3379-3384. doi: 10.1093/bioinformatics/btaa170

Gilberto Pereira & Marco Cecchini. Multi-basin Quasi-Harmonic Approach for the Calculation of the Configurational Entropy of Small Molecules in Solution. *Journal of Chemical Theory and Computation*. **2021**, 17(2), 1133-1142. doi: 10.1021/acs.jctc.0c00978.

Gilberto Pereira, Carlos Kikuti, Diego Gomes, Catherine Guillou, Marco Cecchini, and Anne Houdusse. Identification of hits on the quest for allosteric modulators of the myosin molecular motor. *In Preparation*. **2021**.

List of Talks and Presentations

Gilberto Pereira, Diego E. Gomes & Marco Cecchini. Improving the accuracy of MM/GBSA binding free energy calculations by including the conformational entropy loss of the ligand. *21e congrés du Groupe de Graphisme et Modélisation Moléculaire*, Nice, 3-5 Avril 2019. (Poster)

Gilberto Pereira & Marco Cecchini, M. A multi-basin Quasi-harmonic approach for the calculation of the Configurational Entropy of small molecules in solution. Journées Scientifiques de l'UMR7177, Institut de Chimie, Strasbourg, 2020. (Oral communication)

Gilberto Pereira & Marco Cecchini. A multi-basin Quasi-harmonic approach for the calculation of the Configurational Entropy of small molecules in solution. Lecture for the Post-Graduation program in Computer Modelling of the Federal University of Juiz de Fora, Brazil, 2021. (Lecture)

Table of Contents

Dedication.....	II
Acknowledgements.....	III
Abstract.....	V
Résumé de Thèse	VI
List of Publications.....	XXVII
List of Talks and Presentations	XXVIII
Table of Contents	XXIX
List of Abbreviations.....	XXXIV
List of Figures.....	XXXVII
List of Tables	XXXIX
1. Drug Discovery.....	2
1.1: Introduction.....	3
1.2: What makes a molecule a drug? The concept of drug-likeness.....	6
1.3: The drug discovery pipeline	10
1.3.1: Target identification and validation.....	11
1.3.2: Hit identification phase	12
1.3.3: Lead optimization phase	12
1.3.4: Pre-clinical tests.....	13
1.3.5: Clinical studies	14
1.4: Affinity and activity concepts in protein-ligand binding	14
1.4.1: The difference between ligand activity and ligand binding affinity	15
1.4.2: The Michaelis-Menten model.....	17
1.4.3: The Cheng-Prusoff equation	17
1.5: Experimental determination of ligand binding affinities	18
1.5.1: Isothermal Titration Calorimetry	19
2. Computer Aided-Drug Design.....	22
2.1: Computer-Aided Drug Design methodologies	22
2.2: Molecular Recognition.....	24

2.2.1: Electrostatic Interactions	25
2.2.2: The hydrophobic effect	26
2.3: Structure-Based Drug Discovery - Virtual Screening.....	27
2.3.1: Protein preparation for SBVS	28
2.3.2: Additional steps in protein preparation	33
2.3.3: Binding site identification.....	35
2.3.4: Ligand library preparation.....	36
2.3.5: Molecular Docking.....	38
2.3.6: Scoring of docking poses.....	41
3. Binding Free Energy calculations.....	47
3.1: Introduction.....	47
3.2: Statistical Mechanics definition of the protein-ligand binding affinity	49
3.3: Molecular Dynamics simulations	52
3.3.1: Molecular Mechanics	52
3.3.2: Integrating Newton's Equations of motion.....	55
3.3.3: Thermostat and Barostat for MD simulations	57
3.4: Rigorous free energy methods.....	57
3.4.1: Free Energy Perturbation – Double Decoupling	58
3.5: End-point methods for binding free energy calculations	62
3.5.1: The Linear Interaction Energy method.....	62
3.5.2: The Linear Interaction Energy with Continuum Electrostatics method	64
3.5.3: The Molecular Mechanics Poisson Boltzmann Surface Area method.....	65
3.6 – Recent applications of MM/PB(GB)SA calculations	74
4. Theory of single molecule entropy methods with applications in MM/PB(GB)SA calculations	77
4.1: Introduction.....	77
4.2: A statistical mechanics view of entropy.....	78
4.3: The Harmonic oscillator.....	81
4.3.1: The classical harmonic oscillator	81
4.3.2: The quantum-mechanical harmonic oscillator	83

4.4: Rigid-rotor Harmonic Oscillator-based entropies in flexible molecules	84
4.5: Configurational entropy within the Rigid-Rotor Harmonic Oscillator approximation	86
4.5.1: Normal Mode Analysis	86
4.5.2: Quasi-Harmonic Analysis.....	89
4.6: Going beyond the Rigid-Rotor Harmonic-Oscillator approximation: The Mutual Information Expansion method.....	95
4.7: Merging Rigid Rotor Entropies and non-parametric estimates of conformational entropy: the CENCALC approach to Mutual Information Expansion.....	98
4.8: Benchmarking entropy calculations.....	100
5. A multi-basin quasi-harmonic approach for the calculation of the configurational entropy of small molecules in solution	104
5.1: Introduction.....	104
5.2: Material and Methods	107
5.2.1: Benchmark datasets	107
5.2.2: MD simulations	108
5.2.3: Entropy calculations	109
5.2.4: MM/GBSA binding free energy calculations	110
5.2.5: Clustering algorithms explored.....	111
5.3: Results and Discussion.....	112
5.3.1: Absolute entropy calculations in vacuum for cyclohexanone.....	112
5.3.2: Calculation of absolute entropies in vacuum.....	114
5.3.3: An automatic procedure for QHMB absolute entropy calculations	116
5.3.4: Calculation of absolute entropies in solution	117
5.3.5: Application to binding free energy calculations	118
5.4: Conclusions.....	121
6. Identification of hits on the quest for allosteric modulators of the myosin molecular motor.....	123
6.1: Introduction.....	123
6.1.1: The myosin molecular motor cycle.....	123
6.1.2: Myosin as pharmacological target.....	125

6.1.3: Smooth muscle myosin II – Pharmacological relevance.....	126
6.1.4: Project goal	126
6.2: Methodology: Virtual Screening protocol	127
6.2.1: Protein preparation	127
6.2.2: Ligand library preparation	128
6.2.3: Molecular docking	130
6.2.4: Molecular Dynamics simulations	130
6.2.5: Binding Free Energy calculations	132
6.2.6: Experimental ATPase activity inhibition assay	132
6.3: Results and Discussion.....	133
6.3.1: Structural analysis of the CK-571 binding pocket	133
6.3.2: Results from a previous VS campaign on SMM2	135
6.3.3: Calibration of MM/GBSA setup for the VS campaign on SMM2	138
6.3.4: Virtual Screening campaign.....	141
6.3.5: Structural analysis of the predicted hit compounds.....	149
6.4: Conclusions.....	150
7. Concluding Remarks.....	152
Bibliographic References.....	156
Supplementary Information: Chapter 5	189
Annex S5.1: Dataset for the gas-phase entropy calculations.	189
Annex S5.2: Dataset for the protein-ligand binding free energy calculations.	193
Annex S5.3: Absolute entropy calculations by QHMB in the gas phase and aqueous solution.	195
Annex S5.4: Binding free-energy results from MM/GBSA calculations.....	198
Annex S5.5: Time series of the heavy-atom RMSD of the ligand from its crystallographic binding mode in 21 protein-ligand complexes extracted from the Greenidge dataset.	200
Annex S5.6: Molecular structures of the compounds from the Greenidge dataset.	202
Supplementary Information: Chapter 6	203

Annex S6.1: MM/GBSA Binding Free Energy results for the 60 compounds that were prioritized from CN.....	203
Annex S6.2: Physicochemical properties of the 60 compounds that were prioritized from CN.....	207
Annex S6.3: Distribution plots of the physicochemical properties of the 2300 compounds from CN selected for VS.	212

List of Abbreviations

ACO - Ant Colony Optimization
ADMET - Absorption, Distribution, Metabolization, Excretion and Toxicity
ADS - Asymmetric Double Sigmoidal
AMIE - Approximated Mutual Information Expansion
APBS - Adaptive Poisson Boltzmann Solver
APR - Attach-Pull-Release
AUC - Area Under the Curve
BACE - β -Secretase
BAT - Bond Angle and Torsion
BBB - Brain-Blood Barrier
BEERT - Binding Entropy Estimation of Rotation and Translation Entropy
BeF₃ - Beryllium Trifluoride
BO - Born Oppenheimer
CADD - Computer-Aided Drug Design
CCCBDB - Computational Chemistry Comparison and Benchmark DataBase
CFA - Coulomb Field Approximation
CN - Chimie Nationale du CNRS
CREST - Conformer-Rotamer Ensemble Sampling Tool
CT - Charge Transfer
CV - Collective Variable
Da - Dalton
DDM - Double Decoupling Method
DFT - Density Functional Theory
DUD - Directory of Useful Decoys
EC₅₀ - Compound concentration that enhances the target activity by 50%
EF - Enrichment Factor
EMA - European Medicine Agency
EP - Endothiapepsin
ESP - Electrostatic Potential
FDA - Food and Drug Administration
FEP - Free Energy Perturbation
FNR - False Negative Rate
FPR - False Positive Rate
GB - Generalized Born

GK - Guanylate Kinase
GPCR - G-coupled Protein Receptor
HCM - Hypertrophic Cardiomyopathy
HCT - Hawkins, Cranmer and Thrular
HO - Harmonic Oscillator
HPC - High Performance Computing
HSP-90 - Heat Shock Protein 90
HTS - High Throughput Screening
IC50 - Compound concentration that decreases the target activity by 50%
ITC - Isothermal Titration Calorimetry
LBDD - Ligand-based Drug Discovery
LDH - Lactate Dehydrogenase
LIE - Linear Interaction Energy
LIECE - Linear Interaction Energy with Continuum Electrostatics
M2 - Mining Minima 2
MB - Multiple Basins
MC - Monte Carlo
MD - Molecular Dynamics
MIE - Mutual Information Expansion
MIST - Maximum Information Spanning Tree
MM - Molecular Mechanics
MM/GBSA - Molecular Mechanics - Generalized Born Surface Area
MM/PBSA - Molecular Mechanics - Poisson Boltzmann Surface Area
MUE - Mean Unsigned Error
MW - Molecular Weight
NIST - National Institute of Standards and Technology
NMA - Normal Mode Analysis
NME - New Molecular Entity
OBC - Onufriev, Bashford and Case
PAINS - Pan Assay Interference Compounds
PB - Poisson Boltzmann
PBP2a - Penicillin-binding Protein 2a
PCIP - Pentachloropseudilin
PCM - Polarizable Continuum Model
PDB - Protein Data Bank
PEP - Phosphoenol Pyruvate
PES - Potential Energy Surface
Pi - Inorganic Phosphate

PK - Pyruvate Kinase
PLP - Piecewise Linear Potential
PME - Particle Mesh Ewald
PMF - Potential of Mean Force
PPS - Pre-Powerstroke State
PSA - Polar Surface Area
PSD-95 - Post-Synaptic Density Protein 95 (PSD-95)
PTP1B - Tyrosine phosphatase-1B
QED - Quantitative Estimation of Desirability
QHA - Quasi-Harmonic Analysis
QHMB - Quasi-Harmonic Multi Basin
QM - Quantum Mechanical
QSAR - Quantitative Structure Activity Relationship
REMD - Replica Exchange Molecular Dynamics
RESP - Restrained Electrostatic Potential
RMSD - Root Mean Squared Deviation
RMSE - Root Mean Squared Error
R-o-3 - Rule of Three
R-o-5 - Rule of Five
RR - Rigid Rotor
RRHO - Rigid Rotor Harmonic Oscillator
SA - Surface Area
SAS - Solvent Accessible Surface
SASA - Solvent Accessible Surface Area
SBDD - Structure-based Drug Discovery
SBVS - Structure-based Virtual Screening
SES - Solvent Excluded Surface
SMM2 - Smooth Muscle Myosin II

List of Figures

Figure 1.1 – Some landmark compounds in the drug discovery field.....	5
Figure 1.2 – Example of a drug-likeness analysis carried out in the SwissADME server for acetylsalicylic acid (Aspirin).....	10
Figure 1.3 – Schematic representation of the drug discovery pipeline.....	11
Figure 1.4 – Example of a dose-response curve fitted to the Hill equation	16
Figure 1.5 – Schematic representation of the ITC experimental equipment	20
Figure 2.1 – Schematic representation of the Computer-Aided Drug Discovery workflow typically employed in the early phases of Drug Discovery campaigns	24
Figure 2.2 – General workflow for employed in SBVS campaigns	27
Figure 3.1 – Distribution of methods for binding free energy calculations according to calculation efficiency and accuracy	47
Figure 3.2 – Thermodynamic cycle for a DDM calculation.....	61
Figure 3.3 – Thermodynamic cycle for a binding free energy calculation of a protein-ligand complex.....	67
Figure 3.4 – Illustration of the two surface definitions	69
Figure 4.1 – Comparison between two harmonic oscillators.....	83
Figure 5.1 – Proof-of-concept approach for QHMB using cyclohexanone as a test case	113
Figure 5.2 - Correlation of computed versus experimental absolute standard entropies for 23 small molecules in the gas phase	115
Figure 5.3 – Accuracy of QHMB as a function of the clustering algorithm and the RMSD cutoff	117
Figure 5.4 – Absolute configurational entropies in solution	118
Figure 5.5 – Correlation of experimental versus predicted binding affinities by MMGBSA with (red) and without (empty points) entropy correction by QHMB.	120
Figure 6.1 – Illustration of the myosin molecular motor cycle.	124
Figure 6.2 – Crystallographic structure of the SMM2 motor domain.....	128
Figure 6.3 - Schematic representation of the VS campaign workflow carried out on CN towards the discovery of new SMM2 modulators.	129
Figure 6.4 - Three-dimensional view of the SMM2/inhibitor systems	134
Figure 6.5 – Predicted binding free energies for the subset of prioritized compounds from the previous VS campaign targeting SMM2.....	136
Figure 6.6 – Experimental measurements of ATPase activity.	137

Figure 6.7 – Molecular structures of ligands with the best predicted binding affinities from the VS campaign of 2018, obtained using MarvinSketch from ChemAxon	138
Figure 6.8 - Results arising from the SMM2-ligand complex MM/GBSA calculations using different dielectric constants and the GB model 7 ⁴² on the benchmark dataset.....	140
Figure 6.9 – Binding free energies obtained for the top 20 compounds of the CN at various steps of the screening campaign on the Chimiotèque National.....	142
Figure 6.10 - Virtual screening results of the campaign towards smooth muscle myosin II (SMM2)	143
Figure 6.11 – Results for the SMM2 ATPase inhibition assay carried out at Institut Curie	145
Figure 6.12 – Binding mode of some of the hit compounds found by virtual screening of the Chimiotèque National du CNRS.....	148

List of Tables

Table 1.1 – Different rule-based definitions of drug-likeness and lead-likeness according to Lipinski, Ghose, Veber, Muegge, Oprea <i>et al.</i> , Lovering and the QED paradigm	8
Table 5.1 - Accuracy of various RRHO methods for the calculation of absolute molecular entropies	115
Table 5.2 – Inclusion of the ligand entropy loss in MMGBSA calculations.....	120
Table 6.1 - Benchmark calculations carried out on the dataset of prioritized compounds from the previous screening	139
Table 6.2. 2D chemical structures and % of inhibition of the identified hits by the Houdusse group	146

Part I – Introduction

1. Drug Discovery

The accurate determination of protein-ligand binding affinities by experimental means is possible, although costly and time consuming.^{81,82,380} It requires a significant amount of protein, which can hamper application to proteins which are difficult to purify.^{81,82,380} To decrease costs at the hit identification stage, which would be large if an experimental determination of the binding affinity per ligand would be carried out for thousands of compounds, many research projects employ computational methods.^{27,84} In particular, molecular docking experiments are common approaches to filter out compounds and retrieving a promising subset of ligands for experimental testing.^{26,27,89} Due to the limitations of docking scoring functions, it is often the case that many false positives are found within prioritized subsets of compounds. To decrease the false positive rate, ligand docking poses of interest can be selected and rescored using higher-level of theory binding free energy calculation methods.^{189,201,213} Several methods can be applied at the free energy rescoring stage, ranging from accurate and rigorous calculations, which simulate the path connecting the bound and unbound states, to approximate end-point methods, which focus on estimating the binding free energy by computing the free energy contribution of each integrant of the binding reaction.^{189,201,213}

This dissertation focuses on the accurate calculation of ligand-binding affinities from numerical simulations. In particular, the main goal was the discovery of new allosteric inhibitors of Smooth Muscle Myosin II (SMM2), which is a myosin molecular motor with implications in several pathologies, by computational means. The project aimed at identifying these inhibitors by refining molecular docking predictions with a free energy rescoring step for a prioritized subset of small molecule ligands. To do so, the top ranked compounds arising from molecular docking were rescored using an end-point binding free energy approach coupled to a new entropy calculation method which accounts for the ligand configurational entropy loss upon binding. Following two chapters of introduction (**Chapter 1** and **2**), the theoretical basis of binding free energy calculations and of entropy calculations will be explored (**Chapter 3** and **4**). Then, the Quasi-Harmonic Multi-basin entropy calculation method will be presented, highlighting its accuracy in reproducing experimental gas-phase entropies and in improving the results of end-point binding free energy calculations (**Chapter 5**). Finally, a VS campaign in search of new and potent inhibitors of SMM2 function will be described, employing a multi-layer ligand library filtering and prioritization scheme to narrow down the Chimiotèque National du CNRS to a subset of 26 compounds selected for experimental testing (**Chapter 6**).

1.1: Introduction

From ancient times, humanity has tried to treat illnesses and pathological conditions. The most common therapeutic route in antiquity was through usage of plant extracts.²⁻⁴ Some of the earliest records on medicinal applications of plants and their extracts date back to 2600 BC, where close to 1000 plants were cataloged according to their therapeutic effects.⁴ Early records in Egypt go back to the year 2900 BC. From these, the “Ebers Papyrus” from *circa* 1600 BC, containing about 700 drugs, are one of the best preserved ancient compilation of drugs.⁵ Another example is Chinese Traditional Medicine, whose records date many millennia and are still being used today not only in China, but all around the world.^{4,5} During the period of the Roman empire, Pliny the Elder, a Roman author, army officer and naturalist, wrote the “*Naturalis Historia*” and created the first pharmacopoeia, while Pedanius Dioscorides wrote “*De Materia Medica*”, a highly influential manuscript on drugs and their application which remained a reference up until the 15th century.^{4,6} In the first millennium, the center of medical and pharmaceutical study moved to the Arab World. Critical contributions were given by Abu Bakr Al-Razi, in Bagdad, where he suggested to evaluate the safety of treatments in animals before human administration.⁷ In the 16th century, the Swiss physician Paracelsus defined the concept of dose in his Third Defense, stating that it was the dose of the chemical entity that determined whether its effect was harmful or not.⁸ Paracelsus also advocated for the usage of pure compounds as medicine instead of administrating herbal mixtures.⁷ His contributions were fundamental towards modern-day medical practice. In the 17th century, while on board of the HMS Salisbury, James Lind conducted the first ever recorded clinical trial, to determine whether citrus fruits were able to treat scurvy-afflicted sailors.⁹

Even though many modern analytical methods were not available before, it was understood that something inside herbal extracts and plants had exploitable therapeutic properties. As the field of Chemistry as a whole matured, analytical techniques were developed allowing the isolation of bioactive compounds from these extracts.^{4,10,11} An example was the discovery of morphine, a potent analgesic isolated from opium extracts in 1805 by a German apothecary assistant called Friedrich Sertürner.¹⁰⁻¹² The compound was named by taking inspiration from the god of sleep Morpheus.¹⁰⁻¹² Sertürners’ finding instigated other apothecaries at the time to examine medicinal plants and herbs, which led to successes in the isolation of bioactive natural products such as alkaloids.^{4,11} One example is Emanuel Merck in Darmstadt, Germany, who would become the progenitor of the pharmaceutical company named after him.

In the 19th century, the field of drug discovery saw tremendous progress, brought by the ability to isolate compounds from natural sources. However, since production yields

from natural extracts were suboptimal, the field turned towards finding synthetic routes to produce pure compounds.⁷ By re-using coal tar, waste from the gas industry, William Perkin was able to synthesize the first dye, mauveine, in 1856.⁷ With the growth of the chemical industry, in time chemists were able to synthesize an array of compounds which deviated significantly from dyes.⁷ A landmark achievement at the time was the development of Aspirin in Germany by Bayer and Company, starting from extracts of willow bark in 1897.^{7,387,388} The salicylates found in willow bark extracts were corrosive and caused stomach irritation but the extract itself had antipyretic properties, thus motivating the company into producing a compound with antipyretic and analgesic properties, named acetylsalicylic acid, later known as Aspirin.^{7,387,388}

While there are several success cases, discovery of drugs in the past was mostly due to serendipity^{4,10,11}, as in the discovery of penicillin by Alexander Fleming in 1928, which later on proved to be fundamental in tackling sepsis during World War II.^{4,13} The discovery of penicillin led the field forward, as researchers started looking at microbial organisms as potential sources of new drugs.^{4,11} As the years progressed, derivatives of penicillin were produced, such as cephalosporins.^{4,14} From the 1980s onward, screening methodologies and drug design approaches were proposed and validated. The lack of powerful analysis and screening tools left much of the success of drug design campaigns dependent on chemical intuition and luck, which motivated further research.^{11,15} It was established successful drug design required an open dialogue between chemists and biologists, regarding biochemical mechanisms of action of compounds with potential to become drugs.¹¹ Further, it was recognized that the connection between structure and function was key to create novel chemical modulators.¹¹ With the development of high-throughput screening (HTS) platforms, the evaluation of the activity of several thousand compounds towards a given target in a fast, low-cost and automated manner was made possible.^{16,17} Chemical libraries could now be screened in pursuit of new modulators of receptors with pharmacological and medical relevance, at a reduced cost and in a time-efficient manner, while integrating chemical and biological information from experts.^{16,17} Examples of drugs which were discovered following HTS experiments include drugs for cancer therapy such as Alectinib, an orally available small molecule which blocks the activity of anaplastic lymphoma kinase^{18,19} or Olaparib, a small-molecule approved for ovarian, breast or prostate cancer therapy.²⁰ Other small-molecule drugs approved following HTS include Maraviroc, an antiviral cytokine inhibitor for HIV therapy²¹ or Ceralifimod as a therapeutic avenue for Multiple Sclerosis.²² In **Figure 1.1**, some of the compounds representing landmarks in the drug discovery field are shown. Recently, HTS has also proven to be useful in the combat of the Covid-19 pandemic. Following HTS of a library of FDA-approved drugs, the group of Zhang *et al.* found several small-molecule hits against SARS-Covid-19 and arenaviruses.²³ In particular, mycophenolic acid,

Clofazimine, Dabrafenib, and Apatinib were found to significantly inhibit SARS-Covid-19 infection with an IC_{50} in the low micromolar range.²³ Despite this, the number of new Molecular Entities (NME) launched in the drug market has seen sharp decline in the last years.²⁴

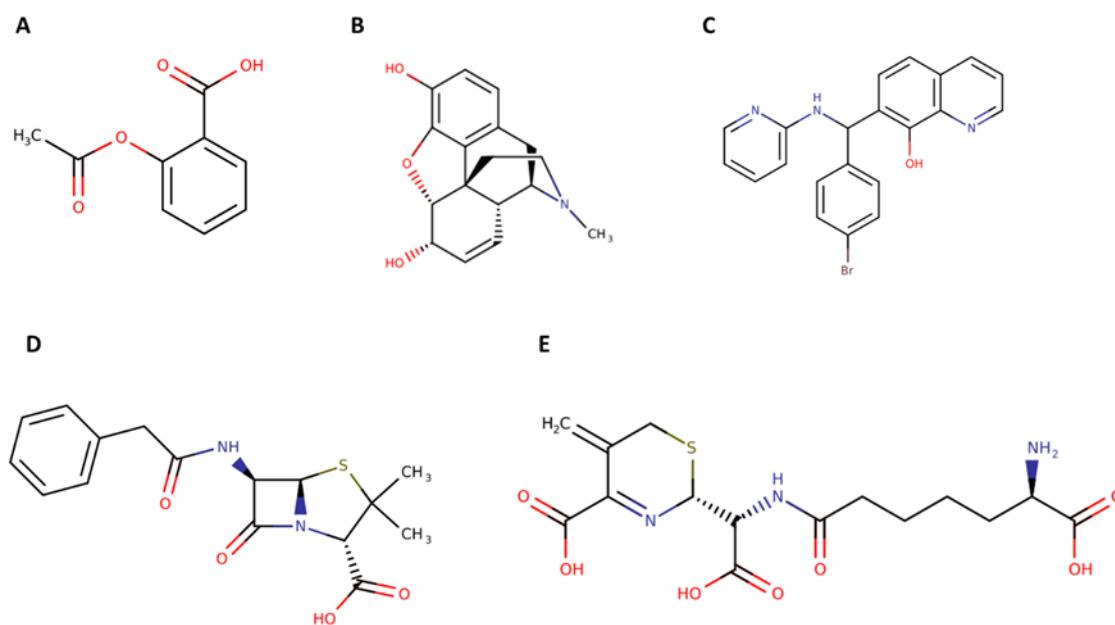


Figure 1.1 – Some landmark compounds in the drug discovery field. A) Molecular structure of Aspirin; B) Molecular structure of Morphine; C) Molecular structure of Mauveine; D) Molecular structure of Penicillin G; E) Molecular structure of Cephalosporin C. Molecular structures were drawn using MarvinSketch from ChemAxon.²⁵

Another significant advance towards lowering resource consumption in drug discovery campaigns was the incorporation of molecular modelling and numerical simulations.^{26,27,29} The inclusion of *in silico* approaches was possible due to the increase in computer power and the development and validation of computational models aiming to accurately describe the nature of interactions between biological entities.^{15,26–28} Thus, the medicinal chemist's knowledge could be combined with molecular modelling and HTS to accelerate the drug discovery pipeline. This new field was named Computer-Aided Drug Discovery (CADD) and hosts a wide array of tools and strategies which can facilitate and accelerate the discovery of new therapeutic agents.^{26,27,29} Methodologies comprised in CADD have played a vital role in the discovery of currently approved and used drugs.²⁷ These methodologies include molecular modelling, quantitative structure-activity relationships (QSAR), virtual screening approaches, molecular docking and machine learning methods, among others.^{27,29–31}

1.2: What makes a molecule a drug? The concept of drug-likeness

A drug is any substance which can elicit a physiological or psychological response in an organism after administration.³² Currently approved drugs spawn a rich and diverse chemical space, going from small molecules, like Apatinib or Penicilin, to peptides or even larger compounds like Mipomersen.³³ Given the large differences in physicochemical profiles among drugs, these entities have different administration methods. Drug administration methods include but are not limited to inhalation, ingestion, injection and dissolution. One of the most well-known online drug repositories is the DrugBank³⁴, which includes 2683 small-molecules approved by the Food and Drugs Administration (FDA), the European Medicine Agency (EMA) and the Canadian agencies as of March 1st 2021. Additionally, DrugBank contains 1463 biologics, 131 nutraceuticals and 6654 experimental compounds. Another useful resource is DrugCentral³⁵, designed at the University of New Mexico. DrugCentral regularly updates its library by monitoring the FDA, EMA and the Pharmaceuticals and Medical Devices Agency in search of newly approved drugs. For each compound, DrugCentral provides information on the active ingredients, drug mode of action, pharmacological action, physicochemical properties, among other properties, extracted from expert-curated resources such as ChemBL³⁶ or KEGG.^{35,37} Focusing on small molecules, which are key to this dissertation, it is important to question what kind of physicochemical properties a molecule should exhibit to be considered a drug. In other words, it is important to define the concept of drug-likeness.³⁸ Defining this concept allows researchers to filter out of their chemical libraries compounds which do not satisfy the requirements, thus saving resources and prioritizing molecules with a more promising profile.

A fundamental contribution to the definition of drug-likeness was the seminal paper of Christopher Lipinski in 1997.³⁹ In his contribution, Lipinski outlined a set of rules, later known as Lipinski's rule-of-five (R-o-5), which were derived from orally bioavailable and human administration-approved drugs at the time.^{29,38,39} The Lipinski R-o-5 states that a compound is likely orally bioavailable if: (1) it has a molecular weight (MW) below 500 Daltons (Da), (2) it has less than 5 hydrogen-bond donor groups, (3) it has less than 10 hydrogen bond acceptor groups, (4) its octanol/water partition coefficient (logP) is less than 5.^{38,39} The connection between drug-likeness and oral bioavailability was then established. At the time, this rule of thumb analysis was a landmark contribution that allowed researchers to efficiently evaluate the potential of chemical libraries and prioritize compounds with the potential to become drugs approved by the FDA. However, the R-o-5 also has limitations.³⁸ It relies on the assumption of passive transport (not considering the existence of transporter proteins), only 50% of currently approved small-molecule drugs comply to the R-o-5 and there are many drugs which do not require oral

bioavailability and thus also do not respect it.³⁸ Other researchers, like Ghose⁴⁰, Veber⁴¹, Egan⁴² and Muegge⁴³, suggested their own drug-likeness evaluation schemes either based on the Lipinski R-o-5 or based on a pharmacophore point definition. More recently, other researchers suggested a new interpretation of drug-likeness based on desirability functions.⁴⁴ In this definition of drug-likeness, the contribution of key ligand molecular properties are combined to compute a desirability score in the scale of 0 to 1, encoding the Quantitative Estimation of Desirability (QED) per compound, by means of double asymmetric sigmoidal functions which were calibrated on a general ligand benchmark dataset.⁴⁴ In this particular study, the some of the queried properties were the number of rotatable bonds, the molecule the number of aromatic rings, the number of hydrogen bond donors and acceptors, the molecular weight and the octanol-water partition coefficient.⁴⁴ A more recent development in the computation of compound desirability was produced by Akyiama and co-workers, where instead of generating a set of general functions whose weights are optimized on ligand datasets targeting several proteins, the researchers opted by developing one QED function per target.⁴⁵ This allows the development of protein-specific ligand filters to prune chemical datasets according to the physicochemical properties of known binders. Further exploration in the field led the group of Lovering^{46,47} to identify that the fraction of sp³ carbons (Fsp³) present in the molecular structure of the ligand could be a useful descriptor to estimate compound success in the drug discovery pipeline.^{48,49} Lovering was able to determine that the fraction sp³ carbon in molecules increased between phases in the pipeline from the initial hit identification step and throughout clinical trials.^{47,48} It was hypothesized that it could be either related to molecular solubility or the fact that the larger ligands were occupying better the binding pocket.^{47,48} A detailed description of the rule-based systems proposed over the years to define drug-likeness is shown in **Table 1.1**.

A contrasting view to the usage of the R-o-5 as a filter for hit selection was put forth by Congreve *et al.*⁵⁰ and further deepened in the work of Tudor Oprea and colleagues^{38,51}, where it is discussed that the R-o-5, while useful, was derived from analyzing drugs, not hit compounds.⁵¹ By employing the R-o-5 in HTS context, it is highly likely that many of the active compounds found correspond to drug-like compounds barely within the ranges defined by the ruleset. Further, in HTS experiments, the activity cut-off appears to be arbitrarily selected, or set from a known active, and the hits retrieved are typically micromolar.⁵² As pointed out in a review, many studies use low to mid-micromolar thresholds, ranging from 1 to 500 μ M.⁵² Further optimization of drug-like compounds, which is a necessary requirement to achieve nanomolar activity compounds, would probably place them outside the of R-o-5 space. As such, stricter rules were designed with the aim of prioritizing smaller and more optimizable compounds. The suggested rules stated that selected compounds should be active hits exhibiting: (1) a MW equal or

below 300 Da, (2) at most 3 rotatable bonds, (3) a logP below 3, (4) at most three hydrogen bond donor groups and (5) at most three hydrogen bond acceptor groups. This rule was termed the rule-of-three (R-o-3)^{35,38,50} and these compounds were named lead-like compounds.³⁸ As such, lead-like hit compounds are significantly smaller, more soluble, less hydrophobic, less complex and less flexible when compared to drug-like compounds.^{38,51}

It is then apparent that to increase the success of drug discovery campaigns, special attention must be given to the filtering steps such that the processed library contains only compounds of high value for the screening campaign and thus minimizes time and resource consumption. Although arguments exist towards enforcing of the R-o-3 in drug discovery projects, the Lipinski R-o-5 remains popular in the drug discovery community, in part for historical reasons.

Table 1.1 – Different rule-based definitions of drug-likeness and lead-likeness according to Lipinski, Ghose, Veber, Muegge, Oprea *et al.*, Lovering and the QED paradigm.

Rule Name	Publication Year	Type of definition	Features
Lipinski's Rule-of-Five	1997	Rule-based	MW ≤ 500 Da; HBA ≤ 10; HBD ≤ 5; logP ≤ 5
Ghose Filter	1998	Rule-based	-0.4 ≤ logP ≤ 5.6; 40 ≤ MR ≤ 130; 180 ≤ MW ≤ 480; 20 ≤ NAT ≤ 70
Veber Filter	2002	Rule-based	NRot ≤ 10; PSA ≤ 140 Å ²
Muegge Filter	2001	Rule-based	200 ≤ MW ≤ 600; -2 ≤ logP ≤ 5; TPSA ≤ 150; NRing ≤ 7; NC > 4; NHA > 1; NRot ≤ 15; HBD ≤ 5; HBA ≤ 10
Egan Filter	2000	Rule-based	logP ≤ 5.88; TPSA ≤ 131.6

Quantitative Estimation of Desirability (QED)	2012	Multivariable optimization	MW, logP, HBA, HBD, NRing, NRot, PSA, ALERTS
Target-specific QED	2019	Multivariable optimization	MW, logP, HBA, HBD, NRing, NRot, PSA, ALERTS
Molecular Complexity	2009, 2012	Rule-based	$0.36 \leq \text{Average Fsp3} \leq 0.46$; $\text{NSte} \geq 1$
Lead-likeness Rule-of-Three	2003	Rule-based	$\text{MW} \leq 300 \text{ Da}$; $\log P \leq 3$; $\text{NRot} \leq 3$; $\text{HBD} \leq 3$; $\text{HBA} \leq 3$

MW: Molecular Weight; logP: octanol-water partition coefficient; HBD: Hydrogen bond donors; HBA: Hydrogen bond acceptor; NRot: Number of rotatable bonds; NRing: Number of rings; Average Fsp3: Average fraction of sp³ carbons; NAt: Number of atoms; PSA: Polar Surface Area; NSte: Number of stereocenters; MR: Molecular Refractivity; NC: Number of carbons; NHA: Number of Heavy Atoms; TPSA: Total Polar Surface Area; ALERTS: Number of structural alerts according to Brenk *et al.*⁵³

Rules describing drug-likeness have been applied in drug discovery campaigns ever since the early 2000s, with significant impact in the development of many drugs. To aid the filtering and analysis of chemical libraries, tools were developed with the aim of computing molecular properties and numerically evaluate the drug-likeness of molecules.

One such tool was developed by the Swiss Institute of Bioinformatics, SwissADME⁵⁴, combining drug-likeness analysis and pharmacological profiling with a user-friendly graphical interface.⁵⁴ In this web-based server, calculation of the physicochemical properties of molecular structures is straightforward and a thorough analysis is carried out and supplied in the form of a visual report, as shown in **Figure 1.2**, and in a .csv format for further analysis. As is shown in **Figure 1.2**, Aspirin fulfills all the requirements for druglikeness comprised in the R-o-5, Egan, Ghose and Veber filters. It does not, however, satisfy the Muegge filtering scheme. In particular, the molecular weight of Aspirin (180.16 Da) falls outside of the range recommended by Muegge ($200 \leq \text{MW} \leq 600$). As such, it is important to question whether rigid application of these filters is reasonable or not. Most researchers, when faced with this situation, will opt by allowing at most the violation of one rule in the filtering scheme, as recommended³⁹ in the work of Lipinski. This was the approach followed in the work presented in **Chapter 6**, where the application of a VS workflow, with ligand library filtering steps comprised within, led to the identification of innovative Smooth Muscle Myosin II (SMM2) allosteric inhibitors. Other tools with the

ability to carry out these analysis include DataWarrior¹ and the tools developed by molinspiration⁵⁵. DataWarrior has the added benefit of allowing the calculation of pharmacophore descriptors for small-molecule ligands, which can then be used for ligand library clustering based on pharmacophore similarity.¹ These pharmacophores can then be used for further library pruning by selecting representative ligands within each molecular cluster.¹ Doing so allows to explore the full chemical diversity of the library at a reduced cost. The reader is encouraged to explore the Swiss Institute of Bioinformatics Click2Drug⁵⁶ web repository for more computational tools with potential applications in drug discovery projects.

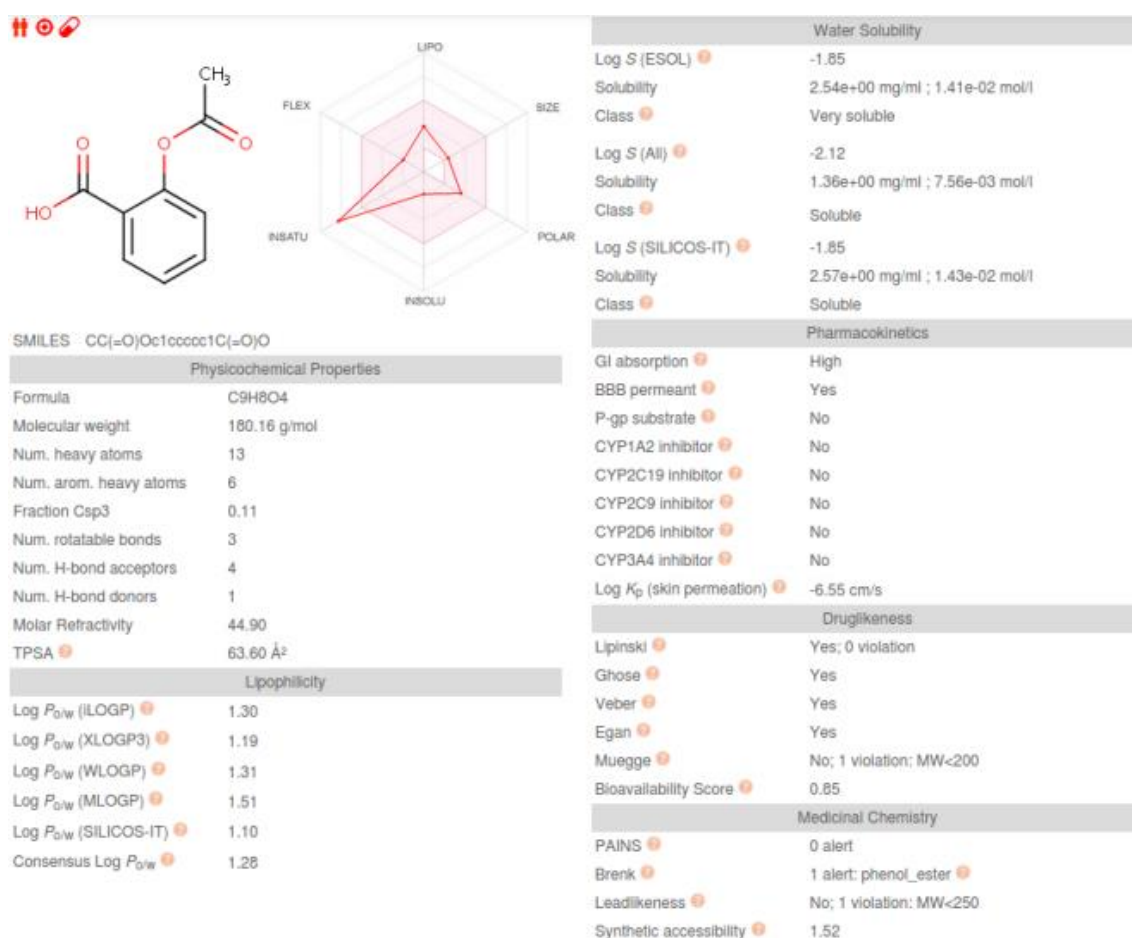


Figure 1.2 – Example of a drug-likeness analysis carried out in the SwissADME server for acetylsalicylic acid (Aspirin). The SwissADME server supplies not only the molecular features but also evaluates the solubility, the druglikeness, the lipophilicity, the lead-likeness and the pharmacokinetic profile of queries compound.

1.3: The drug discovery pipeline

Drug discovery is a long and costly process in which new pharmaceutical compounds are identified, optimized, tested and brought to the market.^{5,27,29,57} A drug discovery program is usually initiated due to the lack of an appropriate therapeutic agent against a

given disease or clinical condition.⁵⁸ It is estimated that until a molecular entity is approved for clinical usage, researchers will have invested around 12 years and on average upwards of US\$2 billion dollars.⁵⁹ However, the cost and time required for the process to be successfully completed depends on a number of factors, including compound safety, synthetic accessibility, potency and intellectual property protection.⁵⁹ A schematic representation of the drug discovery pipeline is shown in **Figure 1.3**. The drug discovery pipeline is comprised of 3 broad stages: target validation and hit discovery, hit development, including pre-clinical and clinical trials, and FDA approval.²⁷ While every year new compounds with drug potential are found, the large majority fail to reach the market. Common reasons for failure in a drug discovery campaign include compound toxicity, undesirable compound side effects, bad ligand pharmacodynamic and pharmacokinetic profiles, low bio-availability, chemical instability, low water solubility and promiscuous binding to other targets.^{27,59}

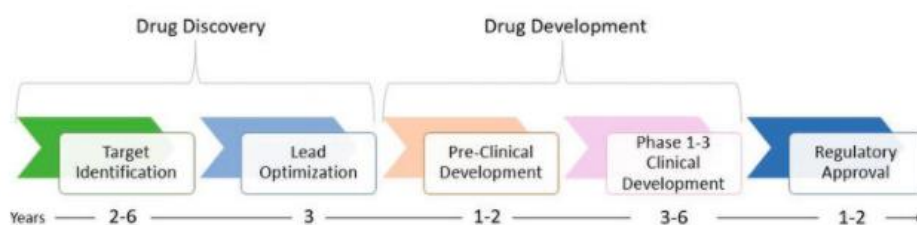


Figure 1.3 – Schematic representation of the drug discovery pipeline. The pipeline contains three main stages: Identification, Development and Approval. Inside the Identification stage, there are two main phases: Target Validation and Hit identification. Inside the Development stage, there are two main phases: Lead optimization and the clinical and preclinical trials. The Approval stage corresponds to the submission and approval by the FDA of the new therapeutic for commercialization. It may also include post-approval phase IV clinical studies. Obtained from “An Overview of AI in Oncology Drug Discovery and Development. In *Artificial Intelligence in Oncology Drug Discovery and Development*”; by Linton-Reid, 2020. Taylor, J. W. C., Taylor, B., Eds.; Copyright @ IntechOpen.⁶⁰

1.3.1: Target identification and validation

All the steps of the drug discovery pipeline present significant challenges for researchers. The selection of a target must obey a series of requirements such as: safety, clinical relevance, efficacy and druggability of the target.⁵⁸ Furthermore, it must be confirmed that perturbing the normal function of the target yields the desired therapeutic. As recently suggested, processing available biomedical data using modern data mining tools appears as a robust approach for target identification and validation.⁵⁸ The validation of a target can be carried out from *in vivo* and *in vitro* studies and while each approach is valid, it is recommended that multiple procedures be combined to maximize confidence.⁵⁸

1.3.2: Hit identification phase

In the hit identification phase, large-scale HTS or virtual HTS campaigns are carried out to screen large chemical libraries in search of small molecules which produce the desired therapeutic effect, be it inhibition or activation, on the desired target.^{29,58,61} The confirmation of the modulatory effect implies the *a priori* existence of a robust and trustable biochemical assay.^{58,389} The development of an assay requires a set of conditions: ability to identify compounds with an appropriate mechanism of action (pharmacological relevance), reproducibility, consideration of the effect of compounds found in the assay (solvents, for example), assay quality and cost.⁵⁸ Within the context of hit identification by HTS, biochemical and cell-based assays have seen routine use.³⁸⁹ Biochemical assays are ascribed to receptor targets or enzymes, being simple and reliable, allowing researchers to evaluate how potent is the compound by evaluating its effect on a given experimental end-point.⁵⁸ Cell-based assays are more complex and can be used to report on compound properties like toxicity.⁵⁸ However, at the hit identification stage it is possible to recover false positives and false negatives due to aggregation and precipitation of compounds, non-specific binding and interference in the assay by interaction between compounds and the constituents of the assay.³⁸⁹ One way to probe for false positives is to use counterscreens, testing the hit compounds against a protein in the same family of the target.³⁸⁹

Typically, many of the recovered hits are not potent enough or exhibit poor physicochemical profiles and toxicity, preventing their direct application as drugs. As such, lead optimization campaigns are carried out to improve both the potency and the physicochemical features of the molecules.^{15,27,29,58,61} Examples of undesirable properties found in these compounds are low water solubility, low lipophilicity or the existence of chemical warhead groups within the molecular structure.^{15,27,29,58,61}

1.3.3: Lead optimization phase

At this stage, both advanced computational techniques and chemical intuition are key ingredients driving the design of derivatives which are potent and whose physicochemical profile is desirable. Often, lead-optimization also focuses on improving the absorption, distribution, metabolism, excretion and toxicity profile (ADMET) of compounds usually through *in vitro* and *in silico* means while not worsening the potency with respect to the lead molecule.^{27,58,59,62} The chemical structure of the identified hits acts as starting points for chemical modification, usually according to the expertise and creativity of the medicinal chemist. While in the past the potency of the compound was a key driver of lead-optimization campaigns, this paradigm shifted and now other properties, like

protein selectivity, solubility, lipophilicity or the compound pharmacokinetics properties, have been raised to the same level of importance as the potency of the hits.³⁸⁹

1.3.4: Pre-clinical tests

The optimized compounds, also termed lead compounds, are then tested *in vitro* or *in vivo*, using animal models, to determine efficacy, safety and ADMET features.⁵⁸ The main goals of this stage are the determination of a safe starting dose for human testing in clinical trials and assessment of the potential toxicity of the compound.^{59,63} Typically, it is at this phase that most potential drugs fail: either due to not being active in the animal models or due to toxicity concerns.^{59,63} It is particularly important to evaluate the toxicological profile of chemicals as early as possible to save both time and resources by avoiding problematic optimization routes or to avoid being stuck in a situation where the compound is not-optimizable.^{27,58,59,63} However, toxicological profiling of compounds is a challenging subject on its own, depending on many factors such as the route of administration, dose and exposure time and the molecular physicochemical properties. During drug development, in particular in the pre-clinical phase, several toxicity endpoints are studied, such as hepatotoxicity, cardiotoxicity, genotoxicity, immunotoxicity and phospholipidosis.^{64,65} Other toxicity endpoints evaluated include carcinogenicity or teratogenicity.^{63,66} If a lead compound still exhibits some undesirable properties after an optimization round, it must be subjected to additional cycles. Importantly, there exist computational methodologies able to predict some of these endpoints based on machine learning models trained on large and diverse chemical datasets containing experimental measurements.^{27,66,67} It is important to note the development of VEGA as a significant contribution to the field of computational toxicity.⁶⁶ This piece of software is a user-friendly tool for *in silico* toxicity evaluation and targets many endpoint, providing a comprehensive analysis of the predicted toxicological profile of candidate compounds in a concise and detailed report and, in a first pass, reducing the number of required animal testing.⁶⁶

Typically, toxicity studies investigate which organs could the compound target as well as the risk of toxicity. Both *in vitro* and *in vivo* tests are required to complete pre-clinical trials and validate *in silico* predictions, through rodent or canine models, although *ex-vivo* testing is also an option.⁵⁸ The goal with the evaluation of the pharmacodynamics profile of the molecule is to know how potent the compound is, the dose at which toxic side-effects appear and the dose which elicits a significant therapeutic effect without significant toxicity arising in the animal models.⁵⁸

1.3.5: Clinical studies

Molecules which pass the pre-clinical phase are then tested in humans. Clinical trials are divided into phases: Phase I, Phase II, Phase III and Phase IV, should the latter be required by the FDA.^{68,390,391,392} In Phase I trials, the objective is to evaluate short-term toxicity and to monitor drug safety in a small, healthy population.^{68,391,392} Additionally, it evaluates the optimal drug administration path, the maximum tolerated dose (MTD) and potential side effects.^{390,392} In Phase II trials, the aim is to evaluate the effectiveness and safety of the lead molecule, while evaluating pharmacokinetics and dynamics. The prototypical drug is administered in larger concentrations to a group of participants exhibiting the disease or condition.^{68,392} During the Phase II trials, one key aspect is the determination of therapeutic dose in humans at which no apparent toxicity is elicited.^{68,390,392} In Phase III trials, the objective is to determine if the new proposed molecule outperforms the ones already available in the market and the incidence of adverse reactions.^{68,392} The group of participants, which is comprised of thousands of individuals, is split into two groups: one is treated with the new compound and one is either treated with the current standard treatment or a placebo.^{68,390} At least two Phase III trials are required for FDA approval.⁷ Finally, Phase IV trials, or post-marketing studies, are performed following FDA approval to identify long-term effects, rare adverse reactions and evaluate drug effectiveness in populations with different features from the original population of the study.^{68,390,392} If a molecule passes the clinical trials, it is eligible to feature a New Drug Application (NDA) submission to the FDA.

1.4: Affinity and activity concepts in protein-ligand binding

Since the 1990s, the usage of HTS in drug discovery pipelines became a cornerstone of pharmaceutical research.^{15,29} A molecule will be considered a hit if, after testing and confirmation assays, it is able to elicit a significant effect on the target in a reproducible manner.^{58,68} During the remaining of this dissertation, the term target will be used as synonymous for protein even if there are drugs targeting DNA, lipid structures and other biomolecules. Known limitations of typical HTS screenings are low hit rates⁵² and the struggle with supplying hits that are easily optimizable into novel therapeutic compounds.³⁸ Furthermore, HTS experiments are typically “hit-or-miss” as compounds are tested at only one concentration⁶⁹ and do not provide a quantitative measure of the activity of the compounds. This is especially important in the context of hit identification, where the aim is to select which molecules should progress towards lead-optimization.⁵⁸ It is fundamental to correctly select the hits to maximize the chances of success in the

subsequent steps of the pipeline.⁵⁸ Following HTS, it is standard to confirm the activity of hits by evaluating the ligand concentration at which 50% of the receptor activity is inhibited or enhanced (IC₅₀ or EC₅₀ determinations, respectively).⁵⁸ These determinations allow the separation between high and low activity compounds and the effective prioritization of the most active hits.

1.4.1: The difference between ligand activity and ligand binding affinity

At equilibrium, the binding reaction of a ligand (L) to a receptor (R), with a 1:1 stoichiometry, is defined by **Equation 1.1**¹⁸⁹ and the equilibrium constant for binding by **Equation 1.2**:



$$K_{\text{eq}} = \frac{[RL]}{[R][L]} = \frac{1}{K_D} = K_A \quad (1.2)$$

where K_{eq} is the equilibrium constant, $[R]$ is the receptor concentration free in solution, $[L]$ is the ligand concentration free in solution, $[RL]$ is the concentration of the receptor-ligand complex, K_A is the association constant and K_D is the dissociation constant, which quantify the ligand binding affinity. Indeed, these constants are effective binding constants, as they depend on the conditions of the assay like pH or ionic strength. The ligand binding affinity is a measure of binding strength, meaning that the smaller the K_D the stronger the affinity with which the ligand binds.³⁸⁰ Looking at **Equation 1.1**, it means that potent binders shift the chemical equilibrium in the direction of complex formation.¹⁸⁹ Following HTS, hit compound activity is quantified in a precise manner through IC₅₀ determinations. The calculation of an IC₅₀ requires construction of a dose-response, or Hill, curve.⁷⁰⁻⁷² The Hill equation is used to describe many non-linear relationships, among which quantitative pharmacology and protein-ligand binding are included.⁷⁰ The Hill equation can be derived from the law of mass action, as in the Michaelis-Menten model, under three main assumptions: (1) Receptors are fully accessible to ligands, (2) protein is either bound to a ligand or free in solution, (3) Ligand-receptor binding is a reversible process.⁷³ During an IC₅₀ determination, the ligand is added at different concentrations and the response that is elicited in the receptor (in this case, protein) is measured. The points are then fit to a Hill equation^{71,72} which can be written in the general form of **Equation 1.3** or in the form typically employed in IC₅₀ determinations (**Equation 1.4**) after some algebra:^{71,72}

$$f_b = \frac{1}{1 + \left(\frac{K_D}{[L]}\right)^n} \quad (1.3)$$

$$\frac{E}{E_{\max}} = \frac{1}{1 + \left(\frac{IC_{50}}{[L]}\right)^n} \quad (1.4)$$

where f_b is the percentage of protein molecules with n ligands bound, $[L]$ the ligand concentration, E is the response and E_{\max} is the maximal response. In the case of activators, the measured end-point is the EC_{50} .^{71,72} An example of a dose-response curve is shown in **Figure 1.4**.

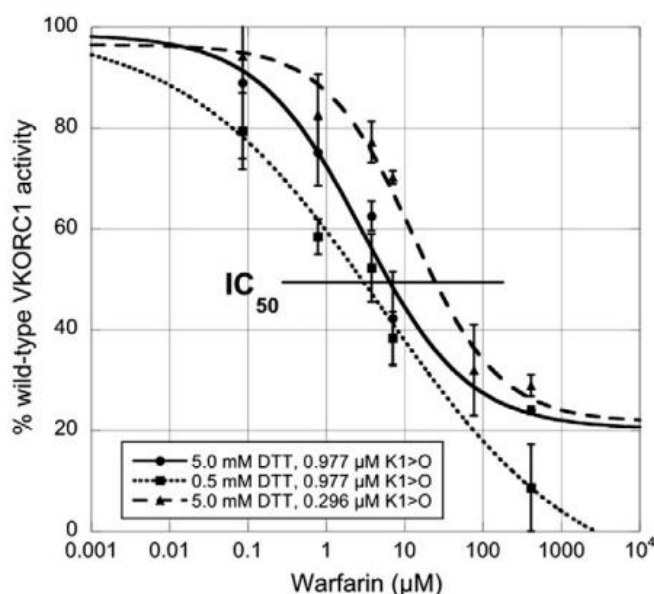


Figure 1.4 – Example of a dose-response curve fitted to the Hill equation. The IC_{50} of warfarin towards vitamin K 2,3-epoxide reductase complex subunit 1 (VKORC1) was determined at different concentrations of vitamin K1 2,3-epoxide in an *in vitro* dithiothreitol (DTT) driven enzymatic assay employing different DTT concentrations. Extracted from “Determination of the warfarin inhibition constant K_i for vitamin K 2,3-epoxide reductase complex subunit 1 (VKORC1) using an *in vitro* DTT-driven assay” by Bevens *et al.*, 2013, *Biochimica et Biophysica Acta – General Subjects*, 1830, 8, 4202-4210.³⁷⁹ Copyright @ 2021, Elsevier B. V.

Although activity and affinity are related, these terms describe different phenomena. While the activity describes the magnitude of the effect elicited upon ligand binding to the target, the ligand-binding affinity (K_D) is related to the strength of the binding reaction, which does not necessarily correlate 1:1 with the observable effect.⁷⁴ Ligands which are highly affine towards a target are not necessarily potent and indeed ligands with the same K_D to a protein can elicit diametrically opposed effects (inhibitors and activators of ion channels serve as example here). Additionally, inhibitors can be partial agonists, where their potency depends on to the percentage of effect they elicit on the

target with respect to the maximal target activity, which does not depend only the affinity with which they bind it.³⁸¹

1.4.2: The Michaelis-Menten model

To determine the rate of a reaction catalyzed by an enzyme as a function of substrate concentration⁷², one typically uses the Michaelis-Menten model (**Equation 1.5**).

$$v = \frac{V_{\max} * [S]}{K_M + [S]} \quad (1.5)$$

Here, v is the reaction rate, V_{\max} corresponds to the maximum rate, K_M is the Michaelis constant and $[S]$ is the substrate concentration.⁷² A low affinity substrate will have a large K_M , meaning that a large $[S]$ is required to elicit 50% of V_{\max} .⁷² To grasp the usefulness of the Michaelis-Menten model, consider two cases: a competitive inhibitor binding assay and a non-competitive inhibitor binding assay. In the former, there is a competition for the active site, meaning that competitive inhibitors will cause a decrease in K_M , as the affinity of the enzyme towards the substrate is reduced because of inhibitor binding. In the first example, the competitive inhibitor prevents the substrate from binding to the enzyme (decrease in K_M), occupying the active binding site and reducing the activity of the enzyme by decreasing its affinity towards the substrate. In the second case, the inhibitor binds to an allosteric site and inhibits protein activity by stabilizing a non-functional conformation without affecting the affinity of the substrate towards the protein (V_{\max} will decrease without affecting K_M).

The Michaelis-Menten model establishes a clear relationship between the activity of an enzyme and substrate concentration.⁷² This model also constituted the basis upon which equations were derived to show the relationship between activity and affinity in a quantitative fashion. In particular, it allows us evaluate the effect an inhibitor has upon enzymatic function.

1.4.3: The Cheng-Prusoff equation

The relationship between activity and affinity was first demonstrated in a quantitative fashion by Cheng and Prusoff in 1973^{75,76}, when they derived **Equation 1.6** to describe the relationship between K_D (or the inhibition constant K_i) and IC_{50} for competitive reversible inhibitors of enzymes as:

$$IC_{50} = K_i \left(1 + \frac{[S]}{K_M} \right) \quad (1.6)$$

In 1975, however, Cha noticed that the original Cheng-Prusoff equation did not take into account tight binding scenarios, and obtained **Equation 1.7**⁷⁷:

$$IC_{50} = K_i \left(1 + \frac{[S]}{K_D} \right) + \frac{E_0}{2} \quad (1.7)$$

where E_0 corresponds to the total concentration of enzyme. These equations will hold as long as the protein target is or behaves like an enzyme.⁷⁷ However, there exist drawbacks to the usage of IC_{50} values as absolute measurements. Comparisons between IC_{50} values are not straightforward.³⁸² As observable from **Equations 1.5 and 1.7**, the IC_{50} determinations depend on the enzyme and substrate concentration, which may change from assay to assay, as well as on the assay type.³⁸² Furthermore, in case of tight binding, the lowest IC_{50} measurable is half of the active enzyme concentration.⁷⁷ Assuming that the assay uses 150 nM of enzyme, picomolar inhibitors would be categorized as low nanomolar inhibitors due to the limitation in the resolution of the measurements. Due to these limitations, comparison of IC_{50} determinations is not straightforward, especially when they arise from different laboratories.³⁸² To circumvent these limitations, one can use the Cheng-Prusoff equation to compute the K_i , which configures a binding affinity, from the determined IC_{50} through **Equations 1.6 or 1.7**. The current gold-standard for binding affinity determinations is Isothermal Titration Calorimetry (ITC)^{81,82,380}, a technique which will be summarized in the following section.

1.5: Experimental determination of ligand binding affinities

The binding of a ligand to a receptor is an equilibrium phenomenon, obeying to thermodynamic laws. The binding free energy of this associative process at the standard state ΔG_{bind}^0 is characterized by thermodynamic observables as shown in **Equation 1.8**.³⁸³

$$\Delta G_{bind}^0 = -RT \ln K_{eq} = \Delta H^\circ - T\Delta S^\circ \quad (1.8)$$

where R is the perfect gas constant, T is the temperature in Kelvin, K_{eq} is the equilibrium constant at the standard state (usually 1 Molar concentration, **Equation 1.2**), ΔH° is the standard enthalpy of binding and $-T\Delta S^\circ$ is the standard entropy of binding. The

dissociative process is described at the same time by plugging **Equation 1.2** into **Equation 1.8**, yielding **Equation 1.9**.³⁸³

$$\Delta G_{bind}^0 = RT \ln K_D \quad (1.9)$$

The two terms that compose ΔG_{bind}^0 , enthalpy and entropy, must be analyzed carefully. In protein-ligand binding, typically the binding enthalpy contribution is negative. A negative enthalpy term means that complex formation implies that the new interactions established between the ligand and the receptor are more favorable than the interactions these species established previously with the solution. A positive enthalpic contribution shows that the interactions established within the protein-ligand complex and between the complex and the solution are not as favorable as those established by the individual species with the solution. However, as complexation occurs there are translational, rotational and vibrational degrees of freedom which become confined in the ligand and receptor, leading to a loss in configurational flexibility and limiting the translational and rotational motions (*i.e.* a loss in entropy) of the unbound state of each species with respect to the bound state characterized by the protein-ligand complex. In other words, as ligand binding occurs, an entropic loss ensues as the conformational space accessible to the bound state of the chemical species is usually smaller than that of the unbound state ($T\Delta S < 0$).⁷⁸ In addition to this entropic penalty, which is essentially a solute entropy contribution, there exist other contributions which can play a major role during binding. One clear example is the contribution of the solvent, such as when the binding reaction requires the displacement of several entropically unfavorable water molecules within the binding site.¹⁰⁸ As these water molecules are expelled from the binding pocket, their translational, rotational and vibrational degrees of freedom become unhindered. Thus, this process results in a favorable entropic contribution arising from the freeing of the water molecules present in the binding site.¹⁰⁸ This contribution is very much system dependent and, in some cases, may even mean that the total entropy change (solvent + solute) turns out to be a favorable for ligand-binding.⁷⁹

1.5.1: Isothermal Titration Calorimetry

There are several experimental techniques which can be used to estimate the thermodynamic parameters of binding reactions, in particular when an accurate measurement of the ligand-binding affinity is desired. The current gold-standard methodology is the Isothermal Titration Calorimetry (ITC) assay.^{80,81} In an ITC experiment, the apparatus comprises an adiabatic box with two recipients (or cells) inside whose temperatures are measured by heaters.^{81,82,380} One of the recipients contains a reference

solution and the other contains a solution with the target receptor (reference and sample cell, respectively). During the experiment, the reference cell is heated to a constant temperature and the sample cell is injected at regular intervals with a solution containing the ligand.^{81,82,380} As a result of the ligand binding to the receptor, the temperature inside the sample cell changes. This change is then detected and the heater associated with the sample cell will either receive more or less voltage from the power source, producing more or less heat, to maintain the temperature equal between the two cells.^{81,82,380} The power needed to maintain both cells at the same temperature is recorded over time. The data points are fit to a sigmoidal curve and the K_D value can be estimated.^{81,82,380} A graphical description of the assay is given in **Figure 1.5**.

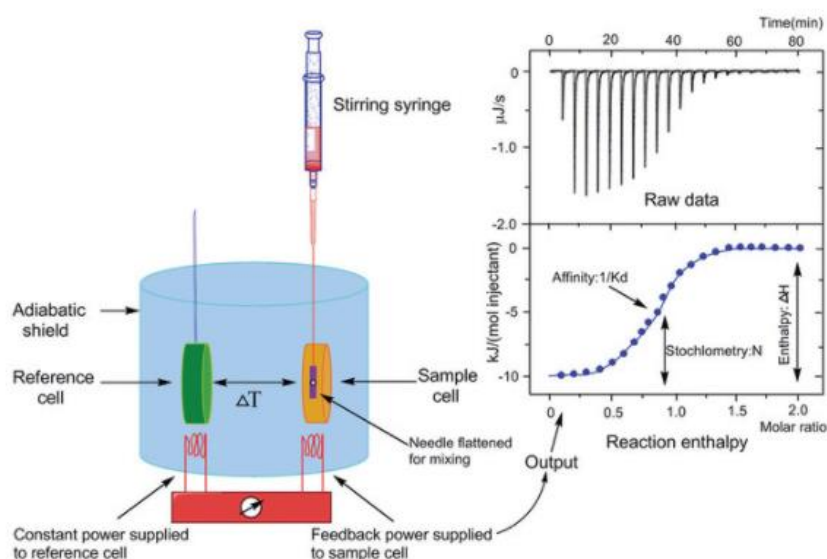


Figure 1.5 – Schematic representation of the ITC experimental equipment. On the left, the two cells inside the adiabatic box, the equipment for ligand aliquot injection and the voltage-controlling apparatus are highlighted. On the right the result of a typical ITC experiment is shown, a curve which is fit with a Hill curve to determine the all thermodynamic parameters. Extracted from “Choosing a suitable method for the identification of replication origins in microbial genomes” by Huang *et al.*, 2015. *Frontiers in Microbiology*, 6, 1049. Copyright © 2015, Song Zhang and Huang.⁸³

The technique allows direct retrieval of the reaction stoichiometry, K_{eq} and ΔH° .^{81,82,380} The remaining parameters can be obtained using **Equation 1.8**, thus allowing the direct evaluation of the ligand-binding affinity.^{81,82} Furthermore, ITC experiments are straightforward to analyze for an experienced manipulator and produce highly accurate data, making them useful in both academic and industrial settings. Some of the disadvantages of ITC are the requirement for a large amount of receptor and ligand, and the time necessary to carry out the experiment (which can go up to several hours).^{81,82,380} As such, ITC is not amenable to screening campaigns due to its costly nature and

alternatives exist. One viable alternative is offered by computational methodologies for binding free energy calculations, which will be discussed in the next chapters.

2. Computer Aided-Drug Design

2.1: Computer-Aided Drug Design methodologies

Computational methodologies have seen growing application in drug discovery campaigns over the last decades,^{27,29,31,84} as highlighted in 1981 by the article “Next Industrial Revolution: Designing Drugs by Computer at Merck” which it underpinned the importance of including numerical approaches in rational drug design pipelines.²⁹ While the computational chemistry field was already established at the time and supplied important contributions to drug discovery projects, this article garnered the interest of the general public in CADD.⁸⁵ Drug discovery projects focused on HTS for hit identification at the time and CADD methodologies were introduced to complement HTS experiments and to aid in lead optimization.²⁹ Employment of CADD by academic institutions and pharmaceutical companies thus became essential for the preliminary stages of drug discovery.²⁴ In particular, CADD-based approaches are mostly applied in hit identification and lead optimization cycles through the application of purpose-built methodologies targeting specific endpoints like binding affinity prediction or ADMET properties.

