Computational exploration of host-guest complexes

Dylan Serillon

UNIVERSITÉ DE STRASBOURG

ÉCOLE DOCTORALE DES SCIENCES CHIMIQUES

Laboratoire de Synthèse des Assemblages Moléculaires Multifonctionnels
Institut de Chimie

THÈSE présentée par :

Dylan SERILLON

soutenue le : **10 Décembre 2021**

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

Discipline/ Spécialité : Chimie / Chimie informatique et théorique

Computational exploration of host-guest complexes

THÈSE dirigée par :

Mme. HEITZ Valérie	Professeur, université de Strasbourg
M. BARRIL Xavier	Professeur, université de Barcelone

RAPPORTEURS :

M. LE QUESTEL Jean-Yves	Professeur, université de Nantes
M. ZANUY GOMARA David	Professeur, université polytechnique de Catalogne

---

AUTRES MEMBRES DU JURY :

M. CURUTCHET Carles	Professeur, université de Barcelone
M. ROGNAN Didier	Directeur de recherche, université de Strasbourg

ACKNOWLEDGEMENTS

First of all, I would like to give infinite thanks to my supervisor Dr. Xavier BARRIL. Thank you so much for being so helpful, for your motivation and for your friendship. During this thesis and the years spent in Barcelona, I was able to find in your research team a place where I could develop myself both humanly and scientifically. Thank you for all the discussions we have had over the years, I feel that I have always grown from them. I am also extremely grateful for your involvement in my arrival at the University of Barcelona after many ups and downs, and I have not forgotten how it has greatly improved my daily life. Thank you very much.

Many thanks to Dr. Carles BO, who agreed to host me in his laboratory in Tarragona for a few months at the beginning of the thesis. Thank you very much for the discussions and the private courses on quantum mechanics. And also, for the discovery of the software that allowed me to finish this thesis.

In Barcelona, I found a human and scientific environment that pushed me to surpass myself daily. This is certainly due to the impact of Xavier, who, in my opinion, has created an environment of unprecedented warmth, but also to the immense impact of each of my colleagues on a daily basis. Who have made, are making and will continue to make this team grow. Your welcome, your kindness, our discussions in the lab or over a drink or at a Halloween party will remain engraved in my memory. Thanks to you: Alvaro, Marina, Moira, Serana, Salvatore, Miriam, Andrea, Beste, Dani, Roger, Ania, Maciej, Carles C. and Carles G.

I would like to thank the jury members for having accepted to participate in this thesis: Dr. Jean-yves LE QUESTEL, Dr. Didier ROGNAN, Dr. Carles CURUTCHET, Dr. David ZANUY GOMARA. It’s a great pleasure from my side to present to you the work done during the past three years.

I would like also to thank all the different partners involved in the NOAH project, both the ESRs and the supervisors, who participated in some way in the realisation of this thesis.

Je tiens également à remercier chaleureusement mon second superviseur : le Dr. Valérie HEITZ, qui a toujours été présente dans ce projet, tant dans les moments difficiles que dans les moments plus paisibles. Merci de ton accueil lors de mon arrivée à Strasbourg. Merci également pour ta patience et tous les conseils avisés.

Henri-Pierre, merci de ton engagement au quotidien dans mon projet lors de mon arrivée à Strasbourg, l'apprentissage des bonnes pratiques de laboratoire ne fut pas toujours aisé, mais tu as toujours su trouver les mots. Merci également pour toutes nos discussions scientifiques et pour le temps considérable passé à me montrer et m'expliquer tout ce qui semble logique pour un chimiste de synthèse.

Merci également à l'accueil de mes collègues strasbourgeois, qui bien que souvent taquin m'ont parfaitement accepté parmi eux ! Merci à vous tous pour nos diverses discussions, pour les repas et les soirées partagées, et pour la bonne humeur et la gentillesse dont vous avez fait preuve envers l'apprenti chimiste que je suis. Merci Charly, Vincent, Sonia, Johnny, Alice, Amy, Etienne, Alexis, Mathieu.

Merci également à vous tous, famille, et amis qui avez pu de près ou de loin participer à cette thèse au cours de ces trois dernières années, Stéphanie, Marie M., Guy, Yvonne pour toutes les fois où j'ai pu vous parler de ce que je faisais. Mes amis pharmaciens et autres, avec qui nous partageons un amour commun de la science : Thomas Y., Jeremy, Charles, Selma, Sébastien, Juliette D., Thomas M., Delphine, Mathilde, Nadia, Jessy, Maud, Nolwenn, Juliette G., Anthony et bien d'autres encore.

Un immense merci à mes parents, vous qui année après année me supportez et me soutenez dans la réalisation de mes projets personnels et professionnels. Je n'en serais pas arrivé là sans vous. Votre soutien indéfectible restera toujours une force fantastique qui m'a conduit là où j'en suis aujourd'hui. Merci Gwenvaël, Merci Remy, Merci Maël.

Enfin merci à toi Marie. Toi qui partages ma vie depuis de nombreuses années à présent. Toi que j'ai emmené à Paris, à Barcelone, à Strasbourg, et avec qui je compte bien partir au bout du monde. Cette thèse n'aura pas été de tout repos, on se souviendra avec humour dans quelques années je l'espère du début mouvementé de ce projet ! Sans toi cette thèse n'aurait pas été pareil, pour ton soutien, tes encouragements, et pour tout le reste merci à toi. Du fond du cœur.

0

RESUME DE THESE

I - INTRODUCTION

I. A - ETAT DE L'ART

La chimie supramoléculaire a connu un énorme essor ces dernières années. Les processus supramoléculaires et, en particulier, les interactions hôte-invité sont étudiées pour leurs implications possibles dans une variété d'applications (vectorisation de molécule d'intérêt, catalyse, modulation des propriétés physicochimiques…). En améliorant la stabilité ou en modifiant les propriétés d'un composé encapsulé, ou même en augmentant la sélectivité réactionnelle, nous prévoyons un large éventail d'applications qui s'étendent des processus industriels au domaine médical.

Actuellement, les progrès en chimie supramoléculaire hôte-invité sont entravés par la complexité de la caractérisation thermodynamique et cinétique des processus d'inclusion/libération, ce qui rend difficile la génération de prédictions utiles sur l'encapsulation moléculaire. La prédiction quantitative des énergies d'interaction est particulièrement difficile mais fondamentale car elle est associée à une perte de temps et d'argent due à l'effort de synthèse et d'essai d'invités non-actifs.

Dans ce contexte, ce projet de thèse se concentre sur le développement d'un outil de prédiction *in silico* pour trouver des invités appropriés pour différents hôtes qui seront développés par les partenaires du projet européen NOAH (Network Of Functional Molecular Containers With Controlled Switchable Abilities H2020-MSCA-ITN-2017).

De cette façon, nous espérons non seulement améliorer les connaissances globales dans le domaine de la chimie supramoléculaire, mais aussi fournir de nouvelles opportunités et applications pour les invités déjà existants et fournir un supplément d'information dans le développement rationnel d'invités porteurs de nouvelles activités.

I. B - BUT GENERAL DU PROJET

Au cours de cette thèse, nous avons comparé plusieurs méthodes pour pouvoir calculer l'enthalpie libre de différents systèmes. Nous avons comparé le calcul DFT à une nouvelle méthode semi-empirique appelée GFN2-xTB qui sera finalement utilisée pour le calcul des paramètres thermodynamique.

Les différents outils informatiques (Figure 0. 1) bien qu'ils soient séparés ici, sont complémentaires et ont été développés simultanément au cours du projet. L'outil de criblage virtuel est utilisé pour paramétrer l'hôte et l'invité, et pour générer le mode d'interaction entre l'hôte et les invités, nécéssaires pour toute analyse ultérieure. Un protocole de simulation est ensuite utilisé afin de prédire les constantes de liaison des complexes dans différents solvants et évaluer le comportement de ces complexes sous différents stimuli. Toutes les énergies de liaison calculées par notre approche et toutes les valeurs expérimentales qui seront fournies par les partenaires du projet seront stockées afin d'être comparées, affinées et servir de base pour une approche faisant intervenir des algorithmes d'intelligence artificielle : « *Machine Learning* ».

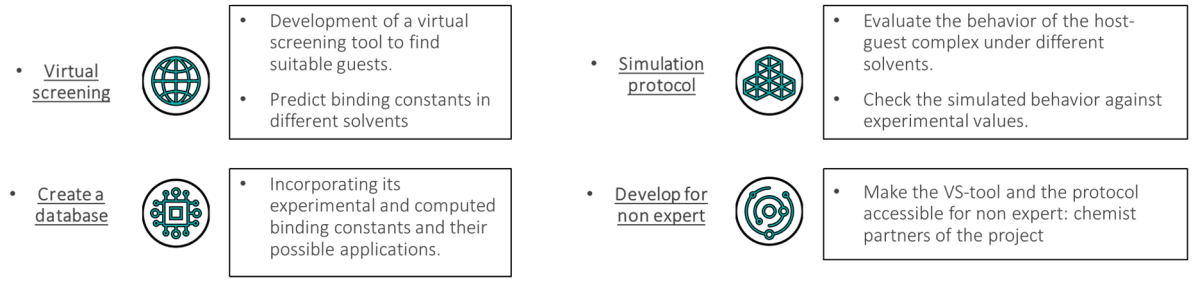


Figure 0. 1 : objectif général du projet : une approche multiple pour prédire le comportement dynamique des systèmes hôte-invité

Nous nous sommes concentrés sur les différents outils informatiques développés, de façon à ce que ceux-ci soient compréhensibles et utilisables par des utilisateurs non-expert, qui sont les utilisateurs attendus de la plateforme.

Durant les dernières années, plusieurs instances du défi SAMPL (Statistical Assessment of the Modeling of Proteins and Ligands) nous ont montré des approches intéressantes pour calculer l'enthalpie libre de liaison des complexes hôte-invité, avec une gamme relativement large de méthodes et de performances. Hélas, aucun des composés étudiés au cours des challenges SAMPL ne concernent des molécules contenant des métaux. Néanmoins, les diverses informations extraites des défis précédents, nous ont permis de développer une méthode

générale de prédiction d'énergie libre de Gibbs, pouvant être utilisée pour la prédiction de molécules possédant des métaux dans leurs structures. Les nouvelles instances du défi ont également pu être utilisé au cours de la thèse de façon à valider nos méthodes à l'aveugle sur un large éventail de système différent, pour lesquels la prédiction d'énergie libre de Gibbs est particulièrement difficile.

Dans le contexte du réseau européen NOAH, notre intérêt a été de développer des méthodes automatisées pour la prédiction du comportement de systèmes hôte-invité dans différentes conditions de solvatation et autres conditions expérimentales pour répondre aux besoins des différents partenaires du projet dans lequel s'inscrit la thèse.

À terme, les utilisateurs de la plateforme sont supposés pouvoir générer des prédictions pour un large éventail de systèmes (décrits ou innovants) sans pour autant avoir une formation en chimie computationnelle. Ce pour quoi notre plateforme se veut automatique, performante et précise : elle doit être capable de fournir des résultats rapidement et de manière fiable, tout en ne nécessitant pour son utilisation que des informations connues des utilisateurs (charge ionique, type et quantité d'ion métallique, modèle de solvatation utilisé…).

II - METHODES

II. A - VUE D'ENSEMBLE DES METHODES UTILISEES

De façon à mesurer notre capacité à prédire les énergies de liaison des composés hôte-invité, deux méthodologies différentes ont été développés lors de cette thèse. Dans les deux cas, le résultat rendu est une valeur numérique correspondant à l'énergie de liaison prédite pour un système hôte-invité.

La figure ci-dessous (Figure 0. 2) résume les deux différentes méthods et leurs spécificités : en bleu à gauche, la méthode basée sur les connaissances, et en rouge à droite la méthode basée sur la détermination des paramètres thermodynamiques.

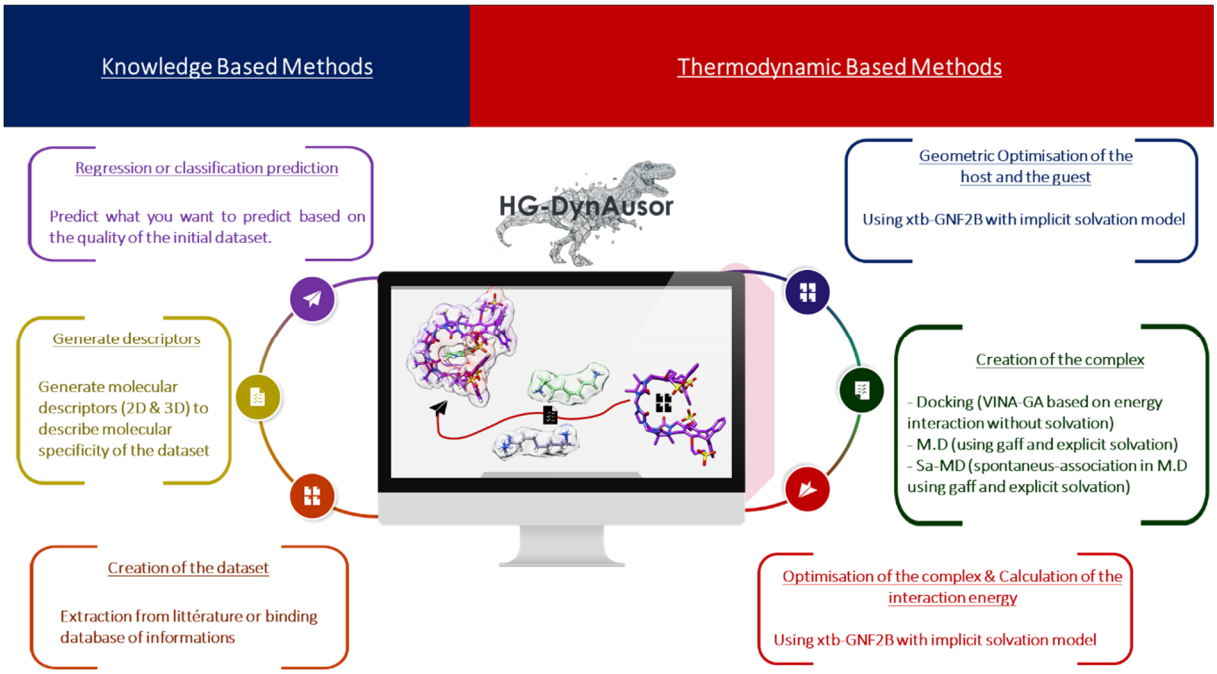


Figure 0. 2 : présentation des deux méthodes utilisées pour le calcul d'enthalpie libre (en bleu la méthode basée sur l'apprentissage automatisé, et en rouge la méthode basée sur la thermodynamique)

II. B - LA METHODE PHYSIQUE OU THERMODYNAMIQUE

La méthode physique ou thermodynamique (en rouge sur la Figure 0. 2) est une méthode basée sur les conformations 3Ds des complexes hôte-invités, qui utilise une méthode semi-empirique pour le calcul de l'enthalpie libre. Celle-ci est calculée à partir de la différence des énergies libres du complexe, de l'hôte et des molécules invitées. A la fois les termes enthalpiques et entropiques sont considérés pour calculer un ΔG^0 numérique.

Les énergies libres de Gibbs (ΔG^0) des géométries optimisées ont été calculées comme étant la somme de l'énergie électronique (E), qui comprend la correction de dispersion D4, les

corrections thermostatiques (G_{RRHOT}) calculées selon une approche couplée à un oscillateur harmonique à rotor rigide, et la contribution de solvatation (G_{solv}) calculée par le modèle de solvatation implicite *GBSA*, tel que :

$$\Delta G = E + G_{RRHOT} + G_{solv} \quad 0.1$$

Avec :

$$\Delta G_{solv} = \Delta G_{born} + \Delta G_{sasa} + \Delta G_{hb} + \Delta G_{shift} \quad 0.2$$

Ainsi, l'énergie libre de Gibbs est calculée à partir de la différence des énergies libres du complexe, de l'hôte et des molécules invitées, tel que :

$$\Delta G_{bind} = \Delta G_{complex} - \Delta G_{host} - \Delta G_{guest} \quad 0.3$$

Compte tenu de la complexité du paysage énergétique conformationnel du complexe et de la molécule hôte, différentes géométries du système sont utilisées comme points de départ pour la minimisation, augmentant ainsi la probabilité de trouver le minimum absolu. Pour cela, de multiples structures sont extraites des simulations classiques de dynamique moléculaire pour effectuer une optimisation géométrique à un niveau semi-empirique, suivie d'un calcul de la matrice hessienne pour confirmer que l'énergie finale est un véritable minimum (c'est-à-dire que toutes les fréquences vibrationnelles sont positives). Bien que les degrés de liberté de l'invité soient beaucoup plus réduits, nous utilisons pour celui-ci un protocole similaire par souci de cohérence.

II. C - LA METHODE BASEE SUR LES CONNAISSANCES

La méthode basée sur les algorithmes d'apprentissage automatisés (en bleu sur la Figure 0. 2), utilise les données disponibles extraites de différentes bases de données ou de la littérature, et utilise des algorithmes d'apprentissage automatique afin de prédire l'enthalpie libre pour un complexe hôte-invité donné. Pour cela, elle utilisera les descripteurs moléculaires fournis par les utilisateurs pour apprendre des données préexistantes afin de prédire l'énergie libre de liaison pour d'autres complexes utilisant les mêmes descripteurs moléculaires.

Le modèle ayant donné les meilleurs résultats se trouve être le réseau de neurone. L'algorithme de réseau neuronal est un algorithme qui appartient à la classe de l'apprentissage profond. L'apprentissage profond est un ensemble de méthodes d'apprentissage qui vont utiliser des

transformations non linéaires pour modéliser un ensemble de données avec des architectures complexes. Le modèle le plus simple de l'apprentissage profond est constitué par les réseaux neuronaux qui sont combinés pour former le réseau neuronal profond. Il existe de multiples architectures de réseaux neuronaux, les perceptions multicouches étant les plus simples, ce sont celles que nous avons utilisées dans cette thèse (Figure 0. 3).

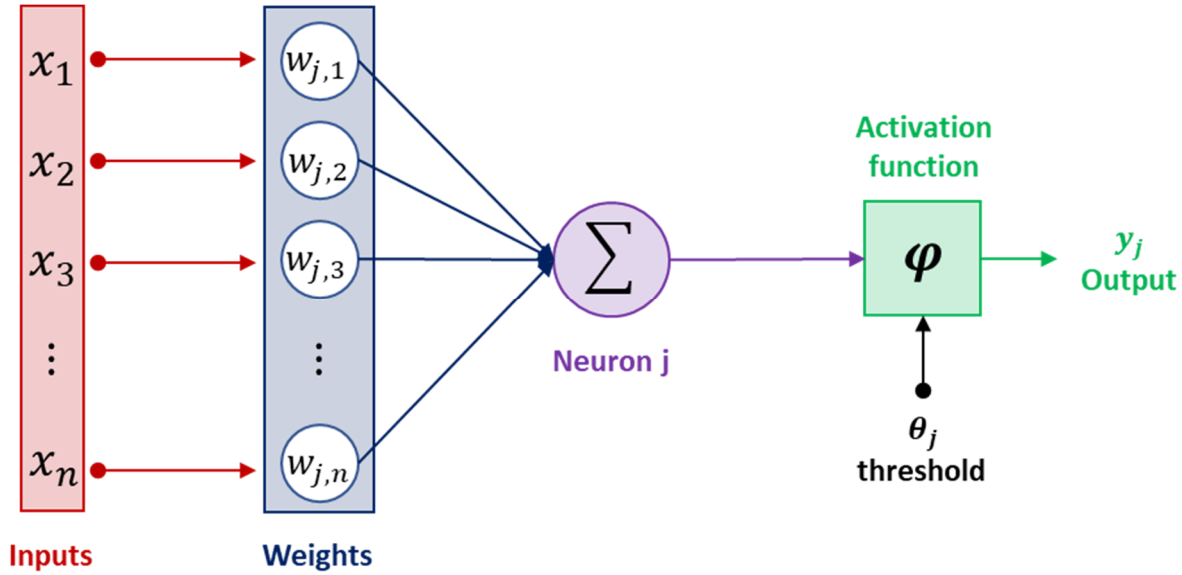


Figure 0. 3: représentation schématique d'un neurone artificiel ou $\sum = \langle w_j, x \rangle + net_j$

Mathématiquement parlant, un réseau de neurones artificiels est une application non linéaire présentant un paramètre θ qui associe à une entrée x, une sortie y telle que $y = f(x; \theta)$. Ceci permet la prédiction directe de y, dans notre cas : la prédiction de valeur numérique correspondant aux énergies libres de Gibbs (ΔG^0) de complexe hôte-invité.

III - LA PLATEFORME HG-DYNAUSOR

III. A - VUE D'ENSEMBLE DE LA PLATEFORME

La plateforme HG-DYNAusor (Host-Guest DYNamic an Automated application) est une application automatisée de manière à faciliter l'exécution de tâches courantes dans la modélisation des complexes hôte-invité, y compris l'utilisation de différentes méthodes pour la détermination de la géométrie, de la dynamique, et de l'énergie des complexes supramoléculaire, ainsi que leurs composants individuels. Une vue d'ensemble de la plateforme HG-DYNAusor est présenté dans la Figure 0. 4 suivante :

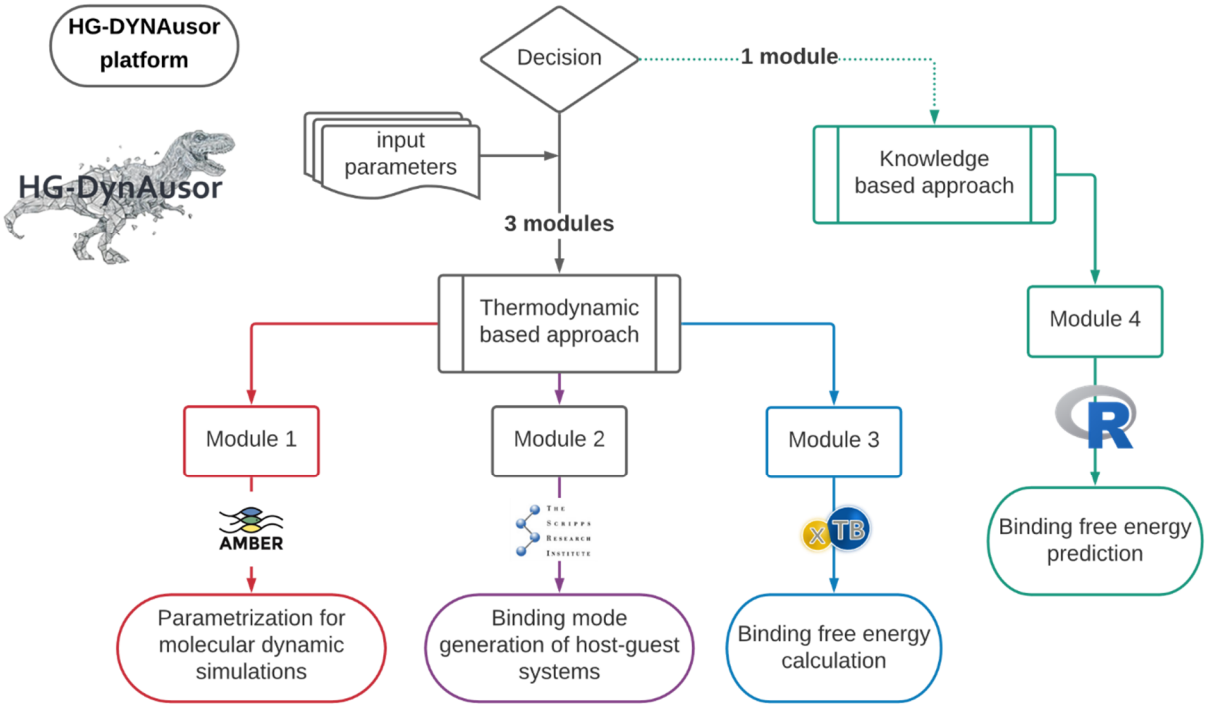


Figure 0. 4 : vue d'ensemble de la plateforme HG-DYNAusor ; les trois premiers modules font partie de la l'approche basée sur la thermodynamique : le premier module est dédié à la génération des paramètres, le second à la prédiction du mode de liaison, alors que le troisième module est utilisé pour la prédiction *stricto sensu* de l'énergie libre de Gibbs, alors que le dernier module est dédié à l'approche basée sur les connaissances (en vert)

Cette plateforme, qui a été conçue et développée au cours de la thèse, est d'ores et déjà opérationnelle bien qu'elle soit toujours en cours de développement. Elle a pu être utilisée pour l'étude de plusieurs complexes hôte-invité qui seront présentés dans les chapitres suivants. La plateforme peut être séparée en quatre modules différents : trois sont dédiés au calcul de l'énergie libre de liaison en utilisant l'approche basée sur la thermodynamique. Un quatrième module est dédié à la prédiction de l'énergie libre de liaison en utilisant une approche basée sur

les connaissances. Le principe de chacun des modèles est présenté brièvement dans les parties suivantes.

III. B - GENERATION DES PARAMETRES POUR LES SIMULATIONS DE DYNAMIQUE MOLECULAIRE

Le module 01 de la plateforme, dédiée à la génération des paramètres pour l'exécution de simulation de dynamique moléculaire, est présentée dans la Figure 0. 5 suivante :

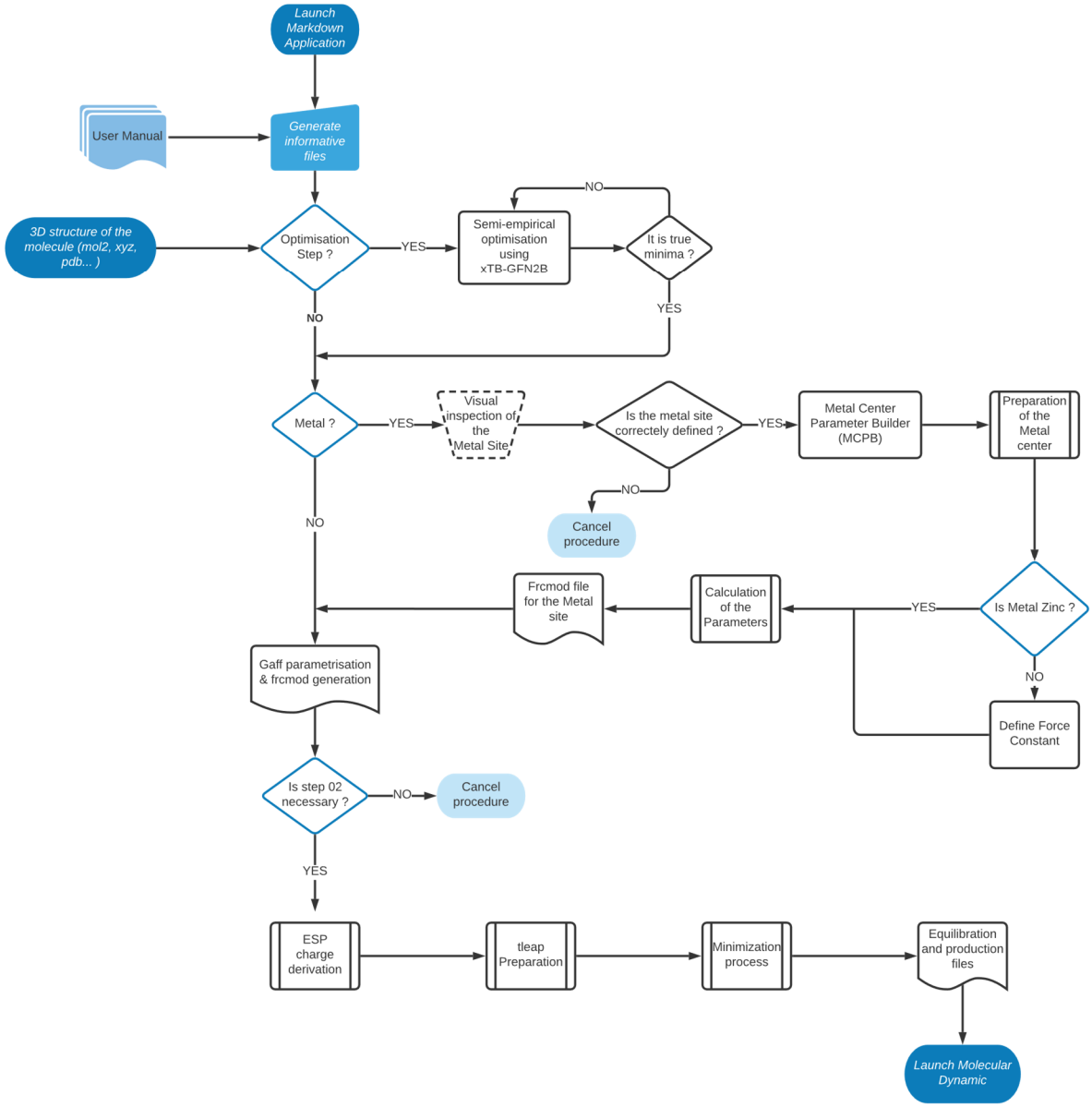


Figure 0. 5 : vue d'ensemble du module 01 de la plateforme HG-DYNAusor dédiée à la génération des paramètres pour les simulations de dynamique moléculaire.

Ce premier module, utilise un fichier d'entrée contenant les informations nécessaires à l'application concernant le système à simuler (charge ionique, nom du fichier, type de métal…). Le premier module de la plateforme HG-DYNAusor peut être séparé en six différentes parties :

1) Génération d'une conformation 3D raisonnable du système hôte, qui sera utilisée dans toutes les étapes suivantes. Cette étape peut être évitée dans le cas où la structure moléculaire considérée provient d'une structure cristalline bien résolue.

2) Paramétrisation du centre métallique. Une fois encore, cette étape est optionnelle si le système hôte ne contient pas de métal.

3) Calcul des charges partielles pour le système d'intérêt en utilisant une approche que nous avons développée basée sur la génération d'une base de données de charges partielles.

4) Génération des fichiers topologiques pour le système considéré. Les fichiers topologiques contiennent tous les paramètres nécessaires à la réalisation des simulations moléculaires.

5) Minimisation du système avant la dynamique moléculaire.

6) Création des fichiers d'équilibrage et de production.

III. C - GENERATION DU MODE DE LIAISON POUR LES SYSTEMES HOTES INVITES

Le second module de la plateforme HG-DYNAusor est dédié à la génération du mode de liaison du complexe hôte-invité. Il existe différentes méthodes de génération du mode de liaison qui dépendent principalement des données existantes sur l'hôte et sa structure moléculaire.

Dans notre application, l'étude du mode de liaison est construite comme un module supplémentaire après le paramétrage du système hôte. Trois approches différentes peuvent être envisagées : (i) Soit le mode de liaison est connu et peut être directement extrait d'une structure cristallographique, (ii) soit le mode de liaison n'est pas connu mais l'hôte est dans une conformation adaptée à la liaison, (iii) l'hôte a une mobilité intrinsèque très élevée et donc une approche de liaison spontanée peut être considérée. Il est admis que l'étape la plus limitante de notre analyse pour les calculs d'énergie libre de Gibbs consiste en la génération du mode de

liaison pour le système hote-invité. Une vue d'ensemble du module 02 de la plateforme HG-DYNAusor est présentée dans la Figure 0. 6 suivante :

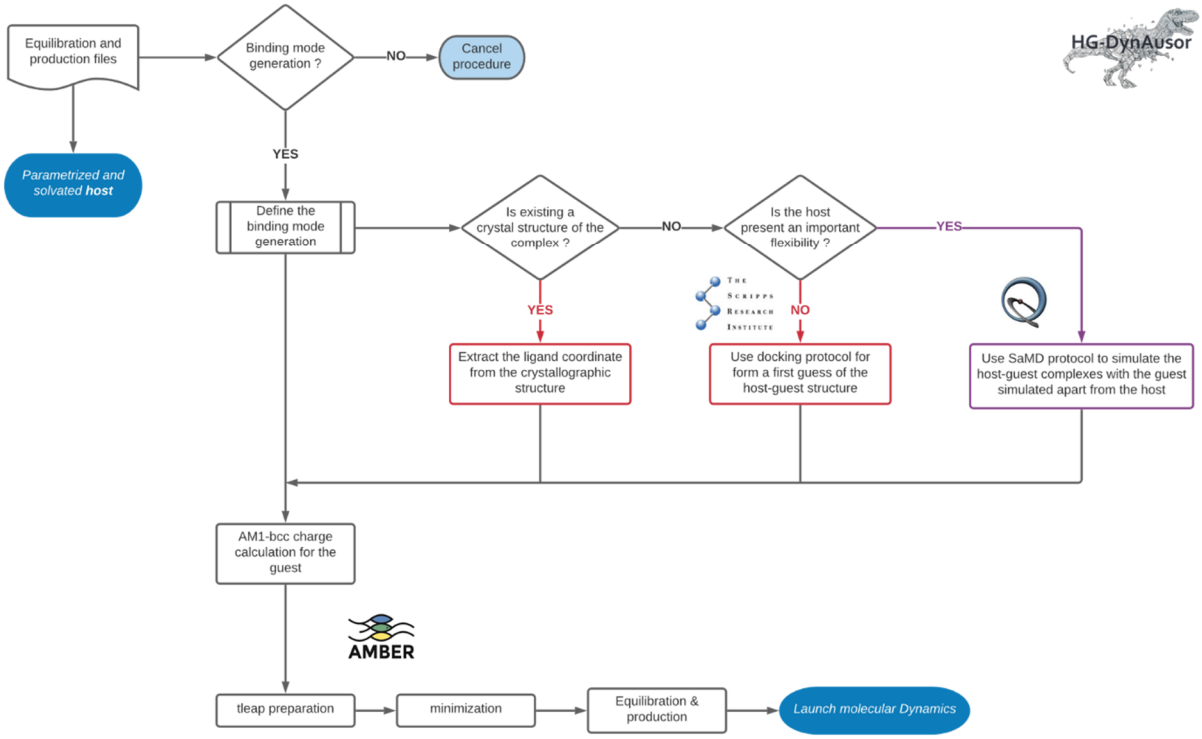


Figure 0. 6 : vue d'ensemble du module 02 de la plateforme HG-DYNAusor dédié à la détermination du mode de liaison

III. D - CALCUL DES ENERGIES LIBRES DE GIBBS PAR LA METHODE THERMODYNAMIQUE

Le module 03 de la plateforme dédiée au calcul des énergies libre de Gibbs, est présenté dans la Figure 0. 7 suivante :

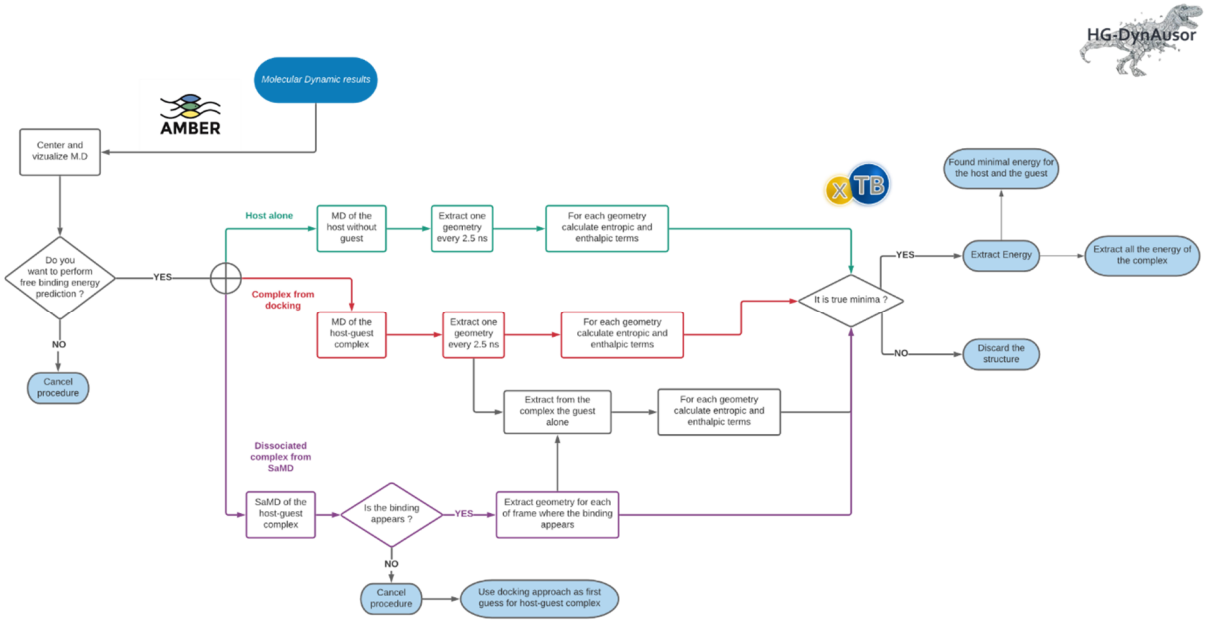


Figure 0. 7 : vue d'ensemble du module 03 de la plateforme HG-DYNAusor dédiée au calcul des énergies libres de Gibbs

Le troisième module réalise des prédictions d'énergie libre de Gibbs en utilisant les résultats des simulations de dynamique moléculaire. Il dépend entièrement des deux modules précédents de la plateforme car il nécessite : une simulation de dynamique moléculaire de l'hôte seul et une dynamique moléculaire du complexe. Si les utilisateurs veulent prédire plusieurs composés sur le même récepteur, une seule simulation de l'hôte est nécessaire, mais chaque système hôte-invité doit être simuler indépendamment.

III. E - PREDICTION DE L'ENERGIE LIBRE DE GIBBS PAR LA METHODE BASEE SUR LES CONNAISSANCES

Le dernier module de la plateform HG-DYNAusor est dédié à la prédiction de l'énergie libre de liaison en utilisant une approche basée sur les connaissances. Pour cela, des données concernant de nombreux système hôte-invité sont extraites de la BindingDataBase (BindingDB). Ces données sont ensuite traitées, et les informations concernant l'hôte et l'invité sont considérées séparément. Environ 200 descripteurs moléculaires décrivant à la fois les

paramètres 2D et 3D des hôtes et des invités sont calculés pour chacun d'entre eux, et les deux jeux de données sont ainsi formés : un décrivant les systèmes hôtes, et un décrivant les invités.