One of the main goals of molecular modelling as a pharmacological and medicinal chemistry tool is to predict novel biologically active compounds ahead of chemical synthesis and experimental testing in the wet-lab.⁸⁶ Early CADD approaches focused on using simple molecular descriptors and topological information from small-molecule ligands aiming to predict different chemical endpoints, a methodology known as quantitative structure-activity relationship (QSAR).⁸⁶ Later on, three-dimensional molecular modelling techniques were introduced, allowing the prediction of realistic chemical structures and their associated biological properties. In the 1980s, drug design methods started using experimental structural data from macromolecular targets like proteins or DNA.⁸⁶

Through CADD approaches, specifically through Virtual Screening (VS), compounds in large chemical libraries are tested towards a given chemical end-point and the non-interesting molecules are filtered out, thus reducing resource consumption at the Hit Discovery phase.²⁷ The best ranked compounds progress to experimental testing and optimization steps.²⁷ In a drug discovery campaign, CADD is usually used for three purposes: library filtering, supplying hit compounds and to guide lead-optimization steps and design new compounds by either fragment-growing techniques or by medicinal chemistry.²⁹ These numerical methodologies can be separated into two categories, with preference of approach given according to the availability or not of a three-dimensional

(3D) structure of the pharmacologically relevant protein conformation. The two categories of methods are Ligand-based and Structure-based Drug Discovery (LBDD and SBDD, respectively).^{27,84} A schematic representation of CADD methodologies employed in the context of drug discovery campaigns is shown in **Figure 2.1**. A successful application of CADD methodologies can yield multiple hit compounds, but only some of them are amenable to optimization and experimental testing.⁸⁷

Throughout the years, CADD methodologies have proven their worth both in performance with respect to HTS and in supplying hits which would later on become drugs approved by the FDA or the EMA. One particular example comparing the performance of HTS vs CADD-based procedures is the work of Doman *et al.*^{29,88} In this study, the goal was to identify tyrosine phosphatase-1B (PTP1B) inhibitors. The researchers employed CADD to screen the ACD database, comprising 235,000 commercial available compounds, against the X-Ray crystal structure of PTP1B, and used HTS to screen a corporate library of 400,000 compounds.^{29,88} Comparing both approaches, it was found that CADD yielded a higher hit rate (127 hits from 365 tested compounds, 35%) than the HTS based approach (81 hit compounds, 0.021% hit rate).^{29,88} A hit compound was defined as a compound for which the experimental IC₅₀ was below 100 µM. From the molecular docking experiment, 21 predicted hits had an IC₅₀ below 10 µM. However, the authors report that correlation between IC₅₀ and the docking scores is poor. Other researchers have also documented that docking scores correlate poorly to experimental binding affinities.^{89–91} This is a feature observed in other studies, which upholds the argument that docking scoring functions are, first and foremost, a screening tool and not accurate enough for quantitative calculations.⁸⁸ Other examples of success stories employing CADD include the work of Chang *et al.*⁹² which was able to find a new series of non-β-lactam antibiotics, the oxadiazoles, following virtual screening of 1.2 million compounds from the ZINC database.⁹³ These compounds showed inhibitory activity targeting the penicillin-binding protein 2a (PBP2a) of methicillin-resistant *Staphylococcus aureus* (MRSA), the cause of many infections in hospitals.^{31,92} Some examples of drugs which were discovered using CADD include Captopril⁹⁴ for hypertension and Saquinavir, for HIV therapy²⁷.

Over the years, Ligand-based (LB) and Structure-based (SB) drug discovery methods evolved separately. However, combination of both strategies yields higher effectiveness since both approaches act in a complementary manner.²⁴ In the following subchapters we will discuss molecular recognition and what interactions drive it, structure-based drug discovery and finalize by describing the steps a structure-based virtual screening workflow must follow. The critical aspects of the workflow will be approached, such are protein and ligand library preparation, and the scoring functions underlying binding score (or affinity) predictions and compound ranking will be analyzed.

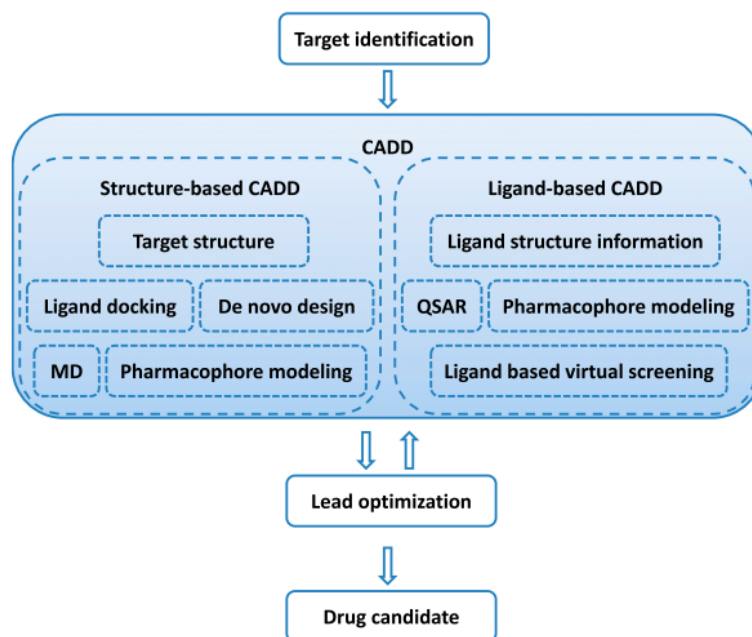


Figure 2.1 – Schematic representation of the Computer-Aided Drug Discovery workflow typically employed in the early phases of Drug Discovery campaigns. QSAR: Quantitative Structure-Activity Relationship; MD: Molecular Dynamics. Extracted from “Computational Methods in Drug Discovery” by Sliwoski et al., 2014. *Pharmacological Reviews*, 66(1), 334-395. Copyright @ 2013 The American Society for Pharmacology and Experimental Therapeutics.²⁹

2.2: Molecular Recognition

For the binding of a ligand to a protein, it is fundamental that the two species recognize each other. The Nobel Prize in Chemistry awarded to Cram, Pederson and Lehn "for their development and use of molecules with structure-specific interactions of high selectivity"⁹⁵ in 1987 shows how much supramolecular chemistry and non-covalent interactions are fundamental for biological processes and technological development. It also highlights supramolecular chemistry as an exciting field, which focuses on studying reversible non-covalent molecular association reactions and whose underlying mechanism is molecular recognition.⁹⁶ Molecular recognition is critical for supramolecular binding and is a symptom of the complementarity between two molecules whose strength of interaction is dictated by several factors, among which are their chemical structures, spatial arrangement, solvent effects and entropy.⁹⁷ Molecular recognition events mediate cellular interactions such as protein-protein interactions or binding and unbinding of small molecule ligands to receptors.⁹⁸

Complete understanding of molecular recognition processes remains a challenge. Nevertheless, there are ways to address this issue through numerical approaches based on MD simulations, as will be illustrated in **Chapter 3**. Knowledge derived from studying

this processes is critical in many areas of science, such as biochemistry, medicine and pharmacology. For perspective sake, there are close to 160 000 X-ray structures of proteins, nucleic acids and associated complexes available in the Protein Data Bank⁹⁹ and more than 800 000 small molecule crystal structures available in the Cambridge Structural Database.¹⁰⁰ These numbers are expected continue to grow, resulting in a wealth of information with the ability to impact pharmaceutical and agrochemical research, among others. In particular, the knowledge extracted from these supramolecular complexes could open new avenues of research to target diseases which, to this day, remain orphan of a safe and effective treatment such as some cancers or neurological disorders. Beyond, it would also impact significantly other research areas like energy storage or molecular bioremediation. One example of the latter is the usage of cyclodextrins¹⁰¹ to bind and chelate pollutants from contaminated water.¹⁰² Thus, it is critical for CADD campaigns to use approaches able to describe numerically the interactions between the binding partners with a high degree of accuracy.

A fundamental ingredient for molecular recognition in the context of protein-ligand binding are the interactions established between the binding partners. Indeed, there are many types of interactions molecules establish between each other. The interactions which do not imply bond formation or breaking are known as non-covalent interactions, which can be modelled as in **Equation 2.9**. Non-bonded interactions include electrostatic forces, in origin due to Coulomb interactions between charges (permanent dipoles, quadrupoles, etc.), polarization forces, which arise from the dipole moment induced in atoms and molecules by the electric fields of nearby charges and/or permanent multipoles, dispersion terms, charge-fluctuation forces and induced dipole-induced dipole forces. The balance of such forces is at the origin of the non-bonded interactions between different molecules and drives molecular recognition processes.

2.2.1: Electrostatic Interactions

Some non-bonded interactions are quantifiable using a Coulomb law and are pairwise additive, with repulsive or attractive character depending on the partial charges of each atom pair.¹⁰³ Examples are hydrogen-bonds (interactions between a partially positive charged hydrogen atom and electronegative and partially negatively charged oxygen, nitrogen, sulfur or fluorine atom¹⁰⁴) and halogen-bonds, which are highly directional. Halogen bonds involve a halogen atom like chlorine which acts as an electrophile and interacts with a nucleophilic atom which is highly electronegative, similar to hydrogen-bonds.⁹⁶

In addition to these, dispersion, induction and steric repulsion, electrostatic (between permanent charge distributions over the interacting partners) and the Charge Transfer (CT) (dynamic redistribution of the electronic density) contributions can play a non-negligible role. Dispersion interactions, also known as London forces in honor of Fritz London, arise due to the highly fluctuating nature of the electron cloud between molecules.¹⁰⁵ These attractive interactions occur when a feeble and temporary dipole is generated. These are also known as induced dipole-induced dipole interactions and are the weakest of the intermolecular forces known.¹⁰⁵ Another type of non-covalent molecular interaction are π -interactions, which arise from the interplay between a molecule and the π -systems of conjugated molecules. One example is the cation- π interaction between the positive charge of a sodium atom and the electron cloud atop a benzene ring.¹⁰⁶ The balance of the attractive terms is taken into account with the exchange repulsion term. It manifests at short distance, as consequence of Pauli's exclusion principle between the electron clouds of molecules, as too much overlap of the electron clouds of two atoms leads to a sharp increase in energy.¹⁰⁷

2.2.2: The hydrophobic effect

Another type of non-bonded interaction manifests as the hydrophobic effect, where non-polar molecules in an aqueous media try to avoid contact with water molecules by aggregating together in order to expose as little surface as possible. Consider a system of two droplets of oil in an aqueous solution. Before these two droplets merge, the water molecules will envelop each oil droplet without being able to establish hydrogen bonds with them. This involves an energetic gain due to the establishment of a hydrogen-bonding network amongst waters and due to the favorable dispersive interactions between water and oil while at the same time an entropic penalty because some of the water molecules are restrained in position without being able to interact with the oil droplet through hydrogen-bonding. When the droplets merge, a re-organization of the water network occurs, which implies breaking the previous hydrogen-bonding network which enveloped these two isolated droplets. Upon aggregation of the two droplets, some of the water molecules which enveloped the individual droplets are released to the bulk, leading to an entropic gain as their translational, rotational and vibrational degrees of freedom are unhindered. By doing so, the oil phase exposes as little surface as possible and. This is known as the hydrophobic effect. The hydrophobic effect is critical in protein-ligand binding, as expulsion of high-energy water molecules in buried binding sites represents an entropic gain to the binding reaction and introduced a favorable contribution.¹⁰⁸ In the bulk, these water molecules can diffuse and re-orient freely, which is not the case within the binding site. There, water molecules may or not establish

hydrogen-bonds with amino acid residues. However, these water molecules will always lose translational and rotational freedom which means that releasing these water molecules into the bulk leads to an entropic gain.

2.3: Structure-Based Drug Discovery - Virtual Screening

The availability of high quality 3D structures of the target protein opens to the application of SBDD techniques, aiming at the discovery of new and innovative modulators of protein function.^{84,90} The scientific community, over the years, has developed many methodologies which profit from the existence of such structures to collect information regarding ligand binding sites, protein conformational dynamics and interactions between proteins and ligands.¹⁰⁹ This knowledge is useful to design new modulators and to assist in the interpretation of experimental.⁹⁰ SBDD is an iterative endeavor which relies on cycles of ligand design, numerical evaluation, compound prioritization and experimental testing. The main techniques employed in SBDD are de novo drug design and Structure-Based VS (SBVS).^{27,91,109,110}

There are many steps which must be carefully carried out for a successful application of an SBVS campaign. These steps include protein preparation, binding site identification, ligand library preparation and design, molecular docking and binding pose scoring.^{31,61,109,111,112} In the following we will analyze each of these steps, describing common approaches and pitfalls. A general workflow for SBVS is shown in **Figure 2.2**, where critical steps such as protein preparation, ligand library preparation, molecular docking and scoring are highlighted.

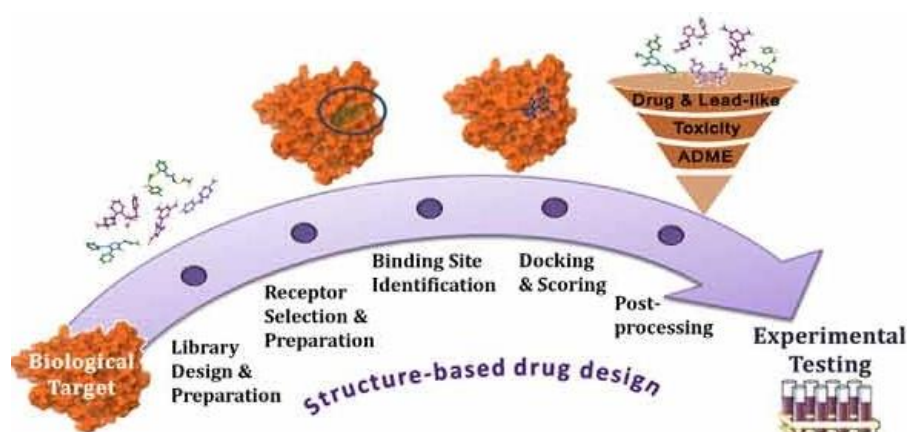


Figure 2.2 – General workflow for employed in SBVS campaigns, obtained from “Structure Based virtual screening for drug discovery: principles, applications and recent advances.”Lionta et al, 2014. Current Topics in Medicinal Chemistry, 14(16), 1923-1938. Copyright @ Bentham Science Publishers.¹¹¹

In *de novo* Drug Design, ligands are either built by fragment-linking or fragment-growing¹⁰⁹ approaches. In the former, the binding site is mapped, key interacting regions are found and the ligand fragments which better interact with those regions are selected, docked and linked together to form a final ligand.^{24,109} For the success of this technique, the selection of the linkers connecting the fragments is critical. In fragment-growing, a central fragment is docked inside the binding cavity and search algorithms are used to place and score small chemical modifications of the ligand, growing it iteratively.^{24,109} An example of a SBVS tool which employs this rationale is FlexX.^{113,114}

The second type of approach is the SBVS campaign, where small-molecule libraries are screened against the target of interest, similarly to HTS, through numerical simulations.^{31,91,115} In SBVS campaigns, hit identification is carried out by first predicting the ligand binding mode within the binding site present in the protein structure and then evaluating the corresponding binding affinity either through docking scoring functions or more accurate and expensive numerical methods.⁸⁴ Among the methodologies within SBDD, particular attention has been given to Molecular Docking, pharmacophore modelling and Molecular Dynamics (MD) coupled to binding free energy calculations as tools for drug design.^{27,29,84}

Within SBVS, *in silico* methodologies are employed to identify potential hits from large chemical libraries. Compounds are prioritized according to their predicted binding affinity and then a subset is selected for experimental testing. If active compounds are found, researchers try to solve the structure of the protein-ligand complex to gain insights into the protein-ligand association mechanism.^{84,112} Some of the information which can be obtained from the structure include the preferred ligand-binding conformations, protein residues which are key for ligand binding, characterization of unknown binding sites and capturing of ligand-induced allosteric effects.¹¹² The information harnessed from this cycle is then used to design new ligands, and the cycle is restarted until ligands with sufficient potency and desirable pharmacological profiles are found.

2.3.1: Protein preparation for SBVS

There is an old adage which states that the quality of the input data conditions the quality of the output. Such adage is verifiable in SBVS, with special emphasis of the protein and ligand preparation stages. Structure-based Virtual Screening starts from a three-dimensional structure of the target protein, either in complex with a known ligand or in the apo state.^{91,115} Structures of protein-ligand complexes are easily found in the PDB database.⁹⁹ The PDB database contains, as of the beginning of 2021⁹⁹, close to 180k structures at varying degrees of resolution (from < 1Å to > 4.6Å) which are readily accessible. From these structures, 157k correspond to 3D protein structures in complex

with a variety of other chemical entities like fragments, drug-like small molecules or peptides.⁹⁹ However, PDB structures typically contain only heavy atoms and do not supply information regarding topology, bond order or formal charges. Additionally, many structures are obtained with bound co-factors, ligands, water molecules or metal ions.¹¹¹ It is also common to find missing portions in the PDB structure, ranging from single amino acid side chains to entire loops whose position could not be determined due to the resolution of the structure in that specific area of the protein. Steric clashes can also be present and again, can be ascribed to structure resolution issues. All of these details play a significant role in the success of SBVS and must be handled with care when preparing the protein.

2.3.1.1: Protonation state of titratable residues

A key aspect of protein preparation is the determination of the protonation state of the side chains of titratable residues. The electrostatic properties of proteins are heavily influenced by the ability of amino acid residues to exchange protons with the environment in a pH-dependent manner¹¹⁶ and it is known that the protonation state of the titratable residues in a binding site play a significant role in stabilizing molecular conformations and in binding ligands.¹¹⁷ Such amino acid residues include aspartic and glutamic acid, cysteine, tyrosine, histidine and lysine.¹¹⁸ There exist different methods to compute the protonation state of titratable residues.^{119–121} Some are based on force-fields at an atomistic level and where a full thermodynamic cycle is built to compute the free energy required to bring the amino acid from solution to the protein environment, mostly through perturbation theory.^{122–124} Other methods, which are more approximated, employ continuum electrostatic models like the Poisson Boltzmann (PB) or the Generalized Born (GB) models¹²⁴, which are faster than perturbation theory-based methods but are less accurate. However, even these methods entail a significant computational cost which motivates the development of empirical methods like PROPKA.¹²¹ While empirical methods are by definition non-rigorous, they are exceptionally fast and easy-to-use and have been shown to be fairly accurate. The PROPKA3¹²¹ method, which was trained and optimized using x-ray crystal structures, is one of the most popular empirical approaches. The relationship between the *pH* of the solution and the ionization state of a group HA in a given residue is given by the Henderson-Hasselbalch equation (**Equation 2.1**)¹²⁵:

$$pH = pK_a - \log \left(\frac{[HA]}{[A^-]} \right) \quad (2.1)$$

where pK_a is negative logarithm of the acidity constant (K_a) of the HA group, $[HA]$ is the concentration of the HA group in the protonated form and A^- is the concentration of HA group in the deprotonated form.^{121,126} When pH is greater than pK_a , the group will be predominantly found in the deprotonated form. Inversely, when pH is smaller than pK_a , it will be predominantly found in the protonated form. When pH is equal to pK_a , both forms co-exist with the same concentration in solution. The free energy change associated with transporting the titratable group from solution to the protein environment can be computed from a thermodynamic cycle (**Equation 2.2**)^{121,126}

$$\Delta G^{Protein}(AH \rightarrow A^- + H^+) = \Delta G^{Water}(AH \rightarrow A^- + H^+) + \Delta G_{Solvation}^{Water \rightarrow Protein}(A^-) - \Delta G_{Solvation}^{Water \rightarrow Protein}(AH) \quad (2.2)$$

where ΔG^{Water} and $\Delta G^{Protein}$ are the pKa value of AH in water and in the protein in free energy units and the last two terms correspond to the solvation free energies of translocating the deprotonated and protonated forms of the group from water to the protein environment.^{121,126} Taking the relationship $\Delta G = 2.30RT * \Delta pK_a$, the pKa of the protein residue in the environment of the protein is given by **Equation 2.3**^{121,126}

$$pK_{a,i}^{Protein} = pK_{a,i}^{Water} + \Delta pK_{a,i}^{Water \rightarrow Protein} \quad (2.3)$$

where $pK_{a,i}^{Water}$ corresponds to the model value of the residue i in water. In PROPKA3¹²¹, the pKa value of a group in a residue or of a ligand is computed by summing the pKa value in water with the pKa shift arising from transporting the residue from water to the protein environment. Thus, the first term is well-known, whereas $\Delta pK_a^{Water \rightarrow Protein}$ is what remains to calculate. PROPKA3 aims at computational efficiency and so the $\Delta pK_{a,i}^{wat-prot}$ is approximated as a sum of contributions per aminoacid residue as in **Equation 2.4**^{121,126}

$$\Delta pK_{a,i}^{wat-prot} = \Delta pK_{a,i}^{desolv} + \Delta pK_{a,i}^{HB} + \Delta pK_{a,i}^{RE} + \Delta pK_{a,i}^{QQ} \quad (2.4)$$

where ΔpK_a^{desolv} is the contribution from desolvation effects, ΔpK_a^{HB} is the contribution arising from hydrogen bond interactions established, ΔpK_a^{RE} is a contribution from unfavorable electrostatic reorganization energies and ΔpK_a^{QQ} is the contribution from coulombic interactions.^{121,126}

The desolvation contribution corresponds to the energetic penalty of making a cavity in the solvent which is occupied by the protein surrounding the ionizable residue. The coulomb term is computed using a distance-dependent weighting scheme based on the

distance r between residue charge centers i and j ($w(r_{ij})$). The intrinsic electrostatic contributions are computed using the two middle terms.^{121,126} The first term corresponds to short range polar interactions like hydrogen bonds ($\Delta pK_{a,i}^{HB}$), which are modelled in a distance and angle dependent manner, requiring that the angle formed between the hydrogen bond donor and acceptor pair is at least 90° .^{121,126} The second term, $\Delta pK_{a,i}^{RE}$, corresponds to a possible hydrogen bond acceptor for acids or a hydrogen bond donor for bases. It is, nonetheless, a rare occurrence and it is unfavorable, characterizing a sort of “reverse hydrogen-bond.”¹²¹ One limitation of PROPKA is that long and short range electrostatic interactions are separated at a distance cutoff at 6\AA , disregarding interactions between residues farther away than this distance.¹²¹ Other freely available tools to compute the protonation state of titratable residues include H++¹¹⁹, SPORES¹²⁰ or Karlsberg¹²⁷, each with its given limitations. For ligand protonation studies, a common reference tool is Epik, developed by Schrödinger.^{128,129}

2.3.1.2: Hydrogen atom assignment and optimization of the hydrogen bond network

Should one open a complex structure obtained from the PDB database, one noticeable aspect is the lack of hydrogen atoms.¹¹¹ Furthermore, often due to the lack of hydrogen atoms, some side chain orientations may not be properly defined in the structure. Indeed, this appears to be the case for the terminal chi angle of Asn, Gln and His residues, which require a careful analysis to determine if a flip is required to produce an optimal hydrogen bonding network.^{130,131} There is a significant body of literature on this topic and several free tools are available. Two of the most well-known are PDB2PQR¹³² and MolProbity¹³⁰, the latter from Duke University. In MolProbity, the biomolecular complex structure is analyzed with regards to all of the atomic contacts taking place within the biomolecular assembly.¹³⁰ It detects local issues such as steric clashes and attempts to resolve them automatically, inspecting and evaluating the geometry of the residues and the dihedral-angle orientations in space.¹³⁰ In addition, MolProbity is also able to add missing hydrogen atoms to protein residues using the Reduce routine.¹³⁰ By taking into account the all-atom contact network, it determines the optimal position of the hydrogen atoms in the tri-dimensional structure while attempting to avoid steric hindrance and facilitating hydrogen bonds.¹³⁰

2.3.1.3: Energy minimization of the structure

After producing a complete PDB, where the protonation state of the titratable residues is assigned and the hydrogen atoms are added, the next step is to minimize the protein

structure such that un-resolved steric clashes can be addressed.^{111,112} One can define energy minimization as the process of finding a three-dimensional arrangement of the atoms of the system where the position found is at stationary point. As an example, a ball which is placed on a hill will roll down, until it meets a crevice or valley at the bottom. In this process the balls rolls until it reaches an energy minimum, which is, more often than not, local but on which the forces acting on the ball add up to zero.¹³³ The same logic is applied in energy minimization, where the final objective is to drive a system of N particles towards the lowest possible energy conformation. Since the local evolution of the potential energy of the system is downhill, local energy minimization routines such as steepest descent (SD) or conjugate gradient (CG) are not able to cross energy barriers and thus can only minimize towards the closest local minima.¹³³ Other methods for global optimization, such as simulated annealing, do not suffer from this. In this dissertation, energy minimization was carried out using a combination of SD and CG algorithms. Briefly, the SD algorithm is an optimization routine which attempts to find a local minimum of a function.¹³³ In an SD iteration, the gradient of the system energy is first determined considering a small step in all possible directions in the potential energy surface (PES).¹³³ Then, the direction in which this gradient is the most negative, and thus yielding the optimal direction, is selected, the positions of the atoms of the system are updated and the cycle repeats itself until the gradient converges or the number of pre-defined steps is reached, using **Equation 2.5**¹³³

$$\vec{r}_{i+1} = \vec{r}_i - \lambda_i \nabla V(\vec{r}_i) \quad (2.5)$$

where \vec{r}_i is the 3N vector containing the positions of all atoms of the system at step i , λ_i is the step size and $\nabla V(\vec{r}_i)$ is the gradient of the energy with respect to position which determines the direction along which the energy is minimized more strongly.¹³³ One interesting feature of this algorithm is that the step size is adjusted at every iteration and is incremented by a small factor when the energy of the new conformation is lower than the energy of the previous one.¹³³ Thus, if the system is evolving towards lower and lower energies, the SD algorithm can afford to increase the step size to reach the energy minimum faster.¹³³ If the energy of the new conformation turns out to be higher, the step size is decreased.¹³³ However, due to the imprecision of the SD algorithm, in most cases it is not able to drive the system to the bottom of the energy well but instead moves around it.¹³³ Nevertheless, this algorithm is useful as a first pass crude energy minimization, because it will bring the system to the vicinity of the energy minimum and fix initial bad contacts.¹³³ Then, to reach the bottom of the energy well, it is possible to apply more refined techniques for energy minimization, such as the CG approach.¹³³ The conjugate descent algorithm takes in account both the gradient of the system in the

current iteration and information about the previous step, which allows the system to converge quicker to the minima the system is close to.¹³³ The master equation (**Equation 2.6**) ruling conjugate gradient is shown below, where an additional term which encodes the memory of the previous minimization step is present such that

$$\vec{r}_i = \vec{r}_{i-1} + \lambda_i(\vec{S}_i) \quad (2.6)$$

with

$$\vec{S}_i = -\nabla V(\vec{x}_i) + b_k \vec{S}_{i-1} \quad (2.7)$$

where the $b_k \vec{S}_{i-1}$ parameter controls how much should the knowledge of the previous step influence the next one.¹³³ In particular, the first step has $b_k \vec{S}_{i-1} = 0$ and as such depends only on the value of the gradient starting from the initial structure. At steps $i > 1$, the direction of the minimization is computed by taking into consideration the direction taken in the previous step.¹³³ During the energy minimization steps carried out within this dissertation, SD was used as a crude energy minimization to alleviate some steric hindrance effects and to lead the system towards a lower energy conformation whereas the CG algorithm was used to further refine the structure obtained by SD into a true energy minima.

2.3.2: Additional steps in protein preparation

Finally, other steps of protein preparation must be considered, such as the treatment of crystallographic water molecules, co-factors and metals, assigning atomic partial charges or modelling into the structure absent amino acid side chains or loops.^{111,112} A strategy to add missing loops and amino acid side chain residues is to use MODELLER.¹³⁴ From the PDB one can obtain a FASTA sequence containing the amino acids of the protein. By comparing it to other structures through bioinformatics tools like BLAST, one can retrieve the sequences with the highest sequence identity. Using that protein structure as a reference, it is possible to reconstruct the missing portions and generate potential models. These models are given a score by MODELLER which are then used as a filter to select the best one.¹³⁴

2.3.2.1: Crystallographic binding site water molecules

A key contribution to the structure and function of proteins is given by the network of water molecules surrounding the binding site. These water molecules interact with the

solute through hydrogen-bonding, solvation and the hydrophobic effect and are important in mediating protein-protein and protein-ligand interactions.^{136,137} Since the resolution of structures determined by x-ray crystallography is too low to map hydrogen atoms, crystallographic water molecules in PDB files are identified by determining the position of the corresponding oxygen atoms.¹³⁵ The water molecules which are free to diffuse are modelled as bulk solvent. However, for the water molecules with ordered behavior, which are close to the binding site, how to model them is an open question. In biomolecular simulations, this decision heavily impacts simulation outcomes because the binding free energy of a ligand to a protein is sensitive to the position of the water molecules surrounding the binding site.^{136,137} Furthermore, the water molecules in x-ray structures only provide static information, obtained within a well-ordered structure. As such, not all crystallographic water molecules need to be kept and thus the question arises: which ones to keep, and how to select them?^{135,136}

The Consolv tool¹³⁷ tries to answer the question above by estimating the degree of conservation of water molecules in x-ray crystal structures.^{136,137} Parameters considered are the structure B-factor, the number of closest protein atoms around each water molecule, the hydrophobicity of the hydration site and the number of water-protein hydrogen bonds.¹³⁷ It was tested on a set of 7 complexes and achieved a prediction accuracy of 75%. Other tools, which aim at quantifying the interactions of the water molecules in a binding site to the protein residues, have been developed by taking inspiration from docking scoring functions. An example is WaterScore¹³⁸, which can estimate which crystallographic water molecules to keep by evaluating structural properties found in crystal structures and converting these observations into a score.¹³⁸ WaterScore has been shown to be moderately accurate in identifying conserved water molecules, highlighting a prediction efficiency between 67.4 and 71.7% in a benchmark study.¹³⁸

It can also be the case that no crystal waters are available but there is the hint that protein-ligand binding in the studied complex is stabilized by water molecules in the binding site. In this case, researchers may want to consider the effect of solvent molecules in the binding site explicitly, something which is achievable for instance by the WaterMap tool. The WaterMap¹³⁹ tool relies on short explicit solvent MD simulations where the protein conformation is held fixed by restraints. The idea behind is to record the position and orientation of each water molecule in the binding site during MD in order to produce a water density profile.¹⁴⁰ Then, the binding site is represented on a grid mesh and water molecules are mapped to grid points where their density is above that of bulk solvent.¹⁴⁰ Limitations of WaterMap, as discussed by the authors, include the short MD simulation time and the use of conformational restraints in the protein.¹³⁹ In particular, the latter prevents the receptor from adapting a fully relaxed conformation and which

may bias the analysis.¹³⁹ Nonetheless, WaterMap has been used to predict accurately the position of water molecules in binding sites before.^{136,139}

In most SBVS campaigns, however, crystallographic water molecules are typically removed. One possible reason for it may be the difficulty in determining their position and/or orientation in the binding pocket, although some tools have been described which facilitate this step. Another possibility may be the case that these crystallographic waters which are not expected to affect the binding reaction. Nonetheless, whenever possible, the treatment of crystallographic water molecules must be carried out carefully and it is advised not to remove them blindly.

2.3.2.2: Co-factors and metals

Another step is the treatment of co-factors and ions. Metal ions and co-factors can be important to stabilize the crystal conformation and, upon their removal, the structure may no longer be stable and/or functional.¹⁴¹ However, these need to be modeled separately from the protein and as such are initially removed from the structure and are added back after being carefully treated. Co-factors can typically be modelled by adding hydrogens, computing their ionization state and assigning atomic partial charges to their atoms. In the case of metals, the common procedure is to produce parameters by applying high-level quantum mechanical calculations.¹⁴¹ One example is the parametrization of the Fe³⁺ containing heme group found in the cytochrome c, whose geometry and partial charges can be derived using density functional theory (DFT) calculations at the B3LYP level of theory using the methodology described by Rarey *et al.*¹⁴¹ The critical point is that the 3D geometry of the co-factors and metal containing groups be correctly assigned, alongside the corresponding atomic partial charges and bond orders¹⁴¹ prior to VS. These tasks are normally carried out with in-house scripts or, more commonly, using the Protein Preparation Wizard tools from Schrödinger's Maestro software. The importance of protein preparation before molecular docking experiments must not be underestimated. It has been shown that SBVS employing proper protein preparation steps yields improved enrichment factors as in the systematic evaluation by Sastry and co-workers⁹¹ using the GLIDE¹⁴² validation set and a series of decoys from the Directory of Useful Decoys (DUD)¹⁴³ database.

2.3.3: Binding site identification

The next task within the workflow is the characterization and/or identification of the ligand binding site, if it is unknown.¹⁰⁹ A binding site can be defined as a small cavity to which a ligand can bind, eliciting an effect on the protein target upon complexation. The

ideal binding site is a small concave cavity comprising different functional groups¹¹¹ with which the ligand can establish favorable interactions as well as having hydrophobic characteristics.¹⁴⁴ When the ligand-binding site is unknown, or when new allosteric modulators are desired, tools which are able to predict potential binding sites on the protein structure can be employed. As described by Lionta *et al.*, here are different avenues in this direction.¹¹¹ One way to approach the problem is to use tri-dimensional structures of the protein of interest, where one can map binding pockets using small organic probes and evaluating the druggability of the cavities found through tools such as FPocket¹⁴⁵ or others.^{111,146,147} Another approach is through flooding simulations using Molecular Dynamics, where several different small-molecule evolve dynamically over time in a simulation box containing the protein structure. The idea is that potential binding sites can be found by probing the protein with a diverse set of chemical probes while taking into account protein flexibility.¹¹¹ Some examples of programs to carry the following approach are MDmix¹⁴⁸ or SILCS.^{149,150} In this context, the MD simulations carried out are long, introducing a large computational overhead, and can help understanding the mechanism behind ligand binding to that protein.¹¹¹ It is also possible to employ water molecules as probes instead of small organic molecules.^{111,139}

It is important to know where the binding site of interest is located in the protein structure before a SBDD campaign, as some of the methods employed in the pipeline require extensive calculations and computational power. Furthermore, fundamental information can be retrieved from structures of proteins co-crystallized with their corresponding ligands, as well as from mutation studies identifying key residues responsible for mediating protein-ligand binding.^{24,109} It is even possible to design ligands based on this information, through a four step workflow as described by Rognan *et al.*¹⁵¹: fragmentation of the reference bound ligand, mapping the fragments which interact with binding site and the type of interactions established, defining the chemical environment of the sub pockets in the binding site based on the reference ligand portions which interact therein, converting this information into a fingerprint descriptor.¹⁵¹ Finally, it is possible evaluate ligands based on their predicted interactions within the binding site and on the ligand-binding site shape complementarity with respect to the reference compound by means of a fingerprint similarity evaluation.¹⁵¹

2.3.4: Ligand library preparation

The chemical space theoretically existent contains 10^{60} possibilities of druglike small-molecule ligands, with 10^{20} to 10^{24} of these corresponding to molecules with up to 30 atoms.^{152,153} However, exploring the full chemical space within a SBDD campaign is an impossible task at the moment.^{152,154,155} Nonetheless, VS campaigns aim at evaluating

large chemical libraries in search of new hit compounds. Databases employed in SBVS are typically composed of drug-like small molecules which are readily purchasable or synthesizable.¹¹¹ In addition, these compounds should also possess some desirable characteristics like solubility in aqueous media and chemical stability, coupled to the absence of toxic moieties in the chemical structure. However, large chemical libraries are typically of the order of millions of compounds, and screening a very large ligand library implies a large computational cost which may be a limit factor in the SBVS campaign. Depending on the computational resources available, a compromise must be reached regarding the ligand library size to be explored. The appropriate selection and preparation of the ligand library thus constitutes a critical step in SBDD and must be carefully tackled. As such, it is often the case that researchers working with very large chemical libraries cluster and filter them using a set of rules before proceeding to the next steps in the SBVS workflow

Typical filters for ligand library preparation include the removal of non-drug-like compounds through the Lipinski R-o-5³⁹ or other measures of druglikeness, removal of compounds containing atoms which are not common in organic molecules, filtering with respect to some structural features, removal of compounds exhibiting substructures known to occur in Pan Assay Interference Compounds (PAINS),^{1,156,157} and removal of compounds whose molecular features are significantly different from those of the known actives. In particular, PAINS compounds are known to give false positive results in HTS by reacting in a non-specific manner with many biological targets.^{156,157} These highly promiscuous compounds can be identified by the presence of some functional groups, which are shared among them. Some examples of PAINS are toxoflavins, isothiazoles, alkylidene barbiturates and quinones.¹⁵⁶ If possible, one should also consider database filtering steps to account for the ADMET profiles of the compounds.^{109,111,159,160} A tool which could be used for step is the FAF-Drugs³¹⁵⁸, a public webserver which can filter compounds according predictions of physicochemical and ADMET properties. These ligands must also be filtered according to additional criteria, some of which are known ADMET end-points, to evaluate their potential to become new drugs.^{109,111,159,160} One example is the ability to go through the brain-blood barrier (BBB)^{161,162}, which is indirectly estimated by evaluating the compound ability to permeate through Caco-2 cells *in vitro*.¹⁶³ *In silico* models for Caco-2 cell permeability have been built and are routinely used in lead-optimization.^{164,165} An additional manner to reduce the computational cost in a prospective VS campaign can be to cluster the ligands according to structural similarity.³⁶⁷ The clustering step allows researchers to identify molecular structures which are representative of a given chemical scaffold. The cluster center is then evaluated and studied and, in case the compound is active, researchers can come back and screen all compounds belonging to that cluster.

Automatic workflows for library filtering play a major role in the preliminary stages of drug discovery and have attracted significant attention from the scientific community. Several examples of tools exist in the literature, such as VSPrep¹⁶⁶, LigPrep¹⁶⁷ and the recently developed PrepFlow.¹⁶⁸ In particular, PrepFlow is able to carry out large chemical library preparation with impressive speed while at the same time being robust to potential errors and readily transferable to high performance computing centers (HPC). It takes advantage of the available resources to parallelize the calculations to even higher efficiency.

Ligand preparation also requires that the compounds must have bond order parameters attributed and their valences filled, alongside atomic partial charges and the appropriate protonation state at the pH of interest.¹⁶⁸ Partial atomic charge calculations can be carried out using empirical schemes, such as ingasteiger charge calculations¹⁶⁹, using semi-empirical methods such as the one employed in the AM1-BCC¹⁷⁰ scheme or using quantum-mechanical methods. In particular, the method employed to compute partial atomic charges over the course of this dissertation was the Restrained-Electrostatic Potential method (RESP)¹⁷¹ as developed by Bayly and co-workers, using the Gaussian09¹⁷² tool at the HF/6-31G* level of theory. In RESP, the goal is to derive atomic charges using *ab initio* calculations. These quantum mechanically-derived atomic partial charges also account from multipole moments.¹⁷¹ The 6-31G* basis set was selected because it is known to overestimate the polarity of molecules by as much as the dipole is enhanced in TIP3P¹⁷³ water over its value *in vacuum* and yields a balanced representation of solvent-solvent, solvent-solute and solute-solute interactions, as described by the Bayly *et al.*¹⁷¹

In some cases, general purpose chemical libraries may not fit the requirements of the SBVS campaign. Indeed, it is often the case that SBVS campaigns require the design of custom-made libraries to target a specific problem.¹¹¹ Thus, customization of ligand libraries may be required to fit the needs of the project. As exemplified by Lionta *et al.*, libraries can be built to conform to some distribution of molecular properties, or be tailored to a given target based on reference compounds.¹¹¹ Additionally, may also be desirable to design a ligand library with the aim of maximizing the sampled chemical space.¹¹¹

2.3.5: Molecular Docking

The most popular technique among SBVS methods is molecular docking. When carrying out a molecular docking experiment, two goals exist: the determination of the optimal position and orientation of the ligand molecules inside the protein cavity delimiting the binding site (*pose prediction*) and the calculation of a score reflecting how favorable that

binding reaction is (*scoring*).^{26,27,89} The first step of molecular docking, pose prediction, can be carried out using different approaches, among them rigid docking, flexible docking or ensemble docking. A host of molecular docking programs have been proposed over the years, including DOCK¹⁷⁴, AutoDock4¹⁷⁵, GLIDE¹⁴², GOLD¹⁷⁶, PLANTS¹⁷⁷, ICM², FlexX¹⁷⁸ and others^{26,112,179,180}. Docking methods employ different philosophies for pose prediction, such as systematic conformational search or stochastic conformational search methods. Systematic methods are those which place the ligand in the binding site after considering rotations and translations of all degrees of freedom. Stochastic methods are those which start from a predicted ligand pose and try to evolve it to a lower energy conformer by stochastic torsional searches based either on Monte Carlo simulations or on Genetic Algorithms. A third method to produce protein-ligand binding poses goes through MD simulations and energy minimization to explore the rough PES of a molecule.^{111,112,181}

2.3.5.1: Rigid docking

Within docking approaches, the focus is on predicting the bound protein-ligand complex configuration. In rigid docking experiments, the protein structure is replaced by a grid representation centered on the binding site.²⁶ Solvent effects are typically neglected or approximated using an implicit solvent model and entropic terms are usually not present or crudely approximated by considering the number of expected frozen torsions.¹⁸¹ Ligand conformations are generated and then optimized to maximize the complementarity between binding site and ligand. The proposed ligand binding poses are then ranked using a scoring function.¹⁸¹ In most docking studies in the past, calculations were carried considering the protein and the ligand as rigid entities.¹⁸² The binding pose would be based on the binding site-ligand shape complementarity by means of a geometrical fit¹⁸², using Fischer's "lock-and-key" theory as the underlying paradigm.¹⁸³ As such, protein and ligand flexibility was overlooked and binding poses were obtained from translations and rotations of the ligand relative to the binding site. However, neither ligands nor proteins are rigid entities and, indeed, exhibit a degree of flexibility.¹⁸⁴ Unfortunately, considering the full protein as a flexible entity is computationally expensive because of the large number of degrees of freedom within the receptor. However, Emil Fischer's theory could not explain the behavior of enzyme noncompetitive inhibition nor allosteric modulation.¹⁸² As such, a new theory which did consider some degree of receptor flexibility was proposed in 1958 by Koshland, the so-called "Induced fit" theory, where the ligand induces protein conformational changes such that upon binding, the protein is able to perfectly accommodate it in the binding site.¹⁸⁴

2.3.5.2: Flexible docking

In flexible docking approaches for SBVS, there are two ways to account for protein side-chain flexibility. On one hand, some methods rely on pre-computed low energy conformations of the ligand which are placed on the binding site and allow the binding site side-chains to wrap around it.^{113,185} Other methods rely on user-supplied information describing which portions of the receptor should be treated as flexible and the search space explored by these methods spans all possible ligand conformations, translations and rotations coupled to the conformational exploration of the flexible parts of the receptor.¹⁸³ Flexible docking methods allow consideration of the receptor flexibility, which is critical in binding reactions, but also exhibit some limitations.¹⁸³ In particular, it is difficult to find the global energy minimum structure in solution spaces which grow exponentially with the number of flexible degrees of freedom and the associated computational cost of may hamper the efficiency of the calculations.

2.3.5.3: Ensemble docking

Given the limitations of rigid and flexible docking, efforts were devoted towards a cost-efficient description of protein flexibility within docking experiments. It was known that neglecting protein flexibility meant that the thermal fluctuations ruling molecular motion were overlooked.^{111,185} These motions lead macromolecules to explore a myriad of conformational states where the shape of the binding site may change significantly, changing the available binding volume. A proposed solution was to carry out docking experiments using many different protein conformations in a methodology known as ensemble docking.^{111,185} Ensemble docking is a technique where multiple conformations of the target protein are used and ligand libraries are docked and scored against each of them. The final score is taken as a consensus between all predictions. The paper originally describing ensemble docking of small molecules was published in 1999¹⁸⁶, targeting the catalytic domain of HIV Integrase. The researchers had carried out extensive MD simulations and noted large binding site fluctuations. This observation prompted them to test and confirm that usage of multiple protein conformations, arising from MD or from x-ray crystallography, yielded better binding affinity predictions than those arising from single structure docking upon production of the consensus scores.¹⁸⁶

2.3.6: Scoring of docking poses

In the second step of molecular docking experiments, multiple ligand binding poses are ranked according to their docking score. The prediction of the docking score is based on an underlying scoring function which belongs to one of four types: knowledge-based¹⁸⁷, force field¹⁴², empirical¹⁷⁷ and machine learning scoring functions.¹⁸⁸ In the following, we will analyze empirical, forcefield and knowledge-based scoring functions. Machine learning-based scoring functions, which have been developed in the last years, will not be reviewed and the interested reader is encouraged to the work of Ballester *et al*¹⁸⁸ for an example of a machine-learning based scoring function.

2.3.6.1: Empirical scoring functions

Empirical scoring functions are used to score a large number of docking based on a weighted sum of molecular features.¹⁸¹ The underlying idea is that the binding free energy of a protein-ligand complex can be correlated to a set of additive and independent variables.¹⁸¹ Each feature is associated with a given weight, which is calibrated by regression analysis using binding affinities from experimentally determined structures of protein-ligand complexes.^{181,189} Each molecular feature is selected by the researcher to represent one or more key intermolecular interaction.^{181,189} Examples of features considered are explicit hydrogen bonds or entropic terms related to the number of ligand rotatable bonds.^{181,189} The first empirical scoring function developed was LUDI by Böhm¹⁹⁰. It was able to supply predictions of the absolute binding free energies starting from 3D structures of protein-ligand complexes. As the field matured as a whole, many empirical scoring functions were proposed. Examples are ChemScore, GlideScore and CHEMPLP.^{142,176,177} The CHEMPLP scoring function is part of the PLANTS software for molecular docking and was used in the VS campaigns described in **Chapter 6**.

In PLANTS¹⁷⁷, the protein-ligand docking problem is treated as a continuous global optimization problem. The dimension of the problem is dependent on the number of considered degrees of freedom.¹⁷⁷ For the ligand, 3 translational, 3 rotational and rl torsional degrees of are considered whereas for protein only rp torsional degrees of freedom for the flexible side-chains or rotatable hydrogen bond donor groups are taken into account.¹⁷⁷ The total number of degrees of freedom is then $n = 6 + rl + rp$. The PLANTS algorithm is based on an ant-colony optimization (ACO) algorithm which is inspired by the real behavior of ants.¹⁷⁷ When ants walk, they deposit pheromones along the paths they take. If an ant is following the track of others and is faced with choosing between two paths, it will most probably take the path with the highest pheromone concentration. Over time, this will lead to increased pheromone concentration on the most travelled

tracks, those which correspond to the optimal paths.¹⁷⁷ For a given docking problem, each virtual ant will assign a value j to a degree of freedom and evaluate how good that move is by evaluating the energy of the conformation.¹⁷⁷ After all ants have carried out the conformational search, by iteratively moving all possible degrees of freedom and finding their optimal values, and improved the resulting protein-ligand complex structure by a local refinement algorithm, the information is used to modify the pheromone trail.¹⁷⁷ In practice, the values associated the most favorable moves for a given degree of freedom have an increased pheromone concentration and the values which led to worse solutions will have a decreased pheromone concentration.¹⁷⁷ The iterative application of the ACO algorithm will produce docking poses which were constructed from the most favorable moves found for each degree of freedom at each iteration, until the algorithm converges on a set of solutions proposed by the ensemble of ants. The functional form of CHEMPLP is shown in **Equation 2.8**:

$$f_{PLANTS_{CHEMPLP}} = f_{plp} + f_{hb} + f_{hb-ch} + f_{hb-CHO} + f_{met} + f_{met-coord} + f_{met-ch} + f_{met-coord-ch} + f_{clash} + f_{tors} + c_{site} \quad (2.8)$$

where the steric complementarity of the protein and ligand is modelled by the piecewise linear potential (PLP), angle-dependent terms for hydrogen-bond interactions and metal binding (f_{hb} and f_{met} terms) are extracted from GOLD's ChemScore¹⁷⁶ and the torsional potential from the TRIPOS forcefield¹⁹¹ (f_{tors}) is included in conjunction with a heavy atom clash term (f_{clash}).¹⁷⁷ The final term, c_{site} , corresponds to a term that is used to guide the search algorithm towards the binding site.¹⁷⁷ This potential penalizes placing ligand heavy atoms outside of the binding site definition, which in PLANTS is encoded by a sphere.¹⁷⁷

These scoring functions are simple and fast to evaluate, which makes them attractive for VS purposes. However, the gain in computational efficiency comes with a cost in the accuracy of the calculations. Furthermore, since these equations require fitting using a training set to determine the weight of each contribution, the models are not transferable and may exhibit system-dependent performances.^{177,180,181}

2.3.6.2: Force field-based scoring functions

Force field-based scoring functions consist of a sum of energy terms which are extracted from classical force fields.¹⁸¹ An example of a widely used force field-based scoring function is DOCK.^{174,192} In DOCK, the protein-ligand interactions are quantified within a Molecular Mechanics formalism as in **Equation 2.9**^{174,192}:

$$\Delta G^\circ = \sum_i \sum_j^{protein\ ligand} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^6} + 332.0 \frac{q_i q_j}{e(r_{ij}) r_{ij}} \right) \quad (2.9)$$

where r_{ij} is the distance between atom j in the ligand and atom i in the protein, A_{ij} and B_{ij} are the van der Waals radii of atoms i and j and q_i and q_j the corresponding atomic partial charges. The above shown equation contains two components: the first one is the 12-6 Lenard-Jones potential, which is used to model van der Waals interactions, and the second is the Coulomb law to model electrostatic interactions.^{174,192} The non-bonded interactions are typically computed *in vacuo* with a distance-dependent dielectric constant to account for solvation effects.^{174,192} However, it is desirable to include a more accurate treatment of the solvation free energy, which can be achieved at a reduced cost through implicit solvent models.¹⁷⁵

One scoring function which employs an explicit treatment of the solvation effects is the AutoDock4¹⁷⁵ scoring function. In AutoDock4, the binding free energy is estimated by considering five free energy contributions: a van der Waals contribution, an electrostatic contribution, a hydrogen-bond contribution, a ligand desolvation contribution in implicit solvent and a torsional contribution which is related to a ligand entropic penalty¹⁷⁵ as in Equation 2.10.

$$\begin{aligned} \Delta G^\circ = & W_{vdw} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^6} \right) + W_{hbond} \sum_{i,j} E(\theta) \left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}} \right) \\ & + W_{elec} \sum_{i,j} \frac{q_i q_j}{e(r_{ij}) r_{ij}} + W_{sol} \sum_{i,j} (S_i V_j + S_j V_i) e^{\frac{-r_{ij}^2}{2\sigma^2}} + W_{conf} N_{tor} \end{aligned} \quad (2.10)$$

where each weight W_i is determined by fitting to experimental binding affinities coming from a training set of crystal structures of protein-ligand complexes.¹⁷⁵ In the above equation, the hydrogen-bond term is a 12-10 Lenard-Jones potential where C and D are equation parameters encoding the well depth.¹⁷⁵ The $E(\theta)$ function accounts for hydrogen-bond directionality based on the angle θ . The desolvation potential is based on the volume of atoms (V) that surround a given atom and protect it from the solvent weighted by a solvation parameter (S) and an exponential term with a distance-weighting factor σ .¹⁷⁵ The entropy term corresponds to the loss of ligand flexibility upon binding and is proportional to the number of ligand rotatable bonds (N_{tor}). The AutoDock4¹⁷⁵ scoring function introduces a significant advantage: consideration of the desolvation effects, even if approximated, is carried out without introducing a significant increase in calculation cost because only one conformation of the complex is taken.¹⁷⁵

2.3.6.3: Knowledge-based Scoring Functions

Knowledge-based scoring functions are constructed from structural information contained in experimentally determined protein-ligand complexes, such as those available on the PDB database. From these structures, pairwise contacts between ligand and protein atoms are identified and used to derive distance-dependent pairwise statistical potentials.^{181,193} As such a “free-energy”-like contribution is computed for each pairwise contact, which can then be summed to yield a score.^{181,193} Examples of currently available knowledge-based scoring functions are DrugScore, PMF and IPMF.^{112,181,187,193} An example of a knowledge-based scoring function is the DrugScore scoring function.¹⁹³

In DrugScore, statistical potentials encoding protein-ligand pairwise atomic interactions were derived from a set of 6026 PDB structures containing ligands whose parametrization was according to the SYBYL convention.¹⁹³ The dataset was pruned by removing crystal structures with: a resolution above 2.5Å, ligands containing less than 6 or more than 50 non-hydrogen atoms and non-druglike or covalently bound ligands.¹⁹³ Solvent contributions were introduced using a potential proportional to the solvent-accessible surface (SAS) of the protein and ligand atoms which become buried upon complexation.¹⁹³ This term is computed using the van der Waals radii of each atom as defined in the TRIPOS forcefield¹⁹¹, with the exception of oxygen and nitrogen, for which the radii were reduced by 0.2Å (**Equation 2.11**).

$$\Delta W = \gamma \sum_i \sum_j \Delta W_{i,j}(r) + (1 - \gamma) * \left[\sum_i \Delta W_i(SAS, SAS_0) + \sum_j \Delta W_j(SAS, SAS_0) \right] \quad (2.11)$$

where the sums run through atoms i in the ligand and j in the protein separated by a interatomic distance between r and $r + dr$, $\Delta W_{i,j}(r)$ is difference distance-dependent potential between atoms i and j with respect to the value in the training set, ΔW_i corresponds to the difference in the one-body potential of the SAS for atom i with respect its value in the training set and γ corresponds to an empirical parameter which was set to 0.5.¹⁹³

2.3.6.4: Small comparison of multiple docking softwares

The ability of DrugScore to produce near native binding poses was compared to that of FlexX¹⁷⁸, a standard docking program at the time. The test cases for DrugScore¹⁹³ were

two: a dataset of 91 protein-ligand complexes extracted from the FlexX docking program validation set and 68 protein-ligand complexes whose ligand properties match those of the ligands from the FlexX validation set. Employing FlexX, to the first test set, in 54% of the cases the RMSD of the first ranked binding pose was found below 2Å of the native pose. When DrugScore was employed, in 73% of the cases the best ranked binding pose had an RMSD below 2Å of the native pose. In the second set, the performance is comparable (93 and 92%, for FlexX and DrugScore respectively). As such, for this dataset the power of DrugScore in identifying native poses is highlighted, when compared to FlexX which is a fragment-based docking algorithm.

A comparison study among several scoring functions including VINA¹⁷⁵, LUDI¹⁹⁰, GLIDE¹⁴² and PLP¹²⁰, as well as knowledge-based scoring functions shows modest predictability in the best case scenario and illustrates that molecular docking scoring functions correlate poorly with experimental binding affinities. Thus, at the molecular docking stage, it appears that production of reliable scores is still a challenge.

Part II – Theory

3. Binding Free Energy calculations

3.1: Introduction

Protein-ligand binding is a complex phenomenon governing many biological processes fundamental for life. Understanding of protein-ligand binding is critical to many areas of science, ranging from structural biology to pharmacology.³⁸⁰ One of the fundamental factors driving protein-ligand binding is the association constant or its inverse, the dissociation constant, K_a and K_d respectively.¹⁹⁴ It is intimately connected to the free energy change upon complexation, informing on the stability of the binding reaction. Accurate calculation of protein-ligand binding affinities, or binding free energies, would open to a detailed understanding of molecular association processes and supply powerful information to guide drug design programs.¹⁹⁵ Many approaches to estimate the binding affinity for a molecular association reaction have been proposed over the last decades.^{61,87,98,115,180,195–200} However, the calculation of the protein-ligand binding affinities remains a fundamental challenge in computational chemistry, from a theoretical and practical point of view.^{189,195}

Class of Methods	1. Absolute Chemical Potentials	2. Ligand Partition Equilibrium	Focus	Context (No. compounds)
Rigorous (week ⁻¹)		DDM FEP/PMF DAM	Full Reaction Path	lead optimization (10 ⁴ -10 ⁵)
End-points (day ⁻¹)	QM/MM quasi-harmonic MM/PBSA MM/GBSA one-average	QMLIECE LIE LIE(α,β)	Bound & Unbound States	hit-to-lead (10 ³ -10 ⁴)
Empirical (sec ⁻¹)		LIECE Dock AutoDock FF Böhm Fresno ES	Bound State	hit identification (10 ⁴ -10 ⁶)

Accuracy ↑ (red arrow) and Efficiency ↓ (green arrow) are indicated on the left and right sides of the table, respectively.

Figure 3.1 – Distribution of methods for binding free energy calculations according to calculation efficiency and accuracy, taken from “Computational Approaches to the Chemical Equilibrium Constant in Protein-Ligand binding” by Montalvo & Cecchini, 2016. Molecular Informatics, 11-12, 555-567.¹⁸⁹ Copyright @ 2016 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim.

The degree of difficulty in achieving accurate calculations is proportional to the complexity of the studied system. For example, proper calculation of protein-ligand binding affinities when a large conformational rearrangement of the protein structure occurs remains problematic with current techniques and hardware. In the case that binding of the ligand does not involve large conformational transitions of the protein,

these calculations can be carried out with remarkable levels of accuracy in some cases.^{159,189,194,201} The existing methods for binding free energy calculations can be spread out in a spectrum which balances calculation accuracy and computational efficiency, as shown in **Figure 3.1**.^{189,380}

On one side of the spectrum, methods relying on empirical scoring functions allow the evaluation of the binding affinity in seconds. These ultra-fast methods are popular because they are efficient and user-friendly. However, their computational efficiency is achieved at the expense of introducing many approximations to simplify the calculation and by only considering the bound complex in the calculation.^{31,111,175,177,180}

On the other side of this spectrum are rigorous free energy methods, where the binding affinity is computed by constructing a path from the bound complex towards the unbound species via the simulation several overlapping intermediate states.^{123,202–205} Some of the most well-known rigorous methods for binding free energy calculations are alchemical transformations by means of Free Energy Perturbation (FEP), as developed by Jorgensen *et al.*,^{204,206,207} and geometrical transformations along a reaction coordinate, also known as Potential of Mean Force (PMF)²⁰⁰, like in the Attach-Pull-Release (APR) method by Gilson *et al.*^{208,209} In the first case, the ligand is alchemically decoupled from the receptor and the solvent by means of perturbation theory whereas in the second case, the ligand is physically pushed or pulled away from the binding site along a path described by a reaction coordinate.^{208,209} Rigorous methods are the most accurate techniques available, exhibiting root mean squared errors (RMSE) to experimental binding affinities around 1 kcal/mol in some cases.^{202,205,210} However, since rigorous methods require extensive sampling they remain computationally costly and unamenable to virtual screening campaigns.

A third type of approach, known as end-point methods, represents a trade-off between accuracy and calculation efficiency.²¹¹ In end-point methods, conformational sampling is carried out only on the end-states of the binding reaction (the complex and the unbound species), and the computational cost is further reduced by computing the solvation free energy of each species of the binding reaction using a continuum solvation model or using the linear response approximation.^{115,189} End-point methods remain popular in the drug discovery community because they often are more accurate than typical docking scoring functions while maintaining a low computational cost when compared to full rigorous binding free energy calculations.^{115,196,201,211–214} However, it has been reported that these methods exhibit target-dependent performances.²⁰¹

Popular end-point methods are those belonging to the Molecular Mechanics (MM) Poisson Boltzmann (PB) Surface Area (SA) (MM/PBSA) family of techniques, proposed by Peter Kollman and collaborators^{89,115,196,201,212,213}, and the Linear Interaction Energy (LIE) method by Åqvist.^{115,189,215–217} In LIE, the solute/solvent interactions are considered

explicitly and the binding free energy is computed by evaluating the changes in electrostatic and van der Waals interactions between the ligand within the protein or solvent environment.^{189,215–217} In MM/PBSA, for each species of the binding reaction three terms must be computed from the MD simulations carried out: A potential energy term which is accessed in vacuum, a solvation free energy term which is estimated by employing an implicit solvent model and an entropic contribution which usually is estimated under the Rigid Rotor Harmonic Oscillator (RRHO) approximation.^{115,196,201,214,218}

In any of the above methods, the fundamental pre-requirement is the existence of an ensemble of molecular configurations from which thermodynamic observables can be quantified. Molecular simulations have a long history of application to biomolecular systems, starting from the MD simulations by Alder and Wainwright in 1957²¹⁹ and 1959²²⁰ until current days, where MD simulations are being used to compute binding affinities, to probe transition events and to study the basis of protein function and dynamics.^{31,115,122,124,218,221} Molecular Dynamics simulations and their application in binding affinity predictions are at the heart of the work produced in this thesis and, as such, will be described in the next subchapters.

Since molecular docking based approaches are known to correlate poorly to the experimental binding affinities⁸⁴, if the objective is to quantify the binding affinity of a ligand to a protein one must select more rigorous methods. The choice of method is carried out by considering the available resources and the type of project. For instance, if one is interested in small modifications of a scaffold leading to a chemical series or in obtaining quantitative agreement with experimental data, then rigorous methods are the best choice. If, on the other hand, the goal is compound ranking (or re-ranking) towards prioritization in a VS campaign context, then end-point methods represent a reasonable trade-off between accuracy and efficiency. Indeed, the approach followed in **Chapter 6** employs a docking plus free energy rescoring by MM/GBSA since the goal of the VS campaign was producing an accurate compound ranking, and not absolute predicted binding affinities.

3.2: Statistical Mechanics definition of the protein-ligand binding affinity

In the context of SBDD campaigns, molecular docking approaches provide scores which can be compared to experimental determinations of protein-ligand binding affinities. However, since these correlate poorly, the docking score is best regarded as a filtering tool as stated before. To produce reasonable predictions of binding affinities, more rigorous methods for their numerical estimation are required.¹⁹⁷ As described in **Chapter**

1, ITC experiments are costly and time consuming. Thus, it is desirable to obtain binding affinities without having to go through experimental determinations, but through numerical approaches. Doing so reduces the financial and human costs compared to experiments while profiting from the tremendous increase in computational power which, over the years, has improved prediction throughput.^{31,194,222} Furthermore, numerical approaches employing molecular modelling allow to obtain an atomistic view of the binding process in addition to an estimation of K_d .²²³ As such, it is important to understand the connection between statistical mechanics and K_d . Statistical mechanics is a branch of science which connects events happening in microscopic world to macroscopic phenomena.²²⁴ It defines a number of rules from which one can derive the thermodynamic properties of a system. In the context of a closed system, which only exchanges heat with the outside world, the temperature, the number of particles and the volume inside it are constant.²²⁴ Starting from this isothermal-isochoric ensemble or *canonical* ensemble (NVT), we can define a fundamental quantity for statistical mechanics, the *partition function* Q .²²⁴ The *partition function* encodes all possible configurations the system may explore in the conditions set, and is shown in **Equation 3.1**.³⁸⁰

$$Q = \frac{1}{N! h^{3N}} \int e^{-\beta H(q,p)} d^3q d^3p \quad (3.1)$$

where h is the Plank constant, β is the Boltzmann factor given by $\frac{1}{k_B T}$, $H(q,p)$ is the Hamiltonian operator of the system encoding momenta (p) and position (q) of all particles in the system, d^3 indicates that q and p are vectors in $3N$ dimensional space and $N!$ is introduced to avoid over-counting the number of microstates.²²⁴ From this equation one can derive the configurational partition function Z_N in Cartesian coordinate space, which is of the form:^{189,380}

$$Z_N = \frac{1}{N! h^{3N}} \int e^{-\beta E_i(q)} d q \quad (3.2)$$

where the integration is carried out over the position of all particles in the q configurational space and $E_i(q)$ is the energy of a given configuration as a function of the atomic positions. The contribution coming from the momenta, in the case that free energy differences are of interest, is expected to cancel out.²²⁵ From the above **Equation 3.2**, one is able to extract important thermodynamic quantities. For example, it is possible to obtain the internal energy of the system (U) by taking the partial derivative of the

natural logarithm of the partition function with respect to β and connect it to the configurational ensemble sampled by MD as in **Equation 3.3**:³⁸⁰

$$\frac{\partial \ln Z_N}{\partial \beta} = \langle E_i(q) \rangle = U = \int E_i(q) P_i(q) dq \quad (3.3)$$

where $\langle E_i(q) \rangle$ is the ensemble average energy of the system extracted from the i configurations sampled by MD or MC simulations in the limit of the ergodic hypothesis.²²⁴ For a system containing many atoms, computing manually the configurational partition function becomes intractable. However, it can be estimated from molecular simulations using **Equation 3.3** where $P_i(q)$ is the probability of configuration i , extracted from the Boltzmann distribution (also known as the *canonical* distribution) as in **Equation 3.4**.^{189,224}

$$P_i(q) = \frac{e^{-\beta E_i(q)}}{Z_N} \quad (3.4)$$

Additionally, from the configurational partition function one can obtain the free energy of the system in the NVT ensemble using the Helmholtz equation (**Equation 3.5**):

$$F = U - TS = -K_B T \ln Z_N \quad (3.5)$$

where U is the system internal energy, TS is the system entropy at temperature T in Kelvin and K_B is Boltzmann's constant.²²⁵ The Helmholtz free energy is quasi-equivalent to the Gibbs free energy, which is derived in the isothermal-isobaric (NPT) ensemble. The enthalpy term (H) in the Gibbs free energy (G) corresponds to the internal energy (U) plus a pV term accounting for the volume (V) variation with respect to the fixed pressure (p), as in $G = H - TS = U + pV - TS$. The pV term, however, can be approximated as having zero contribution to the change in free energy in solution when considering incompressible liquids like water, as pointed out by Shirts and Mobley.²²⁶

The free energy corresponds to the maximum amount of useful work that can be extracted from a closed system.²²⁴ It can be evaluated either at constant temperature and volume (Helmholtz) or at constant temperature and pressure (Gibbs). To compute accurate free energies from molecular simulations, it is first required that all configurations accessible to a system be visited and that their probabilities be converged (*e.g* reversibly sampling all conformational transitions). The sampling of the free energy landscape can be done either through MD simulations or MC methods. Then, we employ **Equation 3.6** (NVT) as⁷⁸:

$$F = -K_B T \ln \sum_i e^{-\beta E_i(q^N)} \quad (3.6)$$

Analysis of **Equations 3.4** and **3.6** allows us to grasp one very important piece of information. The molecular configurations which come from low energy regions of the free energy landscape will contribute more to the free energy than configurations coming from high energy regions because they correspond to the configurations which are sampled more frequently. Indeed, **Equation 3.4** connects the macroscopic with the microscopic world by connecting U to the energy of each individual microstate. However, in most cases perfect sampling is unachievable through simulations since sampling reversibly all high-energy conformations requires simulation time-scales orders of magnitude above what is currently possible. One way to try and circumvent this limitation is through the usage of enhanced sampling techniques such as replica exchange Molecular Dynamics (REMD).^{227,228} The underlying reason is that populating high-energy states in unbiased MD implies crossing large energy barriers and the corresponding time τ for these conformational transitions to happen by thermal fluctuation is large. However, it must also be noted that while enhanced sampling techniques tend to sample the configurational space better than unbiased MD, they can be difficult to apply and may still be plagued with under-sampling issues. Thus, accurate estimation of absolute binding free energies remains a challenge.

Finally, it is often the case that one is not interested in absolute free energies but in the free energy difference between two states, whichever they are. Given the above and that the free energy of binding is obtained from the ratio of concentration of each chemical species, then one can write the K_{eq} as a ratio of the partition functions of the complex and the unbound protein and ligand.⁸⁶ Indeed, molecular simulations are a powerful technique which underlie most free energy calculation methods.^{110,200,201,209,221,226,229,230} However, some ingredients are necessary to compute accurately free energies: a potential energy function, a sampling protocol to obtain the configurational ensemble and a method for binding free energy calculation. In the next subchapter we will discuss the fundamental concepts required for Molecular Dynamics simulations. Then, we will discuss rigorous and end-point binding free energy methods.

3.3: Molecular Dynamics simulations

3.3.1: Molecular Mechanics

Throughout this dissertation, the most used CADD techniques require MD simulations. Particularly, most results were obtained based on unbiased MD using classical forcefields, which rely on a Molecular Mechanics formalism.¹¹⁸ Within a quantum mechanical description of molecules, the behavior of all electrons is explicitly considered, allowing

the study of phenomena involving bond breaking and generation. However, the time-dependent Schrödinger equation²³¹ only has analytical solutions for molecular systems of one electron meaning that numerical approaches within QM have to be pursued to study more complex systems.

The Molecular Mechanics (MM) description of large systems is substantially cheaper than the QM treatment but requires the introduction of several approximations. One approximation is the Born-Oppenheimer²³² (BO) approximation. The BO approximation relies on the fact that the electrons move much faster than nuclei do, adapting almost instantly to any position of the nuclei, thus meaning that MM forcefields focus on the motions of atomic nuclei.²³² In Molecular Mechanics, the potential energy of the system is written as a function of the coordinates of atomic nuclei, and thus most MM approaches are unable to reproduce bond breaking/forming events. An example of a MM forcefield which is able to efficiently describe bond breaking/formation is the ReaxFF.²³³ Other approximations are also introduced, such as the treatment of atoms as hard spheres with fixed volume and the almost-complete neglect of electronic polarization phenomena.¹⁹⁴ To study large molecular systems within Molecular Mechanics, it is necessary to use a potential energy function, also known as a forcefield.¹¹⁸ A forcefield is an equation comprised of bonded and non-bonded terms which permits the calculation of the potential energy of a system in function of its three-dimensional molecular configuration.¹¹⁸ Since forcefields are parametric in nature, it is desirable that their parameters, derived from fitting to reference data, be transferable to larger and more complex systems for which reference data does not exist.¹¹⁸ Because parameters are optimized by fitting to reference data, some forcefields are better at reproducing some properties than others.¹¹⁸ This is the reason why there are forcefields specifically developed for DNA, proteins or lipid simulations. Over the years, many forcefields have been developed with the aim of accurately reproducing the dynamical behavior of biomolecules. Some of the most well-known are AMBER¹¹⁸, CHARMM²³⁴, OPLS²⁰⁶ and GROMOS.²³⁵ Throughout this dissertation we will use the AMBER forcefield.¹¹⁸ The functional form of the forcefield equation for AMBER is described in **Equation 3.7**.

$$\begin{aligned}
 U(r) = & \sum_{bonds} K_b(b - b_0)^2 + \sum_{angles} K_\theta(\theta - \theta_0)^2 + \sum_{dihedral} \frac{K_\chi}{2}(1 + \cos(n\chi - \delta)) \\
 & + \sum_{non-bonded} \left\{ \left[\epsilon_{ij}^{min} \left(\frac{R_{ij}^{min}}{r_{ij}} \right)^{12} - 2 \left(\frac{R_{ij}^{min}}{r_{ij}} \right)^6 \right] + \frac{q_i q_j}{4\pi\epsilon_0\epsilon R_{ij}} \right\} \quad (3.7)
 \end{aligned}$$

where r_{ij} is the distance between atoms i and j , R_{ij}^{min} is the reference distance computed at the energy minimum structure, q_i and q_j are the atomic partial charges of

atoms i and j , ϵ_0 is the dielectric constant in vacuum, ϵ is the solute internal dielectric constant of the system, ϵ_{ij}^{min} is the depth of the well in the van der Waals potential and each K_x parameter corresponds to the force constant applied to that specific term of the forcefield.¹¹⁸ The bonded terms, encoded by the three first terms, describe the dynamical motions of covalently bound atoms. These are ruled by harmonic potentials in the case of bonds and angles, whereas the dihedral angles are described by a sinusoidal function.¹¹⁸ The bonds and angles' associated force constants are high, to ensure that these fluctuate only slightly around the reference value.¹¹⁸ The dihedral angles, however, are allowed to fluctuate and span any value between 0 and 360°, depending on the height of the energy barriers separating each conformation from the others. Since dihedral degrees of freedom (χ) are periodic, a simple harmonic potential is unsuitable to describe them. Thus a sinusoidal function is used to describe torsions, which depends on the periodicity n and the phase δ .¹¹⁸ The phase dictates the location of the maxima in the potential energy surface of the dihedral. The periodicity indicates the number of cycles along a 360° rotation around the dihedral.¹¹⁸

The non-bonded interactions are specified by the final two terms. The first one encodes the van der Waals interactions and takes the shape of a Lennard-Jones potential.²³⁶ This potential is a combination of dipole-dipole interactions of attractive nature (the R^6 term) and repulsive interactions (the R^{12} term).¹¹⁸ The R^{12} term has been shown to reproduce the expected steep increase in energy when two atoms' electronic clouds overlap too much and repel each other.¹¹⁸ As the distance increases between atom pairs, these interactions quickly become negligible and it is customary to truncate them to a given cut-off value¹¹⁸ using a switching function to lead the interaction towards zero as the interatomic distance between the atom pair approaches a threshold.¹¹⁸

The final term, encoding electrostatic interactions, is described using a Coulomb law between particles with fixed atomic charges.^{103,118} Due to the long-range nature of coulombic interactions, introducing truncation thresholds would lead to a significant error which is non-trivial to work around.¹¹⁸ Thus, it would seem that to compute electrostatic interactions one would need to compute them at each integration time step as the position of the particles of the system are updated iteratively during simulation. Since these depend on the distance between each pair of point charges, a list containing all pairwise interatomic distances would need to be computed at each time step and the computational cost would increase significantly.²³⁷ To tackle this problem, typically the Particle Mesh Ewald (PME) algorithm is used under periodic boundary condition, splitting the electrostatic contributions into short-range and long-range interactions.²³⁷ The short-range portion is computed with a direct pairwise summation and the long-range part is computed by the PME algorithm using a distance cut-off.^{118,237}

3.3.2: Integrating Newton's Equations of motion

The Molecular Mechanics formalism allows both to calculate single-structure properties and to simulate the time-evolution of the conformational dynamics of the system by integrating Newton's second law of motion at regular time intervals using Equation 3.8¹³³

$$F = m\ddot{x} = -\frac{\partial U}{\partial x_i} = m_i \frac{d^2 x_i}{dt^2} \quad (3.8)$$

where $-\frac{\partial U}{\partial x_i}$ corresponds to the partial derivative of the energy with respect the position and m_i is the mass of the particle i .¹¹⁸ During a simulation, the forces acting the system degrees of freedom are updated at each time-step, updating the position of the atoms according to particles' velocities¹³³ and thus producing a trajectory.²²⁰ To guarantee that Newton's second law of motion is correctly solved, integrators need to have some properties, among which conservation of energy and time reversibility.¹³³ Time reversibility means that changing the sign of all velocities and momenta of the particles in the system will cause the system to retrace its steps, like it is going back in time.¹³³ It is important to note that no integrator is perfect, and they accumulate errors¹³³ meaning that time reversibility is only possible over short periods of time. Integration of the dynamics is typically carried out using well-established finite-difference algorithms, such as the one by Verlet²³⁸ in 1967 or the Velocity Verlet²³⁹ algorithm, a variation of the original algorithm by Wilson and colleagues.²³⁹ These algorithms are computationally efficient, reversible and respect the law of energy conservation. The Verlet algorithm²³⁸ is one of the most common integrators in MD simulations. It is based on two Taylor expansions, one in the forward and one in the reverse direction such that when these two are combined, it yields^{133,238}

$$r_{n+1} = 2r_n - r_{n-1} + \frac{F_n}{m} \Delta t^2 + O(\Delta t^4) \quad (3.9)$$

where r_n is the position of the atoms at time t , $\left(\frac{F_n}{m}\right) \Delta t^2$ is the acceleration at time t , $O(\Delta t^4)$ are the terms of order $O(\Delta t^n)$ and r_{n+1} and r_{n-1} are the position of the atoms in the next and the previous step at time $t + \Delta t$ and $t - \Delta t$.¹³³ The algorithm proceeds iteratively with the following set of commands:^{133,238} First, from the position r_n , compute the force F_n acting on all atoms. Then, from the positions in the previous and current step, r_n and r_{n-1} , and the just computed force F_n , update the atomic positions to the

new timestep r_{n+1} .¹³³ However, the Verlet algorithm is known to accumulate large errors.¹³³ Thus, some modifications to the Verlet algorithm have been proposed.

The velocity Verlet scheme is one of them and was used during the MD simulations carried out in this dissertation. It can be written as a set of equations¹³³ (**Equations 3.10** to **3.13**).

$$r_{n+1} = r_n + v_n \Delta t + \frac{1}{2} \left(\frac{F_n}{m} \right) \Delta t^2 \quad (3.10)$$

$$v_{n+1} = v_n + \frac{1}{2} \left(\frac{F_n}{m} + \frac{F_{n+1}}{m} \right) \Delta t \quad (3.11)$$

$$v_{n+1/2} = v_n + \frac{1}{2} \left(\frac{F_n}{m} \right) \Delta t \quad (3.12)$$

$$v_{n+1} = v_{n/2} + \frac{1}{2} \left(\frac{F_{n+1}}{m} \right) \Delta t \quad (3.13)$$

The algorithm proceeds in an iterative fashion. First it computes the atomic positions r_{n+1} at time $t + \Delta t$ from the above **Equation 3.10**.¹³³ Then, with **Equation 3.12**, it calculates the velocity at the step $v_{n+1/2}$ from the force F_n acting on the system atoms. From the positions and velocities calculated, the forces acting on the particles are updated to the next step $t + \Delta t$ (F_{n+1}).¹³³ Finally, using the recently computed force, it is able to estimate the new velocities v_{n+1} at time $t + \Delta t$ using **Equation 3.13**.¹³³ The velocity-Verlet algorithm is known to provide an accurate estimate of velocities. Furthermore, it is stable and respects both the time-reversibility of Newtons' equations of motion and the energy conservation law.¹³³

Another critical component for the simulations to be both stable and physically reasonable is the time-step of integration, which must allow capturing the vibrational motions of the fastest degrees of freedom of the system.¹³³ For biomolecular systems such as proteins, this corresponds to the heavy atoms-hydrogen vibrational frequency (X-H stretching), which is approximately 3000 cm^{-1} .¹³³ Thus, an appropriate time step should be between 0.5^{133} and 1 fs .¹¹⁸ In most cases, however, these motions are constrained by algorithms like SHAKE²⁴⁰ or LINCS²⁴¹, leading to a significant gain in simulation efficiency by freezing the X-H stretching. The fastest vibrational mode becomes the X-X stretching, which is on the order of 1500 cm^{-1} ,¹³³ representing a gain in integration time step of 2-fold, pushing to the range to between 1^{133} and 2 fs .¹¹⁸

The SHAKE algorithm compares the length of each X-H bond with the reference value at each time-step and adjusts the position of hydrogen atoms should this deviation be larger than a threshold.²⁴⁰ Other methods which allow going beyond 2 fs integration time-

steps exist, such as the Hydrogen Mass Repartitioning scheme²⁴², where the mass of the heavy atoms covalently bonded to hydrogen atoms is distributed among the hydrogens. By doing so, the vibrational frequency of the hydrogen atoms becomes lower since their molecular weight increases.²⁴² This methodology allows all-atom simulations to be carried out using a 4 fs integration time-step while maintaining a physically sound dynamical behavior.²⁴²

3.3.3: Thermostat and Barostat for MD simulations

Application of MD simulations to estimate binding affinities in biological systems requires the reproduction of experimental conditions. The simulation setup must then be run in the NPT ensemble, meaning that temperature and pressure control must be exerted during dynamics. This is especially important because ITC experiments, which is the current gold standard method for experimental determination of binding affinities, are conducted at a standard of 298.15K and 1 bar pressure. Temperature control during MD is achievable through the introduction of a thermostat, like the one proposed by Langevin²⁴³ or by Nosé-Hoover²⁴⁴, and pressure control through the inclusion of a barostat. Langevin dynamics is a thermostating approach where the equations of motion are modified through a friction term, affecting the atomic velocities, plus a random force. The new dynamics for a system are given by **Equation 3.14**²⁴³

$$m_i \frac{d^2 x_i}{dt^2} = -\nabla x_i U - \gamma_i \frac{dx}{dt} + L(t) \quad (3.14)$$

where γ_i is a friction coefficient applied to atom i of the system and $L(t)$ is the Langevin random force. The friction coefficient will decrease particle velocity, leading to improved numerical stability while the random force is a type of “noise” which could help the sampling of the free energy landscape.²⁴⁵ To maintain constant pressure, one must also introduce a barostat. Commonly used barostats include the Berendsen²⁴⁶ or the Monte Carlo barostat²⁴⁷, recently implemented in the Amber18 simulation package.

3.4: Rigorous free energy methods

Rigorous binding free energy calculation methods are able, at a significantly high computational cost, to produce binding affinity predictions in general agreement with experimental data.^{189,195,197,209} The term rigorous is employed here to define those methods which are based on simulating a path to bring the system from the unbound

protein and ligand towards the bound complex. Among these, two major types of approaches exist: alchemical and geometrical free energy calculations.¹⁸⁹ Free energy calculations are known as alchemical when the free energy difference between two systems is computed by slowly transforming one system into the other by means of not necessarily physical intermediates, evaluating the free energy contribution of each non-physical intermediate along an alchemical path.²¹⁰ One application relates to the study of how the modification of a chemical group in a series of compounds affects the affinity towards the receptor. Geometrical binding free energy calculations, on the other hand, rely on simulating a physical path which connects the bound and unbound states of the protein-ligand complex.^{208,209} The binding partners are then pulled together or pushed away by adding a potential (or force) onto a selected collective variable (CV).^{208,209} The free energy profile, or PMF, of the system with regards to the selected CV is then constructed, allowing one to evaluate the binding affinity.^{208,209} In the following, some of the most popular approaches for binding free energies are explored: Double Decoupling Method (DDM) based on FEP, LIE and MM/PBSA methods. While the focus on this dissertation is not on rigorous binding free energy methods, it was deemed appropriate that the current gold-standard method for rigorous binding, FEP, be described.

3.4.1: Free Energy Perturbation – Double Decoupling

Within alchemical calculations, the protein-ligand binding free energy can be evaluated by the double decoupling, or double annihilation, method as introduced by Jorgensen.⁹⁸ In DDM, the ligand is slowly annihilated, becoming a non-interacting particle in the binding site and in solution and the binding free energy is computed by means of a thermodynamic cycle (**Figure 3.2**).²⁴⁸ To enable these transformations, a parameter, λ , is introduced, building a hybrid Hamiltonian scheme linking the end-states of the alchemical transformation such that

$$H_{\lambda} = (1 - \lambda)H_{\mathbf{A}} + \lambda H_{\mathbf{B}} \quad (3.15)$$

where $H_{\mathbf{A}}$ is the Hamiltonian of the system in the starting state and $H_{\mathbf{B}}$ is the Hamiltonian of the system in the end state. For example, consider the annihilation of the ligand from the protein binding site. In this case, $H_{\mathbf{A}}$ corresponds to the ligand-bound conformation of the system and $H_{\mathbf{B}}$ to the protein in the unbound state, after the ligand has been decoupled. To connect these two end-states, the parameter λ is introduced. It takes values ranging between 0 and 1 and, in practice, permits the transformation by gradually turning off the van der Waals and electrostatic interactions between the protein and the ligand in small, discrete steps known as intermediate states.³⁰ Introducing an additional

intermediate state to transfer the ligand to the gas-phase from either solution or the binding site, one obtains¹⁸⁹

$$K_{eq}C^0 = \frac{\int_{\text{site}} dL \int dR \exp(-\beta U_A)}{\int_{\text{bulk}} dL \delta(r_L - r^*) \int dR \exp(-\beta U_B)} x \frac{\int_{\text{bulk}} dL \delta(r_L - r^*) \int dR \exp(-\beta U_B)}{\int_{\text{bulk}} dL \delta(r_L - r^*) \int dR \exp(-\beta U_A)} \quad (3.16)$$

where the first factor in **Equation 3.16**¹⁸⁹ on the right is the reversible work of decoupling the ligand from the binding site into the gas-phase and the second factor is the reversible work of decoupling the ligand from solution into the gas-phase.¹⁸⁹ The terms L and R correspond to the coordinates of the ligand atoms and the coordinates of the atoms of the remaining elements in the simulation box respectively, β is the inverse of the Boltzmann constant, U_A is the potential energy of the system with a fully coupled ligand and U_B the potential energy of the system with a fully uncoupled ligand.¹⁸⁹

From a statistical mechanics point of view, one can thus write the binding free energy in the NPT ensemble of going from one state A to another B as in Zwanzig's exponential formula²²⁶

$$\Delta G_{AB} = -K_B T \ln \langle e^{-\beta(H_B - H_A)} \rangle_A \quad (3.17)$$

In the above **Equation 3.17** we start from microstates from the conformational ensemble of state A and use it to compute the partition function of B by changing the potential energy function from U_A to U_B . It allows one to connect the change in free energy when the system goes from state A to state B with the small sequential perturbations introduced in the Hamiltonian.²⁵⁰ The limitation associated with this approach is that it requires that B and A be states which are not too different from one another. Otherwise, if the configurations of A are of low probability in B , the difference in free energy between the states will be large and thus attaining convergence is difficult.^{230,250} In other words, it means that the configurations sampled in A have a low probability of being sampled in B , implying poor simulation overlap. In this case, it is possible to exploit the notion that the free energy is a state function and thus build a path connecting the two states using a series of intermediates. The binding free energy is then computed by summing the free energy contribution of each intermediate state introduced along the path.¹⁹⁷ However, if no other terms are considered, then the ligand in bulk solvent or when fully decoupled from the protein is allowed to wander along the simulation box, a phenomenon which is known to significantly hamper convergence of FEP calculations.¹⁸⁹

To improve convergence, modern approaches attach restraints to the ligand such that its accessible configurational space is strongly diminished and statistical convergence is improved.²⁴⁸ In particular, the ligands' position, orientation and configuration must be restrained both within the binding site and in bulk solution.¹⁸⁹ Introducing harmonic restraints on the ligand position (μ_p), orientation (μ_o) and conformation (μ_c) yields a modified version of **Equation 3.17**, where the equilibrium constant, and thus the binding free energy, is accessed by computing the contributions arising from each of the eight integrals in **Equation 3.18** as described by Montalvo-Acosta and Cecchini¹⁸⁹

$$\begin{aligned}
K_{eq}C^0 = & \frac{\int_{\text{site}} dL \int dR \exp(-\beta U_1)}{\int_{\text{site}} dL \int dR \exp[(-\beta(U_1 + \mu_c))]} x \\
& \frac{\int_{\text{site}} dL \int dR \exp[(-\beta(U_1 + \mu_c))]}{\int_{\text{site}} dL \int dR \exp[(-\beta(U_1 + \mu_c + \mu_o))]} x \\
& \frac{\int_{\text{site}} dL \int dR \exp[(-\beta(U_1 + \mu_c + \mu_o))]}{\int_{\text{site}} dL \int dR \exp[(-\beta(U_1 + \mu_c + \mu_o + \mu_p))]} x \\
& \frac{\int_{\text{site}} dL \int dR \exp[(-\beta(U_1 + \mu_c + \mu_o + \mu_p))]}{\int_{\text{bulk}} dL \delta(r_L - r^*) \int dR \exp[(-\beta(U_0 + \mu_c + \mu_o + \mu_p))]} x \\
& \frac{\int_{\text{bulk}} dL \delta(r_L - r^*) \int dR \exp[(-\beta(U_0 + \mu_c + \mu_o + \mu_p))]}{\int_{\text{bulk}} dL \delta(r_L - r^*) \int dR \exp[(-\beta(U_0 + \mu_c + \mu_o))]} x \\
& \frac{\int_{\text{bulk}} dL \delta(r_L - r^*) \int dR \exp[(-\beta(U_0 + \mu_c + \mu_o))]}{\int_{\text{bulk}} dL \delta(r_L - r^*) \int dR \exp[(-\beta(U_0 + \mu_c))]} x \\
& \frac{\int_{\text{bulk}} dL \delta(r_L - r^*) \int dR \exp[(-\beta(U_0 + \mu_c))]}{\int_{\text{bulk}} dL \delta(r_L - r^*) \int dR \exp[(-\beta(U_1 + \mu_c))]} x \\
& \frac{\int_{\text{bulk}} dL \delta(r_L - r^*) \int dR \exp[(-\beta(U_1 + \mu_c))]}{\int_{\text{bulk}} dL \delta(r_L - r^*) \int dR \exp[(-\beta(U_1))]} x
\end{aligned} \tag{3.18}$$

Schematically, one can describe the thermodynamic cycle encoded as follows: gradual confinement of the position (μ_p), orientation (μ_o) and configuration (μ_c) of the ligand in the binding pocket, annihilation of the restrained ligand in the binding site, re-coupling of the restrained non-interacting particle from gas-phase into bulk solvent as fully interacting ligand with conformational restraints, gradual release of the ligand restraints in the solvent (**Figure 3.2**).^{189,248} The binding free energy is thus obtained by summing the free energy contribution arising from each intermediate state simulated along the transition from state A to state B. (**Equation 3.19**).

$$\Delta G_{AB} = -k_B T \sum_{i=1}^{N-1} \ln \left\langle \exp \left(-\frac{U(\lambda_{i+1}) - U(\lambda_i)}{k_B T} \right) \right\rangle \tag{3.19}$$

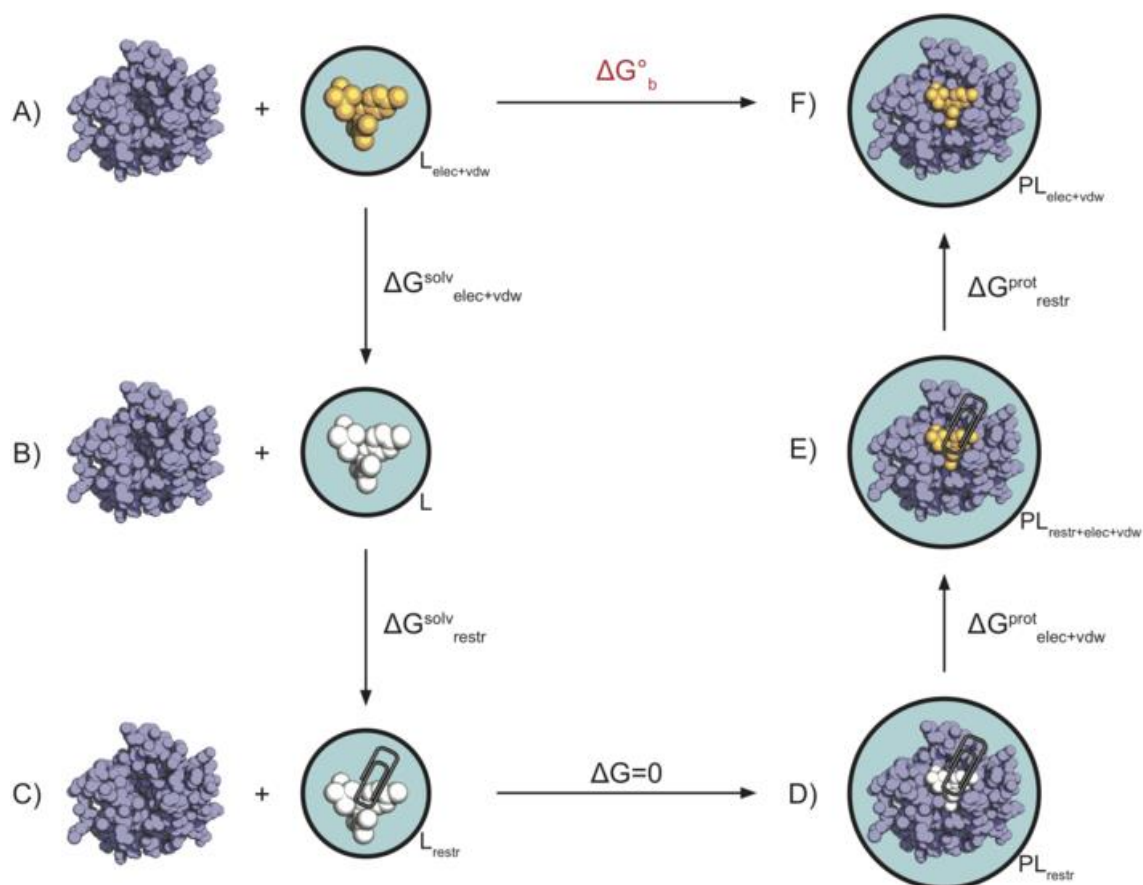


Figure 3.2 – Thermodynamic cycle for a DDM calculation. A) Protein and ligand unbound in solution; B) Protein in solution, ligand transformed into a non-interacting particle; C) Protein in solution, the non-interacting particle is restrained; D) Protein-ligand complex with the ligand as a restrained non-interacting particle; E) Protein-ligand complex with the ligand restrained in the binding site; F) Bound Protein-Ligand complex. Extracted from “Accurate calculation of the absolute free energy of binding for drug molecules” Aldeghi *et al.*, 2016.²⁴⁸ *Chemical Science*, 7, 207. Copyright @ The Royal Society of Chemistry.

A significant drawback of the FEP method is that most of the simulation time is spent on unphysical intermediates, which are non-meaningful but critical to the calculation since they establish the connection between the end-states of the binding reaction.^{189,248} The above mentioned scheme is the one used of *absolute* binding free energy calculations.^{123,210} However, it is also extensible to *relative* binding free energy calculations. To do so, instead of decoupling ligand A in the binding site and solution, the ligand A is alchemically transformed into another ligand, B. This means that as $\lambda_{0 \rightarrow 1}$, H_A approximates 0 and H_B approximates 1 and one ligand is converted in another by annihilating one portion and coupling another to it.²⁵¹ The intermediate states of the transformation contain a hybrid Hamiltonian where part of the interactions between ligand A and the protein are turned on and part of the interactions between ligand B and the protein are also turned on at any given point.²⁵¹ Ideally the ligands share the same chemical scaffold and the perturbation involves a small number of atoms such that the

calculation is computationally affordable.²⁵¹ It is typically useful in lead-optimization problems, where chemical modifications of a scaffold yield a chemical series worth investigating.¹⁹⁷ There are two schemes to carry out FEP calculations: the dual topology scheme or the single topology scheme. In the former, a topology containing atoms of the two ligands is supplied.²²⁶ At $\lambda = 0$, the atoms of ligand A are turned on and the atoms which are unique to ligand B behave as dummy atoms. As the calculation progresses and λ goes from 0 to 1, the interactions with particles which are unique to ligand A are gradually turned off and, at $\lambda = 1$ these atoms behave as dummy atoms. In the latter, the atoms corresponding to ligand A are transformed into the atoms of ligand B by changing the atom type as λ goes from 0 to 1.²²⁶

3.5: End-point methods for binding free energy calculations

In the context of SBDD and VS approaches, rigorous methods remain too costly for routine use in screening of large ligand libraries. Thus, end-point methods are used to provide an evaluation of the protein-ligand binding affinity at a fraction of the computational cost expected for rigorous methods.¹⁸⁹ These methods are simplified approaches with explicit consideration of the conformational dynamics of the protein-ligand complex and the unbound ligand and protein in solution.²⁰¹ Some end-point methods require a training test to build a linear regression model for binding affinity prediction²¹⁶, similarly to machine learning-based approaches, while others are training-set independent and rely on a physics-based functional form.^{211,229} Two of the most prominent end-points methods will be described in this section: the LIE method^{216,252,253} and MM/PBSA, alongside some variants.^{196,211,212,252,254} These are methods which have a long history of application to drug discovery in protein-ligand binding and in host-guest systems, collecting both successes and failures. In particular, the MM/PBSA family of methods can be used as a free energy rescoring scheme following molecular docking experiments as a way to evaluate the binding free energy of a protein-ligand complex in a more accurate manner and including solvation effects in the calculation via the use of an implicit solvent model.

3.5.1: The Linear Interaction Energy method

The original implementation of the LIE method is due to Johan Åqvist and co-workers, who in 1994 aimed at designing an approach to predict fast and accurately protein-ligand binding free energies.²¹⁶ At the time, the rigorous method which was applicable to this sort of problem was FEP but it was limited to very small perturbations because of the

computational overheads.²¹⁶ The idea behind LIE is to compute the absolute binding free energy of a ligand to a protein based on the electrostatic and van der Waals interactions established by the ligand in solution and in the protein environment as in **Equation 3.20**.²¹⁶

$$\Delta G_{bind} = \frac{1}{2} \langle \Delta U_{solvent-protein}^{elec} \rangle + \alpha \langle \Delta U_{solvent-protein}^{vdW} \rangle \quad (3.20)$$

where $\langle \Delta U_{lig-prot}^{elec} \rangle$ is the difference in the electrostatic energy taken from the ligand simulation in solution or in the protein media and $\langle \Delta U_{lig-prot}^{vdW} \rangle$ corresponds to the difference in the van der Waals energy taken from the ligand simulation in solution or in the protein media, α is an empirically determined parameter and $\langle \dots \rangle$ denotes an ensemble average.²¹⁶ Thus, a LIE calculation requires two simulations to be carried out, one of the solvated complex and one of the ligand unbound in solution.²⁵⁵

The $\frac{1}{2}$ weight for the electrostatic term was obtained by comparing the difference in the electrostatic contribution between vacuum and explicit solvent simulations of Na^+ and Ca^{2+} ions ($\langle \Delta U_{vacuum-solution}^{elec} \rangle$) to FEP calculations which computed the free energy of charging these ions in a water simulation box.²¹⁶ A factor of 0.49 and 0.52, respectively, was found between the two calculations, which justifies using a 0.5 weight for the electrostatic part.²¹⁶ For the nonpolar part, Åqvist and co-workers assumed that it could be approximated based on the vdW interaction energies and the empirical parameter α .²¹⁶ To support this approximation, the authors recall that the solvation free energy of hydrocarbons like n-alkanes depends linearly on the chain length both in the liquid form and in water.²¹⁶

In general, the parameter α is empirically determined so as to reproduce the experimental binding free energies of training set compounds.^{216,253} In the original article, benchmark calculations were carried out on a set of endothiapepsin (EP)-inhibitor complexes. The calibration set consisted of 4 complexes, for which the RMS error was determined at 0.39 kcal/mol with a α value of 0.169.²¹⁶ Impressively, the method was able to identify the low affinity and high affinity binders successfully.²¹⁶ To further test the method, Åqvist and co-workers selected a fifth inhibitor with significantly different chemistry from the calibration set compounds. The predicted absolute binding affinity for this compound exhibited quantitative agreement with the experimental data, as the authors report.²¹⁶ However, this study also had limitations which are, nonetheless, addressed by the authors. These mention as potential shortcomings the size of the dataset explored, the bigger weight given to the electrostatic contribution than to the vdW, the short simulation time pursued and the fact that α may not be transferable to other systems.²¹⁶ Furthermore, the $\frac{1}{2}$ factor in the electrostatic contribution term has a

physical meaning but some researchers have decided to treat it also as a free parameter (β), carrying out a two-parameter fit using a calibration set.^{253,256} The original LIE model then assumes that other effects like entropy and intramolecular energies cancel out by fitting,²⁵⁶ allowing to recover absolute binding free energies efficiently.

3.5.2: The Linear Interaction Energy with Continuum Electrostatics method

Almost a decade later, Caflisch and Huang published a variant of the LIE method called the Linear Interaction Energy with Continuum Electrostatics (LIECE) approach.²⁵³ In LIECE, the MD simulations used for sampling are substituted by an energy minimization step and the electrostatic contribution is accessed by numerically solving the Poisson equation.²⁵³ The accuracy of the LIECE method and the transferability of the parameters were evaluated by studying 13 β -Secretase (BACE) and 24 HIV-1 protease inhibitors.²⁵³ The electrostatic contribution was computed by summing a term in vacuum, obtained through Coulombs' law, and the electrostatic part of the solvation free energy, computed by taking the difference between two Poisson calculations for each species of the binding reaction.²⁵³ One calculation is carried out with the external dielectric constant set to 1 and another with it set to 78.5, to evaluate the free energy difference for the polar part. The binding free energy is then computed using either a 2 or 3 parameter model (Equations 3.21 and 3.22)²⁵³

$$\Delta G_{bind} = \alpha \Delta U_{vdW} + \beta \Delta G_{elec} \quad (3.21)$$

$$\Delta G_{bind} = \alpha \Delta U_{vdW} + \beta \Delta G_{elec} + \Delta G_{tr,rot} \quad (3.22)$$

where ΔG_{elec} is the sum of the protein-ligand coulomb interactions in vacuum and the change in the electrostatic contribution to solvation free energy of the protein-ligand system upon complexation.²⁵³ The third term of Equation 3.22 $\Delta G_{tr,rot}$ corresponds to the loss of translational and rotational degrees of freedom upon binding.²⁵³

Remarkable accuracy was achieved for both the 2 and 3 parameter models in the BACE dataset, with a cross-validated RMSE of 1.16 and 1.28 kcal/mol for the 2 and 3 parameter model, respectively.²⁵³ As for the HIV-1 protease systems, it is reported that the 3-parameter model outperforms the 2-parameter one (RMSE of 0.77 kcal/mol and a cross-validated q^2 of 0.77 versus 0.97 kcal/mol and 0.64, respectively).²⁵³ These results are even more impactful when considering that the calibration and test datasets contain chemically diverse compounds which span a wide range of torsional flexibility.²⁵³ However, the value of the α and β parameters were found to not be transferable between the two datasets, which limits the applicability of each model. In the last decade,

LIE has seen applications to drug discovery, some of which required modifications to the LIE scheme. In particular, Oostenbrink and Sterjnschantz in 2010²⁵⁷ developed a scheme to combine the contribution of multiple ligand binding poses arising from molecular docking into a single binding free energy.²⁵⁷ This development appears to be critical for the prediction of the protein-ligand binding affinity when the complex involves highly flexible proteins, such as in cytochrome P450s, where ligands can adapt several different binding poses.²⁵⁷

3.5.3: The Molecular Mechanics Poisson Boltzmann Surface Area method

Within a VS campaign for a complex target, it is often the case that rigorous methods are not applicable at the hit identification stage and molecular docking approaches are commonly employed. During the docking experiment, a crude scoring function is used to evaluate the affinity of small ligands to a protein with a special emphasis on the throughput of the calculation. However, since the scores produced usually correlate poorly to with experimental binding affinities, many researchers opt by using other methods to re-score the docking poses to achieve higher ranking and calculation accuracy.²¹³ A popular method for docking pose rescoring is the MM/PBSA approach, introduced in the 1990s by Peter Kollman^{189,196,258,259}, which combines molecular mechanics energy terms, an implicit solvent Poisson-Boltzmann model to compute the polar contribution to the solvation free energy, a solvent accessible surface area (SASA)-based term accounting for the nonpolar contribution to the solvation free energy and an entropic term typically assessed in the rigid rotor harmonic oscillator approximation by either quasi-harmonic analysis (QHA)^{260,261} or normal mode analysis (NMA)^{262,263}. In the original implementation, MM/PBSA calculations implied carrying out individual MD simulations for the complex, receptor and ligand and extracting average quantities from these trajectories (3-average formalism). However, it is also possible to use MM/PBSA without carrying out MD simulations and instead applying it to an energy minimized structure as a post-docking filter. The MM/PBSA method aims at evaluating the absolute chemical potential of each species of the binding reaction. At chemical equilibrium, one can write that the equilibrium constant is related to the chemical potential by¹⁸⁹

$$K_{eq} = e^{-\beta\Delta\mu_b^0} \quad (3.23)$$

where $\Delta\mu_b^0$ is the difference between the chemical potential of complex and that of the unbound protein and ligand at the standard state.¹⁸⁹ In the limit of infinite dilution and assuming that the solution volume remains unchanged upon ligand binding ($V_{PL} = V_P$)²²⁶,

Equation 3.24^{189,264} shows that the value of the protein-ligand binding affinity is deeply connected to the reversible work of transferring the solute to solution¹⁸⁹

$$\mu_i = \mu_{i,vac} + W_{bulk}(X_0) \quad (3.24)$$

where $\mu_{i,vac}$ is the chemical potential of species i in vacuum, W_{bulk} is the reversible work of transferring the chemical species from vacuum to solution and X_0 is the minimum energy configuration of species i . Decomposing the above equation into entropic and enthalpic terms¹⁸⁹ one arrives at the original formulation of Peter Kollman and Irina Massova for MM/PBSA^{196,201,211} in **Equations 3.25** and **3.26**

$$\mu_i = G_i = \bar{E}_{BAT} + \bar{E}_{vdW} + \bar{E}_{elec} + \bar{G}_{pol} + \bar{G}_{npol} - TS_i(V) \quad (3.25)$$

$$\Delta G_{bind} = \langle G_{PL} \rangle - \langle G_P \rangle - \langle G_L \rangle \quad (3.26)$$

where G_i is the free energy contribution of species i , \bar{E}_{BAT} the potential energy contribution from the bonded terms, \bar{E}_{vdW} the force field vdW energy, \bar{E}_{elec} the electrostatic energy from the force field, \bar{G}_{pol} the polar contribution to the solvation free energy and \bar{G}_{npol} the nonpolar contribution to the solvation free energy.^{115,196,201,211} A schematic representation of the thermodynamic cycle applied in MM/PBSA calculations is shown in **Figure 3.3**. For the standard MM/PBSA approach, it is required to run individual MD simulations for complex, protein and ligand either in implicit or explicit solvent.²⁰¹ This formalism is typically known as the 3-average formalism, as it requires three independent simulations. Other researchers, however, opt by carrying out only one simulation of the protein-ligand complex and extracting the protein and ligand simulations from this one, in a variant called 1-average.¹⁹⁶ The reason for this is two-fold: to save computer time, since only one simulation per complex needs to be carried out, and due to the empirical observation that the simplified version of MM/PBSA yields more accurate results than the 3-average implementation, possibly due to the cancellation of errors from the contribution of the bonded terms.²⁰¹ However, the 1-average approach does not hold when protein-ligand binding events are associated with large conformational changes as the overlap between the configurational space sampled by the complex and the individual binding partners is significantly narrowed.²⁰¹ A third variant was put forth by Swanson *et al.*,²⁶⁵ where both the complex and ligand in solution would be simulated explicitly, so as to account for the ligand reorganization energy.²⁶⁵ Importantly, MM/PBSA can also be used as a scoring function.²⁰¹ In this case, the calculation is carried out on a single conformation of the protein-ligand complex and no

MD simulation is required. However, as in molecular docking, this approach neglects conformational flexibility.²⁰¹

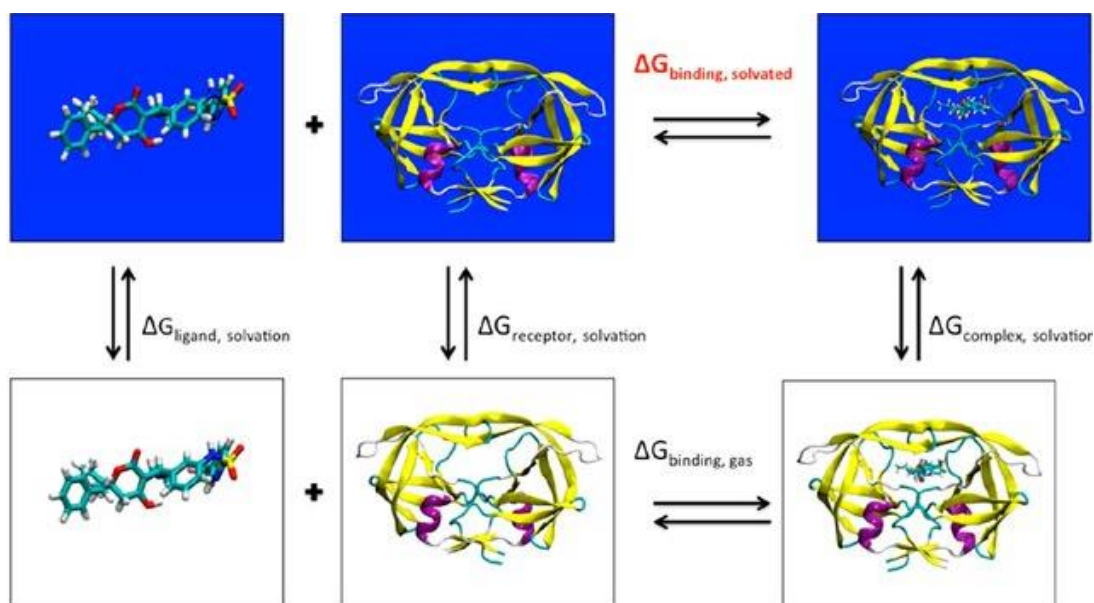


Figure 3.3 – Thermodynamic cycle for a binding free energy calculation of a protein-ligand complex. The solvated systems are illustrated in blue background and the gas-phase is illustrated in blank boxes. In red is highlighted the free energy of interest, computed by computing the free energy contributions in black. Adapted from “MMPBSA.py: An Efficient Program for End-State Free Energy calculations.” Miller III *et al.*²⁶⁷, 2012. *Journal of Chemical Theory and Computation*, 8, 9, 3314-3321. Copyright © 2012 American Chemical Society.

A technical limitation to MM/PBSA calculations is that the appropriate length of MD simulation for MM/PBSA calculations appears to be system-dependent.^{266,386} This limitation is also present in other binding free energy calculation approaches, including FEP calculations. Nonetheless, Genheden and Ryde advise that better results are obtained when the MM/PBSA or MM/GBSA results are drawn from averaging over many small independent MD trajectories as opposed to one single long MD trajectory.^{201,266} One explanation is that with a single long simulation the system initially samples configurations around the local energy minimum. As it progresses, though, it may visit other minima in a non-reversible manner due to difficulties in crossing back the energy barriers when relying solely on thermal fluctuations. This limitation leads to incomplete sampling and may be more serious in unbiased MD, hampering calculation convergence.²⁶⁶ Starting from many independent short simulations would allow better convergence by starting each simulation from different points in phase space while at the same time tackling the errors introduced by the forcefield which arise when carrying out one long MD simulation.^{266,384}

3.5.3.1: The Polar solvation term: Poisson Boltzmann calculation

The solvation free energy, as shown in **Equation 3.25**, is decomposable into two contributions: one polar and one nonpolar. The polar contribution is accessed by numerically solving the PB equation or by using a Generalized Born (GB) model.²⁰¹ In any case, the polar term to the solvation free energy corresponds to the electrostatic contributions arising from solvating a low dielectric material (the molecule) composed of charged particles within a homogenous high dielectric environment (the implicit solvent model).^{201,229,268} In MM/PBSA, this is carried out starting from the Poisson equation, which is a second-order partial differential equation, where the electrostatic potential $\Phi(\vec{r})$ is computed from is the position-dependent dielectric distribution function ($\epsilon(\vec{r})$) and the atomic charge density of the system ($\rho(\vec{r})$). Due to the lack of analytic solutions to the Poisson equation, it is often the case that numerical solutions are pursued, for example using Finite Difference (FD) methods.^{201,269} However, the original Poisson Equation does not consider the contribution of ions.²²⁹ Extending this equation to include the ionic salt concentration, where the distribution of these extra charges is obtained from the Boltzmann distribution, yields the Poisson-Boltzmann Equation (PBE) (**Equation 3.27**)²²⁹

$$\nabla[\epsilon(\vec{r})\nabla\Phi(\vec{r})] + \lambda(\vec{r})f(\Phi(\vec{r})) = -4\pi\rho(\vec{r}) \quad (3.27)$$

Here, the $\lambda(\vec{r})$ function is an ion-exclusion function which takes the value 1 outside the Stern layer. The $f(\Phi(\vec{r}))$ term, which is the term encoding the effect of the salts, is a function of the $\Phi(\vec{r})$ potential, the valence on the ion (z_i) and the bulk concentration (c_i) at a given temperature. At low ionic strength, one can apply a linearized form of the PBE, which is easier to solve numerically^{201,229}

$$\nabla[\epsilon(\vec{r})\nabla\Phi(\vec{r})] = -4\pi\rho(\vec{r}) + \epsilon_v k^2 \Phi(\vec{r}) \quad (3.28)$$

where $k^2 = \frac{8\pi e^2 I}{\epsilon_v K_B T}$, I is the ionic concentration of the solution and ϵ_{sol} is the solvent dielectric constant (78.5 or 80 for bulk water, typically). Once the electrostatic potential is determined in both vacuum and implicit solvent, the polar contribution to the solvation free energy can be computed through **Equation 3.29**²⁶⁸

$$\Delta G_{pol} = \frac{1}{2} \sum_i q_i (\Phi(\vec{r})^{solvent} - \Phi(\vec{r})^{vacuum}) \quad (3.29)$$

for a given set of discrete charges q_i (the solute), where the electrostatic potential in vacuum is computed using an external dielectric constant of 1 and 78.5 in solution.²⁶⁸ A PB calculation demands some steps to be carried out. First, it is necessary to superimpose the system on a cubic grid of a given edge size.²⁰¹ Then, the atomic charges should be mapped in the grid, assigning electrostatic potential, ionic strength and charge density values at each grid sub-cube.^{201,229,268} This implies defining the high and low dielectric regions and defining the values of the dielectric constant at each sub-cube, accounting for the boundaries of high and low dielectric portions.²⁰¹ Finally, one is ready to compute the electrostatic potential, numerically evaluated at each grid point.

Examples of established algorithms for PB calculations include the *pbsa* solver in the Amber software suite¹¹⁸, the ZAP algorithm²⁷⁰ in CHARMM and others.^{201,271,272} Nonetheless, solving this equation still implies a significant computational overhead since it needs to be solved every time the molecular configuration changes.²⁰¹ Additionally, the computational cost of these calculations is also expected to vary depending on the coarseness of the mesh used.²⁷³ One strategy is to apply grid-focusing such that the mesh in the high-dielectric region is coarse and then becomes finer as it approaches the solute, such that the computational cost is only paid on the finer mesh regions mapping the solute.^{201,273}

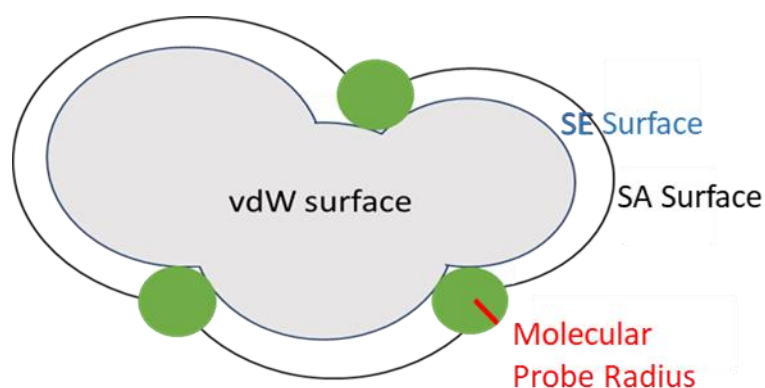


Figure 3.4 – Illustration of the two surface definitions: the SA surface (black) and the solvent excluded surface (blue) and the vdW surface (grey). The overlapping grey circles represent a three atoms of a molecule represented by a circle with a radii equal to their vdW radii. Adapted from “McVol - A program for Calculating Protein Volumes and Identifying Cavities by a Monte Carlo Algorithm.” Till and Ullmann, 2010.²⁷⁴ Journal of Molecular Modeling, 16, 3, 419-429. Copyright @ 2009, Springer-Verlag.

An important consideration for these calculations lies which the choice of the dielectric boundary between the solvent and the solute.²⁷⁵ There are two surface definitions which are typically used: the vdW surface, and the Solvent Excluded (SE) surface. The vdW surface is computed based on the surface union of the vdW sphere of the solute atoms. The SE surface corresponds to the surface obtained by rolling a solvent probe over the

vdW surface and taking the boundary defined by the probe. The solvent accessible (SA) surface is defined similarly, but by considering the center of the sphere as in **Figure 3.4**.²⁷⁵

3.5.3.2: The Polar solvation term: A Generalized Born description

In the context of VS and SBDD, the associated costs of introducing a PBE based method to compute the polar contribution to the solvation free energy can quickly become overwhelming. To address the issue of the computational cost, researchers have developed alternative methods.^{201,386} In particular, the Generalized Born (GB) model was developed aiming at a fast and efficient estimation of the polar contribution to solvation free energies by evaluating the electrostatic potential of one atom according to its local environment within a given distance.^{201,268,385} Within a GB model, the atoms are represented as spheres of a given radius filled with an internal dielectric constant and exhibiting each a given atomic partial charge^{268,385} whose level of exposure to the (implicit) solvent leads to a position-dependent dampening effect on their electrostatics. In practice, the more the atom is surrounded by atoms of the solute, the less its electrostatics will be dampened and the stronger it interacts with other solute atoms.^{201,268,385} This effect is critical for the GB calculation, because at the heart of the GB model is the effective Born radius α , which is dependent on the level of screening experienced by a given atom due to the implicit solvent.^{268,385} A large water screening leads to a small Born radius, because this atom has its interactions with the remaining solute atoms heavily impaired as it is solvent exposed.^{268,385} The GB equation (**Equation 3.30**) is described as follows, for a polyatomic system, and using a function based on a pairwise sum over the interacting charges^{201,277}

$$\Delta G_{solv}^{pol} = - \left(\frac{1}{\epsilon_{in}} - \frac{e^{-kf_{GB}}}{\epsilon_{sol}} \right) \sum_{i,j} \frac{q_i q_j}{f_{GB}} \quad (3.30)$$

with

$$f_{GB} = \sqrt{r_{ij}^2 + \alpha_{ij}^2 + e^{\left(\frac{r_{ij}^2}{4\alpha_{ij}^2}\right)}} \quad (3.31)$$

where q_i is the atomic partial charge of atom i , r_{ij} is the distance between atoms i and j , α_{ij} is the geometrical average between the effective Born radii of atoms i and j , ϵ_{in} is the internal dielectric constant and ϵ_{sol} is the solvent dielectric constant. The above **Equation 3.31** allows the inclusion of ions in the calculation through the Debye-Huckel (k) term.^{201,277}

The original GB model was the Hawkins, Cranmer, Truhlar one (HCT).^{201,281} This model allowed to represent a solute as a set of spheres and was much faster than standard PB calculations.^{280,385} However, it was noticed that this model did not account properly for the spaces left between atom spheres particularly those deep inside the solute.^{281,385} This is because, to prevent overlap between spheres with large Born radii, these radii were scaled down. In doing so, small crevices appeared within the innermost parts of the molecule.^{201,281,268,385} The problem is that these cavities would be treated as if they were filled with solvent and thus the effective Born radii of deeply buried atoms would be significantly reduced.²⁸¹ As such, Onufriev, Bashford and Case developed a new model (OBC), to account for these interstitial cavities by scaling up the Born radii of the buried atoms without affecting the Born radii of the solvent-exposed solute atoms.²⁸¹ More recently, the GBn models were developed aiming to account for a limitation in the OBC models: The small patches between spheres in the boundary with the solvent which inherently appear because of representing atoms as spheres.^{118,201} By accounting for the area between the spheres using a “neck integral”, the solute representation is closer to that of a molecular surface and in the studied systems the accuracy of the GB calculation increases with respect to a PB reference.¹¹⁸ Another model was later introduced by Onufriev *et al.*, which uses a grid-based surface implementation of a R^6 potential²⁸² which represents a successful compromise between speed and calculation accuracy.²⁰¹

3.5.3.3: The Variable Dielectric MM/GBSA model

The use of a continuum model to compute the polar contribution to the solvation free energy is common to both PB and GB calculations. These calculations depend heavily on the solute internal dielectric constant.^{201,386} While some researchers have reported good agreement with experimental data when setting $\epsilon_{in} = 1$, other researchers have shown that setting ϵ_{in} to 2,4 or even higher values, could improve predictions for highly charged protein-ligand complexes.^{115,201,386} It thus opens the discussion: what dielectric constant to select for each system to be studied, since it affects the accuracy of the binding free energy calculation significantly.³⁸⁶ Using a single dielectric constant to describe the electrostatic properties of a solute, especially when talking about protein-ligand systems, is a strong approximation, usually taken for simplicity.²⁰¹ In reality, biomolecular complexes are not uniform and thus it would be beneficial to have a way to assign a value of the dielectric constant as a function of the local electrostatic environment.^{201,283} Thus, studies aiming at a better understanding of how to assign this parameter were carried out either by benchmarking different values of ϵ_{in} and see which value produced the most accurate predictions of solvation free energies or binding free energies or by using variable dielectric models.^{201,283,284,386} In particular, it is important to note that although

the default value of the internal dielectric constant is set to 1, for highly charged binding sites it needs to be set to a larger value.^{201,384,386} An example of the other avenue pursued, the variable dielectric model, is the distance-dependent dielectric model, as proposed by Wang *et al.*²⁸³ In this model, the electrostatic energy and the polar contribution to the solvation free energy are computed while employing a distance threshold d_0 such that

$$E_{elec} = \begin{cases} \sum_k \frac{1}{\epsilon_{in(k)}} \sum_{i \in k, j} \frac{q_i q_j}{r_{ij}}, & d_{rl} < d_0 \\ \sum_{i, j} \frac{q_i q_j}{\epsilon_{in} r_{ij}}, & d_{rl} \geq d_0 \end{cases} \quad (3.32)$$

$$G_{pol} = \begin{cases} -\sum_k \left(\frac{1}{\epsilon_{in(k)}} - \frac{1}{\epsilon_{sol}} \right) \sum_{i \in k, j} \frac{q_i q_j}{f_{GB}} d_{rl} < d_0 \\ -\left(\frac{1}{\epsilon_{in}} - \frac{1}{\epsilon_{sol}} \right) \sum_{i, j} \frac{q_i q_j}{f_{GB}}, & d_{rl} \geq d_0 \end{cases} \quad (3.33)$$

where d_{rl} is the distance between the center-of-mass of residue of type k to the center of geometry of the ligand.²⁸³ In the case that the ligand atom j and the protein residue k are in close proximity, $d_{rl} < d_0$ and the value of the internal dielectric constant for the protein residue is assigned according to its type (charged, polar and non-polar aminoacid, 4, 2, 1 respectively). For residues far away from the ligand, they are assigned a default dielectric constant.²⁸³ Furthermore, the dielectric constant of the ligand is also computed, since the free energy contribution of the ligand to the overall binding free energy is also affected by the choice of the solute dielectric constant.²⁸³

The researchers found that using the VD model led to increased predictivity when compared to calculations carried out using an solute dielectric constant of 1²⁸³ for 130 protein-ligand complexes from the 2013 PDBbind core set.²¹⁴ In practice, however, the improvement was limited to a few percent points²⁸³ and the authors ascribe the modest improvement to the fact that they explored a rather narrow range of dielectric constants for the solute.²⁸³ This could be a problem particularly in highly charged systems, as evoked by the authors as a justification.^{283,386}

3.5.3.4: The Nonpolar solvation term

The nonpolar contribution to the solvation free energy is related to the process of creating a cavity on the solvent with the shape of the solute and filling that cavity with electronic density, which has both repulsive and attractive contributions.^{201,283,285,384} The cavity creating step represents a free energy cost which is then balanced by the gain in

interactions established by the solute with the solvent.^{118,283,285} In most cases, it is estimated as a linear relationship to the SASA as in **Equation 3.34**^{118,283,285}

$$\Delta G_{nonpol} = \gamma * SASA + b \quad (3.34)$$

where $\gamma = 0.00542 \text{ kcal/mol/\AA}^2$ is the surface tension and $b = 0.92 \text{ kcal/mol}$ is a correction term set to be constant in the Amber18 software suite.¹¹⁸ More recently, other models for computing the ΔG_{nonpol} were put forth. On one hand, a modified version of **Equation 3.35** was proposed, where ΔG_{nonpol} is estimated as the sum of two main terms: cavity formation and the dispersion term arising from the vdW interactions between solute and solvent.^{201,229} On the other hand, a polarizable forcefield approach, the Polarizable Continuum Model (PCM)^{201,286} opens, in principle, to the calculation of non-polar solvation free energies with higher accuracy than those obtained by SASA-based methods.²⁸⁷ The matter of fact is that although the non-polar contribution is important for a proper binding affinity prediction, little attention has been given to it in the last years.²⁰¹

3.5.3.5: Entropy in MM/PBSA calculations

Entropy is one of the key thermodynamic properties ruling protein-ligand binding. However, accurate evaluation of entropic contributions is typically difficult and time consuming. One of the most well-known methods used to include entropic contributions in MM/PBSA or MM/GBSA calculations is NMA.^{262,263} However, NMA requires the creation of a Hessian matrix of all 3N-6 internal degrees of freedom of the system, considers only the local displacements around the equilibrium conformation examined and neglects the flexibility the protein-ligand complex beyond harmonic motions, leading it to underestimate the entropy of flexible systems.^{201,218,262} This matrix must then be diagonalized, an endeavor which is computationally costly and must be repeated for a given number of snapshots.²⁰¹ Other methods have also been developed but most of them require very large simulation times to provide converged measurements or are known to produce only upper-bound estimates to the true entropy of a chemical species like in QHA.^{225,260,261} Due to the difficulties in correctly computing these contributions some researchers opt by ignoring entropic terms entirely.^{115,201} An additional reason for ignoring entropic contributions in these calculations stems from the empirical observation that when these terms are considered using NMA or QHA, the correlation between predicted and experimental binding affinities tends to decrease.²⁰¹ However, by ignoring these terms, the comparison between ligands of different sizes with respect to their predicted MM/GBSA binding affinities becomes unreliable³⁸⁴ as the calculation is

biased towards predicting bigger ligands as better binders. In the context of VS, it would lead to the prioritization of many false positives. In recent years, some researchers developed simplified methods for entropy calculations, such as the Truncated NMA entropy method²⁵⁴, the interaction entropy method²¹² or the Binding Entropy Estimation of Rotation and Translation (BEERT) approach.²⁸⁸ In the truncated NMA entropy method, only protein residues under a radius threshold between 8-16Å of the ligand center-of-mass are kept while the remaining protein residues are removed. This truncation greatly reduces the size of the system, accelerating NMA calculations.²⁷³ It was shown by the group of Hou that introducing the truncated NMA entropy estimation increases the accuracy of the absolute binding free energies and the correlation to the experimental data.²⁷³ The interaction entropy calculation introduced by Duan *et al.*, on the other hand, post-processes the MD simulation to evaluate the protein-ligand interactions and calculates the entropy change upon binding.²¹² In particular, it estimates the entropy change upon binding from the fluctuations of the electrostatic and vdW energies between protein and ligand, extracted directly from an MD simulation of the complex by post-processing.^{201,212} Finally, the BEERT method aims at computing the translational and rotational entropy loss of the ligand upon binding starting from the flexible molecule framework.²⁸⁸ All of these methods are significantly more efficient than standard NMA or QHA. Still, many people opt by not including these contributions in their end-point binding free energy calculations, which clearly provides incorrect predicted binding free energies. As pointed out recently by Tuccinardi³⁸⁴, the development of a method allowing accurate and efficient calculation of entropic contributions in MM/PBSA and related methods is envisaged in the future.³⁸⁴ To this end, an interesting avenue would be exploiting the computational power offered by GPUs.³⁸⁴

In **Chapter 4** we will review the statistical mechanics basis for RRHO and beyond-RRHO entropy calculations and highlight the theoretical framework which constitutes the basis for the Quasi-Harmonic Multi-Basin method developed and presented in **Chapter 5**.

3.6 – Recent applications of MM/PB(GB)SA calculations

Over the years, due to the increasing popularity of computational methodologies within academic and pharmaceutical drug discovery projects, MM/PBSA and its variants have enjoyed a spot in the limelight. Since docking scoring functions are highly approximated and most often do not include neither solvation nor entropic terms in an appropriate manner¹⁸⁹, it is often the case that a single scoring function may have limitations when predicting the binding affinity for protein-ligand complexes. Thus, some researchers argue that the combination of docking and MM/PBSA or MM/GBSA rescoring would lead

to better ranking of compounds.²¹³ For docking pose rescoring, these methods are some of the fastest available and are suitable for VS of large chemical libraries. The choice of whether to use MM/PBSA or MM/GBSA is not crystal clear as some benchmark studies report better performances for the former while others for the latter.^{201,229} Considering only computation efficiency, however, leads one to prefer the MM/GBSA method within VS approaches. One study illustrating the usefulness of these methods is the work by Lightstone *et al.*²¹³ where a docking plus free energy rescoring by MM/GBSA scheme was applied to study ligand recognition upon binding to antithrombin, achieving good agreement with experimental data ($R^2 = 0.69$).²¹³ However, a closer look at the predicted binding affinities highlights limitations of the approach: the predicted binding affinities were systematically overestimated by as much as one order of magnitude.²¹³ The authors state that a reason may be the lack of entropic contributions while another potential source of error may be due to the usage of an implicit solvent model.²¹³ Nonetheless, the ranking of compounds agreed with the experimental ranking, which means that the method was able to identify the least and most potent binders.²¹³

The power of end-point methods in discriminating true actives from decoys has also been a large object of study, with benchmark studies highlighting the screening power of these methods.^{289,384} An enlightening example is given by Sgobba *et al.*^{229,289} which assessed the power of these end-point methods in docking pose rescoring for six different drug targets. They reported that MM/GBSA binding affinity calculations yielded predictions which lead to both a better Area Under the Curve (AUC) and enrichment factor (EF) when compared to molecular docking scoring functions.^{229,289} A large scale study by Zhang *et al.*,²²⁹ targeting 38 drug targets with a large ligand library containing more than half a million compounds, including active molecules and decoys, also demonstrated the same trend as in Sgobba *et al.*^{229,289}

More recently, a molecular docking and MM/PB(GB)SA free energy rescoring approach was carried out to identify hit compounds from already approved drugs in a docking plus free energy rescoring approach targeting the SARS-Covid19-Mprotease.²⁹⁰ The author started from a dataset of 2201 approved drugs present in the DrugBank database and carried out molecular docking to the SARS-Covid19-Mprotease using Glide.²⁹⁰ The best ranked compounds from the docking campaign were rescored using MM/GBSA, utilizing MD simulations to generate the conformational ensembles for the rescoring step.²⁹⁰ The author found that several drugs were predicted as potential hits. Furthermore, a MM/GBSA free energy decomposition allowed to identify key residues to SARS-Covid19-Mprotease-ligand binding, an information which facilitates the rational design of novel inhibitors.²⁹⁰ Another recent application of MM/GBSA was reported by Lagarias *et al.*,²⁹¹ where it was applied both to a small scale VS campaign comprised of compounds similar to a potent Adenosine A3 receptor antagonist and to evaluate the effect of mutating

binding site residues on the activity of the antagonist.²⁹¹ A final example highlighting an advanced implementation of MM/PBSA calculations is described by the group of Gohlke.²⁹² The method was used to obtain a per-residue energy contribution to the dimerization process of G-coupled Protein Receptor (GPCR) TGR5 embedded in a lipid membrane using an implicit membrane model.^{229,292} The implicit membrane was divided into five slabs, each with a given internal dielectric constant which increased from the center of the membrane outward.²⁹² The MM/PBSA analysis allowed the discovery of residue hotspots in the interface between TGR5 dimers which could be targeted by small molecules.²⁹²

4. Theory of single molecule entropy methods with applications in MM/PB(GB)SA calculations

4.1: Introduction

During a binding reaction between a protein and a small-molecule ligand, interactions are established between the binding partners while their structures tend to rigidify, meaning some motions become constrained. In other words, the gain in enthalpy which arises from the protein-ligand interactions established is balanced out by an entropic penalty due to the constraining of the external and internal motions of each species of the binding reaction. Thus, entropy can be seen as a fundamental thermodynamic observable which exerts control over binding phenomena between two chemical entities.^{199,218,293} It is also critical in other biomolecular processes, such as surface self-adsorption²⁹⁴ or protein folding²⁹⁵. Nonetheless its pivotal role in these processes, accurate calculations of entropic contributions for flexible molecules, which can span many conformations, remains a challenge. Over the years, many different methods have been proposed to evaluate absolute entropies and entropy differences. Some of them are based on the Rigid Rotor Harmonic Oscillator (RRHO) approximation^{225,261–263,296}, using either the classical or the quantum mechanical (QM) harmonic oscillator, while others are based on information expansion techniques^{293,297–299}, nearest neighbors estimators^{300,301} or even hybrid calculations combining RRHO-based calculation and information expansions approaches.^{302–305}

The scientific community is still in pursuit of a method which allows the proper calculation of absolute molecular entropies in solution accurately and efficiently.²¹⁸ One reason is that unraveling the influence of entropic terms in protein-ligand binding would allow a more complete picture of how the binding reaction happens. Another reason is the fact that due to current limitations in software and hardware, rigorous absolute binding free energy calculations are still inaccessible within VS approaches and thus, it is usually the case that end-point methods are employed for refinement of rankings produced from molecular docking simulations. As stated in **Chapter 3**, some researchers opt by not including entropic contributions in calculations carried out using MM/PBSA and similar methods, as empirically it has been observed that they introduce more noise than signal and they are costly to compute.^{115,229,273} However, the resulting predicted binding free energies are often overestimated and biased to predict bigger ligands as better binders. In this chapter, we will start by describing the statistical mechanics

fundamentals for absolute entropy calculations. Then, we will describe entropy calculation approaches both in and outside the RRHO formalism which can be used in conjunction with MM/PB(GB)SA binding free energy calculations. Finally, we will discuss how to evaluate the accuracy of absolute entropy calculations through benchmark data.