Après une analyse dimensionnelle et une réduction du nombre de descripteurs pour chacune de ses tables, chaque invité est associé à l'hôte correspondant, et l'information concernant l'activité (énergie libre de Gibbs) associée est sauvegardée dans le jeu de données finales. Finalement, notre jeu de données finales prend en compte à la fois les informations moléculaires des invités, mais également des hôtes avec lesquelles ils interagissent. Ainsi deux invités interagissant avec deux hôtes différents et présentant donc deux valeurs différentes d'énergie libre de Gibbs, vont être considérés différement par l'algorithme d'apprentissage automatique.

Ce principe est présenté dans la Figure 0. 8 suivante :

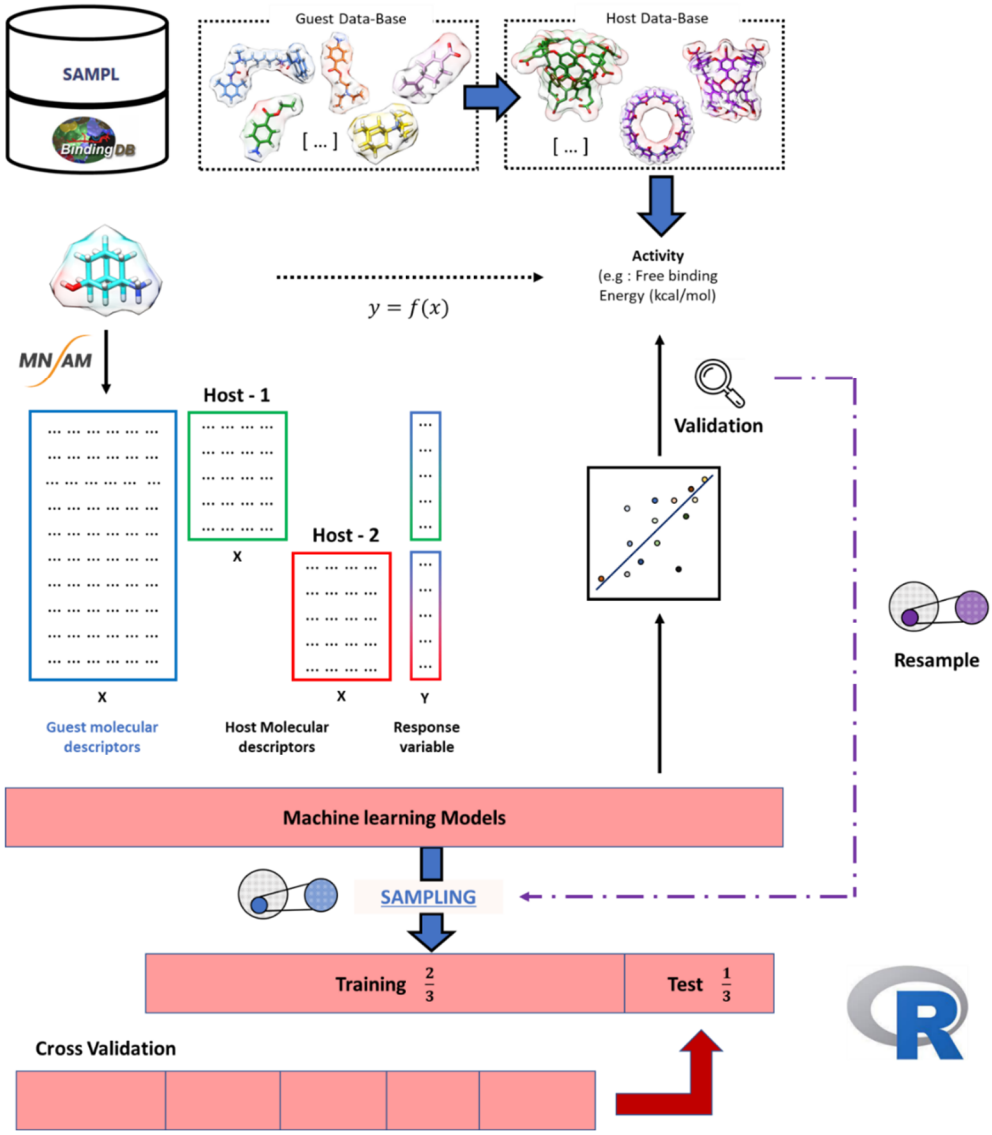


Figure 0. 8 : principe de fonctionnement du module 04 de la plateforme HG-DYNAusor dédiée à la prédiction des énergies libres de Gibbs

IV - EXEMPLE D'UTILISATION DE LA PLATEFORME : LE SAMPL7 CHALLENGE

IV. A - ÉCHANTILLONAGE CONFORMATIONNEL

Compte tenu de la complexité du paysage énergétique conformationnel du complexe et de la molécule hôte, nous avons utilisé plusieurs géométries du système hôte libre comme point de départ de la minimisation, augmentant ainsi la probabilité de trouver le minimum absolu. Pour ce faire, le protocole présenté dans la Figure 0. 9 est utilisé : nous extrayons environ 15 structures des simulations classiques de dynamique moléculaire et effectuons une optimisation géométrique à un niveau semi-empirique, suivie d'un calcul de la matrice hessienne pour confirmer que l'énergie finale est un véritable minimum (c'est-à-dire que toutes les fréquences vibrationnelles soient positives).

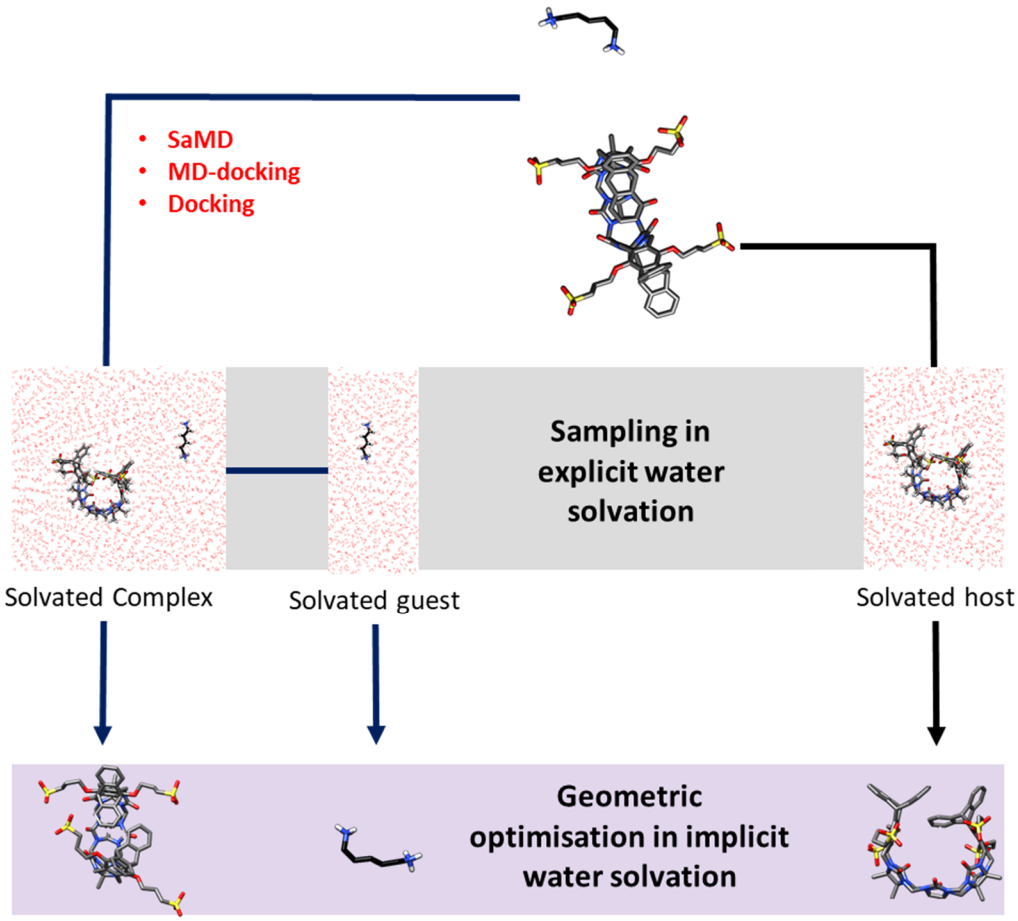


Figure 0. 9 : protocole utilisé pour générer des conformations de l'hôte, de l'invité et des systèmes hôte-invité. Différentes méthodes ont été testées pour générer des modèles initiaux du complexe hôte-invité. La dynamique moléculaire est réalisée dans l'eau et un modèle de solvatation explicite est utilisé pour échantillonner l'espace conformationnel. Ensuite, pour les calculs des paramètres thermodynamiques, les molécules de solvants sont supprimées et la géométrie est minimisée à l'aide du logiciel xTB en utilisant un modèle de solvatation aqueuse implicite (GBSA).

Les différentes géométries issues de cette analyse présentent jusqu'à 10 kcal/mol de variations, confirmant l'importance fondamentale de l'échantillonnage conformationnel. La structure globale de plus basse énergie est définie comme référence pour le calcul de l'énergie libre. Bien que les degrés de liberté de l'invité soient beaucoup plus réduits, nous utilisons un protocole similaire par souci de cohérence.

I.A. 1 - ANALYSE RETROSPECTIVE SUR UN SYSTEME SIMILAIRE

Comme preuve de concept pour notre méthodologie, nous avons utilisé les données du défi SAMPL3. Cet hôte est similaire mais plus simple que celui de SAMPL7. Une procédure d'amarrage (docking) est réalisée en considérant une grande boîte de simulation (15 $Å^3$) conduisant à la formation de complexes prédit par la fonction de score (scoring) comme possédant une énergie de liaison négative, cependant une analyse des interactions révèle que l'invité n'a formé que des interactions de surface avec l'hôte. Cela nous a conduit à tester deux autres conditions d'amarrage où l'espace d'amarrage (la taille de la boite) est progressivement réduit. Les géométries résultantes possèdent cette fois des scores positifs, indiquant des possibles problèmes dans la conformation du système hôte-invité, mais dans ce cas, l'invité s'insère dans la cavité de l'hôte. Trois à cinq modes de liaisons différentes ont été sélectionnées pour chaque protocole d'amarrage. La minimisation à l'aide du logiciel CHIMERA a permis une relaxation du système avant l'étape de minimisation et le calcul de l'énergie libre au niveau semi-empirique.

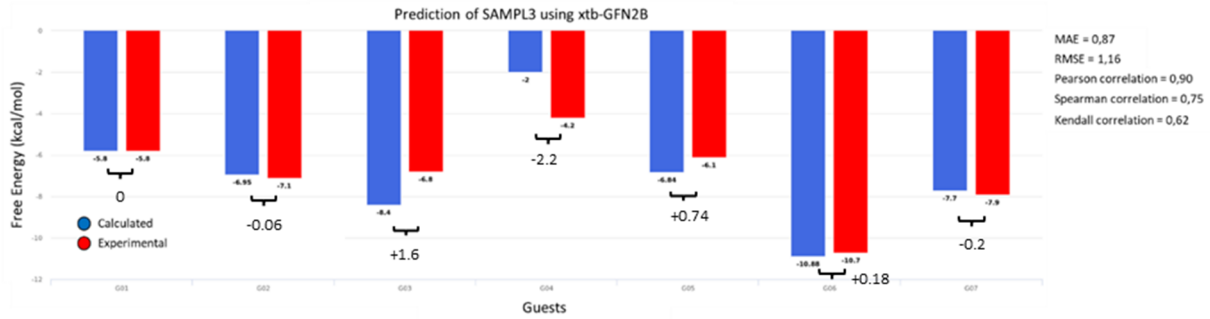


Figure 0. 10 : résultat de l'analyse rétrospective sur le système hôte-invité issu du challenge SAMPL3. Les prédictions sont montrées en bleu, et comparées aux valeurs expérimentales en rouge.

Comme le montre la Figure 0. 10, les énergies libres de liaison prédites sont en excellent accord avec l'expérience (RMSE = 1,16 kcal/mol ; MAE = 0,87 kcal/mol ; corrélation de Pearson (r) = 0,90 ; corrélation de rang de Spearman (ρ) = 0,75, corrélation tau de Kendall = 0,62(τ)). En fait, dans quatre des sept cas d'essai, nous obtenons un accord quantitatif, et l'erreur est inférieure à 1 kcal/mol, dans les deux autres cas, les erreurs sont respectivement de 1,6 kcal/mol et 2,2 kcal/mol. Cela nous a conduit à penser que, si le mode de liaison est correct, la méthode

semi-empirique GFN2B-xTB est susceptible de fournir des résultats d'une niveau de précision semblable à la mécanique quantique pour seulement une partie du coût de calcul. Pour cette analyse rétrospective réalisée sur les données du challenge SAMPL3, la prédiction d'énergie libre d'interaction montre des résultats très précis par rapport à ceux qui ont été publiés initialement.

I.A. 2 - PREDICTION D'ENERGIE LIBRE DE GIBBS SUR LE SYSTEME TRIMERTRIP ISSU DU CHALLENGE SAMPL7

En ce qui concerne le jeu de donnée du challenge SAMPL7, pour chaque complexe, nous extrayons 5 à 10 modes de liaison différents générés avec les protocoles décrits ci-dessus. Chacune des géométries est ensuite minimisée individuellement au niveau semi-empirique, et seules celles pour lesquelles toutes les fréquences vibratoires sont positives sont considérées. Le complexe de plus basse énergie est considéré comme le minimum de référence, sauf dans quelques cas où une inspection visuelle a permis l'identification de problèmes structurels avec la géométrie correspondante, vraisemblablement liés à un filtrage inadéquat des charges lié à la méthode de solvatation implicite.

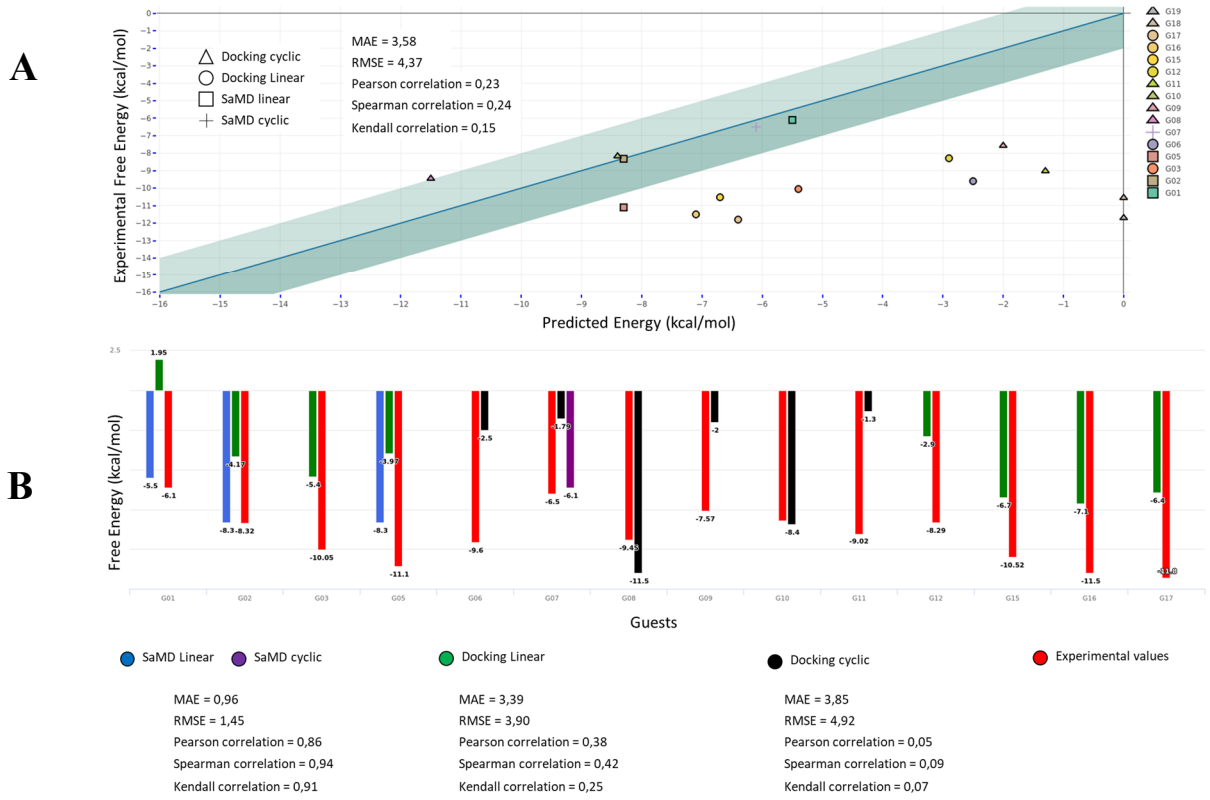


Figure 0. 11 : comparaison entre valeurs expérimentales et valeurs prédites pour le challenge SAMPL7. (A) Graphique de corrélation, la zone en verte, représentant le seuil de +1/-1 kcal/mol. Les symboles indiquent la nature de la méthode utilisée pour la prédiction du mode de liaison. (B) Histogramme de la prédiction, chaque couleur correspondant à la méthode utilisée pour la prédiction du mode de liaison.

Pour les invités G18 et G19, nous n'avons pas pu trouver un mode de liaison correct, et les résultats de l'amarrage moléculaire ont donné une énergie de liaison positive. Comme les deux protocoles ont échoué pour ces deux invités cycliques (vraisemblablement en raison de leurs grands volumes), nous avons renoncé à faire des prédictions pour ceux-ci.

Dans le cas du SAMPL7 challenge, les modes de liaison, ont été générés de trois façons différentes de la moins précise à la plus précise : (i) l'amarrage moléculaire (Docking), (ii) l'amarrage moléculaire suivi d'une dynamique moléculaire (MD-Docking), (iii) l'association spontanée de l'hôte et l'invité par dynamique moléculaire non biaisée (SaMD).

D'une manière générale, nos prédictions peuvent être séparées en trois classes dépendant de la précision de nos prédictions :

- (i) Les prédictions présentant un excellent accord avec les données expérimentales (< 2 kcal mol). 5 systèmes hôtes-invités se trouvent dans cette catégorie, trois d'entre eux extrait du protocole d'association spontanée (G01, G02, et G07), deux d'entre eux issus de l'amarrage moléculaire suivi d'une dynamique moléculaire (G08 et G10).

- (ii) Les prédictions incorrectes mais qui restent proches des valeurs expérimentales (3 à 5 kcal/mol d'erreurs). Ces complexes (G03, G05, G15, G16, G17) sont principalement des molécules présentant une structure linéaire, et les résultats proviennent des résultats de l'amarrage moléculaire, à l'exception de G05, qui provient du protocole d'association spontanée.

- (iii) Les prédictions avec de grandes erreurs (> 4 kcal/mol). 6 systèmes hotes-invités appartiennent à cette catégorie, y compris les G18 et G19 (pour laquelle aucune énergie de liaison négative n'a été trouvée). La plupart d'entre eux correspondent à des hôtes cycliques, et les erreurs peuvent être attribuées principalement à notre incapacité à trouver des modes de liaison raisonnables dans le délai du défi.

Dans la Figure 0. 11B, nous montrons que pour les complexes pour lesquels SaMD fournit un mode de liaison correct, les prédictions d'énergie libre de liaison sont bien plus précises que pour les résultats obtenus à partir des poses d'amarrage moléculaire. En fait, la plupart des cas (G01, G02, G05, G07) sont en accord quantitatif avec l'expérience (+/- 1kcal/mol) et les statistiques de performance globale sont excellentes : pour RMSE = 1,45 kcal/mol ; MAE = 0,96 kcal/mol ; corrélation de Pearson (r) = 0,86 ; corrélation de rang de Spearman (ρ) = 0,94, corrélation de rang de Kendall = 0,91(τ). Par rapport à la méthode SaMD, les résultats du docking sous-estiment l'énergie libre de liaison, ce qui suggère que les conformations à faible énergie du complexe hôte-invité peuvent être échantillonnées avec la méthode de la dynamique moléculaire, mais que l'amarrage moléculaire seul se révèle insuffisant.

I.A. 3 - APPROCHE BASEE SUR LA CONNAISSANCE

Pour la prédiction du GDCC, comme il y avait une quantité importante de données préexistantes provenant de défis précédents, nous avons décidé d'essayer une approche orthogonale basée sur le ML. L'ensemble de données comprend 35 composés au total, appartenant à trois classes de systèmes hôtes qui sont similaires en termes de structure et de composition chimique : OA, TEMOA et exoOA. Les valeurs d'énergie libre de liaison sont comprises entre -3,73 kcal/mol et -8,38 kcal/mol. Le modèle utilisé est un réseau neuronal, utilisant 90 descripteurs moléculaire généré à l'aide de la web-plateforme CORINA (60 décrivant l'invité et 30 décrivant le système hôte). Les prédictions pour l'ensemble d'apprentissage sont très précises, avec RMSE = 0,92 kcal/mol et toutes les valeurs prédites dans une fourchette de 1 kcal/mol par rapport aux valeurs expérimentales (Figure 0. 13). Pour l'ensemble de test, toutes les valeurs prédites sont proches des valeurs expérimentales, avec des erreurs maximales et minimales de -1,49 kcal/mol et +0,22 kcal/mol, respectivement.

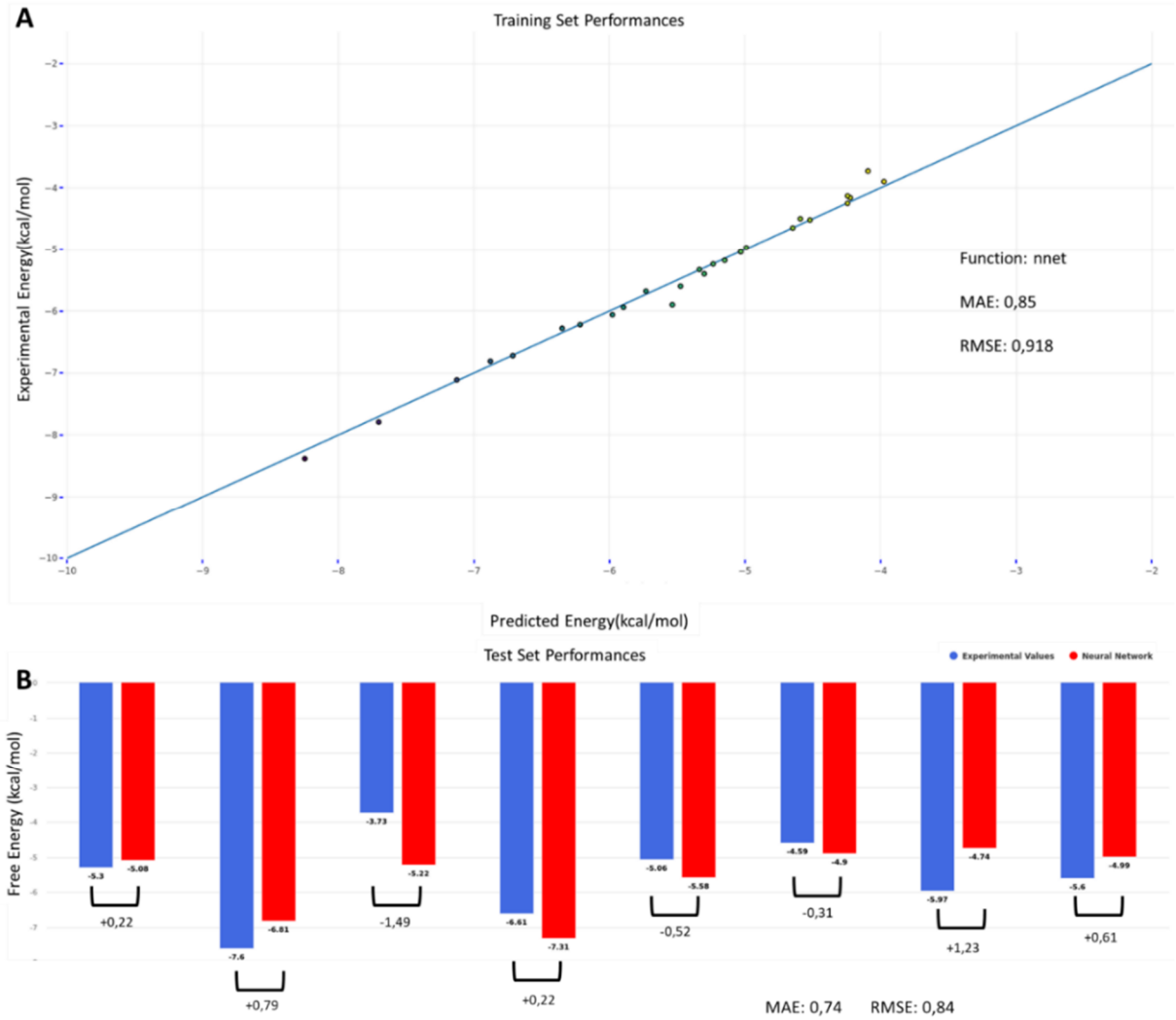


Figure 0. 12 : (A) Performance du jeu d'entrainement incluant 27 invités différents interagissant avec deux hôtes différents. (B) the jeu de test inclus 8 invités différents et l'énergie libre de Gibbs est prédite par le modèle utilisant le jeu d'entrainement.

L'ensemble de données GDCC à prédire consistait en 8 composés invités (quatre chargés positivement et quatre neutres) se liant à deux systèmes hôtes apparentés. Après les diverses optimisations du modèle, il ne faut que 10 secondes pour calculer l'énergie libre de liaison des huit invités dans les deux hôtes. Avec des valeurs RMSE et MAE de 1,67 kcal/mol et 1,21 kcal/mol, respectivement, les performances globales sont plutôt satisfaisantes. Il est intéressant de noter que pour les quatre invités présentant des charges négatives, les prédictions ne sont pas optimales, ce qui peut s'expliquer par les limites du modèle imposé par la composition de l'ensemble d'apprentissage : comme la valeur d'énergie libre de liaison la moins favorable est de -3,73 kcal/mol, le modèle ne peut pas prédire des valeurs plus positives. Cependant, même dans ce cas, la hiérarchie entre les valeurs des invités est respectée (G4 < G3 < G2). Il n'y a pas de valeur expérimentale pour G1, il n'a donc pas été pris en compte dans cette analyse. Si nous

appliquons la même analyse à chaque sous-groupe (en fonction de la charge positive ou négative et de l'hôte avec lequel ils interagissent), nous obtenons une prédiction hiérarchique presque parfaite. La seule exception est le complexe OA-G7, qui a été prédit plus bas que OA-G6 en raison du fait que OA-G7 a été sous-estimé (-5,67 kcal/mol au lieu de -6,98 kcal/mol) alors que OA-G6 a été prédit très proche de ses valeurs expérimentales (-5,92 pour -5,83 valeurs expérimentales). Tous les systèmes, à l'exception des quatre composés négatifs interagissant avec le système exo-OA, sont prédits à moins de 1 kcal/mol des valeurs expérimentales (Figure 0. 13). Pour les complexes impliquant le système OA, qui figure en bonne place dans l'ensemble d'apprentissage, les prédictions sont encore meilleures, avec MAE = 0,55 kcal/mol et RMSE = 0,85 kcal/mol.

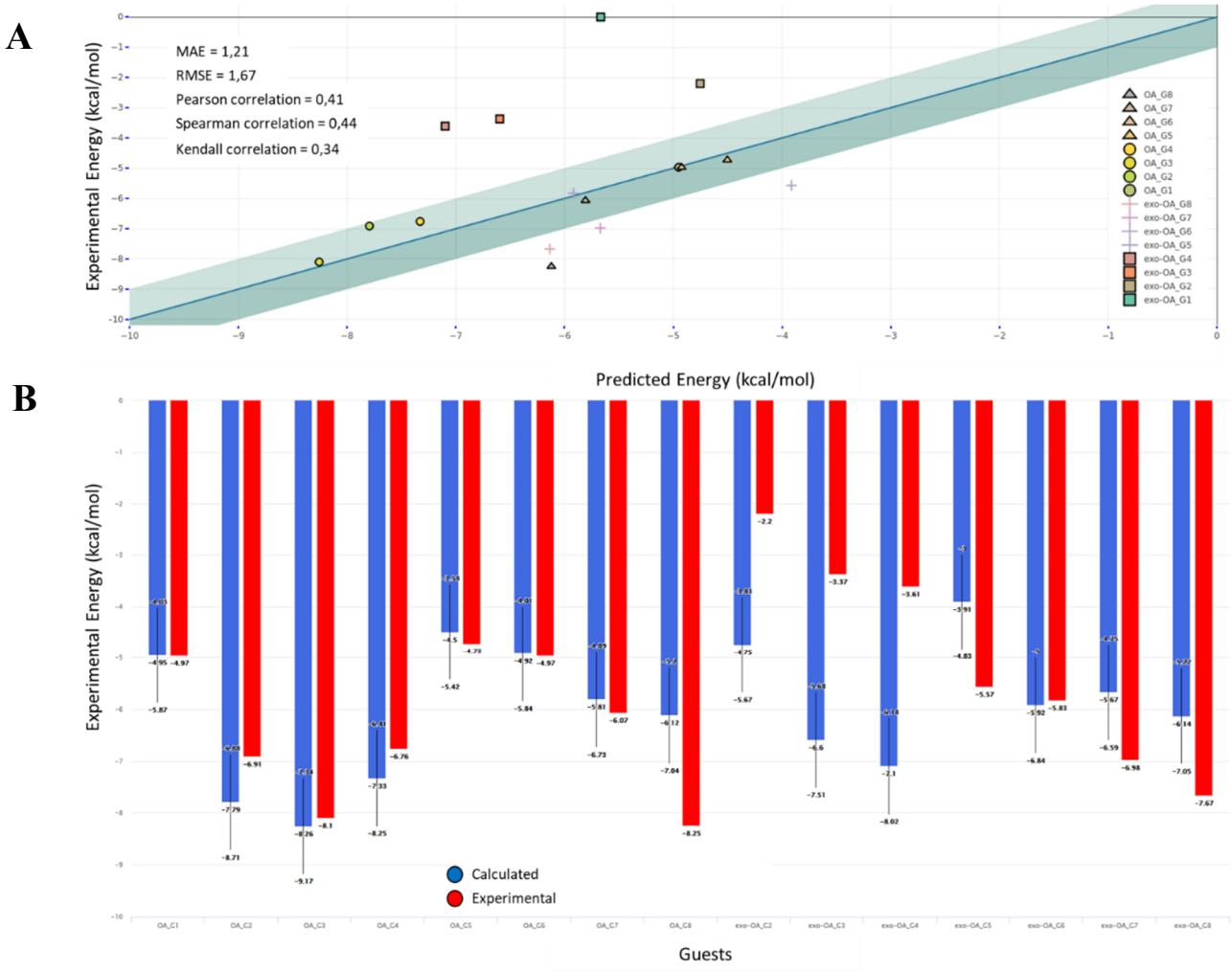


Figure 0. 13 : comparaison des valeurs expérimentales et prédites pour les énergies libres de Gibbs. (A) Graphique de corrélation, la zone en verte, représentant le seuil de +1/-1 kcal/mol. Les symboles indiquent la nature des invités et chaque prédiction possède une couleur différente. (B) Histogramme de la prédiction, avec en bleu les valeurs prédites et en rouge les valeurs expérimentales. Les barres d'erreurs reflètent l'erreur moyenne du modèle (RMSE).

I.A. 4 - CONCLUSION SUR LE DEFI SAMPL7

La participation à SAMPL7 nous a permis de tester deux approches orthogonales pour calculer les énergies libres de liaison hôte-invité, en identifiant dans chaque cas les forces et les limites qui ont été prises en compte pour la conception finale de la plateforme automatisée HG-DYNAusor.

L'approche basée sur la thermodynamique est absolument générale et peut être utilisée, en principe, sur n'importe quel système hôte-guide. L'utilisation d'un ensemble de base semi-empirique avancé (GFN2B-xTB) pour calculer les énergies et les corrections thermostatiques offre des performances accrues par rapport aux approches de mécanique moléculaire avec un coût de calcul modéré et élimine la dépendance vis-à-vis des champs de force des petites molécules, qui sont souvent imprécis. Différents aspects critique pouvant conduire à conduire à des prédictions incorrectes ont cependant pu être identifié :

Le premier est une dépendance critique de la structure du complexe hôte-invité utilisé pour la génération du mode de liaison. Pour les systèmes présentant une flexibilité importante de l'hôte, l'arrimage rigide du récepteur peut être inapproprié, et un échantillonnage conformationnel est alors nécessaire. L'observation directe de la formation de la paire hôte-invité en utilisant la dynamique moléculaire dans un modèle de solvatation explicite représente une solution optimale en termes de qualité des prédictions de l'énergie libre de liaison, mais peut être limité par les temps de simulation, qui augmentent avec le nombre de degrés de liberté du système. Dans le cas du système dit « Trimertrip », nous avons identifié une transition lente entre la conformation fermée et ouverte de l'hôte comme étant le goulot d'étranglement du processus d'association. Dans ce cas, le fait de commencer les simulations de dynamique moléculaire à partir d'une conformation ouverte de l'hôte peut donner d'excellents résultats pour une fraction du coût de la simulation.

La deuxième limite de notre approche est la méthode de solvatation implicite (GBSA) qui peut sous-estimer le coût de désolvatation des espèces ioniques en solvatation aqueuse, conduisant à la formation de paires ioniques dont la contribution est surévaluée. Nous n'avons pas observé avec notre méthode de biais systématique, cependant le modèle de solvatation implicite reste l'une des faiblesses de l'approche. Des versions plus récentes du logiciel xTB ont remplacé le formalisme GB pour un modèle de Poisson-Boltzmann linéarisé analytiquement (ALPB). Il sera intéressant de vérifier les performances du modèle ALPB dans les futures éditions de SAMPL. Dans tous les cas, la solvatation explicite dans les simulations MD est mieux adaptée.

Ainsi, l'utilisation d'instantanés MD comme géométries d'entrée dans les calculs GFN2B-xTB semble fournir de meilleurs résultats que l'échantillonnage conformationnel exhaustif avec solvatation implicite.

L'utilisation de méthodes basées sur la connaissance peut être très avantageuse lorsqu'il existe suffisamment de données préexistantes. Contrairement aux complexes protéine-ligands, pour lesquels il existe un grand nombre de données, les systèmes hôte-invité ne peuvent pas bénéficier d'ensembles d'entraînements massifs. Ainsi, nous étions particulièrement intéressés par l'examen de l'adéquation des approches d'apprentissage automatique, avec une attention particulière sur le risque de sur-apprentissage. Les résultats obtenus sur le système GDCC sont vraiment encourageants et nous ont incité à construire une base de données de systèmes hôte-invité avec leurs énergies libres de liaison correspondantes de manière à améliorer la prédiction par des méthodes d'apprentissage automatisées.

V - SYNTHESE ET CARACTERISATION D'UNE NOUVELLE PINCE MOLECULAIRE Zn(II)-PORPHYRINE-ACRIDINIUM

V. A - SYNTHESE ET CARACTERISATION

Dans la dernière partie de cette thèse, nous avons réalisé la synthèse et la caractérisation d'un nouveau récepteur Zn(II)-porphyrine-acridinium suivant la voie de synthèse présentée ci-dessous sur la Figure 0. 14.

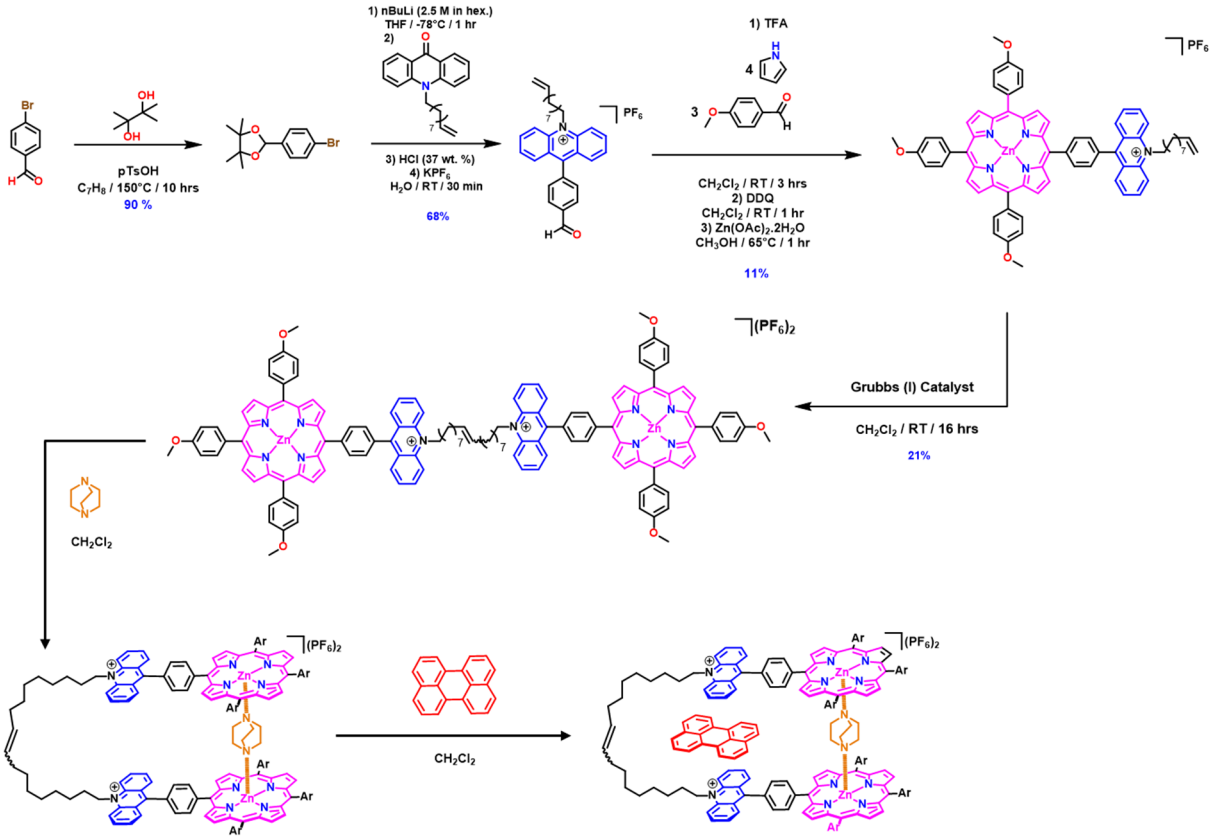


Figure 0. 14 : synthèse d'un nouveau récepteur Zn(II)-porphyrine-acridinium, utilisant le DABCO pour coordiner les deux Zincs ions métalliques zinc(II). Le récepteur ainsi formé peut encapsuler le perylène.