4.2: A statistical mechanics view of entropy

As illustrated in **Chapter 3**, the partition function connects the micro and the macroscopic worlds. It does so by relating a macroscopically measurable quantity, like the free energy of a system or the difference in free energy between two states of a system, to the assessable microstates at a given set of conditions. In the limit of classical mechanics, the *partition function* Q of a system can be written as in **Equation 3.1**. Within this framework, the entropy of that system can be computed from the probability distribution $p(q, p)$ as²²⁵

$$S = -K_B T \int \int p(q, p) * \ln(p(q, p)) dp dq \quad (4.1)$$

which is known as the Shannon-Gibbs entropy.²²⁵ In particular the entropy above described contains two contributions: one arising from the momenta p , and one arising from the position q , of the systems' degrees of freedom.^{225,29} After factorizing the above integral, two contributions are obtained: one from the momenta and one from the coordinates.^{225,29} Since the momenta contribution cancels out in entropy differences, often the real challenge is estimating the configurational integral (**Equations 4.2**) as stated by Díaz and Suárez²²⁵ Hence, the entropic contributions of interest are usually coordinate-dependent and assigned the name of configurational entropies.^{225,293,298} The equation to compute the configurational entropy of a system is then written down as^{225,293,306}

$$S_{config} = -K_B T \int p(q) * \ln(p(q)) dq \quad (4.2)$$

The above equation provides access to the absolute entropy of a molecular system through the probability distribution function of its configurational space $p(q)$, which can be sampled by molecular simulations.³⁰⁴ Another way of obtaining the absolute entropy of a system arises by taking the derivative of the free energy with respect to the temperature, which arrives at the following **Equation 4.3**

$$S = - \left(\frac{\partial F}{\partial T} \right) = - \left(\frac{\partial K_B T \ln Z}{\partial T} \right)_{N,V} \quad (4.3)$$

providing access to the absolute entropy through the *canonical partition function* by evaluating its temperature dependence. The above **Equation 4.3** can also be written at constant temperature and pressure by taking the derivative of the Gibbs free energy ($\frac{\partial G}{\partial T}$) with respect to temperature as opposed to the derivative of the Helmholtz free energy ($\frac{\partial F}{\partial T}$). By taking **Equation 4.3** and realizing the third law of thermodynamics, which states that the entropy of a perfect crystal at zero K is zero, *i.e* $S(0K) = 0$, one arrives at the calorimetry definition of entropy, **Equation 4.4**

$$S(T) = S(0K) + \int_{0K}^T \frac{\delta q_{rev}}{T} \quad (4.4)$$

where δq_{rev} is the heat exchanged during a reversible transformation from **0K** to the target temperature at constant volume. At a constant 1 bar pressure, which is typically how experimental gas-phase entropies are measured³⁰⁷, **Equation 4.4** becomes **4.5**

$$S(T) = \int_{0K}^T \frac{dH}{T} \quad (4.5)$$

Through the ideal gas approximation, which assumes that the particles are sufficiently far away such that they do not interact with each other and are indistinguishable, the partition function Z of a system composed by N molecules is readily computed from the product of the individual molecular partition functions (z), in the limit of high temperature²²⁵

$$Z = \frac{z^N}{N!} \quad (4.6)$$

where $N!$ is included to avoid overcounting due to the indistinguishable nature of the molecules in the system.²²⁵ This equation can then be plugged into **Equation 4.3** to yield the single molecule contribution to the absolute entropy of a system with N molecules as²²⁵

$$\frac{S}{N} = \left[\frac{\partial K_B T \ln \left(\frac{z}{N e} \right)}{\partial T} \right]_V \quad (4.7)$$

The entropy per molecule of the system is then accessed from the molecular partition function. Employing the Born-Oppenheimer approximation, we can separate translational from internal motions.²²⁵ The rigid rotor (RR) approximation then allows one to further decouple two types of internal motions: rotations and vibrations^{224,225}, although there may still exist some rotation-vibrational coupling. The vibrational partition function then is accessed in the limit of the Harmonic Oscillator approximation (HO).^{78,225} As such, under the RRHO approximation, the single molecule partition function can then be written as a product of the partition functions encoding each type of motion (translations, rotations and internal vibrations) and the electronic partition function.^{78,225} However, this last term is generally irrelevant for classical MD simulations in the ground state, and thus^{218,308}

$$Z = Z_{trans}Z_{rot}Z_{vib} \quad (4.8)$$

where Z_{trans} is the translational partition function, Z_{rot} is the rotational partition function and Z_{vib} is the vibrational partition function of the $3N-6$ internal degrees of freedom of a molecule. Following **Equation 4.8** and **Equation 4.7**, they yield

$$S = S_{RRHO} = S_{trans} + S_{rot} + S_{vib} \quad (4.9)$$

where for each entropic contribution there is an analytical expression arising from statistical mechanics such that²²⁴

$$S_{trans} = NK_B \ln \left(\frac{V e^{5/2}}{N} \left(\frac{2\pi m K_B T}{\hbar^2} \right)^{3/2} \right) \quad (4.10)$$

$$S_{rot} = NK_B \left[\ln \left(\frac{\sqrt{\pi I_A I_B I_C}}{\sigma} \left(\frac{8\pi^2 K_B T e}{\hbar^2} \right)^{3/2} \right) \right] \quad (4.11)$$

$$S_{vib} = NK_B \sum_{i=1}^{3N_{at}-6} \left[\frac{\frac{\hbar \nu_i}{K_B T}}{\left(e^{\frac{\hbar \nu_i}{K_B T}} - 1 \right)} - \ln \left(1 - e^{-\frac{\hbar \nu_i}{K_B T}} \right) \right] \quad (4.12)$$

where V is the volume, I_A, I_B, I_C are the three principal moments of inertia of the molecule, σ is the rotational symmetry number and ν_i the harmonic vibrational frequencies of the $3N_{at} - 6$ internal degrees of freedom.²²⁴ From **Equations 4.10**, **4.11** and **4.12**^{224,225,261,296}, one can see that these entropic contributions depend only on a couple of terms. The translational contribution depends on the standard state of the

system $\frac{V}{N}$, the volume occupied per N molecules, and on the mass of the system. The standard volume for a molecule can be 22.4 mol/L in the case of the vacuum (1 bar pressure, 273.15K), 24.78 mol/L in the case of an ideal-gas at 1 bar pressure and 298.15K, and 1 mol/L in solution. The rotational contribution depends on the symmetry number σ and the moments of inertia, which encode the geometry of the molecule. The vibrational contributions depends solely on the harmonic vibrational frequencies of the internal degrees of freedom, which describe the width and depth of the potential energy wells explored by these degrees of freedom of the system.²²⁵ Thus, the RRHO approximation allows the calculation of absolute entropies and entropy differences in gas-phase or in solution, this last one commonly done using implicit solvent models, in combination with MM forcefields typically used in biomolecular simulations. It comes as no surprise that RRHO-based methods for entropy calculations are commonly applied within MM/PBSA and MM/GBSA approaches^{201,229,273}, although with system-dependent success rates. Indeed, the RRHO approximation as it is normally applied tends to produce either upper or lower bound estimates to the true entropy, depending if QHA or NMA is used. The harmonic oscillator model is, however, central to these calculations and is briefly described below.

4.3: The Harmonic oscillator

The harmonic oscillator is a physical model used to study the motions of pendulums, masses connected to springs and acoustical systems.³⁰⁹ It also has applications in other fields, such as molecular simulations²²⁹. It is fundamental within RRHO entropy calculations, as the solution to the HO equation yields the vibrational frequencies which are to be plugged into **Equation 4.12**.^{224,310}

4.3.1: The classical harmonic oscillator

The harmonic oscillator model is used to describe the movement of a one-dimensional particle attached to a spring which behaves according to Hooke's law.^{224,309} In the model, a force acts to displace the particle from the equilibrium position by stretching the spring.³⁰⁹ This force is typically proportional to the displacement and depends on the stiffness of the spring.³⁰⁹ A stiffer spring will be more difficult to stretch. The sign of the force is typically negative because the restoring force produced by the spring is of the inverse sign of the force used to stretch it. In the uni-dimensional case, this force is described by³⁰⁹

$$F = -kx \quad (4.13)$$

where F is the force, k is the force constant and x is the displacement of the particle from the equilibrium, or reference, position. A strong force constant implies that the displacements away from equilibrium are more strongly acted upon and more difficult, meaning the stiffness of the spring is stronger than for a smaller value of k . The equation of motion of the particle is given by

$$\ddot{x}(t) + \omega^2 x(t) = 0 \quad (4.14)$$

where $\omega^2 = k/m$ and m is the mass of the particle attached to the spring. The oscillatory motion is thus described by a sinusoidal function³⁰⁹

$$x(t) = A \sin \omega t + B \cos \omega t \quad (4.15)$$

The function shown in **Equation 4.15** describes the motion of a particle with harmonic behavior when a given force F acts upon it. This particle is called a harmonic oscillator and the associated angular frequency of oscillation is given by

$$\omega = \sqrt{\frac{k}{m}} \quad (4.16)$$

with ω measured in radians per second. The vibrational frequency of the oscillator ν is then computed as

$$\nu = \frac{\omega}{2\pi} = \frac{1}{2\pi} \sqrt{\frac{k}{m}} \quad (4.17)$$

and has units of Hertz (Hz). There is also a relationship between the force acting on the harmonic oscillator and the potential energy $U(q)$ of the system of the particle attached to the spring.³¹⁰ It can be realized by first writing the equation of the potential energy of the oscillator (**Equation 4.18**) and then taking the derivative of the energy with respect to the coordinates q of the system (**Equation 4.19**), such that

$$U(q) = \frac{1}{2} k(q - \bar{q})^2 \quad (4.18)$$

$$F = \frac{-dU(q)}{dq} \quad (4.19)$$

The potential energy is thus described by a quadratic function with the oscillator moving around the equilibrium position. As an example, **Figure 4.1** illustrates two oscillators: one with a lower and one with a larger value of k . From the HO, one can estimate the entropy of that system by **Equation 4.20** where ω is the angular frequency²⁹⁶

$$S_{classical} = K_B + \frac{K_B}{2} \ln \left(\frac{1}{\left(\frac{\hbar \omega}{K_B T} \right)^2} \right) \quad (4.20)$$

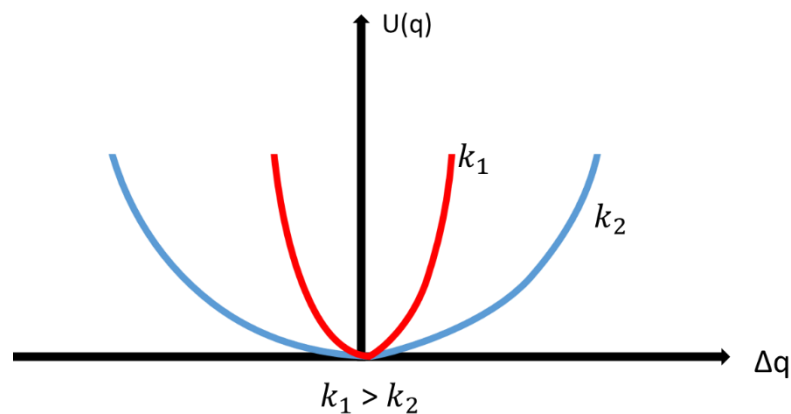


Figure 4.1 – Comparison between two harmonic oscillators. In red, harmonic oscillator number 1 is represented, with its force constant k_1 . In blue is shown the same oscillator but with a lower force constant, k_2 .

4.3.2: The quantum-mechanical harmonic oscillator

The quantum-mechanical HO is the quantum version of the classical HO and represents a key model system for which an analytical solution to the Schrödinger equation is attainable.^{224,310} The Hamiltonian of the QM oscillator is given by

$$\hat{H} = \frac{\hat{p}^2}{2m} + \frac{1}{2} k \hat{q}^2 = \frac{\hat{p}^2}{2m} + \frac{1}{2} m \omega^2 \hat{q}^2 \quad (4.21)$$

where \hat{p} is the momenta operator, m is the particle mass, \hbar corresponds to the reduced plank's constant, $\omega^2 = k/m$ and \hat{q} is the position operator of the coordinates q . One of the key results that appear from solving **Equation 4.21** are the allowed energy levels with energy E , through **Equation 4.22**^{224,310}

$$E_n = (2n + 1) \frac{\hbar}{2} \omega \quad (4.22)$$

A consequence of **Equation 4.22** is that it shows the energy levels of the HO to be quantized: they exist as discrete energy values which are equally spaced, given by $\frac{\hbar}{2} \omega$ times an integer number. It also implies that the lowest energy state, with $n = 0$, is given by $\frac{\hbar}{2} \omega$ which means it is a non-zero energy level. The entropy of a QM oscillator is then obtained by combining **Equation 4.12 and 4.17**.^{225,261,296} A clear distinction between the two models is that in the classical treatment, high vibrational frequencies do not correspond to an entropic contribution of zero, whereas in the quantum limit they do.²⁹⁶ In the classical limit, these fast vibrational motions tend to produce negative vibrational entropy contributions. On the other hand, as demonstrated by Karplus and Andricioaei²⁹⁶ using the QM oscillator these frequencies have close to zero or even zero contribution to the vibrational entropy.²⁹⁶ For this reason, the QM HO model was used when computing vibrational frequencies in the entropy calculations carried out for this dissertation.

4.4: Rigid-rotor Harmonic Oscillator-based entropies in flexible molecules

The expression in **Equation 4.9** can be used to compute accurately absolute entropies for rigid molecules in the ground state, where all of their internal degrees of freedom behave harmonically and only one potential energy well exists in the landscape. An example one could think is benzene or propiolactone. However, these represent a small fraction of the small-molecule chemical space. In practice many molecules are anharmonic in nature and can explore many different configurations in gas-phase or in solution. Indeed, the accuracy of RRHO-based entropies breaks down when flexible molecules are considered. Or does it? The RRHO approximation provides the absolute entropy of a molecule for a given configuration or ensemble of configurations. Thus, should many configurations be visited, one RRHO calculation should be carried out per configuration. For a given flexible molecule, its RRHO entropy is not given only by computing the absolute entropy for one configuration but instead by considering all sampled configurations which may be obtained by MD simulations. In this case, we write that the RRHO entropy of one configuration is^{225,296}

$$S_{rrho}^i = S_{trans,i} + S_{rot,i} + S_{vib,i} \quad (4.23)$$

where S_{rrho}^i is the entropic contribution of configuration i in the RRHO approximation. It may be tempting to compute S_{rrho}^i for all configurations sampled by MD simulations for instance, and just average their contribution to obtain the absolute entropy. However a second contribution related to the fact that the system can visit any of this configurations at a given time must also be considered.^{101,296,303} This is known as the mixing entropy, entropy of the landscape, or the conformational entropy, and is closely related to the Boltzmann formula for entropy^{101,296,303}

$$S = K_B \ln \Omega \quad (4.24)$$

where Ω corresponds to the number of microstates accessible to the system. The above **Equation 4.24**, however, carries an implicit assumption: that of equal probability of sampling each system configuration or microstate. A more general equation, which is fundamental for absolute entropies, is given when this assumption is not present and carries the name of the entropy of mixing (**Equation 4.25**).^{78,311,312}

$$S_{mix} = -K_B \sum_i p_i \ln(p_i) \quad (4.25)$$

in which p_i is the probability of conformer i . Considering **Equation 4.25**, two issues arise. First, how to get those probabilities. Second, how does that affect S_{rrho}^i ? The answer for the first question comes by considering that the probability of sampling conformer i is given by **Equation 3.5**. Another approach to it is to carry out a clustering procedure on the molecular snapshots. The molecular snapshots obtained by MD are grouped together based on a similarity threshold and each cluster corresponds to an individual basin in the PES describing one of the sampled microstates. As such, the cluster occupancy turns out to be the probability of that microstate assuming that we have converged on sampling the PES of that molecule. As for the second question, it is apparent that if conformers are sampled with different probabilities, then the simple averaging scheme should be turned into a probability-weighted average^{78,225}

$$S_{rrho} = \bar{S}_{rrho,i} = \sum_i p_i (S_{trans,i} + S_{rot,i} + S_{vib,i}) \quad (4.26)$$

which in turns means that under the RRHO approximation, the entropy of a molecule can be computed by: (1) enumerating all relevant microstates at a given temperature, (2) computing the corresponding probabilities, (3) computing a probability-weighted RRHO entropy and (4) adding the entropy of mixing to the RRHO entropy. The underlying assumption here is that the potential energy surface of the molecule is decomposable into a sum of individual harmonic wells where the entropy of mixing is indeed the

entropic contribution arising from populating multiple energy wells. This approach is named the “mixture of conformers” method and its master equation (Equation 4.27)
 jS^{78,225,296,313}

$$S = \sum_i p_i S_{rrho,i} - K_B \sum_i p_i \ln(p_i) \quad (4.27)$$

Through the “mixture of conformers” approach it is possible to compute absolute entropies. However, an underlying requirement is the proper enumeration of all relevant conformers which can be populated at a given set of conditions of temperature and pressure. The exhaustive search is non-feasible for large macromolecules, since these populate an incredibly high number of microstates at the conditions in which we wish to study protein-ligand binding. Thus, while formally correct, this approach is limited to small molecules unless significant improvements in either direct counting approaches or sampling methods are introduced.²²⁵ For a flexible small molecule in either gas-phase or solution which undergoes conformational transitions and explores different microstates, the configurational entropy is given by Equation 4.28.^{225,296,313}

$$S_{config} = \sum_i p_i S_{vib,i} - K_B \sum_i p_i \ln(p_i) = \bar{S}_{vib} + S_{mix} \quad (4.28)$$

Within the RRHO approximation, the vibrational frequencies are typically accessed using normal mode analysis or quasi-harmonic analysis. This approach is at the heart of the newly developed Quasi-Harmonic Multi-basin method (QHMB), where the vibrational frequencies are accessed through QHA for each microstate.²¹⁸ The QHMB method relies on the “mixture of conformers” framework to compute the absolute entropy of small molecules in solution or in gas-phase.

4.5: Configurational entropy within the Rigid-Rotor Harmonic Oscillator approximation

4.5.1: Normal Mode Analysis

The normal mode analysis is a technique that can be used to probe the flexibility of a molecule around an equilibrium conformation. An equilibrium system is defined as a system at the bottom of a potential energy well where the forces acting on it sum up to zero.²⁶³ Mathematically, we start by writing the potential energy function of the system

in terms of a Taylor expansion around the configuration at the bottom of the potential energy well like^{263,314}

$$U(q) = U(q^{eq}) + \left(\frac{\partial V}{\partial q_i}\right)^{eq} (q_i - q_i^{eq}) + \frac{1}{2} \sum_{i,j} \left(\frac{\partial^2 V}{\partial q_i \partial q_j}\right)^{eq} (q_i - q_i^{eq}) * (q_j - q_j^{eq}) + \dots \quad (4.29)$$

where the eq superscript indicates the equilibrium conformation and $\sum_{i,j} \left(\frac{\partial^2 V}{\partial q_i \partial q_j}\right)^{eq}$ are the second derivatives of the potential energy with respect to each pair of components i and j .^{263,314} The first term of **Equation 4.29** is assumed to be zero as it corresponds to the minimum value of the potential. The second term is also zero. The third term is where the expansion is truncated, such that **Equation 4.30** becomes³¹⁵

$$U(x) = \frac{1}{2} \sum_{i,j} \left(\frac{\partial^2 V}{\partial q_i \partial q_j}\right)_{eq} (q_i - q_i^{eq}) * (q_j - q_j^{eq}) = \frac{1}{2} \sum_{i,j} \Delta q^T \mathbf{H} \Delta q \quad (4.30)$$

where \mathbf{H} is the Hessian matrix. The Hessian matrix is a matrix built of the second derivatives of the potential energy with respect to the position of the degrees of freedom. It encodes the degree of correlation between every two degrees of freedom in the off-diagonal terms and the force constant applied to each internal degree of freedom is stored in the diagonal.^{263,315} Once the Hessian matrix is diagonalized, the eigenvalues and eigenvectors of the matrix are obtained and thus the angular frequency of each of the harmonic oscillators, which are now independent from each other, is retrieved. The eigenvalues are closely related to the vibrational frequencies with which the particles of a normal mode oscillate whereas eigenvectors describe the direction and the magnitude of the displacement each particle experiences with respect to the remaining particles.³¹⁵ In Hessians there are 6 zero eigenvalues, corresponding to rigid-body translations and rotations of the molecule.^{263,315,316} For a proper NMA calculation, it is important that the system be at an energy minimum. To check for it, following one NMA calculation the minimum energy conformation will either have zero or positive curvature in all directions of motion (*i.e* all the harmonic oscillators excluding the first six will have positive eigenvalues). If this is not the case, the conformation obtained is not at the energy minimum and must be further subjected to energy minimization steps.^{263,314} As shown by Bahar *et al.*,^{263,314} the connection between the Hessian and the angular frequency is observable through **Equation 4.31**

$$\mathbf{H}\mathbf{u}_i = \omega_i^2 \mathbf{M}\mathbf{u}_i \quad (4.31)$$

where \mathbf{u}_i is a 3N-dimensional vector accounting for the phase and the amplitude of each oscillator. Each eigenvector is connected with a normal mode coordinate and ω_i^2 is the eigenvalue (the square of the angular frequency).¹³³ Along the vibrational modes found for a given system, some will have lower and other higher frequencies. High frequency modes then to be rapid-oscillating modes with small displacements around the equilibrium position and involving few atoms.³¹⁰ These oscillatory modes contribute very little to the vibrational entropy, an example being carbon-hydrogen stretching. Low frequency modes are those which oscillate slowly and can deviate significantly from the equilibrium position, contributing the most towards the vibrational entropy.³¹⁷ The lowest frequency modes describe motions of larger groups of atoms and thus allow us to probe large-scale conformational transitions.³¹⁵ Applications of NMA are found throughout the literature. Within the context of binding free energy calculations, this method is routinely used in MM/PBSA calculations and other variants.^{115,201,273}

To carry out a NMA calculation, (1) molecular configurations are subjected to an energy minimization procedure. This is a crucial step, as for NMA to yield proper vibrational frequencies the conformation used must be at an energy minimum. Following energy minimization, (2) the Hessian matrix is built and diagonalized, producing the angular frequencies. From these, (3) the vibrational frequencies of each mode are obtained and plugged into **Equation 4.12**. To compute accurately absolute entropies or even entropy differences using NMA for large macromolecules, it is required that exhaustive sampling be achieved such that all relevant molecular configurations are enumerated. Converging on sampling the conformational landscape of these complexes imposes a significant cost. This cost is further increased because each snapshot must be energy minimized and the Hessian obtained per snapshot must be constructed and diagonalized. Furthermore, even if the computational cost is affordable, the NMA calculations of biomolecular complexes in solution are carried out following energy minimization in implicit solvent while most likely MD simulations will be carried out in explicit solvent. As such, the energy minimization step will most likely change the conformational ensemble sampled by explicit solvent MD and it is difficult to quantify how much does this affect the accuracy of the calculated entropy. Thus, for NMA calculations applied to an ensemble of molecular snapshots obtained, for example, by MD simulations in explicit solvent, the computational cost can become intractable depending on system size and the number of MD snapshots, while the accuracy of the calculation can also be questioned. As a result, standard MM/PBSA and related approaches in VS tend to run short MD simulations and when NMA-based entropy calculations are carried out, use only some of the sampled snapshots to maintain computational efficiency.

4.5.2: Quasi-Harmonic Analysis

The other well-known method for computing absolute and relative entropies in the RRHO approximation was the one developed by Karplus and Kushick in 1981, called quasi-harmonic analysis.²⁶¹ The QHA method, as opposed to NMA, does not impose the harmonic approximation to the internal degrees of freedom of the system. Instead, by evaluating the atomic fluctuations of the system along the course of a molecular simulation, it is possible to build a matrix containing the variance of each internal degree of freedom on the diagonal and the covariance between each two degrees of freedom in the off-diagonal terms.^{260,304,317}

4.5.2.1: The classical Quasi-Harmonic Analysis approach

The original method starts by connecting the configurational entropy difference between conformations of a molecule with the evaluation of the configurational integral for each of those conformations. This integral is evaluated in internal coordinates, also known as bond-angle-torsion (BAT), plus six external coordinates, which means that a transformation from Cartesian coordinates is required, carried out by introducing a Jacobian matrix.²⁶¹ The configurational integral then reads²⁶¹

$$Q_C = 8\pi^2 V \int_C e^{-\beta U(q')} J(q') dq' \quad (4.32)$$

where $8\pi^2 V$ corresponds to the contributions from the external degrees of freedom, three translational and three rotational and $J(q')$ is the Jacobian term which is a function of the internal coordinates q' .^{78,298} From the internal coordinates, the most important ones correspond to soft degrees of freedom, like dihedral angles. The motions of dihedral angles explore many configurations, opposed to those of angles and bonds which fluctuate very little around the equilibrium value.²⁶¹ Thus, the configurational integral is only evaluated on the soft degrees of freedom while treating the contribution from the hard degrees of freedom as a constant value.²⁶¹ Thus, the configurational entropy associated with a given conformation is given by²⁶¹

$$S_C^q = \frac{\langle V \rangle}{T} + K_B \ln Q_C^q \quad (4.33)$$

where $Q_C^q = Q_C / \text{const}$, const is a constant value which comes from the hard degrees of freedom and $\langle V \rangle$ is the average potential energy considering only the soft degrees of freedom. The entropy ΔS_C^q is a physical quantity which takes into consideration

rotational, vibrational and vibration-rotation contributions and measures the entropy difference between two molecular conformations, like the open and closed states of a protein.²⁶¹ Considering that the PDF of the soft degrees of freedom is of the same form as **Equation 4.2** and that the entropy S_C^q can also be estimated from this PDF, the QHA for this subset of degrees of freedom boils down to the assumption that PES of the soft degrees of freedom of the system can be approximated as a single Gaussian function.²⁶¹ Thus the full PES can be approximated as a single, multi-dimensional, Gaussian function which has the shape of a harmonic potential.^{225,261} In the above case, the full PES is constructed based on the soft degrees of freedom in BAT coordinates. A implementation in Cartesian coordinates was also developed, in which it is possible to carry out QHA calculations²⁹⁶ without the Jacobian transformation. In any of the cases above, since this Gaussian function is dependent on the atomic fluctuations of the systems' degrees of freedom, it means that the PDF of the coordinates $p(q)$ can be written as a function of the covariance matrix^{225,261}

$$p(q) = \frac{1}{2\pi^{\frac{n}{2}} \sigma^{\frac{1}{2}}} e\left[-\frac{1}{2}(q' - \bar{q}')^T \sigma^{-1} (q' - \bar{q}')\right] \quad (4.34)$$

where n is the number of degrees of freedom and σ is the covariance matrix where the atomic fluctuations, sampled by MD or MC simulations around the average structure of the system, are stored.²⁶¹ The terms in the covariance matrix are computed by^{225,260,261,296,304,317}

$$\sigma_{i,j} = \overline{\left((q'_i - \bar{q}'_i)(q'_j - \bar{q}'_j)\right)} \quad (4.35)$$

The QH approximation can be used to extract absolute entropies from ensembles of molecular configurations generated by numerical simulations. To do so, the RRHO-based translational and rotational entropy contributions are added to the vibrational contribution computed using vibrational frequencies extracted by diagonalizing the covariance matrix of atomic fluctuations in QHA.^{225,296} This is possible due to the connection between the Hessian matrix of force constants and the covariance matrix, where $H_{ij} = K_B T (\sigma^{-1})_{ij}$ and σ^{-1} is the inverse of the covariance matrix.^{225,296} The eigenvectors of this pseudo-Hessian matrix describe the QH modes. The approximated QH probability density function in Cartesian coordinates is then written following Karplus and Andricioaei²⁹⁶

$$\tilde{p}(q) = \frac{e^{\left(\frac{-(q-\bar{q})^T \mathbf{H} (q-\bar{q})}{2K_B T}\right)}}{Z} \quad (4.36)$$

To carry out a QHA calculation, (1) the MD snapshots must be centered and superimposed on top of a reference (or average) molecular structure before building the covariance matrix, to decouple translational and rotational contributions from the 3N-6 vibrational degrees of freedom.²⁹⁶ Then, (2) the covariance matrix is built from the atomic fluctuations of these degrees of freedom sampled through numerical simulations with respect to an average structure.²²⁵ Upon diagonalizing the covariance matrix (3), the independent eigenvectors describing the displacements of each degree of freedom around the average configuration are obtained, along with the eigenvalues λ_i which can be plugged into **Equation 4.12** to obtain the vibrational entropy contribution of each QH oscillator within the QM treatment ($v_i = \sqrt{K_B T / \lambda_i}$).^{225,261,296} As for the rotational and translational degrees of freedom, their contribution to the absolute entropy is computed using **Equations 4.10** and **4.11**.²²⁴ From here, absolute molecular entropies may be accessed. However, approximating the rugged PES using a multi-dimensional Gaussian function means that the obtained entropies are in fact strictly upper bounds to the true entropy.^{261,296} Furthermore, these calculations converge slowly and require extensive sampling, which increases with the complexity of the system studied.^{225,260} Finally, while no energy minimization is required prior to the building of the covariance matrix, diagonalizing it for a large macromolecular system is still a computationally intensive operation. As such, one can enumerate three main limitations of QHA: The need for extensive sampling (high number of molecular snapshots) which may become prohibitive for very large systems, the lack of high-order correlation terms and the overall accuracy of the method due to the approximation of the PES a single Gaussian well.

One advantage of QHA over NMA is that since there is no need to perform energy minimization, when the water molecules are removed from the simulation box their effect on the solute is still implicitly captured in the fluctuations of the solute degrees of freedom.²¹⁸ As such, QHA opens to the calculation of entropies in solution without any additional computational overhead. However, while in NMA one single energy-minimized snapshot is needed for the calculation in QHA the minimum number of snapshots required varies with system size but is always more than one. Recently, QHA has been used to compute the entropy difference between two conformations of the Heat Shock Protein 90 (HSP90), the loop-in conformation and the helical conformation.³¹⁸ The entropy of binding for a series of 20 compounds was estimated for each conformation by considering the solvent entropy contribution and the solute entropy differences between

compounds binding to either conformation by QHA or a version of the MLA method.³¹⁸ Another example of application of QHA is in the study of the entropy of stapled peptide inhibitors. In this work, researchers studied the binding between postsynaptic density protein 95 (PSD-95) and the SAPAP/Shank complex with the aim of inhibiting the phosphorylation ability of the guanylate kinase (GK) domain of PSD-95. Several peptides were developed and simulated using MD. Their binding affinities towards the GK domain of PSD-95 were estimated using the MM/PBSA method and the entropy was estimated via QHA.³¹⁹

4.5.2.2: Schlitter's method

Another approach within the RRHO approximation which allows the calculation of the vibrational contribution to the absolute entropy using the QH approximation and the QM oscillator is known as Schlitter's formula³²⁰. Within the RRHO formalism described in **Equation 4.12**, the entropy of a QM harmonic oscillator is given by **Equation 4.37**^{225,296,304,321}. The Schlitter formula assumes that the vibrational entropy of a molecular system in Cartesian coordinates can be approximated as a sum of individual contributions from QM oscillators while introducing an approximation to **Equation 4.12**. The contribution of one QM oscillator according to this formalism is given by **Equation 4.38**^{296,320}

$$S_{ho} = \frac{K_B \alpha}{e^\alpha - 1} - \ln(1 - e^{-\alpha}) \quad (4.37)$$

$$S_{ho,schlitter} = \frac{K_B}{2} \ln \left(1 + \frac{e^2}{\alpha^2} \right) \quad (4.38)$$

with $\alpha = \frac{\hbar v_i}{K_B T}$. According to Schlitter, the variance of one degree of freedom $\langle \Delta q^2 \rangle = \frac{K_B T}{m v^2}$, is related its vibrational frequency as per the equipartition theorem.^{225,296,320} Thus, Schlitter³²⁰ and later Karplus and Andricioaei²⁹⁶ realize that α can be written as a function of the variance. Thus, we arrive at the master equation for Schlitter's method for one QM harmonic oscillator³²⁰

$$S_{ho,schlitter} = \frac{K_B}{2} \ln \left(\frac{K_B T}{\hbar^2} m e^2 \langle \Delta q^2 \rangle + 1 \right) \quad (4.39)$$

The above **Equation 4.39** is then extended to all degrees of freedom of the system, by introducing a mass-weighted covariance matrix σ_m and the expression then becomes

$$S_{sch} = \frac{K_B}{2} \ln \left(\left| \frac{K_B T}{\hbar^2} e^2 \boldsymbol{\sigma}_m + \mathbf{1} \right| \right) \quad (4.40)$$

as described by Suárez and Díaz.²²⁵ The above **Equation 4.40** is analogous to the one described by Karplus and Andricioaei, which is written as²⁹⁶

$$S_{sch} = \frac{K_B}{2} \ln \det \left(\frac{K_B T}{\hbar^2} e^2 \mathbf{M} \boldsymbol{\sigma} + \mathbf{1} \right) \quad (4.41)$$

where $\mathbf{1}$ is the unity matrix and \mathbf{M} is the mass matrix. In any case, because a covariance matrix must be built and diagonalized to obtain vibrational frequencies, the QH approximation is present in the Schlitter method. As in the QHA method when the QM oscillator is used, **Equations 4.39** and **4.40** obtain the correct QM limit for entropic contributions arising from high-frequency motions, meaning that the vibrational contributions of these degrees of freedom are negligible. However, it also keeps the limitations of the original QHA method such as the slow convergence, the large overestimation of the absolute entropy and the need for extensive sampling. Furthermore, as noted by Karplus and Andricioaei, the Schlitter formula is an upper bound to the true entropy as in QHA but QHA is a stricter upper bound estimate.²⁹⁶

4.5.2.3: Boltzmann Quasi-Harmonic method

A third method based on QHA is the method developed by DiNola³⁰⁵ and further expanded by Sharp and Harpole³⁰⁴, named Boltzmann Quasi-Harmonic (BQH). The BQH calculation is carried out in BAT coordinates following configurational sampling by MD simulations. From the QH expression for the configurational entropy difference³⁰⁴

$$\Delta S_{config} = \frac{K_B}{2} \ln((2\pi e)^n |\boldsymbol{\sigma}|^2) \quad (4.42)$$

then Sharp and Harpole factor the determinant of the covariance matrix $|\boldsymbol{\sigma}|^2$ into a product of two terms: The matrix containing the diagonal elements \mathbf{D} and a determinant of the matrix \mathbf{C}_{ij} , such that³⁰⁴

$$|\boldsymbol{\sigma}|^2 = \prod_i^n \sigma_{ij}^2 |\mathbf{C}_{ij}| = \mathbf{D} |\mathbf{C}_{ij}| \quad (4.43)$$

where the elements of \mathbf{C}_{ij} are the covariances in the covariance matrix $\boldsymbol{\sigma}$. Substituting Equation 4.43 in 4.42, one arrives at

$$\Delta S_{config} = \frac{K_B}{2} \ln((2\pi e)^n \mathbf{D} |\mathbf{C}_{ij}|) \quad (4.44)$$

where the $\frac{K_B}{2} \ln((2\pi e)^n \mathbf{D})$ terms corresponds to the correlation-free entropy contribution from the elements in the diagonal matrix and the term containing \mathbf{C}_{ij} is a term accounting for the pairwise correlations, obtained by QHA. In the QH approach, both terms are computed by approximating the PES of a molecular system as a multidimensional Gaussian. In the BQH model, the diagonal matrix is replaced by the first-order marginal entropies computed through the configurational integral in Equation 4.2^{225,246,304} for all internal degrees of freedom. The evaluation of the integral to compute the marginal entropy of each internal degree of freedom implies that their continuous probability distribution of must be discretized by, for example, histogramming methods. In principle, the BQH method is expected to be more accurate than standard QHA as the diagonal terms are evaluated properly based on the probability distributions obtained by MD and pairwise corrections due to correlations are still considered.³⁰⁴ However, it still lacks high-order correlation terms which can play a significant role in large biomolecular systems. Recently, a benchmark study on the accuracy of different methods based on QHA was carried out by the group of Ikeguchi.³⁰² One of the aims of the study was evaluating the accuracy of different methods to compute the configurational entropy for systems of increasing complexity, from butane to protein A, following Replica-Exchange MD simulations carried out from 260 to 600K in implicit solvent.³⁰² The calculated entropies were estimated at 300K. The reference entropies were obtained using an equation derived from Clausius method as Harpole and Sharp.³⁰⁴ It was found that BQH configurational entropies deviated the least from the reference data in implicit solvent, partly due to a correction added by authors such that contributions from improper dihedrals were also considered.³⁰² It was also found that in the other methods tested, which employ the QM oscillator, the results show poor quantitative agreement with the reference entropies.³⁰² The authors explain that the application of the QM oscillators to study protein A and the Trp cage, is not proper because the treatment of protein degrees of freedom as harmonic oscillators is a severe oversimplification of the dynamical motion of the system in which the Gaussian approximation breaks down.³⁰²

4.6: Going beyond the Rigid-Rotor Harmonic-Oscillator approximation: The Mutual Information Expansion method

The RRHO approximation works properly for rigid and flexible small molecules, using the formalism of the “mixture of conformers” approach, in the cases where higher-order correlations do not come into play. However, to access absolute or relative entropies for larger molecules, other methodologies are more suitable due to the sampling limitations discussed before. These methods, known as non-parametric methods²²⁵, are able to estimate configurational entropies of more complex systems and do not assume anything about the shape of the PDF of the atomic coordinates of the molecules.²²⁵ Instead, these approaches focus approximating the configurational integral (**Equation 4.1**) by computing lower-dimensional integrals which encode the entropies of subsets of internal degrees of freedom, truncated up to a given order. One such example is the Mutual Information Expansion (MIE) method, developed in 2007 by Killian *et al.*, within the group led by Michael Gilson.²⁹³ It approximates the full configurational integral over the 3N-6 internal coordinates by writing the PDF in terms of one-dimensional, two-dimensional or up to n-dimensional probability distributions.^{293,298,322} In particular, the full entropy is approximated as a sum of first-order entropies and mutual information terms which correct for correlation between groups of internal degrees of freedom.³⁰³ Starting from the expression in **Equation 4.4**, and transforming the systems coordinates from Cartesian to BAT coordinates, Killian *et al.*²⁹³ write the configurational integral as

$$S = -R \int p(q'_1, q'_2, q'_3 \dots q'_i) \ln(p(q'_1, q'_2, q'_3 \dots q'_i)) J(q'_1, q'_2, q'_3 \dots q'_i) dq'_1 \dots dq'_i \quad (4.45)$$

where the internal coordinates defined by bond length (b), bond-angle (θ) and dihedral angle (φ) as in Killian *et al.*²⁹³ are replaced by q' . However, evaluating this integral requires a significant amount of sampling which is prohibitively expensive for large macromolecular systems.²⁹³ Thus, it must be approximated by relying on converged lower-dimensionality PDFs and using Kirkwoods' generalized algorithm (GSKA)³²³ to include mutual information terms between internal degrees of freedom. A first-order approximation is given by **Equation 4.46**^{225,293}

$$S^1(q'_i) = -R \int p(q'_i) \ln(p(q'_i)) J(q'_i) dq'_i \quad (4.46)$$

where correlations between degrees of freedom are not considered and the entropy is computed by summing the marginal entropies of all internal degrees of freedom,

$$S^1 = \sum_{i=1}^m S^1(q'_i) \quad (4.47)$$

meaning that the simplest manner to approximate the expensive full PDF is by a sum of individual one-dimensional PDFs.^{225,293} However, since correlations between motions do play a role in large biomolecular systems, the above expression is incorrect for all but the simplest systems.^{225,297,322} To include second-order correlation terms within the configurational entropy estimation, first the marginal PDF of each degree of freedom in the pair is computed and then the joint PDF of the pair is estimated.^{293,297,298,322} From these PDFs the marginal and the pair-wise entropies are calculated. As such, Killian *et al.* arrive at **Equation 4.48**^{293,297,298,322}

$$S^2(q'_1, q'_2, q'_3) = S^2(q'_1, q'_2) + S^2(q'_1, q'_3) + S^2(q'_2, q'_3) - S^1(q'_1) - S^1(q'_2) - S^1(q'_3) \quad (4.48)$$

where $S^2(q'_1, q'_2)$ is the joint entropy of the pair of internal degrees of freedom q'_1, q'_2 . Introducing the mutual information expression among pairs of degrees of freedom (**Equation 4.49**)^{276, 281,310}

$$I^2(q'_i, q'_j) = S^1(q'_i) + S^1(q'_j) - S^2(q'_i, q'_j) \quad (4.49)$$

allows re-writing **Equation 4.48** into **Equation 4.50**, where $I^2(q'_i, q'_j)$ describes the degree of correlation between any two degrees of freedom.²⁹³ The degree of correlation between internal degrees of freedom is a measure of how much knowledge can be gained about j when the distribution of the i^{th} degree of freedom is completely known.

$$S^2(q'_1, q'_2, q'_3) = S^1(q'_1) + S^1(q'_2) + S^1(q'_3) - I^2(q'_1, q'_2) - I^2(q'_1, q'_3) - I^2(q'_2, q'_3) \quad (4.50)$$

For expansions of higher order, Killian *et al.*^{293,281} provide a general expression

$$S^{m-1} = \sum_{i=1}^m S^1(q'_i) - \sum_{C_2^m} I^2(q'_i, q'_j) + \sum_{C_3^m} I^3(q'_i, q'_j, q'_k) - \sum_{C_4^m} I^4(q'_i, q'_j, q'_k, q'_l) \dots \quad (4.51)$$

where the expansion is truncated at a given order $m-1$. The general expression in **Equation 4.51** can be applied to estimate the configurational entropy of a molecule at any truncation order, including correlations between motions up until that point²⁹³ In practice, going to higher order correlations is complicated because the higher the order of the truncation, the more sampling is required to build the multi-dimensional PDFs, whose cost increases with the combinatorial explosion in explorable molecular configurations.^{225,293} When the internal degrees of freedom are assumed as independent, the computational cost of populating the one-dimensional PDFs is small and highly tractable.²⁹⁷ When this is not the case, and thus correlation between motions must be accounted for, it is necessary to truncate the expansion above to avoid the skewed histogram problem which arises when populating histograms built from high-dimensional PDFs. Thus, the truncation order must be selected to balance out the accuracy of the calculation and the available computational resources.²⁹³ Nonetheless, it is a useful approach to estimate configurational entropies without any assumption on the shape of the configurational PDF.

One interesting observation is that while second-order correlations are always correcting the S^1 estimates to lower values, the higher-order terms either increase or decrease the entropy depending on the sign²²⁵, and thus the calculation may not converge. Additionally, application of the MIE in a truncated form is bound to introduce an error which must also be evaluated.³⁰³ The MIE approach was demonstrated in the ACCENT-MM tool by the Gilson group³²². For a set of small-molecules in gas-phase, the entropy was estimated by the MIE in internal coordinates until third-order and compared to reference data from the Mining Minima method (M2).³²² The first step in the MIE calculation is the discretization of the PDFs by histogramming. Then, one to three-dimensional PDFs were built from the histograms and the mutual information terms for pairs and triads of internal degrees of freedom were computed.^{225,322} It was demonstrated that the second-order entropies converge quickly for a series of alkanes, in 50ns MD simulations, whereas the third-order entropies converged slowly and required much more sampling.²⁹³ The first-order results for the alkanes were significantly different from the M2 predicted entropies. However, the second-order entropies agreed fairly well with M2 results ($T\Delta\Delta S$ between 1.0 and 2.0 kcal/mol) which highlights the important contributions given by the correlation corrections at second order.²⁹³ The third-order entropies, on the other hand, deviate significantly from the M2 data. For the protein-urea system, it was found that while the second order MIE expansion agreed with the M2 results, with a $T\Delta\Delta S$ of 1.0 kcal/mol, the third-order estimates differed significantly from M2 data.²⁹³ In both cases, the third-order entropy estimations are plagued with sampling problems, which is what prevents the convergence of the entropy

estimations at that level and is the a major concern in MIE calculations when the expansion is led towards higher-order terms.²⁹³ Application of ACCENT-MM to evaluate the configurational entropy difference at the second order for a protein-peptide system required 2 million snapshots collected, from 200 10ns MD simulations, which meant that 2 μ s of MD were necessary for this entropy calculation.³²² Thus, a limitation of MIE is highlighted: The need for extensive sampling, which increases significantly with truncation order.

4.7: Merging Rigid Rotor Entropies and non-parametric estimates of conformational entropy: the CENCALC approach to Mutual Information Expansion

Within the “mixture of conformers” approach, the absolute entropy of a molecule can be obtained by a sum of the probability-weighted RRHO entropy and the entropy of mixing. While the first term is readily obtainable, the entropy of mixing is either obtained by clustering the MD trajectory and extracting cluster occupancies or by expansion approaches. One strategy to compute S_{mix} without relying on clustering methodologies is to try and approximate the full conformational integral by a series of lower dimensionality integrals of dihedral degrees of freedom through discretization and histogramming like in MIE.³⁰³ As explained by Suarez *et al.*,³⁰³ the conformations a molecule can explore within its PES can be enumerated by discretizing the time series of the dihedral degrees of freedom.³⁰³ In CENCALC, the PDF for each torsion is estimated from the MD simulation using a von Mises kernel in which $\theta \in [0, 2\pi]$.³⁰³ From the PDF, one has to find the basins in the range of 0 to 2π which correspond to energy minima in the PES.³⁰³ These correspond to the conformational states populated by each torsion which must then transformed into a set of integer numbers which label the conformational states a given torsion explores, thus discretizing the PDF.³⁰³ However, due to the computational cost associated with extensively sampling the free energy landscape to be able to populate adequately the higher-order histograms³⁰³, employing a pure MIE to calculate S_{mix} is not recommended.³⁰³ Truncating the expansion alleviates this overhead but at the same time introduces a truncation error. Furthermore, it is also reported that even in the case of perfect sampling, the truncation errors are difficult to handle and do not decrease as the calculation goes to higher and higher expansion orders.³⁰³ One alternative approach is to introduce a distance threshold such that only the dihedral angles which are within a certain distance are included in the mutual information expansion around a given torsion at a specific truncation order.³⁰³ Thus, a

more efficient version of the MIE method can be written as in **Equation 4.52** for any order of truncation and employing a distance cut-off scheme as

$$S^n(q') = \sum_{k=1}^n (-1)^{k-1} \sum_{\substack{\mathcal{J} \subset \mathcal{C}(R) \\ |\mathcal{J}|=k}} I_k(\mathcal{J}) \quad (4.52)$$

where the ensemble of dihedral angles considered for the expansion calculations is restricted to the closest ones by the $\mathcal{C}(R)$ function, which is based on the Euclidean distance between two dihedral angles and a distance threshold R . However, employing the cut-off based MIE approach implies determining the cut-off, which is an arbitrary parameter, and still calculations may remain expensive.³⁰³

Thus, other approaches were explored, including a variant of MIE which is much more efficient called Approximated-MIE (AMIE). In the AMIE method, **Equation 4.52** is modified such that the mutual information expansion is computed for a non-redundant list L .³⁰³ This list is constructed by a series of lists (L_i), one per each dihedral degree of freedom where the first element is the dihedral i itself and the other elements of the list are selected to be the dihedrals in which the distance between i and any j is smaller than R and which are after the dihedral i .³⁰³ For example, if dihedral i is numbered 2, the correlation will be computed at the second order between dihedral 2 and all dihedrals from 3 to N whose Euclidean distance to dihedral 2 falls within R .

The AMIE formulation allows significant speed-up in the calculations because the expansion is only computed once between sets of dihedrals and only for those dihedrals whose Euclidean distance falls within a given threshold from a list is ordered in such a way that each calculation is carried out only once and stored. The AMIE method is one of the methods implemented in CENCALC to estimate conformational entropies in an accurate and efficient manner.^{225,308,325} Other methodologies which are used to compute the entropy of mixing, or conformational entropy, are those based on the Multibody Local Approximation (MLA).³²⁵ While the “mixture of conformers” approach looks appealing for small-molecule systems, it may quickly become intractable to compute S_{mix} for large and even moderate-size molecular systems due to the combinatorial explosion of explorable molecular configurations.³⁰³ The CENCALC approach, on the other hand, provided that sufficient sampling is available, is able to reach systems of that size and open to more accurate calculations of larger systems by using a combination of NMA-based RRHO entropies and efficient mutual information-based expansion approaches for the S_{mix} .

4.8: Benchmarking entropy calculations

Usually, RRHO-based molecular entropy calculations are validated by benchmarking against experimental entropies in gas-phase.^{225,313} These experimental data, however, are only available for small molecules which are mostly rigid and do not exhibit any significant higher-order correlations among their internal degrees of freedom.²²⁵ Thus, for more flexible small molecules in solution, which are typically the object of study in VS campaigns, appropriate absolute entropy data is lacking.²²⁵ Nonetheless, the gas-phase benchmark data can still be useful and, in the past, has been used to probe the accuracy of different methods, ranging from those based on MD simulation data to QM calculations. The experimental entropies in gas-phase are obtained from calorimetric experiments.²⁹¹ To compare between computed absolute entropies and experimental data, it is important to account in the theoretical calculations for the entropy arising from the rotation of dihedral angles containing terminal symmetric groups such as methyls.^{225,326} One way to do so is by clustering MD simulation using a small cut-off such that the microstates that are generated by this torsional motion are separated. For example, the rotation of a dihedral ending in a terminal methyl group produces three microstates if we consider each hydrogen as distinguishable from the others. Thus, hydrogen-1 can be in three different positions.²¹⁸ Accounting for this behavior is critical for the accurate calculation of absolute entropies, even more so in QHA calculations where the PES of this torsion would be approximated by a single state and introduce a significant error in the calculation. Another important point to raise is the proper determination of the moments of inertia of the molecule and the symmetry number, as these impact directly the rotational contribution to the absolute entropy and thus the predicted absolute entropy.²⁹² A very useful resource where experimental gas-phase entropies are available is the Computational Chemistry Comparison and Benchmark DataBase (CCCBDB) from the National Institute of Standards and Technology (NIST)³⁰⁷, a repository for experimental and *ab initio* thermochemical properties for small molecules.³⁰⁷ The experimental absolute entropies there available correspond to calorimetric entropies estimated using Clausius' formula for small molecules which contain the following chemical groups: alkanes, alcohols, ethers, acids, esters, ketones, aldehydes, epoxides, chalcogens, nitrites, nitrates, amides, amines, aromatic rings, phenyl rings, among others.³⁰⁷ The newest version also includes bromine-containing molecules.³⁰⁷ The experimental data available in NIST is drawn from multiple sources such as JANAF³²⁷, Gurvich³²⁸ and the TRC data series.³²⁹

Recently, Stefan Grimme and Philip Pracht³¹³ have developed a methodology to obtain absolute molecular entropies and heat capacities, using QM calculations at various levels of theory to generate the conformational ensemble for each molecule and employing

RRHO-based entropy calculations for the vibrational contribution as part of the CRESSET software.³¹³ They selected a set of 39 organic compounds compiled by Head-Gordon *et al.*³³⁰ and, since the aim was to have a representative dataset with flexible molecules, the dataset was further extended by merging with a dataset named A23, containing 23 molecules which are larger than those in the Head-Gordon dataset.³¹³ The range of compounds queried was from ethane to n-dodecane, and experimental absolute gas-phase entropies were retrieved for these compounds.³¹³ Additionally, another set of linear alkanes up to 18 carbon atoms ($C_{18}H_{38}$) was also selected to evaluate method limitations.³¹³ The method proved to be highly accurate, with an RMSD of $0.84 \text{ cal mol}^{-1}K^{-1}$, with the conformational entropy computed using the GFN-FF model, a semi-empirical forcefield which is accurate but amenable to study large macromolecular systems.³¹³ Similar accuracy values were reported by Guthrie for a set of 128 organic compounds at the B3LYP/6-31** level of theory (RMSD = $1.24 \text{ cal mol}^{-1}K^{-1}$), although many of these compounds were rigid molecules where we have already described that simple RRHO calculations achieve sufficient accuracy.^{311,313} Analogously, DeTar studied the heat capacity and entropy of a set of representative hydrocarbons, including compounds containing only methyl group rotations, compounds spanning multiple conformers arising from internal dihedral rotations and compounds without methyl groups.³¹² The approach used by DeTar, which combines RRHO calculations and QM geometry optimizations, generating vibrational frequencies for each conformer of each compound at different theory levels, yielded quantitative agreement with the reported experimental data (RMSD = $0.36 \text{ cal mol}^{-1}K^{-1}$).

On the other hand, absolute entropies of molecules in solution are scarce. However, what is possible to measure is the entropy difference upon binding of a ligand to a protein in solution. These entropies of binding can be accessed through ITC.⁷⁸ The entropy change upon binding which is evaluated by these means is not a pure solute entropy change and instead contains both solute entropy and a contribution from the solvent degrees of freedom.²²⁵ This is because as a molecule is solvated, the solvent degrees of freedom in the first solvation shells around the molecule will orient themselves to interact with the solute. Establishing these interactions implies a loss of configurational freedom and thus is an entropic cost. It is also possible to obtain entropy changes upon binding using NMR relaxation data through the order parameter.^{78,225} When appropriate data is not available, it is also possible to compare predicted entropies with theoretical benchmarks coming from rigorous free energy calculations.²²⁵ However, these methods are typically very expensive and can include numerical errors and errors due to finite sampling^{331,332}. Nonetheless, these free energy methods present an attractive alternative to obtain benchmark data in solution by alchemically decoupling the ligand in vacuum

and in solution, and then computing the free energy of solvating the ligand, and from the free energy estimating the entropic term.²²⁵

Part III – Results

5. A multi-basin quasi-harmonic approach for the calculation of the configurational entropy of small molecules in solution

Disclaimer: Most of this chapter was taken from Pereira & Cecchini. 2021. Journal of Chemical Theory and Computation, 17,2, 1133-1142. It corresponds to an original work developed by Gilberto Pereira under the supervision of professor Marco Cecchini at the University of Strasbourg.

5.1: Introduction

The free energy is widely acknowledged as the driving force in fundamental biological processes¹¹⁵, such as protein folding and protein-ligand binding. In drug discovery, it is often unfeasible to determine (experimentally) the binding affinity of all entries of a chemical library for a given target. As such, there has been a tremendous effort to develop computational methods able to estimate ligand-binding free energies accurately and efficiently from first principles.¹⁸⁹ Currently available methodologies encompass a broad spectrum where both the computational cost and the quality of the predictions vary considerably.¹⁸⁹ In between rigorous binding free-energy calculations and semi-empirical approaches like docking, an intermediate class of methods, referred to as *end-point*, has attracted significant interest. Among them, the popular Molecular Mechanics/Poisson-Boltzmann Surface Area or the Molecular Mechanics Generalized-Born Surface Area methods aim at the ligand-binding affinity via quantification of the absolute chemical potentials of the protein, the ligand, and the protein-ligand complex,¹⁹⁶ which requires the numerical evaluation of absolute molecular entropies in solution. Because of the computational burden and the empirical observation that introducing entropy contributions often worsens the correlation with experiments,^{211,273} many researchers do not include entropy in their MM-PB(GB)SA calculations.

To evaluate the configurational entropy of single molecules, several methodologies exist; they have been thoroughly reviewed elsewhere.²²⁵ In the limit of the ideal gas and the RRHO approximation, absolute molecular entropies can be quantified by statistical mechanics formulas²²⁴. Although approximated, this approach is efficient and relies on a small number of molecular parameters that can be easily accessed by modeling, typically geometry optimization and NMA. The validity of these results, however, is limited to rigid molecules populating a single conformational state and to the gas phase. To go beyond these approximations, a significant effort has been made. A popular approach that

extends the scope of *ab initio* entropy calculations to solution conditions is the QHA of the internal molecular motions, introduced by Karplus and Kushick forty years ago²⁶¹. In this approach, the vibrational frequencies are obtained from the analysis of the room temperature atomic fluctuations sampled by MD. Since the simulations can be carried out with an explicit treatment of the solvent, these calculations incorporate not only part of the anharmonicity but also solvent contributions. Nonetheless, standard QHA is prone to fail in the context of flexible molecules that exist in equilibrium between multiple conformers and was shown to provide at best an upper bound estimate to the absolute molecular entropy.²⁶⁰

Molecular flexibility has been tackled by the *mixture of conformers* approach, also introduced by Karplus and coworkers to quantify the entropy of denaturation in proteins.³²¹ Assuming that a flexible molecule has N distinct conformations and that each of them can be treated as a disjoint multi-dimensional harmonic well, this model includes two contributions to the total configurational entropy: one that is due to the local fluctuations in the neighborhood of a well-defined molecular structure, and another one corresponding to the existence of more than one structure or energy well on the landscape.²²⁵ In this approach, the former, which is often referred to as the *per-basin entropy*, is determined in the RRHO approximation as an ensemble average over multiple conformers, whereas the latter, which is referred to as the *entropy of the landscape* or the *entropy of mixing*, is evaluated using Gibbs's formula of entropy from the equilibrium probabilities of the accessible states.²²⁵ Using this approach in conjunction with DFT calculations at the B3LYP/6-31G* level of theory, Guthrie predicted the gas-phase entropy of 128 organic compounds with up to 10 carbon atoms for which calorimetric entropies were available.³¹¹ Despite the use of an approximated formula for the entropy of the landscape, i.e. all conformers were assumed as equally probable, and the significant amount of manual work for both conformer enumeration and symmetry determination, the numerical results were extremely accurate with a standard deviation from the experiments of 0.38 kcal/mol at 298.15K. Using the same approach, DeTar analyzed a dataset of eighteen hydrocarbons in the gas phase featuring freely rotating groups, internal rotation and/or multiple conformational states.³¹² By computing the vibrational frequencies at the Hartree-Fock level of theory with various basis sets (3-21G, 6-31G*) or MP2/6-31G(d,p) and evaluating the entropy of the landscape via Boltzmann probabilities from quantum mechanical energies, an impressive RMSE of 0.1 kcal/mol was obtained.³¹² More recently, absolute entropy calculations for a dataset of eight alkanes, five dipeptides and the PFG hexapeptide were reported by Suarez et al.³⁰⁸ In this work, an original strategy that combines NMA of a large number of molecular snapshots from MD with a mutual information treatment of the n -order correlations between the torsional degrees of freedom was developed to estimate the entropy of the landscape.

³⁰⁸ Suarez *et al.* report that such methodology is efficient, in particular when the vibrational analysis is done by molecular mechanics.³⁰⁸ This method is also automatic, since conformer enumeration and weighting are obtained by clustering of the MD trajectory, and accurate, exhibiting a mean unsigned error (MUE) of 0.31 kcal/mol. Importantly, the method was shown to cover a large chemical space, i.e. from small hydrocarbons to large and flexible peptides, and to provide numerical estimates of the absolute entropy in solution in combination with an implicit solvent model.³⁰⁸

Alternative approaches that go beyond the single-well harmonic model have been explored.³²⁶ One of them is the second generation M2 method by Gilson and coworkers that provides absolute molecular entropies from the calculation of configurational integrals for a large set of conformers sampled by an automated configurational search.³²⁶ These configurational integrals provide a numerical estimate of the molecular partition function, which is used to quantify the absolute molecular entropy indirectly as a difference between the system free energy and its internal energy.¹⁹⁹ In this method, solvent effects are included via an implicit solvent model and the anharmonicity within the well is accounted for by numerical integration of the lowest-frequency modes.¹⁹⁹ Another example is the mutual information expansion also by Gilson in which absolute molecular entropies are accessed by an expansion of mutual information terms to the n^{th} order that account for the correlation between the internal degrees of freedom (bonds, angles and torsions) of the molecule.²⁹³ Interestingly, MIE never imposes the harmonic approximation but uses a truncated form of the expansion to ensure convergence of the calculation. In addition, it provides absolute molecular entropy estimates directly from the analysis of a single MD trajectory, possibly including solvent effects. To the best of our knowledge, none of these methods was benchmarked against calorimetric entropies.

Here, we present a novel approach for the numerical evaluation of absolute entropies of small molecules in vacuum or in solution. Our method, named quasi-harmonic multi-basin (QHMB), improves the accuracy of standard QHA calculations by using a multi-basin decomposition scheme based on clustering of MD simulations. The absolute entropy of the molecule is then evaluated through the “mixture of conformers” approach. This way, the potential energy surface is no longer approximated by a single multivariate Gaussian distribution and the contributions from all conformers visited during MD are correctly taken into account. By comparing with calorimetric data, we show that QHMB is able to predict absolute molecular entropy in the gas phase with an almost perfect correlation with experiments. Furthermore, the introduction of a QHMB correction in MM/PBSA to account for the configurational entropy loss of the ligand upon binding is shown to improve the correlation between calculated and experimental binding affinities.

In the following, we review the theory beyond the QHMB method, we present the datasets used for tests calculations, we compare the performances of QHMB relative to

standard RRHO-based entropy calculation approaches, and discuss the significance of these results in the context of protein-ligand binding affinity calculations.

5.2: Material and Methods

5.2.1: Benchmark datasets

5.2.1.1: Dataset for gas-phase entropy calculations.

To evaluate the performance of the QHMB approach, a dataset of 22 small molecules with experimentally determined gas-phase entropies were selected from the Computational Chemistry Comparison Benchmark DataBase³⁰⁷ from NIST; see **Table S5.1**. In addition to those, heptane was included in the dataset, whose gas-phase entropy was extracted from the TRC data series.³²⁹ Overall, our dataset includes several linear alkanes, and both aromatic and halogenated compounds with up to 10 heavy atoms. In addition, they feature a different number of rotatable bonds and span a wide range of experimental gas-phase entropies. For comparison, a subset containing only the most flexible compounds, i.e. those with more than three rotatable bonds, was also considered.

5.2.1.2: Dataset for ligand-binding free energy calculations.

To quantify the impact of the QHMB entropy correction on the performance of endpoint binding free energy calculations, a dataset of 21 protein-ligand complexes was selected from the Greenidge dataset.²¹⁴ This dataset spans a wide range of ligand binding affinities (pKi from 2.48 to 10.49), molecular weights (MW from 101.1 to 376.43 g/mol), and number of rotatable bonds (from 1 to 8); see **Table S5.2**. Although not thorough, this collection is diverse and features rather flexible ligand flexibilities, which makes it challenging for any computational method. In addition, all protein-ligand complexes but one were stable in explicit-solvent MD, i.e. the protein backbone RMSD was lower than 2.5 Å over 100 ns for 20 out of 21 complexes (**Figure S5.1**), which ensures convergence of the MM/GBSA calculations. Marginal stability was only observed for 1P1N, whose RMSD from the crystallographic binding mode rises to 3Å.

5.2.2: MD simulations

5.2.2.1: Gas-phase Molecular Dynamics simulations

MD simulations in the gas phase were carried out using Amber18 and the GAFF2⁵¹ force field in a vacuum. All compounds were first drawn and 3D coordinates generated using MarvinSketch from ChemAxon.²⁵ The resulting sdf file was then submitted to antechamber³³³ for the calculation of partial atomic charges using the RESP method¹⁷¹ at the HF/6-31G* level of theory³³⁴ in Gaussian09.¹⁷² Finally, force-field parameters were obtained using tleap in Amber18.¹¹⁸ MD simulations were carried out using sander for 300 ns, at 298.15K in the NVT ensemble using a Langevin thermostat. Molecular snapshots were saved each 5ps for a total 60000 molecular configurations for further analysis. The sander code was employed for these simulations.