Concernant la modélisation du système Zn(II)-porphyrine-acridinium complexant le DABCO, trois dynamiques moléculaires différentes sont lancées à partir de trois points de départ différents. Comme nous l'avons fait précédemment, un ensemble de descripteurs moléculaires décrivant la déviation, le rayon de giration et la surface du récepteur sont calculés. En outre, plusieurs autres descripteurs numériques décrivant plus en détail la structuration des récepteurs sont ajoutés à l'analyse. Ces descripteurs sont séparés en trois types : les descripteurs de distance, les descripteurs d'angles et les descripteurs d'angles dièdres et décrivent principalement la chaîne alcène et la dynamique de l'acridinium. Au total, 33 descripteurs

supplémentaires sont ajoutés. Comme le jeu de données contient beaucoup plus d'informations que les fois précédentes, la variabilité est plus diluée dans les différentes composantes. Avant l'analyse, un processus de réduction dimensionnelle est effectué, 12 variables sont sélectionnées et utilisées pour une analyse plus approfondie.

Les résultats sont présentés dans la Figure 0. 15 : : ~80% de la variabilité est expliquée par les quatre premières composantes. En ce qui concerne "MD2", et"MD3" respectivement colorés en vert et en rouge dans le graphique, ils échantillonnent un espace conformationnel très similaire mis en évidence par le fait qu'il se chevauche. "MD1" quant à lui présente une certaine variation. Bien que la plus grande partie de la dynamique se chevauche avec les deux autres, certaines géométries ne sont échantillonnées que par ce système, ce qui représente probablement une conformation particulière associée à un événement rare. Certaines de ces géométries spécifiques sont mises en évidence dans la partie inférieure de la Figure 0. 15. En bleu, une conformation rare mise en évidence par *1 dans le graphique représente une conformation du récepteur ou la chaîne alcène entre dans la cavité, générant un encombrement stérique important, et bloquant l'entrée d'un éventuel invité dans le site de liaison.

La majeure partie des géométries présente une cavité de liaison accessible où le récepteur est dans une configuration ouverte permettant de complexer un invité entre les deux acridiniums (*2 en bleu et *1 et *2 en orange dans la Figure 0. 15). Cette forme est prédominante dans toutes les simulations, les points sont donc pour la plupart situés dans la même zone et se chevauchent. Pour MD3 (vert), deux géométries spécifiques peuvent être extraites : la première (*1) montre la conformation fermée du récepteur avec les deux acridiniums interagissant entre eux. En revanche, la seconde (*2) représente la conformation très ouverte, où les deux acridiniums ont tourné de 90° et ne se font plus face. La chaîne alcène semble être suffisamment flexible pour permettre ce changement de conformation dans la structure.

D'après les observations de la dynamique moléculaire, la conformation ouverte prédominante peut prendre plusieurs orientations au fil du temps. D'après ce que nous avons vu dans les simulations, la conformation ouverte peut être considérée comme un point d'équilibre à partir duquel la géométrie est susceptible de diverger vers une autre conformation (fermée, semi-fermée, tournée...). Nous nous attendons à ce que la géométrie avec la configuration ou la chaîne alcène entre dans la cavité, ait une énergie absolue plus faible que la configuration trans, qui est généralement celle utilisée pour lier un invité à l'intérieur de la cavité de liaison.

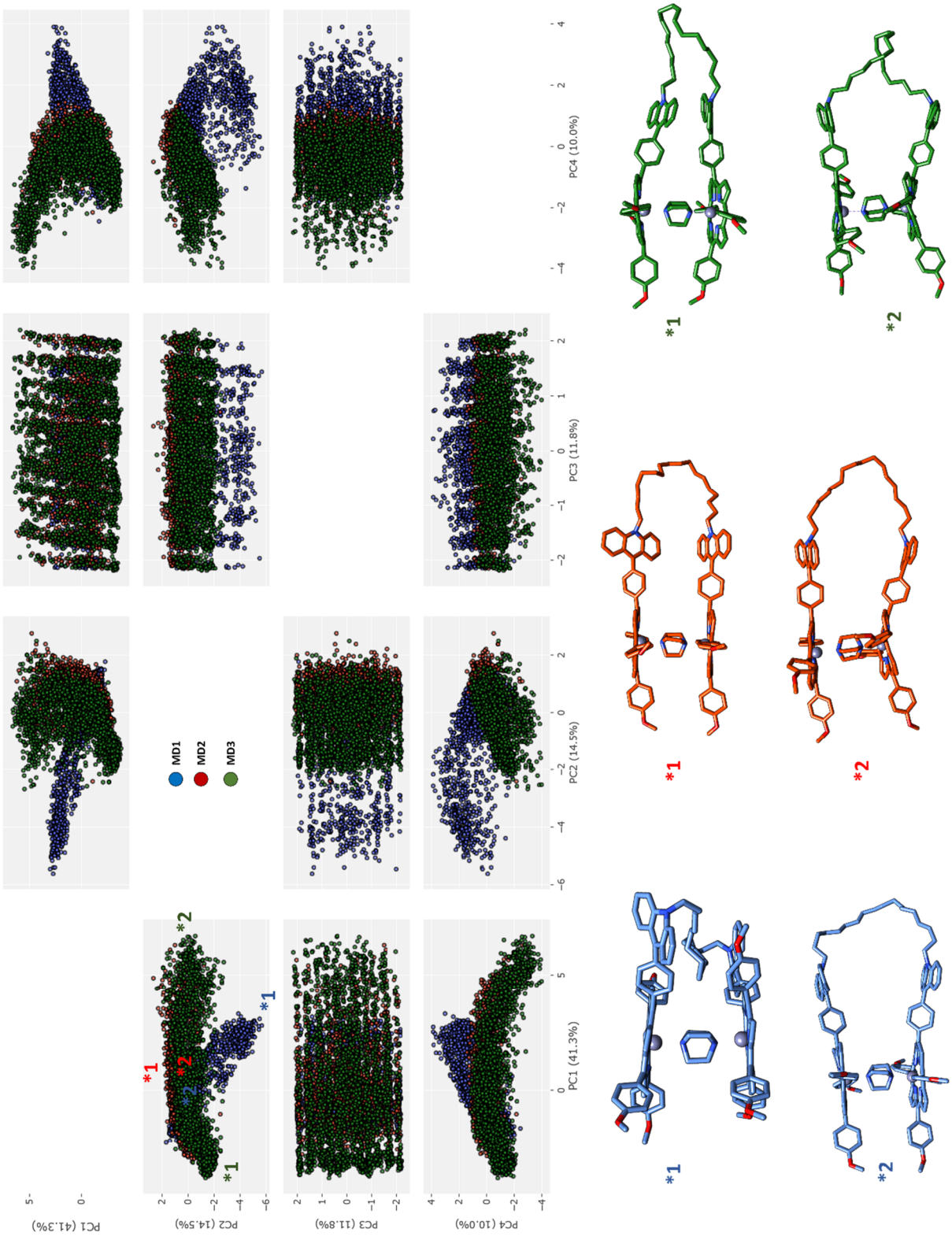


Figure 0. 15 : analyse en composante principale du nouveau récepteur Zn(II)-porphyrine-acridinium décrit par un ensemble de descripteurs moléculaires calculé à partir des simulations de dynamique moléculaire. L'espace formé par la combinaison des quatre premières composantes explique ~80% de la variabilité de l'échantillon. Chaque point représente une géométrie, et chaque couleur représente une simulation différente de dynamique moléculaire.

V. B - CRIBLAGE VIRTUEL DE NOUVEAUX INVITES POTENTIEL

L'identification de nouveaux invités potentiels par criblage virtuel de base de données d'intérêt (Drug Bank et T3DB) a été réalisé, de façon à identifier de possibles applications du système : vectorisation de molécule d'intérêt thérapeutique, recapture de polluant… Un premier aperçu des résultats est présenté dans les deux tables suivantes :

Table 0. 1 : vue d'ensemble des 10 meilleurs résultats d'amarrage moléculaire sur la base de données T3DB

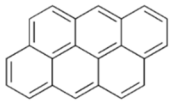
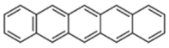
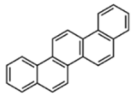
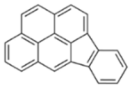
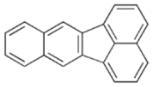
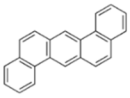
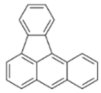
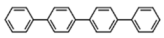
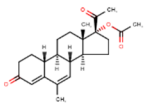
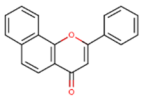
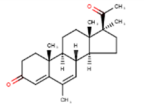
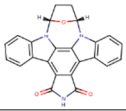
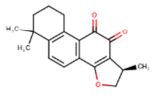
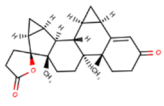
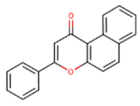
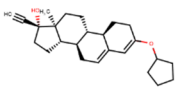
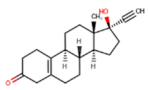
Score	Name	structure	Score	Name	structure
-14.9	Ovalene		-11.9	Anthanthrene	
-12.3	Pentacene		-11.9	Benzo[ghi]perylene	
-12.3	Coronene		-11.8	Picene	
-12.2	Indeno(1,2,3-cd)pyrene		-11.7	Benzo[k]fluoranthene	
-12.1	Dibenzo[a,h]anthracene		-11.5	Benzofluoranthene	

Table 0. 2 : vue d'ensemble des 10 meilleurs résultats d'amarrage moléculaire sur la base de données DrugBank

Score	Name	structure	Score	Name	structure
-11.3	pQuaterphenyl **Investigational**		-10.6	Nomegestrol **Approved**	
-11.0	alphaNaphthoflavone **Experimental**		-10.6	Medrogestone **Approved**	
-10.9	DB08683 **Experimental**		-10.6	Cryptotanshinone **Experimental**	
-10.8	Drospirenone **Approved**		-10.5	betaNaphthoflavone **Experimental**	
-10.8	Quingestanol **Experimental**		-10.5	Norethynodrel **Approved**	

Les dix meilleurs résultats de la procédure de l'amarrage moléculaire sont présentés avec le score associé pour chacune des bases de données. Chaque base de données a été considérée séparément, et l'amarrage moléculaire est réalisé une fois pour chacune d'entre elles dans un récepteur extrait des dynamiques moléculaires précédentes et présentant une cavité de liaison ouverte. Sur les vingt meilleurs résultats, toutes les molécules sont situées dans la cavité.

Les dix meilleures molécules extraites de la base de données T3DB sont principalement des polluants. Elles peuvent toutes être classées comme des hydrocarbures aromatiques polycycliques (HAP), susceptibles d'interagir avec une bonne affinité avec la pince moléculaire en utilisant des interactions π-π. Si l'on considère les scores des composés de la T3DB, ils sont globalement élevés, ce qui suggère que ces molécules sont de bons substrats pour le récepteur. Concernant la DrugBank, seul le meilleur résultat est un HAP, mais la majorité des molécules classées peuvent interagir avec le récepteur avec des interactions π-π. Il est très intéressant de souligner que parmi les dix meilleures molécules, quatre d'entre elles présentent un squelette stéroïdien (Figure 0. 16). Cela pourrait suggérer une particularité intéressante du récepteur pour lier ces types de molécules.

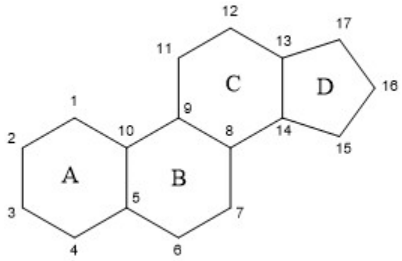


Figure 0. 16 : représentation du squelette stéroïdien

L'une des principales limites du processus d'amarrage moléculaire est la géométrie initiale du récepteur. Contrairement aux molécules de l'invité, l'hôte, ne peut pas être considéré comme totalement flexible dans l'algorithme d'amarrage moléculaire. Cela signifie que la structure initiale a été extraite de la simulation précédente de dynamique moléculaire. Pour l'instant, les propriétés thermodynamiques de l'hôte ne sont pas encore calculées. Cependant, compte tenu du peu de temps nécessaire à l'exécution de la procédure d'amarrage moléculaire, il est toujours possible d'envisager d'exécuter un nouvel amarrage moléculaire sur un nouvel ensemble de conformation basé sur plusieurs géométries différentes présentant une faible énergie.

D'une manière générale, les résultats de l'amarrage moléculaire présentent des ligands qui sont presque tous des molécules capables d'interactions π-π avec le récepteur. Le mode de liaison et la dynamique de ces dix molécules sera étudiée dynamique moléculaire en suivant le protocole

décrit précédemment concernant l'utilisation de la plateforme HG-DYNAusor. En conclusion, les cinq meilleures molécules de chaque amarrage moléculaire seront extraites, et l'énergie libre de liaison sera mesurée pour ces dix molécules.

V. C - ECHANTILLONAGE PHARMACOPHORIQUE

Enfin, une approche d'échantillonnage pharmacophorique est réalisée de façon à identifier dans ces mêmes bases de données, certains composés pouvant directement agir comme ligand ditopique et se lier aux porphyrines de Zn(II), permettant d'envisager de nouvelles applications possibles pour ce système innovant.

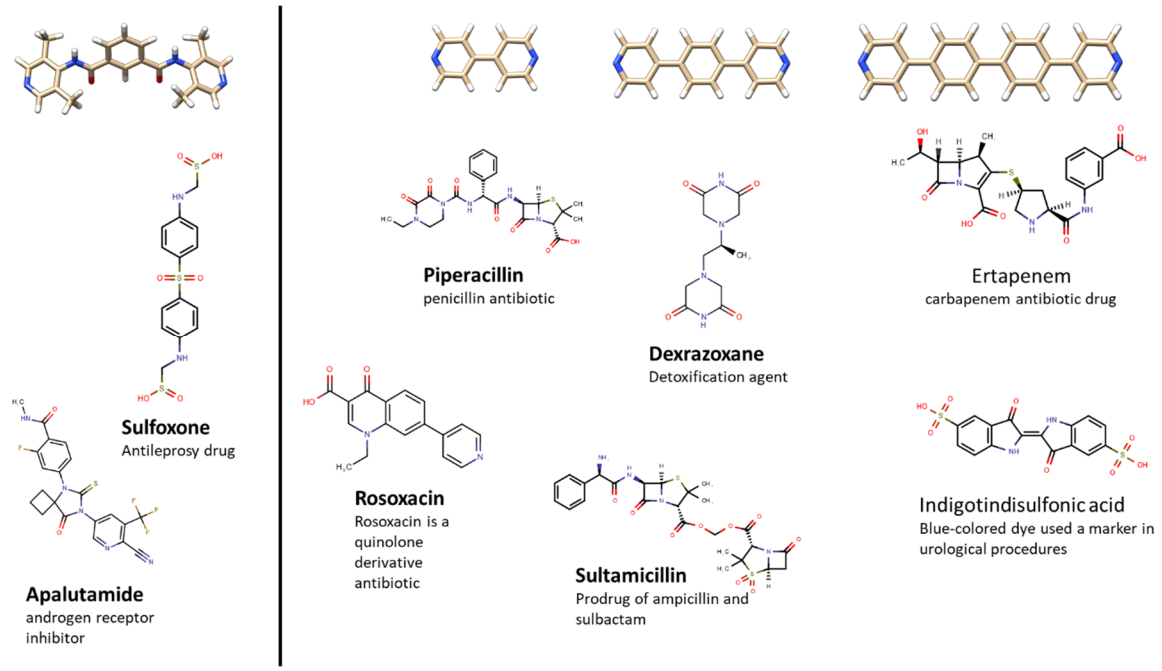


Figure 0. 17: vue d'ensemble des résultats préliminaire de l'approche pharmacophorique

La recherche pharmacophorique est effectuée avec le logiciel MOE en utilisant les bases de données 3D construites. Quatre pharmacophores différents sont étudiés (partie supérieure de la 0. 17). Seule la distance entre l'azote et le caractère donneur ou accepteur est considérée pour chacun d'entre eux. La banque de médicaments est utilisée, et les différents filtres sont appliqués. Un aperçu de certains résultats est présenté dans la Figure 75. L'idée est de sélectionner des molécules qui coordonneraient les deux Zn(II) des porphyrines. Malheureusement, la plupart des molécules extraites présentent des groupes près de l'azote qui sont susceptibles de réduire ou d'inhiber l'affinité de cet invité ditopique pour le Zn(II), rendant leur utilisation compliquée dans ce contexte. En conclusion, il est nécessaire de générer des filtres plus détaillés pour éviter les collisions stériques ou, au contraire, moins restrictifs mais limitant la base de données en ne considérant que les molécules « approuvées ».

VI - CONCLUSION GENERALE

En conclusion, au cours de cette thèse, nous avons construit une plateforme automatisée qui : (i) paramètre rapidement et avec précision des systèmes supramoléculaires hôte-invité pouvant être modélisés et simulés en mécanique moléculaire classique, et ce, sans nécessité de connaissances particulièrement poussées en informatique ; (ii) pour lesquels l'enthalpie libre avec d'éventuels invités peut-être mesurée via deux différentes méthodes fonctionnant indépendamment l'une de l'autre.

Nous avons synthétisé un nouveau récepteur Zn(II)-porphyrine-acridinium, préorganisé par la complexation du DABCO, pour lequel le comportement dynamique et énergétique a été analysé, conduisant à un échantillonnage conformationnel des différentes orientations géométriques du système au cours du temps, ainsi qu'à une analyse de son comportement dans les solvants organiques. De cet échantillonnage, la structure de plus basse énergie a pu être mis en évidence. Dans un second temps, nous avons pu extraire de nos analyses computationnelles un échantillon représentatif de composés susceptibles d'interagir de façon forte avec notre hôte, et pour lesquels l'enthalpie libre sera mesurée et testée ultérieurement.

CONTENTS

List of the abbreviations **1**

Chapter 1: Introduction **3**

I - Computational chemistry 4

II - Supramolecular chemistry 6

II. A - General definition 6

II. B - The intermolecular forces in molecular association 8

II. C - Host-guest chemistry 10

III - Computational methods for the determination of binding free energy in host-guest complexes 12

III. A - State of art of the computational methods used for binding free energy prediction 12

III. B - Molecular modelling approaches 15

III. C - Knowledge-based approaches 17

IV - Computational exploration of host-guest complexes 19

IV. A - Presentation of the NOAH project 19

IV. B - General aim of the thesis project 19

Chapter 2: Computational methods **21**

I - Simulations methods 22

I. A - Quantum mechanics (QM) 22

I. B - Semi-empirical quantum mechanic (SQM) 27

I. C - Molecular mechanics (MM) 30

II - Solvation models 44

II. A - Introduction 44

II. B - Explicit solvation 44

II. C - Implicit solvation 46

III - Machine learning methods 49
III. A - Unsupervised methods 49
III. B - Supervised methods 55
IV - Binding Free energy determination 67
IV. A - Principle of the thermodynamic based method 67
V - Docking 69
V. A - Principle 69
V. B - Programs 69
Chapter 3: The HG-DYNAusor platform 72
I - Introduction 73
I. A - Capabilities 73
II - Proof of concept using acridinium tweezer 74
II. A - Introduction 74
II. B - Generation of parameters 74
II. C - Association in water: 75
II. D - Behaviour analysis of host-guest system in solution 77
II. E - DFT and GFN2-xTB comparison: 80
III - The Thermodynamic based approach of the HG-DYNAusor platform 83
III. A - Required software 83
III. B - Parametrisation of the Host system (module 01) 84
III. C - Parametrisation of the guest system and binding mode generation (module 02) 100
III. D - Thermodynamic based approach for binding free energy prediction (module 03) 103
IV - Behaviour analysis 105
V - The knowledge-based approach of the HG-DYNAusor platform (Module 04) 106
V. A - Overview of the Knowledge-based methods 106
V. B - Optimisation of the ML algorithm 114

V. C - Dimensional reduction procedure: 115
VI - HG-DYNAusor platform: future directions 126
VI. A - Clustering methods 126
VI. B - Thermodynamic based approach 126
VI. C - Knowledge-based approach 126
Chapter 4: Application of de HG-DYNAusor platform 128
I - Investigated systems: 129
I. A - Gibb Deep Cavity Cavitand (GDCC) 129
I. B - Cucurbituril CB[8] 131
I. C - Trimertrip 134
I. D - Calix[4]-Pyrrole 135
II - SAMPL challenges as a validation step for the platform 136
II. A - The SAMPL7 challenge 136
II. B - SAMPL8: CB[8] drug abuse challenge 148
II. C - SAMPL8 GDCC challenge 177
III - Solvent exchange analysis in Calix[4]pyrrol capsule: 180
III. A - Presentation of the host system 180
III. B - Simulations of the hosts: 181
Chapter 5: Computational analysis, synthesis, and characterisation of novel Zn(II)-porphyrin-acridinium receptors 185
I - Zn(II) bisporphyrin-acridinium scaffold 186
I. A - Porphyrin receptor 186
I. B - Allosterism 187
II - Presentation of the Zn(II)-porphyrin receptor 191
II. A - General structure 191
III - Generation of parameters for Zn(II)-porphyrin receptor 194
IV - Computational analysis of ditopic ligand binding 195

IV. A - Conformational Analysis of Zn(II)-porphyrin Receptors 195

IV. B - Binding free energy prediction of ditopic ligands 203

V - Synthesis and characterisation of a new Zn(II)-porphyrin acridinium receptor 205

V. A - Generality 205

V. B - Synthesis and characteriSation 206

V. C - Characterisation 210

VI - Identification of potential binders for the new Zn(II)-porphyrin acridinium: a future perspective 214

VI. A - Protocol 214

VI. B - Behaviour of new Zn(II)-porphyrin acridinium host 215

VI. C - Virtual screening 217

VI. D - Perspective 220

Experimental part 221

I - Synthesis of 2-(4-bromophenyl)-4,4,5,5-tetramethyl-1,3-dioxolane 222

II - Synthesis of 10-(but-en-1-yl)acridin-9(10H)-one 223

III - synthesis of the 10-allyl-9-(4-formylphenyl)acridin-10-ium 224

IV - Synthesis of porphyrin-acridinium conjugate ($\mathbf{4} \cdot \boldsymbol{PF6}$) 225

V - Synthesis of $\mathbf{1} \cdot (\boldsymbol{PF_6})_2$ 227

VI - Synthesis of $\mathbf{1} \cdot (\boldsymbol{PF_6})_2 \boldsymbol{\bullet DABCO}$ 228

VII - Synthesis of $[\mathbf{1} \cdot (\boldsymbol{PF_6})_2 \boldsymbol{\bullet DABCO}] \supset \boldsymbol{Perylene}$ 229

List of the figures 230

Bibliography 239

LIST OF THE ABBREVIATIONS

SAMPL = Statistical Assessment of the Modelling of Proteins and Ligands

D3R = Drug Design Data Resource

MM = Molecular mechanic

QM = Quantum Mechanics

SQM = Semi-empirical Quantum Mechanics.

GBSA = Generalized Born Surface Area

PBSA = Poisson–Boltzmann Surface Area

FEP = Free Energy Perturbation

MD = Molecular dynamics

RESP = Restrained Electrostatic Potential

AI = Artificial intelligence

ML = Machine Learning

DL = Deep Learning

NOAH = Network Of Functional Molecular Containers With Controlled Switchable Abilities

HG-DYNAusor = Host-Guest DYNamic an Automated application

HF = Hartree Fock

SCF = Self-Consistent Field

DFT = Density Functional Theory

NDDO = Neglect of Differential Diatomic Overlap

ZDO = Zero Differential Overlap

DFTB = Density-Functional Tight-Binding

XRD = X-ray crystallography

NRM = Nuclear Magnetic Resonance

OBC = Onufriev–Bashford–Case

PCA = Principal Component Analysis

QSAR = Quantitative Structure-Activity Relationship

SVM = Support Vector Machines

RF = Random Forest

Knn = K-nearest neighbours

NNET = Neural Network

SaMD = Spontaneous association Molecular Dynamics

RMSD = Root Mean Square Deviation

Rg = Radius of Gyration

SASA = Surfaces Accessible Solvent Area

ALPB = Analytical Linearized Poisson-Boltzmann

MPD = Molprint2D

BindingDB = Binding Database

GDCC = Gibb Deep Cavity Cavitand

OA = Octa-Acid

exo-OA = exo-Octa-Acid

TeMOA = Tetra-endoMethyl Octa-Acid

TeeTOA = Tetra-endoEthyl Octa-Acid

CB[n] = CucuBirt[n]uril

DBI = Davies-Bouldin Index

psF = pseudo-F statistic

THF = TetraHydroFurane

NDI = Naphthalenetetracarboxylic diimide

DABCO = 1,4-diazabicyclo[2.2.2]octane

1

INTRODUCTION

I - COMPUTATIONAL CHEMISTRY

In the last decades, the important development of informatics has led to a revolution in many fields of science, including chemistry and, specifically, the study of molecular association processes.[1,2] The modelling of any chemical process between molecules is generally very informative and can be used first to explain chemical phenomena and secondly to predict the outcome of a specific reaction or process. All of this falls into the realm of computational chemistry. So, we can ask ourselves, what is computational chemistry?

In 2001, in his book "computational chemistry", David C. Young stated two definitions[3]: "The term theoretical chemistry may be defined as the mathematical description of chemistry. The term computational chemistry is generally used when a mathematical method is sufficiently well developed that it can be automated for implementation on a computer." In summary, computational chemistry can be defined as a branch of chemistry that uses computer simulation to assist in solving chemical problems. It uses methods of theoretical chemistry, incorporated into efficient computer programs, to calculate the structures and properties of molecules and solids.[4,5]

It is interesting to highlight the fact that these definitions say nothing about the accuracy of the prediction. Any mathematical operation that leads to the description of a chemical process has limitations due to the extreme complexity of the natural process. Even though the mathematical algorithms have found very intuitive and provable results, the basis of these equations is based on approximations, which can lead to unrealistic results in some cases. Of particular note, predicting the changes of entropy and enthalpy upon the formation of a molecular complex is a very challenging problem for a computational chemist, even though this process is theoretically well understood. Determining the equilibrium constant of a binding event (or the binding free energy, which is equivalent) is of fundamental interest in chemistry, but its computational prediction is most difficult.[6] A large proportion of my thesis deal with this problem.

In his book "Essentials of Computational Chemistry", Cramer[7] defined the theory as one or more rules postulated to govern the behaviour of physical systems. For him, the role of the computational chemist is not devoted to the chemical aspects of the problem but more to the computer-related aspects (writing improved algorithms or developing new ways to visualize data…). What is interesting in this approach is the idea that behind the multidisciplinary aspect of computational science, all biological processes are modelled through the prism of

approximations, with the idea that in many cases, the real complexity of the system is not precisely measurable and must therefore be approximated.

II - SUPRAMOLECULAR CHEMISTRY

II. A - GENERAL DEFINITION

In 1978, J-M Lehn defined supramolecular chemistry as the chemistry interested in a new chemical entity, more complex than molecules: supermolecules.[8] A supermolecule represent a complex of molecules held together by noncovalent bonds, generally referred to as a supramolecular complex or supramolecular assembly. The complexed elements are associated according to the principles of molecular association. These supramolecular complexes can perform multiple functions, separated into three main groups: (i) molecular transformation, (ii) molecular vectorization, and (iii) molecular association.[9]

The molecular transformation represents the capabilities for some supramolecular complexes to act as a catalyst (supramolecular catalysis).[10]

Molecular vectorization can be defined as the process of association between a molecular substance and a molecular carrier to improve biological distribution, prevent potential degradation, decrease toxicity or improve the Physico-chemical properties of the molecular substance.[11]

The molecular association process is the most important one. It represents one of the most important events in chemical and biological processes, and both molecular transformation and molecular vectorization depend on it.[12–14] In supramolecular chemistry, the analysis of these events gives a lot of information about the chemical diversity and functions during the assembling of the host and the guest, particularly how they interact with each other. Figure 1 shows a classification of the possible combination of host and guest in the supramolecular field, depending on the dimension of the guest and the host jointly. Following the assembling process between the guest and the host, we can define three different families of supramolecular complexes based on the scaling of the complex: the molecular scale, the mesoscale, and the nanoscale. At the molecular scale, the complex is mainly formed by a 1:1 combination of a macrocycle or cavitand host with a relatively small guest.

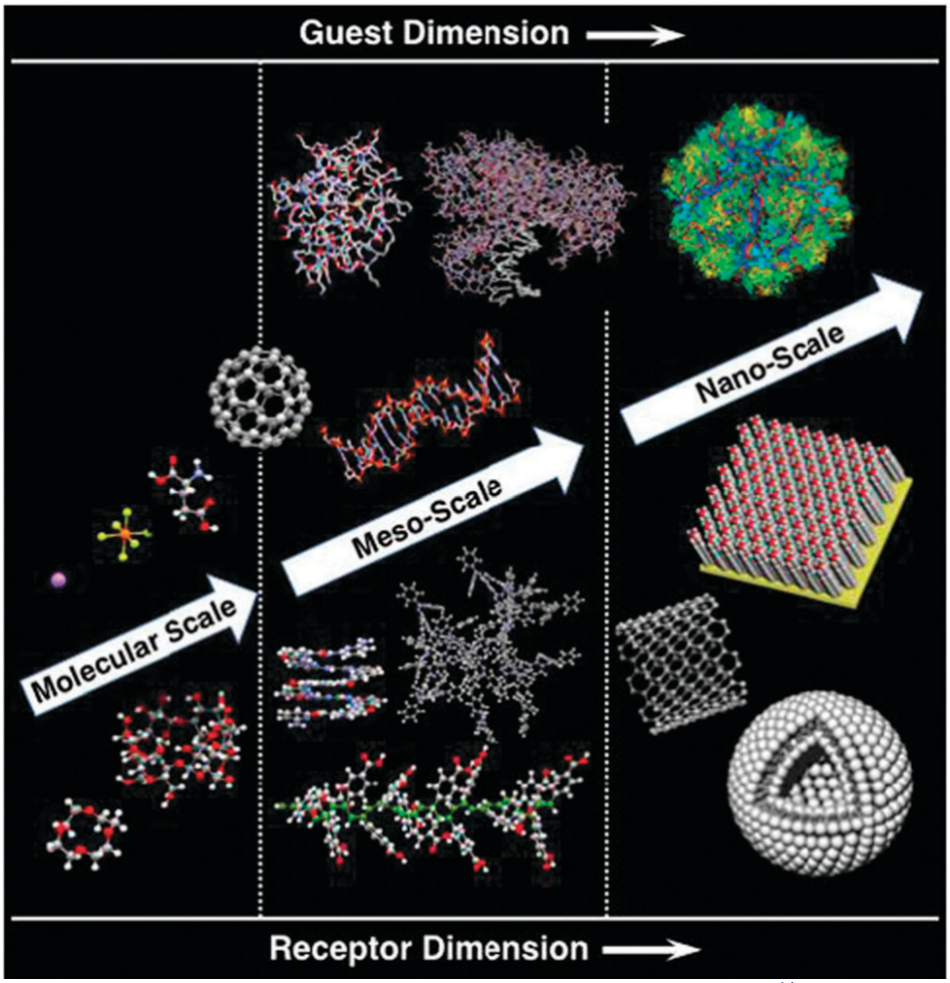


Figure 1: Classification of the supramolecular complex[14]

The meso-scale groups are steadily increasing, and in recent years host having a larger cavity (> 1nm) have been designed to interact with guest presenting a high diversity of shape and molecular function, with different stoichiometry: 1:1, 1:2, … (n-guest for one host). The molecular capsules for drug delivery enter this category. At the nano-scale level, the host and the guests are assembled with dimensions around hundreds of nanometers. In recent years, interesting approaches have been published to interact with biological systems at the nanoscale level.[15] In all these fields, many systems have been studied these past years for many possible applications.[16–20]

With the Nobel prize of chemistry of 2016 awarded jointly to Jean-Pierre Sauvage, Sir J. Fraser Stoddart, and Bernard L. Feringa "for the design and synthesis of molecular machines", the interest in the supramolecular chemistry field and its possible applications have grown significantly.[21–24]

II. B - THE INTERMOLECULAR FORCES IN MOLECULAR ASSOCIATION

The molecular association is one of the key concepts in chemistry. It deals with the affinities between types of atoms and the interactions they form. The association of two or more molecules implies the formation of multiple favourable interactions between them. The precise geometric constraints of the interactions also imply that small chemical or geometrical changes result in loss of binding. This gives rise to the selectivity of the molecular association, a phenomenon of special interest in biology. This is explained by the rather dogmatic model of lock and key, proposed by E. Fisher in 1894: for any molecular interaction, the substrate must present the correct topology to form a complex with the receptor.[25] This topology allows any molecular receptor to recognize its specific substrate. But the molecular flexibility of many ligands (free rotation of some bonds, angles, torsion…) and particularly of the receptor invalidates this simple model. Any theory of binding must also consider the ability of ligand and receptor to adapt to each other.

Molecular receptors are host molecules that contain a binding site or a cavity available to bind a smaller guest molecule. These hosts systems are capable of binding smaller molecules using reversible non-covalent interactions: interactions that do not involve the sharing of electrons but rather involve a more dispersed variation of electromagnetic interactions.[26] To form a host-guest complex, there must exist molecular complementarity between the host and the guest, and this complementarity must translate into a negative binding free energy.[27]

The non-covalent interactions are also called the “weak forces” due to their low energy values compared to the formation of covalent bonds. These non-covalent interactions are multiples, and most of them come from the electrostatic interaction among particles. Historically, they are classified into two main parts: the long-range and the short-range interactions. A brief overview of the forces is presented in Table 1:

Table 1: Non-covalent interactions (NCI) classification for the molecular association.[28]

Contribution	Additive?	Sign	Comment
Long-range ($U \sim R^{-n}$)			
Electrostatic	Yes	±	Strong orientation dependence
Induction	No	−	
Dispersion	approx.	−	Always present
Resonance	No	±	Degenerate states only
Magnetic	Yes	±	Very small
Short-range ($U \sim e^{-\alpha R}$)			
Exchange-repulsion	approx.	+	Dominates at very short range
Exchange-induction	approx.	−	
Exchange-dispersion	approx.	−	
Charge transfer	No	−	Donor-acceptor interaction

I.A. 5 - THE LONG-RANGE INTERACTIONS

The long-range interactions are mainly composed of electrostatic, inductive, and dispersive contributions. These interactions are involved when the interacting particles are separated by long distances. Coulomb's law quantifies the electrostatic effect, pairwise additive and can be attractive or repulsive depending on the charge of the particles. The induction (polarization) effect is the resultant of a net electric field on the atom by the environment. It is an attractive effect but non-additive because the electric field of the atoms in the environment can cancel each other out. The last interaction (dispersion effect) has an additive effect, which mainly takes place when the two molecules move closer to each other. This effect, also known as the instantaneous dipole–induced dipole effect, is due to fluctuations in the electron distribution over time, which are the nearby molecule feels and responds to.

The resonance and magnetic contributions are very small, and for practical reasons, they can be neglected because they are not involved in the molecular association process of most binding events at room temperature.

I.A. 6 - THE SHORT-RANGE INTERACTIONS

Short-range interactions are dominated by the exchange-repulsion force: a quantum mechanical effect resulting from two opposing effects, one attractive and the other repulsive. When the distance between two atoms decreases, their electron clouds approach each other, and their charge distributions gradually overlap. The Pauli exclusion principle prohibits all the electrons

from occupying the overlap region and reduces the electron density in this region. The positively charged nuclei of the atoms are then incompletely shielded from each other and therefore exert a repulsive force on each other. Thus, the electron overlap increases the system's total energy and gives a repulsive contribution to the interaction. In practice, the repulsive component is greater than the attractive component, so the resultant of the exchange-repulsion interaction has a net repulsive effect. Compared to the exchange-repulsion, the exchange-induction, the exchange-dispersion, and the charge transfer have a neglected effect.

II. C - HOST-GUEST CHEMISTRY

Host-guest chemistry is a concept related to supramolecular chemistry, where complexes are composed of at least two molecules (host + guest), for which the theory of non-covalent interactions describes the strength of the interaction.[29] Host-guest chemistry is strongly linked to the concept of molecular association, where the two components of molecular complexes are held together by non-covalent interactions (hydrogen bonds, π-π interactions...) and for which the binding mode is specific to its molecular interactions.[30]

In supramolecular chemistry, the hosts have a well-defined classification. They can be (i) constituted by a monomeric scaffold that provides the information about the family of the host: like the cucurbituril[n] where n represents the numbers of monomers or (ii) constituted by a particular chemical scaffold that gives the identity of the host: like the metalloporphyrin complexes.[31]

In general, and compared to proteins, molecular hosts have a low molecular weight and exhibit a much lower degree of freedom. For this reason, host-guest systems were initially used as a representative model to evaluate computational methods for predicting ligand-protein binding.[32,33] Indeed, although they present an important geometrical difference, they have nevertheless a certain number of common critical points known to be fundamental in the binding mode, such as the flexibility of the receptor, the cost of solvation/desolvation, the change of protonation state, the consideration of hydrophobic effects, or tautomerism...[6,34] Due to their size, these systems are – in principle – well designed for testing computational methods owing to their comparatively low computational cost, but also due to the reproducibility of the experiments where the uncertainties in host-guest protonation are more controlled than in protein-ligands systems. Following that thought, a group of scientists composed of experimentalists and computational scientists created a challenge called SAMPL[35] (Statistical Assessment of the Modelling of Proteins and Ligands) and D3R[36] (Drug Design Data

Resource) that used during many years host-guest system to try several computational methods, comparing to experimental data blindly. The SAMPL project, initiated in 2008, has traditionally included challenges based on small molecular systems, such as the hydration-free energies of small molecules and the binding thermodynamics of host-guest systems. Since 2018 the SAMPL challenge is funded by the National Institutes of Health (NIH).[37] These past years, a broad range of biologically relevant systems with different sizes and levels of complexities, including different host-guest complexes, has been selected by the SAMPL challenges. The aim of the challenge is to test the latest modelling methods and force fields and how they perform predicting blind data. The experimental data, such as binding affinity and hydration free energy, are withheld from participants until the prediction submission deadline so that the predictive method is ranked and compared based on the performance of the systems.

These challenges provide a fair assessment of state of the art, an objective comparison of methods, and, over the years, a large available dataset that can be used for testing new methodology.[34,35,38–42] Unfortunately, the SAMPL challenges are not using supramolecular complexes, including metal compounds in various solvated environments, but more biologically related and water-soluble systems. As this is one of the only significant knowledge bases for the computational prediction of the binding free energy of non-protein molecular complexes, we consider that the analysis of previous SAMPL methodologies could give us a good overview of the accurate methods used for prediction. For the aim of the thesis, the methods will need to be adapted to suit supramolecular complexes, including simulated metal complexes in a non-aqueous environment.

III - COMPUTATIONAL METHODS FOR THE DETERMINATION OF BINDING FREE ENERGY IN HOST-GUEST COMPLEXES

III. A - STATE OF ART OF THE COMPUTATIONAL METHODS USED FOR BINDING FREE ENERGY PREDICTION

Retrospective analysis based on previous SAMPL challenges, and specifically the SAMPL6[42] challenge, gives us some interesting information about the state of the art of binding free energy prediction in host-guest complexes. Before presenting several computational methods, we must highlight that the method can depend on the system used. As we stated in the molecular association part, the binding between host and guest is realised using non-covalent interactions. The non-covalent interactions are diverse, and depending on the size of the cavity and the type of the interactions, the accuracy of the resulting prediction with one method or another can vary. Considering the computational approach for predicting host-guest binding free energy, we decided to mainly focus on what was described in the previous SAMPL challenges, as they represent – in theory – the more recent advances in terms of computational techniques. In addition to the fundamental differences between the methods, there are numerous variations for each of them, such as the force field, the partial-charge model, the solvation models used, or the sampling method.

I.A. 7 - METHODS

The different methods can be separated into two main classes: the classical molecular mechanics (MM) and the Quantum Mechanics (QM) / Semi-empirical Quantum mechanics (SQM). The classical mechanics can be used to calculate binding free energies following three approaches: End state methods[43] (MM/GBSA and MM/BPSA), the Umbrella Sampling[44], and the Perturbation methods[45] (FEP). For the end-state methods, the application of the MM consists of using the Molecular Dynamics (MD) simulations in addition to Poisson–Boltzmann or generalized Born and surface area continuum solvation (MM/PBSA and MM/GBSA) methods to predict the difference in binding free energy between the bound and unbound states (i.e. only the end states are considered). These methods are fast and moderately accurate, highly dependent on the system, and are known to overestimate the binding free energy of the host-guest complexes.

Umbrella sampling is a free energy method that allows one to probe regions of the free energy curve that would not be available using simple MD. The general idea of the method is to restrict

the reaction coordinate without constraining it by applying a bias potential to explore the dissociation process of the ligand-receptor and measure the binding free energy.[46,47]

The perturbation methods use molecular simulations (MD or Monte Carlo simulations) to compute the free energy difference between two molecular structures. The perturbation theory is an old technique that was initially used in physics in celestial mechanics to analyse the effect of bodies on the elliptical orbits of planets. These techniques have been adapted to computational chemistry to measure the free energy difference between a reference system and a target system, with the conditions that the target system is sufficiently similar to the reference system. It uses an alchemical transformation to go from molecule A (reference) to molecule B (target).

The last classes consist of the usage of QM (mixed with the MM or performed at SQM level) for the binding free energy prediction. In theory, the quantum mechanical formulation is almost exact and describes the underlying physics of the system, including all the energetical contributions. Multiple interactions or phenomena are missing in a force field and can be considered using a quantum mechanical framework (such as electronic polarization, charge transfer…). Another advantage of QM is the absence of pre-parameterisation of the system that accurately describes the chemical space. Unfortunately, practically speaking, the QM methodology faces some limitations, especially considering (i) the solvent's effect on the host-guest system, (ii) the calculation of entropic terms, and (iii) the computational cost compared to other methods.[48] In practice, QM methods can only be used to calculate binding free energies following an end-state formalism. Recent implementations of SQM allow sufficient sampling to make the Umbrella Sampling approach accessible. Perturbation methods, particularly those involving no-physical states, are not amenable to QM formalisms.

I.A. 8 - FORCE FIELDS

For host and guest structures, as they are not composed of amino acids, it is necessary to generate the parameters to represent them in a force field. Multiple different force field exists. In the SAMPL6 challenge, two different classes of force field have been used: the classical force-field (Gaff / Gaff2, CGenFF, OPLS, …) and the polarizable force field (AMOEBA).[49–52] The representation of the charge of the host and guest molecules is also a challenging problem. Multiple partial-charge models exist. The simplest ones are the Gasteiger-Marsilli charges[53], this is one of the fastest methods based on the electronegativities and connectivity of the atoms, but unfortunately, known to be too inaccurate to describe a molecule when

parameters have to be added for the force field. In the AMBER[54,55] force-field, the RESP[56] charges (or the AM1bcc[57] ones, which are considered to be a faster approximation) are preferred for the partial-charge calculation of the host-guest system for the calculation of the binding free energy. The RESP charges are derived at the Hartree-Fock level of theory and are supposed to be more accurate. Other methods exist and depend on the force field, such as the AMOEBA charges or the CgenFF charges.[58]

I.A. 9 - METHOD FOR BINDING MODE PREDICTION

As we already mentioned, to predict binding free energy, the binding mode of the host-guest complexes has to be generated. In recent years, multiple sampling methods have been tested to generate the initial estimate of the molecular complex. Most prominently, the complexes can be generated using a docking protocol.

Docking was born in 1982, and with the development of informatics capabilities associated with the important growth of co-crystallized ligand-protein in the PDB (Protein Data Bank), docking has been more often used and can now be considered as a standard approach in the panel of the computational approaches.[59]

I.A. 10 - GENERATION OF AN ENSEMBLE OF CONFORMATIONS

In most cases, docking protocol cannot accurately describe the molecular environment between host and guest. For that, some sampling methods are considered to generate an ensemble of different host-guest conformations. One of the most used techniques for sampling is MD simulations. Using the docking to generate a first guess of the host-guest complexes, their behaviour in a solvated environment is then analysed by MD simulations. Two main outcomes can be expected from the sampling procedure:

- After a few nanoseconds, the guest is stabilized in the host cavity, the molecular complex is considered stable, and the representative conformation is extracted.

- When the host-guest equilibrium is complex, multiple structures differing by their geometries are extracted corresponding to the different representative conformations of the complex.

I.A. 11 - SOLVATION MODELS

The solvation effect is known to have considerable importance in the determination of the non-covalent interactions. In MD simulations, the solvent is considered explicitly through the

inclusion of water molecules. Several water models exist (e.g., TIP3P, TIP4P[60], SPC/E[61]) that should be used with compatible force fields. Because the solvent can have long-range effects and has a multitude of degrees of freedom, to obtain equilibrium properties, it is necessary to consider multiple layers of solvation and a very large ensemble of the solvent's configurational states. When this is not possible due to the computational cost, one can use implicit solvation models (COSMO-RS[62], SMD[63], KMTISM[64]), where the solvent is treated as a continuous polarizable medium.

III. B - MOLECULAR MODELLING APPROACHES

One of the solutions for the binding free energy calculation is the thermodynamic-based approach using a mix of molecular modelling approach with the usage of SQM[65] for the binding free energy calculation. In this context, the accurate calculation of the binding free energy of host-guest systems depends on several parameters: (i) the potential energy surface of the system, (ii) the flexibility of the host, (iii) the non-covalent interaction type realised by the guest with the host, (iv) the importance of the solvent molecules in the dynamic behaviour of the complex and (v) the implication of solvent molecules in the binding.[66]

Considering the outcome of the thermodynamic based approach (i.e., binding free energy prediction), three possible cases can be highlighted:

- (i) The host-guest complex structure is known, and the binding mode can be extracted from the crystallographic data.

- (ii) The geometry of the host is known, but the location and the dynamic of the guest are not. This includes specific cases where the binding mode of several guests is known and where the investigation concerns the binding of new guests in a well-described system.

- (iii) The impact of the guest in a host changes the global dynamic of the system, and the host in the complex has a different equilibrium state than the host alone, including qualitative change in geometric structure.

The geometry of the complex is fundamental for free energy prediction.[67] Thus, the protocol for the binding-mode prediction has to be based on different algorithms that are system-dependent and correlated to the three cases presented above. An overview of the different cases is presented in the following Figure 2:

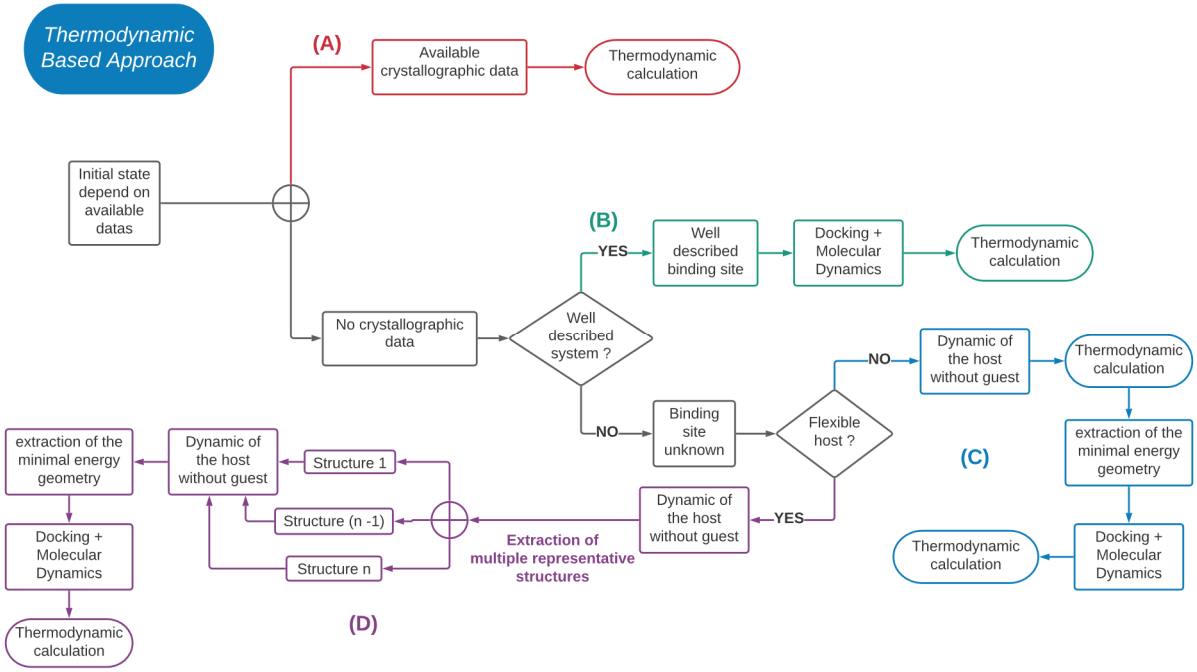


Figure 2: Flowchart of the molecular modelling approaches

When the binding is observed by crystallographic data, precise information about the binding mode and the representation of the solvent in the cavity are known. In that case, the SQM approach can be directly used to calculate the entropic and enthalpic contribution of the complex and thus determine the binding free energy. If some solvent molecules are involved in the binding, they can be considered explicitly at the SQM level to improve the prediction (Figure 2A).

In both other cases, the precise binding mode is unknown and must be determined using molecular modelling methods. In these cases, different options based on the considered system can be investigated. When the studied system is well described and then precise information about the binding site is known, only the binding mode of the guest in the host cavity has to be studied. In that case, a docking protocol can generate the first guess of the host-guest complex, followed by an MD simulation to sample the binding. Multiple representative structures of the complex can then be extracted from the simulations, and the SQM approach can be used on all the extracted structures leading to the calculation of multiple energy corresponding to multiple different complexes (Figure 2B).

In the last case, the binding mode has to be found. If the information about the geometry of the host in solvated media is missing, and the host can be considered flexible enough to change its geometry over time, a sampling procedure using MD of the host alone has to be considered, and two main cases can be highlighted: (i) In the first case, the thermodynamic properties are

calculated, and the minimal energy is extracted for the binding. However, this assumes that the minimum energy of the host without a guest is in an accessible conformation that allows the binding. This is equivalent to saying that the host does not have very important flexibility, susceptible to converge on a geometry distant from the binding geometry (Figure 2C). In the second case (ii), from an MD simulation of the host alone, multiple representative conformations can be extracted with the idea that the host alone has superior flexibility without a guest and takes on different conformations over time, one of which represents the binding conformation. A docking protocol can then be done in all the different conformations leading to multiple complexes that could be simulated by MD simulations, and followed by thermodynamic calculation at the SQM level. The major problem of this protocol is that a large number of simulations are launched on all the structures extracted from the previous step, but finally, only one represents the binding, resulting in a significant increase in the calculation time (Figure 2D).

III. C - KNOWLEDGE-BASED APPROACHES

Knowledge-based approaches offer an orthogonal strategy to predict molecular properties. As the name implies, in this case, one aims to learn from pre-existing data, hoping that it can be used to extrapolate properties of molecules for which data does not exist yet.[68] Humans are well adapted to extrapolate from specific examples but lack the ability to deal with large volumes of data. Since man started to have scientific reasoning, they tried to invent a machine able to imitate human reasoning. But the term artificial intelligence is a very recent concept born in the mind of the mathematician Alan Turing in 1950.[69] Alan Turing, in his book "Computing Machinery and Intelligence", mentioned for the first time the notion of artificial intelligence. He described a test known as the Turing test, in which a subject interacts with another human and then with a machine programmed to formulate a meaningful response. If the subject is unable to make the difference between the machine and the human, then the machine passes the test and can be considered “intelligent”. A more recent definition of artificial intelligence stated: The artificial intelligence (AI) is a process of imitating human intelligence based on the creation and the application of algorithms. The final goal of AI is to enable computers to think and act like humans being.[70,71]

AI is a very general concept in which we can find the scientific algorithms: Machine Learning (ML) and Deep Learning (DL).[72] ML is an evolving branch of computational algorithms based on mathematical and statistical approaches designed for the emulation of human intelligence

by using the surrounding environment as a learning step. It gives the computers the capabilities to learn from pre-existing data and finally increase their performances to solve processes without being explicitly programmed to do so.[73] DL can be considered as a subtype of ML where the learning algorithms represent a specific type of ML algorithms based on artificial Neural Networks (NNET) where the machine can train itself and improve the prediction based on this self-learning.[74]

An overview of the previous definitions is represented in the Figure 3:

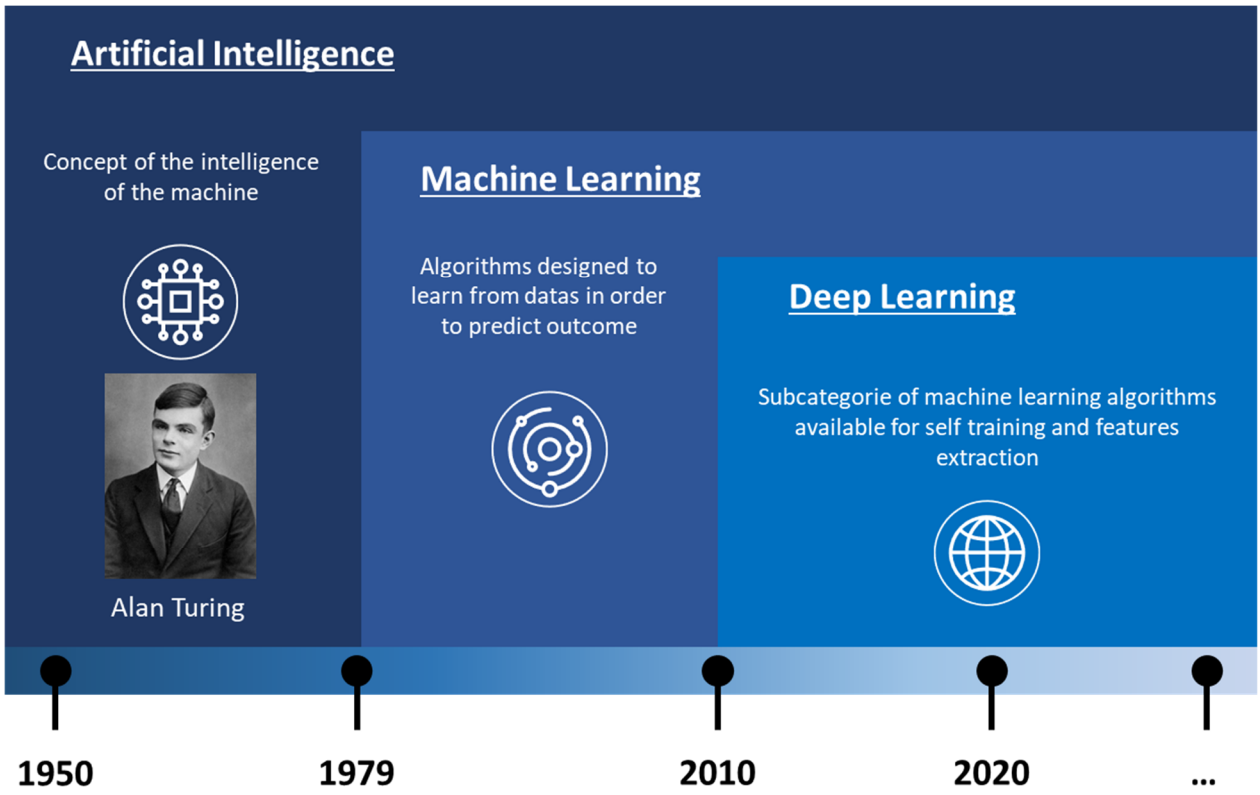


Figure 3: Artificial intelligence history and description inspired by nvidia[75]

In the field of host-guest chemistry, the artificial intelligence techniques are not well used, but as we said in the previous part, the supramolecular science is expanding, and the quantity of available data is also growing, giving a nice opportunity to develop a knowledge-based method for the explorations of the host-guest complexes.

IV - COMPUTATIONAL EXPLORATION OF HOST-GUEST COMPLEXES

IV. A - PRESENTATION OF THE NOAH PROJECT

This thesis is a part of the NOAH (Network Of functional molecular containers with controlled switchable Abilities) European consortium. The NOAH project is a European project involving ten early-stage researchers (ESRs) in a multidisciplinary chemical research program in the area of functional molecular containers, encapsulation processes, and their applications. This multidisciplinary project involves academic teams and industrial partners across Europe, and except for the University of Barcelona that is carrying the computational part of the project, each of the involved groups has expertise in the synthesis and experimental study of a specific type of molecular container, and skills in specialized techniques for their characterization and an interest in a specific type of functional behaviour.

IV. B - GENERAL AIM OF THE THESIS PROJECT

Because of the increasing complexity of supramolecular systems, the computational analysis of the host-guest complexes has become indispensable in the field of supramolecular research to understand the complex behaviour of these systems under different stimuli or to select from databases suitable guests for a given host system. At the moment, breakthrough discoveries in molecular host-guest chemistry are hampered by the complexity of the thermodynamic and kinetic characterization of the inclusion/release processes, which make it difficult to generate useful predictions about molecular encapsulation.[76] Quantitative prediction of binding energies is particularly difficult but fundamental because they are associated with a loss of time and money due to the effort to synthesize, characterize and test several compounds that are finally not active. We aim to develop computational tools to explore the interactions between host and guest molecules.

A brief overview of the steps done in this thesis can be shown in the Figure 4:

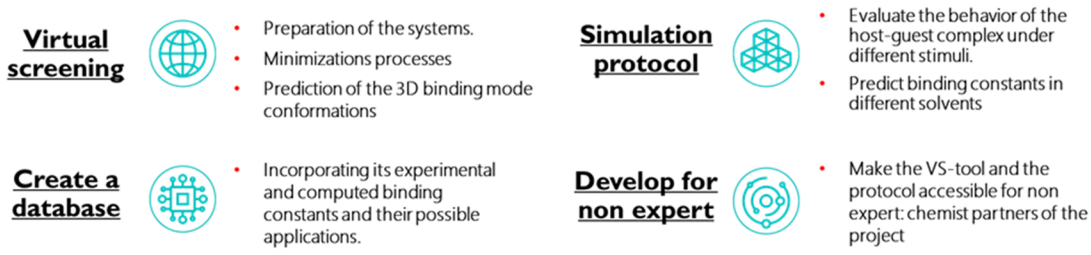


Figure 4: General aim of the project: multiple approaches we used for the exploration of the host-guest complexes

Although these steps are separated here, they were carried out simultaneously. It led us to create a computational platform we called HG-DYNAusor (Host-Guest DYNamic and Automated application for binding free energy prediction). In general, the virtual screening tool is used to prepare the host and the guest and to generate the binding mode interaction between the host and the guest systems, which are necessary for almost all further analysis. We developed a simulation protocol to try to predict the binding constants of the complexes in different solvents and evaluate the behaviour of those complexes under different stimuli. All the computed binding free energies associated with the experimental values provided by the project partners have been stored to be compared and refined during the PhD. The knowledge-based methods mainly depend on the databases, thus can be considered independent of the screening and simulation steps.

This thesis takes place in a multidisciplinary project involving mainly synthetic chemists who work on different systems. In this context, the project focused on developing a computational platform for the analysis of a broad range of host-guest systems, including the ones synthesized by the academic and industrial partners of the project. In this way, we aim to improve the global knowledge in the field of supramolecular chemistry and provide new opportunities and applications for existing containers and provide direction in the rational development of containers with new activities.[77]

At the end of the thesis project, several months have been dedicated to the chemical synthesis of supramolecular complexes, with the aim of measuring binding free energy of host-guest complexes and comparing them with the computational prediction. After a brief presentation of the computational methods used in Chapter 2, the computational platform we developed during the thesis will be presented in Chapter 3. Then several important results obtained with the platform using SAMPL challenges and molecular structure provided by the NOAH partners will be presented (Chapter 4). In the end, the synthetic chemical procedure, including the computational protocol and the perspective, will be presented in Chapter 5.

2

COMPUTATIONAL METHODS

I - SIMULATIONS METHODS

In this part, we will describe the methods used during the thesis.

I. A - QUANTUM MECHANICS (QM)

QM was initially developed at the beginning of the 20th century by several scientists (Heisenberg, De Broglie, Einstein, Bohr …). At this moment, QM is the most precise theory to describe molecular systems. The application of QM is based on the Schrödinger equation that is described in the Born-Oppenheimer approximation.[78] For a system of N electrons and A nuclei, the equation is presented such as:

$$H.\psi(\boldsymbol{r}) = E.\psi(\boldsymbol{r}) \qquad 1$$

where $\psi(\boldsymbol{r})$ represents the polyelectronic wave function.

With:

$$\widehat{H} = -\sum_{i=1}^{N}\frac{1}{2}\nabla_i^2 + \sum_{i=1}^{N}\sum_{j>i}^{N}\frac{1}{r_{ij}} - \sum_{A=1}^{M}\sum_{i=1}^{N}\frac{Z_A}{r_{iA}} \qquad 2$$

In the equation (1), H is the Hamiltonian operator, E is the total electronic energy of the system, and ψ is the wave function associated, describing the exact behaviour of the electron of the system. The Hamiltonian development (equation 2) can be separated into three different terms:

- The first term corresponds to the mono-electronic Hamiltonian that is calculated for every electron of the system.
- The second term corresponds to the bi-electronic Hamiltonian, used to determine the interaction between the electron.
- The third and last term corresponds to the electrostatic interaction between the nuclei and electrons.

The electron belongs to the fermion family, implying the fact that following the Pauli-exclusion principle, and in consequence, the wave function has to be antisymmetric. The slater formalism allows us to define the polyelectronic wave function ($\psi(\boldsymbol{r})$) as the determiners of orthogonal monoelectronic wave functions ($\varphi_i(\boldsymbol{r_i})$):

$$\psi(\boldsymbol{r}) = \frac{1}{\sqrt{N!}}\begin{vmatrix}\varphi_1(\boldsymbol{r_1}) & \cdots & \varphi_N(\boldsymbol{r_1})\\ \vdots & \ddots & \vdots\\ \varphi_1(\boldsymbol{r_N}) & \cdots & \varphi_N(\boldsymbol{r_N})\end{vmatrix} \qquad 3$$

The analytic resolution of this system corresponding to the second term of the equation 2 (i.e., interaction between the electrons) cannot be calculated in most cases, leading to an inability to solve the equation for almost all of the non-hydrogens-like systems. For that reason, to solve this equation, we need to use approximations given by *ab-initio* methods (Hartree Fock (HF) and post-HF) and electronic density functional methods (DFT).

I. A. 1 - HARTREE FOCK (HF) METHOD

The HF method is used to solve Schrödinger's equation approximatively, assimilating the bi-electronic integrals in a coulombic interaction integral (J) and an exchange integral (K). This method describes a novel operator (F) called Fock operator such as:

$$F.\varphi = E.\varphi$$

$$\hat{F} = -\sum_{i=1}^{N}\frac{1}{2}\nabla_i^2 - \sum_{A=1}^{M}\sum_{i=1}^{N}\frac{Z_A}{r_{iA}} + \sum_{i=1}^{N}\sum_{j>i}^{N} J_{ij} - K_{ij} \qquad 4$$

The integral forms are known, so it is possible to solve the equation. But the wave function is correlated to the J and K parameters, and J and K also depend on the wave function directly. For this reason, a linear resolution of the equation could not be envisaged. The self-consistent field (SCF) method is used to resolve the equation based on a set of basis functions called basis-set.

This initial basis-set is used to calculate for the first time the J and K parameters, and these parameters are used to recalculate the orbitals and so on until the convergences of the energetical parameters. The initial basis set could be more or less developed based on the expectations. The more the basis is developed, the more accurate the results will be, but the calculation will take longer.

This approximation led us to resolve with accuracy the Schrödinger equation. However, the previous approximation made the calculated energy larger than the exact solution (this difference is called correlation energy). From a physical point of view, the electron's position in a given timeframe is not independent. We say then the electrons are correlated, which can induce a considerable difference between the HF energy and the real energy of a system. For that reason, several *ab-initio* methods called post-HF can lead to a better approximation of the energy, and therefore to get closer to the exact solution, but they increase the computational time exponentially.

I. A. 2 - DENSITY FUNCTIONAL THEORY (DFT) METHOD

In the past decades, DFT methods have become very popular for calculating the energies and the properties of the molecules. The theory behind the DFT developed by Hohenberg and Kohn in 1964[79] states:

- The electron density of a system in its ground states is sufficient to determine the energy: the energy can be described as a density function. The exact energy of a system corresponds to the global minima of the function of density.
- For an external potential (Vext), any system composed of multiple interactive particles will have a unique electronic density.

In theory, DFT is an *ab-initio* method, but it could be considered as an empirical method as it is necessary to approximate the energy functional.

As the energy of the system is a function of the density, $E[\rho]$ can be written as:

$$\begin{aligned} E[\rho] &= V_{ne}[\rho] + T[\rho] + V_{ee}[\rho] \\ &= \int \rho(\boldsymbol{r})v(\boldsymbol{r})d^3\boldsymbol{r} + T[\rho] + V_{ee}[\rho] \end{aligned} \qquad 5$$

In the equation (5), $T[\rho]$ represent the kinetic energy associated with the given electron density, $V_{ee}[\rho]$ is the electron-electron interaction energy, and $V_{ne}[\rho]$ is the interaction energy between the electron and the "external" field. The formal problem is that the functionals $T[\rho]$ and $V_{ee}[\rho]$ are not known, and it is impossible to determine them using the actual knowledge. For that reason, in 1965, Kohn and Sham[80] developed an iterated (auto-coherent) equation describing the energy of a system as a function of density in the presence of an external potential. In the equation below (6), the starting point is a deterministic wave function for N non-interacting electrons in N orbitals φ_i. For this system, the electron density and the kinetic energy are exact:

$$\rho(r) = \sum_i |\varphi_i(\boldsymbol{r})|^2 \ , T_s[\rho] = -\frac{1}{2}\sum_i \langle\varphi_i|\nabla^2|\varphi_i\rangle \qquad 6$$

While the coulomb (classical) part $J[\rho]$ of the electron repulsion energy can also be calculated easily. The energy functional now takes the form of the equation (7):

$$E[\rho] = V_{ne}[\rho] + T_s[\rho] + J[\rho] + E_{xc}[\rho] \qquad 7$$

Where $T_s[\rho]$ is called the Kohn–Sham kinetic energy and $E_{xc}[\rho]$ is the *exchange-correlation functional* that is described in the following equation (8):

$$E_{xc}[\rho] = T[\rho] - T_s[\rho] + V_{ee}[\rho] - J[\rho] \quad (8)$$

The orbitals φ_i, satisfy the Kohn and Sham equations:

$$K\varphi_i = \varepsilon_i\varphi_i \quad (9)$$

Where:

$$K = -\frac{1}{2}\nabla^2 + v(\boldsymbol{r}) + \int \frac{\rho(r')}{|\boldsymbol{r} - \boldsymbol{r}'|} d^3\boldsymbol{r}' + v_{xc}(\boldsymbol{r}) \quad (10)$$

And $v_{xc}(\boldsymbol{r})$, the exchange-correlation potential is a functional derivative of $E_{xc}[\rho]$. $E_{xc}[\rho]$ Characterizing the correlation and exchange (directly calculated by the DFT methods, unlike HF methods). The exchange-correlation functional $E_{xc}[\rho]$ and its potential $v_{xc}(\boldsymbol{r})$ are still unknown and need to be approximated to solve the Kohn and Sham equation.

Several functionals have been developed these past years to represent the exchange-correlation contributions. These functionals can be separated for practical reasons into three parts: LDA, GGA, and hybrid-GGA.

- LDA (*Local density approximation*) functional is based on the theory of the uniform electron gas; the inhomogeneity of the system is neglected. In that case, the energy of the exchange-correlation $E_{xc}^{LDA}[\rho]$ can be written like that:

$$E_{xc}^{LDA}[\rho] = \int d^3\boldsymbol{r}\rho(\boldsymbol{r})\varepsilon_{xc}[\rho(\boldsymbol{r})] \quad (11)$$

  With $\varepsilon_{xc}[\rho(\boldsymbol{r})]$ representing the energy of exchange-correlation per electron. The LDA approximation is well adapted for small isolated molecules, which present low variation in their density. Still, in most cases, this method is too inaccurate to be useful for chemistry.

- The GGA (*Generalized gradient approximation*) functional has been developed to deal with the inaccuracy of the LDA. This time the GGA functionals take into account a density gradient such as:

$$E_{xc}^{GGA}[\rho] = \int d^3\boldsymbol{r}\rho(\boldsymbol{r})\varepsilon_{xc}^{GGA}[\rho(\boldsymbol{r}), \nabla\rho(\boldsymbol{r})] \qquad 12$$

Hybrid-GGA functionals have been introduced to improve the description of exchange-correlation energy. These functional results to the combination between the HF exchange and the LDA/GGA functionals. The logic behind this combination is based on the fact that as the HF calculations are treating the exchange energy correctly, they can be combined empirically with LDA or GGA density functionals. The best-known and most widely used of this hybrid-GGA is the B3LYP functional developed in 1993 by Becke[81] based on the adiabatic connection between the real system and the fictive system introduced in the Kohn-Sham approach (described in the previous part). For the initial work, the B3LYP functional has been used in this thesis.

I. A. 3 - BASIS SET

As previously stated, the numerical resolutions of the Kohn-Sham equation (9) need the usage of a basis-set. The choice of the basis set is a very complex question, and generally, one wishes to use the largest possible basis-set to improve the results. Still, the cost of the calculation increases very sharply with the size of the basis set. The main reason explaining that is the two-electron integral calculations that have the form of:

$$(ij|kl) = \int\int \chi_i(1)^*\chi_j(1)\frac{1}{r_{12}}\chi_k(2)^*\chi_l(2)\, d\tau_1\, d\tau_2 \qquad 13$$

Where χ_i corresponds to the basis functions. Almost $\frac{1}{8}N^4$ two-electrons integrals are considered for a calculation involving N basis function. These functions χ_i are centred on every atom of the system and define what we call a basis set.

They are generally composed of two parts:

- (i) The radial part:
    - Which can be either orbital slater type:

$$\chi = R_{lk}(\boldsymbol{r})\, exp^{-\varsigma r}) \qquad 14$$

- Or gaussian type:

$$\chi = R_{lk}(\boldsymbol{r}) \sum_i C_i N_i \, exp^{-\alpha_i^2} \qquad 15$$

- (ii) The angular part $R_{lk}(\boldsymbol{r})$ is a type of a spherical harmonic function.

Concerning all these basis sets, and the spherical harmonic functions are always used to describe the angular part of the orbital, meaning that only the radial part of the orbital changes.

I. B - SEMI-EMPIRICAL QUANTUM MECHANIC (SQM)

I. B. 1 - THEORY BEHIND

The SQM methods are the simplest variant of electronic structure theory. These methods involve approximations that could limit their accuracy but make them very efficient and able to be used as screening methods. The SQM methods cannot be described as *ab-initio*, practically they are simplified versions of HF theory.[82] As presented in the previous part, the computation of the two-electrons integrals is the most time-consuming step of the true *ab-initio* SCF calculations. The SQM methods use empirical corrections derived from experimental data in order to neglect most of them and approximate most of the rest. The one-electron integrals are also approximated.

These methods are usually referred to by acronyms. The most frequently used methods (AM1[83], PM6[84]) are all based on the Neglect of Differential Diatomic Overlap (NDDO[85]) integral approximation that belongs to the class of Zero Differential Overlap (ZDO[86]) methods, in which all two-electron integrals involving two-centre charge distributions are neglected.

These SQM methods are widely used to calculate structures and charges distributions for molecules that are too large for standard *ab-initio* methods. They are also often used to obtain a first geometrical structure of a large molecule in order to save time in a subsequent geometric optimisation using *ab-initio* methods.[87,88] Additionally, these past years, SQM methods have reached a really good level of sophistication, and they are now providing really good results.[89,90] However, historically, SQM methods were generally using heats of formations data to be calibrated, supposing their inaccuracy for the calculation of small energy differences of molecular complexes such as binding free energy.[48] The recent growth in those SQM methods, with the usage of different methods for the calibration, lead us to study these new methods for binding free energy prediction. Particularly, the development of new SQM methods has

recently seen a renewed interest triggered primarily by the advent of the density-functional tight-binding (DFTB) method pioneered by Seifert, Elstner, and Frauenheim.[91]

We made the hypothesis that the new SQM methods, because of their computational performances, will allow us to carry out thermodynamic calculations with an acceptable precision on systems which, because of their size, could not have been studied via *ab-initio* methods considering the calculation times.

I. B. 2 - THE GFNn-*xTB* METHODS

Initially, the GFN family of methods was designed as a special-purpose tool focusing on molecular properties. The idea behind these methods is to be able to describe with relatively good accuracy the **G**eometries, the **F**requencies, and the **N**on-covalent interactions leading to the acronym **GFN**.[92] These past years, these methods have grown really fast, the first version (called GFN1-*xTB*) employs the same approximations for the Hamiltonian and for the description of the electrostatic energy as the DFTB3[93] basis without relying on an atom pair-wise parametrisation. Instead, the ZDO type method is used. One of the advantages of the GFN-*xTB* method is the covering of a large part of the periodic table (up to Z =86). For that reason, the GFN-*xTB* methods have been considered in that thesis because they are known to be fast, robust, reasonably accurate, and works for many metallic systems.

In early 2019, the GFN2-*xTB*[65] method was released with the *xTB* software leading us to use this method in this thesis work. Figure 5 shows an overview of the GFN family:

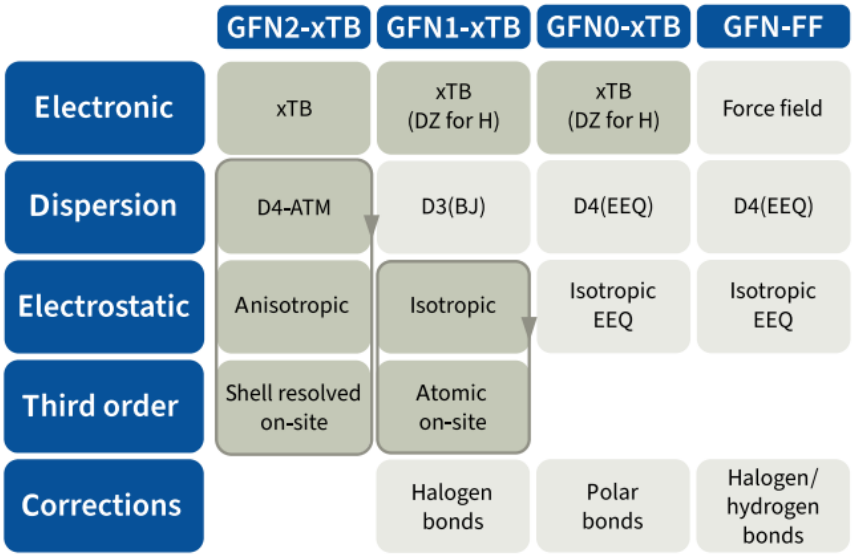


Figure 5: Overview of the GFN family of methods with main components and classification of the most important terms. Dark grey shaded areas denote a quantum mechanical description, while light grey parts indicate a classical or semi-classical description[94]

I. B. 3 - TIGHT-BINDING THEORY

xTB methods, like all the related DFTB methods, are rooted in the Kohn–Sham density-functional theory and formally represent an SQM approximation to the latter. In the following part, the connection of the *xTB* methods to DFT, DFTB, and classical FFs will be highlighted.