5.2.2.2: Solution Molecular Dynamics simulations

Starting from optimized geometries in vacuum, all compounds from NIST were embedded in a periodic octahedral TIP3P water box that extended 18Å from the solute. Before running MD, the solvent molecules were relaxed by three cycles of energy minimization (1000 steps each). Then, the full system was energy minimized (5000 steps) using a combination of the steepest descent and conjugate gradient algorithms. The molecular systems were gradually heated up to 298.15K over 2ns in the NVT ensemble using Langevin dynamics and an integration time step of 2 fs. SHAKE was used to constrain all covalent bonds involving hydrogens and particle-mesh-Ewald for the treatment of the long-range electrostatic interactions. Then, the system was equilibrated over the course of 2ns employing Langevin dynamics in the NPT ensemble, using a Monte Carlo barostat, with 1 atm pressure at the temperature of 298.15K. Production runs were carried out for 300ns in the NPT ensemble, and molecular snapshots were saved every 5ps for a total of 60000 molecular snapshots for further analysis. The pmemd.cuda code was employed for these simulations. For the simulations of the protein-ligand complexes from the Greenidge dataset²¹⁴, the same protocol was applied, adding counter ions to the simulation box to ensure charge neutrality and Na⁺ and Cl⁻ ions at 0.15M to mimic the physiological conditions.

5.2.3: Entropy calculations

For the absolute molecular entropy calculations, several approaches were explored. They are all based on the RRHO approximation and use NMA and QHA to evaluate the intramolecular entropy along with a number of variants to account for the anharmonicity of the underlying potential energy surface including the existence of multiple energy wells. For each compound, the molecular weight and the symmetry number were obtained from VMD³³⁵, whereas the moments of inertia and the vibrational frequencies for the calculation of the per-basin entropy were determined using AmberTools18.¹¹⁸ All entropy calculations (**Equation 5.4** and **Equation 5.10**) were carried out using Thermo.²⁹⁴ For the calculations in vacuum the standard volume was set to 24.78L. For those in solution, a standard 1M concentration was considered.

5.2.3.1: Normal-Mode Analysis.

NMA calculations were performed using a deeply optimized molecular geometry. For this purpose, molecular coordinates were energy minimized until the root mean square gradient of the energy was $< 10^{-5}$ kcal/mol/Å. The vibrational frequencies were then determined from the eigenvalues of the mass-weighted Hessian in Cartesian coordinates (NMA). Alternatively, NMA was carried out starting with the structure of the most populated cluster based on the statistical distribution of the non-redundant dihedral angles (NMA-clust). The coordinates of the cluster center were energy minimized and vibrational frequencies collected as described above. Last, another variant that evaluates the intramolecular entropy by averaging the RRHO entropy over a series of molecular configurations was also considered (NMA-multi). For this purpose, ten molecular snapshots were extracted from an MD trajectory, energy minimized and analyzed.

5.2.3.2: Quasi-harmonic analysis.

QHA was carried out using the molecular snapshot closest to the average structure in MD as reference. For this purpose, the mass-weighted covariance matrix was determined in Cartesian coordinates after optimal superimposition of the MD trajectory on the reference and diagonalized to obtain the vibrational frequencies (QHA). One advantage of this approach is that some anharmonicity and solvent effects are effectively captured in the entropy calculation. Alternatively, a QHA variant that uses the structure of the most populated cluster in dihedral space as reference (QHA-clust) was also explored.

5.2.3.3: Quasi-Harmonic Multi-basin approach (QHMB).

In QHMB, the anharmonic nature of the underlying potential energy surface is captured by a divide-and-conquer approach, which aims at describing such anharmonicity via a series of non-overlapping harmonic wells whose configurational entropy is captured in the RRHO approximation. For this purpose, the number of conformers per compound and their equilibrium probability (p_i) were determined by clustering a converged MD simulation via the average linkage algorithm,³³⁶ as implemented in AmberTools18. Each cluster center was then used as a reference to extract molecular snapshots within the basin and evaluate the per-basin entropy by standard QHA. Finally, by introducing quasi-harmonic entropies and equilibrium probabilities into Eq.10 a multi-basin estimate of the entropy of the landscape was obtained. Note that the use of QHA for the per-basin entropy not only accounts for part of the anharmonicity but opens to the numerical evaluation of absolute molecular entropies in solution.

5.2.4: MM/GBSA binding free energy calculations

Protein structures were prepared using the Schrodinger suite⁹¹ adding missing atoms and assigning standard protonation states for all titratable residues at pH 7.0. For the ligand, initial coordinates were extracted from the PDB of the complex and converted into an sdf file. Atomic partial charges were assigned using the RESP methodology¹⁷¹ upon geometry optimization at the HF/6-31G* level of theory. The tleap utility was used to build the protein-ligand complexes and FF14SB⁵⁰ and GAFF2 force fields to assign the protein and ligand atom types, respectively. The system was then solvated in a octahedral TIP3P¹⁷³ water box with a 14 Å water layer around the protein. Counter ions were added to ensure charge neutrality. Simulations of the protein-ligand complexes were carried out as described above with production runs of 100ns. By saving molecular snapshots every 2 ps, a total of 50000 molecular configurations were collected for further analysis. Production runs of the free ligand in solution were 300ns long, saving snapshots every 2 ps for a total of 150000 molecular configurations.

Protein-ligand binding free energies in solution were evaluated using the MM/GBSA methodology as implemented by the MMPBSA.py script in AmberTools18¹¹⁸. The polar contribution to the solvation free energy was computed using the OBC gb model²⁸¹ with mbondi2 radii, a protein internal dielectric constant of 4, and an external dielectric constant of 80. The non-polar contribution was evaluated from the Solvent Accessible Surface (SAS) computed using the LCPO algorithm.^{285,337} A frame-skip of 5 was employed to reduce the number of snapshots to be processed to 10000. For the binding affinity

calculations, the standard single-trajectory setup was used; i.e. the trajectories of the receptor and the ligand in solution were extracted from the MD simulation of the complex.

5.2.5: Clustering algorithms explored

5.2.5.1: Hierarchical clustering

Hierarchical clustering algorithms focus on finding relationships within data by partitioning hierarchically the data points in a tree-like representation. These can be either bottom-up (agglomerative) algorithms, where all clusters are initiated with only one data point and then the closest points are iteratively merged together until all points are grouped into a single cluster, or top-down (divisive) algorithms, where the algorithm starts from one single cluster and iteratively separates the most dissimilar points from each other. Within any of the schemes, there must be a threshold measuring cluster similarity, like the root mean squared distance (RMSD), to allow cluster identification. The difference between different flavors of hierarchical algorithms is then just the way in which this criterion is used.³³⁶

In single-linkage, the distance between clusters is that given by the smallest intercluster distance between a point in cluster 1 and a point in cluster 2.³³⁸ Based on this distance matrix, the two closest clusters are merged together. In the complete-linkage method, the distance matrix is built taking into account the maximal intercluster distance between clusters and then taking the pair which exhibits the minimal distance in the matrix.³³⁹ Finally, in average-linkage the cluster-to-cluster distance matrix is built by taking the average of all pairwise intercluster distances between all pairs of clusters. The clusters with the smallest average distance are then merged together.³³⁶ In any of the cases, this procedure is done iteratively until either no new clusters are found with distances below the threshold or until the maximal number of iterations is achieved.

5.2.5.2: Density-based Spatial Clustering of Applications with Noise

Another approach for clustering molecular snapshots relies on the density (e.g amount of points) around a given point in space, such as the Density-based Spatial Clustering of Applications with Noise (DBSCAN) method. In DBSCAN, two parameters should be set: the number of points required to define a region as dense (*minPoints*) and a measure of point closeness based on a distance criterion (*eps*). For the distance criterion, typically the Euclidean distance is used and if the distance between points is smaller or equal to the *eps* threshold, they are considered as part of a cluster. It is important to remember

that the successful application of DBSCAN is dependent on setting appropriate values for the *eps* and the *minPoints* parameters.³⁴⁰ In particular, should the *eps* be too small, no clustering would happen because the distances between points would be too large and many points discarded as noise. On the other hand, if *eps* is too large, then all points would be clustered together. This parameter must be calibrated, usually with a k-distance graph (*eps* vs number of clusters). As for the *minPoints*, it is important to use larger numbers as a small *minPoints* value would produce many small, isolated clusters. In DBSCAN, the points are mapped into a hypersphere and *eps* becomes the radius of the hypersphere. It then follows that a core point is a point which has at least *minPoints* within the hypersphere created with itself at the center and a given *eps* radius. A border point is one which has less than *minPoints* around itself but is close to a core points. A noise point is a point with no core points close-by and less than *minPoints* around itself.^{336,340}

The algorithm progresses by picking a random point in space and evaluating the number of *minPoints* close-by within a given *eps* radius, clustering all these points together around a core. Then, the calculation is iteratively performed, expanding the cluster or initiating new clusters altogether until all points are visited.^{336,340} In this work, we set *minPoints* to 5 and varied the *eps* threshold as described previously.

5.3: Results and Discussion

5.3.1: Absolute entropy calculations in vacuum for cyclohexanone

To provide a proof-of-concept calculation before pursuing a full benchmark for the accuracy of QHMB for small-molecules in gas-phase, cyclohexanone was selected. It was selected because this compound is relatively rigid save for a dihedral angle (C6-C5-C1-O1) ending in a carbonyl group. The movement of this dihedral angle yields two equally probable states (oxygen up or down), which is expected to cause QHA to fail in reproducing the absolute entropy of this compound in gas-phase (**Figure 5.1A**). However, should the multi-basin decomposition scheme be valid, it would be possible to reproduce the experimental gas-phase entropy of cyclohexanone (23.9 kcal/mol) following the decomposition of the MD trajectory into the two individual states. A MD trajectory of 300ns in gas-phase was post-processed by QHA and QHMB, and the results are shown in **Figure 5.1B**.

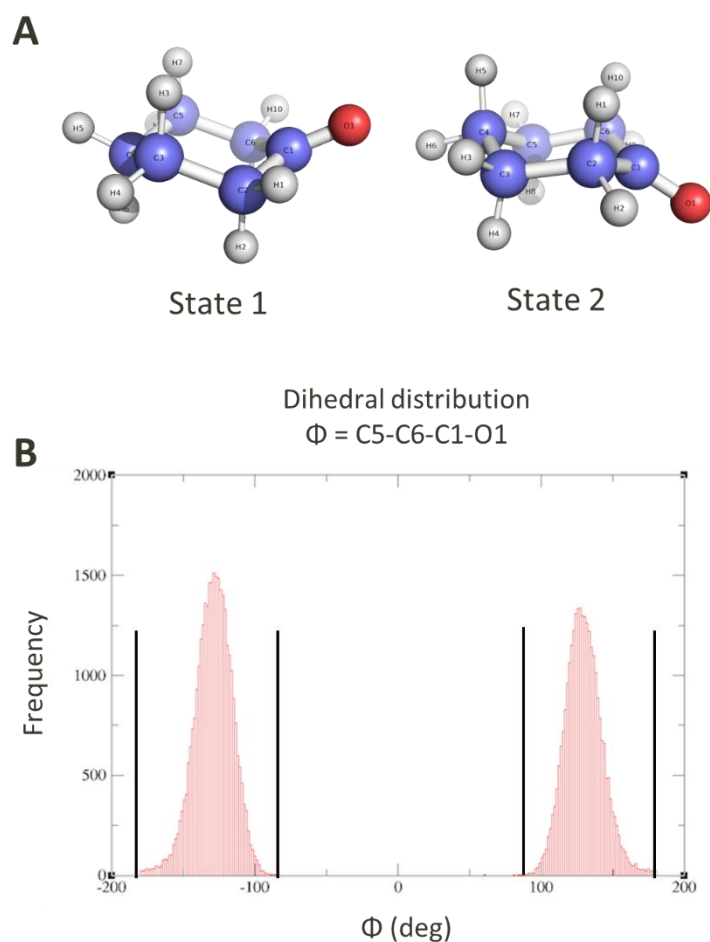


Figure 5.1 – Proof-of-concept approach for QHMB using cyclohexanone as a test case. **A)** The two states sampled by cyclohexanone due to the rotation of the C6-C5-C1-O1 dihedral angle. **B)** Distribution of the C6-C5-C1-O1 captured from the 300ns MD simulation of cyclohexanone in gas-phase.

As is shown in **Figure 5.1B**, the dihedral angle populates two different states and thus the approximation of the PES using a single, multidimensional, Gaussian distribution by QHA is likely to introduce an underestimation on the vibrational frequencies of the molecule which in turn cause the predicted absolute entropy of cyclohexanone to be overestimated. In QHMB the PES is decomposed into two different states which are analyzed separately. Thus, we expect the clustering step to produce two pure clusters of molecular configurations which would allow us to compute the absolute entropy of cyclohexanone based on the “mixture of conformers” theory. We find that QHA indeed fails to reproduce the experimental entropy and instead overestimates it by ~ 2 kcal/mol ($TS = 26.0$ kcal/mol). Using QHMB we find that the multi-basin decomposition scheme corrects this overestimation error and manages to reproduce the experimental entropy 0.2 kcal/mol of error (23.7 kcal/mol). Thus, it appears that indeed a possible approach to correct the large overestimation error which is commonly found in QHA calculations is to

decompose the PES into individual microstates within which the RRHO approximation is valid.

5.3.2: Calculation of absolute entropies in vacuum

For all compounds in the NIST dataset, absolute entropies in the gas phase were evaluated using three NMA variants, two QHA variants, and QHMB; see *Methods*. Linear regressions for the NMA, QHA and QHMB results are shown in **Figure 5.2**. The data show that standard harmonic analyses provide results that are strongly correlated with the experiments, i.e. R^2 for NMA and QHA are 0.96 and 0.86, respectively. However, the slope for NMA is less than one, whereas that for QHA is more than double; see **Table 5.1**. These data indicate that neglecting the contribution arising from the existence of multiple basins, as done by NMA, produces a systematic underestimation of the entropy. However, approximating the complex potential energy surface by a single, multi-dimensional harmonic well, as done by QHA, results into a much larger overestimation and yields at best an upper bound to the absolute entropy. Both conclusions are consistent with previous reports in the literature.^{263,225}

Analysis of the deviation from experiments provides further insights. The results in **Table 5.1** show that the RMSE for the full data set is 1.1 kcal/mol for NMA, whereas that for QHA is 8.1 kcal/mol. Although this suggests that NMA provides, on average, accurate entropy predictions, this is not always the case. In fact, by looking at the subset containing the most flexible molecules (i.e. compounds with > 3 rotatable bonds), the RMSE increases from 1.1 to 1.7 kcal/mol. Together with the large deviation of the QHA predictions, these data consistently point to the fact that standard approaches based on the RRHO approximation do not provide accurate entropy estimates even for small molecules in the gas phase. In sharp contrast, the results of QHMB are in quantitative agreement with the experiments and show a MUE of 0.28 kcal/mol and a slope of 1.02 for this dataset; see **Table 5.1**.

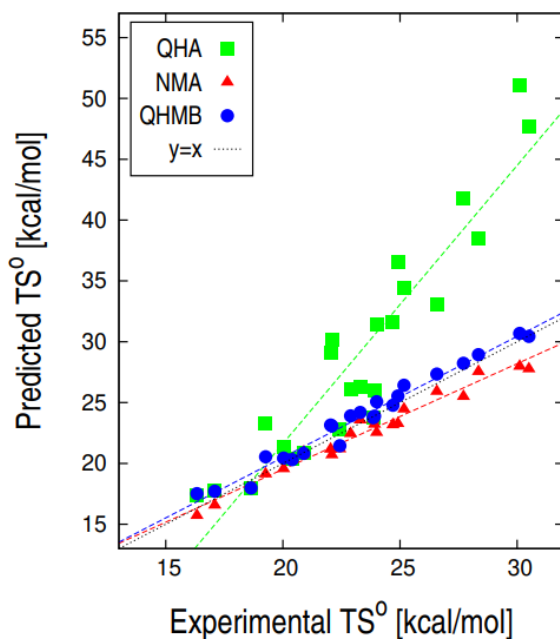


Figure 5.2 - Correlation of computed versus experimental absolute standard entropies for 23 small molecules in the gas phase. Experimental entropies at 298K were taken from NIST. The data show the performances of the newly introduced multi-basin approach (QHMB) relative to popular single-well approaches based on the RRHO approximation, i.e. NMA (red) and QHA (green).

Table 5.1 - Accuracy of various RRHO methods for the calculation of absolute molecular entropies.

Experiment	RMSE ^[*]	RMSE ^[+]	R ²	Slope	MUE ^a
NMA ^b	1.12	1.67	0.96	0.86	0.93
NMA-multi ^b	1.17	1.75	0.96	0.86	0.97
NMA-clust ^b	1.13	1.64	0.96	0.87	0.96
QHA ^b	8.09	13.36	0.86	2.24	5.87
QHA-clust ^b	8.42	13.95	0.86	2.30	6.07
QHMB ^c	0.36	0.24	0.99	1.02	0.28

All RMSE values are given in kcal/mol. [*] – RMSE for the full benchmark set. [+] – RMSE for the subset of ligands with 3 or more non-redundant torsions. [a] – Mean Unsigned Error (MUE) between predicted and experimental absolute entropies. [b] – Results obtained by enforcing the symmetry number to one. [c] – Results using appropriate symmetry numbers.

We conclude that the large overestimation by standard QHA can be corrected to a remarkable accuracy by an efficient multi-basin decomposition of a converged MD simulation. Perhaps surprisingly, these results highlight that the strong correlation observed with QHA ($R^2 = 0.86$) does not correspond to accurate absolute entropy predictions and cannot be used standalone for validation purposes. Indeed, both the MUE (5.9 kcal/mol) and RMSE (8.1 kcal/mol) for QHA indicate that these predictions are completely off.

5.3.3: An automatic procedure for QHMB absolute entropy calculations

A QHMB calculation encompasses three steps: i. the determination of the stable conformers along with their equilibrium probability from a converged MD; ii. the calculation of the per-basin entropy per conformer by QHA (**Equation 5.4**); and iii. the calculation of the entropy of the landscape (**Equation 5.9**). Although a manual implementation of QHMB based on the analysis of dihedral distributions and visual refinement may be accurate enough, this procedure is impractical, if the aim is to apply it to hundreds or thousands of compounds. For this purpose, an automatic QHMB procedure was developed. Following Suarez *et al.*,^{225,308} the implementation aims at: i. identify stable conformers by RMSD clustering of an extended MD trajectory; ii. extract a series of sub-trajectories corresponding to each of them; and iii. analyze those sub-trajectories by QHA automatically. For this purpose, several hierarchical algorithms including the Average-Linkage³³⁶, the Single-Linkage³³⁸ and the Complete-Linkage,³³⁹ and the density-based DBSCAN³⁴⁰ were considered; note that all clustering methods are part of the Amber18 software suite¹¹⁸. In addition, since a proper decomposition of the configurational space into distinct potential energy wells is critical for a proper evaluation of the absolute entropy, the QHMB analysis was repeated by varying the RMSD cutoff for clustering from 2.0Å to 0.1Å in decrements of 0.1Å. To validate the procedure, the dataset from NIST for which experimental entropies are available (see above) was used as benchmark. The results are shown in **Figure 5.3**. At large cutoffs, i.e. RMSD $\sim 2\text{Å}$, all methodologies are equivalent and yield entropy results with a systematic error as large as that of standard QHA. The lower the cutoff, the smaller the error of the entropy calculation. Interestingly, by decreasing the RMSD cutoff below 0.5Å both the Average linkage (green) and the complete-linkage (black) algorithms improve the QHMB predictions quite significantly and plateau to an RMSE below 1 kcal/mol. On the other hand, neither the Single linkage nor the DBSCAN algorithms achieved satisfactory accuracies in the investigated range of cutoffs. Based on these results, we conclude that the Average linkage or the complete-linkage algorithms are essentially equivalent and produce accurate and stable QHMB results at reasonably low clustering cutoffs. Since

clustering based on Average linkage produced the lowest RMSE at cutoffs below 0.5 Å, this protocol was selected for all subsequent studies. We note in passing that at too large cutoffs (≥ 2 Å) all clustering algorithms fail because they mix conformations belonging to different basins and QHMB reduces to standard QHA.

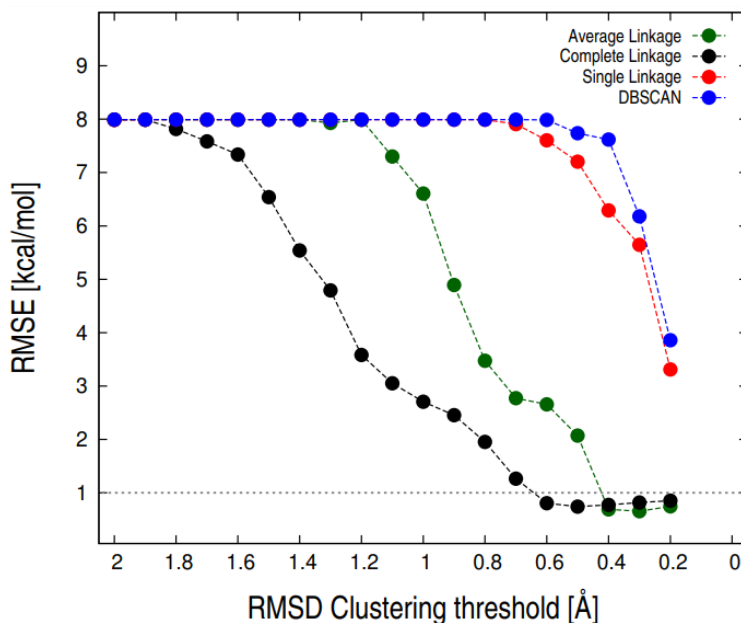


Figure 5.3 – Accuracy of QHMB as a function of the clustering algorithm and the RMSD cutoff. The accuracy is evaluated by the root-mean-squared error (RMSE) from experimental entropies in the gas phase for the NIST dataset; see Main Text. The dashed line at 1 kcal/mol illustrates the boundary of chemical accuracy.

5.3.4: Calculation of absolute entropies in solution

The remarkable accuracy of QHMB along with its straightforward application to explicit-solvent MD trajectories opens to the evaluation of absolute molecular entropies in solution. This latter is particularly interesting because absolute entropies in solution remain experimentally inaccessible. Here, we use QHMB to evaluate standard molecular entropies in solution for 23 compounds from the NIST dataset (**Annex S5.1**) and compare them with results obtained in the gas phase. The results in **Figure 5.4** (and **Annex S5.3**) show a systematic difference between gas-phase and solution of about 2 kcal/mol; i.e. the standard entropies in vacuum are systematically higher than those in solution. This large difference is primarily due to the translational entropy contribution and a different definition of the standard state; the standard volume per mole amounts to 24.78L in the gas-phase and 1L in solution, which corresponds roughly to a 1.9 kcal/mol translational entropy change at 298.15 K. Accounting for the change in the standard state definition, the difference in intramolecular entropy between the gas phase and the solution is much smaller and in most cases within 0.1 kcal/mol; see **Figure 5.4**. The largest changes are

found for butane, butanoic acid, di-n-propylether and pentane, whose conformational entropy appears to be slightly more favorable in water.

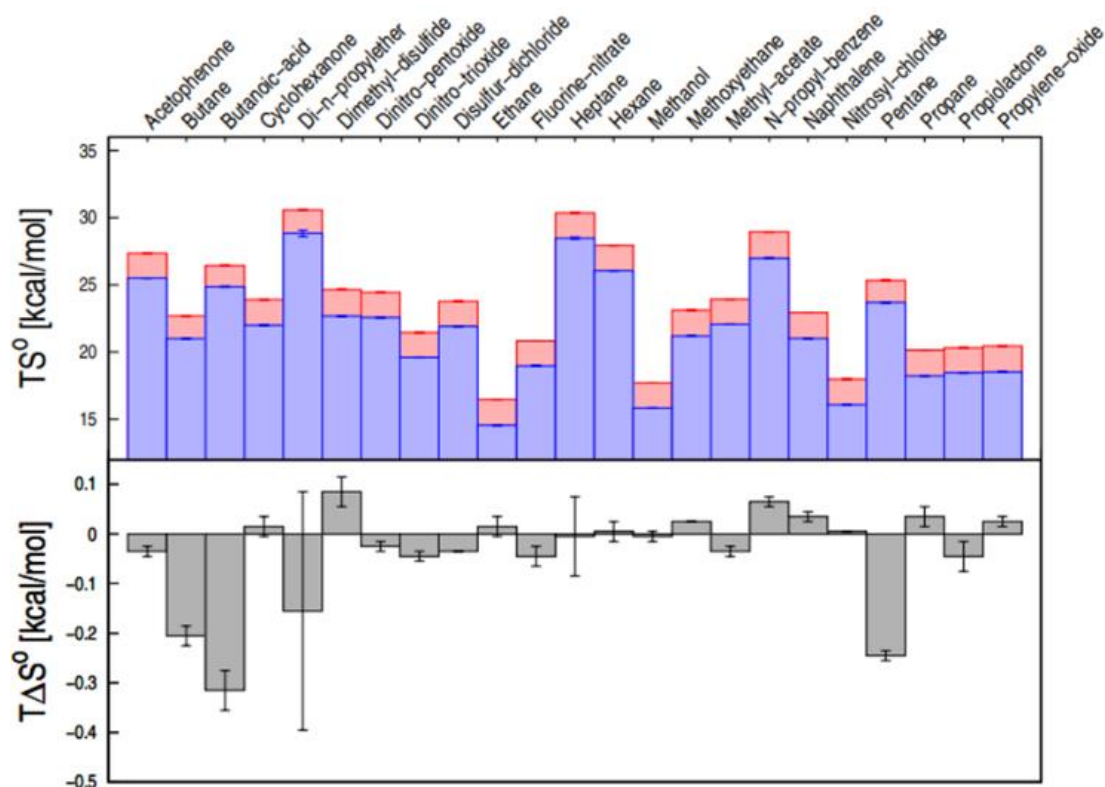


Figure 5.4 – Absolute configurational entropies in solution. On the top, absolute QHMB entropies for 23 small molecules from the NIST dataset in aqueous solution (blue) are shown and compared with the corresponding entropies in the gas phase (red). In solution, a standard concentration of 1M was used. On the bottom, the difference in configurational entropy in the gas phase versus solution is shown upon correction for the standard state definition. Error bars were estimated from standard error of the mean (S.E.M) by block analysis.

Based on these results, we conclude that for small and apolar molecules the water environment does not affect the accessible configurational space dramatically, so that gas-phase entropies provide reasonable estimates for the solution environment. Nonetheless, for larger and more polar solutes, water is expected to produce stronger effects e.g. by stabilizing configurations inaccessible in gas-phase, which could be quantified by QHMB.

5.3.5: Application to binding free energy calculations

MM/PBSA and MM/GBSA are very popular approaches for the calculation of relative ligand binding affinities.^{89,110,115,201,212,213,341} However, it has been noticed that their performance are highly system dependent.^{211,214} In addition, many researchers prefer not to include entropy contributions in their MM-PB(GB)SA calculations due to the additional

computational cost and previous observations that an explicit inclusion of entropy may worsen the correlation of the predicted binding affinities with experiments.^{201,211,214,225} Others, to overcome the limitations of entropy calculations in binding reactions have proposed alternative strategies that do not rely on the harmonic approximation like the interaction entropy method.⁴⁹ Motivated by the accuracy of the QHMB results in the gas phase (see above), we selected a dataset of 21 protein-ligand complexes from the Greenidge dataset²¹⁴ and used QHMB to quantify the entropy loss on binding. Specifically, QHMB was used to evaluate the configurational entropy of the ligand in its bound and unbound states from independent MD simulations, so as to estimate the entropy of binding from the difference between the two. This term was then introduced as an entropy correction to standard MM/GBSA calculations. Binding free energy results from MM/GBSA with and without the QHMB entropy correction are shown in **Figure 5.5**. For comparison, the ligand entropy loss was also accessed by standard QHA²⁹⁶ and the performances of the two protocols compared; see **Table 5.2**. The data show that standard MM/GBSA calculations as implemented in AmberTools18¹¹⁸ do a reasonable job with this dataset yielding a $R^2 = 0.67$. Application of a ligand entropy correction based on QHA introduces a larger error and the correlation decreases by 17% ($R^2 = 0.5$), consistent with previous reports.^{211,254} In sharp contrast, the introduction of the ligand entropy correction by QHMB increases the correlation by 11% and yield a final $R^2 = 0.78$; see **Table 5.2**. These results lead to the following observations. First, the QHMB entropy correction introduces a penalty in the calculated ΔG° , which accounts for the restriction of the configurational volume accessible to the ligand in its bound state. Second, the size of the correction is strongly ligand-dependent and introduces a larger penalty for big and flexible ligands; i.e. the QHMB correction is > 9 kcal/mol for four ligands in the dataset, while being 6 kcal/mol on average (**Annex S5.4**). In addition, the correction relies on configurational sampling by MD, which allows for direct probing of the change in the configurational volume of the ligand independently of its intrinsic flexibility. Taken together, these results suggest that the introduction of a ligand-dependent entropy correction based on QHMB increases the accuracy of relative binding-affinity calculations.

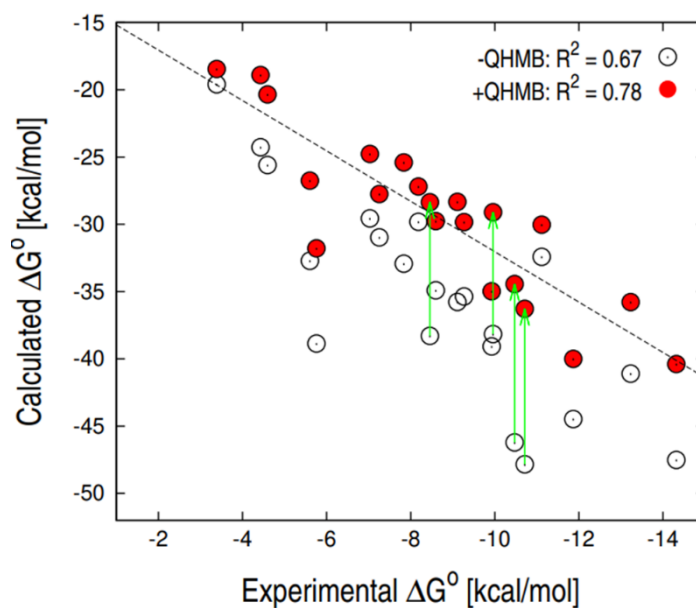


Figure 5.5 – Correlation of experimental versus predicted binding affinities by MMGBSA with (red) and without (empty points) entropy correction by QHMB. The data show that the introduction of the entropy correction increases the correlation with experiments by 11%. For some large and flexible ligands, the correction is as large as 10 kcal/mol (green arrows).

Table 5.2 – Inclusion of the ligand entropy loss in MMGBSA calculations.

Experiment	R	R ²	slope	$\rho^{[4]}$
MM- GBSA ^[1]	0.82	0.67	2.15	0.79
MM-GBSA +QHA ^[2]	0.71	0.50	1.33	0.66
MM-GBSA +QHMB ^[3]	0.88	0.78	1.87	0.88

[1] Single-trajectory MMGBSA results. [2] MMGBSA results with entropy correction by QHA. [3] MMGBSA results with entropy correction by QHMB. [4]-Spearman's correlation coefficient. All simulations were performed at 298.15 K and 1M.

5.4: Conclusions

By using a basin decomposition scheme via RMSD clustering of converged MD trajectories, we have demonstrated how to improve the accuracy of absolute molecular entropy calculations relative to standard approaches based on the RRHO approximation; i.e. normal-mode and quasi-harmonic analyses. The proposed numerical strategy, here referred to as quasi-harmonic multi-basin, includes two stages. The first one involves the calculation of a probability-weighted RRHO entropy per basin. The second one evaluates the entropy of the landscape via the Gibbs formula. In the current formulation, the per-basin entropy is accessed by QHA of molecular snapshots sampled by MD, which not only accounts for the anharmonicity of the basin but also opens to the evaluation of molecular entropies in solution. By using QHMB in test calculations, we were able to reproduce experimental gas-phase entropies for 23 small molecules from NIST with an RMSE below chemical accuracy (i.e. 0.36 kcal/mol) and without too much computation. In addition, QHMB was used to estimate the entropy loss upon ligand binding for a set of 21 protein-ligand complexes from the Greenidge dataset, which was shown to improve the accuracy of standard MM/GBSA calculations. Since this entropy correction penalizes more strongly large and flexible ligands, its use is likely to alleviate the typical bias towards larger and larger compounds, thus reducing the false-positive rate in simplified binding affinity calculations. In addition, the availability of an automatic procedure to compute the QHMB entropy correction opens to applications in virtual screening campaigns.

When compared to existing absolute entropy approaches that go beyond the RRHO approximation, such as the work of DeTar³¹², Guthrie³¹¹ and Suarez *et al*,³⁰⁸ the QHMB method has also some advantages. The method of DeTar³¹² evaluates the per-basin contribution by NMA using vibrational frequencies from *ab initio* calculations. When applied to a series of 18 alkanes, the difference between predicted and experimental $T\Delta S$ was around 0.1 kcal/mol. While accurate, these calculations are computationally very intensive, they require an a priori knowledge of all relevant conformational states, and they are limited to the gas phase. In the work of Guthrie³¹¹, a simplified mixture of conformers approach was developed that relies on vibrational frequencies from DFT and assumes equal probabilities for all conformers. When applied to an extended dataset of 128 organic compounds, the reported MUE on predicted absolute entropies in the gas phase was 0.26 kcal/mol. Although more efficient, these calculations are still computationally demanding, they rely on an approximated formula for the entropy of the landscape, and require manual intervention to determine the number of conformers and the correct symmetry number. In the work of Suárez *et al*³⁰⁸, the per-basin contribution was evaluated from the RRHO entropy of a number of molecular snapshots sampled by MD, while the entropy of the landscape was obtained via the mutual information

expansion²⁹³ upon discretization of the torsional degrees of freedom. Using this strategy, calculated absolute entropies on test systems including hydrocarbons in vacuum showed a MUE of 0.31 kcal/mol. In the gas phase, the RRHO entropies are computed by NMA using DFT, which is computationally intensive and solvent effects on the entropy of the PFG hexapeptide in solution could be captured only indirectly, e.g. using an implicit solvent model. Last, convergence of the conformational entropy via MIE introduces an additional cost, which requires long simulation trajectories if high-order corrections are needed. The QHMB methodology suffers from none of the above limitations. In fact, energy minimization is not required, conformational sampling is provided by Molecular Dynamics, the vibrational frequencies per conformer are accessed from the analysis of the atomic fluctuations, which effectively accounts for solvent effects, and all calculations are carried out by molecular mechanics. Most importantly, the MUE of the QHMB entropies in the gas phase was 0.28 kcal/mol, which outperforms the calculations by Suarez³⁰⁸, and is comparable to the DFT calculations by Guthrie³¹¹ and the *ab initio* calculations from DeTar³¹². Thus, QHMB is able to provide absolute molecular entropies accurately and efficiently both in vacuum and in solution.

However, QHMB has its own limitations. Currently, we have tested the protocol only with molecules with up to 8 rotatable bonds, for which conformational sampling converges in < 300ns of MD. For more flexible molecules with hindered conformational transitions, this simulation time might be too short.³⁰⁸ Thus, the strongest limitation of QHMB is the need for sufficient sampling, which becomes increasingly costly as molecular size and flexibility increase. This is even more so for the evaluation of a QHMB entropy correction for MM/PBSA, where sampling of the configurational space in the bound state may include slow rotameric transitions of the amino-acid side chains in the protein binding site. The combination of QHMB with enhanced sampling techniques such as REMD²²⁸ is proposed as a possible strategy to alleviate this problem.

The original combination of basin decomposition via clustering of Molecular Dynamics trajectories with quasi-harmonic analyses, here termed QHMB, opens to accurate absolute entropies calculations of small molecules in solution with possible implications on binding free energy calculations. The availability of an automatic procedure to compute QHMB entropies makes it a new available tool in the field of drug discovery.

6. Identification of hits on the quest for allosteric modulators of the myosin molecular motor

6.1: Introduction

Many cellular functions are dependent on the polymerization of actin and its interaction with myosin molecules.³⁴² Myosins are a family of molecular motors with the ability to hydrolyze ATP, harnessing the energy arising from its hydrolysis to perform mechanical work.³⁴³ As an example, if the actin filament is a road then myosin would be the car and ATP the fuel. There are many classes in the myosin superfamily, of which myosins V³⁴⁴ and II³⁴⁵ are examples of processive and non-processive myosins respectively. Processive myosins are those able to transport cargo within the cell and non-processive myosins are those involved in muscle contraction.^{344,345} In a coarse-grained view, myosin motors are composed of light and heavy chains. Within the heavy chains, three domains are found: the head (or motor) domain, containing the ATP and actin binding sites, the neck region, and the tail domain, the latter determining the functional properties of these molecular motors.³⁴³ A fundamental feature of the myosin motor domain is that there is a wide cleft delimited by the U50 and the L50 domains which harbors the actin binding site.³⁴⁶ Myosins have the ability to move along the actin filament, a motion which was studied extensively by means of *in vitro* motility assays.³⁴⁷ Interestingly, most processive myosins move towards the plus end of the filament, while myosin VI walks towards the minus end.³⁴⁸ Defective myosins are known to be implicated in the physiopathology of many conditions, including heart conditions³⁴⁹, chronic obstructive pulmonary disorder³⁵⁰ and cancer.³⁵¹

6.1.1: The myosin molecular motor cycle

The myosin motor cycle is composed by two main phases: the powerstroke, which is the force-generation step while actin-bound, and the recovery stroke, which happens after force generation and culminates with a myosin motor ready to bind actin.^{350,352} The recovery stroke encompasses several conformational transitions which re-prime the lever arm in an actin-unbound state.^{352,353} At the start of the recovery stroke, ATP binds to the actin-bound myosin motor, found in the rigor state, and promotes actin unbinding.³⁵³ Then, ATP hydrolysis fuels the resetting of the position of the lever arm.^{352,353} At the end of the recovery stroke, the myosin motor is bound to ADP and

inorganic phosphate (Pi), and the lever arm is in the correct position for the force generation phase. In this state, myosin binds to actin, initiating the powerstroke.³⁵³ Along this phase, which is composed of major conformational transitions, Pi, ADP and Mg²⁺ are released in discrete steps as the myosin motor produces mechanical work due to a swing of the lever arm.^{350,354} This swinging motion happens because conformational changes occurring the ATP and actin binding sites are then transmitted to the lever arm, leading to a 60° rotation of the converter subdomain.^{346,352} As myosins' affinity towards actin increases, the interactions between SMM2 and actin become stronger and mechanical tension is accumulated.³⁴⁶ The release of this tension occurs by the swinging of the lever arm, and is known as the powerstroke.³⁴⁶ At the end of the powerstroke, the myosin enters into the rigor state by releasing ADP, achieving maximum affinity towards actin.³⁴⁶ Upon ATP binding to SMM2, the myosin motor loses affinity towards actin and unbinds from it, restarting the cycle.^{349,352} A schematic representation of the myosin motor cycle is illustrated in **Figure 6.1**.³⁴⁹ Along the cycle, many unstable or transiently stable intermediate states may be populated as the myosin undergoes conformational transitions between stable states (rigor, post-stroke, pre-powerstroke, power-stroke and force holding state). Some of these intermediates have potential for pharmacological targeting *en route* to the development of new therapeutic approaches. Indeed, several studies have reported the modulation of myosin activity by small-molecule ligands.^{345,346,350,353,355,356}

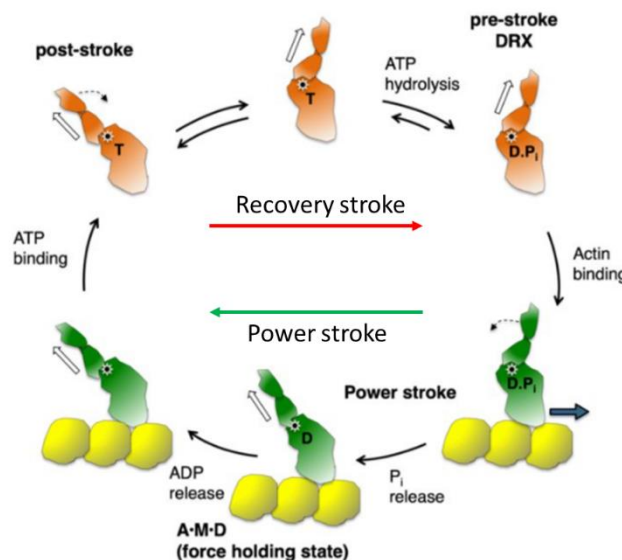


Figure 6.1 – Illustration of the myosin molecular motor cycle. In green are highlighted the steps carried out in the powerstroke phase and in red the steps of the recovery stroke. The post-stroke step is when ATP binds to the unbound myosin motor which populates rigor states. Adapted from “Three perspectives on the molecular basis of hypercontractility caused by hypertrophic cardiomyopathy mutations.” Spudich, 2019.³⁴⁵ Pflügers Archiv – European Journal of Physiology, 471, 701-717. Copyright © 2019, James Spudich.

6.1.2: Myosin as pharmacological target

Recently, myosins have drawn attention for their potential as targets with therapeutic applications such as cancer³⁵¹, heart failure³⁴⁹ or neurodegeneration.³⁴⁶ However, targeting the actin binding site using small-molecule inhibitors is not a reliable solution because the active site is highly conserved across different myosins, G-proteins or microtubule-based motor proteins.³⁴⁶ Since allosteric pockets are not as well conserved, they represent a better strategy to target myosin motor activity modulation.³⁴⁶ Studies into the existence of allosteric pockets in myosin molecular motors highlight 4 non-overlapping accessible sites for small drug-like ligands.³⁴⁶ One of the first well-characterized small molecule myosin modulators was blebbistatin.³⁴⁵ Blebbistatin is a small-molecule inhibitor of myosin II which binds an allosteric pocket close to the interface between the U50 and L50 domains.^{345,346} It traps the motor cycle in a pre-powerstroke state with low actin affinity, where ADP and P_i are bound, by preventing P_i release.^{345,346} While potent, with an IC₅₀ of 0.5–2 μM for skeletal and nonmuscle myosin-2 respectively³⁴⁶, blebbistatin is photolabile and becomes cytotoxic when irradiated with blue light.^{345,355} Another example of an important small molecule modulator of myosin is mavacamten, a cardiac myosin small-molecule inhibitor.³⁵³ This compound binds to an allosteric pocket within cardiac myosin and inhibits the force production phase with an IC₅₀ of 0.71 μM in human tissue.³⁵³ It provides an exciting avenue towards improving the contractile properties of a heart which has developed hypertrophic cardiomyopathy (HCM), as shown in rat model systems.³⁵³ More recently, another allosteric cardiac myosin inhibitor was announced by Cytokinetics, CK-274.³⁵⁶ A final example are pseudilins, in particular pentachloropseudilin (PCIP), a low micromolar myosin V inhibitor binding at the edge of the 50 kDa cleft on the motor domain.³⁵⁷ Thus, it emerges that targeting allosteric pockets which open and close during the myosin motor cycle is a reasonable approach *en route* to the discovery and development of innovative myosin modulators.³⁴⁶ However, these proteins remain difficult targets for drug design, as some of the allosteric pockets found in myosin motor domains, like the one of PCIP, are only transiently open in low-populated intermediate states. Furthermore, some of these states can only be crystallized when the inhibitor is bound because the conformational equilibria is shifted upon ligand binding, stabilizing the intermediate state.³⁵⁰ Thus, innovative and efficient approaches for allosteric drug discovery and design are required to study and propose new allosteric modulators of myosin function.

6.1.3: Smooth muscle myosin II – Pharmacological relevance

Smooth muscle cells are a fundamental part of hollow organs such as the bladder, the gastrointestinal tract, the uterus, the airways or vasculature.³⁵⁸ To induce SM relaxation there are two main paths: inhibiting the contractile mechanism or removing the contractile stimulus.^{358,359} Smooth muscle contractility is fundamental in pathologies such as asthma³⁶⁰, prostatic hyperplasia³⁶¹ and chronic obstructive pulmonary disease.³⁵⁰ While there exist smooth muscle airway relaxants, notably β -adrenergic agonists and muscarinic antagonists, these compounds inhibit the activity of SMM2 in a non-specific manner.³⁵⁰ It would be desirable to have a SMM2-specific inhibitor, allowing the modulation of the smooth muscle contractility of these tissues for a fast and efficient relaxation of the contracted muscle.^{350,362} Recently, the discovery of a potent and highly selective inhibitor (CK-571) of SMM2 has been reported by the group of Dr. Anne Houdusse in partnership with Cytokinetics.³⁵⁰ The inhibition of SMM2 by CK-571 was found to be elicited by binding to an allosteric pocket whose location could not have been guessed from X-ray crystallographic structures solved in the absence of the bound inhibitor. The pocket targeted by CK-571 is an allosteric pocket which opens in a short-lived intermediate state during the recovery stroke.³⁵⁰ Inhibition by CK571 results from the stabilization of an intermediate myosin conformation preceding the pre-powerstroke state (PPS), thus blocking the motor in a state of low actin affinity and preventing the cycle from continuing by hampering the re-binding of SMM2 to actin. High-resolution structural information of the co-crystal is therefore fundamental for the discovery of new potent and selective modulators of allosteric proteins like myosin. They provide a starting point for allosteric drug design which is amenable to be explored by means of computational structure-based drug discovery approaches.

6.1.4: Project goal

The aim of this project was to undertake a VS campaign and identify innovative SMM2 hit compounds from virtual chemical libraries. We first studied the MM/GBSA^{196,201} methodology and devised a calculation setup for the VS campaign based on results from a previous *in-house* campaign on the SMM2/CK-571 pocket. It was understood from that screening that parameters like the solute internal dielectric constant and the GB model required optimization. These were optimized through benchmarking MM/GBSA calculations arising from implicit solvent MD data to reference data from the previous screening. It was also realized that entropy contributions had been neglected in that campaign. To address this issue, a term accounting for the ligand configurational entropy loss upon binding²¹⁸ was added using QHMB. The VS campaign was carried out targeting

the SMM2/CK-571 X-ray crystal structure using a docking and free energy rescoring approach, implementing the optimized rescoring setup, through our *in-house* developed software *ChemFlow*. The selected library was the Chimiotèque National du CNRS (CN).³⁶³ The 3D structure of the protein-ligand complexes were obtained following docking of ligands to the CK-571 binding pocket, selecting the best scored binding mode per ligand. Two free energy rescoring steps were carried out by MM/GBSA calculations, the first one in implicit solvent for a fast evaluation of the binding affinity, aiming at compound prioritization. Prioritized compounds were then simulated in explicit solvent and results from MM/GBSA calculations arising from the explicit solvent MD simulations were coupled to the QHMB penalty to produce the final compound ranking. The subset of best ranked compounds was selected from the prioritized set and these ligands were experimentally tested for their ability to decrease SMM2 ATPase activity.

6.2: Methodology: Virtual Screening protocol

6.2.1: Protein preparation

The crystallographic structure of SMM2 was retrieved from the Protein Data Bank (PDB), deposited under the PDB-id **5M05**. It represents an intermediate state within the recovery stroke phase of the myosin motor cycle which is stabilized by CK-571 binding (**Figure 6.2**).³⁵⁰ Within this structure, the inhibitor CK-571, ADP and a magnesium ion (Mg^{2+}) are found bound to SMM2. A similar structure of SMM2, though bound with ADP, Mg^{2+} and beryllium trifluoride (BeF_3^-), was also crystallized (PDB **5T45**) previously. The ADP and the Mg^{2+} are found in the ATP binding site of the motor domain and CK-571 is found in the allosteric binding pocket to be targeted. Crystallographic water molecules were removed from the structure prior to protein preparation. Then, missing loop portions were constructed using MODELLER.¹³⁴ MolProbity was used to add missing atoms, including hydrogens, and to optimize the geometry of amino acid side chains by comparison to an internal rotamer library whenever necessary.¹³⁰ The protonation state of the titratable residues was computed using SPORES³⁶⁴ at pH 7.4. After protein preparation, a short energy minimization was carried out using a combination of steepest descent and conjugate gradient descent.

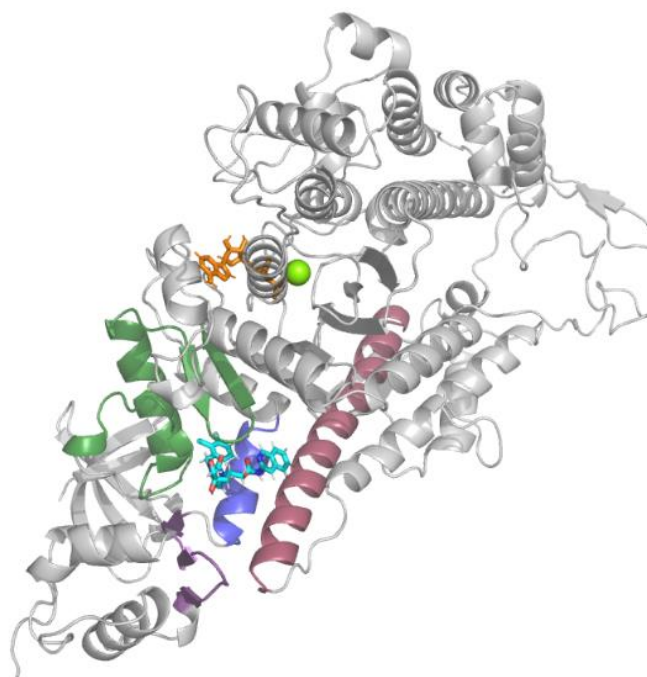


Figure 6.2 – Crystallographic structure of the SMM2 motor domain. The structure ADP (orange), CK571 (cyan) and a magnesium ion (green) bound (PDB code **5M05**), displayed in cartoon. Highlighted are the portions of the SMM2 motor domain lining the binding site: SH1 helix (blue), relay helix (dark red), N-terminal domain (green) and part of the converter domain (light purple).

6.2.2: Ligand library preparation

To carry out the screening campaign, the CN virtual library³⁶³ was selected. This library contained at the time approximately 60 thousand compounds, which are accessible through a network of collaborations within the French scientific community at a reduced cost. Furthermore, although it is a medium-sized library, it contains many different chemical scaffolds and thus a large degree of chemical diversity. The CN library was initially prepared using PrepFlow.¹⁶⁸ A schematic representation of the library screening workflow is shown in **Figure 6.3**. Ligands were extracted in SMILES format³⁶⁵ and converted to 2D SDF format. Chemical entities containing unknown or uncommon atoms, too many double bonds and too large cyclic structures were filtered out.¹⁶⁸ Then, a filtering step was applied to remove salts, solvent molecules and other chemical structures which do not correspond to ligands using the *standardizer* tool from ChemAxon.³⁶³ The remaining ligands were converted into 3D format using the ChemAxon *molconvert* tool²⁵ and a subsequent tautomer enumeration was carried out at pH 7.4, keeping only the most probable tautomer, using the *cxcalc* tool.²⁵ Conformer and stereoisomer enumeration was then performed, keeping the all forms of each ligand with a probability above 10%. The ligand library was then loaded into DataWarrior¹ for additional filtering steps which consisted of evaluating different ligand physicochemical

properties. In ligands were filtered according to the well-known Lipinski Rule-of-Five³⁹ and all compounds which broke more than one of the rules were excluded (See **Table 1** in **Chapter 1**).

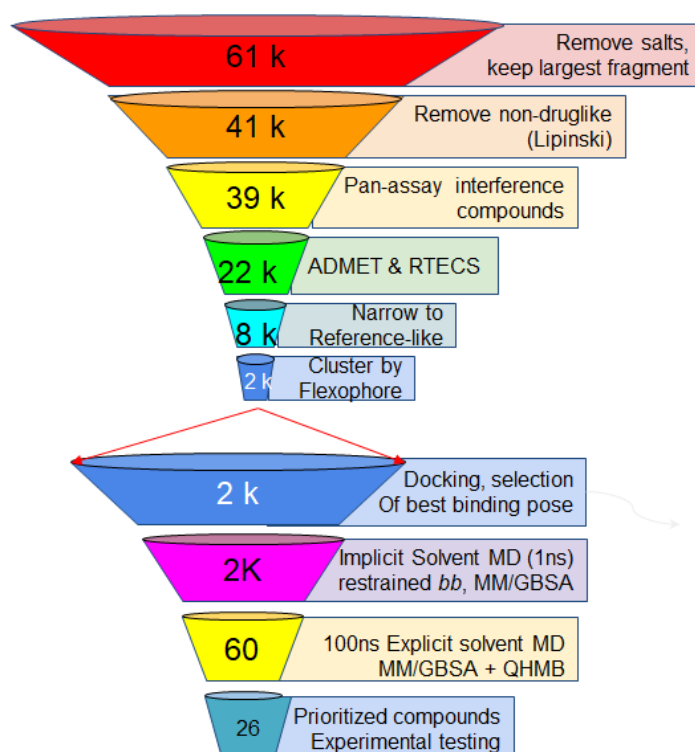


Figure 6.3 - Schematic representation of the VS campaign workflow carried out on CN towards the discovery of new SMM2 modulators.

After the library preparation and R-o-5 filtering, 39 thousand compounds were retained. These were then filtered out according to the presence of specific PAINS substructures using a KNIME³⁶⁶ node.¹⁵⁶ The PAINS-filtered dataset contained 22 thousand compounds which were further pruned based on physicochemical properties such as the Polar Surface Area (PSA), the computed solubility (clogS), the number of rotatable bonds and the presence of some toxic warhead groups in their molecular structures (RTECS) using DataWarrior.¹ The aim of this step was to focus the library on drug-like compounds with physicochemical features similar to CK-571 (see **Annex S6.3**) The size of the dataset decreased significantly at this stage, for a total of 8K compounds. Finally, a flexophore descriptor for each ligand was computed.³⁶⁷ The library was clustered using a 3D flexophore descriptor, further narrowing down the dataset to about 2300 chemically diverse structures where each structure represents a cluster of compounds, grouped based on 3D similarity.

6.2.3: Molecular docking

A molecular docking campaign was performed to produce 3D structural models of the protein-ligand complexes. The prepared ligand library (see **Chapter 6.2.2**) containing 2300 drug-like compounds was docked to the allosteric pocket of CK-571 in the SMM2 structure using the PLANTS software.¹⁷⁷ The ligands were considered as flexible and the protein was treated as rigid. The protein structure contained the ADP and Mg²⁺ co-factors in the ATP-binding site of the myosin motor domain. The search space for PLANTS docking was set as a sphere with a radius of 12Å, centered on the center of mass of CK-571 in the crystal structure. For the PLANTS docking protocol, the initial number of ants was set as 20, and the pheromone evaporation rate was set at 0.15 (see **Chapter 2**). All other parameters were set as default, producing a maximum of 10 binding poses per compound. The scoring function employed for the docking experiment was the ChemPLP scoring function described in **Chapter 2**. For each compound, the binding modes were ranked according to their score and the best-scored binding mode per compound was kept for further study.

6.2.4: Molecular Dynamics simulations

Following molecular docking, the 2300 protein-ligand complexes were scored using the MM/GBSA method.^{196,211,214} To carry out the free energy rescoring campaign, it was necessary to run molecular dynamics simulations. We initially performed short implicit solvent MD simulations and coupled them to MM/GBSA binding free energy calculations to perform a fast free-energy rescoring of the complexes obtained by molecular docking. A subset of complexes from the 2300 investigated by molecular docking was prioritized based on the ranking from the implicit solvent MM/GBSA calculations. The 60 top-ranked complexes were then simulated in explicit solvent for a longer period of time and an additional simulation of each of the ligands in solution was also carried out. The simulations of the bound ligand, extracted from the complex simulation, and the unbound ligand in solution were used to compute the ligand configurational entropy loss upon binding.²¹⁸ The simulation of the ligand unbound in explicit solvent was carried out because it is to estimate the ligand absolute entropy in solution.

6.2.4.1: Molecular Dynamics simulations: Implicit solvent

Implicit solvent MD simulations were carried using the FF14SB³⁶⁸ and the GAFF2 forcefields in the Amber18 simulation suite¹¹⁸, the implicit solvent Generalized Born model⁷³⁶⁹ and a solute internal dielectric constant of 4. The radii set used for these

simulations was the bondi radii set as suggested in the Amber18 package. For electrostatics, an infinite cutoff was employed. For the polar contribution to the solvation free energy, the pairwise summation of the effective Born radii calculations were truncated to atom pairs at a maximum distance of 15Å for calculation efficiency. Partial atomic charges were computed using the RESP method¹⁷¹, through Gaussian09¹⁷² at the HF/6-31G* level of theory. Parameters for ADP were extracted from Meagher *et al.*³⁷⁰, OPLS²⁰⁶, Veenstra³⁷¹, Weiner *et al.*³⁷² and Cornell *et al.*³⁷³ whereas parameter for Mg²⁺ were obtained from Allner *et al.*³⁷⁴ Before carrying out MD simulations, each system was energy minimized in implicit solvent for 1000 cycles with an energy convergence criterion of 1.0×10^{-4} kcal/mol-Å. Then, implicit solvent MD simulations were carried out for 1ns using the Langevin thermostat²⁴³, to set the temperature at 298.15K, with a collision frequency of 1 ps^{-1} and an integration timestep of 2fs. The SHAKE algorithm was used to constrain covalent bonds between heavy atoms and hydrogen atoms. Harmonic restraints with a force constant of 10 kcal/mol-Å^2 were enforced on the backbone atoms of the protein. The long-range electrostatic interactions were treated using the PME method. Molecular snapshots were collected every 2ps, for a total of 500 molecular snapshots.

6.2.4.2: Molecular Dynamics simulations: Explicit solvent

Starting from the structures produced by molecular docking experiments, the prioritized subset of protein-ligand complexes were embedded in octahedral TIP3P¹⁷³ water box extending 14Å from the edge of the solute molecule for a total of about 190K atoms. Sodium and chloride ions were added to the simulation box to ensure net charge neutrality and to set the salt concentration to 0.15M so that physiological conditions were reproduced Protein, ligand and co-factor parameters were obtained as described in **Section 6.2.4.1**. Before running the simulation the systems were subjected to three cycles of energy minimization, each comprising 1000 steps of steepest descent.¹³³ Then, the full system was energy minimized for 5000 steps using a combination of steepest descent and conjugate descent algorithms.¹³³ The biomolecular systems were then heated gently to 298.15K over 5ns in the NVT ensemble using Langevin Dynamics²⁴³ and an integration time step of 2 fs. The SHAKE algorithm²⁴⁰ was used to constrain all heavy atom-hydrogen covalent bonds and the PME scheme was used to treat long-range electrostatics.²³⁷ The system was then equilibrated in the NPT ensemble, using the Monte Carlo barostat²²³ as implemented in the Amber18 software suite for pressure control and the Langevin thermostat²⁴³ for temperature control. Equilibration and production runs were carried out at 1 atm pressure and 298.15K, the former running for 5ns and the latter for 100ns with an integration timestep of 2 fs and employing the pmemd.cuda code. Molecular

snapshots were collected every 2ps, in a total of 50000 snapshots. For the MM/GBSA calculations, a frameskip of 5 was used, processing 10000 equally spaced snapshots. Simulations of the unbound ligands in explicit solvent were carried out using the same setup as the protein-ligand complex simulations, running for 400ns and yielding a total of 200000 molecular snapshots.

6.2.5: Binding Free Energy calculations

To prioritize compounds for experimental testing, the trajectories arising from implicit and explicit solvent MD simulations were post-processed and used in end-point binding free energy calculations through the MM/GBSA method.^{196,201} The MM/GBSA calculations were carried out as implemented in the MMPBSA.py²⁶⁷ tool from AmberTools18¹¹⁸ using the 1-average approach.^{201,229} In this study, the potential energy of each species was evaluated using the Amber forcefield. The polar contribution is related to the interactions established by the particles due to their charges and was evaluated using the GB model of Simmerling *et al.*³⁶⁹ with bondi radii. The solute internal dielectric constant was set as 4 and the external dielectric constant was set at 80. The non-polar contribution, which should account for both the free energy cost of making a cavity in the solvent and the free energy gain in filling the cavity with electron density due to the dispersive interactions between solute and solvent, was evaluated from the Solvent Accessible Surface (SAS) computed using the LCPO algorithm.²⁸⁵ The entropic terms, which are usually accessed in the limit of the RRHO approximation^{115,225,229} by Normal Mode Analysis^{262,263} or Quasi-Harmonic Analysis^{261,296,317}, were initially neglected. However, the ligands' configurational entropy loss upon binding was included in the MM/GBSA calculations for the prioritized subset of complexes simulated in explicit solvent. This entropic term was computed by QHMB (see **Chapter 5**).²¹⁸

6.2.6: Experimental ATPase activity inhibition assay

The experimental assays were carried out using the method described by De La Cruz and Ostap for measuring the actin-activated Mg²⁺-ATPase activity of myosin.³⁷⁵ In short, the enzyme Pyruvate Kinase (PK) synthesizes ATP from the ADP and Pi released during the ATPase cycle of actomyosin while phosphoenol pyruvate (PEP) is transformed into pyruvate. Pyruvate is then used as a substrate for lactate dehydrogenase (LDH) while NADH is oxidized to NAD⁺ by LDH. From the reaction chain, one NADH molecule is consumed per ATP regenerated.³⁷⁵ Since NADH absorbs light at 340 nm and NAD⁺ does not, changes in NADH concentration are measurable through absorbance-based experiments.³⁷⁵ By carrying this assay over time, the change in absorbance, and thus

NADH consumption, in the presence of SMM2 and actin is obtained and then fit to a linear function.³⁷⁵ The slope of the fit provides the steady-state ATPase activity of myosin.³⁷⁵ If SMM2 is inhibited, NADH oxidation is slowed down due to the lack of ADP as ATP hydrolysis is prevented. Thus, the slope of the reaction (A340/s) is less inclined than in the case of uninhibited myosin, where NADH consumption is faster.³⁷⁵

For the assays, 5 mg of each compound was diluted in 100,3 μL of DMSO. Following, 6 serial dilutions of 10 μL were carried out. An activity mix was prepared containing 918,4 μL of KMg50 buffer, 8 μL of 100 mM NADH, 7,50 μL of 4,000 U/ml LDH, 12,50 μL of 10,000 U/ml PK, 6,25 μL of 100 mM PEP and 173,41 μL actin (25 μM) for a total volume of 1161 μL . The assay is conducted as follows: (1) in a 384-well plate, add 38.7 μL of activity mix to three wells (controls). (2) Then, 31.4 μL of myosin (3.5 μM) is added to the activity cocktail and mixed. (3) Add 38.7 μL of the SMM2-containing cocktail mix to the other wells (7x3 wells). (4) Using a robot, add 0.5 μL of compound to the activity cocktail and mix. After a 10 minute wait, add 0.8 μL of 100 mM buffered ATP and mix again. The A340 decays are measured at 25°C. Actin was purified from rabbit skeletal muscle as described in Sirigu *et al.*³⁵⁰ The extinction coefficient of NADH ($\epsilon_{340} = 6220 \text{ M}^{-1} \text{ cm}^{-1}$) is used to convert the absorbance at 340 nm to an ADP concentration.³⁷⁵

6.3: Results and Discussion

6.3.1: Structural analysis of the CK-571 binding pocket

The availability of a crystallographic structure of the SMM2/CK-571 complex is a landmark achievement for the development of novel therapeutic avenues targeting SMM2. In particular, this structure shows a previously unknown allosteric pocket where ligand binding inhibits SMM2 activity by stabilizing an intermediate state with low actin affinity and trapping the myosin motor domain. The co-crystal inhibitor was found to have an IC_{50} of 12 nM, making it a highly potent SMM2 inhibitor. More recently, the crystal structures of two other inhibitors from cytokinetics were solved (unpublished data) in complex with SMM2. Interestingly, these compounds target the same allosteric pocket in the same state, and have very similar binding modes to CK571. For these ligands, known as CK-144 and CK-903, no IC_{50} value is available but they are inhibitors of SMM2. The inhibitor CK-571 binds to SMM2 by inserting a hydrophobic moiety in either side of the SH1 helix, with the isoquinoline carbamate portion (Pocket 1; P1) held fixed between the relay helix and the SH1 helix, a hydrophobic moiety on the other side (Pocket 2; P2) and an N-terminal extending tail (Pocket 3; P3) (**Figure 6.4A** and **4B**). As noted by Sirigu *et al.*³⁵⁰, CK-571 does not establish any direct hydrogen bonds with protein residues and its

interactions are mostly hydrophobic, relying on a network of van der Waals contacts to interact with SMM2 (**Figure 6.4A** and **4B**). By binding to the SMM2 allosteric pocket in this pincer-like manner, CK-571 prevents the repriming of the lever arm and arrests the cycle in this intermediate state between the rigor and the pre-powerstroke states (**Figure 6.4B**).