Starting from a non-local DFT energy expression, we get:

$$E_{tot} = E_{nn} + \sum_{i}^{N_{MO}} n_i \int \psi_i(\boldsymbol{r}) \left[\hat{T}(\boldsymbol{r}) + V_n(r) + \varepsilon_{XC}^{LDA}[\rho(r)] + \frac{1}{2}\left(\frac{1}{|r-r'|} + \Phi_C^{NL}(r,r') \right) \rho(r') dr' \right] \psi_i(\boldsymbol{r}) dr \quad 16$$

With ψ_i, the molecular spatial orbitals with occupation n_i. The kinetic operator is represented by $\hat{T}(\boldsymbol{r})$ and $V_n(\boldsymbol{r})$ is the coulombic operator due to the interaction with the nuclei. $\varepsilon_{XC}^{LDA}[\rho(r)]$ is the expression of the predefined exchange-correlation (XC) function. Finally, the kernel $\Phi_C^{NL}(r,r')$ is used to obtain the inner integral over r' containing the nonlocal (NL) correlation and the interelectronic Coulomb term. With this term, we can find that in the same case as the intermolecular force fields methods, the dispersion interaction naturally occurs. The idea using the GFN-*xTB* method is to use a system of formally independent particles, from which the density can obtain as:

$$\rho(r) = \sum_{i}^{N_{MO}} n_i \int \psi_i^*(\boldsymbol{r}) \psi_i(\boldsymbol{r}) dr \quad 17$$

Then the total energetical term is reformulated in terms of (i) a reference density ρ_0 (ideally close to the final converged density ρ) and (ii) a density difference $\Delta\rho$ with $\rho = \rho_0 + \Delta\rho$. Allowing us to decompose the energy in the form of a Hartree energy at the reference density:

$$E_{tot} = E_0^H + \Delta E^H + E_{XC}^{LDA}[\rho] + E_C^{NL}[\rho, \rho'] \quad 18$$

This equation (18) is totally equivalent to the equation (16) just reformulated in terms of the difference of density $\Delta\rho$. In DFTB methods, the total energy is expanded using Taylor series around $\Delta\rho = 0$ such as:

$$E[\rho] = E^{(0)}[\rho] + E^{(1)}[\rho_0, \delta\rho] + E^{(2)}[\rho_0, \delta\rho^2] + E^{(3)}[\rho_0, \delta\rho^3] + \cdots \quad 19$$

The GFN2-*xTB* approach truncate this expansion after the third-order term given, and for the second term, a self-consistent version of D4 based on the Mulliken charges derived from GFN2-*xTB* are used.

I. C - MOLECULAR MECHANICS (MM)

I. C. 1 - MOLECULAR MODELISATION

MM represent a further step in the simplification of the molecular systems, where the electrons are removed altogether, and their effects are represented by simple mathematical functions. In this manner, each atom is represented as a particle, and the molecular system can be described using classical newton mechanics, where there is a straightforward relationship between atomic coordinates and the energy of the system. Carrying out this calculation for the different conformations of the considered structure allows the characterization of other points of the potential energy surface associated with the system. Therefore, each point of the surface is associated with a probability of sampling and induces a statistical procedure for the simulated molecular conformations. The potential energy of a system can be described as follow:

$$E_{potential} = E_{\text{bonded}} + E_{\text{non-bonded}} \quad (20)$$

With:

$$E_{\text{bonded}} = E_{bonds} + E_{angle} + E_{dihedrals} \quad (21)$$

And:

$$E_{\text{non-bonded}} = E_{electrostatic} + E_{vdW} \quad (22)$$

The bonded terms correspond to the contributions of the bonds, the angles, and the dihedrals. The atoms are considered as a non-deformable sphere described by a charge and a Van der Walls radius. The bonds that link the atoms and associated angles are represented by a harmonic potential. Dihedrals angles are, as far as they are concerned, modelized by a sinusoidal function that allows the characterization of all the states that can be sampled by a torsion of those angles.

$$E_{bonds} = k(x - x_{eq})^2 \quad (23)$$

$$E_{angle} = k(\theta - \theta_{eq})^2 \quad (24)$$

$$E_{dihedral} = \frac{V_n}{2}(1 + \cos[i\varphi - \gamma]) \quad (25)$$

E_{bonds}, E_{angle}, and $E_{dihedral}$ represent the energy of deformations of the considered bonds and angles and also the rotation of the dihedral angle. x_{eq} and θ_{eq} from the equation 23 and 24 are respectively the distance and the angle at the equilibrium of the considered atoms. The force constant potential is given by k and (V_n for the dihedrals). More thereof, the higher the binding (or angle) that represents this potential will be strong, involving a drastic increase in energy in the event of removal of the equilibrium value. For the dihedral angle, φ represent the values of the angle while γ represents the phase of the angle.

To control the physics of the system, it is necessary to consider the intermolecular terms to represent the interactions between two distant atoms. As presented above, these terms are the electrostatic interaction term and Van der Waals interactions term. Such as:

$$E_{electrostatic} = \sum_{j>i}^{i} \frac{q_i q_j}{\varepsilon d_{ij}} \quad 26$$

And:

$$E_{vdW} = \sum_{j>i} \left(\frac{\sigma_{ij}}{d_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{d_{ij}}\right)^{6} \quad 27$$

σ is called vdW radius and is equal to the distance at which the intermolecular potential between the two particles is zero. It gives an idea of how close two nonbonding particles are. The Van der Waals bond potential is often approximated by a Lennard-Jones potential of the form:

$$E_{LJ} = 4\varepsilon \left[\left(\frac{\sigma_{ij}}{d_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{d_{ij}}\right)^{6}\right] \quad 28$$

The electrostatic energy is calculated based on the coulombic potential and is dependent on the dielectric constant. In equation (26), i and j represent two atoms with a respective charge q_i and q_j, separated by a distance of d_{ij} and a dielectric epsilon (values >1 can be used as a crude implicit solvation method).

For the vdW energetic term, the equation considers (a) the interaction between atoms distanced by three bonds at a minimum and (b) the interaction between non-bonded atoms. It allows to take into account the weak interactions that can be separated into three-part:

- The London dispersion force is the dominant contribution of the vdW term and describes the instantaneous dipole-induced dipole attraction. Due to the permanent motion of the electron, a molecule can develop and temporary (instantaneous) dipole when the distribution of the electrons is unsymmetric about the nucleus.
- The Debye force represents the contribution between the induced dipole and permanent dipole.
- The Keesom force represents the contributions between permanent dipoles (dipolar molecules).

The three forces that are involved above are both attractive where they represent the term to the six in equation 26 ($\left(\frac{\sigma_{ij}}{d_{ij}}\right)^6$) and repulsive, described by the term to the twelve $\left(\frac{\sigma_{ij}}{d_{ij}}\right)^{12}$. σ_{ij} represent the equilibrium distance between two atoms *i* and *j*.

These parameters used for these methods have been determined experimentally (using XRD structures) or using QM and are regrouped in a force field. There are many force fields (CHARMM, Martini, AMBER, OPLS, GROMOS…). In our case, the central force field we used is the AMBER ff14SB that is applicable for the simulation of the molecules that present a biological interest.

All the difficulty of the molecular simulations lies in the sampling of the conformational geometries. This sampling has to be representative to describe the conformational space of the system in a reasonable time. Practically, two methods exist to generate the conformations of a system: the MD and the Monte-Carlo simulations. The Monte Carlo simulation is a stochastic method that explores the configuration of the system. The MD simulation is a technique based on the analysis of the system's particles temporal evolution. During this thesis, we have used MD to explore the configurations of the systems.

I. C. 1. A - ENERGETICAL MINIMIZATION

In molecular modelling, to simulate a system, a pre-minimization is required in the preparation phase of the system. This energy minimization makes it possible to correct any defects in the initial structures, adapt the system to the force field, and then find the minimum of a novel surface of potential energy associated with the used method. For this reason, the initial structure is fundamental. In the best cases, we start from the crystallographic structures (obtained by XRD or NMR).

The potential energy surface is highly correlated with the energy of the system and depends on the geometry. The general principle of minimization is to modify the geometry of the system by changing some atoms to decrease its energy and force it into a potential well corresponding to a possible energetic minimum. Figure 6 shows the minimization principle in a schematic one-dimensional energy surface and the minima that would be obtained starting from three different geometries A, B, and C.

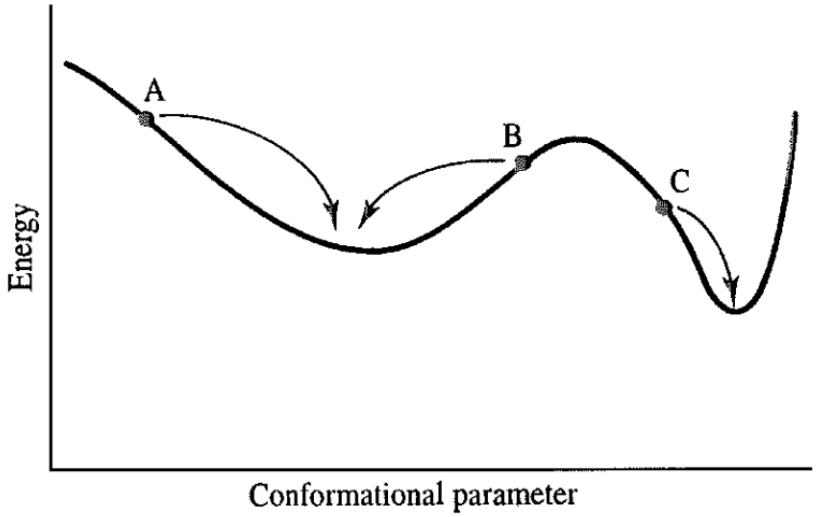


Figure 6: A schematic one-dimensional energy surface: starting from three different geometries, the minimization methods move downhill to the nearest minimum[95]

Different algorithms exist to perform the energetic minimization: the steepest-descent[96] and conjugated-gradient[97] are the most used algorithms and the ones used in this thesis.

I. C. 1. A. (I) - THE STEEPEST-DESCENT ALGORITHMS

This method is called "steepest-descent" because it consists of finding the most significant slope, from which the function representing the energy is likely to decrease the most.

The steepest descent method moves in the direction parallel to the net force. For 3N Cartesian coordinates, the direction is most conveniently represented by a 3N-dimensional unit vector ($\vec{D}_k$) such as:

$$\vec{D}_k = -\nabla\vec{E}_{(r_k)} \qquad 29$$

The variable α defines the calculation step representing the gap between two calculations and is adjusted for each iteration. α will be increased if an energetical reduction follows the energy's direction and decreases in the opposite case. Each direction considered for approaching the

energetical minima is orthogonal to the previous one. This new vector $\vec{r}_{k+1}$ is calculated using the $\vec{r}_k$ vector from the previous state, the α variable, and the direction $\vec{D}_k$:

$$\vec{r}_{k+1} = \vec{r}_k + \alpha_k \vec{D}_k \qquad 30$$

This algorithm has multiple advantages, it is fast and can be used on a local computer with high performance. Unfortunately, he is known to have convergence problems when the geometry is close to an energetic minimum where geometric oscillations can appear around the energetical minima.

I. C. 1. A. (II) - THE CONJUGATED-GRADIENT ALGORITHM

The conjugated-gradient method is different: it produces directions that do not show the oscillatory behaviour or the steepest descent methods in narrow valleys. In the previous process, both the gradients and the direction of all the successive steps are orthogonal. In contrast, in conjugated gradient, the gradients at each point are orthogonal, but the direction is called conjugate. That means this method takes into account all the previous steps to define more precisely the direction of the vector $\vec{D}_k$. This vector is defined using the direction $\vec{D}_{k-1}$ following the equation (31):

$$\vec{D}_k = -\nabla\vec{E}_{(r_k)} + \frac{\nabla\vec{E}_{(r_k)}^{\,T} \cdot \nabla\vec{E}_{(r_k)}}{\nabla\vec{E}_{(r_{k-1})}^{\,T} \cdot \nabla\vec{E}_{(r_{k-1})}} \vec{D}_{k-1} \qquad 31$$

The main interest of this algorithm is to avoid the convergence problems due to the geometric oscillations that can appear around the energetical minima with the steepest-decent method. This algorithm improves the accuracy because the direction is adjusted at each step to optimize the energetical minimum search. Regrettably, this one also has two primary deficiencies: first, it has a low efficiency correcting structural failure in the initial geometry of the system. Secondarily, the conjugated-gradient approach is highly dependent on the initial structure because all the new directions are dependent on the previous ones, implying that if the initial geometry has defaulted, the results with the conjugated-gradient algorithm will be unreliable. In conclusion, one of the best solutions would be to use the steepest-descent algorithm to correct the initial geometry and contacts between atoms and then use the conjugated-gradient algorithm to precisely find the structure corresponding to an energetical minimum.

I. C. 2 - MOLECULAR DYNAMICS (MD)

MD is a molecular simulation method that allows sampling the phases of the system starting from an initial configuration. This method can predict a huge number of phenomena (host-guest receptor interaction, host behaviour under different solvents).

The MD follows the Boltzmann law, meaning that the probability of sampling a particular conformation bears an exponential relation with its relative energy. As the passage between two wells of potential energy can be considered a rare event, we need to extend the simulation (i.e., sample a larger number of configurations) or start from another geometry to observe these transitions. The ergodic hypothesis states that the average of a process parameter over time and the average over the statistical ensemble are the same. Assuming this hypothesis, we could determine the interaction free energy between a host and a guest and compare it to the experimental values. Of course, this will only be valid if we have observed the binding and unbinding event a sufficiently large number of times. Additionally, the MD allows us to study the dynamic behaviour under different stimuli (solvation type…).

MD follows Newton's second law of motion, which relates the sum of the forces applied to the system to its acceleration, such as:

$$\sum \vec{F} = m \times \vec{a} \qquad 32$$

Where F represents the sum of the force exerted on the atom, of mass m and acceleration $\vec{a}$. As we stated in the previous part, the force on an atom is obtained by calculating the energy of the system. And as we explained in the last part, the energy of the system (E_{system}) is a summation of the calculated $E_{potential}$ and the kinetic energy:

$$E_{system} = E_{\text{potential}} + E_{\text{kinetic}} \qquad 33$$

Considering the energy of the system, if we derive the energy as a function of the position of atom i, we can say that:

$$F_i = -\frac{dE}{dr_i} \qquad 34$$

Following that transformation, we can rewrite the equation (34) simply by rewriting the acceleration as the second derivative of the position with respect to time, giving the following equation (35):

$$-\frac{dE}{dr_i} = m_i.a_i = m_i\frac{d^2r_i}{dt^2} \qquad 35$$

This transformation gives the relation between the energy of the system and the coordinates. Practically, It is impossible to calculate the coordinates directly as there is no analytical resolution of this equation. Then the coordinates are going to be approximated using a polynomial function. First: the accelerations of the atoms are calculated using the newton law. The integration of the accelerations allows to estimate the velocity, and finally, the position is obtained by the integration of the velocity: the Taylor expansion (equation 36):

$$\begin{aligned} r(t+\delta t) &= r(t) + v(t)\delta t + a(t)\frac{\delta t^2}{2} + \cdots \\ v(t+\delta t) &= v(t) + a(t)\delta t + b(t)\frac{\delta t^2}{2} + \cdots \\ a(t+\delta t) &= a(t) + b(t)\delta t + c(t)\frac{\delta t^2}{2} + \cdots \end{aligned} \qquad 36$$

The integration step δt defined as the gap between the initial values at time t and the values at the time t_1 (equal at $t + \delta t$). Thus, the position at the time t are used to determine the positions at the time $t + \delta t$, themselves then used to obtain the positions at time $t + 2\delta t$ and so on.

The Taylor expansion allowing to reach these values necessarily leads to an approximation in the obtained results, as truncated at the 2nd term. The validity of this development is all the more significant as performed on a small value of δt. Moreover, since the objective of this algorithm is to correctly represent the motion of the atoms with respect to each other, it must be able to describe the fastest molecular motions. This constraint exists in all cases where the molecular motion must be represented by an algorithm leading to discrete values. This notion is called Nyquist frequency and is defined as the maximal frequency that a signal must contain to allow its unambiguous description by sampling at regular intervals. In the case of the molecular systems, the Nyquist frequency corresponds to the frequency of the fastest interatomic vibration and is equal to 1fs. For that reason, δt cannot be higher than this value.

Multiple algorithms that allow this estimation of the new coordinates exist. The two well-known are the Verlet[98] and Leap-frog[99] algorithms and are presented in the following part.

I. C. 2. A - VERLET ALGORITHM

The verlet algorithm takes into account the coordinates at the time $t - \delta t$ and t and the acceleration at time t to calculate the position at time $t + \delta t$.

$$\begin{aligned} r(t+\delta t) &= r(t) + v(t)\delta t + a(t)\frac{\delta t^2}{2} \\ r(t-\delta t) &= r(t) - v(t)\delta t + a(t)\frac{\delta t^2}{2} \end{aligned} \qquad 37$$

The addition of the two previous equations gave us the position at the time $t + \delta t$:

$$r(t+\delta t) = 2r(t) - r(t-\delta t) + a(t)\delta t^2 \qquad 38$$

This algorithm does not use the velocity to calculate the new positions of the atoms. It is a very stable method that has to advantage of being reversible.

I. C. 2. B - LEAP-FROG ALGORITHM

The leap-frog algorithm is divided into two integration steps in order to be more precise in the calculation of the new coordinates. The velocities are calculated at the time $t + \frac{1}{2}\delta t$ and they are used to calculate the positions at the time $t + \delta t$ such as:

$$\begin{aligned} r(t) &= r\left(t + \frac{\delta t}{2} - \frac{\delta t}{2}\right) = r\left(t + \frac{\delta t}{2}\right) - v\left(t + \frac{\delta t}{2}\right)\frac{\delta t}{2} \\ r(t+\delta t) &= r\left(t + \frac{\delta t}{2} + \frac{\delta t}{2}\right) = r\left(t + \frac{\delta t}{2}\right) + v\left(t + \frac{\delta t}{2}\right)\frac{\delta t}{2} \end{aligned} \qquad 39$$

By subtracting the two equations, we obtain the expression of the position at the time $t + \delta t$:

$$r(t+\delta t) = r(t) + v\left(t + \frac{\delta t}{2}\right)\frac{\delta t}{2} \qquad 40$$

The velocity at the time $t + \frac{1}{2}\delta t$ are calculated using the velocity at the time $t - \frac{1}{2}\delta t$ and the acceleration at the time t:

$$v\left(t+\frac{\delta t}{2}\right)=v\left(t-\frac{\delta t}{2}\right)+a(t)\delta t \qquad 41$$

As the Verlet algorithm, the Leap-frog algorithm presents as well stability and reversibility properties. The Leapfrog one is generally more precise than the Verlet integration by using a smaller integration step. It is the one that we use with the AMBER software for MD.

I. C. 2. C - THE BERENDSEN THERMOSTAT AND BAROSTAT

The microcanonical statistical ensemble (N, V, E) for which the system's total energy is conserved is the natural study set of the classical MD. But practically, our interest is to study systems close to the experimental condition, implying a temperature or a pressure defined by the environment. For that reason, which is led by the experimental condition, It is possible to introduce a thermostat and a barostat in the modelling that will fix the temperature and pressure of the system, respectively. Multiple thermostats and barostats exist.[100] One of the most known is the Berendsen one.[101]

I. C. 2. C. (I) - BERENDSEN THERMOSTAT

The Berendsen thermostat introduces the energy exchanges in the simulation to maintain constant the temperature of the system. It is a method of weak coupling that is modifying the equation of motion to introduce the first-order relaxation from the temperature T to the reference temperature T_0:

$$\frac{dT(t)}{dt}=\frac{T_0-T(t)}{\tau_T} \qquad 42$$

The relaxation of the temperature is controlled by a constant that depends on the thermostat: τ_T. In each step of the simulation, the particles' velocity is corrected to add or remove energy in the system. The correction factor $\lambda_{(t)}$ is written:

$$\lambda_{(t)}=\left[1+\frac{\Delta_t}{\tau_T}\left(\frac{T_0}{T(T)}-1\right)\right]^{\frac{1}{2}} \qquad 43$$

The constant τ_T have to be well controlled, this parameter is dependent on the system and needs to be smaller enough to maintain the temperature close to the reference temperature (T_0) but sufficiently big to avoid any perturbation of the dynamic.

I. C. 2. C. (II) - BERENDSEN BAROSTAT

Like the thermostat, the barostat of Berendsen is a weak coupling method that will dictate to the system external mechanical constraints assimilated to the external pressure. This constraint is generally isotropic but, in some particular cases, can be non-isotropic. The principle of the barostat algorithm is very close to the thermostat one. Still, in that case, the corrected parameters are the atomic positions and the mesh of the simulation box, such as:

$$\frac{dP(t)}{dt} = \frac{P_0 - P(t)}{\tau_P} \qquad 44$$

Where τ_P correspond to the time constant associated with the relaxation of the barostat. As the pressure at constant temperature is linked to the volume by the isothermic compressibility (κ_T), the coupling is carried out by correcting the coordinates of the particles and the size of the simulation box. In the case of an isotropic system in a cubic box, the correction factor $\mu(t)$ can be written:

$$\mu(t) = 1 - \frac{\kappa_T \Delta_t}{3\tau_P}\left(P_0 - P(t)\right) \qquad 45$$

All the simulations have been done at constant pressure in that thesis. The Berendsen thermostat is extremely efficient for relaxing a system to the target temperature, but once your system has reached equilibrium, it might be more important to probe a correct canonical ensemble. For that reason, we decided to move to newer barostats and thermostats in some systems, and we made a try with Nose-hoover[102] thermostat and Langevin[103] thermostat to better interact with the solvents. The principle of these newer thermostats compared to the previous algorithm is basically to extend the real system by the addition of an artificial dynamical variable (associated with a “mass" and a velocity that plays the role of a time-scaling parameter.

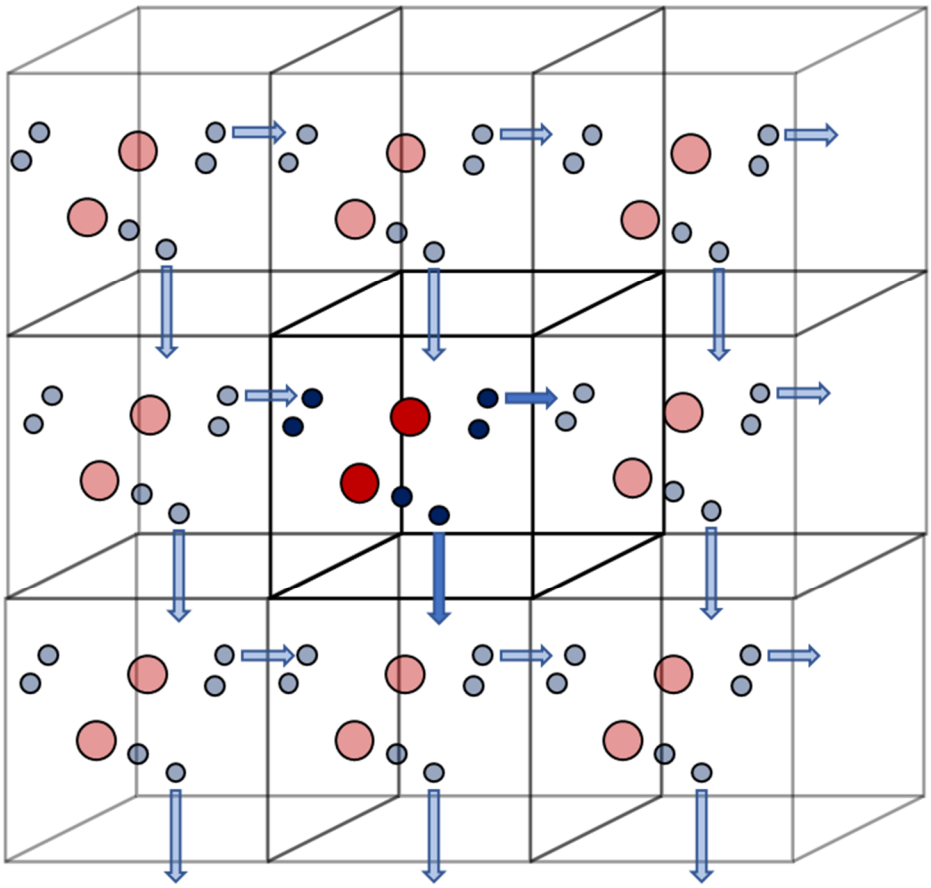


Figure 7: Schematic view of the periodic boundary conditions inspired by ISAAC program[104]

The purpose of the MD is to simulate a molecular system in order to propose a model that relates the experimental measurement with a possible explanation at the atomic level considering the motion of the different moieties of the system. To do that, we have to reproduce as precisely as we can the biological conditions, implying the solvents that are known to be involved in the structuration of the molecules and conditionate the interactions between host and guest molecules. For that, the studies of a condensed phase are realised considering a simulation box of volume V containing N particles. For that box, the energy potential of the system can be written as a double summation of their constituents:

$$E = \sum_{\alpha} \sum_{\beta>\alpha} V_{\alpha\beta}\left(R_\alpha - R_\beta\right) = \frac{1}{2} \sum_{\alpha} \sum_{\beta\neq\alpha} V_{\alpha\beta}\left(R_\alpha - R_\beta\right) \qquad 46$$

Then this system is going to be simulated in the periodic boundary conditions: the solvent box is replicated to infinity in the three directions of space. Figure 7 schematize this concept. This allows us to neglect the border effect and the dispersion of the solvent molecules: if a molecule

goes out of the box, it will reappear on the opposite side, which allows keeping a constant number of molecules in the simulation box. Then we can rewrite the previous equation to take into account the interaction between particles of the box and with all the periodic images:

$$E = \frac{1}{2}\sum_{n}\sum_{\alpha}\sum_{\substack{\beta \\ \beta \neq \alpha \; if \; n=0}} V_{\alpha\beta}\left(R_\alpha - R_\beta + n\right) \qquad 47$$

Where n is a translational vector between the simulation box and one of the images. If we consider a cubic box of size L, we have: $n = \left(n_x L, n_y L, n_z L\right)$ with $\left(n_x, n_y, n_z\right) \in \mathbb{Z}^3$. As we stated before, most of the potential is isotropic and has long-distance dependence of the form: $V(r) \sim r^{-m}$. Allow us to distinguish short-range potential ($m > 3$) from the long-range potential ($m \leq 3$).

The periodicity of the box does not cause any problems for the short-term interactions like the Lennard-Jones potential. And the equation (47) converges quickly, a truncation radius (r_{cut}) can be introduced in order to limit this summation for the terms that respect the relation: $\left\|\left(R_\alpha - R_\beta + n\right) \leq r_{cut}\right\|$. Then the successive terms corresponding

to non-zero values of n, will then be neglected. The mean error introduced by this approximation is written using a radial distribution function ($g_{\alpha\beta}(r)$) of the form:

$$\langle \Delta E \rangle = 2\pi \frac{N_\alpha N_\beta}{V} \int_{r_{cut}}^{\infty} dr \; r^{2 g_{\alpha\beta}(r) V_{\alpha\beta}(r)} \qquad 48$$

With the assumption that $g_{\alpha\beta}(r) \rightarrow 1$ at very long range, we can introduce a corrective term for the equation (48) such as:

$$E_{\alpha\beta}^{corr.} = 2\pi \frac{N_\alpha N_\beta}{V} \int_{r_{cut}}^{\infty} dr \; r^{2 V_{\alpha\beta}(r)} \qquad 49$$

For the Lennard Jones potential, this correction term is expressed as follows:

$$E_{\alpha\beta}^{LJ,corr.} = 8\pi \frac{N_\alpha N_\beta}{V} \varepsilon_{\alpha\beta} \sigma_{\alpha\beta}^3 \left[\frac{1}{9}\left(\frac{\sigma_{\alpha\beta}}{r_{cut}}\right)^9 - \frac{1}{3}\left(\frac{\sigma_{\alpha\beta}}{r_{cut}}\right)^3\right] \# \qquad 50$$

But for the long-range energetic interaction such as the electrostatic one, It is necessary to introduce some specific method to be able to obtain the energy. The descriptions of the electrostatic interaction between two charges are usually done with long-range potentials such as $V(r) \sim r^{-1}$. The electrostatic energy of the system can be written like that:

$$E_{electrostatic} = \frac{1}{8\pi\varepsilon_0} \sum_{n} \sum_{\alpha} \sum_{\substack{\beta \\ \beta \neq \alpha \; if \; n=0}} \frac{q_\alpha q_\beta}{\left\| R_\alpha - R_\beta + n \right\|} \qquad 51$$

The idea is to perform the summation over the different images (the sum on n) by order of increasing distance with the simulation box. Unfortunately, the sum converges too slowly for its direct evaluation to be possible. For that, we have to use the Ewald particle mesh method that differentiates the short and the long electrostatic interaction by using the Ewald summation. The Ewald summation will modify the equation (51) that converge too slowly by a summation of two terms that converge quickly: the charge is defined as the summation of two density: the real space + the reciprocal space (obtained by a Fourier transformation of the electrostatic potential). The following Figure 8 represents this schematic transformation:

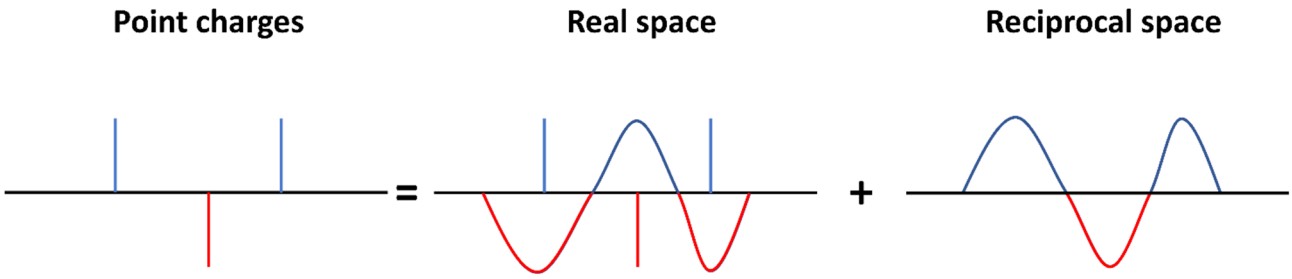


Figure 8: The schematic transformation of the Ewald summation method

The electrostatic potential of the system can be written as follow:

$$\phi(r) = \phi_{real}(r) + \phi_{reciprocal}(r) \qquad 52$$

And the electrostatic energy such as:

$$E_{electrostatic} = \frac{1}{2} \sum_{\alpha} q_\alpha \left(\phi_{real}(R_\alpha) + \phi_{reciprocal}(R_\alpha) - E_{self} \right) \qquad 53$$

With E_{self} called the self-term and representing a corrective term introduced to remove the interaction of each charge with itself. Considering explicitly the potential, the expression of the electrostatic energy of the previous equation (53) became:

$$E_{electrostatic} = \sum_{n}\sum_{\alpha}\sum_{\beta}^{*}\frac{q_\alpha q_\beta}{8\pi\varepsilon_0}\frac{erfc\left(\alpha_e\left\|R_\alpha - R_\beta + n\right\|\right)}{\left\|R_\alpha - R_\beta + n\right\|} + \frac{2\pi}{V}\sum_{k\neq 0}\sum_{\alpha}\sum_{\beta}\frac{q_\alpha q_\beta}{4\pi\varepsilon_0}\frac{exp\left(-\frac{k^2}{4\alpha_e^2}\right)}{k^2}e^{ik.(R_\alpha - R_\beta)} - \sum_{\alpha}\frac{2\alpha_e q_\alpha^2}{\sqrt{\pi}} - \frac{1}{2}\sum_{\alpha}\sum_{\beta}'\frac{q_\alpha q_\beta}{4\pi\varepsilon_0}\frac{erfc\left(\alpha_e\left\|R_\alpha - R_\beta\right\|\right)}{\left\|R_\alpha - R_\beta\right\|} \quad 54$$

Where k is a vector of the reciprocal space. The sum $\sum_{\beta}'$ relates to all atoms β that belong to the same molecule as atom α, while the sum $\sum_{\beta}^{*}$ instead relates to all atoms β that do not belong to the same molecule as atom α.

In practical terms, we introduce a truncation radius r_{cut} because the sum in the real space has a fast convergence. In the same way, the sum of the reciprocal space is dependent on the values of $\|k\|$ and It is faster when $\|k\|$ increase. The choice of the Ewald parameters $(\alpha_e, \|k\|, r_{cut})$ determine the precision and efficacy of the summation method.

II - SOLVATION MODELS

II. A - INTRODUCTION

In the molecular processes, most of the biological processes take place in a solvated environment. Solvation is known to be fundamental in molecular association and takes a very important place in the binding free energy. But finally, what is solvation? The IUPAC defines solvation as an interaction of a solute with the solvent, which leads to the stabilization of the solute species in the solution.[105]

The solvent can be classified by their dielectric constant in vacuum (ε):

- A solvent is considered as polar if $\varepsilon > 15$ (Water, Acetonitrile, Ethanol…).
- A solvent is considered as non-polar if $\varepsilon \leq 15$ (Toluene, Chloroform, Dichloromethane…).

For the polar solvent, we can add a subclassification between protic and aprotic: a solvent is called protic if it contains a labile proton that can be taken by another molecule. Aprotic solvents cannot donate protons. Three main models of solvation can be defined:

- The explicit solvation model: where the molecular details of each solvent molecule are defined in the medium.
- The implicit solvation model: where the solvents are treated as a continuous medium.
- The Hybrid model: where the two previous models are used to simulate a molecular environment, one solvation sphere is treated explicitly while the rest of the simulation box is treated using an implicit solvation model.

II. B - EXPLICIT SOLVATION

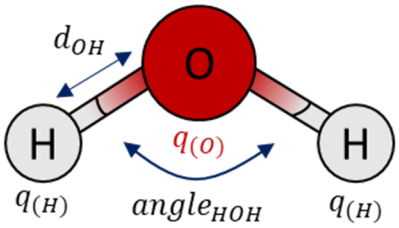
As explained before, in the explicit solvation model, the solvent is considered in the simulation as a set of molecules with all the molecular details and the possibility of the solvent molecule to interact with themselves and with the solute. For the molecular simulations, this means the solute molecule needs to be parametrised by the force field to be simulated. We know that the intramolecular and intermolecular terms will be considered all along with the simulation. The free energy of solvation will be calculated by simulating the solute-solvent interactions. We clearly understand at this step that the initial configuration is fundamental and determine the quality of the simulation.

II. B. 1 - TIP3P WATER MODEL

A solvent model is defined by its geometry: with the addition of other parameters such as the Coulombic potential and Lennard-Jones parameters.

A water molecule can be schematically represented as follow in Figure 9:

TIP3P	
d_{OH}	0.9572 Å
$angle_{HOH}$	104.52°
σ_{OH}	3.15061 Å [6]
ε	0.6364 $kJ.mol^{-1}$
$q_{(O)}$	- 0.8340e
$q_{(H)}$	+ 0.4170e



$$E_{LJ} = 4\varepsilon\left[\left(\frac{\sigma_{OH}}{d_{OH}}\right)^{12} - \left(\frac{\sigma_{OH}}{d_{OH}}\right)^{6}\right]$$

$$E_{elec.} = 4\pi\varepsilon\frac{q_{(O)}q_{(H)}}{d_{OH}}$$

Figure 9: Representation of a water molecule inside the TIP3P model with the values extracted from idc technical reference[106]

Considering the Water solvent, there exist more than 46 different models, the one presented in the figure corresponds to the most commonly used water model: the TIP3P[107]. In the TIP3P three-site are considered, and they have three interaction points (one for each of the three atoms of the water molecule).

II. B. 2 - OTHER SOLVENT MODELS

For other solvents, the principle is exactly the same, and the equivalent parameters have to be defined related to the geometry of the solvent molecules. For some of the solvent molecules, the parameters are not implemented in the AMBER suite that we used for MD simulation. In those cases (chloroform, Acetonitrile…), the molecules were optimised at QM level using DFT calculation, and RESP charges were derived, then replicated in a box and equilibrated using the AMBER MD package. The parameters and coordinates were then saved for their use in other simulations.

In conclusion, the explicit solvation model considers solvent-solvent interactions and is known to treat better the solvent-solute interaction, but they are mainly used at the MM level. To go into the SQM model and the DFT calculation, we have to take into account solvation implicitly.

II. C - IMPLICIT SOLVATION

In the implicit solvation, the solvent is treated as a continuous polarizable medium, the dielectric constant ε is fixed while the solute is simulated in the cavity in the medium.

The free energy of the solvation is given by the general equation (55):

$$G_{solv} = G_{cav} + G_{disp} + G_{elec} + G_{hb} \qquad 55$$

With G_{cav} representing the free energy required to form the solute cavity (there is an entropic cost due to the reorganization of the molecules of solvents around the solute), G_{disp} represent the Van der Waals interaction between the solute and the solvent. G_{elec} represent the electrostatic interaction term between the solute and the solvent (i.e., coulombic component) while G_{hb} represent the hydrogen bonding term. Depending on the particular implicit solvation model, these terms may differ, and not all of them need to be considered explicitly.

The way the implicit models can be implemented is represented in the following schematic Figure 10.