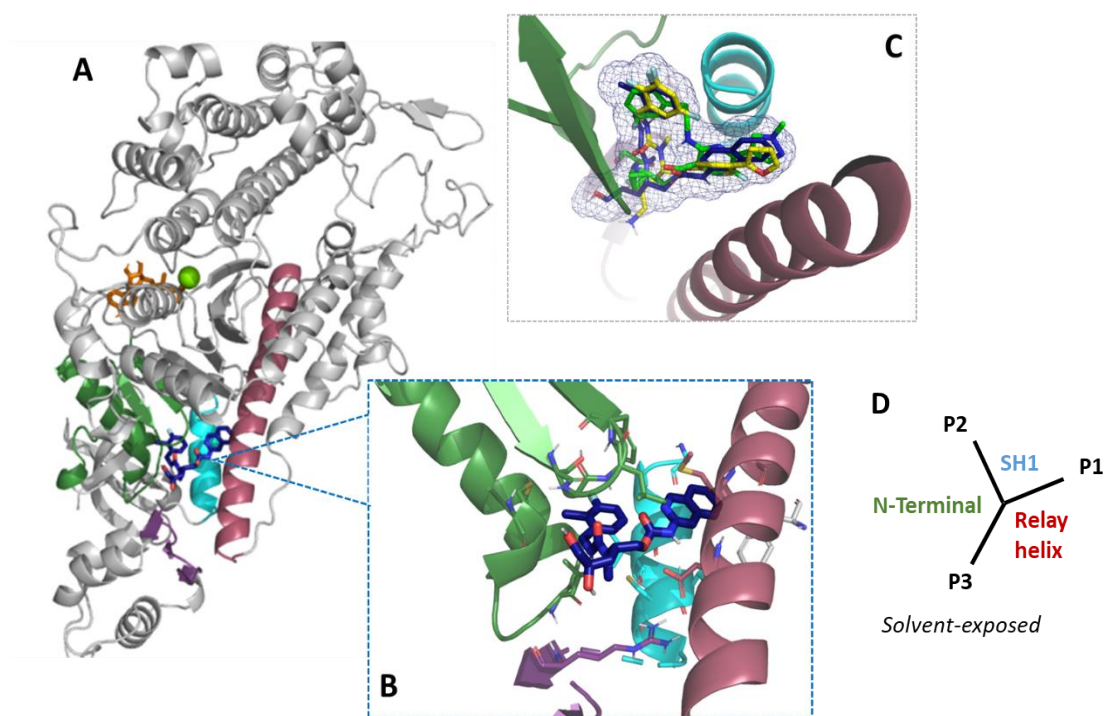


Figure 6.4 - Three-dimensional view of the SMM2/inhibitor systems. The N-terminal domain is shown in dark green, the Relay helix is shown in red, the SH1 helix is highlighted in cyan and the three β -sheets of the converter domain are shown in purple. A) Structure of the SMM/CK-571 biomolecular complex. In dark blue is CK-571, ADP is shown in orange and the magnesium ion is shown as a green sphere. B) Zoom into the binding site of CK-571, showing the residues lining the binding site as sticks and colored according to atom type. The top-view allows one to see the carbamate moiety of CK-571 inserted between the Relay and the SH1 helix and the chloro-fluoro-phenyl moiety of CK-571 inserted on the other side of the SH1 helix. It also highlights the presence of a polar tail extending towards the N-terminal domain. C) Superimposition of the three SMM2 cytokinetics inhibitors in the binding site. A mesh representation of the CK-571 volume is shown in blue, the molecular structure of CK-144 is shown in yellow and the molecular structure of CK-903 is shown in light green. D) Illustrative scheme of the binding mode of the CK inhibitors in the SMM2 allosteric pocket.

The binding mode of CK-571 is shared by the other two cytokinetics inhibitors, where the SH1 helix is surrounded from either side and a polar tail extends outward into the N-terminal domain. However, CK-903 envelops the SH1 helix without completely exploring the available volume on the left side of the SH1 helix instead of inserting its chemical groups deeply in P2 (**Figure 6.4C**). For CK-144, the binding mode is similar to that of CK-571, inserting a moiety in P2 and another large, hydrophobic moiety in P1, between the SH1 and the Relay helix. Another important point to be raised is that the interactions

between CK-144 and SMM2 are mostly nonpolar and, similarly to CK-571, it appears that CK-144 does not establish any direct hydrogen bonds with SMM2. From **Figure 6.4A** we can see that the allosteric binding site of CK-571 is far away from the nucleotide binding site and from the ATP binding domain. However, by limiting the movement of the Relay helix, it prevents ATP hydrolysis and arrests the cycle at the beginning of the conformational transition towards the pre-powerstroke state. It is apparent that while the pocket does not appear to be fully filled by the CK ligands, the CKs fill the large majority of the pocket, establishing many van der Waals contacts, and thus it comes as no surprise that the nonpolar interactions seem to be the main driving force stabilizing CKs binding. Further, the fact that all three ligands bind in the pocket with very similar binding modes implies that a particular ligand geometry is necessary to inhibit SMM2 activity by targeting this allosteric pocket. The information from the crystal structure highlights that potential binders must have two hydrophobic moieties to envelop the SH1 helix by either side and an outward extending polar tail towards the N-terminal domain for appropriate binding in the pocket, in a Y shape as shown in **Figure 6.4D**.

6.3.2: Results from a previous VS campaign on SMM2

A prospective VS campaign on SMM2 was carried out in the past, using ChemFlow to target the SMM2-CK571 allosteric pocket. This campaign was performed on the CN database by our group in 2018. Two free energy rescoring steps were performed to refine the compound ranking obtained by molecular docking. Because implicit solvent MD simulations are significantly cheaper than explicit solvent ones, binding free energy calculations using MM/GBSA were first performed on configurational ensembles obtained from implicit solvent MD simulations. A subset of compounds was then prioritized and later simulated in explicit solvent. Again, binding free energy calculations were carried out using MM/GBSA for this subset and some compounds were acquired for experimental testing by collaborators in the Houdusse group at Institut Curie, Paris. The predicted binding free energies for this subset, along with the predictions of the known inhibitors of SMM2, are shown in **Figure 6.5**. In both instances where MM/GBSA was employed, the GB model used to compute the polar contribution to the solvation free energy was the one of Onufriev, Bashford and Case (OBC, GB2), which corrects the Hawkins, Cranmer and Thrular (HCT, GB1) GB model by rescaling the effective Born radii and accounting for the interstitial spaces between atom spheres.²⁰¹ The solute internal dielectric constant was set to 1 (permittivity of vacuum), the external dielectric constant set to 78.5, the non-polar contribution to the solvation free energy was evaluated from the Solvent Accessible Surface (SAS) computed using the LCPO algorithm and no entropy terms were considered in the final ranking of CN compounds.

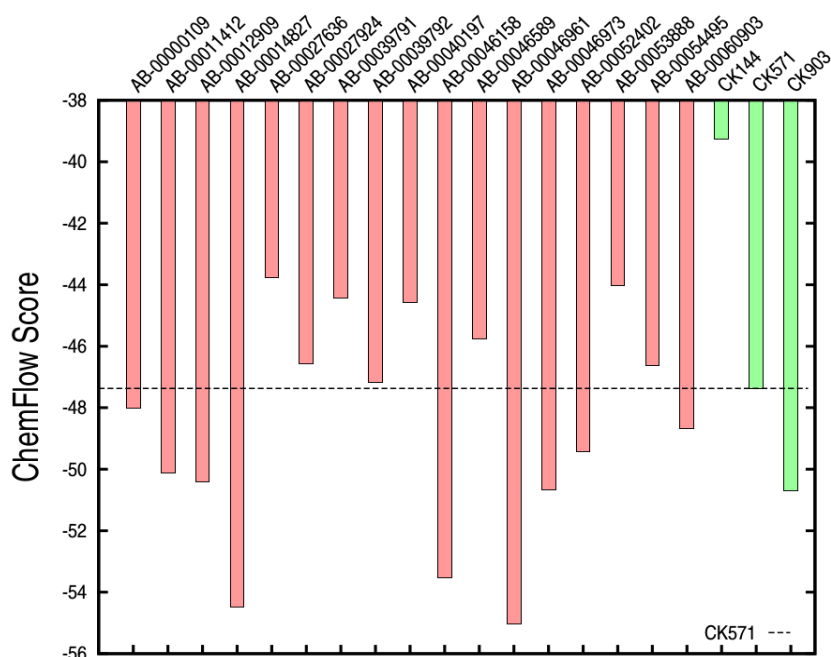


Figure 6.5 – Predicted binding free energies for the subset of prioritized compounds from the previous VS campaign targeting SMM2.

From this screening campaign it emerges that CK-571 was not predicted as the most affine of the CK compounds. Furthermore, several compounds were predicted as better than CK-571. It was ranked 11th in the final ranking of this screening behind CK-903, which is another SMM2 inhibitor, and nine CN compounds. From these nine compounds, seven had MM/GBSA scores which were significantly better than that of CK-571 ($\Delta\Delta G^\circ > -2.5$ kcal/mol), which is surprising because CK-571 is a 12 nM inhibitor of SMM2. Furthermore, the prediction for the other inhibitor, CK-144, is rather poor when compared to all the other compounds shown. This was unexpected, because CK-144 is known to inhibit SMM2 activity and is the one whose predicted binding affinity is the highest (more positive). The above results may indicate the presence of some errors in the calculations. Nonetheless, these compounds were acquired and experimentally tested by collaborators at Institut Curie to assess their ability to decrease SMM2 ATPase activity. Many of these compounds were promising when evaluating their binding modes on the computer screen. However, the experiments showed that none of the prioritized compounds was active (**Figure 6.6**).

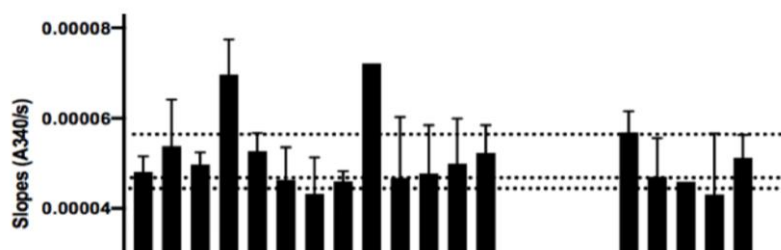


Figure 6.6 – Experimental measurements of ATPase activity. Assays were carried out on CK-571 (positive control), DMSO (negative control) and the prioritized compounds from CN. The assay monitors absorbance changes at 340 nanometers (A340), which are coupled to NADH oxidation through a series of coupled enzymatic reactions. The dashed lines are given by the A340 measurements on the negative control experiment, which is carried out using DMSO and no inhibitor.

From the figure above (**Figure 6.6**) none of the tested compounds from the CN library shows a significant decrease in SMM2 ATPase activity, which means that ATP hydrolysis is occurring normally in the presence of these compounds (see **Chapter 6.2.6**). Considering the error bars of the slopes of the fit applied to the A340 decays (A340/s), no CN compound inhibits SMM2 activity as the values of slopes for those experiments fall within the range obtained for the negative control experiment using DMSO. In the case of CK-571, which binds to SMM2 and arrests the myosin motor cycle, there is a significant decrease in the slope of the fit, meaning that ATP hydrolysis is being hampered. The decrease in the slope indicates a slower ATP hydrolysis rate because SMM2 is unable to continue with the force generating cycle. Obtaining a ligand exhibiting the effect observed with CK-571 was the objective of the VS campaign but such was not possible. While these results were disappointing, because they show that the VS approach was not able to find any active compounds, they allowed to reflect on the MM/GBSA setup employed. Upon inspecting the parameters used in the calculations, three possible terms that could be optimized in a future campaign were found: the solute internal dielectric constant (ϵ), the GB model and inclusion of entropic terms within the MM/GBSA calculations. The final aspect comes from the realization that most of the best ranked ligands were large and/or flexible molecules (**Figure 6.7**). Restraining large and flexible ligands in a binding site entails an entropic cost which, when not paid, leads to biased MM/GBSA calculations where bigger ligands are predicted as better binders. Thus, the likelihood of finding false positives within the subset of prioritized compounds increases when entropic terms are disregarded.

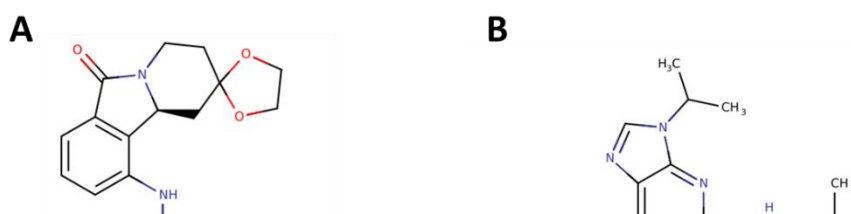


Figure 6.7 – Molecular structures of ligands with the best predicted binding affinities from the VS campaign of 2018, obtained using MarvinSketch from ChemAxon. A) AB-000014827; B) AB-00046961; C) AB-00046158; D) AB-00046973.

6.3.3: Calibration of MM/GBSA setup for the VS campaign on SMM2

Before undertaking a new VS campaign, the VS workflow required tuning to try and prevent a second unsuccessful round of experiments. The parameters for the MM/GBSA calculations were calibrated based on reference data, where the dataset was composed of the three known inhibitors and the subset of prioritized compounds found to be inactive in the previous VS campaign. The MM/GBSA binding free energies obtained from explicit solvent MD simulations for this dataset were used as reference data and correlated with MM/GBSA results obtained from configurational ensembles sampled by implicit solvent MD. The effect of ϵ and the GB model on the correlation between predictions from implicit and explicit solvent MD ensembles was queried simultaneously. The MM/GBSA setup selected for the new VS campaign was the one which exhibited the highest correlation between these calculations (**Table 6.1**). Correlation was assessed by evaluating the squared Pearson's correlation coefficient (R^2) and the Spearman correlation of ranks (ρ).

Table 6.1 - Benchmark calculations carried out on the dataset of prioritized compounds from the previous screening. In red is highlighted the setup used in the first screening whereas in green is highlighted the setup for future studies.

GB Model	R^2 ($\epsilon = 1$)	ρ ($\epsilon = 1$)	R^2 ($\epsilon = 2$)	ρ ($\epsilon = 2$)	R^2 ($\epsilon = 4$)	ρ ($\epsilon = 4$)	R^2 ($\epsilon = 10$)	ρ ($\epsilon = 10$)
GB1	0.54	0.62	0.63	0.75	0.64	0.75	0.64	0.74
GB2	0.37	0.45	0.59	0.69	0.64	0.74	0.64	0.76
GB5	0.29	0.43	0.55	0.65	0.62	0.76	0.64	0.75
GB7	0.35	0.47	0.59	0.68	0.64	0.76	0.64	0.76
GB8	0.07	0.40	0.48	0.65	0.60	0.76	0.63	0.76

The benchmark shows that the setup used in the previous screening (GB = 2, $\epsilon = 1$) exhibited poor correlation between calculations carried on implicit and explicit solvent MD simulation data ($R^2 = 0.37$, $\rho = 0.45$). The implicit solvent MM/GBSA calculations configure a filtering step in the VS workflow with the aim of selecting a subset of compounds to be investigated with longer and thus more expensive explicit solvent MD simulations. Thus, the fact that the correlation is low decreases our confidence in the filtering step and raises the question of whether the prioritized compound subset is meaningful. The final aim of this benchmark was to find a MM/GBSA setup with the highest possible agreement between the calculations ran on the two types of simulations, as it would mean that the free energy rescoring steps would be consistent with each other. Higher degree of correlation was always obtained when ϵ was set to either 4 or 10, as shown in **Table 6.1**.

The solute internal dielectric constant is fundamental to compute the electrostatic terms and the polar contribution to the solvation free energy, meaning that fine tuning of this parameter is critical. In the calculations which set ϵ to 4 or 10, some small variations in the predictability of the MM/GBSA calculations were observed when employing different GB models. Considering the Spearman coefficient and the squared Pearson's coefficient, it was found that GB = 7 and $\epsilon = 4$ or 10 were the most predictive combinations. By tuning both the GB model and ϵ simultaneously, we improved the correlation between MM/GBSA calculations carried out in implicit and explicit solvent, increasing from $R^2 = 0.37$ to 0.64 and from $\rho = 0.45$ to 0.76. To select whether to set ϵ to 4 or to 10, we evaluated the degree of separation between CK-571 and similar compounds which are known actives (CK-144 and CK-903) and inactives from the previous VS campaign. In particular, the setup which was kept was the one which better separated actives from inactives (**Figure 6.8**).

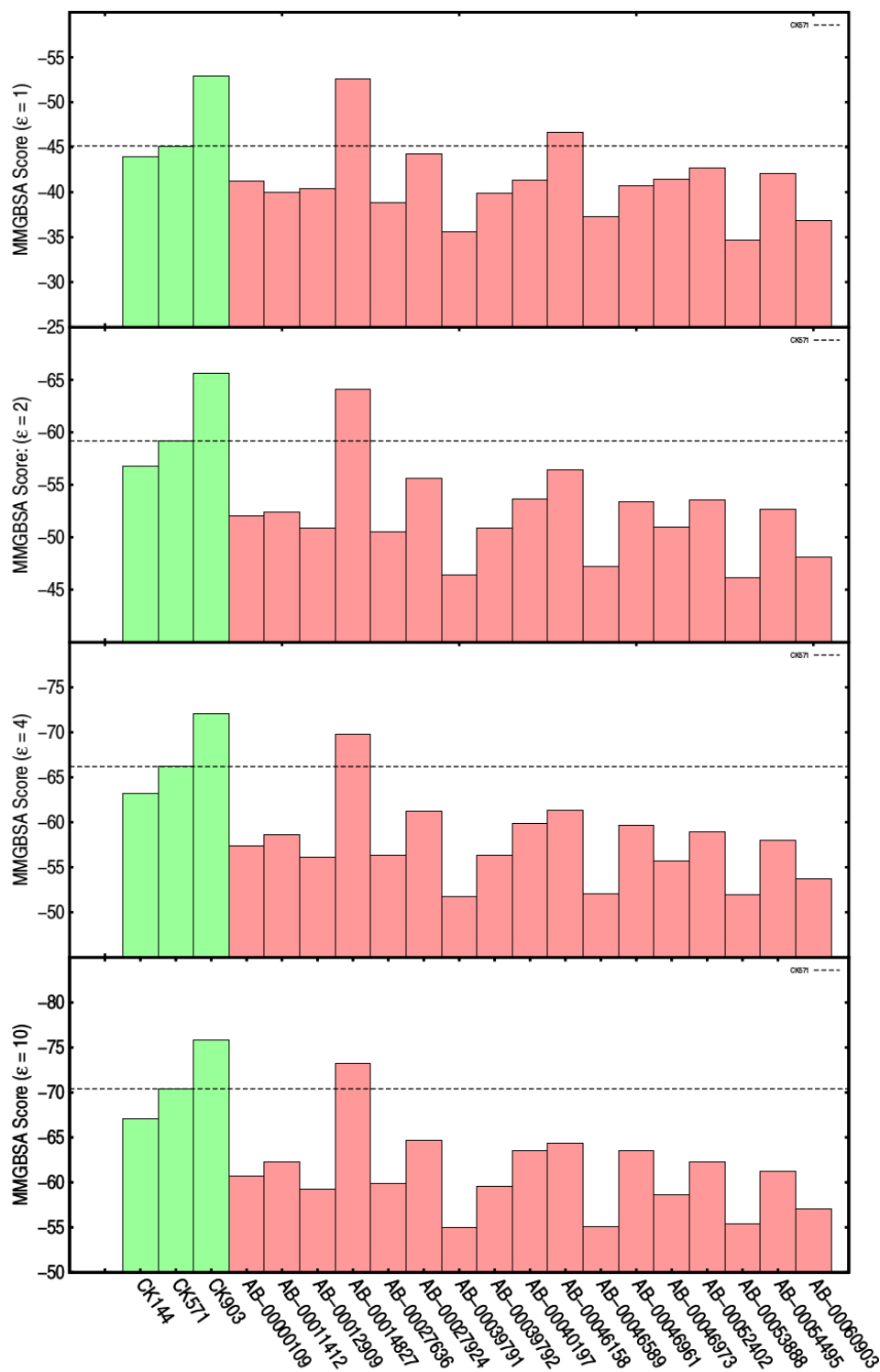


Figure 6.8 - Results arising from the SMM2-ligand complex MM/GBSA calculations using different dielectric constants and the GB model ⁷³⁶⁹ on the benchmark dataset. Compound AB-00014827 is still predicted as better than all CKs but CK903, as previously observed, however a more accurate ranking of the ligands, in line with what was observed experimentally, is obtained when the polarizability effects are implicitly accounted for through an increase in the dielectric constant. These results are obtained from the explicit solvent MD simulations.

It emerges from the data that using $\epsilon = 4$ or 10 improves separation between the actives and the inactives with respect to $\epsilon = 1$ or 2, with only one compound better predicted than two of the actives and no compound predicted better than all three CKs. However, no significant differences are observed between the two calculation setups. Thus, $\epsilon = 4$ was selected since it is a value usually used according to the literature for protein-ligand systems.^{201,229} Taken together, the data indicates that in the first screening the setup used for MM/GBSA free energy rescoring was suboptimal and thus produced many false positives. Introducing a higher dielectric constant and optimizing the GB model used to compute the polar contribution to the solvation free energy appears to address this issue. The optimized setup allows differentiation between the known active compounds and the inactives, which previously was not possible. To address the issue of the bias due to the neglect of entropic contributions, a correction term accounting for the ligand configurational entropy loss upon binding using QHMB²¹⁸ (see **Chapter 5**) was added in the final rescoring step of the newest VS campaign (**Chapter 6.2.2**). By introducing this ligand-dependent correction, the bigger and more flexible ligands are expected to be more penalized than small and rigid compounds.²¹⁸ Thus, the bias towards more flexible ligands found in the initial VS campaign is also expected to be accounted for.

6.3.4: Virtual Screening campaign

The most recent virtual screening campaign carried out implements an original docking plus free-energy rescoring approach. The methodology proposed to rescore molecular docking binding poses using end-point binding free energy calculations, which rely on configurational sampling of the chemical species by all-atom Molecular Dynamics. An original entropy correction based on a multi-basin decomposition of the ligand configurational space in the bound and unbound states was also added to correct the final MM/GBSA binding free energy estimates. We expect to increase the accuracy of the ranking obtained from binding-affinity predictions and reduce the false-positive rate by including our entropy correction. This strategy is fully automated and implemented by the *ChemFlow* software (Gomes *et al.*, **in progress**).

The prepared library was docked to the allosteric pocket of CK-571 in the SMM2 structure through PLANTS¹⁷⁷, and the best scored binding mode per compound was retained. The 2300 complexes were then simulated in implicit solvent for 1 ns, using the Generalized Born model 7³⁶⁹ and an solute internal dielectric constant of 4 (see **Chapter 6.3.2** and **6.3.3**).

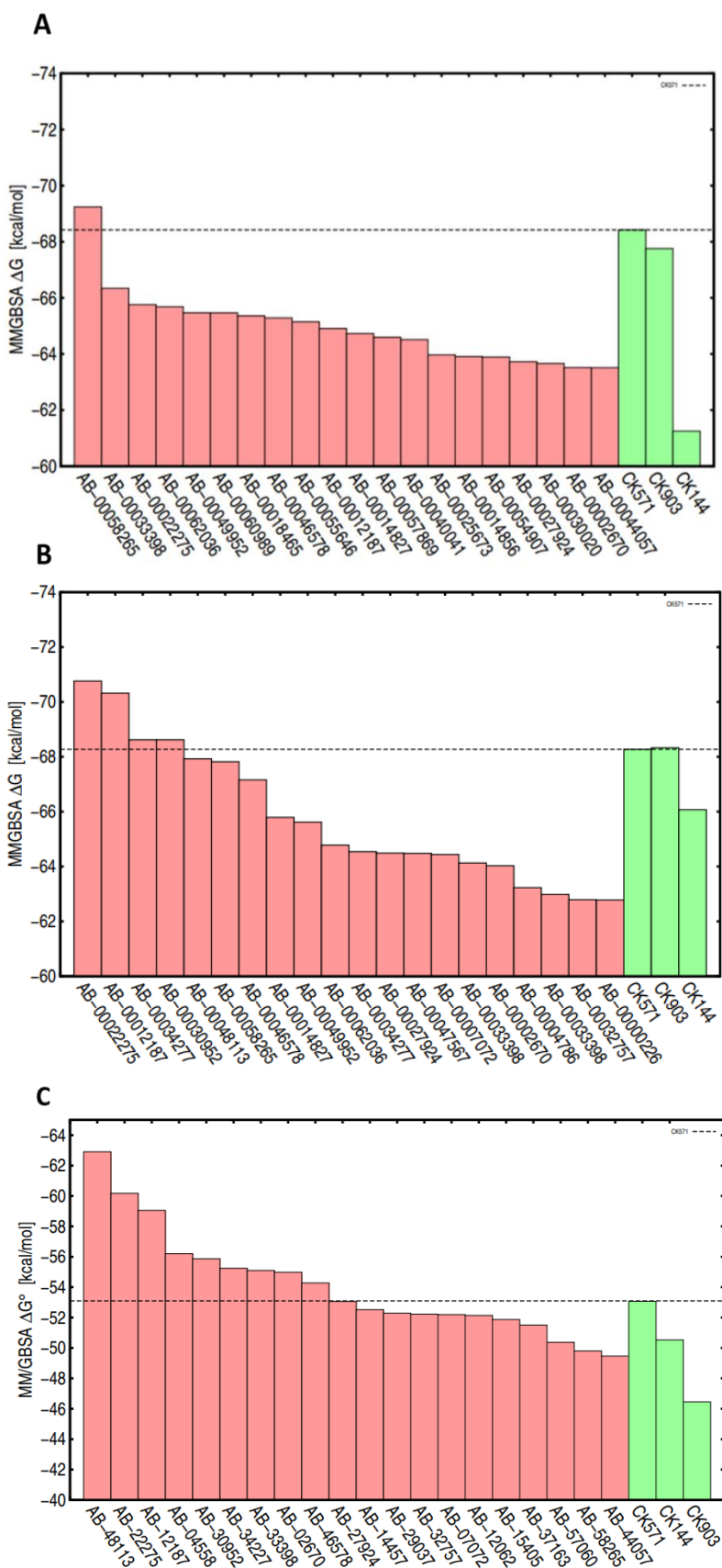


Figure 6.9 – Binding free energies obtained for the top 20 compounds of the CN at various steps of the screening campaign on the Chimiotèque National. The red bars refer to the predicted binding free energies

for compounds of the Chimiotèque and green bars correspond to the predicted binding free energies of the reference CK compounds. The dashed line highlights the predicted binding free energy of CK-571, which we use as a reference. A) Top 20 compounds obtained following the MM/GBSA calculations using implicit solvent trajectories. B) Top 20 compounds obtained following the MM/GBSA calculations using explicit solvent trajectories. C) Top 20 compounds obtained following the MM/GBSA calculations using explicit solvent trajectories and including the QHMB entropy correction.

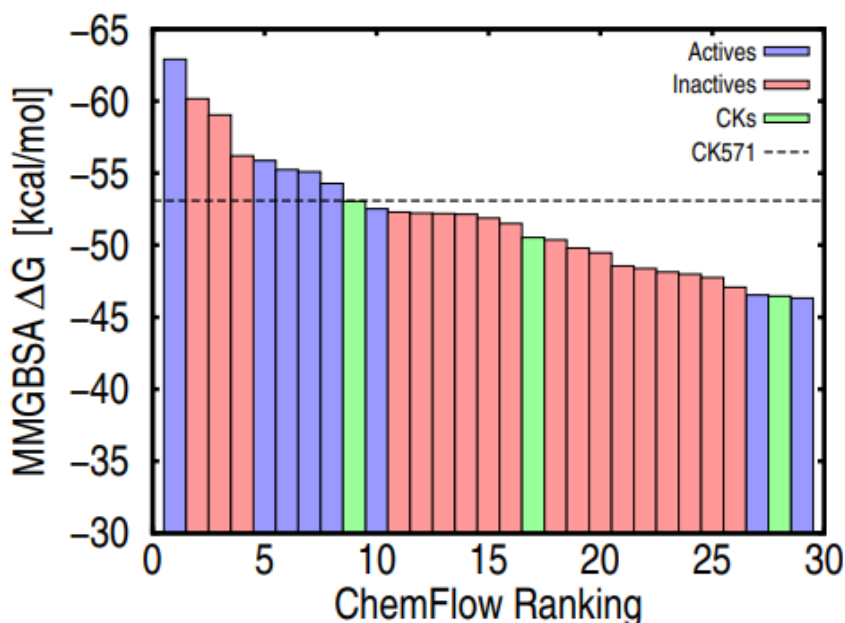
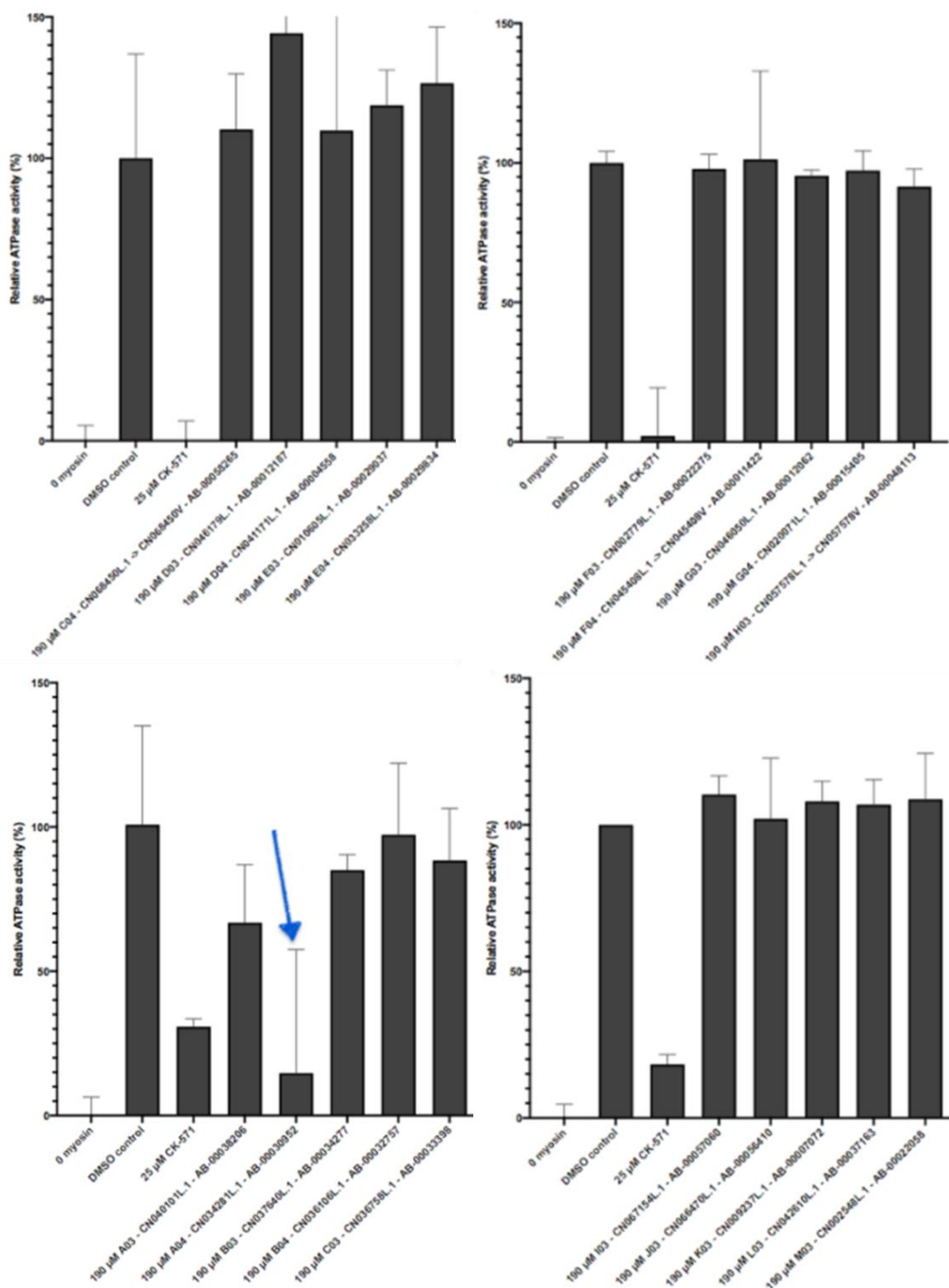


Figure 6.10 - Virtual screening results of the campaign against Smooth Muscle Myosin II (SMM2). The predicted affinity of experimentally inactive compounds prioritized from the Chimiotèque Nationale is shown in red. The predicted affinity of three known inhibitors from Cytokinetics is shown in green. Compounds with detectable activity in in-vitro tests (see below) or *hits* are shown in blue.

Throughout the simulations in implicit solvent, the backbone of the protein was restrained with a moderately high force constant (See **Chapter 6.2**) to maintain the bound conformation of the complex. The MM/GBSA method^{196,201,211} was used to compute the binding free energies of the protein-ligand complexes simulated in implicit solvent employing the same GB model and solute internal dielectric constant that were used to run the simulation. A subset containing the top ranked 60 compounds from the calculations carried out at the implicit solvent step was selected for further studies using explicit solvent MD. At the implicit solvent stage, only one ligand, AB-00058265, was predicted with a better binding affinity than CK-571. Following 100ns explicit solvent MD simulations, the binding free energy of the compounds was computed using MM/GBSA and the compounds were ranked again. Some re-ranking was observed, but in general the implicit and the explicit solvent results were well correlated (**Figure 6.9A** and **9B**). However, it was noticed that the best ranked compounds were in most cases big and flexible ligands, which is a known artifact of MM/GBSA calculations when entropic terms are neglected, as was the case of our calculations to this point. Thus, these calculations

were complemented using the QHMB correction²¹⁸ to account for the entropic cost of restraining the ligand within the binding site, computed by taking the difference between the QHMB entropy of the ligand in the bound state, extracting the ligand coordinates from the complex simulation, and in the unbound state in solution.²¹⁸ The emerging protocol is as follow: (1) use docking to produce initial coordinates of the complex; (2) use implicit-solvent MD for a fast ranking and filtering; (3) complete by explicit-solvent MD and QHMB for the final ranking.



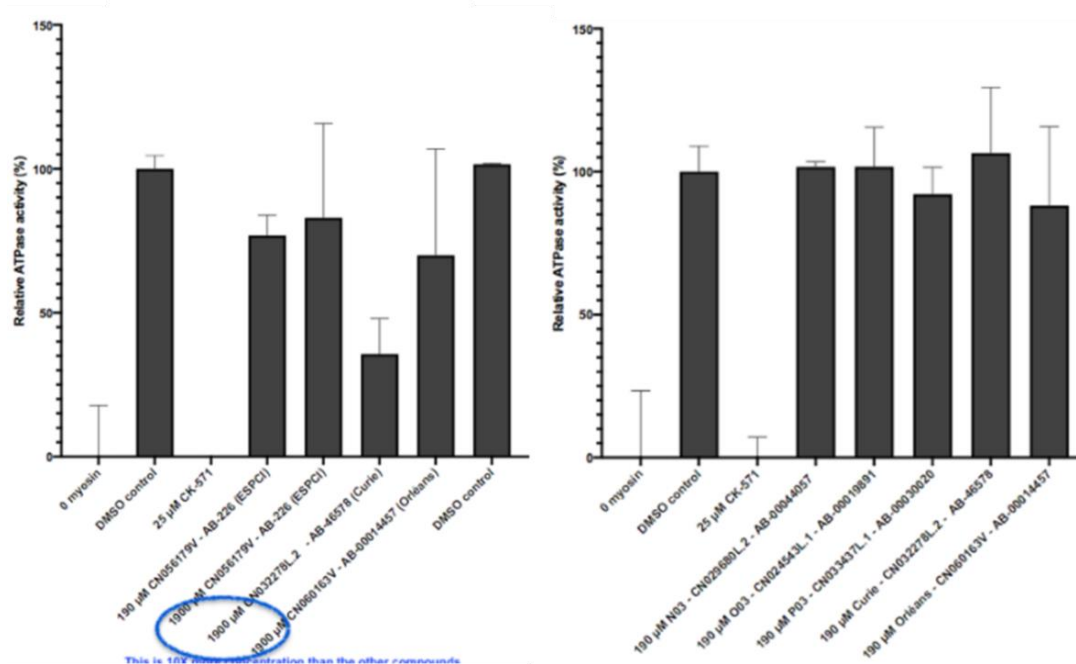


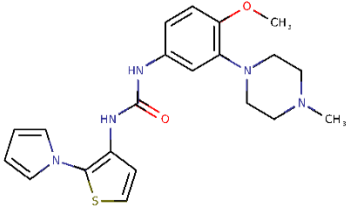
Figure 6.11 – Results for the SMM2 ATPase inhibition assay carried out at Institut Curie. In the y axis is reported the relative SMM2 ATPase activity in the presence of DMSO, CK-571 or each ligand from the CN that was acquired at 190 µM. The first column corresponds to the ATPase activity in absence of myosin. The blue arrow highlights compound AB-00030952, which is the most active compound. However, this compound has a very large error bar. Three compounds are shown with a blue circle around, corresponding to compounds whose activity was assessed at very high concentration (1.9 mM).

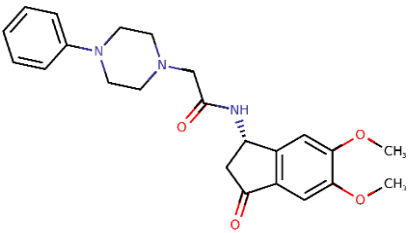
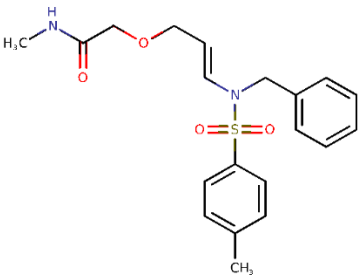
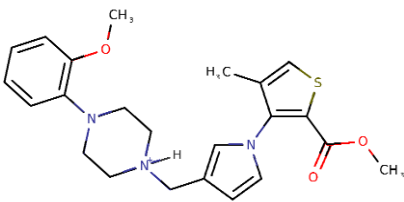
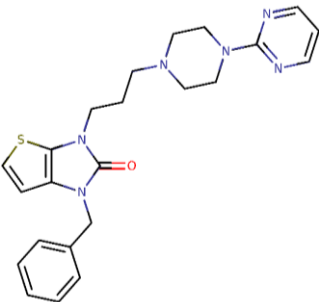
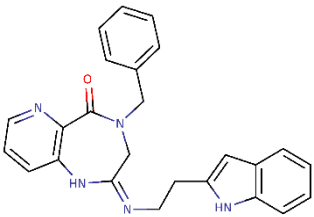
Following QHMB correction²¹⁸, a significant re-ranking of the compounds was observed. We found several compounds with predicted binding free energies similar to that of CK-571 (**Figure 6.9B** and **9C**). These compounds were neither the largest nor the most flexible compounds in the CN library and exhibit different chemotypes from CK compounds (**Table 6.2**), which is attributable to the ligand-dependent character of the QHMB correction. In particular, highly flexible ligands were penalized more strongly than more rigid ligands by QHMB. As an example, compound AB-00048113 was ranked 7th before QHMB correction. After the correction was applied, it became the top ranked compound whereas the compound which was ranked 8th (AB-00058265) before QHMB correction is now ranked 21st. The results in **Figure 6.9** illustrate the effect of moving from implicit solvent to explicit solvent MD simulations, as well as the effect of including the QHMB correction in MM/GBSA binding free energy predictions. Further, CK-571 and CK-903 were predicted with similar binding free energies before the correction, while CK-144 is less affine. Following QHMB correction, CK-903 was heavily penalized and became the worst among the three inhibitors. Among the top-ranked compounds, three are predicted to have a binding affinity stronger than CK571, and about 20 compounds present predicted affinities comparable to the known SMM2 inhibitors from Cytokinetics.

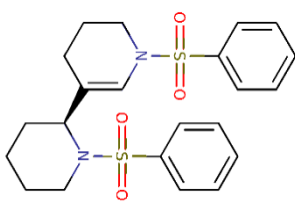
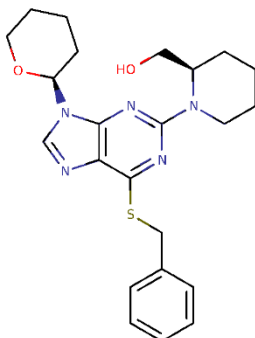
From the total of ~60 compounds, 26 have been acquired and sent for experimental testing. The experimental results were collected by our collaborators in Institut Curie at 190 μM concentration in the presence of 2 μM SMM2 and 25 μM or 40 μM actin using an ATPase inhibition assay (**Figure 6.11**).

By setting as zero of inhibition the signal collected in the presence of DMSO, corresponding to the control experiment, the data show 8 compounds with detectable inhibition. The chemical structures of these compounds along with the % inhibition at 190 μM are given in **Table 6.2**. Most of these compounds have mild but detectable activity ($\text{IC}_{50} > 100 \mu\text{M}$) and are unique chemical entities. Interestingly and despite a large error bar, AB-00030952 shows 85% of SMM2 inhibition at 190 μM . Considering these compounds as hits, the hit rate of the screening protocol is approximately 30% (8 actives/26 tested). These compounds are mild inhibitors and thus it could be argued that the activity captured is so mild that they should not be considered as hit at all. However, as reported in the review by Hevener *et al.*,⁵² the threshold used to define active or inactive compounds in HTS campaigns varies within the literature, with several researchers considering even compounds with a very high IC_{50} (i.e. $\text{IC}_{50} > 400 \mu\text{M}$) as hits.⁵² In particular, Hevener *et al.* detailed at the time that 56 studies used an activity cut-off between 100-500 μM and 25 studies used a criterion above 500 μM . Their justification for defining as hits such low activity compounds is to enrich the hit library in terms of structural diversity.⁵² Among the compounds with detectable activity *in vitro*, 5 out of 8 are in the top-10 predictions; see blue bars in **Figure 6.10**. Tables summarizing the results obtained for the subset of the top-60 compounds and their physicochemical properties are given in **Annexes S6.2** and **S6.3**.

Table 6.2. 2D chemical structures and % of inhibition of the identified hits by the Houdusse group. The ranking of compounds is given with respect to the prioritized subset of virtually screened compounds, not only to those tested experimentally.

CN Number	Structure	Ranking	SMM2 Inhibition @190 μM
AB-00030952		5	85%

<p>AB- 00038206</p>		<p>36</p>	<p>35%</p>
<p>AB- 00000226</p>		<p>38</p>	<p>20%</p>
<p>AB- 00034277</p>		<p>6</p>	<p>15%</p>
<p>AB- 00033398</p>		<p>7</p>	<p>10%</p>
<p>AB- 00048113</p>		<p>1</p>	<p>10%</p>

AB-00014457		12	10%
AB-00046578		9	65% (@1.9 mM)

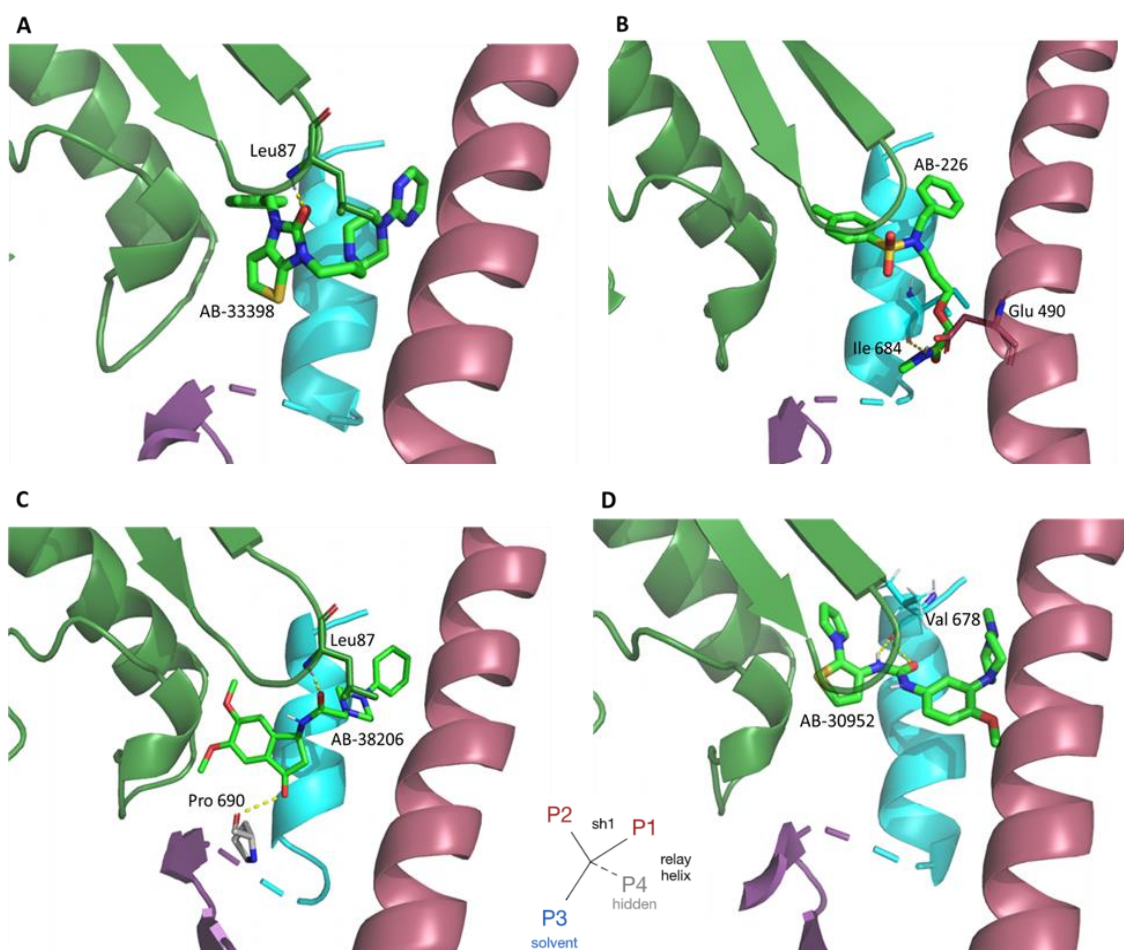


Figure 6.12 – Binding mode of some of the hit compounds found by virtual screening of the Chmiotèque

National du CNRS. The hydrogen bonds established between the ligands (bright green) and SMM2 are shown in yellow alongside the residue name and number. A) Binding mode of AB-00033398, ranked 6th by our VS approach. B) Binding mode of AB-0000226, ranked 37th by our VS approach. C) Binding mode of AB-38206, ranked 40th by our VS approach. D) Binding mode of AB-00030952, the most active compound, in experimental assays ranked 5th by our VS approach.

6.3.5: Structural analysis of the predicted hit compounds

Following the *in silico* screening campaign, eight new SMM2 inhibitors were found. While these show mild activity, they can be useful to understand how to better explore the binding pocket and how plastic and adaptable this pocket is. Comparison of the binding mode of the CK compounds to the hit compounds shows that P1 is only partially occupied by the CKs and could be better filled (**Figures 6.4** and **6.12**). The binding mode of some of the hits, notably AB-00033398, explores P1 by inserting a piperazine+aromatic terminus deep in the pocket between the SH1 and the Relay helices (**Figure 6.12A**). This ligand also establishes a direct hydrogen bond with the protein in leucine 87, a residue which is part of a flexible loop. However, since the loop is flexible this hydrogen bond is not maintained throughout the simulation and thus the magnitude of its contribution is likely negligible. Other compounds, like AB-0000226 (**Figure 6.12B**), explore both P1 and P2 in a shallower manner. However, AB-0000226 also explores longitudinally the crevice along the SH1 and Relay helices, inserting itself between them and possibly preventing their movements this way. Additionally, by exploring this crevice extending downward from P1, effectively occupying a fourth pocket (P4), it is able to maintain two hydrogen bonds with isoleucine 684 and glutamine 490. We also find that P2 could also be filled better. In particular, it looks like it can accommodate larger and bulkier groups, such as in the binding mode of AB-00038206 (**Figure 6.12C**). In addition, AB-00038206 further establishes two hydrogen bond interactions, with leucine 87 and proline 690. In similar fashion with AB-00033398 (**Figure 6.12D**), the hydrogen bond with leucine 87 may not be so relevant due to the intrinsic flexibility of the loop portion. However, the presence of the hydrogen bond at proline 690 is likely to contribute to the stabilization of the binding mode.

In general, the binding modes of the hit compounds seem to suggest that it is possible to explore deeper both P1 and P2. However, as shown by the CK compounds, occupying simultaneously both pockets seems to be critical for potent inhibition of SMM2 and it is not clear how deep these pockets should be filled in order to improve the potency of compounds. For most of the hits, only P1 or P2 are filled completely while the other pocket is left mostly unoccupied. By analyzing the binding mode of the mild hit compounds, we have also found that along the z-axis in P1 another pocket can be found, named P4. This pocket is explored by some of the hit compounds discovered. Finally,

comparing the binding mode of AB-00030952, the most active compound, and CK-571 highlights how crucial it is that P1 and P2 be both explored deeply and simultaneously. Both of these compounds explore the two pockets around the SH1 helix, which is central to arrest the conformational transition to the pre-powerstroke state. We find that the predicted binding mode of AB-00030952 explores deeply both P1 and P2, which could be the reason why it is much more active than the other hits where only one of these two pockets is completely filled. However, AB-00030952 does not have an N-terminal extending polar tail, which is characteristic of the CK compounds. There needs to be caution when inferring on how to improve binding potency from mildly active compounds. Nonetheless, comparison between the predicted binding mode of the hit compounds and the known inhibitors illustrates that the allosteric pocket exhibits a degree of plasticity and can accommodate chemical scaffolds which are structurally different from the CKs. Thus, it seems that the pocket can be further explored *en route* to new SMM2 modulators with therapeutic goals.

6.4: Conclusions

In this study, the SMM2/CK-571 allosteric pocket was described and shown to be targeted by two other inhibitors sharing the CK-571 binding mode (**Figure 6.4**). The structures of these ligands within the allosteric pocket were obtained from a collaboration with the Houdusse group in Institut Curie. Together with the high-quality SMM2/CK-571 crystal structure (PDB **5M05**), the ligands provide the precise location of the allosteric pocket of interest and highlight the existence of three main sub-pockets. Two of the sub-pockets are fundamental for SMM2 inhibition and are located on either side of the SH1 helix. The third sub-pocket is in fact occupied by a polar tail which extends towards the SMM2 N-terminal domain. A previous VS campaign carried out in our group towards the SMM2/CK-571 allosteric pocket highlighted issues in the free energy rescoring step. In particular, it was hypothesized that the MM/GBSA calculation setup applied was suboptimal as none of the prioritized compounds which were experimentally tested was able to affect significantly SMM2 ATPase activity. As such, all prioritized compounds turned out to be false positives. After verifying the MM/GBSA calculation setup, it was found that some parameters could be optimized. Notably, the solute internal dielectric constant and the GB model could be tuned and the lack of entropic terms had to be addressed. Using the three SMM2 inhibitors and the inactive compounds from the 2018 VS campaign as a reference dataset, benchmark calculations were carried out to optimize the solute internal dielectric constant and the GB model for future usage in SMM2 VS campaigns. To account for the ligands configurational entropy loss upon

binding, an entropic correction was introduced by QHMB calculations. Thus, in the 2021 VS campaign this allosteric site was targeted by means of a virtual screening campaign which combined molecular docking with an SMM2-tuned free energy rescoring step implemented by *ChemFlow*, along with an original entropy contribution to the binding free energy. This screening setup allowed the identification of several hits with mild but detectable inhibition in biochemical assays measuring the degree of inhibition of the ATPase function of SMM2. Although only mildly active, these hits provide chemically diverse molecular scaffolds that are privileged starting points for chemical optimization by both *in silico* methodologies and/or medicinal chemistry. Following experimental testing, a hit rate of 30% was obtained on a very difficult target partly due to the fact that inclusion of a ligand entropy correction removes the typical bias towards larger compounds and decreases the false positive rate. Although the large majority of the hits found are mildly active, there have been reports in the literature of hit compounds whose activity was around 500 μM or beyond.⁵² Analysis of the binding mode of the hits found by our *in silico* approach highlights that the binding site volume appears to be extensively explored by CKs but not completely filled. It also highlights the existence of a fourth binding pocket, here P4, which is not exploited by any the known inhibitors from Cytokinetics. The modeling results thus suggest that myosin's flexibility in this region of the motor domain could be better exploited for the design of inhibitors by attempting to fill this volume with bulkier groups in P2, a deeper exploration of P1 and occupying the crevice between the SH1 and Relay helices which configures P4. The original combination of X-ray crystallography, accurate *in silico* screening, and rapid *in vitro* testing which was carried out in this study emerges as an effective strategy for the identification of hits for the allosteric regulation of highly flexible and functional proteins such as myosin motors.

7. Concluding Remarks

Humanity has long tried to exploit the therapeutic properties of plants and other organisms to treat illnesses.^{2,10} With advances in Chemistry and Medicine, the development of compounds which are both potent and specific towards a given target became possible. The workflow through which these discoveries are produced is called the drug discovery pipeline. It is composed of many phases, such as hit identification, lead-optimization, pre-clinical and clinical trials and finally approval for commercialization.^{7,111,150,159,376} The traditional drug discovery pipeline relies on HTS for hit identification.¹⁵ Hit optimization is then carried out by introducing chemical modifications based on chemical intuition, compound synthesis and experimental testing in an iterative cycle until potency, ADMET properties and intellectual property issues have been resolved. However, the average cost of discovering and leading a compound until the market is around 2 billion dollars and requires on average 15 years.⁵⁹

The introduction of computational tools in drug discovery pipelines has grown over the years^{84,87,90,112,115,197} and has the potential to shorten the time required to put a drug in the market and reduce the economic costs of drug discovery. Computational methodologies have provided important contributions in the discovery of several approved drugs, like in the case of HIV-1 protease inhibitors^{30,207,377,378} or Captopril⁹⁴ to treat hypertension. Nonetheless, the accurate calculation of protein-ligand binding affinities in an efficient manner remains a grand challenge in computational chemistry and drug discovery.

In **Chapter 1** we presented the history of drug compounds and drug discovery as a whole, with special emphasis on the concept of drug-likeness, its origins and what it entails when applied in screening campaigns. The difference between affinity and activity was discussed and the gold-standard method for protein-ligand binding affinity determinations, ITC, was illustrated.

In **Chapter 2** we discussed some of the contributions introduced by computational methodologies when applied to drug discovery campaigns. A description of the steps to be carried out in a VS campaign was given, alongside some guidelines on how to carry each of them out. We then discussed the theory behind molecular docking and different types of scoring functions, which try to estimate the protein-ligand binding affinity in an efficient manner by introducing several approximations such as only considering the bound complex and near the neglect of entropic terms. Since molecular docking scores correlate poorly with experimental binding affinities, compound ranking is sometimes refined by a free energy rescoring approach on a subset of the best scored compounds

from docking. These calculations often require sampling of the dynamical behavior of the protein-ligand complexes either by Monte Carlo or Molecular Dynamics simulations.

In **Chapter 3** we first discussed the theory behind MD simulations and the ingredients necessary to carry them out. We then briefly described the rigorous FEP method, which is the gold-standard method for binding free energy calculations, and its limitations when applied to large systems such as allosteric proteins in VS campaigns. A common alternative to the application of FEP is to use end-point methods for free energy rescoring. Thus, most of **Chapter 3** is devoted to describe different end-point binding free energy calculation methods such as the Linear Interaction Energy, the Linear Interaction Energy with Continuum Electrostatics and the Molecular Mechanics – Poisson Boltzmann Surface Area family of methods. In MM/PBSA the binding free energy is computed from configurational sampling of the end-states of the binding reaction and is a sum of three terms: a potential energy term in gas-phase, a solvation free energy term and entropy. In particular, entropic terms in MM/PBSA are typically assessed through the RRHO approximation by NMA or QHA, which are both techniques whose accuracy breaks down when considering large and flexible molecules.

In **Chapter 4** we review some of the most well-known methodologies for computing single molecule entropies both within and outside of the RRHO approximation. We start by laying the ground using statistical mechanics, describing the HO model and its quantum-mechanical version. Then, the RRHO approximation is introduced and we describe the “mixture of conformers” approach. We then arrive at RRHO-based methods for single molecule configurational entropy calculations, where we describe NMA, QHA and QHA variants. Both advantages and disadvantages of each method are presented. We then continue by describing methods which go beyond the RRHO approximation. This Chapter concludes by discussing briefly how to use gas-phase entropy data for method benchmarking and how to produce benchmark entropy data in solution through computational methods.

In **Chapter 5**, we report the development of a novel single-molecule entropy calculation method. We first show that this method, named QHMB, achieves quantitative agreement with experimental gas-phase entropies from small-molecules in gas-phase (RMSE = 0.36, slope = 1.02). We then report the development of an automatic protocol to compute absolute ligand entropies by QHMB which is still able to achieve quantitative agreement with experimental data (RMSE = 0.6 kcal/mol, slope = 1.01). At the heart of the method, the vibrational frequencies are computed based on QHA. Thus, no energy minimization is necessary prior to the QHMB calculation. The method is readily extensible to calculations in solution because the effect of the solvent molecules is implicitly captured on the fluctuations of the solute degrees of freedom during Molecular Dynamics. We coupled QHMB to MM/GBSA calculations to account for the ligand configurational

entropy loss upon binding and saw the correlation to experimental binding affinities of 21 protein-ligand complexes improve significantly, by about 10%. On the other hand, computing entropic terms based on QHA led to a correlation decrease. Furthermore, we observed that the correction was ligand-dependent and stronger for larger and more flexible ligands. Considering these results, we thought of devising a VS workflow where the final ranking of compounds would be carried out by combining MM/GBSA calculations with the QHMB correction.

In **Chapter 6** we described a VS campaign where QHMB is combined with MM/GBSA calculations. Within the context of a collaboration with the Houdusse group at Institut Curie in Paris, we decided to apply our new VS protocol to the discovery of new inhibitors of SMM2. Smooth muscle contractility is implied in pathologies such as asthma³⁶⁰, prostatic hyperplasia³⁶¹ and chronic obstructive pulmonary disease.³⁵⁰ However, there exist no approved drug which specifically targets SMM2. Currently, there are three known specific inhibitors of SMM2 function but crystallographic structures are not published for two of them. For the other, CK-571, the crystal structure (PDB 5M05) shows that CK-571 binds to an allosteric pocket which opens in a short-lived intermediate state during the recovery stroke.³⁵⁰ A VS campaign was carried out on the CN library where the SMM2-ligand complex 3D structures were obtained by molecular docking on the allosteric pocket. Following application of a docking plus free energy rescoring workflow, a subset of 26 compounds was selected, acquired and experimentally tested. From the experimental assays we found 8 compounds which could inhibit SMM2 activity although mildly. Furthermore, 5 out of the 8 compounds were found in the first 10 ranked compounds by *ChemFlow* (Gomes *et al.*, **in progress**). Considering only the compounds tested experimentally, we find 6 active compounds within the top 10 (**Figure 6.10**). We discussed the binding mode of the new hits and also showed that they possess a different chemical scaffold from the known inhibitor, which could provide hints into the critical interactions ruling the binding towards this pocket.

To conclude, the work presented in this thesis highlights: (1) The development of a new entropy calculation method which is able to reproduce experimental gas-phase entropies and can be directly coupled to end-point binding free energy calculations; (2) The usage of QHMB to compute the ligands entropy loss upon binding and how it improves the agreement between MM/GBSA data and experimental binding affinities; (3) The development of an optimized docking plus free energy rescoring workflow where the final ranking step is carried out by MM/GBSA coupled to QHMB calculations. Application of this workflow selected several mild inhibitors targeting an allosteric pocket in SMM2; (4) The establishment of a VS setup for future VS campaigns aiming at the discovery of myosin inhibitors.

Additional VS campaigns have been carried out using other chemical libraries but are not reported here. Nonetheless, we have found additional SMM2 inhibitors from these screenings which are now being characterized. Moving forward, we aim at using the chemical information contained in the discovered compounds to pursue a lead-optimization campaign in close collaboration with our collaborators at Institut Curie and the group of Dr. Catherine Guillou at the Institut de Chimie des Substance Naturelles. The final objective is to propose new, highly potent SMM2 inhibitors with different chemistry than CK-571. Other future work to be published includes the study of the binding reaction between cytochrome c and calixarene molecules by the Attach-Pull-Release method and MM/PBSA coupled to QHMB and benchmark calculations to evaluate the effect QHMB has on the false positive rate when comparing molecular docking to MM/PBSA, MM/GBSA, MM/PBSA + QHMB and MM/GBSA + QHMB calculations.

Bibliographic References

- (1) Sander, T.; Freyss, J.; Kor, M. Von; Rufener, C. DataWarrior: An Open-Source Program For Chemistry Aware Data Visualization And Analysis. *J. Chem. Inf. Comput. Sci.* **2015**, *55* (2), 460–473.
- (2) Kinghorn, D. A.; Pan, L.; Fletcher, J. N.; Chai, H. The Relevance of Higher Plants in Lead Compound Discovery Programs. *J. Nat. Prod.* **2011**, *74* (6), 1539–1555.
- (3) Newman, D. J.; Cragg, G. M. Natural Products as Sources of New Drugs over the 30 Years from 1981 to 2010. *J. Nat. Prod.* **2012**, *75* (3), 311–335.
- (4) Atanasov, A. G.; Waltenberger, B.; Pferschy-wenzig, E.; Linder, T.; Wawrosch, C.; Uhrin, P.; Temml, V.; Wang, L.; Schwaiger, S.; Heiss, E. H.; et al. Discovery and Resupply of Pharmacologically Active Plant- Derived Natural Products : A Review. **2015**, *33* (8), 1582–1614.
- (5) Sneader, W. *Drug Discovery: A Review*; John Wiley and Sons Ltd.: Southern Gate, Chichester, West Sussex, England, 2005.
- (6) Dioscorides, P. *De Materia Medica. - Five Books in One Volume: A New English Translation by T.A.Osbaldeston.*, First Edit.; Osbaldeston, T. A., Ed.; IBIDIS PRESS: Johannesburg, South Africa, 2000.
- (7) Zanders, E. D. *The Science and Business of Drug Discovery*, Second Edi.; Springer International Publishing: Cambridge, UK, 2020.
- (8) Grandjean, P. Paracelsus Revisited : The Dose Concept in a Complex World. *Basic Clin. Pharmacol. Toxicol.* **2016**, *119*, 126–132.
- (9) Byard, R. W.; Maxwell-stewart, H. Scurvy - Characteristic Features and Forensic Issues. **2019**, *40* (1), 43–46.
- (10) Pina, A. S.; Hussain, A.; Roque, A. C. A. An Historical Overview of Drug Discovery. In *Ligand-Macromolecular Interactions in Drug Discovery*; Ana Cecília A. Roque, Ed.; Humana Press: Totowa, NJ, 2010; pp 3–12.
- (11) Drews, J. Drug Discovery : A Historical Perspective. *Science (80-.)*. **2000**, *287* (5460), 1960–1965.
- (12) Sertürner, F. W. A. Ueber Das Morphinum, Eine Neue Salzfähige Grundlage, Und Die Mekonsäure, Als Hauptbestandtheile Des Opiums. *Ann. d. Phys.* **1817**, *55*, 56–

89.

- (13) Fleming, A. On the Antibacterial Action of Cultures of a *Penicillium*, with Special Reference to Their Use in the Isolation of *B. Influenzae*. *Br. J. Exp. Pathol.* **1929**, *10* (3), 226–236.
- (14) Dash, C. H. Penicillin Allergy and the Cephalosporins. *J. Antimicrob. Chemother.* **1975**, *1* (suppl 3), 107–118.
- (15) Bajorath, J. Integration of Virtual and High-Throughput Screening. *Nature* **2002**, *1* (11), 882–894.
- (16) Khaled, H. G.; Feng, H.; Hu, X.; Sun, X.; Zheng, W.; Li, P. P.; Rudnicki, D. D.; Ye, W.; Chen, Y. C.; Southall, N.; et al. A High-throughput Screening to Identify Small Molecules That Suppress Huntingtin Promoter Activity or Activate Huntingtin- Antisense Promoter Activity. *Nat. Sci. Reports* **2021**, *11* (6157).
- (17) Flobak, Å.; Niederdorfer, B.; Nakstad, V. T.; Thommesen, L. A High-Throughput Drug Combination Screen of Targeted Small Molecule Inhibitors in Cancer Cell Lines. *Nat. Sci. Data* **2019**, *6* (237), 1–10.
- (18) Masuda, N.; Ohe, Y.; Yamada, I.; Ishii, T. Safety and Effectiveness of Alectinib in a Real- - World Surveillance Study in Patients with ALK-- Positive Non-Small-Cell Lung Cancer in Japan. *Cancer Sci.* **2019**, *110* (February), 1401–1407.
- (19) Coussens, N. P.; Braisted, J. C.; Peryea, T.; Sittampalam, G. S.; Simeonov, A.; Hall, M. D. Small-Molecule Screens : A Gateway to Cancer Therapeutic Agents with Case Studies of Food and Drug Administration – Approved Drugs. *Pharmacol. Rev.* **2017**, *69* (October), 479–496.
- (20) Menear, K. A.; Adcock, C.; Boulter, R.; Cockcroft, X.; Copsey, L.; Cranston, A.; Dillon, K. J.; Drzewiecki, J.; Garman, S.; Gomez, S.; et al. 4-[3-(4-Cyclopropanecarbonylpiperazine-1-Carbonyl)-4-Fluorobenzyl]-2 H-Phthalazin-1-One : A Novel Bioavailable Inhibitor of Poly (ADP-Ribose) Polymerase-1. *J. Med. Chem.* **2008**, *51*, 6581–6591.
- (21) Abel, S.; Ryst, E. Van Der; Rosario, M. C.; Ridgway, C. E.; Medhurst, C. G.; Taylor-worth, R. J.; Muirhead, G. J. Assessment of the Pharmacokinetics, Safety and Tolerability of Maraviroc, a Novel CCR5 Antagonist, in Healthy Volunteers. *Br. J. Clin. Pharmacol.* **2008**, *65* (Suppl. 1), 5–18.
- (22) Kurata, H.; Kusumi, K.; Otsuki, K.; Suzuki, R.; Kurono, M.; Komiya, T.; Hagiya, H.; Mizuno, H.; Shioya, H.; Ono, T.; et al. Discovery of a 1-Methyl-3,4-Dihydronaphthalene-Based Sphingosine1-Phosphate (S1P) Receptor Agonist

- Ceralifimod (ONO-4641). A S1P1 and S1P5 Selective Agonist for the Treatment of Autoimmune Diseases. *J. Med. Chem.* **2017**, *60*, 9508–9530.
- (23) Wan, W.; Zhu, S.; Li, S.; Shang, W.; Zhang, R.; Li, H.; Liu, W.; Xiao, G.; Peng, K.; Zhang, L. High-Throughput Screening of an FDA-Approved Drug Library Identifies Inhibitors against Arenaviruses and SARS-CoV-2. *ACS Infect. Dis.* **2020**, No. Antiviral Therapies Special Issue.
- (24) Joy, S.; Vijayakumar, Y. M.; Sunhye, G.; Choi, S. Role of Computer-Aided Drug Design in Modern Drug Discovery. *Arch. Pharm. Res.* **2015**, *38* (9), 1686–1701.
- (25) ChemAxon. Calculator Plugins Were Used for Structure Property Prediction and Calculation. 2019, p Calculator, Version 19.4.0, ChemAxon.
- (26) Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and Scoring in Virtual Screening for Drug Discovery: Methods and Applications. *Nat. Rev. Drug Discov.* **2004**, *3* (11), 935–949.
- (27) Prieto-Martínez, F. D.; López-López, E.; Juárez-Mercado, K. E.; Medina-Franco, J. L. Computational Drug Design Methods—Current and Future Perspectives. In *In silico Drug Design*; 2019; pp 19–44.
- (28) Medina-Franco, J. L.; Giulianotti, M. A.; Welmaker, G. S.; Houghten, R. A. Shifting from the Single to the Multitarget Paradigm in Drug Discovery. *Drug Discov. Today* **2013**, *18* (9–10), 495–501.
- (29) Sliwoski, G.; Kothiwale, S.; Meiler, J.; Lowe, E. W. Computational Methods in Drug Discovery. *Pharmacol. Rev.* **2014**, *66* (1), 334–395.
- (30) Jorgensen, W. L. Computer-Aided Discovery of Anti-HIV Agents. *Bioorg. Med. Chem.* **2016**, *24* (20), 4768–4778.
- (31) Yu, W.; MacKerell Jr., A. D. Methods in Computer-Aided Drug Design. *Methods Mol. Biol.* **2017**, *1520*, 85–106.
- (32) Rang, H. P.; Dale, M. M.; Flower, R. J.; Henderson, G. *Rang & Dale's Pharmacology*, Seventh.; Elsevier Churchill Livingstone: St. Louis, MO, 2011.
- (33) Hair, P.; Cameron, F.; Mckeage, K. Mipomersen Sodium: First Global Approval. *Drugs* **2013**, *73* (5), 487–493.
- (34) Wishart, D. S.; Feunang, Y. D.; Guo, A. C.; Lo, E. J.; Marcu, A.; Grant, R.; Sajed, T.; Johnson, D.; Li, C.; Sayeeda, Z.; et al. DrugBank 5.0: A Major Update to the DrugBank Database for 2018. *Nucleic Acids Res.* **2018**, *46*, 1074–1082.
- (35) Avram, S.; Bologna, C. G.; Holmes, J.; Bocci, G.; Wilson, B.; Nguyen, D.; Curpan, R.; Halip, L.; Bora, A.; Yang, J.; et al. DrugCentral 2021 Supports Drug Discovery

- and Repositioning. *Nucleic Acids Res.* **2021**, *49*, 1160–1169.
- (36) Bento, A. P.; Gaulton, A.; Hersey, A.; Bellis, L. J.; Chambers, J.; Davies, M.; Krüger, F. A.; Light, Y.; Mak, L.; McGlinchey, S.; et al. The ChEMBL Bioactivity Database: An Update. *Nucleic Acids Res.* **2014**, *42* (Database issue), D1083-1090.
- (37) Kanehisa, M.; Goto, S.; Sato, Y.; Furumichi, M.; Tanabe, M. KEGG for Integration and Interpretation of Large-Scale Molecular Data Sets. *Nucleic Acids Res.* **2012**, *40* (Database issue), D109-114.
- (38) Ursu, O.; Rayan, A.; Goldblum, A.; Oprea, T. Understanding Drug-likeness. *WIREs Comput. Mol. Sci.* **2011**, *1* (5), 760–781.
- (39) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Deliv. Rev.* **1997**, *23*, 3–25.
- (40) Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J. A Knowledge-Based Approach in Designing Combinatorial or Medicinal Chemistry Libraries for Drug Discovery . 1 . A Qualitative and Quantitative Characterization of Known Drug Databases. *J. Comb. Chem.* **1998**, *1*, 55–68.
- (41) Veber, D. F.; Johnson, S. R.; Cheng, H.; Smith, B. R.; Ward, K. W.; Kopple, K. D. Molecular Properties That Influence the Oral Bioavailability of Drug Candidates. *J. Med. Chem.* **2002**, *45*, 2615–2623.
- (42) Egan, W. J.; Merz, K. M.; Baldwin, J. J. Prediction of Drug Absorption Using Multivariate Statistics. *J. Med. Chem.* **2000**, *43*, 3867–3877.
- (43) Muegge, I.; Heald, S. L.; Brittelli, D. Simple Selection Criteria for Drug-like Chemical Matter. *J. Med. Chem.* **2001**, *44* (12), 1841–1846.
- (44) Bickerton, G. R.; Paolini, G. V.; Besnard, J.; Muresan, S.; Andrew, L. Quantifying the Chemical Beauty of Drugs. *Nat. Chem.* **2012**, *4* (2), 90–98.
- (45) Mochizuki, M.; Suzuki, S. D.; Yanagisawa, K.; Akiyama, Y. QEX: Target-Specific Druglikeness Filter Enhances Ligand-Based Virtual Screening. *Mol. Divers.* **2019**, *23* (1), 11–18.
- (46) Lovering, F.; Bikker, J.; Humblet, C. Escape from Flatland : Increasing Saturation as an Approach to Improving Clinical Success. *J. Med. Chem.* **2009**, *52*, 6752–6756.
- (47) Lovering, F. Escape from Flatland 2: Complexity and Promiscuity. *Med. Chem. Commun.* **2013**, *4*, 515–519.
- (48) Wei, W.; Cherukupalli, S.; Jing, L.; Liu, X.; Zhan, P. Fsp3: A New Parameter for

- Drug-Likeness. *Drug Discov. Today* **2020**, 25 (10), 1839–1845.
- (49) Für, C. S.; Bölcskei, H. New Spiro[Cycloalkane-Pyridazinone] Derivatives with Favorable Fsp³ Character. *Chemistry (Easton)*. **2020**, 2, 837–848.
- (50) Congreve, M.; Carr, R.; Murray, C.; Jhoti, H. A ‘Rule of Three’ for Fragment-Based Lead Discovery? *Drug Discov. Today* **2003**, 8 (19), 876–877.
- (51) Teague, S. J.; Davis, A. M.; Leeson, P. D.; Oprea, T. The Design of Leadlike Combinatorial Libraries. *Angew. Chemie Int. Ed.* **1999**, 38 (24), 3743–3748.
- (52) Zhu, T.; Cao, S.; Su, P.-C.; Patel, R.; Shah, D.; B. Chokshi, H.; Szukala, R.; Johnson, M. E.; Hevener, K. E. Hit Identification and Optimization in Virtual Screening: Practical Recommendations Based Upon a Critical Literature Analysis. *J. Med. Chem.* **2013**, 56 (17), 6560–6572.
- (53) Brenk, R.; Schipani, A.; James, D.; Krasowski, A.; Gilbert, I. H.; Frearson, J.; Graham, P. Lessons Learnt from Assembling Screening Libraries for Drug Discovery for Neglected Diseases. *ChemMedChem* **2008**, 3 (3), 435–444.
- (54) Daina, A.; Michielin, O.; Zoete, V. SwissADME: A Free Web Tool to Evaluate Pharmacokinetics, Drug-Likeness and Medicinal Chemistry Friendliness of Small Molecules. *Nat. Sci. Reports* **2017**, No. 7, 42717.
- (55) Molinspiration Cheminformatics Free Web Services, <https://www.molinspiration.com>, Slovensky Grob, Slovakia.
- (56) <https://www.click2drug.org/>.
- (57) Katsila, T.; Spyroulias, G. A.; Patrinos, G. P.; Matsoukas, M. T. Computational Approaches in Target Identification and Drug Discovery. *Comput. Struct. Biotechnol. J.* **2016**, 14, 177–184.
- (58) Hughes, J. P.; Rees, S.; Kalindjian, S. B.; Philpott, K. L. Principles of Early Drug Discovery. *Br. J. Pharmacol.* **2011**, 162, 1239–1249.
- (59) Mohs, R. C.; Greig, N. H. Drug Discovery and Development: Role of Basic Biological Research. *Alzheimer’s Dement. Transl. Res. Clin. Interv.* **2017**, 3, 651–657.
- (60) Linton-Reid, K. Introduction: An Overview of AI in Oncology Drug Discovery and Development. In *Artificial Intelligence in Oncology Drug Discovery and Development*; Taylor, J. W. C., Taylor, B., Eds.; IntechOpen, 2020.
- (61) Yan, X. C.; Sanders, J. M.; Gao, Y.; Tudor, M.; Haidle, A. M.; Klein, D. J.; Converso, A.; Lesburg, C. A.; Zang, Y.; Wood, H. B. Augmenting Hit Identification by Virtual Screening Techniques in Small Molecule Drug

- Discovery. *J. Chem. Inf. Model.* **2020**, *60* (9), 4144–4152.
- (62) Issa, N. T.; Wathieu, H.; Ojo, A.; Byers, S. W.; Dakshanamurthy, S. Drug Metabolism in Preclinical Drug Development: A Survey of the Discovery Process, Toxicology, and Computational Tools. *Curr. Drug Metab.* **2018**, *18* (6), 556–565.
- (63) Jr, L. G. V. In Silico Toxicology for the Pharmaceutical Sciences ☆. *Toxicol. Appl. Pharmacol.* **2009**, *241* (3), 356–370.
- (64) Polson, A. G.; Fuji, R. N. The Successes and Limitations of Preclinical Studies in Predicting the Pharmacodynamics and Safety of Cell-Surface-Targeted Biological Agents in Patients. *Br. J. Pharmacol.* **2012**, *166*, 1600–1602.
- (65) Langhof, H.; Wei, W.; Chin, L.; Wieschowski, S.; Federico, C.; Kimmelman, J.; Strech, D. Preclinical Efficacy in Therapeutic Area Guidelines from the U. S. Food and Drug Administration and the European Medicines Agency: A Cross-Sectional Study. *Br. J. Pharmacol.* **2018**, *175*, 4229–4238.
- (66) Benfenati, E.; Manganaro, A.; Gini, G. VEGA-QSAR: AI inside a Platform for Predictive Toxicology. *CEUR Workshop Proc.* **2013**, *1107*, 21–28.
- (67) Yang, H.; Sun, L.; Li, W.; Liu, G.; Tang, Y. In Silico Prediction of Chemical Toxicity for Drug Design Using Machine Learning Methods and Structural Alerts. *Front. Chem.* **2018**, *6* (30).
- (68) Umscheid, C.; Margolis, D.; Grossman, C. Key Concepts of Clinical Trials: A Narrative Review. *Postgrad. Med.* **2012**, *123* (5), 194–204.
- (69) Inglese, J.; Auld, D. S.; Jadhav, A.; Johnson, R. L.; Simeonov, A.; Yasgar, A.; Zheng, W.; Austin, C. P. Quantitative High-Throughput Screening : A Titration-Based Approach That Efficiently Identifies Biological Activities in Large Chemical Libraries. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103* (31), 11473–11478.
- (70) Gadagkar, S. R.; Call, G. B. Computational Tools for Fitting the Hill Equation to Dose-Response Curves. *J. Pharmacol. Toxicol. Methods* **2015**, *71*, 68–76.
- (71) Hill, A. V. The Possible Effects of the Aggregation of the Molecules of Haemoglobin on Its Dissociation Curves. *J. Physiol.* **1910**, *40*, iv–vii.
- (72) Silverstein, T. P. CoViD-19 Epidemic Follows the “Kinetics” of Enzymes with Cooperative Substrate Binding. *Biochem. Mol. Biol. Educ.* **2020**, *48* (5), 452–459.
- (73) Motulsky, H. J.; Mahan, L. C. The Kinetics of Competitive Radioligand Binding Predicted by the Law of Mass Action. *Mol. Pharmacol.* **1984**, *25* (1), 1–9.
- (74) Neubig, R. R.; Spedding, M.; Kenakin, T.; Christopoulos, A. International Union of Pharmacology Committee on Receptor Nomenclature and Drug Classification.

- XXXVIII. Update on Terms and Symbols in Quantitative Pharmacology. *Pharmacol. Rev.* **2003**, *55* (4), 597–606.
- (75) Cheng, Y.-C.; Prusoff, W. Relationship between the Inhibition Constant and the Concentration of Inhibitor Which Causes 50 per Cent Inhibition of an Enzymatic Reaction. *Biochem. Pharmacol.* **1973**, *22*, 3099–3108.
- (76) Craig, D. A. The Cheng-Prusoff Relationship: Something Lost in the Translation. *Trends Pharmacol. Sci.* **1993**, *13* (3), 89–91.
- (77) Cha, S. Tight-Binding Inhibitors-I: Kinetic Behavior. *Biochem. Pharmacol.* **1975**, *24*, 2177–2185.
- (78) Zhou, H.; Gilson, M. K. Theory of Free Energy and Entropy in Noncovalent Binding. *Chem. Rev.* **2009**, *109*, 4092–4107.
- (79) Caro, J. A.; Harpole, K. W.; Kasinath, V.; Lim, J.; Granja, J.; Valentine, K. G. Entropy in Molecular Recognition by Proteins. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114* (25), 6563–6568.
- (80) Bonifacino, J. S.; Dasso, M.; Harford, J. B.; Lippincott-Schwartz, J.; Yamada, K. M. Isothermal Titration Calorimetry. *Curr. Protoc. Cell Biol.* **2004**, 17.8.1–17.8.24.
- (81) Archer, W. R.; Schulz, M. D. Isothermal Titration Calorimetry: Practical Approaches and Current Applications in Soft Matter. *Soft Matter* **2020**, *38*, 8760–8774.
- (82) Trani, J. M. Di; Cesco, S. De; Leary, R. O.; Plescia, J.; Jorge, C.; Moitessier, N.; Mittermaier, A. K. Rapid Measurement of Inhibitor Binding Kinetics by Isothermal Titration Calorimetry. *Nat. Commun.* **2018**, *9* (893).
- (83) Song, C.; Zhang, S.; Huang, H. Choosing a Suitable Method for the Identification of Replication Origins in Microbial Genomes. *Front. Microbiol.* **2015**, *6* (1049).
- (84) Caflisch, A.; Sledz, P. Protein Structure-Based Drug Design: From Docking to Molecular Dynamics. *Curr. Opin. Struct. Biol.* **2018**, *48*, 93–102.
- (85) Brown, F. K.; Sherer, E. C.; Johnson, S. A.; Holloway, M. K.; Sherborne, B. S. The Evolution of Drug Design at Merck Research Laboratories. *J. Comput. Aided. Mol. Des.* **2016**, *31* (3), 255–266.
- (86) Tropsha, A.; Zheng, W. Computer Aided Drug Design. In *Computational Biochemistry and Biophysics*; Becker, O., MacKerell, A., Roux, B., Watanabe, M., Eds.; Marcel Dekker Inc: New York, 2001; pp 351–369.
- (87) Jorgensen, W. L. The Many Roles of Computation in Drug Discovery. *Science* (80-

- .). **2004**, *303* (19), 1813–1818.
- (88) Doman, T. N.; MCGovern, S. L.; Witherbee, B. J.; Kasten, T. P.; Kurumbail, R.; Stallings, W. C.; Connolly, D. T.; Shoichet, B. K. Molecular Docking and High-Throughput Screening for Novel Inhibitors of Protein Tyrosine Phosphatase-1B. *J. Med. Chem.* **2002**, *45*, 2213–2221.
- (89) Greenidge, P. A.; Kramer, C.; Sherman, W. Improving Docking Results via Reranking of Ensembles of Ligand Poses in Multiple X - Ray Protein Conformations with MM-GBSA. *J. Chem. Inf. Model.* **2014**, *54* (10), 2697–2717.
- (90) Fischer, A.; Smiesko, M.; Sellner, M.; Lill, M. A. Decision Making in Structure-Based Drug Discovery: Visual Inspection of Docking Results. *J. Med. Chem.* **2021**, *64*, 2489–2500.
- (91) Sastry, G. M.; Adzhigirey, M.; Sherman, W. Protein and Ligand Preparation : Parameters , Protocols , and Influence on Virtual Screening Enrichments. *J. Comput. Aided. Mol. Des.* **2013**, *27*, 221–234.
- (92) Daniel, P. I. O.; Peng, Z.; Pi, H.; Testero, S. A.; Ding, D.; Spink, E.; Leemans, E.; Boudreau, M. A.; Yamaguchi, T.; Schroeder, V. A.; et al. Discovery of a New Class of Non- β -Lactam Inhibitors of Penicillin- Binding Proteins with Gram-Positive Antibacterial Activity. *J. Am. Chem. Soc.* **2014**, *136*, 3664–3672.
- (93) Sterling, T.; Irwin, J. J. ZINC 15 – Ligand Discovery for Everyone. *J. Chem. Inf. Model.* **2015**, *55*, 2324–2337.
- (94) Cushman, D. W.; Cheung, H. S.; Sabo, E. F.; Ondetti, M. A. Design of Potent Competitive Inhibitors of Angiotensin-Converting Enzyme. Carboxyalkanoyl and Mercaptoalkanoyl Amino Acids? *Biochemistry* **1977**, *16* (25), 5484–5491.
- (95) The Nobel Prize in Chemistry 1987. NobelPrize.Org. Nobel Media AB 2021. Wed. 16 Jun 2021. <<https://www.nobelprize.org/prizes/chemistry/1987/summary/>>.
- (96) Persch, E.; Dumele, O.; Diederich, F. Molecular Recognition in Chemical and Biological Systems. *Angew. Chemie Int. Ed.* **2015**, *54* (11), 3290–3327.
- (97) Yan, C.; Zou, X.; States, U. *Modeling Protein Flexibility in Molecular Docking*, Third Edition.; Elsevier, 2017; Vol. 3.
- (98) Lamb, M.; Jorgensen, W. Computational Approaches to Molecular Recognition. *Curr. Opin. Chem. Biol.* **1997**, *1* (4), 449–457.
- (99) Burley, S. K.; Bhikadiya, C.; Bi, C.; Bittrich, S.; Chen, L.; Crichlow, G. V.; Christie, C. H.; Dalenberg, K.; Costanzo, L. Di; Duarte, J. M.; et al. RCSB Protein Data Bank: Powerful New Tools for Exploring 3D Structures of Biological

- Macromolecules for Basic and Applied Research and Education in Fundamental Biology, Biomedicine, Biotechnology, Bioengineering and Energy Sciences. *Nucleic Acids Res.* **2021**, *49* (November 2020), 437–451.
- (100) Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C. The Cambridge Structural Database. *Acta Crystallogr.* **2016**, *B72*, 171–179.
- (101) Suárez, D., Díaz, N. Affinity Calculations of Cyclodextrin Host-Guest Complexes: Assessment of Strengths and Weaknesses of End-Point Free Energy Methods. *J. Chem. Inf. Model.* **2018**, *59* (1), 421–440.
- (102) Sikder, T.; Rahman, M.; Hosokawa, T.; Kurasaki, M.; Saito, T. Remediation of Water Pollution with Native Cyclodextrins and Modified Cyclodextrins: A Comparative Overview and Perspectives. *Chem. Eng. J.* **2019**, *355*, 920–941.
- (103) Falconer, I. Charles Augustin Coulomb and the Fundamental Law of Electrostatics. *Metrologia* **2004**, *41*, S107-114.
- (104) Raevsky, O. A.; Skvortsov, V. S. Quantifying Hydrogen Bonding in QSAR and Molecular Modeling. *SAR QSAR Environ. Res.* **2007**, *16* (3), 287–300.
- (105) Liptrot, D. J.; Power, P. P. London Dispersion Forces in Sterically Crowded Inorganic and Organometallic Molecules. *Nat. Rev. Chem.* **2017**, *1* (0004).
- (106) Meyer, E. A.; Castellano, R. K.; Diederich, F. Interactions with Arenes Interactions with Aromatic Rings in Chemical and Biological Recognition. *Angew. Chemie Int. Ed.* **2003**, *42* (11), 1210–1250.
- (107) Kollman, P. Non-Covalent Forces of Importance in Biochemistry. In *The Chemistry of Enzyme Action*; Elsevier, 1984; pp 55–71.
- (108) Chandler, D. Interfaces and the Driving Force of Hydrophobic Assembly. *Nature* **2005**, *47*, 640–647.
- (109) Batool, M.; Ahmad, B.; Choi, S. A Structure-Based Drug Discovery Paradigm. *Int. J. Mol. Sci.* **2019**, *20* (11), 2783.
- (110) Yang, T.; Wu, J. C.; Yan, C.; Wang, Y.; Luo, R.; Gonzales, M. B.; Dalby, K. N.; Ren, P. Virtual Screening Using Molecular Simulations. *Proteins Struct. Funct. Bioinforma.* **2011**, *79* (6), 1940–1951.
- (111) Lionta, E.; Spyrou, G.; Vassilatis, D. K.; Cournia, Z. Structure-Based Virtual Screening for Drug Discovery: Principles, Applications and Recent Advances. *Curr. Top. Med. Chem.* **2014**, *14*, 1923–1938.
- (112) Ferreira, L. G.; Santos, R. N.; Oliva, G.; Andricopulo, A. D. Molecular Docking and Structure-Based Drug Design Strategies. *Molecules* **2015**, *200*, 13384–13421.

- (113) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A Fast Flexible Docking Method Using an Incremental Construction Algorithm. *J. Mol. Biol.* **1996**, *261* (3), 470–489.
- (114) Warren, G. L.; Andrews, C. W.; Capelli, A.; Clarke, B.; Lalonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; et al. A Critical Assessment of Docking Programs and Scoring Functions. *J. Med. Chem.* **2006**, *49* (20), 5912–5931.
- (115) Genheden, S.; Ryde, U. The MM/PBSA and MM/GBSA Methods to Estimate Ligand-Binding Affinities. *Expert Opin. Drug Discov.* **2015**, *10* (5), 449–461.
- (116) Kim, M. O.; Blachly, P. G.; McCammon, J. A. Conformational Dynamics and Binding Free Energies of Inhibitors of BACE-1: From the Perspective of Protonation Equilibria. *PLOS Comput. Biol.* **2015**, *11* (10), e1004341.
- (117) Guzman-Ocampo, D. C.; Aguayo-Ortiz, R.; Cano-González, L.; Castillo, R.; Hernández-Campos, A.; Dominguez, L. Effects of the Protonation State of Titratable Residues and the Presence of Water Molecules on Nocodazole Binding to β -Tubulin. *ChemMedChem* **2018**, *13*, 20–24.
- (118) D.A. Case, I.Y. Ben-Shalom, S.R. Brozell, D.S. Cerutti, T.E. Cheatham, III, V.W.D. Cruzeiro, T.A. Darden, R.E. Duke, D. Ghoreishi, M.K. Gilson, H. Gohlke, A.W. Goetz, D. Greene, R. Harris, N. Homeyer, S. Izadi, A. Kovalenko, T. Kurtzman, T.S. Lee, S. LeGra, D. M. Y. and P. A. K. Amber 2018. *Univ. California, San Fr.* **2018**.
- (119) Anandakrishnan, R.; Aguilar, B.; Onufriev, A. V. H++ 3.0: Automating PK Prediction and the Preparation of Biomolecular Structures for Atomistic Molecular Modeling and Simulations. *Nucleic Acids Res.* **2012**, *40*, 537–541.
- (120) ten Brink, T.; Exener, T. E. PKa Based Protonation States and Microspecies for Protein – Ligand Docking. *J. Comput.* **2010**, *24*, 935–942.
- (121) Olsson, M. H. M.; Søndergaard, C. R.; Rostkowski, M.; Jensen, J. H. PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical PKa Predictions. *J. Chem. Theory Comput.* **2011**, *7*, 525–537.
- (122) Heinzelmann, G.; Kuyucak, S. Molecular Dynamics Simulations Elucidate the Mechanism of Proton Transport in the Glutamate Transporter EAAT3. *Biophys. J.* **2014**, *106* (12), 2675–2683.
- (123) Razavi, A. M.; Delemotte, L.; Berlin, J. R.; Carnevale, V.; Voelz, V. A. Molecular Simulations and Free-Energy Calculations Suggest Conformation-Dependent

- Anion Binding to a Cytoplasmic Site as a Mechanism for Na⁺ / K⁺ -ATPase Ion Selectivity. *J. Biol. Chem.* **2017**, 292 (30), 12412–12423.
- (124) Dobrev, P.; Phani, S.; Vemulapalli, B.; Nath, N.; Griesinger, C.; Grubmuller, H. Probing the Accuracy of Explicit Solvent Constant PH Molecular Dynamics Simulations for Peptides. *J. Chem. Theory Comput.* **2020**, 16, 2561–2569.
- (125) Po, H. N.; Senozan, N. M. The Henderson–Hasselbalch Equation: Its History and Limitations. *J. Chem. Educ.* **2001**, 78 (11), 1499–1503.
- (126) Søndergaard, C. R.; Olsson, M. H. M.; Rostkowski, M.; Jensen, J. H. Improved Treatment of Ligands and Coupling Effects in Empirical Calculation and Rationalization of p K_a Values. *J. Chem. Theory Comput.* **2011**, 7, 2284–2295.
- (127) Rabenstein, B.; Knapp, E. Calculated PH-Dependent Population and Protonation of Carbon-Monooxy-Myoglobin Conformers. *Biophys. J.* **2001**, 80, 1141–1150.
- (128) Driessche, G. Van Den; Fourches, D. Adverse Drug Reactions Triggered by the Common HLA-B*57: 01 Variant: A Molecular Docking Study. *J. Cheminform.* **2017**, 9 (13), 1–17.
- (129) Shelley, J. C.; Cholleti, A.; Frye, L. L.; Greenwood, J. R.; Timlin, M. R.; Uchimaya, M. Epik: A Software Program for PK_a Prediction and Protonation State Generation for Drug-like Molecules. *J. Comput. Aided. Mol. Des.* **2007**, 21, 681–691.
- (130) Williams, C. J.; Headd, J. J.; Moriarty, N. W.; Prisant, M. G.; Videau, L. L.; Deis, L. N.; Verma, V.; Keedy, D. A.; Hintze, B. J.; Chen, V. B.; et al. MolProbity: More and Better Reference Data for Improved All-Atom Structure Validation. *Protein Sci.* **2017**, 27, 293–315.
- (131) De Vita, S.; Lauro, G.; Ruggiero, D.; Terracciaano, S.; Riccio, R.; Bifulco, G. Protein Preparation Automatic Protocol for High-Throughput Inverse Virtual Screening: Accelerating the Target Identification by Computational Methods. *J. Chem. Inf. Model.* **2019**, 59 (11), 4678–4690.
- (132) Dolinsky, T. J.; Czodrowski, P.; Li, H.; Nielsen, J. E.; Jensen, J. H.; Klebe, G.; Baker, N. A. PDB2PQR: Expanding and Upgrading Automated Preparation of Biomolecular Structures for Molecular Simulations. *Nucleic Acids Res.* **2007**, 35, 522–525.
- (133) Becker, O. Conformational Analysis. In *Computational Biochemistry and Biophysics*; Becker, O., MacKerell Jr., A. D., Roux, B., Watanabe, M., Eds.; New York, 2001; pp 69–90.

- (134) Webb, B.; Sali, A. Comparative Protein Structure Modeling Using MODELLER. *Curr. Protoc. Bioinforma.* **2017**, *54* (5.6), 1–37.
- (135) Caldararu, O.; Ignjatovic, M. M.; Oksanen, E.; Ryde, U. Water Structure in Solution and Crystal Molecular Dynamics Simulations Compared to Protein Crystal Structures. *RSC Adv.* **2020**, *10*, 8435–8443.
- (136) Hu, X.; Maffucci, I.; Contini, A. Advances in the Treatment of Explicit Water Molecules in Docking and Binding Free Energy Calculations. *Curr. Med. Chem.* **2019**, *26* (42), 7598–7622.
- (137) Raymer, M. L.; Sanschagrin, P. C.; Punch, W. F.; Venkataraman, S.; Goodman, E. D.; Kuhn, L. A. Predicting Conserved Water-Mediated and Polar Ligand Interactions in Proteins Using a K-Nearest-Neighbors Genetic Algorithm. *J. Mol. Biol.* **1997**, *265*, 445–464.
- (138) García-Sosa, A.; Mancera, R.; Dean, P. WaterScore: A Novel Method for Distinguishing between Bound and Displaceable Water Molecules in the Crystal Structure of the Binding Site of Protein-Ligand Complexes. *J. Mol. Model.* **2003**, *9*, 172–182.
- (139) Young, T.; Abel, R.; Kim, B.; Berne, B. J.; Friesner, R. A. Motifs for Molecular Recognition Exploiting Hydrophobic Enclosure in Protein – Ligand Binding. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104* (3), 808–813.
- (140) Benet, L. Z.; Broccatelli, F.; Oprea, T. I. BDDCS Applied to Over 900 Drugs. *AAPS J.* **2011**, *13* (4), 519–547.
- (141) Seebeck, B.; Reulecke, I.; Kämper, A.; Rarey, M. Modeling of Metal Interaction Geometries for Protein – Ligand Docking. *Proteins Struct. Funct. Bioinforma.* **2008**, *71* (3), 1237–1254.
- (142) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; et al. Glide: A New Approach for Rapid , Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* **2004**, *47*, 1739–1749.
- (143) Huang, N.; Shoichet, B. K.; Irwin, J. J.; Francisco, S. Benchmarking Sets for Molecular Docking. *J. Med. Chem.* **2006**, *49*, 6789–6801.
- (144) Anderson, A. C. The Process of Structure-Based Drug Design. *Chem. Biol.* **2003**, *10*, 787–797.
- (145) Guilloux, V. Le; Schmidtke, P.; Tuffery, P. Fpocket: An Open Source Platform for Ligand Pocket Detection. *BMC Bioinformatics* **2009**, *10* (168).

- (146) Schmidtke, P.; Bidon-Chanal, A.; Luque, F. J.; Barril, X. MDpocket: Open-Source Cavity Detection and Characterization on Molecular Dynamics Trajectories. *Bioinformatics* **2011**, *27* (23), 3276–3285.
- (147) Halgren, T. A. Identifying and Characterizing Binding Sites and Assessing Druggability. *J. Chem. Inf. Model.* **2009**, *49*, 377–389.
- (148) Ghanakota, P.; Carlson, H. A. Moving Beyond Active-Site Detection: MixMD Applied to Allosteric Systems. *J. Phys. Chem. B* **2016**, *120* (33), 8685–8695.
- (149) Faller, C. E.; Raman, E. P.; MacKerell, A. D.; Guvench, O. Site Identification by Ligand Competitive Saturation (SILCS) Simulations for Fragment-Based Drug Design. *Methods Mol. Biol.* **2016**, *1289*, 75–87.
- (150) Yu, W.; Lakkaraju, S.; Raman, E. P.; MacKerell, A. D. Site-Identification by Ligand Competitive Saturation (SILCS) Assisted Pharmacophore Modeling. *J. Comput. Aided. Mol. Des.* **2015**, *28* (5), 491–507.
- (151) Eguida, M.; Rognan, D. A Computer Vision Approach to Align and Compare Protein Cavities: Application to Fragment-Based Drug Design. A Computer Vision Approach to Align and Compare Protein Cavities: Application to Fragment-Based Drug Design. *J. Med. Chem.* **2020**, *63* (13), 7127–7142.
- (152) Reymond, J.; Deursen, R. Van; Blum, L. C.; Ruddigkeit, L. Chemical Space as a Source for New Drugs. *Med. Chem. Commun.* **2010**, *1*, 30–38.
- (153) Reymond, J. The Chemical Space Project. *Acc. Chem. Res.* **2015**, *48* (3), 722–730.
- (154) Kireeva, N.; Baskin, I. I.; Gaspar, H. A.; Horvath, D.; Marcou, G.; Varnek, A. Generative Topographic Mapping (GTM): Universal Tool for Data Visualization, Structure-Activity Modeling and Dataset Comparison. *Mol. Inform.* **2012**, *31* (3–4), 301–312.
- (155) Medina-Franco, J.; Martinez-Mayorga, K.; Giulianotti, M.; Houghten, R.; Pinilla, C. Visualization of the Chemical Space in Drug Discovery. *Curr. Comput. Aided-Drug Des.* **2008**, *4* (4), 322–333.
- (156) Baell, J. B.; M. Nissink, J. W. Seven Year Itch: Pan-Assay Interference Compounds (PAINS) in 2017 - Utility and Limitations. *ACS Chem. Biol.* **2018**, *13*, 35–44.
- (157) Baell, J. B.; Holloway, G. A. New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays. *J. Med. Chem.* **2010**, *53*, 2719–2740.
- (158) Lagorce, D.; Sperandio, O.; Baell, J. B.; Miteva, M. A.; Villoutreix, B. FAF-

- Drugs3 : A Web Server for Compound Property Calculation and Chemical Library Design. *Nucleic Acids Res.* **2015**, *43*, W200–W207.
- (159) Cournia, Z.; Allen, B.; Sherman, W. Relative Binding Free Energy Calculations in Drug Discovery: Recent Advances and Practical Considerations. *J. Chem. Inf. Model.* **2017**, *57* (12), 2911–2937.
- (160) Dearden, J. C. In Silico Prediction of ADMET Properties: How Far Have We Come? *Expert Opin. Drug Metab. Toxic.* **2007**, *3* (5), 635–639.
- (161) Dearden, J. C. In Silico Prediction of Drug Toxicity. *J. Comput. Aided. Mol. Des.* **2003**, *17*, 119–127.
- (162) Liebert, M. A. Predicting CNS Permeability of Drug Molecules. *J. Comput. Biol.* **2002**, *9* (6), 849–864.
- (163) van Breemen, R.; Li, Y. Caco-2 Cell Permeability Assays to Measure Drug Absorption. *Expert Opin. Drug Metab. Toxic.* **2005**, *1* (2), 175–185.
- (164) The, H. P.; Gonzalez-, I.; Bermejo, M.; Sanjuan, V. M.; Centelles, I.; Garrigues, T. M.; Cabrera-Perez, M. A. In Silico Prediction of Caco-2 Cell Permeability by a Classification QSAR. *Mol. Inform.* **2011**, *30*, 376–385.
- (165) The, H. P.; Cabrera-Pérez, M. A.; Nam, N.-H.; Garit, J. A. C.; Rasulev, B.; Huong, L.-T.-T.; Casañola-Martin, G. In Silico Assessment of ADME Properties: Advances in Caco-2 Cell Permeability Modeling. *Curr. Top. Med. Chem.* **2018**, *18* (26), 2209–2229.
- (166) Gally, J.-M.; Bourg, S.; Do, Q.; Aci-Sèche, S.; Bonnet, P. VSPrep: A General KNIME Workflow for the Preparation of Molecules for Virtual Screening. *Mol. Inform.* **2017**, *36* (1700023).
- (167) Chen, I.; Foloppe, N. Drug-like Bioactive Structures and Conformational Coverage with the LigPrep / ConfGen Suite: Comparison to Programs MOE and Catalyst. *J. Chem. Inf. Model.* **2010**, *50*, 822–839.
- (168) Sisquellas, M.; Cecchini, M. PrepFlow: A Toolkit for Chemical Library Preparation, Management, and Profiling. *Mol. Inform.* **2021**, *40* (2100139).
- (169) Gasteiger, J. Chemoinformatics: Achievements and Challenges, a Personal View. *Molecules* **2016**, *21* (2).
- (170) Jakalian, A.; Jack, D. B.; Bayly, C. I. Fast , Efficient Generation of High-Quality Atomic Charges. AM1-BCC Model: II. Parameterization and Validation. *J. Comput. Chem.* **2002**, *23* (16), 1623–1641.
- (171) Bayly, C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. A. A Well-Behaved

- Electrostatic Potential Based Method Using Charge Restraints for Deriving Atomic Charges: The RESP Model. *J. Phys. Chem.* **1993**, *97* (40), 10269–10280.
- (172) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenb, D. J. Gaussian09 Revision A02. 2009.
- (173) Price, D. J.; Brooks III, C. L. A Modified TIP3P Water Potential for Simulation with Ewald Summation A Modified TIP3P Water Potential for Simulation with Ewald Summation. *J. Chem. Phys.* **2004**, *121* (20), 10096–10103.
- (174) Allen, W. J.; Balius, T. E.; Mukherjee, S.; Brozell, S. R.; Moustakas, D. T.; Lang, P. T.; Case, D. A.; Kuntz, I. D.; Rizzo, R. C. DOCK6: Impact of New Features and Current Docking Performance. *J. Comput. Chem.* **2015**, *36*, 1132–1156.
- (175) Morris, G. M.; Huey, R.; Lindstrom, W.; Sanner, M. F.; Belew, R. K.; Goodsell, D. S.; Olson, A. J. AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility. *J. Comput. Chem.* **2009**, *30* (16), 2785–2791.
- (176) Liebeschuetz, J. W.; Cole, J. C.; Korb, O. Pose Prediction and Virtual Screening Performance of GOLD Scoring Functions in a Standardized Test. *J. Comput. Aided. Mol. Des.* **2012**, *26* (6), 737–748.
- (177) Korb, O.; Stüzle, T.; Exner, T. E. Empirical Scoring Functions for Advanced Protein - Ligand Docking with PLANTS. *J. Chem. Inf. Model.* **2009**, *49* (1), 84–96.
- (178) Kramer, B.; Rarey, M.; Lengauer, T. Evaluation of the FlexX Incremental Construction Algorithm for Protein – Ligand Docking. *Proteins Struct. Funct. Bioinforma.* **1999**, *37*, 228–241.
- (179) Cross, J. B.; Thompson, D. C.; Rai, B. K.; Baber, J. C.; Fan, K. Y. Comparison of Several Molecular Docking Programs : Pose Prediction and Virtual Screening Accuracy. *J. Chem. Inf. Model.* **2009**, *49*, 1455–1474.
- (180) Li, J.; Fu, A.; Zhang, L. An Overview of Scoring Functions Used for Protein – Ligand Interactions in Molecular Docking. *Interdiscip. Sci. Comput. Life Sci.* **2019**, *11*, 320–328.
- (181) Guedes, I. A.; Pereira, F. S. S.; Dardenne, L. E. Empirical Scoring Functions for Structure-Based Virtual Screening: Applications, Critical Aspects, and Challenges. *Front. Microbiol.* **2018**, *9* (1089), 1–18.
- (182) Salmaso, V.; Moro, S. Bridging Molecular Docking to Molecular Dynamics in

- Exploring Ligand-Protein Recognition Process : An Overview. *Front. Pharmacol.* **2018**, 9 (923), 1–16.
- (183) Fischer, E. Einfluss Der Configuration Auf Die Wirkung Der Enzyme. *Chem. Eur.* **1894**, 27 (3), 2985–2993.
- (184) Koshland, D. E. The Key-Lock Theory and the Induced Fit Theory. *Angew. Chemie Int. Ed.* **1994**, 33, 2375–2378.
- (185) Pagadala, N. S.; Syed, K.; Tuszynski, J. Software for Molecular Docking: A Review. *Biophys. Rev.* **2017**, 91–102.
- (186) Amaro, R. E.; Baudry, J.; Chodera, J.; Demir, Ö.; Mccammon, J. A.; Miao, Y.; Smith, J. C. Ensemble Docking in Drug Discovery. *Biophys. J.* **2018**, 114, 2271–2278.
- (187) Shen, Q.; Xiong, B.; Zheng, M.; Luo, X.; Luo, C.; Liu, X.; Du, Y.; Li, J.; Zhu, W.; Shen, J.; et al. Knowledge-Based Scoring Functions in Drug Design: 2. Can the Knowledge Base Be Enriched ? *J. Chem. Inf. Model.* **2011**, 51, 386–397.
- (188) Ain, Q. U.; Aleksandrova, A.; Roessler, F. D.; Ballester, P. J. Machine-Learning Scoring Functions to Improve Structure-Based Binding Affinity Prediction and Virtual Screening. *WIREs Comput. Mol. Sci.* **2015**, 5, 405–424.
- (189) Montalvo-Acosta, J. J.; Cecchini, M. Computational Approaches to the Chemical Equilibrium Constant in Protein-Ligand Binding. *Mol. Inform.* **2016**, 35 (11–12), 555–567.
- (190) Böhm, H. The Computer Program LUDI : A New Method for the de Novo Design of Enzyme Inhibitors. *J. Comput. Aided. Mol. Des.* **1991**, 6, 61–78.
- (191) Clark, M.; Cramer, R. D.; Opdenbosch, N. Van. Validation of the General Purpose Tripos 5.2 Force Field. *J. Comput. Chem.* **1989**, 10 (8), 982–1012.
- (192) Ewing, T. J. A.; Makino, S.; Skillman, A. G.; Kuntz, I. D. DOCK 4.0 : Search Strategies for Automated Molecular Docking of Flexible Molecule Databases. *J. Comput. Aided. Mol. Des.* **2001**, 15, 411–428.
- (193) Gohlke, H.; Hendlich, M.; Klebe, G. Knowledge-Based Scoring Function to Predict Protein-Ligand Interactions. *J. Mol. Biol.* **2000**, 295, 337–356.
- (194) Durrant, J. D.; McCammon, J. A. Molecular Dynamics Simulations and Drug Discovery. *BMC Biol.* **2011**, 9 (71).
- (195) Mobley, D. L.; Gilson, M. K. Predicting Binding Free Energies: Frontiers and Benchmarks. *Annu. Rev. Biophys.* **2017**, 46, 531–558.
- (196) Massova, I.; Kollman, P. A. Combined Molecular Mechanical and Continuum

- Solvent Approach (MM-PBSA / GBSA) to Predict Ligand Binding. *Perspect. Drug Discov. Des.* **2000**, *18*, 113–135.
- (197) Wang, L.; Wu, Y.; Deng, Y.; Kim, B.; Pierce, L.; Krilov, G.; Lupyán, D.; Robinson, S.; Dahlgren, M. K.; Greenwood, J.; et al. Accurate and Reliable Prediction of Relative Ligand Binding Potency in Prospective Drug Discovery by Way of a Modern Free-Energy Calculation Protocol and Force Field. *J. Am. Chem. Soc.* **2015**, *137* (7), 2695–2703.
- (198) Habib, E.; Maia, B.; Assis, L. C.; Oliveira, T. A. De; Marques, A.; Taranto, A. G. Structure-Based Virtual Screening: From Classical to Artificial Intelligence. *Front. Chem.* **2020**, *8* (343).
- (199) Chen, W.; Gilson, M. K.; Webb, S. P.; Potter, M. J. Modelling Protein-Ligand Binding by Mining Minima. *J. Chem. Theory Comput.* **2010**, *6* (11), 3540–3557.
- (200) Gilson, M. K.; Given, J. A.; Bush, B. L.; McCammon, J. A. The Statistical-Thermodynamic Basis for Computation of Binding Affinities : A Critical Review. *Biophys. J.* **1997**, *72*, 1047–1069.
- (201) Wang, E.; Sun, H.; Wang, J.; Wang, Z.; Liu, H.; Zhang, J. Z. H.; Hou, T. End-Point Binding Free Energy Calculation with MM / PBSA and MM / GBSA : Strategies and Applications in Drug Design. *Chem. Rev.* **2019**, *119* (16), 9478–9508.
- (202) Wang, L.; Deng, Y.; Knight, J. L.; Wu, Y.; Kim, B.; Sherman, W.; Shelley, J. C.; Lin, T.; Abel, R. Modeling Local Structural Rearrangements Using FEP/REST: Application to Relative Binding Affinity Predictions of CDK2 Inhibitors. *J. Chem. Theory Comput.* **2013**, *9*, 1282–1293.
- (203) Khalili-Araghi, F.; Jogini, V.; Yarov-Yarovoy, V.; Emad, T.; Roux, B.; Schulten, K. Calculation of the Gating Charge for the Kv1.2 Voltage-Activated Potassium Channel. *Biophys. J.* **2010**, *98* (2189–2198).
- (204) Wang, L.; Wu, Y.; Deng, Y.; Kim, B.; Pierce, L.; Krilov, G.; Lupyán, D.; Robinson, S.; Dahlgren, M. K.; Greenwood, J.; et al. Accurate and Reliable Prediction of Relative Ligand Binding Potency in Prospective Drug Discovery by Way of a Modern Free-Energy Calculation Protocol and Force Field. *J. Am. Chem. Soc.* **2015**, *137*, 2695–2703.
- (205) Wang, L.; Berne, B. J.; Friesner, R. A. On Achieving High Accuracy and Reliability in the Calculation of Relative Protein – Ligand Binding Affinities. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109* (6), 1937–1942.

- (206) Maxwell, D.; Tirado-Rives, J.; Jorgensen, W. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.* **1996**, *118* (45), 11225–11236.
- (207) Jorgensen, W. L.; Ruiz-caro, J.; Tirado-rives, J.; Basavapathruni, A.; Anderson, S.; Hamilton, A. D. Computer-Aided Design of Non-Nucleoside Inhibitors of HIV-1 Reverse Transcriptase. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 663–667.
- (208) Heinzelmann, G.; Henriksen, N. M.; Gilson, M. K. Attach-Pull-Release Calculations of Ligand Binding and Conformational Changes on the First BRD4 Bromodomain. *J. Chem. Theory Comput.* **2017**, *13*, 3260–3275.
- (209) Heinzelmann, G.; Gilson, M. K. Automation of Absolute Protein - Ligand Binding Free Energy Calculations for Docking Refinement and Compound Evaluation. *Sci. Rep.* **2021**, *11*, 1116–1134.
- (210) Fratev, F.; Sirimulla, S. An Improved Free Energy Perturbation FEP+ Sampling Protocol for Flexible Ligand- Binding Domains. *Sci. Rep.* **2019**, *9*, 16829–16842.
- (211) Hou, T.; Wang, J.; Li, Y.; Wang, W. Assessing the Performance of the Molecular Mechanics / Poisson Boltzmann Surface Area and Molecular Mechanics / Generalized Born Surface Area Methods . II . The Accuracy of Ranking Poses Generated From Docking. *J. Comput. Chem.* **2010**, *32* (5), 866–877.
- (212) Sun, H., Duan, L., Chen, F., Liu, H., Wang, Z., Pan, P., Zhu, F. Zhang, J., Tingjun, H. Assessing the Performance of MM/PBSA and MM/GBSA Methods. 7. Entropy Effects on the Performance of End-Point Binding Free Energy Calculation Approaches. *Phys. Chem. Chem. Phys.* **2018**, *20* (21), 14450–14460.
- (213) Zhang, X., Perez-Sanchez, H., C. Lightstone, F. A Comprehensive Docking and MM/GBSA Rescoring Study of Ligand Recognition upon Binding Antithrombin. *Curr. Top. Med. Chem.* **2017**, *17* (14), 1631–1639.
- (214) Greenidge, P. A.; Kramer, C.; Mozziconacci, J.; Wolf, R. M. MM/GBSA Binding Energy Prediction on the PDBbind Data Set: Successes, Failures, and Directions for Further Improvement. *J. Chem. Inf. Model.* **2013**, *53* (1), 201–209.
- (215) Carlsson, J.; Boukharta, L.; Åqvist, J. Combining Docking , Molecular Dynamics and the Linear Interaction Energy Method to Predict Binding Modes and Affinities for Non-Nucleoside Inhibitors to HIV-1 Reverse Transcriptase. *J. Med. Chem.* **2008**, *51*, 2648–2656.
- (216) Aqvist, J.; Medina, C.; Samuelsson, J.-E. A New Method for Predicting Binding Affinity in Computer-Aided Drug Design. *Protein Engineering* **1994**, *7* (3), 385–

391.

- (217) Sham, Y. Y.; Chu, Z. T.; Tao, H.; Warshel, A. Examining Methods for Calculations of Binding Free Energies: LRA, LIE, PDL-D-LRA, and PDL-D/S-LRA Calculations of Ligands Binding to an HIV Protease. *Proteins Struct. Funct. Bioinforma.* **2000**, *39*, 393–407.
- (218) Pereira, G.; Cecchini, M. A Multi-Basin Quasi-Harmonic Approach for the Calculation of the Configurational Entropy of Small Molecules in Solution. *J. Chem. Theory Comput.* **2021**, *17* (2), 1133–1142.
- (219) Alder, B. J.; Wainwright, T. E. Phase Transition for a Hard Sphere System. *J. Chem. Phys.* **1957**, *27*, 1208–1209.
- (220) Alder, B.; Wainwright, T. Studies in Molecular Dynamics. I. General Method. *J. Chem. Phys.* **1959**, *31* (2), 459–466.
- (221) Karplus, M.; McCammon, J. A. Molecular Dynamics Simulations of Biomolecules. *Nature* **2002**, *9* (9), 646–652.
- (222) Pereira, G.; Szwarz, B.; Mondragão, M. A.; Lima, P. A.; Pereira, F. A Ligand-Based Approach to the Discovery of Lead-Like Potassium Channel K_v 1.3 Inhibitors. *ChemistrySelect* **2018**, *3* (5), 1352–1364.
- (223) Aqvist, J.; Wennerström, P.; Nervall, M.; Bjelic, S.; Brandsdal, B. O. Molecular Dynamics Simulations of Water and Biomolecules with a Monte Carlo Constant Pressure Algorithm. *J. Chem. Phys. Lett.* **2004**, *384*, 288–294.
- (224) McQuarrie, D. A. *Statistical Mechanics*; Books, U. S., Ed.; Sausalito, California, 2000.
- (225) Suárez, D.; Díaz, N. Direct Methods for Computing Single-Molecule Entropies from Molecular Simulations. *WIREs Comput. Mol. Sci.* **2014**, *5* (1), 1–26.
- (226) Shirts, M. R.; Mobley, D. L. *An Introduction to Best Practices in Free Energy Calculations*; 2012; Vol. Chapter 11.
- (227) Liu, P.; Kim, B.; Friesner, R. A.; Berne, B. J. Replica Exchange with Solute Tempering: A Method for Sampling Biological Systems in Explicit Water. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102* (39), 13749–13754.
- (228) Sugita, Y.; Okamoto, Y. Replica-Exchange Molecular Dynamics Method for Protein Folding. *Chem. Phys. Lett.* **1999**, *314*, 141–151.
- (229) Wang, C.; Greene, D. A.; Xiao, L.; Qi, R.; Luo, R. Recent Developments and Applications of the MMPBSA Method. *Front. Mol. Biosci.* **2018**, *4* (87).
- (230) Shirts, M. R.; Pande, V. S. Comparison of Efficiency and Bias of Free Energies

- Computed by Exponential Averaging, the Bennett Acceptance Ratio, and Thermodynamic Integration. *J. Chem. Phys.* **2010**, *144*107 (2005).
- (231) Schrödinger, E. An Undulatory Theory of the Mechanics of Atoms and Molecules. *Phys. Rev.* **1926**, *28* (6), 1049–1070.
- (232) Born, M.; Oppenheimer, R. On the Quantum Theory of Molecules. *Ann. d. Phys.* **1927**, *20*, 457–484.
- (233) Senftle, T. P.; Hong, S.; Islam, M.; Kylasa, S. B.; Zheng, Y.; Shin, Y. K.; Junkermeier, C.; Engel-herbert, R.; Janik, M. J.; Aktulga, H. M.; et al. The ReaxFF Reactive Forcefield: Development, Applications and Future Directions. *Nat. Publ. Gr.* **2016**, *2*, 15011.
- (234) Huang, J.; Mackerell Jr., A. D. CHARMM36 All-Atom Additive Protein Force Field: Validation Based on Comparison to NMR Data. *J. Comput. Chem.* **2013**, *34* (25), 2135–2145.
- (235) Christen, M.; Hünenberger, P.; Bakowies, D.; Baron, R.; Bürgi, R.; Geerke, D.; Heinz, T.; Kastholz, M.; Kräutler, V.; Oostenbrink, C.; et al. The GROMOS Software for Biomolecular Simulation: GROMOS05. *J. Comput. Chem.* **2005**, *26* (16), 1719–1751.
- (236) Jones, J. E. On the Determination of Molecular Fields. 1. From the Variation of the Viscosity of a Gas with Temperature. *Proc. R. Soc. London. Ser. A, Contain. Pap. a Math. Phys. Character.* **1924**, *106* (738), 441–462.
- (237) Salomon-Ferrer, R.; Goetz, A. W.; Poole, D.; Grand, S. Le; Walker, R. C. Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald. *J. Chem. Theory Comput.* **2013**, *9* (9), 3878–3888.
- (238) Verlet, L. Computer “Experiments” on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. *Phys. Rev.* **1967**, *159* (1), 98–103.
- (239) Swope, W. C.; Andersen, H. C.; Berens, P. H.; Wilson, K. R. A Computer Simulation Method for the Calculation of Equilibrium Constants for the Formation of Physical Clusters of Molecules: Application to Small Water Clusters. *J. Chem. Phys.* **1982**, *76* (1), 637–649.
- (240) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of n-Alkanes. *J. Comput. Phys.* **1977**, *23*, 321–341.
- (241) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. LINCS : A Linear Constraint Solver for Molecular Simulations. *J. Comput. Chem.* **1997**, *18* (12),

- 1463–1472.
- (242) Balusek, C.; Hwang, H.; Lau, C. H.; Lundquist, K.; Hazel, A.; Lynch, D. L.; Reggio, P. H.; Wang, Y.; Gumbart, J. C.; Kong, H. Accelerating Membrane Simulations with Hydrogen Mass Repartitioning Curtiss. *J. Chem. Theory Comput.* **2019**, *15* (8), 4673–4686.
- (243) Loncharich, R.; Brooks, B.; Pastor, R. Langevin Dynamics of Peptides: The Frictional Dependence of Isomerization Rates of N-Acetylalanine-N'-Methylamide. *Biopolymers* **1992**, *32* (5), 523–535.
- (244) Evans, D. J.; Holian, B. L. The Nose–Hoover Thermostat. *J. Chem. Phys.* **1985**, *83* (8), 4069–4074.
- (245) Blanc, F. Exploration de La Transduction Chimio-myosine Par Simulations Numériques, 2018, Thèse de Doctorat.
- (246) Berendsen, H. J. C.; Postma, J. P. M.; Gunsteren, W. F. Van; DiNola, A.; Haak, J. R. Molecular Dynamics with Coupling to an External Bath. *J. Chem. Phys.* **1984**, *81* (8), 3684–3690.
- (247) Cave-Ayland, C. I.; Skylaris, C.; Essex, J. W. A Monte Carlo Resampling Approach for the Calculation of Hybrid Classical and Quantum Free Energies. *J. Chem. Theory Comput.* **2017**, *13* (2), 415–424.
- (248) Aldeghi, M.; Heifetz, A.; Bodkin, M. J.; Biggin, P. C. Accurate Calculation of the Absolute Free Energy of Binding for Drug Molecules. *Chem. Sci.* **2016**, *7*, 207–218.
- (249) Jorgensen, W. L.; Buckner, J. K.; Boudon, S.; Tirado-Rives, J. Efficient Computation of Absolute Free Energies of Binding by Computer Simulations. Application to the Methane Dimer in Water. *J. Chem. Phys.* **2011**, *89* (6), 3742–3746.
- (250) Zwanzig, R. W. High Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. *J. Chem. Phys.* **1954**, *22* (8), 1420–1426.
- (251) Chen, W.; Deng, Y.; Russell, E.; Wu, Y.; Abel, R.; Wang, L. Accurate Calculation of Relative Binding Free Energies between Ligands with Different Net Charges. *J. Chem. Theory Comput.* **2018**, *14* (12), 6346–6358.
- (252) Rifai, E. A.; Dijk, M. Van; Vermeulen, N. P. E.; Yanuar, A.; Geerke, D. P. A Comparative Linear Interaction Energy and MM/PBSA Study on SIRT1 – Ligand Binding Free Energy Calculation. *J. Chem. Inf. Model.* **2019**, *59*, 4018–4033.
- (253) Huang, D.; Caflisch, A. Efficient Evaluation of Binding Free Energy Using

- Continuum Electrostatics. *J. Med. Chem.* **2004**, *47* (23), 5791–5797.
- (254) Rastelli, G., Del Rio, A., Degliesposti, G., Sgobba, M. Fast and Accurate Predictions of Binding Free Energies Using MM-PBSA and MM-GBSA. *J. Comput. Chem.* **2010**, *31*, 797–810.
- (255) Aqvist, J.; Luzhkov, V. B.; Brandsdal, B. O. Ligand Binding Affinities from MD Simulations. *Acc. Chem. Res.* **2002**, *35* (6), 358–365.
- (256) Rifai, E. A.; Dijk, M. Van; Geerke, D. P. Recent Developments in Linear Interaction Energy Based Binding Free Energy Calculations. *Front. Mol. Biosci.* **2020**, *7* (114).
- (257) Stjernschantz, E.; Oostenbrink, C. Improved Ligand-Protein Binding Affinity Predictions Using Multiple Binding Modes. *Biophys. J.* **2010**, *98* (11), 2682–2691.
- (258) Massova, I.; Kollman, P. A. Computational Alanine Scanning To Probe Protein - Protein Interactions: A Novel Approach To Evaluate Binding Free Energies. *J. Am. Chem. Soc.* **1999**, *121* (36), 8133–8143.
- (259) Srinivasan, J.; Cheatham, T. E.; Cieplak, P.; Kollman, P. A.; Case, D. A. Continuum Solvent Studies of the Stability of DNA, RNA, and Phosphoramidate - DNA Helices. *J. Am. Chem. Soc.* **1998**, *120* (37), 9401–9409.
- (260) Chang, C.; Chen, W.; Gilson, M. K. Evaluating the Accuracy of the Quasiharmonic Approximation. *J. Chem. Phys.* **2005**, *1* (5), 1017–1028.
- (261) Karplus, M.; Kushick, J. N. Method for Estimating the Configurational Entropy. *Macromolecules* **1981**, *14*, 325–332.
- (262) Brooks, B.; Karplus, M. Harmonic Dynamics of Proteins: Normal Modes and Fluctuations in Bovine Pancreatic Trypsin Inhibitor. *Proc. Natl. Acad. Sci. U. S. A.* **1983**, *80*, 6571–6575.
- (263) Bahar, I.; Lezon, T. R.; Bakan, A.; Shrivastava, I. H. Normal Mode Analysis of Biomolecular Structures : Functional Mechanisms of Membrane Proteins. *Chem. Rev.* **2010**, *110*, 1463–1497.
- (264) Montalvo-Acosta, J. J.; Pacak, P.; Barreto Gomes, D. E.; Cecchini, M. A Linear Interaction Energy Model for Cavitand Host-Guest Binding Affinities. *J. Phys. Chem. B* **2018**, *122* (28), 6810–6814.
- (265) Swanson, J. M. J.; Henchman, R. H.; McCammon, J. A. Revisiting Free Energy Calculations : A Theoretical Connection to MM/PBSA and Direct Calculation of the Association Free Energy. *Biophys. J.* **2004**, *86* (1), 67–74.
- (266) Genheden, S.; Ryde, U. L. F. How to Obtain Statistically Converged MM/GBSA

- Results. *J. Comput. Chem.* **2010**, *31* (4), 837–846.
- (267) Miller, B. R.; Mcgee, T. D.; Swails, J. M.; Homeyer, N.; Gohlke, H.; Roitberg, A. E. MMPBSA.Py : An Efficient Program for End-State Free Energy Calculations. *J. Chem. Theory Comput.* **2012**, *8*, 3314–3321.
- (268) Feig, M.; Onufriev, A.; Lee, M. S.; Im, W.; Case, D. A.; L Brooks III, C. Performance Comparison of Generalized Born and Poisson Methods in the Calculation of Electrostatic Solvation Energies for Protein Structures. *J. Comput. Chem.* **2003**, *25* (2), 265–284.
- (269) Nicholls, A.; Honig, B. A Rapid Finite Difference Algorithm, Utilizing Successive Over-Relaxation to Solve the Poisson-Boltzmann Equation. *J. Comput. Chem.* **1991**, *12* (4), 435–445.
- (270) Prabhu, N. V.; Panda, M.; Yang, Q.; Sharp, K. A. Explicit Ion, Implicit Water Solvation for Molecular Dynamics of Nucleic Acids and Highly Charged Molecules. *J. Comput. Chem.* **2008**, *29*, 1113–1130.
- (271) Li, L.; Li, C.; Sarkar, S.; Zhang, J.; Witham, S.; Zhang, Z.; Wang, L.; Smith, N. DelPhi: A Comprehensive Suite for DelPhi Software and Associated Resources. *BMC Biophys.* **2012**, *5* (9).
- (272) Jurrus, E.; Engel, D.; Star, K.; Monson, K.; Brandi, J.; Felberg, L. E.; Brookes, D. H.; Wilson, L.; Chen, J.; Liles, K.; et al. Improvements to the APBS Biomolecular Solvation Software Suite. *Protein Sci.* **2018**, *27* (1), 112–128.
- (273) Hou, T.; Wang, J.; Li, Y.; Wang, W. Assessing the Performance of the MM/PBSA and MM/GBSA Methods. 1. The Accuracy of Binding Free Energy Calculations Based on Molecular Dynamics Simulations. *J. Chem. Inf. Model.* **2011**, *51* (1), 69–82.
- (274) Till, M. S.; Ullmann, G. M. McVol - A Program for Calculating Protein Volumes and Identifying Cavities by a Monte Carlo Algorithm. *J. Mol. Model.* **2010**, *16*, 419–429.
- (275) Sørensen, J.; Fenley, M. O.; Amaro, R. E. A Comprehensive Exploration of Physical and Numerical Parameters in the Poisson – Boltzmann Equation for Applications to Receptor – Ligand Binding A Practical Guide to PB. In *Computational Electrostatics for Biological Applications*; 2015; pp 39–70.
- (276) Sigalov, G.; Scheffel, P.; Onufriev, A. Incorporating Variable Dielectric Environments into the Generalized Born Model. *J. Chem. Phys.* **2005**, *122* (094511).

- (277) Srinivasan, J.; Trevathan, M. W.; Beroza, P.; Case, D. A. Regular Article Application of a Pairwise Generalized Born Model to Proteins and Nucleic Acids : Inclusion of Salt Effects. *Theor. Chem. Acc.* **1999**, *101*, 426–434.
- (278) Bashford, D.; Case, D. A. Generalized Born Models of Macromolecular Solvation Effects. *Annu. Rev. Phys. Chem.* **2000**, *51*, 129–152.
- (279) Wojciechowski, M.; Lesyng, B. Generalized Born Model: Analysis, Refinement, and Applications to Proteins. *J. Phys. Chem. B* **2004**, *108* (47), 18368–18376.
- (280) Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. Parametrized Models of Aqueous Free Energies of Solvation Based on Pairwise Descreening of Solute Atomic Charges from a Dielectric Medium. *J. Phys. Chem. B* **1996**, *100* (51), 19824–19839.
- (281) Onufriev, A.; Bashford, D.; Case, D. A. Exploring Protein Native States and Large-Scale Conformational Changes With a Modified Generalized Born Model. *Proteins Struct. Funct. Bioinforma.* **2004**, *55*, 383–394.
- (282) Forouzesh, N.; Izadi, S.; Onufriev, A. V. Grid-Based Surface Generalized Born Model for Calculation of Electrostatic Binding Free Energies. *J. Chem. Inf. Model.* **2017**, *57* (10), 2505–2513.
- (283) Wang, E.; Liu, H.; Wang, J.; Weng, G.; Sun, H.; Wang, Z.; Kang, Y.; Hou, T. Development and Evaluation of MM/GBSA Based on a Variable Dielectric GB Model for Predicting Protein-Ligand Binding Affinities. *J. Chem. Inf. Model.* **2020**, *60* (11), 5353–5365.
- (284) Genheden, S.; Ryde, U. Comparison of End-Point Continuum-Solvation Methods for the Calculation of Protein-Ligand Binding Free Energies. *Proteins Struct. Funct. Bioinforma.* **2012**, *80* (5), 1326–1342.
- (285) Weiser, J.; Shenkin, P. S.; Still, W. C. Approximate Atomic Surfaces from Linear Combinations of Pairwise Overlaps. *J. Comput. Biol.* **1999**, *20* (2), 217–230.
- (286) Barone, V.; Cossi, M.; Tomasi, J.; Barone, V.; Cossi, M. A New Definition of Cavities for the Computation of Solvation Free Energies by the Polarizable Continuum Model. *J. Chem. Phys.* **1997**, *107* (8), 3210–3221.
- (287) Genheden, S.; Kongsted, J.; Sönderhjelm, P.; Ryde, U. Nonpolar Solvation Free Energies of Protein-Ligand Complexes. *J. Chem. Theory Comput.* **2010**, *6* (11), 3558–3568.
- (288) Ben-shalom, I. Y.; Pfeiffer-marek, S.; Baringhaus, K.; Gohlke, H. Efficient Approximation of Ligand Rotational and Translational Entropy Changes upon

- Binding for Use in MM-PBSA Calculations. *J. Chem. Inf. Model.* **2017**, *57* (2), 170–189.
- (289) Sgobba, M.; Caporuscio, F.; Anighoro, A.; Portioli, C.; Rastelli, G. Application of a Post-Docking Procedure Based on MM-PBSA and MM-GBSA on Single and Multiple Protein Conformations. *Eur. J. Med. Chem.* **2012**, *58*, 431–440.
- (290) Wang, J. Fast Identification of Possible Drug Treatment of Coronavirus Disease-19 (COVID-19) through Computational Drug Repurposing Study. *J. Chem. Inf. Comput. Sci.* **2020**, *60*, 3277–3286.
- (291) Lagarias, P.; Barkan, K.; Tzortzini, E.; Vrontaki, E.; Ladds, G.; Kolocouris, A. Insights to the Binding of a Selective Adenosine A3 Receptor Antagonist Using Molecular Dynamic Simulations, MM-PBSA and MM-GBSA Free Energy Calculations and Mutagenesis. *J. Chem. Inf. Model.* **2019**, *59* (12), 5183–5197.
- (292) Wäschenbach, L.; Gertzen, C. G. W.; Keitel, V.; Gohlke, H. Dimerization Energetics of the G-Protein Coupled Bile Acid Receptor TGR5 from All-Atom Simulations. *J. Comput. Chem.* **2020**, *41*, 874–884.
- (293) Killian, B. J.; Kravitz, J. Y.; Gilson, M. K.; Killian, B. J.; Kravitz, J. Y.; Gilson, M. K. Extraction of Configurational Entropy from Molecular Simulations via an Expansion Approximation. *J. Chem. Phys.* **2007**, *127* (024107).
- (294) Conti, S.; Cecchini, M. Predicting Molecular Self-Assembly at Surfaces: A Statistical Thermodynamics and Modeling Approach. *Phys. Chem. Chem. Phys.* **2016**, *18* (46), 31480–31493.
- (295) Makhatadze, G. I.; Privalov, P. L. On the Entropy of Protein Folding. *Protein Sci.* **1996**, *5*, 507–510.
- (296) Andricioaei, I.; Karplus, M. On the Calculation of Entropy from Covariance Matrices of the Atomic Fluctuations. *J. Chem. Phys.* **2001**, *6289* (115), 1–5.
- (297) Goethe, M.; Fita, I.; Rubi, J. M. Testing the Mutual Information Expansion of Entropy with Multivariate Gaussian Distributions. *J. Chem. Phys.* **2017**, *147* (224102), 1–9.
- (298) Hnizdo, V.; Tan, J.; Killian, B. J.; Gilson, M. K. Efficient Calculation of Configurational Entropy from Molecular Simulations by Combining the Mutual-Information Expansion and Nearest-Neighbor Methods. *J. Comput. Chem.* **2008**, *29* (10), 1605–1614.
- (299) Fogolari, F.; Esposito, G. Optimal Relabeling of Water Molecules and Single-Molecule Entropy Estimation. *Biophysica* **2021**, *1*, 279–296.