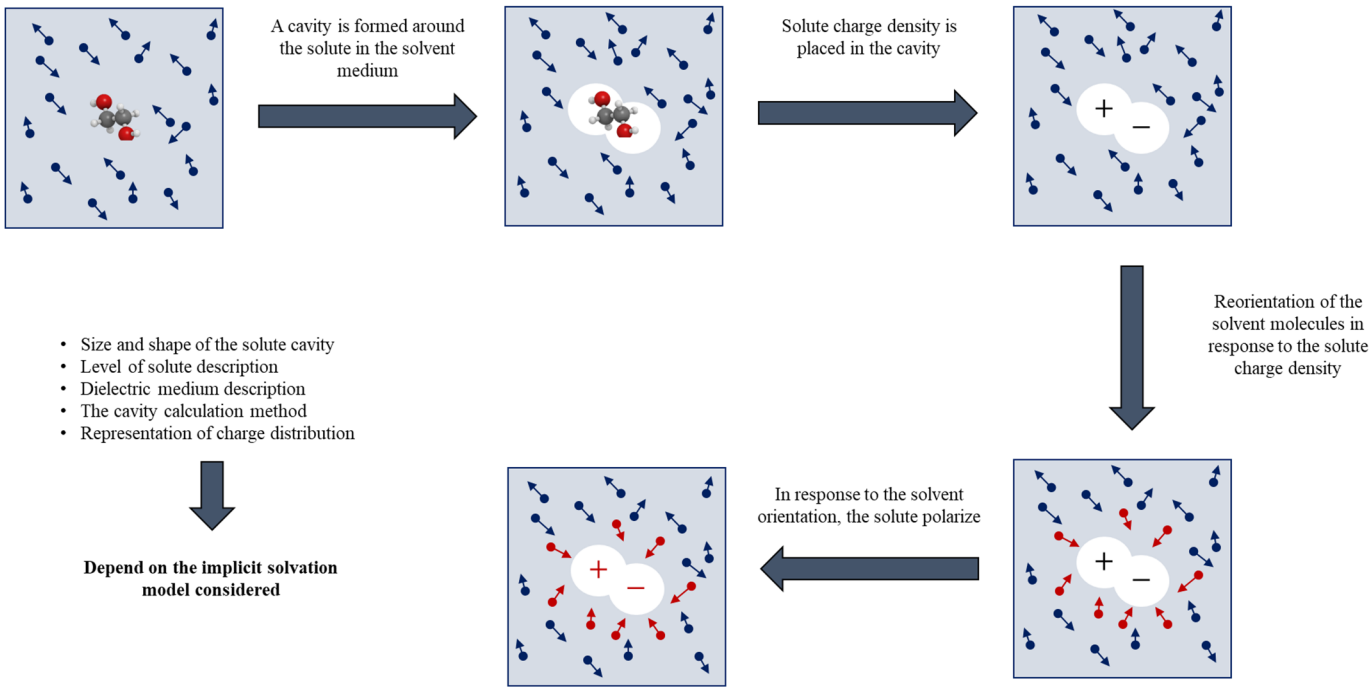


Figure 10: Implementation of implicit solvation on a solute

II. C. 1 - GENERALIZED BORN SOLVATION AREA (GBSA)

The GBSA implicit solvation model is implemented in the SQM program *xTB*. In the GBSA[94] model, a solute is considered as a continuous region with a fixed dielectric constant ε_{in} and

surrounded by infinite solvent medium with a different dielectric constant ε_{out}. In the presence of the polarized solvent, the electrostatic interaction can be expressed as follow:

$$\Delta G_{GB} = -\frac{1}{2}\left(\frac{1}{\varepsilon_{in}} - \frac{1}{\varepsilon_{out}}\right)\sum_{A=1}^{N}\sum_{B=1}^{N}\frac{q_A q_B}{\left(R_{AB}^2 + a_A a_B exp\left[-\frac{R_{AB}^2}{4a_A a_B}\right]\right)^{\frac{1}{2}}} \quad 56$$

Where $a_{A/B}$ is expressed as the effective Born radii of the respective atoms A and B. In the *xTB* Hamiltonian, the GB model is defined as second-order fluctuation in the charge density and can be described by the atomic potential V_A^{GB} such as:

$$V_A^{GB} = \left(\frac{1}{\varepsilon_{in}} - \frac{1}{\varepsilon_{out}}\right)\sum_{B=1}^{N}\frac{q_B}{\left(R_{AB}^2 + a_A a_B exp\left[-\frac{R_{AB}^2}{4a_A a_B}\right]\right)^{\frac{1}{2}}} \quad 57$$

The born radii that are required to measure the ΔG_{GB} are estimated by Onufriev–Bashford–Case (OBC) corrected pairwise approximation of the given volume as follow:

$$\frac{1}{a_A} = \frac{1}{a_{scale}}\left(\frac{1}{R_A^{cov} - R_{offset}} - \frac{1}{R_A^{cov}} \cdot \tanh[b\Psi_A - c\Psi_A^2 + d\Psi_A^3]\right) \quad 58$$

Where R_A^{cov} define the covalent radius of atom A, a_{scale} and R_{offset} are global parameters, and b, c, and d are the parameters for the OBC equation and respectively equal to 1.0, 0.8, and 4.85. We saw in this equation that the OBC correction increases the Born radii for atoms buried deep inside a molecular cavity, implying an underestimation of them logically. Ψ_A is the pairwise approximation of the given volume integral given by:

$$\Psi_A = \frac{R_A^{cov} - R_{offset}}{2}\sum_{B}\Omega(R_{AB}, R_A^{cov}, S_B R_B^{cov}) \quad 59$$

With Ω, the pairwise function is used to approximate the volume integral that mainly depends on the distance and the covalent radii. The $S_B$term is here to correct the overestimation of the volume generated by this approach by correcting the covalent radius of the second atom.

In addition to the polar contribution to the solvation energy, a non-polar surface area contribution that depends on the solvent-accessible surface area (SASA) is given by:

$$\Delta G_{SASA} = \sum_{A=1}^{N} \gamma_A \sigma_A \qquad 60$$

With γ_A representing the surface tension and σ_A the SASA of the atom A. The SASA approach is also used for the hydrogen-bond contributions such as:

$$\Delta G_{GB+HB} = \Delta G_{GB} - \sum_{A} g_A^{HB} q_A^2 \frac{\sigma_A}{A_A} \qquad 61$$

Where g_A^{HB} represent the strength of the hydrogen bonds between the considered atom and the solvent molecules and A_A represent the surface area of the free atom. Practically speaking, in *xTB*, the hydrogen-bond correction enters the Hamiltonian as a potential due to the charge dependency.

Finally, the total solvation free energy is given by:

$$\Delta G_{solv} = \Delta G_{GB+HB} + \Delta G_{SASA} + \Delta G_{shift} \qquad 62$$

An additional ΔG_{shift} is also included to correct the equation (62) (implicitly taking into account the G_{cav} and G_{disp} terms). Finally, this solvation free energy is fitted by four different parameters: the Born radius offset, the Born radius scaling, the probe radius of the solvent molecule, and the value of the ΔG_{shift}. The *xTB* implementation of the GBSA implicit solvation model also adds three parameters that are specific to the considered element; the descreening value, the surface tension, and the hydrogen bond strength.[108]

III - MACHINE LEARNING METHODS

III. A - UNSUPERVISED METHODS

Unsupervised ML uses algorithms to analyse and cluster a dataset based on labelled data. The idea behind these algorithms is to find hidden patterns to group data without the need for human interventions. The unsupervised ML methods are not designed to predict but to cluster.

III. A. 1 - PRINCIPAL COMPONENT ANALYSIS (PCA)

III. A. 1. A - INTRODUCTION

The problem of the PCA came when we studied a dataset with an important number of quantitative variables, how to plot them on a global graphic? The difficulties come from the fact that the studied individuals are not anymore represented one two-dimensional plot, but in an n-dimensional plot where n is related to the number of quantitative variables. The objective of the PCA is to reduce the dimensional space without deforming the reality of the sampling of the dataset. Mathematically speaking, the PCA consists of going from a representation in the canonical basis of the initial variables to a representation in the basis of the factors defined by the eigenvectors of the correlation matrix. The PCA has two main interpretations: (a) the statistical interpretation and (b) the geometrical interpretation.[109,110]

III. A. 1. A. (I) - THE STATISTICAL INTERPRETATION

Using the matrix of the correlation tables, we can have the coefficient of the linear correlation of the variables taken two by two: a succession of bivariate analyses. The diagonalization of that matrix gives the variances of the considered variables. The variance can be used to take into account the dispersion of a quantitative variable.

In the end, each line of the matrix corresponds to a virtual variable (called a factor) whose column gives the variance. One factor is a linear combination of the initial variables in which the coefficient is given by the coordinates of the eigenvectors. Then the PCA can be defined as the search for the linear combinations of the greatest variance of the initial variables (the eigenvalues).

III. A. 1. A. (II) - THE GEOMETRICAL INTERPRETATION

Another interpretation of the PCA is geometrical. Each individual x_i can be considered as a vector of p components in the vectorial space. The PCA is the search of the best plane of projection: the closest one considering the generalized least-squares method in order to obtain the best representation of the individuals in a reduced subspace. This concept can be visualized in Figure 11:

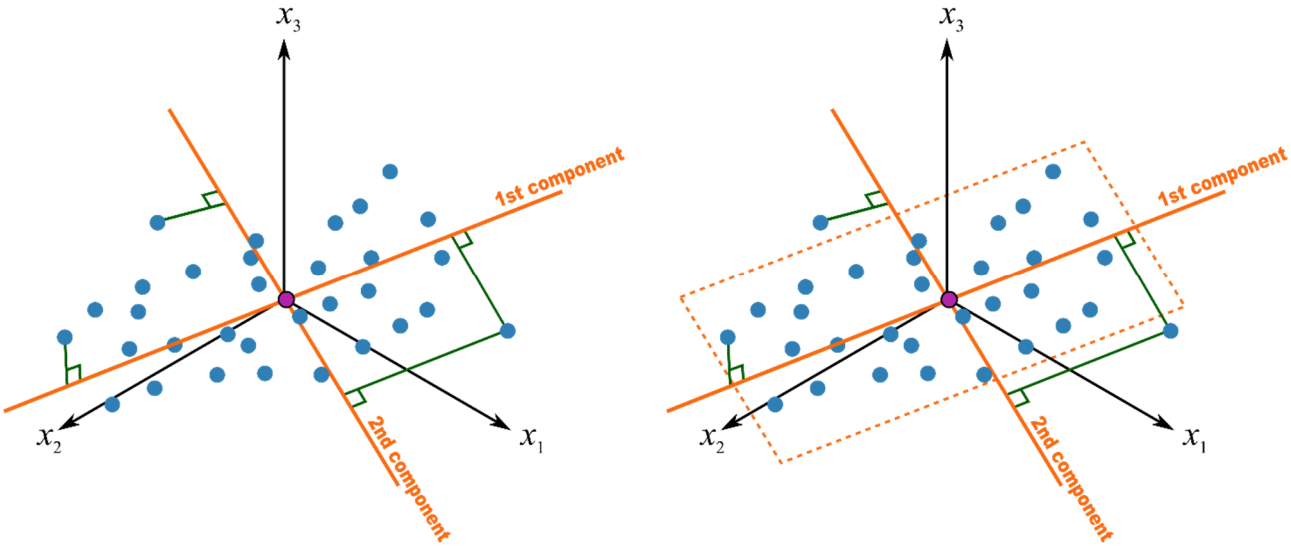


Figure 11: Geometric interpretation of PCA as the search for the best representation subspace of the considered sample[111]

III. A. 1. B - THE VECTORIAL SPACE OF THE PCA

If we consider a p number of real variables X^j (j=1, …, p) observed in n individuals affected by their respective statistical weight w_i such as:

$$\begin{aligned} \forall_i = 1, \dots, n : w_i &> 0 \; et \sum_{i=1}^{n} w_i = 1 \\ \forall_i = 1, \dots, n : x_i^j &= X^j(i), measure\ of\ X^j\ on\ the\ individual\ i \end{aligned} \qquad 63$$

These measures can be grouped on a $(n \times p)$ order matrix:

$$\mathcal{M} = \begin{pmatrix} & X^1 & \cdots & X^j & \cdots & X^P \\ 1 & x_1^1 & \cdots & x_1^j & \cdots & x_1^P \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ i & x_i^1 & \cdots & x_i^j & \cdots & x_i^P \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ n & x_n^1 & \cdots & x_n^j & \cdots & x_n^P \end{pmatrix} \qquad 64$$

For each individual i, a vector x_i is associated containing the i^{th} line of X. It is an element of the vectorial space E with p dimension. With $\mathbb{R}^p$, the canonical basis ε and a matrix M giving it a Euclidean space structure we have E isomorph at $(\mathbb{R}^p, \varepsilon, \mathcal{M})$. E is then describing the individual space. While foreach variable X^j a vector x^j is associated containing the j^{th} centred column (where the mean of the column is subtracted for all the columns). It is an element of the vectorial space noted F of dimension n. With $\mathbb{R}^n$ the canonical basis $\mathcal{F}$ and a matrix D that is the diagonalize if the weight w_i with $D = diag(w_i, \ldots, w_n)$. $\mathcal{F}$ is isomorph at $(\mathbb{R}^n, \mathcal{F}, D)$ and represent the space of the variable.

In general, a model is written as: observation = signal + noise. In the PCA, the matrix of the data comes from the observation of n independent vector $\{x_i, \ldots, x_n\}$, with the same covariance matrix $\sigma^2\Gamma$ but with different expected values z_i, contained in a sub-ensemble of dimension q (with $q \prec p$) of E. Finally, the PCA corresponds to the approximation of a matrix $(n;p)$ by a matrix of the same dimensions but of rank $q \prec p$.

III. A. 1. C - OBJECTIVE AND CONCLUSION

The objectives of a PCA are multiples:

1. The optimal graphical representation of the individuals minimizing the deformations of the scatter graph in a subspace E_q of dimension q $(q \prec p)$.
2. The optimal representation of the variables in the subspace F_q making explicit the initial links between these variables.
3. The dimensional reduction of X by a table of rank q $(q \prec p)$.

As an unsupervised method, the PCA cannot be used for prediction, but it is a fundamental tool for the clustering analysis and as well to prepare the dataset in order to be used for proper ML analysis. In order to analyse the PCA, several graphics can be made with the methods and will be presented in the next part.

III. A. 1. D - GRAPHICAL OUTCOME

Several graphics can be generated for the analysis of the PCA: the projection of the individuals, the projection of the variables, and the dimensional analysis (*screeplot*) will be presented here.

III. A. 1. D. (I) - INDIVIDUALS' ANALYSIS

Each individual *i* represented by x_i are approached by their projection $\mathcal{M}$-orthogonal $\widehat{z_i^q}$ in the subspace $\widehat{E_q}$ created by the first q principal vector $\{v^1, \ldots, v^q\}$. With e_i a vector of the canonical basis of E, the coordinate of the individual *i* in v^k is given by:

$$\langle x_i - \bar{x}, v^k \rangle_M = (x_i - \bar{x})' M v^k = e'_i \bar{X} M v^k = c_i^k \qquad 65$$

In that type of graphic, the quality of the representation is measured by the explained dispersion such as:

$$r_q = \frac{trSM\widehat{P_q}}{trSM} = \frac{\sum_{k=1}^{q} \lambda_k}{\sum_{k=1}^{p} \lambda_k} \qquad 66$$

The dispersion of a one-dimensional scatterplot with respect to its mean is measured by the variance. In the multidimensional case, the dispersion of the scatterplot N with respect to its barycentre x is measured by the inertia, generalization of the variance:

$$I_g(N) = \sum_{i=1}^{n} w_i \|x_i - \bar{x}\|_M^2 = \|\bar{X}\|_{M,D}^2 = tr(\bar{X}' D \bar{X} M) = tr(SM) \qquad 67$$

Then the quality of the representation of each x_i is given by the cosinus square of the angle it forms with its projection:

$$[\cos(x_i - \bar{x}, \widehat{z_i^q})]^2 = \frac{\left\|\widehat{P_q}(x_i - \bar{x})\right\|_M^2}{\|x_i - \bar{x}\|_M^2} = \frac{\sum_{k=1}^{q} (c_i^k)^2}{\sum_{k=1}^{p} (c_i^k)^2} \qquad 68$$

This concept is illustrated in the following Figure 12:

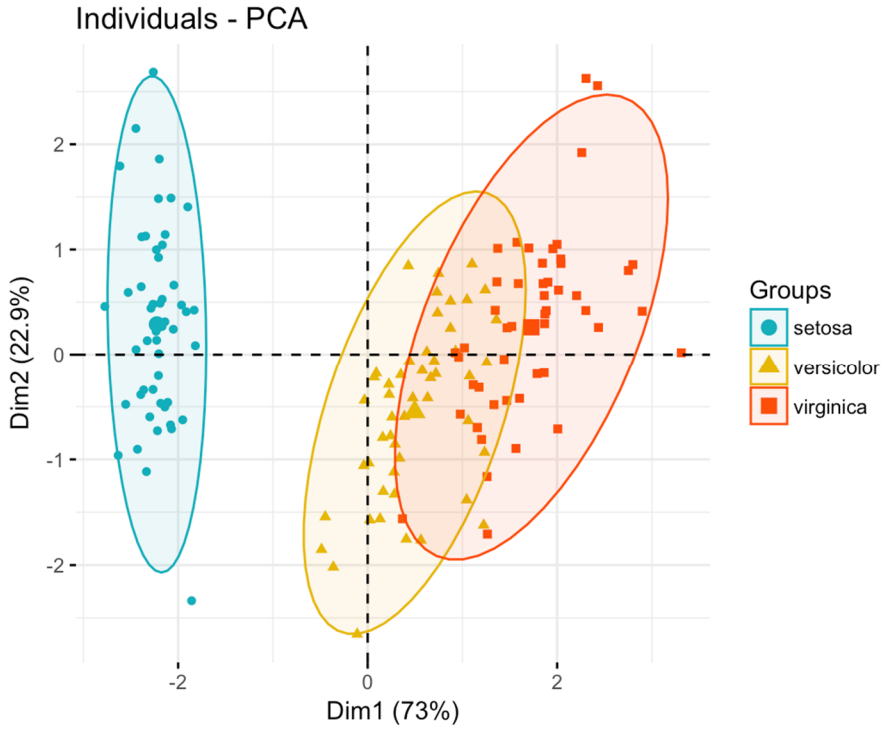


Figure 12: Individuals analysis using the iris dataset provided by R and the *factoextra* package[112]

III. A. 1. D. (II) - VARIABLES ANALYSIS

A variable X^j is represented by the D-orthogonal projection $\widehat{Q_q}x^j$ in the subspace F_q created by the q first factorial axes. The coordinate of x^j on u^k is:

$$\langle x^j, u^k \rangle = x^{j\prime} D u^k = \frac{1}{\sqrt{\lambda_k}} x^{j\prime} D \hat{X} M v^k = \frac{1}{\sqrt{\lambda_k}} e^{j\prime} \hat{X}' D \hat{X} M v^k = \sqrt{\lambda_k v_j^k} \quad (69)$$

And the quality of the representation of each x^j is given in a same relative way as presented before, by the cosinus square of the angle it forms with its projection:

$$\left[\cos\theta\left(x^j, \widehat{Q_q}x^j\right)\right]^2 = \frac{\left\|\widehat{Q_q}x^j\right\|_D^2}{\|x^j\|_D^2} = \frac{\sum_{k=1}^{q} \lambda_k \left(v_j^k\right)^2}{\sum_{k=1}^{p} \lambda_k \left(v_j^k\right)^2} \quad (70)$$

The variable analysis is illustrated in the following Figure 13:

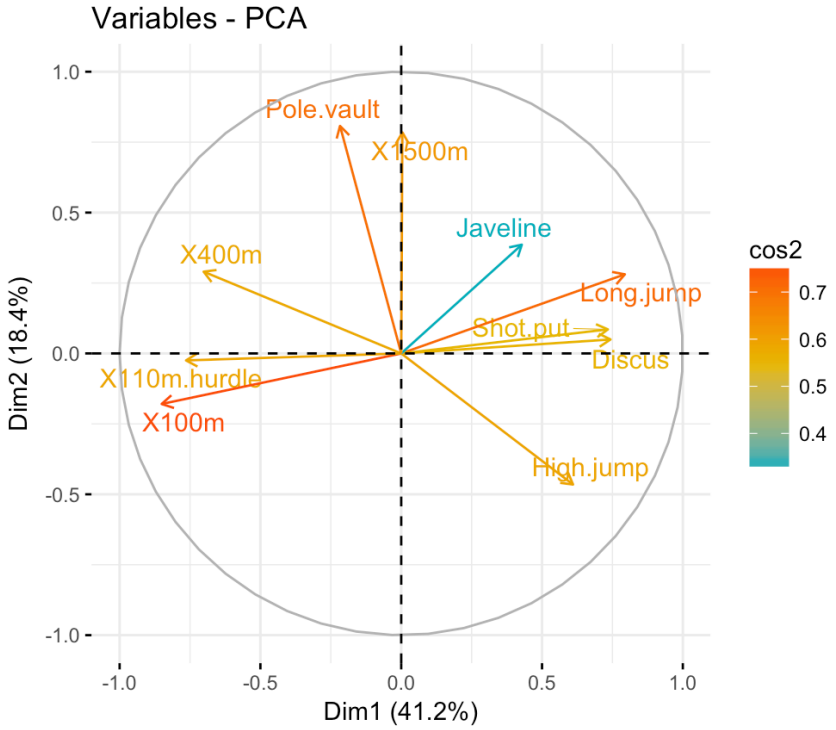


Figure 13: Variables analysis using the decathlon dataset provided by R and the factoextra package[112]

III. A. 1. D. (III) - SCREEPLOT

The screeplot (Figure 14) represents the decrease of the eigenvalues. In principle, we can search which components are important for the prediction using this graphic. Intuitively, the larger the difference $(\lambda_q - \lambda_{q+1})$ is significantly large, and the more we can be sure of the stability of $\widehat{E_q}$.

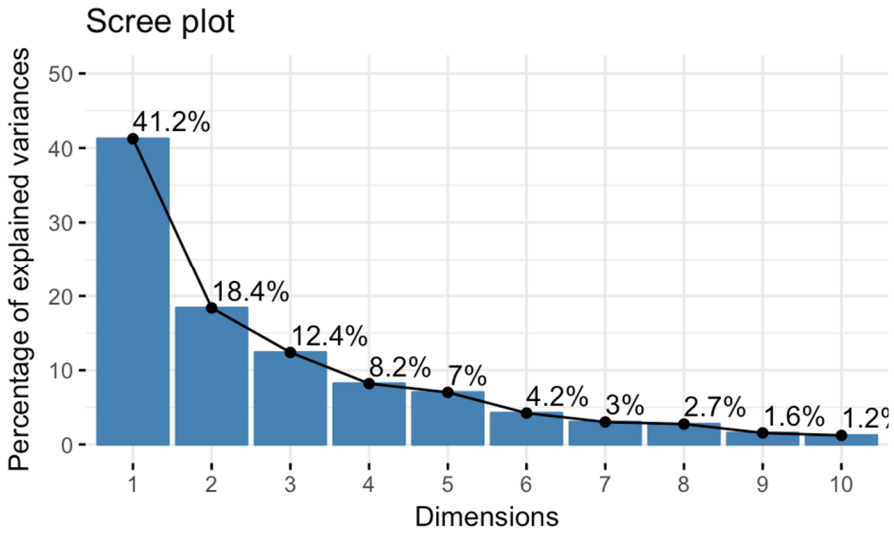


Figure 14: Screeplot using the decathlon dataset provided by R and the factoextra package[112]

As an example, in Figure 14, we might want to stop at the fifth principal component. 87% of the variances contained in the data are retained by the first five principal components. The minimal required values to stop the analysis is different by the dataset considered, but in general, you can consider you explained enough variability if you are > 80%.

III. B - SUPERVISED METHODS

III. B. 1 - INTRODUCTION

QSAR (Quantitative Structure-Activity Relationship) is a method to predict molecular properties using mathematical models.[113] Practically speaking, a QSAR model is a statistical model used to predict a function f using a panel of molecular descriptors[114,115] (x) and a labelled biological activity (Y) (such as binding free energy) such as $Y = f(x)$. The objective of the model is to capture the relationship between molecular descriptors and biological activity. The elaboration of a QSAR model requires several components: (i) a dataset with experimental measurements of the biological activity (input data), (ii) a dataset of molecular descriptors that describing the studied molecules, and (iii) ML models to identify the relationship between the experimental values and the molecular descriptors.

In this thesis, the molecular descriptors are calculated using the CORINA web platform[116]. The data pre-processing and the details about molecular descriptors will be given in the next part of the thesis. Unlike unsupervised ML, a supervised ML algorithm is an algorithm that relies on labelled input data to learn and predict a function producing a decent output when given new unlabelled data. This class of ML algorithm can be used for two different problems linked to the type of the labelled input:

- The classification problem concerns the prediction of a class label represented in the learning dataset by a qualitative variable. In the classification problem, the learning algorithms will produce a function $f: \mathbb{R}^n \rightarrow \{1, \dots, k\}$ with k representing the different categories.
- The regression problem concerns the prediction of a numerical variable. The regression problems represent a very challenging problem in the field of numerical prediction. In the regression problem, the learning algorithm must define a unique function using a vector input to a categorical output. When some inputs are missing rather than providing a single classification function, the learning algorithm must learn a set of

functions. Each of the new functions will be used to classify the descriptors (x) with a different subset of its input missing.

III. B. 2 - K-NEAREST NEIGHBOURS (KNN)

III. B. 2. A - INTRODUCTION

The Knn algorithm belongs to the discriminant analysis class.[117] The idea behind the discriminant analysis is to model a variable Y of m conditions using p quantitative variable $X^j, j = 1, \dots, p$ observed on the same sample Ω of size n. The general principle is to use individuals for whom the X^j are known but not the Y and to decide of the modality of T_l of Y of these individuals. As the variables have to be quantitative, the qualitative variables are replaced by indicators variables.

The final objective is to define some decisions rules: $x = \{x^1, \dots, x^p\}$ designing the observation of explanatory variables on an individual, with $\{g_l, l = 1, \dots, m\}$ the barycentres of the classes calculated in the sample and $\bar{x}$ the global barycentre.

The empirical covariance matrix can be decomposing such as:

$$S = s_e + S_r \qquad 71$$

Where s_e is called the explained variance and S_r is called the residual variance:

$$\begin{aligned} S_r &= \bar{X}_r{}' D \bar{X}_r = \sum_{l=1}^{m} \sum_{i \in \Omega_l} w_i (x_i - g_l)(x_i - g_l)' \\ S_e &= \bar{G}' \overline{DG} = \overline{X_e}' \overline{DX_e} = \sum_{l=1}^{m} \overline{w_l} \, (g_l - \bar{x})(g_l - \bar{x})' \end{aligned} \qquad 72$$

III. B. 2. B - NON-PARAMETRIC ESTIMATION

The estimation is defined as non-parametric when the number of parameters to estimate is infinite. Then the statistical object became a regression function: $y = f(x)$ or a density of probability h. In these cases, we can consider that the density is following a Gaussian distribution whose parameters are estimated. In practice that the density h that is estimated: $\forall x \in \mathbb{R}, h(x)$ is estimated by $\hat{h}(x)$. To use this relationship with relatively good accuracy, you need to use it on large samples: this concept is called the "*curse of dimensionality*".

The Knn algorithm is one of the approaches that estimate the density of $h_l(x)$. This method is described in the following Algorithm 1:

ALGORITHM 1: ***The*** Knn ***algorithm***

for *a sample* Ω of size n, to predict a variable Y of m modalities, **do**

1. Choice of an integer k: $1 \geq k \geq n$
2. Calculation of the distances $d_M(x, x_i), i = 1, \dots, n$ where M is the Mahalanobis metric (the inverse of the variance matrix).
3. Save the k observations $x_{(1)}, \dots, x_{(k)}$ for which the distances are smaller.
4. Count how many times these k observation: $k_{(1)}, \dots, k_{(m)}$ appear in each of the classes.

done

Algorithm 1

The number of clusters you want to define for a classification problem is defined by the user.

III. B. 3 - SUPPORT VECTOR MACHINES (SVM)

III. B. 3. A - INTRODUCTION

The SVM is a class of powerful and flexible modelling techniques.[118] The SVM is a class of learning algorithms initially defined for discrimination, i.e., the prediction of a binary qualitative variable. They were then generalized to the prediction of a quantitative variable. SVM can both be used for classification and regression problems but has only been used for regression problems in the thesis.

III. B. 3. B - SVM FUNCTIONS

SVM for regression use a function similar to the Huber function (i.e., a loss function that uses the squared residuals when the residuals are "small" and use the absolute residuals when the residuals are large), with one important difference. A threshold called ε is set by the user and separates the data points into two classes: those below the threshold that will not contribute to the regression fit and data points that are larger than the threshold that will contribute to a linear-scale amount. Two main consequences can be extracted from this approach, first, the large outliers have a limited effect on the regression equation due to the fact that we do not use the squared residuals. Second, small residuals that the model fits perfectly have no effect on

the regression equation. To estimate the model parameters, SVM uses the predefined **loss function** and also adds a penalty such as:

$$Cost \sum_{i=1}^{n} L_{\varepsilon}(y_i - \hat{y_i}) + \sum_{j=1}^{P} \beta_j^2 \qquad 73$$

Where $L_{\varepsilon}(\cdot)$ represent the ε-intensive function, and the Cost parameters is a user parameter that penalizes the large residuals. Remember that the linear regression model predicts new samples using linear combinations of the dataset constituents. For a given sample μ, the general equation is written:

$$\begin{aligned} \hat{y} &= \beta_0 + \beta_1\mu_1 + \cdots + \beta_P\mu_P \\ &= \beta_0 + \sum_{j=1}^{P} \beta_j\mu_j \end{aligned} \qquad 74$$

For the SVM, the equation is similar, and the estimated parameters can be written as a function of a set of unknown parameters (α_i) and the training set data points so that:

$$\begin{aligned} \hat{y} &= \beta_0 + \beta_1\mu_1 + \cdots + \beta_P\mu_P \\ &= \beta_0 + \sum_{j=1}^{P} \beta_j\mu_j \\ &= \beta_0 + \sum_{j=1}^{P}\sum_{i=1}^{n} \alpha_i x_{ij}\mu_j \\ &= \beta_0 + \sum_{i=1}^{n} \alpha_i \left(\sum_{j=1}^{P} x_{ij}\mu_j \right) \end{aligned} \qquad 75$$

We can highlight from this equation the fact that there are as many α parameters as there are data points (then considered overparameterized from the point of view of the regression model). In general, it is preferable to estimate fewer parameters than data points. Fortunately, the Cost function regularizes the model to help alleviate this problem. Second, the x_{ij} (training set data points) are needed for the new predictions. This can theoretically be a problem when the training set is large. But in fact, many of these data points will have no impact on the regression. Only a few of the data points that are judged to be used will be used for the

regression. Since the regression line is determined using these samples, they are called support vectors because they support the regression line. Finally, in the previous equation, the new samples correspond to a scalar product ($x'u$), then the equation can be rewritten as:

$$f(u) = \beta_0 + \sum_{i=1}^{n} \alpha_i K(x_i, u) \qquad 76$$

Where $K(\cdot)$ is called the Kernel function, the kernel function can have multiple forms, and in this thesis, we used the polynomial kernel that can be used to generalize the regression model and encompass nonlinear functions of the predictors:

$$polynomial_{kernel} = (\phi(x'u) + 1)^{degree} \qquad 77$$

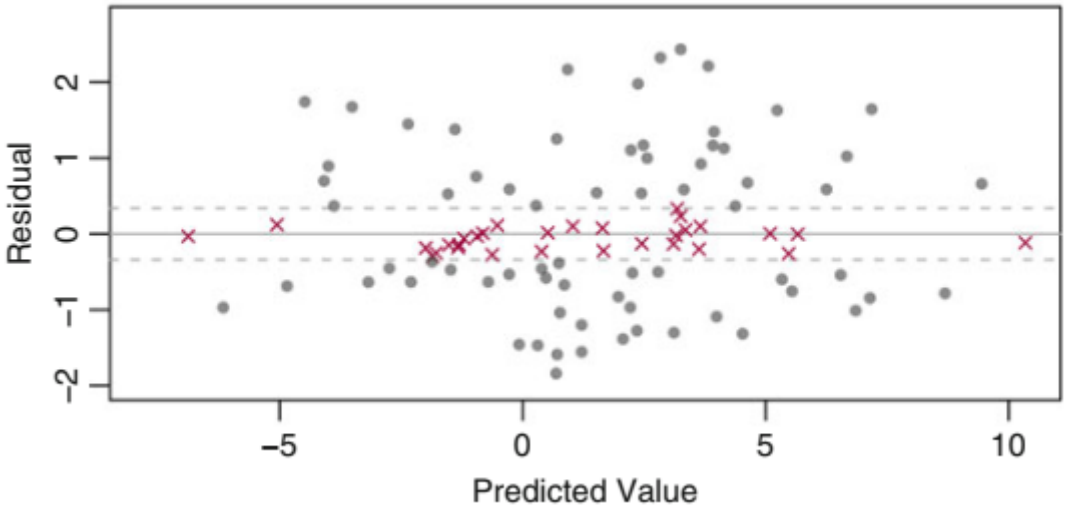


Figure 15: The SVM residuals versus the predicted values[119]

III. B. 4 - RANDOM FOREST (RF)

III. B. 4. A - INTRODUCTION

The RF method is an ML method that is based on model aggregation.[120] Two types of algorithms can be refined, the bagging method[121] (bootstrap aggregation) and then the RF algorithm that is an improvement of the bagging by adding one random component. These algorithms are behind the ML algorithms and the statistics.

The bagging principle can be adapted to every modelling method, but their interest is really highlighted in the case of "unstable" models (a model is defined as unstable if a slight change in the input has an important effect on the hypothesis). For that reason, the usage of these algorithms does not have a lot of sense in the case of multilinear regression and discriminant

analysis. In the case of the tree's algorithms, the known instability of the trees appears as a necessary property for the reduction of the variance by model aggregation. These algorithms are based on the bootstrap principle initially.

III. B. 4. B - PRINCIPLE OF THE BOOTSTRAP

The bootstrap is defined as a way to replace the probabilistic hypothesis with simulations and calculations.[121] The idea is to predict the distribution of an estimator when the real distribution of the sample is not known or, more often, when it cannot be assumed to be Gaussian.

Starting with the Plug-in principle: if we consider $x = \{x_1, \ldots, x_n\}$, an n-sized sample following an unknown distribution F on (Ω, A). The empirical distribution $\hat{F}$ represent the discrete probability of the $\{x_1, \ldots, x_n\}$ affected by the weight 1/n such as:

$$\hat{F} = \sum_{i=1}^{n} \delta x_i \qquad 78$$

Considering θ a parameter that is a function of the $\hat{F}$ distribution. We can write $\theta = t(F)$, and θ correspond then to:

$$\hat{\theta} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} \delta x_i \qquad 79$$

Where $\bar{x}$ is the estimator of θ, It is called a plug-in estimator.

Definition 1: The estimator obtained by replacing the distribution F by the empirical distribution is called the plug-in estimator of a parameter θ of F: $\hat{\theta} = t(\hat{F})$

Bootstrap is used to estimate the standard deviation, such as Considering $\hat{\theta} = s(x)$ any estimator of θ for a given x sample. One seeks to appreciate the precision of $\hat{\theta}$ and thus to estimate its standard deviation.

Definition 2: we called a bootstrap sample of x, an n-sized sample written: $x^* = \{x_1{}^*, \ldots, x_n{}^*\}$ and following the $\hat{F}$ distribution, and with x^* defined as a resampling of x with replacement.

Definition 3: We called bootstrap estimation of the standard deviation $\widehat{\sigma_F}(\hat{\theta})$ of $\hat{\theta}$, his plugin estimation: $\sigma_{\hat{F}}(\hat{\theta})$.

Unfortunately, considering the definition 3, apart from in the case where θ is a mean, there is no simple way to define this estimator explicitly. For that reason, an approximation of the bootstrap estimator is made by a Monte-Carlo like simulation described in Algorithm 2**:**

ALGORITHM 2: ***standard deviation estimation***

Considering x a sample and θ a parameter.

$\boldsymbol{for}\ b = 1\ to\ B, \boldsymbol{do}$

$select\ 1 :\ x^{*b} = \{x_1^{*b}, \dots, x_n^{*b}\}$ by sample with replacement in x.

$Estimate\ in\ this\ sample$: $\widehat{\theta^*}(b) = s(x^{*b})$

$\boldsymbol{end\ for}$

Algorithm 2

$$\hat{\sigma}_B^2 = \frac{1}{B-1}\sum_{b=1}^{B}\left(\widehat{\theta^*}(b) - \widehat{\theta^*}(\cdot)\right)^2$$

$$with:\ \widehat{\theta^*}(\cdot) = \frac{1}{B}\sum_{b=1}^{B}\left(\widehat{\theta^*}(b)\right)$$

With $\hat{\sigma}_B$defined as the bootstrap approximation of the desired plug-in estimate of the standard deviation of $\hat{\theta}$.

As a conclusion, we can say that the bootstrap relies on a very basic assumption: $\widehat{\theta^*}$ behaves with respect to $\hat{\theta}$ as $\hat{\theta}$with respect to θ. The knowledge of $\widehat{\theta^*}$ (distribution, variance, bias, …) then informs about the knowledge of θ.

III. B. 4. C - PRINCIPLE OF THE BAGGING

Considering Y, the quantitative or qualitative variable we want to predict and $\{X^1, \dots, X^p\}$ the molecular descriptors (the variables) and $f(x)$ a model following the x function with $x = \{x^1, \dots, x^p\} \in \mathbb{R}^p$. With n defined as the number of observations and $z = \{(x_1, y_1), \dots, (x_n, y_n)\}$ a sample of the distribution of F.

Considering B independent sample such as: $\{z_b\}_{b=1,B}$, a prediction using a model aggregation method depends on the class of Y:

- If Y is quantitative: $\widehat{f_B}(\cdot) = \frac{1}{B}\sum_{b=1}^{B}\widehat{f_{Z_b}}(\cdot)$ a simple mean of the results for the models associated with each sample.
- If Y is qualitative: $\widehat{f_B}(\cdot) = \arg max_j \, card\{b \mid \widehat{f_{Z_b}}(\cdot) = j\}$, in this second case, an ensemble of models is constituted to predict the most probable statistical response.

Behind these two differences, the principle is simple: by averaging the forecasts of several independent models, it is possible to reduce the variance and thus to reduce the error.