- (300) Hnizdo, V.; Darian, E. V. A.; Fedorowicz, A.; Demchuk, E.; Li, S.; Singh, H. Nearest-Neighbor Nonparametric Method for Estimating the Configurational Entropy of Complex Molecules *. *J. Comput. Chem.* **2006**, *28* (3), 655–668.
- (301) Hnizdo, V.; Misra, N. Nearest Neighbor Estimates of Entropy. *Am. J. Math. Manag. Sci.* **2003**, *22*, 301–321.
- (302) Hikiri, S.; Yoshidome, T.; Ikeguchi, M. Computational Methods for Configurational Entropy Using Internal and Cartesian Coordinates. *J. Chem. Theory Comput.* **2016**, *12*, 5990–6000.
- (303) Suárez, E.; Díaz, N.; Méndez, J.; Suárez, D. CENCALC: A Computational Tool for Conformational Entropy Calculations from Molecular Simulations. *J. Comput. Chem.* **2013**, *34* (23), 2041–2054.
- (304) Harpole, K. W.; Sharp, K. A. Calculation of Configurational Entropy with a Boltzmann-Quasiharmonic Model: The Origin of High-Affinity Protein-Ligand Binding. *J. Phys. Chem. B* **2011**, *115*, 9461–9472.
- (305) Nola, A. Di; Berendsen, H. J. C.; Edholm, O. Free Energy Determination of Polypeptide Conformations Generated by Molecular Dynamics. *Macromolecules* **1984**, *17* (10), 2044–2050.
- (306) Hnizdo, V.; Gilson, M. K. Thermodynamic and Differential Entropy under a Change of Variables. *Entropy* **2010**, *12*, 578–590.
- (307) Russell D. Johnson III. NIST Computational Chemistry Comparison and Benchmark Database <http://cccbdb.nist.gov/>.
- (308) Suárez, E.; Díaz, N.; Suárez, D. Entropy Calculations of Single Molecules by Combining the Rigid-Rotor and Harmonic-Oscillator Approximations with Conformational Entropy Estimations from Molecular Dynamics Simulations. *J. Chem. Theory Comput.* **2011**, *7* (8), 2638–2653.
- (309) Garrett, S. L. The Simple Harmonic Oscillator. In *Understanding Acoustics: An Experimentalist's view on Sound and Vibration*; Springer International Publishing: Pine Grove Mills, 2020; pp 59–131.
- (310) Ling, S.; Sanny, J.; Moebs, W. *University Physics - Volume 3*; OpenStax: Houston, Texas, 2016.
- (311) Guthrie, J. P. Use of DFT Methods for the Calculation of the Entropy of Gas Phase Organic Molecules: An Examination of the Quality of Results from a Simple Approach. *J. Phys. Chem. A* **2001**, *105*, 8495–8499.
- (312) Detar, D. F. Calculation of Entropy and Heat Capacity of Organic Compounds in

- the Gas Phase. Evaluation of a Consistent Method without Adjustable Parameters. Applications to Hydrocarbons. *J. Phys. Chem. A* **2007**, *111*, 4464–4477.
- (313) Pracht, P.; Grimme, S. Calculation of Absolute Molecular Entropies and Heat Capacities Made Simple. *Chem. Sci.* **2021**, *12*, 6551–6568.
- (314) Zhang, Y.; Zhang, S.; Xing, J.; Bahar, I. Normal Mode Analysis of Membrane Protein Dynamics Using the Vibrational Subsystem Analysis. *J. Chem. Phys.* **2021**, *154* (195102), 1–9.
- (315) Bauer, J.; Pavlovic, J.; Bauerova-Hlinkova, V. Normal Mode Analysis as a Routine Part of a Structural Investigation. *Molecules* **2019**, *24* (3293), 1–20.
- (316) National, S.; Miyashita, O.; Tama, F.; Chacón, P. *Normal Mode Analysis Techniques in Structural Biology*; John Wiley and Sons Ltd.: Chichester, 2014.
- (317) Baron, R.; Hu, P. H. Absolute Single-Molecule Entropies from Quasi-Harmonic Analysis of Microsecond Molecular Dynamics: Correction Terms and Convergence Properties. *J. Chem. Theory Comput.* **2009**, *5* (12), 3150–3160.
- (318) Amaral, M.; Kokh, D. B.; Bomke, J.; Wegener, A.; Buchstaller, H. P.; Eggenweiler, H. M.; Matias, P.; Sirrenberg, C.; Wade, R. C.; Frech, M. Protein Conformational Flexibility Modulates Kinetics and Thermodynamics of Drug Binding. *Nat. Commun.* **2017**, *8* (2276).
- (319) Unarta, I. C.; Xu, J.; Shang, Y.; Pui, H. Entropy of Stapled Peptide Inhibitors in Free State Is the Major Contributor to the Improvement of Binding Affinity with the GK Domain. *RSC Chem. Biol.* **2021**, *2*, 1274–1284.
- (320) Schlitter, J. Estimation of Absolute and Relative Entropies of Macromolecules Using the Covariance Matrix. *Chem. Phys. Lett.* **1993**, *215* (6), 617–621.
- (321) Karplus, M.; Ichiye, T.; Pettitt, B. M. Configurational Entropy of Native Proteins. *Biophys. J.* **1987**, *52*, 1083–1085.
- (322) Killian, B.; Kravitz, J. Y.; Somani, S.; Dasgupta, P.; Pang, Y.-P.; Gilson, M. K. Configurational Entropy in Protein-Peptide Binding. Computational Study of Tsg101 UEV Domain with an HIV-Derived PTAP Nonapeptide. *J. Mol. Biol.* **2009**, *389* (2), 315–335.
- (323) Singer, A. Maximum Entropy Formulation of the Kirkwood Superposition Approximation. *J. Chem. Phys.* **2004**, *121* (8), 3657–3666.
- (324) Panday, S. K.; Ghosh, I. Application and Comprehensive Analysis of Neighbor Approximated Information Theoretic Configurational Entropy Methods to Protein–Ligand Binding Cases. *J. Chem. Theory Comput.* **2020**, *16* (12), 7581–

7600.

- (325) Suárez, E.; Suárez, D. Multibody Local Approximation: Application to Conformational Entropy Calculations on Biomolecules. *J. Chem. Phys.* **2012**, *2012* (084115).
- (326) Chang, C.; Gilson, M. K. Tork : Conformational Analysis Method for Molecules and Complexes. *J. Comput. Chem.* **2003**, *24*, 1998–2003.
- (327) Malcolm W. Chase, J. *NIST-JANAF Thermochemical Tables*, Fourth edi.; American Chemical Society, American Institute of Physics for the National Institute of Standards and Technology, 1998.: Washington, DC, 1998.
- (328) Gurvich, L. V.; Veyts, I. V.; Alcock, C. B. *Thermodynamic Properties of Individual Substances*, 4th ed.; V.; Hemisphere, New York, 1989., 1989.
- (329) Frenkel, M., Kabo, G., Marsh, N., Roganov, G., Wilhoit, R. *TRC Data Series: Thermodynamics of Organic Compounds in the Gas Phase*, Volume 2.; CRC Press: Boca Raton, FL, 1994.
- (330) Li, Y.; Bell, A. T.; Head-Gordon, M. Thermodynamics of Anharmonic Systems : Uncoupled Mode Approximations for Molecules. *J. Chem. Phys.* **2016**, *12*, 2861–2870.
- (331) Jorge, M.; Garrido, N. M.; Queimada, J.; Economou, I. G.; Macedo, E. A. Effect of the Integration Method on the Accuracy and Computational Efficiency of Free Energy Calculations Using Thermodynamic Integration. *J. Chem. Theory Comput.* **2010**, *6* (4), 1018–1027.
- (332) Peter, C.; Oostenbrink, C.; Dorp, A. Van; Gunsteren, W. F. Van. Estimating Entropies from Molecular Dynamics Simulations. *J. Chem. Phys.* **2004**, *230* (6), 2652–2661.
- (333) Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A. Automatic Atom Type and Bond Type Perception in Molecular Mechanical Calculations. *J. Mol. Graph. Model.* **2006**, *25*, 247–260.
- (334) Petersson, G. A.; Bennett, A.; Tensfeldt, T. G.; Allaham, M. A.; Shirley, W. A.; Mantzaris, J. A Complete Basis Set Model Chemistry. I. The Total Energies of Closedshell Atoms and Hydrides of the Firstrow Elements. *J. Chem. Phys.* **1988**, *89* (4), 2193–2218.
- (335) Humphrey, W.; Dalke, A.; Schulten, K. VMD : Visual Molecular Dynamics. *J. Mol. Graph.* **1996**, *14*, 33–38.
- (336) Shao, J.; Tanner, S. W.; Thompson, N.; Cheatham, T. E. Clustering Molecular

- Dynamics Trajectories: 1. Characterizing the Performance of Different Clustering Algorithms. *J. Chem. Theory Comput.* **2007**, *3* (6), 2312–2334.
- (337) Tan, C.; Tan, Y.; Luo, R. Implicit Nonpolar Solvent Models. *J. Phys. Chem. B* **2007**, *111*, 12263–12274.
- (338) Sibson, R. SLINK: An Optimally Efficient Algorithm for the Single-Link Cluster Method. *Comput. Journal. Br. Comput. Soc.* **1976**, *16* (1), 30–34.
- (339) Defays, D. An Efficient Algorithm for a Complete Link Method. *Comput. Journal. Br. Comput. Soc.* **1977**, *20* (4), 364–366.
- (340) Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *KDD'96 Proc. Second Int. Conf. Knowl. Discov. Data Min.* **1996**, No. KDD'96, 226–231.
- (341) Lyne, P. D.; Lamb, M. L.; Saeh, J. C. Accurate Prediction of the Relative Potencies of Members of a Series of Kinase Inhibitors Using Molecular Docking and MM-GBSA Scoring. *J. Med. Chem.* **2006**, *49*, 4805–4808.
- (342) Lodish, H.; Berk, A.; Zipursky, S. L.; Matsudaira, P.; Baltimore, D.; Darnell, J. Myosin: The Actin Motor Protein. In *Molecular Cell Biology*; Freeman, W. H., Ed.; New York, 2000; p Section 18.3.
- (343) Alberts, B.; Johnson, A.; Lewis, J.; Raff, M.; Roberts, K.; Walter, P. Molecular Motors. In *Molecular Biology of the Cell*; Science, G., Ed.; New York, 2002.
- (344) Cappello, G.; Pierobon, P.; Symonds, C.; Busoni, L.; Gebhardt, J. C. M.; Rief, M.; Prost, J. Myosin V Stepping Mechanism. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104* (39), 15328–15333.
- (345) Kovács, M.; Tóth, J.; Hetényi, C.; Málnási-Csizmadia, A.; Seller, J. R. Mechanism of Blebbistatin Inhibition of Myosin II. *J. Biol. Chem.* **2004**, *279* (34), 35557–35563.
- (346) Preller, M.; Manstein, D. J. Myosin Structure, Allostery, and Mechano-Chemistry. *Structure* **2013**, *21* (11), 1911–1922.
- (347) Connell, C. B. O.; Tyska, M. J.; Mooseker, M. S. Myosin at Work: Motor Adaptations for a Variety of Cellular Functions. *Biochim. Biophys. Acta* **2007**, *1773*, 615–630.
- (348) Wells, A. L.; Lin, A. W.; Chen, L.; Safer, D.; Cain, S.; Hasson, T.; Carragher, B.; Milligan, R.; Sweeney, H. L. Myosin VI Is an Actin-Based Motor That Moves Backwards. *Nature* **1999**, *401*, 505–508.
- (349) Spudich, J. A. Three Perspectives on the Molecular Basis of Hypercontractility

- Caused by Hypertrophic Cardiomyopathy Mutations. *Pflügers Arch. - Eur. J. Physiol.* **2019**, *471*, 701–717.
- (350) Sirigu, S.; Hartman, J. J.; Planelles-Herrero, V. J.; Ropars, V.; Clancy, S.; Wang, X.; Stura, E. A.; Malik, F. I.; Houdusse, A. M. Highly Selective Inhibition of Myosin Motors Provides the Basis of Potential Therapeutic Application. *Proc. Natl. Acad. Sci. U. S. A.* **2016**, *113* (47), E7448–E7455.
- (351) Hindman, B.; Goeckeler, Z.; Sierros, K.; Wysolmerski, R. Non-Muscle Myosin II Isoforms Have Different Functions in Matrix Rearrangement by MDA-MB-231 Cells. *PLoS One* **2015**, *10* (7), 1–26.
- (352) Blanc, F.; Isabet, T.; Benisty, H.; Sweeney, H. L.; Cecchini, M.; Houdusse, A. An Intermediate along the Recovery Stroke of Myosin VI Revealed by X-Ray Crystallography and Molecular Dynamics. *Proc. Natl. Acad. Sci. U. S. A.* **2018**, *115* (24), 6213–6218.
- (353) Kawas, R. F.; Anderson, R. L.; Ingle, S. R. B.; Song, Y.; Sran, A. S.; Rodriguez, H. M. A Small-Molecule Modulator of Cardiac Myosin Acts on Multiple Stages of the Myosin Chemomechanical Cycle. *J. Biol. Chem.* **2017**, *292* (40), 16571–16577.
- (354) Sweeney, H. L.; Houdusse, A. Structural and Functional Insights into the Myosin Motor Mechanism. *Annu. Rev. Biophys.* **2010**, *39*, 539–557.
- (355) Rauscher, A. Á.; Gyimesi, M.; Kovács, M.; Málnási-csizmadi, A. Targeting Myosin by Blebbistatin Derivatives: Optimization and Pharmacological Potential. *Trends Biochem. Sci.* **2018**, *43* (9), 700–713.
- (356) Stump, W.; Blackwell, T.; Clippinger, S. R.; Greenberg, M. J.; Hartman, J. J.; Hwee, D. T.; Wang, J.; Wu, Y.; Schaletzky, J.; Paliwal, P.; et al. Characterization of the Cardiac Myosin Inhibitor CK-3773274: A Potential Therapeutic Approach for Hypertrophic Cardiomyopathy. *Biophys. J.* **2020**, *118* (3), 596a.
- (357) Chinthalapudi, K.; Taft, M. H.; Martin, R.; Heissler, S. M.; Preller, M.; Hartmann, F. K.; Brandstaetter, H.; Kendrick-Jones, J.; Tsiavaliaris, G.; Gutzeit, H. O.; et al. Mechanism and Specificity of Pentachloropseudilin-Mediated Inhibition of Myosin Motor Activity. *J. Biol. Chem.* **2011**, *286* (34), 29700–29708.
- (358) Sun, J.; Qiao, Y.-N.; Tao, T.; Zhao, W.; Wei, L.-S.; Li, Y.-Q.; Wang, W.; Wang, Y.; Zhao, Y.-W.; Zheng, Y.-Y.; et al. Distinct Roles of Smooth Muscle and Non-Muscle Myosin Light Chain-Mediated Smooth Muscle Contraction. *Front. Physiol.* **2020**, *11*, 593966.

- (359) Webb, R. C. Smooth Muscle Contraction and Relaxation. *Adv. Physiol. Educ.* **2003**, 27 (4), 201–206.
- (360) Berair, R.; Hollins, F.; Brightling, C. Airway Smooth Muscle Hypercontractility in Asthma. *J. Allergy* **2013**, 2013, 185971.
- (361) Chen, P.; Yin, J.; Guo, Y.; Xiao, H.; Wang, X.; Disanto, M. E.; Zhang, X. The Expression and Functional Activities of Smooth Muscle Myosin and Non-Muscle Myosin Isoforms in Rat Prostate. *J. Cell. Mol. Med.* **2018**, 22 (1), 576–588.
- (362) Sweeney, H. L. Regulation and Tuning of Smooth Muscle Myosin. *Am. J. Respir. Crit. Care Med.* **1998**, 158, 95–99.
- (363) Mahuteau-Betzer, F. Chimiothèque Nationale. *Médecine/Sciences* **2015**, 31, 417–422.
- (364) Brink, T.; Exner, T. E. Influence of Protonation, Tautomeric, and Stereoisomeric States on Protein - Ligand Docking Results. *J. Chem. Inf. Model.* **2009**, 49, 1535–1546.
- (365) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, 28 (1), 31–36.
- (366) Berthold, M. R.; Cebon, N.; Dill, F.; Gabriel, T. R.; Meinl, T.; Ohl, P.; Sieb, C.; Thiel, K. KNIME: The Konstanz Information Miner. In *Studies in Classification, Data Analysis and Knowledge Organization*; 2008; pp 319–326.
- (367) Korff, M. Von; Freyss, J.; Sander, T. Flexophore, a New Versatile 3D Pharmacophore Descriptor That Considers Molecular Flexibility. *J. Chem. Inf. Model.* **2008**, 48 (4), 797–810.
- (368) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K.; Simmerling, C. Ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from Ff99SB. *J. Chem. Theory Comput.* **2015**, 11 (8), 3696–3713.
- (369) Mongan, J.; Simmerling, C.; McCammon, J. A.; Case, D. A.; Onufriev, A. Generalized Born Model with a Simple, Robust Molecular Volume Correction. *J. Chem. Theory Comput.* **2007**, 3, 156–169.
- (370) Meagher, K. L.; Redman, L. T.; Carlson, H. A. Development of Polyphosphate Parameters for Use with the AMBER Force Field. *J. Comput. Chem.* **2003**, 24, 1016–1025.
- (371) Veenstra, D. L.; Ferguson, D. M.; Kollman, P. A. How Transferable Are Hydrogen

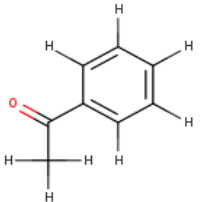
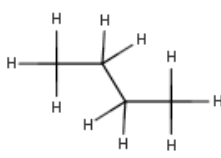
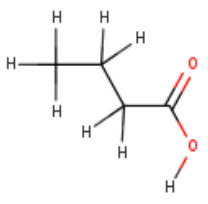
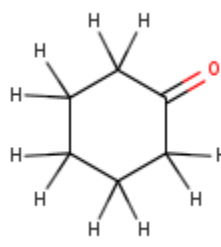
- Parameters in Molecular Mechanics Calculations? *J. Comput. Chem.* **1992**, *13* (8), 971–978.
- (372) Weiner, S. J.; Kollman, P. A.; Nguyen, D. T.; Case, D. A. An All Atom Force Field for Simulations of Proteins and Nucleic Acids. *J. Comput. Chem.* **1986**, *7* (2), 230–252.
- (373) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.
- (374) Allner, O.; Nilsson, L.; Villa, A. Magnesium Ion – Water Coordination and Exchange in Biomolecular Simulations. *J. Chem. Theory Comput.* **2012**, *8*, 1493–1502.
- (375) De la Cruz, E. M.; Ostap, M. Kinetic and Equilibrium Analysis of the Myosin ATPase. *Methods Enzymol.* **2009**, *455*, 157–192.
- (376) Bocci, G.; Carosati, E.; Vayer, P.; Arrault, A.; Lozano, S.; Cruciani, G. ADME-Space: A New Tool for Medicinal Chemists to Explore ADME Properties. *Sci. Rep.* **2017**, *7* (1), 25–27.
- (377) Jenwitheesuk, E.; Horst, J. A.; Rivas, K. L.; Van Voorhis, W. C.; Samudrala, R. Novel Paradigms for Drug Discovery: Computational Multitarget Screening. *Trends Pharmacol. Sci.* **2008**, *29* (2), 62–71.
- (378) Bollini, M.; Domaoal, R.; Thakur, V.; Gallardo-Macias, R.; Spasov, K.; Anderson, K.; Jorgensen, W. Computationally-Guided Optimization of a Docking Hit to Yield Catechol Diethers as Potent Anti-HIV Agents. *J. Med.* **2011**, *54* (24), 8582–8591.
- (379) Bevans, C., Krettler, C., Reinhart, C., Tran, H., Kobmann, K., Watzka, M. and Oldenburg J. Determination of the warfarin inhibition constant K_i for vitamin K 2,3-epoxide reductase complex subunit 1 (VKORC1) using an in vitro DTT-driven assay. *Biochimica et Biophysica Acta – General Subjects.* **2013**, *1830* (8), 4202–4210.
- (380) Montalvo-Acosta, J. Computational approaches to molecular recognition: from host-guest to protein-ligand binding. **2018**. Thèse de Doctorat.
- (381) Atanasov, A., Blunder, M., Fakhrudin, N., Liu, X., Noha, S., Malainer, C., Kramer, M., Cocic, A., Kunert, O., Schinkovitz, A., Heiss, E., Schuster, D., Dirsch, V. and Bauer, R. Polyacetylenes from *Notopterygium incisum* – New selective Partial

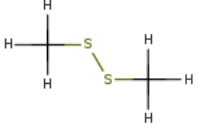
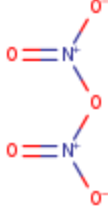
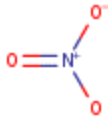
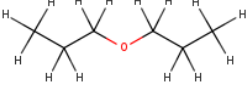
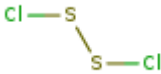

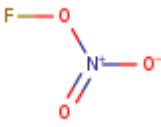
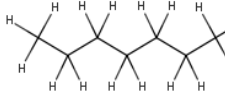
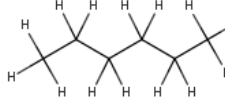
- Agonists of Peroxisome Proliferator-Activated Receptor-Gamma. *PLoS One*. **2013**, 8 (4), e61755.
- (382) Kalliokoski, T., Kramer, C., Vulpetti, A. and Gedeck, P. Comparability of Mixed IC₅₀ Data – A Statistical Analysis. *PLoS One*. **2013**, 8 (4), e61007.
- (383) Borea, P., Varani, K., Gessi, S., Gilli, P. And Dalpiaz, A. Receptor Binding Thermodynamics as a tool for linking drug efficacy and affinity. *Il Farmaco*. **1998**, 53 (4), 249-254.
- (384) Tuccinardi, T. What is the current value of MM/PBSA and MM/GBSA methods in drug discovery? *Exp Op on Drug Disc*. **2021**, 16(11), 1233-1237.
- (385) Feig, M. Modeling Solvent Environments: Applications to Simulations of Biomolecules. Chapter 6: Continuum Electrostatics Solvent Modelling with the Generalized Born model. Onufriev, A. 2009. Wiley-VCH, Verlag GmbH & KGaA, Weinheim.
- (386) Poli, G., Granchi, C., Rizzolio, F. and Tuccinardi, T. Application of MM-PBSA methods in Virtual Screening. *Molecules*. **2020**, 25(8), 1971.
- (387) Vane, J and Bottin, R. The mechanism of action of Aspirin. *Thromb Res*. **2003**, 110(5-6), 255-8.
- (388) Miner, J. and Hoffhines, A. The discovery of aspirin's antithrombotic effects. *Tex Heart Inst J*. **2007**, 34(2), 179-186.
- (389) Keseru, G. and Makara, G. Hit discovery and hit-to-lead approaches. *Drug Disc Tod*. **2006**, 11 (15-16).
- (390) Mahan, V. Clinical Trial Phases. *Int. J. Clin. Med*. **2014**, 5, 1374-1383.
- (391) Van Norman, G. Phase II Trials in Drug Development and Adaptive Trial Design. *JACC. Basic to translational science*. **2019**, 4(3), 428–437.
- (392) Brown, S., Gregory, W., Twelves, C., Buyse, M., Collinson, F., Parmar, M., Seymour, M., and Brown, J. Designing phase II trials in cancer: a systematic review and guidance. *Br J Cancer*. **2011**, 105, 194–199.

Supplementary Information: Chapter 5

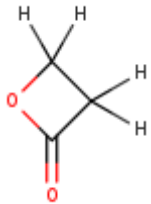
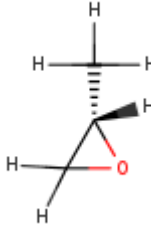
Annex S5.1: Dataset for the gas-phase entropy calculations.

Absolute entropies were extracted from the Computational Chemistry Comparison Benchmark DataBase from the National Institute of Standards and Technology (NIST).³⁰⁷ Experimental entropies in kcal/mol were provided at 298K and 1bar pressure. The number of non-redundant torsions and molecular conformers were evaluated from the analysis of corresponding Molecular Dynamics simulations in vacuum. Molecular structures were drawn using MarvinSketch from ChemAxon.²⁵

Molecule	TS [kcal/mol]	#Torsions	#Microstates	Molecular Structure
Acetophenone	26.57	2	6	
Butane	22.09	3	27	
Butanoic Acid	25.17	3	27	
Cyclohexanone	23.91	1	2	

Dimethyl disulfide	24.00	3	18	
Dinitrogen pentoxide	24.70	1	4	
Dinitrogen trioxide	22.43	0	1	
Di-n-propylether	30.11	6	729	
Disulfur dichloride	23.30	1	6	
Ethane	16.33	1	9	
Fluorine nitrate	20.89	0	0	
Heptane	30.50	6	729	
Hexane	27.71	5	243	

Methanol	17.09	1	3	
Methoxyethane	22.04	3	27	
Methylacetate	22.89	2	9	
Naphttalene	23.86	0	1	
Nitrosyl chloride	18.64	0	1	
Propylbenzene	28.35	3	27	
Pentane	24.91	4	81	
Propane	19.26	2	9	

Propiolactone	20.40	0	1	
Propylene-oxide	20.03	1	3	

The number of the torsions also includes torsions encompassing terminal methyl, hydroxyl and/or amine groups.

Annex S5.2: Dataset for the protein-ligand binding free energy calculations.

The experimental binding affinity expressed as $pK_i = -\log_{10} K_i$, the number of residues per protein, the ligand molecular weight, the total number of torsions per ligand, and the average RMSD with respect to the X-ray structure or the first snapshot of the production run over 100ns of MD are given. The number of torsions and the ligand MW were computed using DataWarrior.¹ RMSD values are given in Å

Complex	pKi	#Residues	Ligand MW ^a	#Torsions ^b	RMSD ^c	RMSD ^d
1CEB	6.00	80	157	3	1.21	1.22
1DF8	9.70	476	243	5	1.45	1.56
1GYX	2.48	152	121	1	1.21	1.23
1LAF	7.85	238	175	8	1.13	1.13
1O2Q	7.68	223	369	6	1.14	1.19
1O33	5.74	233	253	4	0.97	1.07
1O3I	7.30	223	343	5	1.46	1.56
1OWE	6.20	245	290	6	1.13	1.13
1P1N	6.80	516	212	6	2.87	3.00
1PB8	5.15	282	105	4	1.62	1.63
1UA4	4.22	454	345	6	1.01	1.02
1V2J	3.25	223	121	3	0.98	1.07
1V2U	3.37	223	121	3	1.05	1.17
1WDN	6.30	516	146	6	1.31	1.35
1Y20	5.15	223	101	3	2.05	2.04
2EXM	5.32	298	203	6	1.46	1.44

2FQW	6.68	316	269	7	1.08	1.08
2PQL	7.28	144	161	3	1.17	1.28
3BRN	8.70	153	177	4	0.97	0.95
3BU1	8.15	144	112	3	1.10	1.12
5STD	10.49	492	376	5	1.48	1.71

[a] – The molecular weight of the ligand in Daltons.

[b] – Torsions include rotatable bonds comprising terminal groups like methyl, hydroxyl and amine.

[c] – Average RMSD of the ligand from the X-ray binding mode.

[d] – Average RMSD of the ligand from the first snapshot of the MD production run.

Annex S5.3: Absolute entropy calculations by QHMB in the gas phase and aqueous solution.

The calculations were carried out using the Average Linkage algorithm for clustering and an RMSD-cutoff of 0.2Å. Upon clustering, the symmetry number and the molecular weight were obtained using VMD, the moments of inertia and the quasi-harmonic vibrational frequencies were accessed using CPPTRAJ.¹¹⁸ All entropy values are given in kcal/mol. Error bars for the QHMB calculations were estimated from the standard error of the mean (S.E.M.) in a standard block analysis. Absolute molecular entropies were computed using Thermo (available on the GitHub link: <https://github.com/SimoneCnt/thermo>).²⁹⁴ The systematic difference between the gas-phase and solution results is related to the definition of the standard state, which corresponds to a volume of 24.78L at 298.15K in vacuum and 1L in solution; see *Main Text*.

Molecule	TS°	QHA-clust	NMA-clust	Gas phase	Water
Acetophenone	26.57	33.05	25.91	27.34 ± 0.001	25.49 ± 0.006
Butane	22.09	30.17	20.69	23.09 ± 0.013	20.99 ± 0.010
Butanoic-acid	25.17	34.38	24.46	26.42 ± 0.008	24.85 ± 0.043
Cyclohexanone	23.91	26.03	23.21	23.88 ± 0.019	21.99 ± 0.004
Di-n-propylether	30.11	51.06	27.98	30.68 ± 0.050	28.83 ± 0.238
Dimethyl-disulfide	24.00	31.46	22.54	25.08 ± 0.002	22.67 ± 0.031
Dinitro-pentoxide	24.70	31.62	23.16	24.79 ± 0.012	22.56 ± 0.001

Dinitro-trioxide	22.43	22.80	21.19	21.45	±	19.60	±
				0.002		0.005	
Disulfur-dichloride	23.30	26.29	23.56	24.19	±	21.91	±
				0.001		0.001	
Ethane	16.33	17.38	15.72	17.52	±	14.55	±
				0.012		0.016	
Fluorine-nitrate	20.89	20.84	20.92	20.82	±	18.97	±
				0.005		0.020	
Heptane	30.50	47.69	27.76	30.44	±	28.47	±
				0.013		0.076	
Hexane	27.71	41.79	25.50	28.23	±	26.04	±
				0.011		0.013	
Methanol	17.09	17.72	16.56	17.71	±	15.82	±
				0.007		0.004	
Methoxyethane	22.04	29.11	21.16	23.15	±	21.19	±
				0.003		0.001	
Methyl-acetate	22.89	26.06	22.44	23.91	±	22.06	±
				0.007		0.007	
N-propyl-benzene	28.35	38.49	27.54	28.94	±	26.99	±
				0.003		0.014	
Naphthalene	23.86	23.73	23.61	23.76	±	21.00	±
				0.006		0.003	
Nitrosyl-chloride	18.64	17.98	17.98	17.98	±	16.08	±
				0.002		0.001	
Pentane	24.91	36.57	23.27	25.54	±	23.68	±
				0.007		0.012	
Propane	19.26	23.31	19.12	20.55	±	18.22	±
				0.010		0.016	
Propiolactone	20.40	20.33	20.32	20.31	±	18.45	±
				0.022		0.015	

Propylene-oxide	20.03	21.34	19.56	20.44	±	18.52	±
				0.012		0.003	

Annex S5.4: Binding free-energy results from MM/GBSA calculations.

The standard single trajectory approach was used. Entropy corrections by QHMB were computed by taking the difference between the absolute entropy of the ligand in the bound states minus that in the unbound state. The gb5 model was used for the polar contribution to the solvation free energy and the LCPO model was used for the nonpolar contribution to the solvation free energy.^{196,281,285} Experimental binding affinities at 298K were extracted from K_i values reported in Greendige et al.²¹⁴ All free energy values are given in kcal/mol. Error bars were estimated from the standard error of the mean (S.E.M) in standard block analysis.

Complex	Exp ΔG°	Ligand MW ^a	Torsions ^b	-QHMB ^c	+QHMB ^d	Correction
1CEB	-8.18	157	3	-29.82 ± 0.03	-27.19 ± 0.10	2.63 ± 0.09
1DF8	-13.23	243	5	-41.10 ± 0.03	-35.78 ± 0.26	5.32 ± 0.25
1GYX	-3.39	121	1	-19.60 ± 0.02	-18.45 ± 0.15	1.15 ± 0.15
1LAF	-10.71	175	8	-47.85 ± 0.03	-36.28 ± 0.24	11.56 ± 0.24
1O2Q	-10.47	369	6	-46.22 ± 0.04	-34.44 ± 0.26	11.78 ± 0.26
1O33	-7.84	253	4	-32.92 ± 0.03	-25.40 ± 0.16	7.52 ± 0.15
1O3I	-9.96	343	5	-38.16 ± 0.03	-29.10 ± 0.14	9.06 ± 0.14
1OWE	-8.46	290	6	-38.29 ± 0.03	-28.36 ± 0.19	9.93 ± 0.19
1P1N	-9.27	212	6	-35.36 ± 0.05	-29.84 ± 0.10	5.53 ± 0.09

1PB8	-7.03	105	4	-29.56 ± 0.03	-24.77 ± 0.09	4.80 ± 0.08
1UA4	-5.76	345	6	-38.86 ± 0.03	-31.79 ± 0.28	7.08 ± 0.28
1V2J	-4.43	121	3	-24.27 ± 0.02	-18.91 ± 0.16	5.36 ± 0.16
1V2U	-4.60	121	3	-25.59 ± 0.03	-20.35 ± 0.12	5.24 ± 0.11
1WDN	-8.60	146	6	-34.91 ± 0.03	-29.76 ± 0.15	5.15 ± 0.15
1Y20	-7.26	101	3	-30.97 ± 0.02	-27.75 ± 0.06	3.22 ± 0.06
2EXM	-5.60	203	6	-32.72 ± 0.02	-26.76 ± 0.06	5.96 ± 0.06
2FQW	-9.11	269	7	-35.78 ± 0.03	-28.33 ± 0.18	7.45 ± 0.17
2PQL	-9.92	161	3	-39.08 ± 0.02	-34.98 ± 0.21	4.09 ± 0.20
3BRN	-11.87	177	4	-44.47 ± 0.02	-40.00 ± 0.13	4.47 ± 0.12
3BU1	-11.12	112	3	-32.41 ± 0.02	-30.03 ± 0.26	2.38 ± 0.26
5STD	-14.32	376	5	-47.52 ± 0.03	-40.38 ± 0.16	7.13 ± 0.16

a – Ligand molecular weight in Daltons.

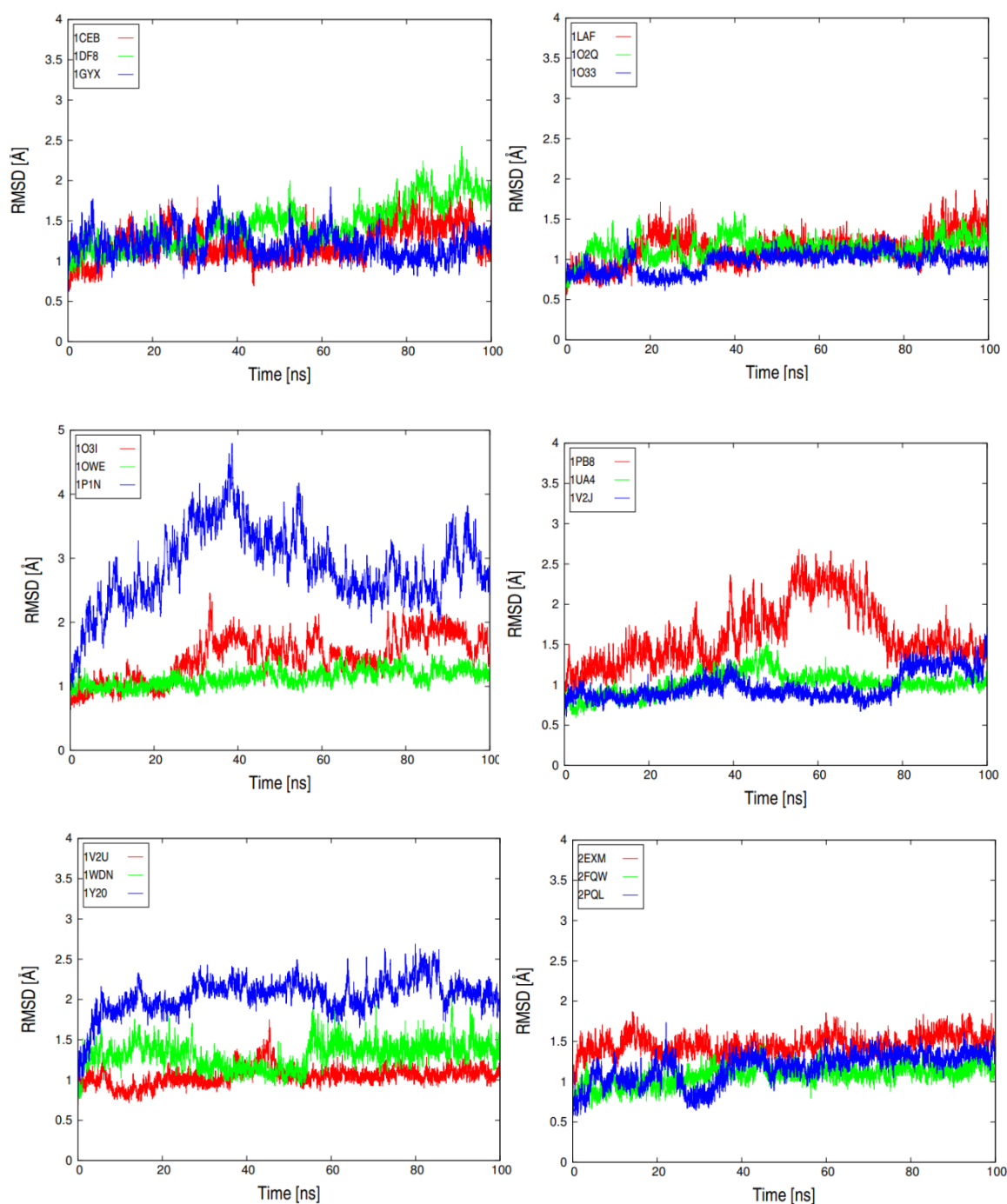
b – The number of torsions include torsions connected to terminal groups like methyl, hydroxyl and amine.

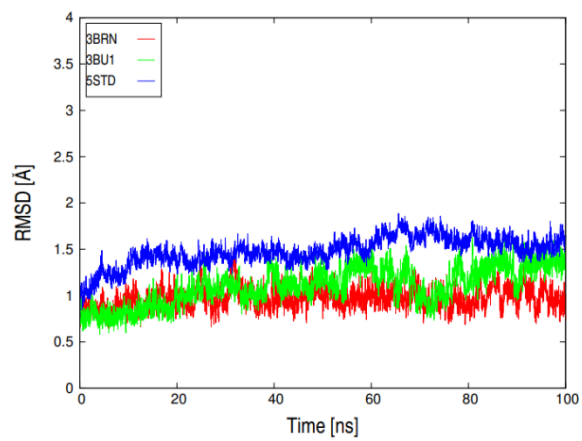
c – MMGBSA calculation carried out in the single trajectory formalism, using 10000 molecular snapshots.

d – MMGBSA calculation carried out in the single trajectory formalism, using 10000 molecular snapshots, augmented with the QHMB entropy correction for the ligand loss of conformational entropy upon binding.

Annex S5.5: Time series of the heavy-atom RMSD of the ligand from its crystallographic binding mode in 21 protein-ligand complexes extracted from the Greenidge dataset.

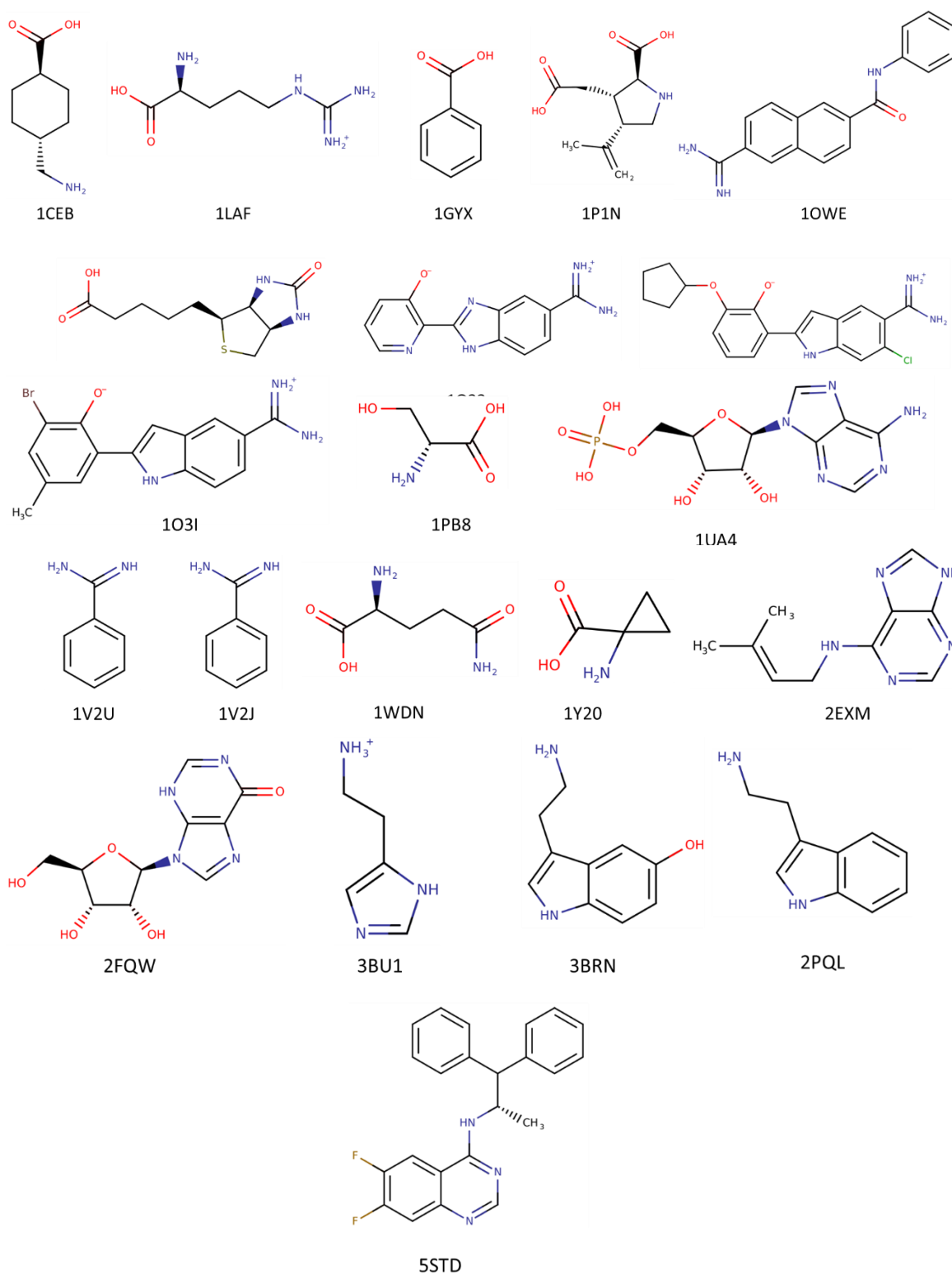
The molecular dynamics simulations were carried out at 298.15K in the NPT ensemble and superimposed to the protein backbone of the reference structure. Most simulations remain close to the initial X-ray structure. In one case, i.e. 1P1N, the ligand changes its binding mode.





Annex S5.6: Molecular structures of the compounds from the Greenidge dataset.

Structures were extracted from the Protein Data Bank and drawn with MarvinSketch from ChemAxon.



Supplementary Information: Chapter 6

Annex S6.1: MM/GBSA Binding Free Energy results for the 60 compounds that were prioritized from CN.

Complex	MMGBSA	-TΔS	ΔG	Bound S.E.M	Unbound S.E.M	MMGBSA S.E.M	ΔG S.E.M
AB-00000226	-62.78	16.46	-46.32	0.31	0.16	0.03	0.35
AB-00002083	-61.27	16.72	-44.55	0.09	0.10	0.03	0.14
AB-0000267	-64.03	9.05	-54.98	0.11	0.06	0.03	0.13
AB-00004558	-61.95	5.75	-56.20	0.03	0.17	0.03	0.18
AB-00007072	-64.44	12.25	-52.19	0.26	0.07	0.03	0.27
AB-00010786	-55.16	10.38	-44.79	0.10	0.12	0.03	0.16
AB-00011412	-59.57	11.96	-47.60	0.64	0.03	0.04	0.65
AB-00011422	-61.76	13.21	-48.55	0.23	0.05	0.03	0.24
AB-00012062	-61.11	8.98	-52.14	0.09	0.03	0.03	0.10
AB-00012187	-70.32	11.27	-59.05	0.15	0.03	0.03	0.16
AB-00014457	-61.81	9.28	-52.52	0.14	0.25	0.03	0.29
AB-00015405	-62.34	10.47	-51.87	0.20	0.09	0.03	0.23
AB-00019891	-62.59	14.61	-47.98	0.57	0.04	0.03	0.57
AB-00022058	-60.03	12.96	-47.07	0.04	0.19	0.03	0.19
AB-00022275	-70.76	10.58	-60.17	0.15	0.17	0.03	0.22

AB-00025673	-56.77	15.24	-41.52	0.42	0.26	0.03	0.49
AB-00026302	-61.63	12.95	-48.68	0.10	0.07	0.03	0.13
AB-00026964	-59.85	12.40	-47.45	0.25	0.70	0.06	0.74
AB-00027532_2_2	-56.72	14.24	-42.48	0.36	0.06	0.04	0.37
AB-00027924	-64.49	11.43	-53.06	0.12	0.01	0.03	0.12
AB-00027945	-56.66	9.01	-47.65	0.07	0.02	0.03	0.08
AB-00029037	-61.88	9.58	-52.29	0.17	0.25	0.02	0.30
AB-00029834	-54.85	6.48	-48.38	0.08	0.01	0.02	0.09
AB-00030020	-61.51	13.77	-47.74	0.30	0.32	0.03	0.44
AB-00030952	-68.62	12.75	-55.87	0.07	0.22	0.03	0.23
AB-00032675	-52.56	14.42	-38.14	0.16	0.05	0.03	0.17
AB-00032757	-62.79	10.57	-52.22	0.08	0.06	0.03	0.11
AB-00033398	-62.98	7.88	-55.10	0.18	0.10	0.03	0.21
AB-00033745	-51.28	12.63	-38.65	0.28	0.11	0.04	0.31
AB-00033828	-55.82	12.87	-42.95	0.29	0.06	0.03	0.29
AB-00034277	-68.62	13.38	-55.24	0.22	0.05	0.03	0.23
AB-00036645	-61.71	12.31	-49.40	0.09	0.04	0.03	0.11
AB-00037163	-61.68	10.18	-51.51	0.08	0.05	0.03	0.10
AB-00038206	-59.46	12.93	-46.53	0.35	0.02	0.03	0.35
AB-00040041	-61.76	18.58	-43.18	0.59	0.90	0.04	1.08

AB-00044057	-58.77	9.29	-49.47	0.29	0.19	0.03	0.35
AB-00044787	-59.53	18.07	-41.46	0.22	0.16	0.04	0.27
AB-00045588	-60.38	16.23	-44.15	0.27	0.02	0.03	0.28
AB-00046158	-57.71	15.24	-42.47	0.23	1.03	0.04	1.05
AB-00046578	-67.16	12.88	-54.28	0.06	0.24	0.03	0.25
AB-00046961	-61.83	19.33	-42.51	0.55	0.18	0.03	0.58
AB-00047567	-64.48	20.92	-43.56	0.17	0.95	0.03	0.96
AB-00048033	-53.05	12.75	-40.30	0.14	0.22	0.04	0.26
AB-00048113	-67.93	5.02	-62.91	0.07	0.17	0.03	0.19
AB-00051118	-61.05	21.68	-39.38	0.29	0.09	0.03	0.30
AB-00054907	-58.94	15.34	-43.60	0.19	0.30	0.04	0.36
AB-00055646	-57.61	12.18	-45.43	0.15	0.33	0.03	0.36
AB-00056410	-53.87	5.74	-48.13	0.06	0.11	0.03	0.13
AB-00057060	-61.61	11.24	-50.37	0.12	0.29	0.03	0.31
AB-00057869	-55.07	12.46	-42.61	0.58	0.19	0.04	0.61
AB-00058265	-67.82	18.02	-49.80	0.20	0.24	0.03	0.32
AB-00058994	-54.03	9.01	-45.03	0.21	0.03	0.04	0.22
AB-00060989	-59.81	13.07	-46.74	0.27	0.27	0.03	0.39
AB-00062036	-64.78	16.09	-48.69	0.85	0.05	0.03	0.85
CK144	-65.41	14.88	-50.53	0.13	0.61	0.02	0.62

CK571	-68.27	15.20	-53.07	0.10	0.28	0.03	0.29
CK903	-68.33	21.87	-46.46	0.19	1.27	0.03	1.28

*All values are shown in kcal/mol

Annex S6.2: Physicochemical properties of the 60 compounds that were prioritized from CN.

Complex	MW	cLogP	cLogS	H-Acceptor	H-Donor	PSA	Rotatable Bonds
AB-00002226	388.49	1.84	-2.60	6	1	84.1	8
AB-00002083	428.39	3.08	-4.80	9	0	106.6	8
AB-00002670	459.50	2.37	-4.75	8	0	79.4	7
AB-00004558	430.48	3.13	-7.08	6	0	84.8	5
AB-00004786	464.53	3.77	-4.14	6	2	109.1	7
AB-00007072	428.47	1.50	-3.99	9	2	127.8	7
AB-00010786	386.52	1.29	-3.01	6	1	87.0	6
AB-00011412	410.43	2.52	-3.85	8	1	102.0	10
AB-00011422	423.49	2.95	-5.03	7	1	90.8	5
AB-00012062	409.44	1.76	-3.80	8	3	109.5	10
AB-00012187	472.52	1.98	-6.19	7	0	94.1	5
AB-00014457	446.59	3.25	-3.16	6	0	91.5	3

AB-00014827	475.50	2.73	-4.62	9	1	99.2	5
AB-00015405	396.50	3.68	-5.77	6	2	75.6	7
AB-00019891	413.47	3.76	-3.68	7	0	74.3	9
AB-00020709	415.48	2.97	-3.04	7	0	82.1	10
AB-00022058	444.58	3.43	-4.06	6	2	109.1	6
AB-00022275	495.37	3.47	-4.66	6	1	77.5	7
AB-00025673	440.54	3.88	-4.35	6	2	134.8	9
AB-00026302	445.50	2.97	-7.39	8	2	135.3	7
AB-00026964	443.59	2.31	-4.23	7	4	99.7	0
AB-00027532	476.53	3.48	-4.88	8	2	113.0	6
AB-00027924	422.48	3.49	-4.36	7	0	68.3	4
AB-00027945	418.46	1.83	-5.16	9	2	108.7	4
AB-00029037	385.47	3.57	-4.67	5	1	78.1	7
AB-00029834	392.48	3.14	-6.08	6	0	70.5	3

AB-00030020	458.61	3.35	-7.39	7	2	121.3	9
AB-00030952	411.53	2.83	-6.18	7	2	90.0	5
AB-00032675	420.50	1.85	-3.07	8	2	97.3	7
AB-00032757	406.51	2.32	-6.49	6	3	107.2	7
AB-00033398	434.57	3.19	-4.26	7	0	84.1	7
AB-00033745	404.49	3.94	-5.17	7	2	116.3	5
AB-00033828	416.50	3.91	-5.27	7	2	124.2	6
AB-00034277	425.55	3.09	-5.37	6	0	75.2	7
AB-00036645	452.00	3.41	-5.27	6	2	107.3	4
AB-00037163	415.53	2.74	-3.48	5	1	80.9	5
AB-00038206	409.48	1.94	-2.92	7	1	71.1	6
AB-00040041	448.52	3.34	-6.02	8	4	116.3	10
AB-00044057	418.45	2.98	-4.47	7	1	84.9	7
AB-00044787	418.50	3.51	-4.77	7	1	70.4	8

AB-00045588	432.55	1.12	-4.12	7	2	63.8	4
AB-00046158	393.53	3.90	-4.09	6	2	75.9	7
AB-00046578	439.58	3.89	-6.25	7	1	101.6	6
AB-00046961	396.49	2.61	-3.20	8	2	94.0	8
AB-00047567	453.58	3.17	-3.85	7	2	87.7	9
AB-00048033	399.51	1.94	-5.71	6	1	79.8	7
AB-00048113	409.49	2.57	-3.69	6	2	73.4	5
AB-00051118	460.48	2.86	-4.79	9	0	98.8	10
AB-00054907	434.50	3.19	-7.07	9	3	110.3	7
AB-00055646	452.53	3.66	-3.76	7	1	102.4	5
AB-00056410	412.28	3.12	-3.57	6	1	79.7	5
AB-00057060	408.49	2.52	-6.92	8	3	121.0	6
AB-00057869	430.51	3.73	-3.82	8	0	79.9	6
AB-00058265	435.48	1.65	-3.45	9	2	107.6	7

AB-00058994	496.34	2.32	-5.55	8	0	125.9	5
AB-00060989	428.50	2.47	-7.60	10	2	139.2	7
AB-00062036	469.59	3.77	-5.58	7	2	140.2	8
CK571	504.94	3.04	-5.88	9	4	124.0	10
CK144	499.46	0.06	-5.88	7	2	95.2	10
CK903	482.49	0.80	-3.74	9	3	112.2	11

PSA: Polar Surface Area, defined as the sum over all polar atoms and their attached hydrogens. Molecules with a PSA above 140Å are typically not good at permeating the cell wall.

cLogP: Calculated octanol/water partition coefficient. High logP means low adsorption and thus low permeation into membranes. A maximum value of 5 is commonly accepted, and molecules with cLogP < 5 are typically membrane permeable.

cLogS: Calculated aqueous solubility. Low solubility is typically correlated with bad adsorption, so the aim is avoiding insoluble compounds. According to DataWarrior, 80% of marketed drugs have a logS > -4.

MW: Molecular weight of the compound.

Annex S6.3: Distribution plots of the physicochemical properties of the 2300 compounds from CN selected for VS.

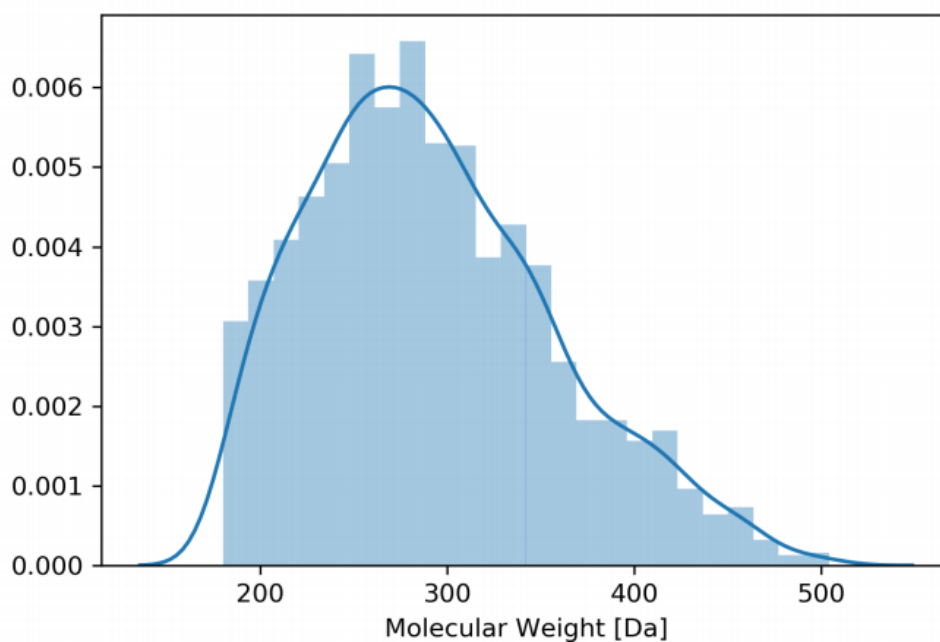


Figure S6.3 – Distribution plot of the molecular weight of the 2300 compounds from CN screened by *ChemFlow*

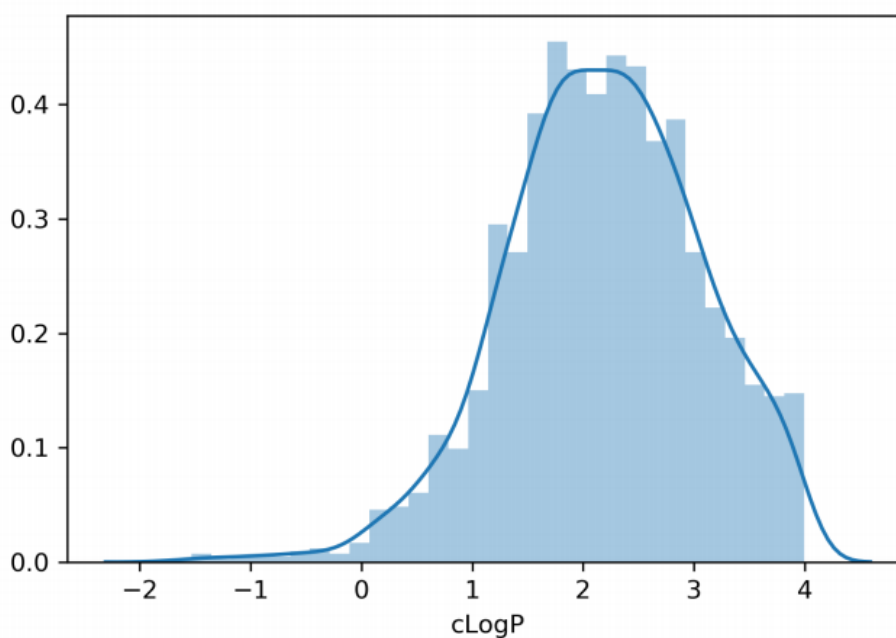


Figure S6.4 – Distribution plot of the calculated logP of the 2300 compounds from CN screened by *ChemFlow*

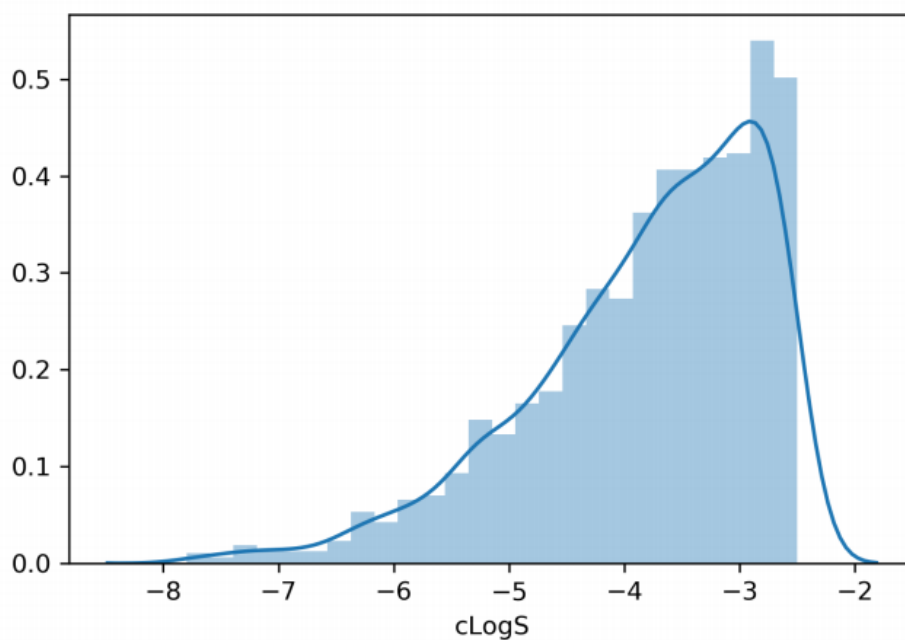


Figure S6.5 – Distribution plot of the calculated logS of the 2300 compounds from CN screened by *ChemFlow*

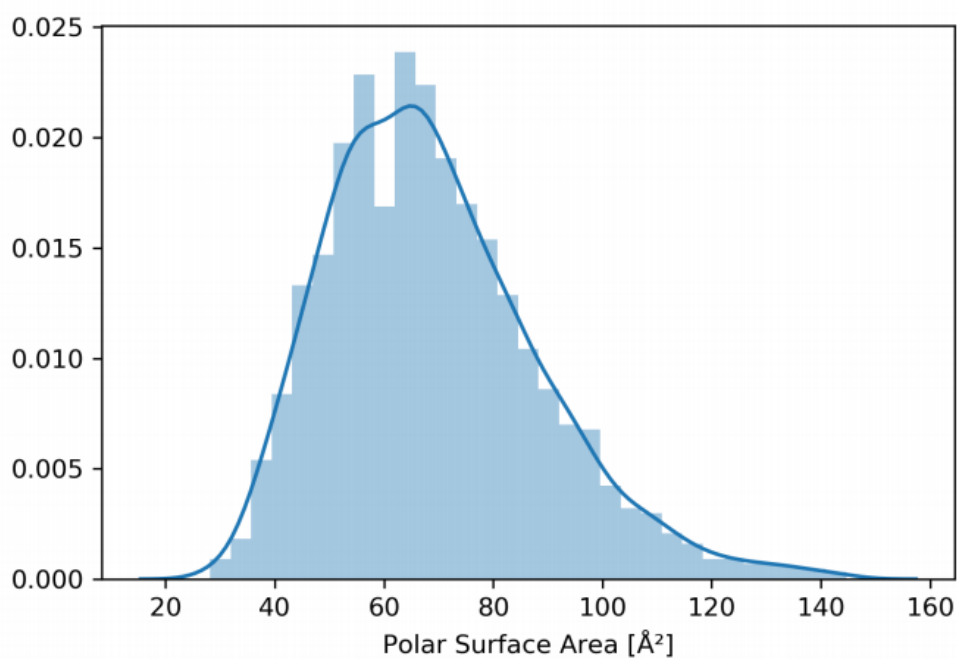


Figure S6.6 – Distribution plot of the calculated Polar Surface Area of the 2300 compounds from CN screened by *ChemFlow*.

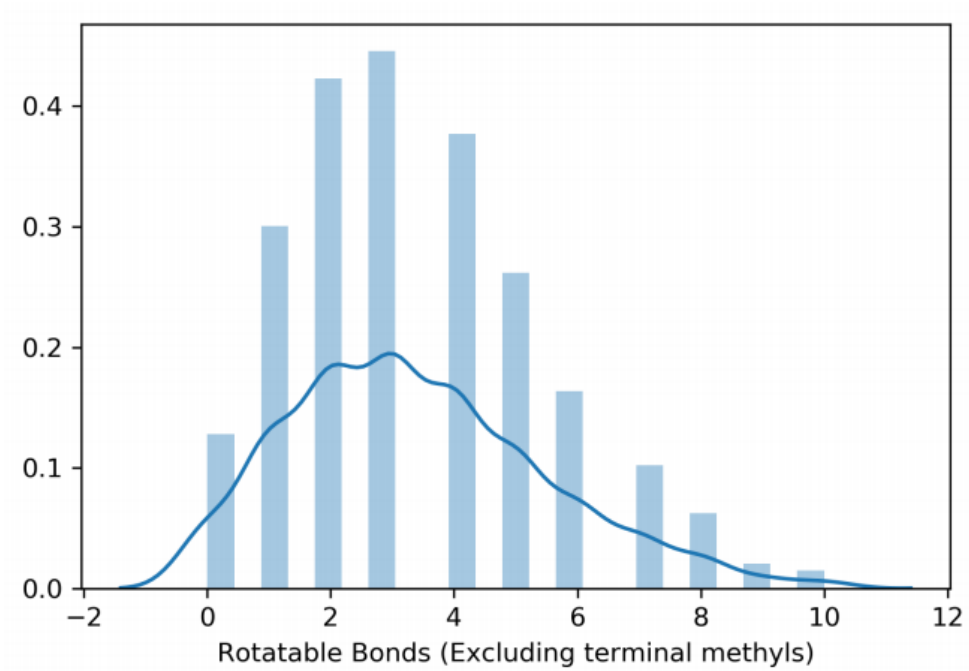


Figure S6.7 – Distribution plot of the number of rotatable bonds of the 2300 compounds from CN screened by *ChemFlow*