Unfortunately, due to the calculation cost, it is unrealistic to consider B independent sample, and for that reason, the samples are replaced by B replication of bootstrap sample obtained by n draws with replacement following the empirical measurement $\hat{F}$. This concept is described below in Algorithm 3:

ALGORITHM 3: ***Bagging***

Considering $x_0$the variable to predict and:

$z = \{(x_1, y_1), \dots, (x_n, y_n)\}$ a sample

for $b = 1$ *to* B, ***do***

Draws a bootstrap sample z_b^*

Estimate $\widehat{f_{Z_b}}(x_0)$ on the bootstrap sample

end for

Calculate the average estimate $\widehat{f_B}(x_0) = \frac{1}{B}\sum_{b=1}^{B}\widehat{f_{Z_b}}(x_0)$ for the quantitative variable or the results of the probabilistic response for the qualitative model.

Algorithm 3

Naturally, this algorithm gives an easy way to calculate the error of the prediction: the out-of-bag error (o.o.b): for each observation (x_1, y_1) you can consider the estimated model on a bootstrap sample that does not contain this observation. The values of $\hat{y}$ is predicted following the algorithm of the bagging (Algorithm 3), and the calculation of the associated prediction error, averaged over all observations, gives an estimate of the o.o.b.

III. B. 4. D - THE RF ALGORITHM

The principle of the RF is a specific case of the CART (Classification and Regression Trees) algorithm that is a bagging algorithm improved by adding a random variable.[122] In the CART algorithms, the final tree is constructed along with several optimisations that we call "split". Multiple-way to split the trees existing: the most known are (a) the Gini criteria that organize the separation of the leaves of a tree by focusing on the most represented class in the data set

with the idea that the separation has to be fast as possible and (b). the entropic criteria where the construction is based on the reduction of the entropic disorder of the considered sub-dataset at each leaf of the tree. By adding a random variable, the idea is to make the aggregation trees more independent by adding some flexibility in the choice of the variable that constitutes the models.

One known limitation of the bagging algorithms is the case of correlated variables. Let's take the case of the B independent variables identically distributed, each with variance σ^2. The variance of the mean of B is $\frac{\sigma^2}{B}$ and if these variables are identically distributed and correlated two by two with a correlation = ρ, the variance of the mean can be rewritten

$$\rho\sigma^2 + \frac{1-\rho}{B}\sigma \qquad 80$$

In this case, the second term is decreasing with B, but the first one is limiting the interest of the bagging if the correlation is high. This reason motivates the introduction of the randomization to introduce in the RF algorithms in order to decrease ρ between the forecasts provided by each model. The *RF* algorithms are presented below in algorithm 4, where the bagging is applicated with binary trees, by adding a random draw of m explanatory variables among the p:

ALGORITHM 4: ***RF***

Considering $x_0$the variable to predict and:

$z = \{(x_1, y_1), \dots, (x_n, y_n)\}$ a sample

for $b = 1$ *to* B, ***do***

Draws a bootstrap sample z_b^*

Estimate a tree on this sample with randomization of the variables: the search of each optimal spl preceded by a random draw of a subset of m predictors.

end for

Calculate the average estimate $\widehat{f_B}(x_0) = \frac{1}{B}\sum_{b=1}^{B}\widehat{f_{Z_b}}(x_0)$ for the quantitative variable or the results of the probabilistic response for the qualitative model.

Algorithm 4

Considering that, there are different parameters of the RF that you can modify in order to improve the prediction: (1) the ***mtry*** input setting that represent the number of variables randomly sampled as candidates at each split, (2) the ***ntree*** input setting that represents the

number of trees to grow. Including the general cross-validation parameters, the way you reduce the initial variables of the dataset, and the way you separate your sample into training-set / test-set, there are billions of possibilities.

III. B. 5 - NEURAL NETWORK (NNET)

III. B. 5. A - INTRODUCTION

The NNET algorithm is an algorithm that belongs to the deep-learning class.[123]

DL is a set of learning methods that will use non-linear transformation to modelized a dataset with complex architectures. The simplest model of DL algorithm is the NNET that are combined to form the deep NNET.[124,125] It exists multiple architectures of an NNET, the multilayer perceptrons are the simplest one and the one we used in this thesis.

Mathematically speaking, an artificial NNET is a nonlinear application presenting a parameter θ that associate to an entry x, an output y such as $y = f(x;\theta)$. The parameter θ is estimated from a learning dataset. The virtual NNET can be used for both classification and regression problems. In 1989, Cybenko and Hornik[126] defined the universal approximation theorem that made the triumph of the method.

III. B. 5. B - THE ARTIFICIAL NEURONS

An artificial neuron is a function f_j of the input $x = (x_1, \dots, x_n)$, weighted by a vector of connection weights $w_j = (w_{j,1}, \dots, w_{j,n})$, completed by a neuro bias net_j and associated to an activation function called φ such as:

$$y_j = f_j(x) = \varphi\left(\langle w_j, x\rangle + net_j\right) \qquad 81$$

It exists multiple activation function φ, the most used one is the sigmoid function:

$$\varphi(x) = \frac{1}{1 + \exp(-x)} \qquad 82$$

The structure of the artificial neuron is represented in Figure 16 below:

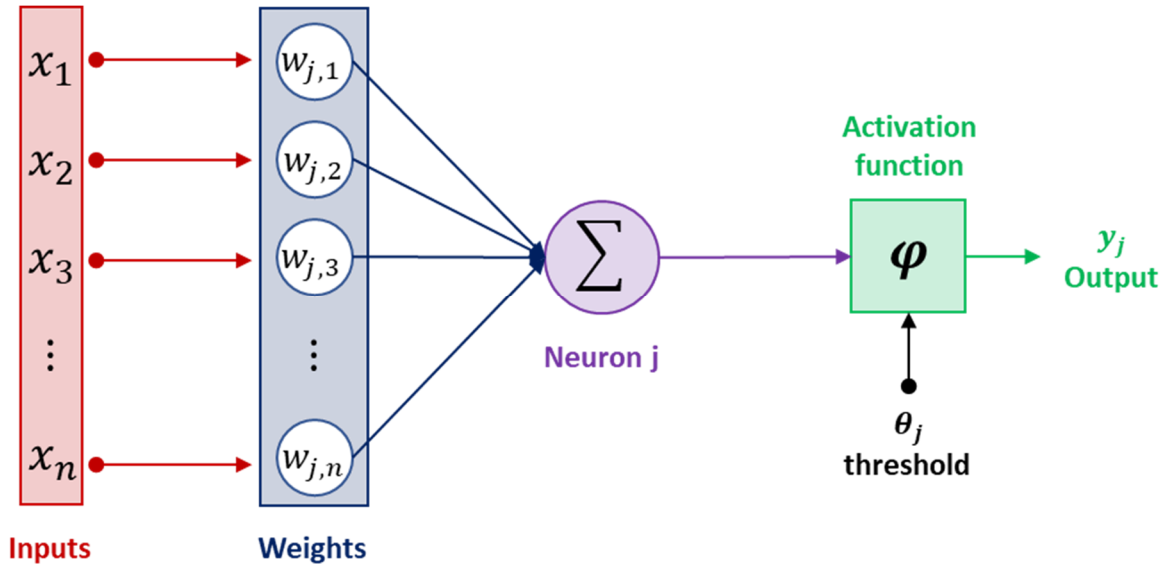


Figure 16: Schematic representation of an artificial neuron where $\sum = \langle w_j, x \rangle + net_j$

III. B. 5. C - THE PERCEPTRON LAYER

In the simple case, when implementing an NNET to a dataset, the nodes of the input layer consist of the n molecular descriptors that describe the sample (extracted from previous steps). The output layer consists of one or several nodes, depending on the classes of the prediction (classification or regression). Linking those two nodes will be a number of hidden layers composed of a number of hidden units. In the simple case, we only have one hidden layer. The prediction process will begin by assigning random weights to the connections between nodes, which are then iteratively updated as predictions are verified against the experimental data provided in the initial dataset, and the error is back-propagated. In an attempt to limit the overtraining of the dataset, another free parameter: the weight decay, is used to apply penalties with the aim of limiting such overfitting and controlling the quality of the prediction.

In a more complex case, the multilayer perceptron (or NNET) is a structure composed of several hidden layers of neurons for which the output of a neuron of a layer becomes the input of a neuron in the next layer. Figure 17 below represents an NNET with three input variables in the input layer, one output variable expected in the output layer, and two hidden layers composed respectively by 4 and 5 hidden units.

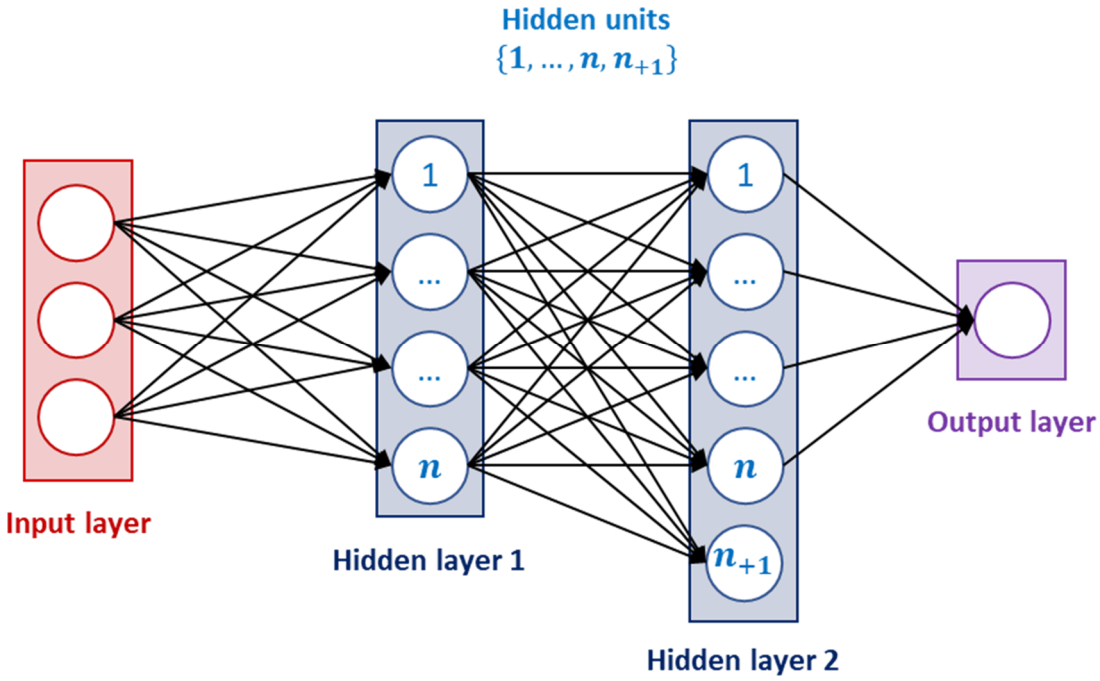


Figure 17: Representation of an NNET

III. B. 5. D - UNIVERSAL APPROXIMATION THEOREM

In 1991, Hornik[127] stated the fact that any bounded and regular function $\mathbb{R}^n \to \mathbb{R}$ can be approximated at any given precision by a NNET with one hidden layer containing a finite number of neurons. More precisely, the Hornik theorem can be stated as follows:

Theorem 1: Let φ be a bounded, continuous and non-decreasing function. Let $K_n$be some compact set in $\mathbb{R}^n$ and $\mathcal{C}(K_n)$ the set of a continuous function on K_n. Let $f \in \mathcal{C}(K_n)$. Then for all $\varepsilon > 0$, there exists $N \in \mathbb{N}$, real numbers $v_i$and net_j and $\mathbb{R}^n$-vector w_i such that, if we define:

$$F(x) = \sum_{i=1}^{N} v_i \, \varphi\left(\langle w_i, x\rangle + net_j\right) \qquad 83$$

Then we have:

$$\forall x \in K_n, |F(x) - f(x)| \leq \varepsilon \qquad 84$$

From a theoretical point of view, this theorem is interesting, and even it is not really useful because the number of neurons in the hidden node may be very large, it is a stated fact that the strength of DL lies in the deep of the network.

IV - BINDING FREE ENERGY DETERMINATION

IV. A - PRINCIPLE OF THE THERMODYNAMIC BASED METHOD

IV. A. 1 - GIBBS FREE ENERGY

Considering the general form of the association between a ligand and its receptor, respecting a 1:1 stoichiometry, we can define the association process as follows:

$$R + L \rightleftharpoons RL \quad (85)$$

The equilibration reaction of this process depends on the thermodynamics properties of the Ligand (guest), the Receptor (host) and the complex (RL). The Gibbs free energy can be defined as follows:

$$\Delta G^0 = -RT \ln(K_{dissociation} C^0) = \Delta H^0 - T\Delta S^0 \quad (86)$$

With T corresponding to the experimental temperature, and R is the gas constant. $K_{dissociation}$ represents the equilibrium dissociation constants, and can be defined as the ratio between the concentration of the product, and the concentration of the reactants at equilibrium condition at 1M concentration (C^0). ΔH^0 represent the enthalpic term, while ΔS^0 represent the entropic term.

A negative value for ΔG^0 indicates that the binding reaction is favourable under standard conditions, and the process is exergonic and spontaneous toward the complex formation. A positive value for ΔG^0 is described as the endergonic change, and the binding process is not spontaneous. In general, if binding occurs, the expected binding free energy prediction is supposed to be negative.

IV. A. 2 - PROTOCOL USED DURING THE THESIS

The Gibbs free energies of the optimised geometries were calculated as the sum of the total energy, which includes the D4 dispersion correction, thermostatistical corrections calculated following a coupled rigid-rotor-harmonic-oscillator approach (GRRHOT), and the solvation contribution (Gsolv) calculated by the implicit solvation model GBSA.

$$\Delta G = E + G_{RRHOT} + G_{solv} \quad (87)$$

With:

$$\Delta G_{solv} = \Delta G_{born} + \Delta G_{sasa} + \Delta G_{hb} + \Delta G_{shift} \quad 88$$

The association Gibbs free energy is calculated from the difference of the free energies from the complex, host, and guest molecules, each on their respective conformational minimum.

$$\Delta G_{bind} = \Delta G_{complex} - \Delta G_{host} - \Delta G_{guest} \quad 89$$

Considering the complexity of the conformational energy landscape of the complex and host molecule, we used multiple geometries of the unbound host system as starting points for minimization, thus increasing the probability of finding the absolute minimum. For that, multiple structures are extracted from the classical MD simulations to and carry out a geometric optimisation at an SQM level, followed up by calculation of the hessian to confirm that the final energy is a true minimum (i.e., all vibrational frequencies are positive). Though the degrees of freedom of the guest is much reduced, we use a similar protocol for consistency.

V - DOCKING

V. A - PRINCIPLE

The docking is generally composed of two components: phases: (i) a search algorithm where the ligand is positioned in the defined cavity, then (ii) an evaluation of the energetical interaction called scoring. In the thesis, molecular docking is used to generate a first guess of the Host-Guest structure we chose AUTODOCKVINA[128] for our assessment for several reasons. It (i) is faster and generally performs better than AUTODOCK itself, (ii) is freely available and competitive with commercial tools.

V. B - PROGRAMS

V. B. 1 - AUTODOCK4

AUTODOCK4 is a freely available program that has been developed by Arthur J. Olson from the "*Scripps Research Institute*". AUTODOCK has its own scoring function and uses a genetic algorithm for the search. A genetic algorithm is an algorithm inspired by the evolution process to improve a population of solutions iteratively. In the case of the docking, each solution corresponds to a possible conformation between the ligand (the guest) and the receptor (the host). The first population is randomly chosen. Then coming from that first population, all the solutions are considered like a "*chromosome*" and can randomly undergo mutation and genetic events (i.e., mutation type or "cross-over"). The mutation event corresponds to a modification of the structure of the given chromosome, whereas the cross-over event corresponds to the exchange of equivalent genetic material between two different chromosomes. The new conformations are then evaluated using the scoring function and replace the initial confirmation if they are better. The mutations are randomly chosen, but as the scoring function is used to select the best poses, the notion of selective pressure is introduced. In conclusion, the genetic algorithm is a stochastic algorithm that gives the opportunity to study several different conformations of the guest molecules. The host conformation, on the other hand, is rigidly fixed on the initially provided geometry. This concept is illustrated in the following Figure 18:

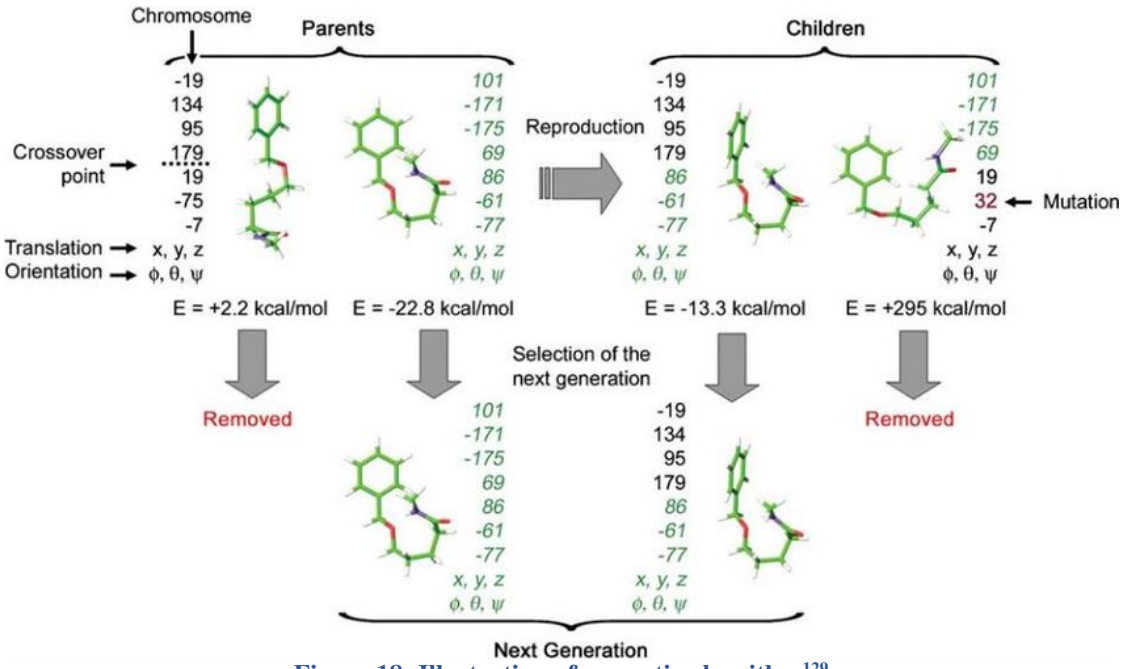


Figure 18: Illustration of a genetic algorithm[129]

Though AUTODOCK is initially is based on a genetic algorithm, three different search methods can be used, (i) the pre-explained classic genetic algorithm where the "chromosome" is composed by the three cartesian coordinates representing the translation of the guest followed by four variables defining the rotation of the guest in the space and finally values that correspond to the possible torsion angles of the guest ; (ii) a local search based on simulated annealing method of type Monte Carlo can be added to minimize the energy ; (iii) a mixed method using both previous approaches in addition to a Genetic Lamarckian algorithm. The Lamarckian algorithm principle states the fact that the phenotypes (the individual characteristics) can modify the genotypes. In the case of the Docking, the genotype represents the ensemble of the genetic operations presented before: mutation, and cross-over, while the phenotype represents the score of the guest in the conformation of the genotype. For the Lamarckian algorithm, the local research based on the simulated annealing method can modify the coordinates of the guest (phenotype) that will be transferred to the corresponding chromosome (genotype) and thus to the descendants.

AUTODOCK4 uses a free energy evaluation score function based on a force field using SQM parameters. This force field has been calibrated using a large heterogeneous group of experimental data of protein-ligand complexes.

V. B. 2 - AUTODOCK-VINA

AUTODOCK-VINA is a docking program entirely derivate from AUTODOCK4 with an improvement of the performance by using multithreading and multi-core calculations. At the

same time, the vina algorithm is close to the AUTODOCK4 ones. The Vina scoring function uses a hybrid function (empirical + knowledge-based) based on the X-Score function that has been calibrated on the “PDBbind” database (a database that gathers the experimental affinity values for which the structures of the complexes are known in the PDB but unfortunately not well documented).

3

THE HG-DYNAUSOR PLATFORM

I - INTRODUCTION

I. A - CAPABILITIES

The HG-DYNAusor platform is an automated platform to facilitate the execution of common tasks in the modelling of host-guest complexes, including several methods to investigate the geometry, dynamics and energy of supramolecular complexes, as well as their individual components. An overview of the HG-DYNAusor platform is presented in the following Figure 19:

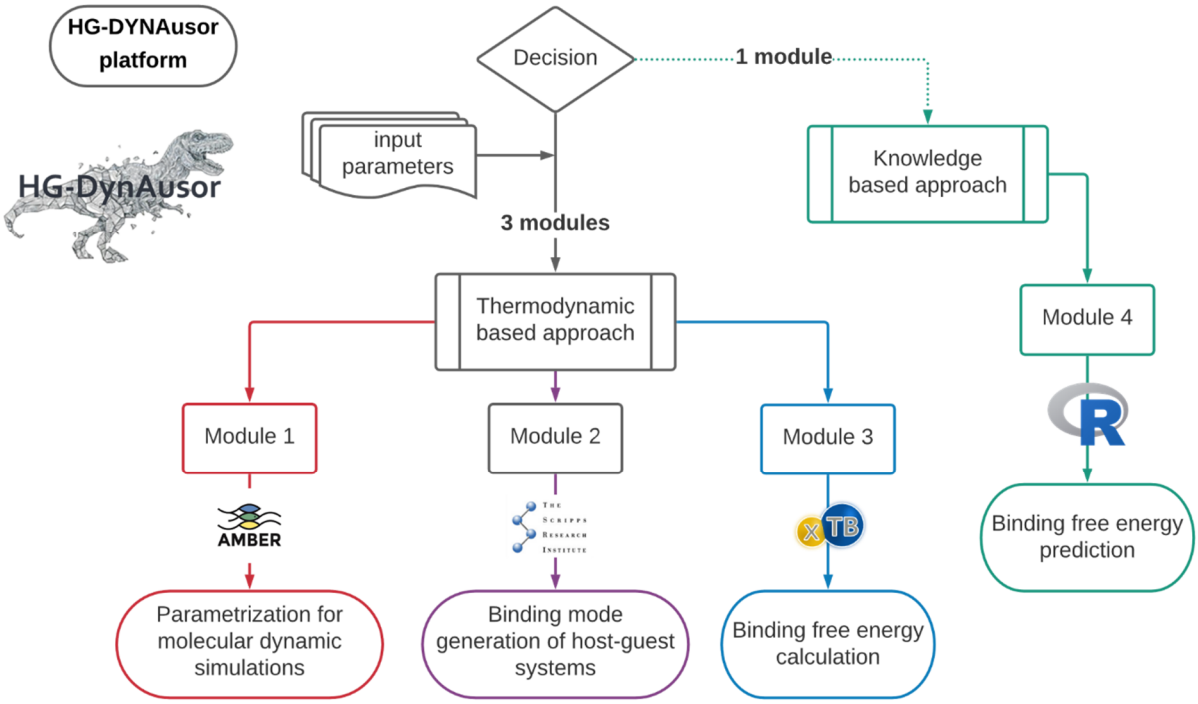


Figure 19: An overview of the modules of the HG-DYNAusor platform; the three first modules are part of the thermodynamic based approach: (in red) module 1 dedicated to the parametrisation, (in purple) the module 2 dedicated to the binding mode generation for host-guest complexes, (in blue) the module 3 dedicated to the binding free energy calculation and in green the knowledge-based approach with the module 4 dedicated to the binding free energy prediction.

The platform, designed and developed during the thesis, is operational but remains under development. It has been used in all the studies of molecular containers presented in the next chapters. The platform can be separated into four different modules: three dedicated to binding free energy calculation using a thermodynamic based approach. And one module dedicated to the binding free energy prediction using a knowledge-based approach.

II - PROOF OF CONCEPT USING ACRIDINIUM TWEEZER

II. A - INTRODUCTION

As a proof-of-concept, we used a bis-acridinium molecular tweezer. In supramolecular chemistry, tweezer is a term defined for the first time in the late 70s by Whitlock and Chen[130] to indicate molecules capable of producing a 1:1 sandwich complex with a planar aromatic guest or dimerizing spontaneously in water. It forms a supramolecular object that looks like a pincer holding an object. The folding of this pincer and the formation of the binary complex with the guest are mainly intermolecular forces (π-π interactions and hydrophobic interactions). Practically speaking, a molecular tweezer consists of two binding units separated by a flexible conformational spacer. The conformational spacer is fixing the binding unit with a limited distance, thus holding the binding units in a specific conformational space. When the two binding units converge, they form a cavity, opened at three sides: a molecular cleft, allowing a guest to bind. This molecular tweezer is a part of the final porphyrin-receptor[131]. This molecular tweezer is known to realise self-assembly and narcissistic self-sorting in water: meaning it spontaneously organize in 1:1 dimer in water. It will be considered as a simplified model of receptor based on a similar scaffold[132].

The goal of the proof of concept can be separated into three parts:

- Find a methodology for the **determination of the binding mode** of host-guest complexes
- Find a way to **analyze the behaviour of host-guest** systems using MD simulations.
- Validate a way to **calculate thermodynamics properties** in a relatively short computational time, allowing us to calculate the binding free energy for many different complexes.

II. B - GENERATION OF PARAMETERS

There are two ways to generate the parameters for the MD simulations: extracting the structure from a reference crystal or constructing the molecule from the 2D information. In our case, both options were tested.

When there is no reference crystal, a first guess of the molecule's geometry is constructed from 2D information. In our case, the SMILES code of the molecule is extracted, and a first estimation of the 3D structure is made using *open-babel*, following the *gen3d* protocol. Because the geometry is built from 2D information, some conformational problems may remain in the

structure. For that reason, DFT was used for better structural optimisation (using B3LYP, 6-31G**). The atoms of the molecular tweezer are typed with *gaff* force field using the *antechamber* module from the *AMBERTOOLS* package. The missing parameters are generated using the *parmchk* module on the optimised geometry. The partial charge of the molecule is calculated using the AM1-bcc charge, implemented in *antechamber* using the *AMBERTOOLS* package.

II. C - ASSOCIATION IN WATER:

Two different methods are used to predict the binding mode: (i) the Spontaneous association Molecular Dynamics (SaMD) protocol, (ii) docking followed by MD. As there is an existing crystal of the dimer, the obtained structure using the MD simulations was compared to the crystal one in terms of geometry and energy. For the generation of the binding mode, unbiased MD were used in a protocol we called SaMD. In what we refer to at SaMD, the host and the guest are simulated in a TIP3P water box, starting from a dissociated configuration (distance ≥ 8 Å), extending the simulation until the binding is observed. We believe that this method could be a good way to identify the bound state. Identifying several chemical states behind the binding mode allows us to understand the binding mechanism better and sample several geometries that can be used for binding free energy prediction.

The other way to generate the binding mode consists of using AUTODOCK-Vina (ADV) to generate the first complex. For that specific case, it consists of using the molecular tweezer as a guest and host. As it is supposed to dimerize, the guest-tweezer will be docked in the host-tweezer. The main difference consists in the flexibility of the tweezer. In the case of the host-tweezer, the used geometry is the one optimised at the DFT level, and as It is the "receptor" for the docking, the structure is considered rigid. In contrast, for the guest-tweezer, the docking protocol considers the "ligand" fully flexible on his rotatable bonds.

An overview of the docking and SaMD results is presented in Figure 20: The crystal structure (Figure 20A) present a closed-closed conformation. Docking can be done on the closed-tweezer, the only difference between the host and the guest for the docking protocol will be the free rotation of the bonds highlighted in red for the guest (Figure 20B). The two best docking poses in the closed-tweezer shows two different conformations (Figure 20C): the best pose is close to the crystal one and present the best docking-score (-14.9 kcal/mol), the second pose present as well a closed-closed conformation with different conformation and a lower docking score (-12.8 kcal/mol). The DFT geometric optimisation of the tweezer led to the formation of

an open-tweezer that could be used both for SaMD and docking leading to the formation of the closed-open complex in both cases (Figure 20E).

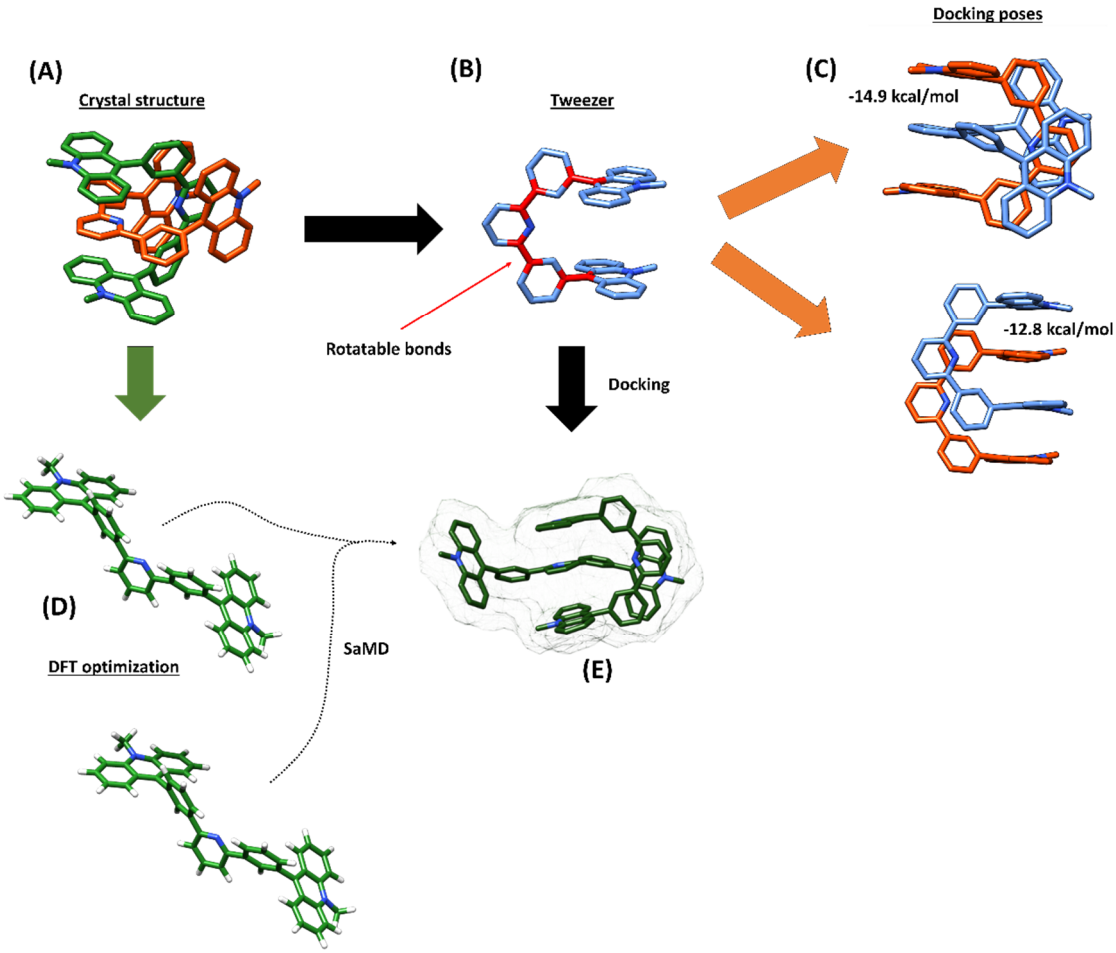


Figure 20: An overview of the docking results using AUTODOCKVina, starting from the crystal structure (A), for the docking of the crystal, one of the tweezer will be the host and considered as rigid while the other one considered as the guest will be allowed to rotate on some bonds (B), two different docking poses representing two different complexes can be extracted from the docking results (C), and one tweezer can be extracted, and his geometry optimised at DFT level (D), both using SaMD protocol and docking the DFT-optimised structure in the previous host lead to the formation of the similar complex (E).

In the case of the open-configuration of the tweezer (Figure 20D), the dimerization process is favoured, but the structured dimer does not present the closed-closed crystal conformation. Suppose the switch from any conformation to the closed-closed conformation represents a break of an energetical well. It can be related to a rare event in the dynamic and may need an extension of the simulation times to see this conformation.

II. D - BEHAVIOUR ANALYSIS OF HOST-GUEST SYSTEM IN SOLUTION

To study the self-assembly and narcissistic self-sorting processes, we investigated the dynamics of the molecular tweezer, first as a monomer and then as a dimer. Our main goal was to prove with a simple system that it was possible to study the assembly of the molecular tweezer computationally. Several MD simulations of 500ns each was carried out with GROMACS and AMBER: for the monomer alone in the water and two monomers in a cubic water box where they are expected to assemble. From the output of MD simulations, we made a protocol based on the generation of several molecular descriptors to describe the variability of the geometry over time (Root Mean Square Deviation (RMSD), Radius of Gyration (Rg), Surfaces Accessible Solvent Area (SASA), N-N distances…). These descriptors are then visualized with a PCA and used to define relevant clusters inside the MD simulations (Figure 21 and Figure 22). These clusters represent the chemical space explored along with MD simulations with the idea that different clusters represent different types of geometry reached along with the simulations. In this graphic, all points correspond to a specific frame of the MD simulations: every point corresponds to a specific geometry at a specific time. The clustering of the monomer is presented in the following Figure 21: three different geometries could be extracted based on the orientation of the acridinium groups: the closed-form (Figure 21A), the semi-open or semi-closed form (Figure 21B), and the open-form (Figure 21C).

Interestingly, these three geometries are represented throughout the simulations, suggesting that we sufficiently sampled the conformational space for the monomer. The conformational space could be defined as the space encompassing all possible positions of the molecule. We cannot say that we sample the whole conformational space. These graphics allow us to say that we sampled the conformational space sufficiently because we reach similar positions in different MD.

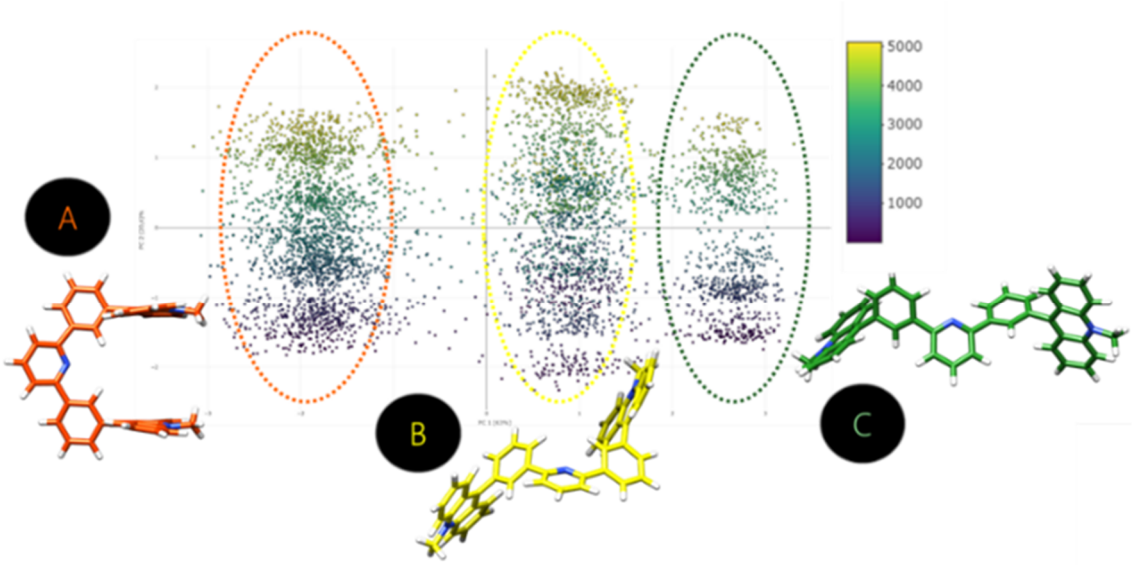


Figure 21: Clustering of the MD of the tweezer-monomer, (A) Clustering of the closed-tweezer in orange, (B) Clustering of the semi-open or semi-closed form in yellow, and (C) clustering of the open-tweezer in green, each point represent an individual geometry, the points are colored by the frames of the simulations

For the second simulation, we built two monomers in a cubic box to study their assembly. The assembly process was realised very fast (few ~ ns), with disassembly not seen, suggesting the dimer once formed is very stable. As previously done for the monomer, the dynamics are clusterised using PCA (Figure 22). What is interesting to note is that only certain combinations of geometries were extracted from the monomer analysis. The closed-form in association with the open form (Figure 22 A-C), the semi-open or semi-closed form dimer in association with itself (Figure 22 B-B), and the open form in association with the semi-open / closed form (Figure 22 B-C).

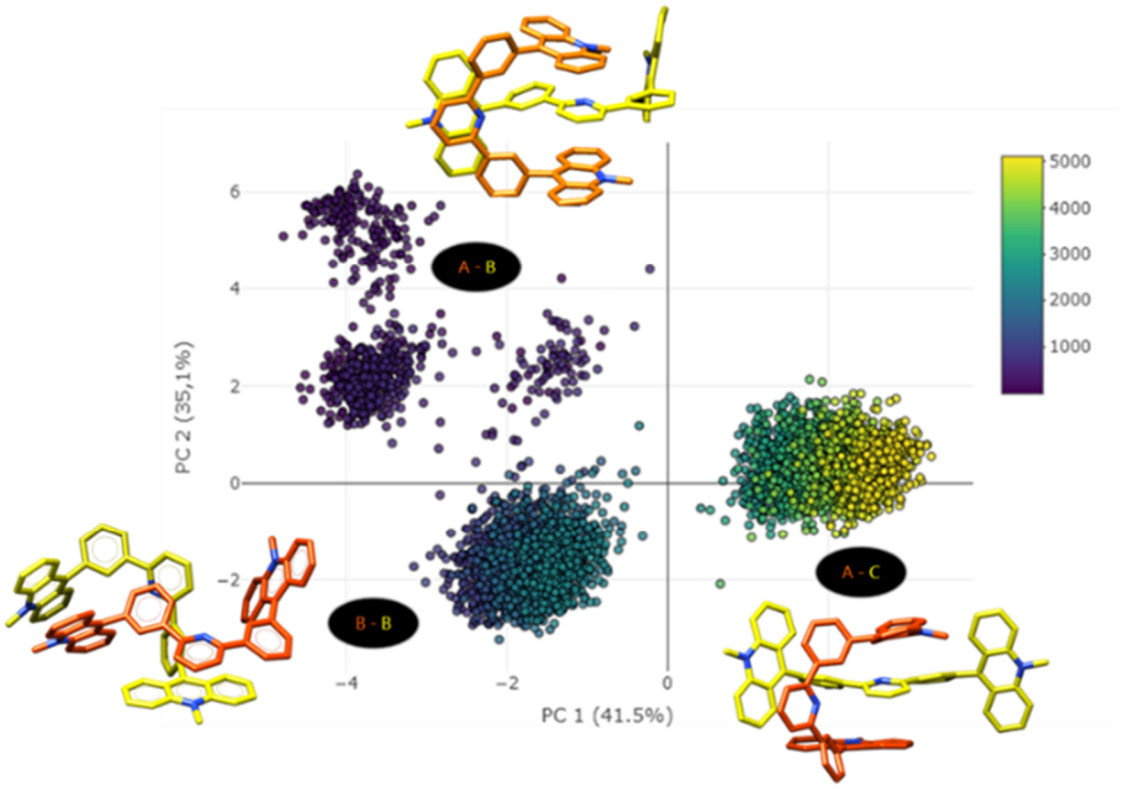


Figure 22: Clustering of the MD of the tweezer-dimer, (A-B) Clustering of the closed-semi-open tweezer, (B-B) Clustering of the semi-open-semi-closed dimer, and (A-C) clustering of the open-closed dimer, each point represents an individual geometry, the points are colored by the frames of the simulations

We know from the crystallographic data that the crystallographic form should be a closed-closed dimer. An MD simulation starting from the crystal structure is also performed, and the structure remains stable during the whole simulation. Suggesting that the MD was not long enough to carry out the closed–closed conformation. Alternatively, a closed-closed conformation is a rare event that heavily depends on how our two monomers initially assemble themselves. A new dissociated MD was run after that and presented similar results with the additional formation of a closed-closed dimer with the same orientation of the previously presented second docking poses (Figure 20C).

Some structures of the closed-closed dimer were extracted from the simulations, and a geometrical optimisation was performed using three different levels of theory: the DFT, the GFN2-xTB and PM6 SQM methods. Additionally, a representative structure of all the dimer conformations (A-B, B-B, A-C, and C-C) were extracted and compared in terms of energies with the other methods.

II. E - DFT AND GFN2-xTB COMPARISON:

In Figure 23, several structures were extracted from the previous MD. These structures were then optimised in geometry and energy using three different methods: the Density-Functional Theory (using B3LYP basis set), the PM6 SQM method and the GFN2-xTB SQM method. All of them were optimised in water implicit solvation models. In this figure, the geometries are comparable, and only the geometry using the PM6 methods differed a bit from the two others. In most cases, the optimised geometries obtained with the GFN2-xTB method were similar to those obtained with the B3LYP method. The GFN2-xTB method, compared to the others, uses a much lower computational time.

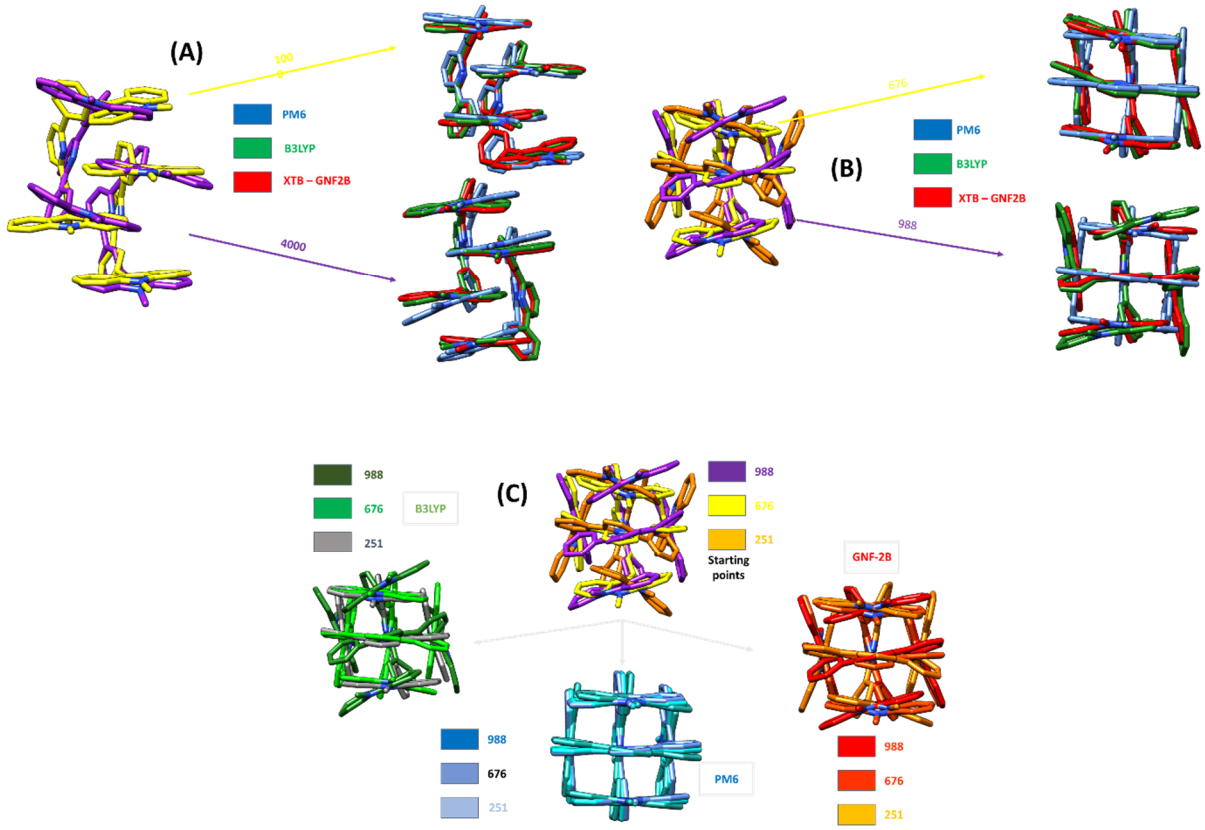


Figure 23: Geometrical comparison between QM level (B3LYP) and two SQM methods (PM6 and GFN2-xTB), the number represents the frame of the MD from which the representative structure is extracted.

Several geometries were extracted from the MD, and the thermodynamic properties were calculated using the DFT, GFN2-xTB and PM6 methods in the following Figure 24. Different methods used different ways to compute the thermodynamic properties, so the energy cannot be compared directly between the methods. What is comparable is the difference between the referenced crystal energy and the others. For these, the energetical difference was comparable between *GFN2-xTB* and the *DFT*. Only the energy calculated with the *GFN2-xTB* method is presented in the graphic, the others are omitted for clarity.

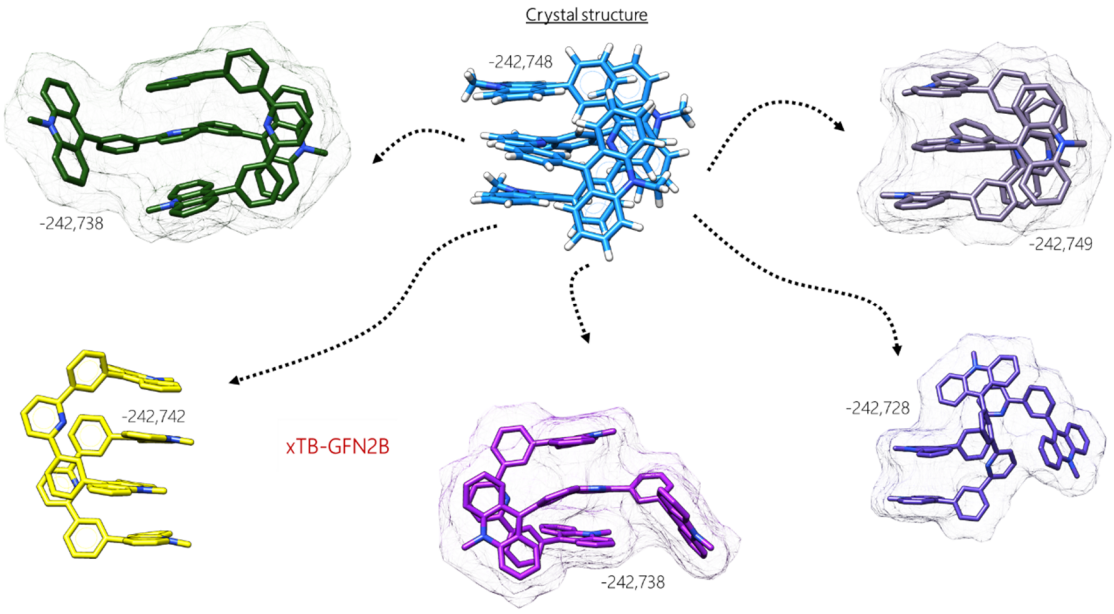


Figure 24: Energetical analysis of the molecular tweezer

Analysis of the acquired energies showed that:

- The MD of the crystal structure shows several geometries that have more favourable energies than the reference crystallographic structure with very similar geometry.
- The SQM method GFN2-xTB results are comparable geometry with the DFT-B3LYP,6-31** with a much lower-computational time.
- None of the geometries obtained by the unbiased MD obtains lower energy than the crystallographic structure even though they are close in energy (~3-6 kcal/mol).

At this step, we find a way to **determine the binding mode** of host-guest complexes using several methods from the less accurate to the more accurate: (i) docking, (ii) docking followed by MD simulations, and (iii) unbiased MD leading to the spontaneous formation of the host-guest complex.

The **behaviour** of the system in a solvated environment have been analysed by an unsupervised ML method (PCA) using a set of automatically and manually generated molecular descriptors. This analysis has led to the identification of several possible conformational states from which the thermodynamic properties have been calculated. Also, molecular descriptors can be useful to clustering and discriminating the geometries of the host-guest complexes during MD simulations leads us to think that the development of ML methods for directly predicting the binding free energy could be an interesting perspective.

The **geometric optimisation and the thermodynamic properties** of host-guest systems have been calculated using the GFN2-xTB methods, with low computational times and interesting accuracy than the DFT standard method.

The protocol used during the analysis of the molecular tweezer was extracted and used to develop an automated platform called HG-DYNAusor. The current platform uses an updated version of the previously presented protocol to determine the binding free energy of host-guest complexes. The clustering step is not well automatable, as it needs manual intervention and is not sufficiently automatized to be used by non-experts, but as it needs the MD simulation, it has been added to this chapter. The HG-DYNAusor platform is still in development and what is presented in this chapter represents the platform in its most successful form. The next chapter will be dedicated to the usage of the platform.

III - THE THERMODYNAMIC BASED APPROACH OF THE HG-DYNAUSOR PLATFORM

III. A - REQUIRED SOFTWARE

III. A. 1 - *OPEN BABEL*

OPEN BABEL is an open-source software package used for cheminformatics applications. Its purpose is to provide users with programs and software libraries designed for molecular modelling, file and data format conversion. OPEN BABEL is a free and open-source software released under a GNU General Public License (GPL) 2.0.[133] OPEN BABEL was used in the platform to calculate the MolPrint2D (MPD) codes and file conversions.

III. A. 2 - UCSF *CHIMERA*

UCSF *CHIMERA* is a molecular visualizer program that can be used to generate high-quality images and animations. *CHIMERA* is free for academic, government, nonprofit, and personal use.[134] CHIMERA was used to visualise and export the PDB and mol2 files of the host-guest complex.

III. A. 3 - AMBER & AMBERTOOLS

AMBER is a suite of biomolecular simulation programs. The term "AMBER" refers to a set of MM force fields for the simulation of biomolecules and a package of molecular simulation programs. AMBER is developed in an active collaboration of David Case at Rutgers University[135]

AMBERTOOLS consists of independently developed packages that work with Amber20. The suite can also carry out complete MD simulations.

The AMBERTOOLS suite is free, and its components are mostly released under the GNU General Public License (GPL).[136]

The HG-DYNAusor platform used several packages of AMBERTOOLS:

- *antechamber* and *MCPB.py*: programs to create force fields for general organic molecules and metal centres.
- *tleap* and *parmed*: basic preparatory tools for AMBER simulations.
- *sander*: workhorse program for MD simulations.
- *cpptraj*: tools for analyzing structure and dynamics in trajectories.

III. A. 4 - VMD (VISUAL MD)

VMD is another molecular visualizer program that can generate high-quality images and animations and export structures. The Theoretical and Computational Biophysics group has developed this software.[137] We are using it to parse the xyz/mol2 files into a proper format that conserves all the pieces of information and coordinates and the usage of *CHIMERA*.

III. A. 5 - GFN2-XTB

The *xTB* software was developed by the Grimme group in Bonn. *xTB* is used for geometric optimisation and hessian calculation of the initial structure. This software is also used for the determination of the thermodynamic properties (enthalpy and entropy) of the guest, the host and the host-guest complexes.[138]

III. A. 6 - R SOFTWARE

R is a free software environment for statistical computing and graphics. It was used for both the development and the optimisation of the knowledge-based method along with the CARET package.[139]

III. B - PARAMETRISATION OF THE HOST SYSTEM (MODULE 01)

III. B. 1 - GRAPHICAL OVERVIEW

Module 01 of the HG-DYNAusor platform is presented in the following Figure 25. It uses an input file containing the necessary information concerning the system to simulate. The first module of the HG-DYNAusor platform can be separated into six parts that will be presented in detail in the next sections:

1) Generation of a reasonable 3D conformation of the host system, which will be used in all subsequent steps. This step can be avoided in case the considered molecular structure comes from a well-defined crystal.
2) The parametrisation of the metal centre. Once again, in case the host system does not contain metal, this step can be avoided.
3) Calculation of the partial charges for the system of interest using an approach that we have developed.
4) Generation of the topological files for the considered system. Topology files contain all the parameters necessary to carry out molecular simulations.
5) Minimization of the system before the MD.

6) Creation of the equilibration and production files.

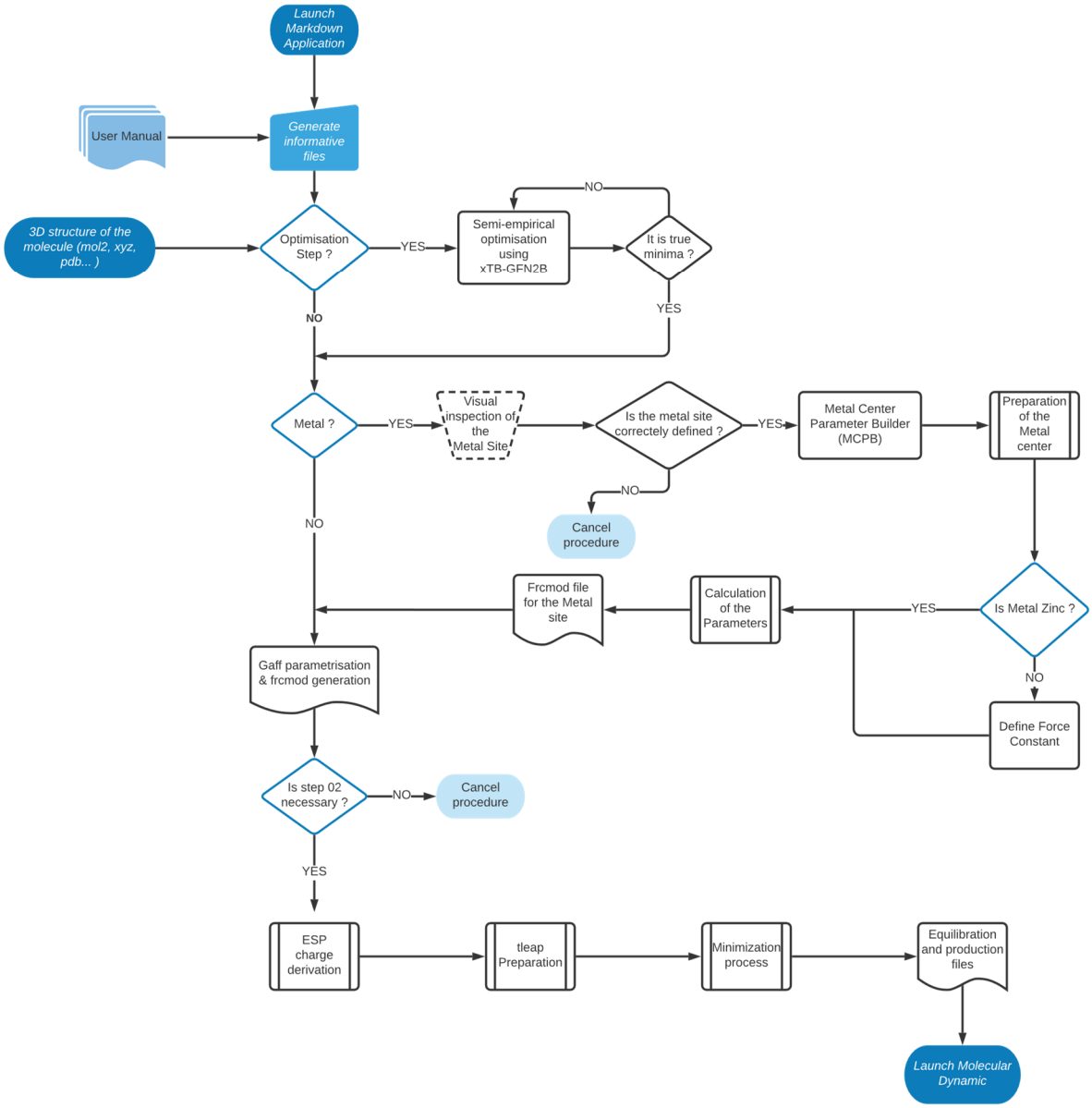


Figure 25: Overview of the first module of the HG-DYNAusor platform

III. B. 2 - HG-DYNAUSOR INPUT-FILE

As shown in Figure 25, the HG-DYNAusor platform needs two inputs: a 3D structure (optimised or not) and an input file containing the information defined by the users concerning the modelling information. This input-file contains two binary questions to let the users decide if he wants to launch (i) only the geometric optimisation and the parametrisation of the metal centre (*frcmod file*) or (ii) the full parametrisation of the system for MD simulation (including partial charges calculation and all following steps to the equilibration & production).

Concerning the geometric optimisation, the users can define in the input files the parameters: (i) the level of optimisation for the GFN2-xTB SQM method (presented in Table 2), and (ii) the solvation model that has to be used (both for SQM and MM modelling) and the available grid for implicit solvation procedure (presented in Table 3). Several structural information is also necessary for the modelling procedure: (i) the molecular charge of the host system, (ii) which metal has to be parametrised (Fe, Co, Zn…), and (iii) the number of metals in the system.

Users will also need to define a "name", which will be used to name all files created during the modelling procedure by the HG-DYNAusor platform.

The last necessary input contains the information about the substructures contained in the host. As the parametrisation of the metal centre requires the extraction of the metal from the host, at some point, it is possible to break certain bonds in the initial host leading to the formation of independent units. These independent units formed must be considered independent molecules in the modelling process, and therefore, they must be named and defined. An example of the procedure for defining how many substructures you will present in your host system is presented in the following Figure 26:

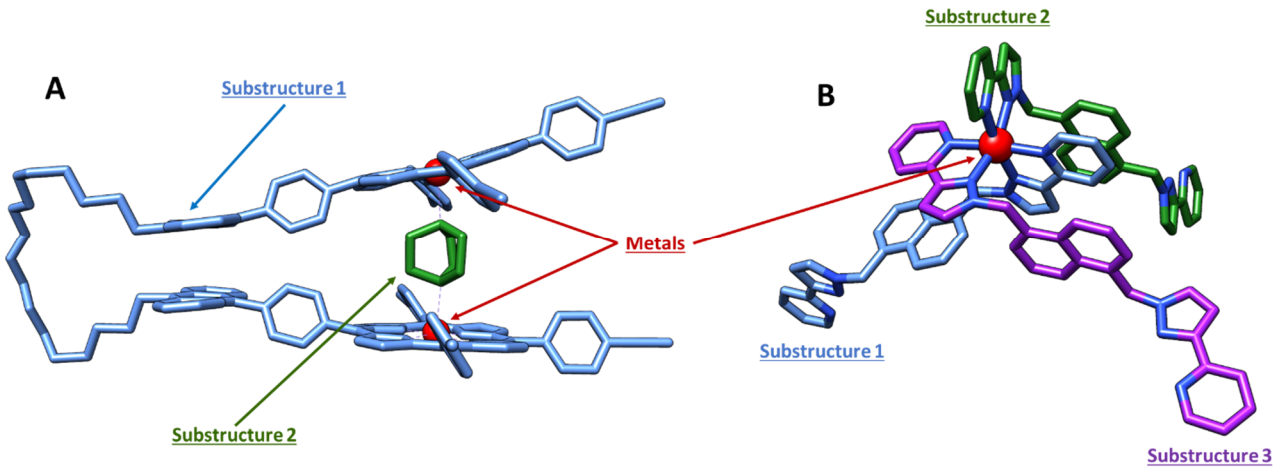


Figure 26: Example of the substructure determination

- In the case of the porphyrin-receptor (Figure 26A), when the metal (in red) is extracted, we realise that the 1,4-diazabicyclo[2.2.2]octane (DABCO, in green) does not make any bonds anymore with the system, meaning it became an independent substructure. In this specific case, by extracting the metal, we create two independent substructures that are parts of the receptor: substructure one corresponds to the porphyrin scaffold (in blue), and substructure two corresponds to the DABCO (in green).

- In the case of the other metal-cage (Figure 26B), the metal is coordinating three structures, i.e., by extracting the metal (in red): three independent substructures will be created: substructure one (in blue), the substructure two (in green) and the substructure three (in purple).

To help users fulfil all the requested information about the input files, a dynamic shiny document have been created where the users fulfil the asked information and press the button "generate the input-files" to get a file that the system will properly read and that contains all the requested variable for the modelling procedure.

III. B. 3 - GEOMETRICAL OPTIMISATION OF THE HOST

The parametrisation of the metal centre is part of module one of the HG-DYNAusor platform, mainly dedicated to the geometrical optimisation of the host structure and the parametrisation of the metal centre. The geometrical optimisation uses the GFN2-xTB method to optimize the given 3D structure that may present some structural error. The geometrical optimizer develops in the *xTB* software called an approximate normal coordinate rational function optimizer (*ANCopt*), which uses a Lindh-type model Hessian to generate an approximate normal coordinate system. It exists multiple levels of optimisation (presented in Table 2) depending on the allowed change in the total energy at convergence (*Econv*) and the allowed change in the gradient norm at convergence (*Gconv*).

Table 2: optimisation level of the *xTB* software[140]

Level	$Econv/Eh$	$Gconv/Eh \cdot \alpha^{-1}$	Accuracy
crude	5×10^{-4}	1×10^{-2}	3.00
sloppy	1×10^{-4}	6×10^{-3}	3.00
loose	5×10^{-5}	4×10^{-3}	2.00
lax	2×10^{-5}	2×10^{-3}	2.00
normal	5×10^{-6}	1×10^{-3}	1.00
tight	1×10^{-6}	8×10^{-4}	0.20
vtight	1×10^{-7}	2×10^{-4}	0.05
extreme	5×10^{-8}	5×10^{-5}	0.01

The accuracy is handed to the single-point calculations for integral cut-offs and SCF convergence criteria. Adjusted to fit the geometry convergence thresholds automatically. The user of the platform is allowed to select the convergence criteria. Still, if the structure cannot

reach the converge criteria at the maximum level, the computational time increases dramatically. The maximum step of minimization authorized is a function of the number of atoms. The geometrical optimisation is done in the implicit solvation model defined by the user to be consistent with the chemical reality. There are two different models of implicit solvation that can be used in *xTB* when we wrote the thesis: the GBSA and the ALPB (Analytical Linearized Poisson-Boltzmann). For each model, several levels of the possible grid are available and can also be chosen by the users (Table 3).

Table 3: different grid available for the calculation of the SASA term[140]

Gridlevel	Grid points
normal	230
tight	974
verytight	2030
extreme	5810

Larger grids increase computational time and reduce numerical noise in the energy. They can help converge geometry optimisations for large molecules that would otherwise not converge due to numerical noise.

Once the geometry is optimised in implicit solvation using the requested level of optimisation and grid-level, a vibrational frequency calculation is performed on the structure to verify that the initial starting point corresponds to a true minimal. If some vibrational modes are found, it does not mean the structure cannot be simulated, but strong attention has to be paid to the metal centre. Supposing the metal centre is not deformed and presents a good conformation, we can launch the MD from a starting point that does not represent a true minimum, with the idea that along the equilibration and the production, the structure is likely to change sufficiently during the simulation so that in the final sample, several structures correspond to a true energy minimum.

III. B. 4 - PREPARATION & PARAMETRISATION OF THE METAL CENTER

III. B. 4. A - PREPARATION

After the geometric optimisation has been done, the resulting structure will be parsed to form (i) the *frcmod* and the mol2 files of the substructure files and (ii) the final PDB file used during the parametrisation using the *MCPB.py* module of AMBER[141]. The initial protocol designed by AMBER for the metal site parametrisation uses QM level and B3LYP basis set to optimize the structure. As we differ in protocol from the *MCPB.py* module (because we are not doing

the QM optimisation), we are only using the *MCPB.py* to generate the *frcmod* of the metal centre.

A file containing the necessary information will be automatically created by the platform to be used by the *MCPB.py* module. The input files are a form of the following:

```
original_pdb [NAME.pdb]
group_name [NAME]
cut_off 2.8
ion_ids [ID1] [ID2] [ID3] […]
ion_mol2files [METALS.mol2]
naa_mol2files [SUBSTRUCTURE.mol2]
frcmod_files [SUBSTRUCTURE.frcmod]
large_opt 0
```

There should be no blank lines in the input file. The values or parameters should follow the variables separated by a blank space.

The required variables are the following:

1) *original_pdb*: represent the file name of the final PDB file created at the previous step, which should have only one chain. The PDB file has to contain hydrogen atoms and metal ions. Users are advised to use the geometrical optimisation steps before performing the modelling in *MCPB.py*.
2) *ion_id*s: represent the PDB atom Identification of the metal ion(s). If there is only one metal ion in the metal site (or only one metal site), you need to put its PDB atom Identification after the variable. If there are multiple metal ions in the metal site, you need to put the PDB atom I.D.s of all these metal ions separated by space after the variable.
3) *ion_mol2files:* represent the names of the ions mol2 files. Depending on how many metals are included in the host, this can be one or several names. In most cases, we encountered cases where multiple metal centres were in the host, but all of them were equivalent.
4) *naa_mol2files:* the variables used to indicate non-amino acid mol2 file(s) in the host system. Practically this represents all the substructures named by the users. These files are typed with gaff atom initially and at this step does not contain any charge as we are only generating the *frcmod* (and not modelling the system).

5) *frcmod_files*: represent the variable used to indicate the parameter modification file(s) for the host's nonstandard residue(s). It corresponds again to all the substructures named by the users.

Concerning points (4) and (5), It is fundamental that the users define precisely the correct number of the substructure to achieve the metal centre modelling without any issues. Assume that users initially define n substructure in the host (in the HG-DYNAusor input files). In that case, the platform will automatically create at the end of the geometrical optimisation an input-file for the *MCPB.py* containing n mol2 files typed with gaff for the substructures + the n associated *frcmod* files.

There are also optional variables in our input files that represent:

1) *The group name*: represent the name the user has specified in the HG-DYNAusor input files. The group name will be used as a prefix to name all the files generated with the *MCPB.py* module.
2) *cut_off*: the cutoff value is used to indicate there is a bond between the metal ion and the surrounding atoms. The default is 2.8Å. In the current form of the platform, the user is not allowed to change this value.
3) *Large_opt:* a variable used to indicate whether to do a geometry optimisation in the Gaussian input file. Several options are available. The 0 means no optimisation (as the optimisation were realised in the previous steps).

III. B. 4. B - PARAMETRISATION OF THE METAL CENTER

Once the *MCPB.py* input files have been created, the module will be launched. Three different files will be created: (i) the MCPBY_1.out file containing pieces of information about the metal-binding that will be used in the next part for the preparation of the metal centre (during module-two of the HG-DYNAusor platform) using *tleap*, (ii) the MCPB_2.out file containing all the information about the nonstandard residue(s) of the metal centre (angles, improper, bonds, …) and (iii) the *frcmod* files containing the pieces of information extracted from the MCPB_2.out file built in a way that AMBER can interpret them.

Unlike the classical way of using the *MCPB.py*, we are not using QM to reduce the computational cost of the protocol because both the optimisation and the determination of the thermodynamic properties are done using the *xTB*-software. In addition, the modelling of such

host will be realised in the HG-DYNAusor module-02, apart from the *MCPB.py*. In conclusion, the *MCPB.py* module is used in our protocol only to generate the *frcmod* file because it contains an automatic way to extract the information about the metal-binding site.

III. B. 5 - PARTIAL CHARGES DERIVATION

III. B. 5. A - GRAPHICAL OVERVIEW

An overview of the partial-charges calculation process is presented in the following figure:

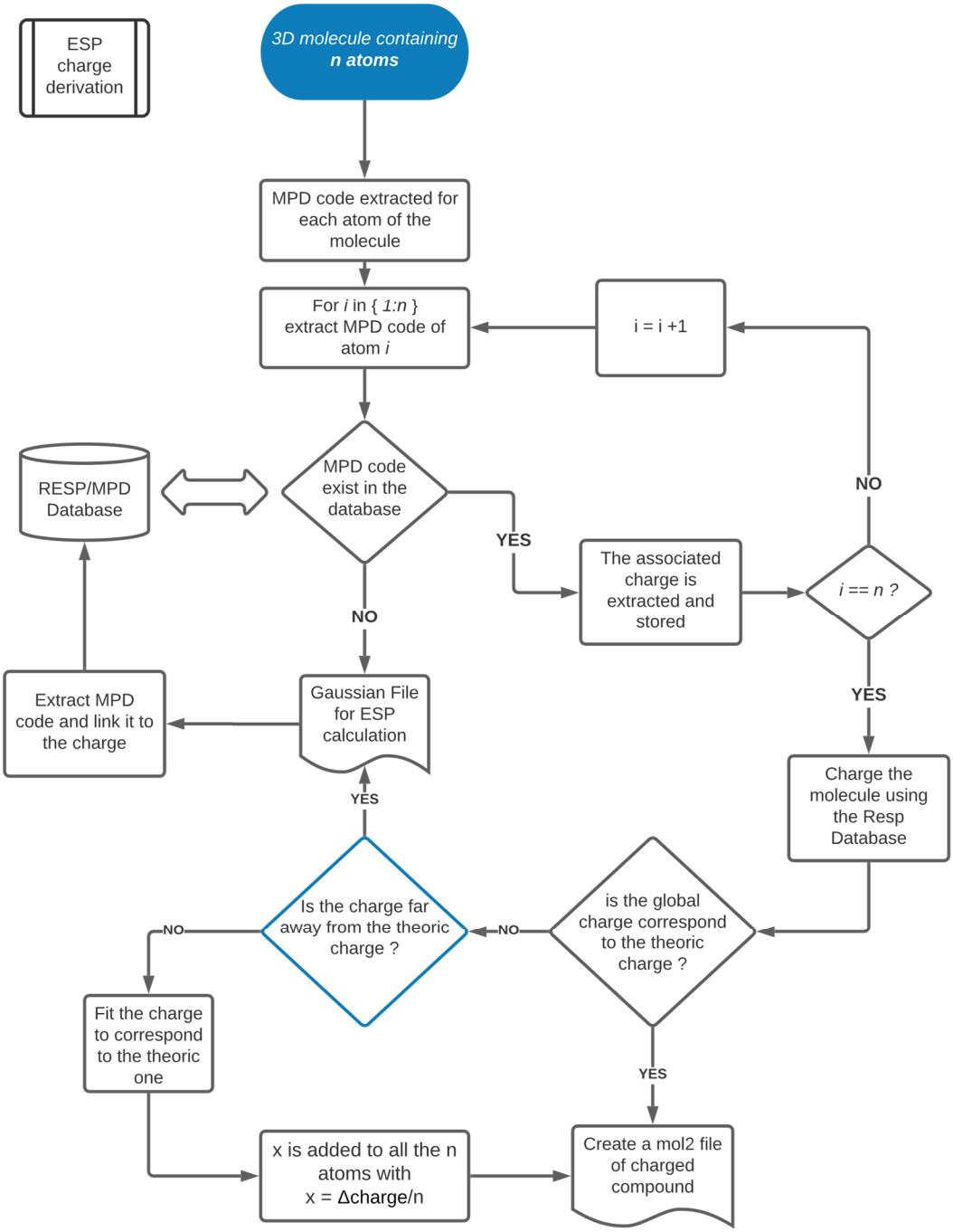


Figure 27: Overview of the MD-RESP protocol for the generation of the atomic partial charges

III. B. 5. B - MODEL OF CHARGES

As we presented in the introduction, it exists several models of partial charges. A partial charge is always inferior in absolute value to the elementary charge. These partial charges are fundamental for the parametrisation of host and guest systems. Some of these are easy to calculate:

- Gasteiger-Marsilli partial charge[53]: One of the simplest partial charge models used for the Docking protocol. Generally, these charges are calculated in two steps: first, the charges are assigned to each atom in the molecule, then in the second step, the initial charges are shared across the bonds, moving a certain amount of charges from one atom to another determined by the difference in term of electronegativities of the atoms at the end of each bond. These charges are mainly depending on the difference in terms of electronegativity.

- *xTB*-GFN charges[65]: the *xTB*-GFN method has its own possibilities to calculate charges derivated from the Mulliken and Charge Model 5 (CM5) charges, taking into account the chemical specificities of the considered molecules. They are really easy to calculate, and it only takes seconds on a local computer to calculate the partial atomic charge of the constituents of a molecule. These charges can be considered as more efficient than the gasteiger one.

The Gasteiger-Marsilli and the *xTB*-GFN charges may be too simple to accurately represent the host and guest system with metallic compounds, especially MD. And while the gasteiger charges are only used by certain docking programs, the *xTB* ones are used during the geometric optimisation steps. For more accuracy and to generate proper charge models for measuring the interaction between host and guest, SQM AM1-bcc and RESP charge will be considered and compared during the thesis work.

- The RESP[142] (Restrained Electrostatic Potential) method is a multistage approach that ensures atoms with free rotation are considered equivalent in charge (like hydrogen or methyl groups). It uses the HF/6.31G* QM calculation to generate the electrostatic potential. As we already said in the introduction, the RESP charge leads to better accuracy in the binding free energy prediction. Unfortunately, for the resp calculation, we need to use licensed software (*GAUSSIAN*), and it takes a larger simulation time than the other methods. Furthermore, as RESP calculations are geometry dependent,

for flexible systems, some artefacts are likely to occur during QM calculations for which RESP fits may be stuck in high energy minima. One way to reduce these artefacts is to use smaller molecules.[143]

- AM1-bcc[57]: the AM1-BCC charges are composed of two parts: the SQM (AM1) with bond charge correction (BCC). AM1BCC charges start with Mulliken-type partial charges derived from the AM1 SQM wave-function. In a second stage, bond-charge corrections (BCCs) are applied to the partial charges on each atom to generate new partial charges. AM1-BCC charges can be calculated with antechamber, part of the AMBERTOOLS package. The main advantage of the AM1-bcc charges is the low computational cost (can be launched on a local computer) and that no licensed software is needed.

In theory, the RESP charge calculation should be restrained to small molecules. In our case, the final expected receptor synthesized by the different partners of the project can be up to > 400 atoms. In the course of our research, we faced several different problems regarding the use of RESP charge for MD simulations: (i) A high computational cost to the calculation, (ii) the 3D conformation of the hosts often includes atoms that are not solvent-exposed. As such, the location of the points where the electrostatic potential is calculated can be quite distant from those atoms, and the derived ESP charges may not reflect the true nature of the atom, and (iii) since the receptors under consideration are mostly still in the research phase, some structural changes may occur due to changes in the chemical procedure of synthesis, and unfortunately, in this case, given the RESP procedure, the charge must be recalculated for the entire structure in case of change.

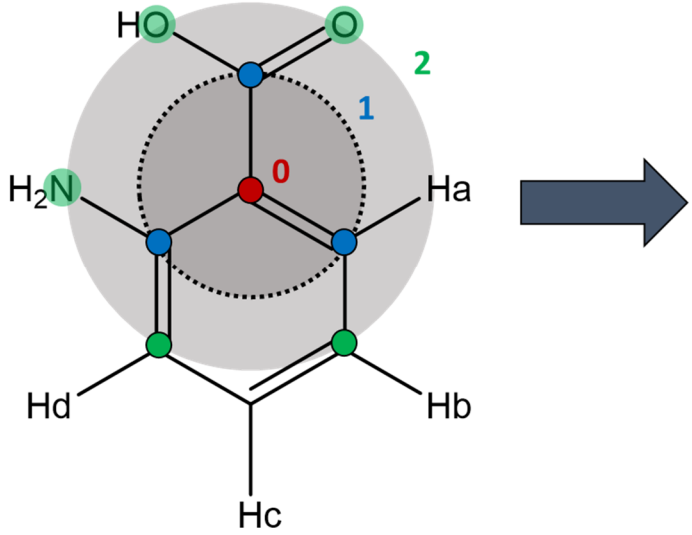
For these reasons, we designed a protocol where each atom of the molecules is described by a 2D code describing the molecular environment (neighbour). For each considered atom, the 2D code is associated with the RESP charge. We made the assumption that the partial charge of an atom of a huge receptor depends almost completely on its nearest neighbours and not on the entire receptor. Thus, being able to describe each atom according to its type and chemical environment, we can calculate the partial charges for a new structurally close receptor. For this purpose, the constant part of the receptor (which has not undergone any structural change on its closest neighbour) can be extracted from a database mapping the 2D code to the charge. Thus, the new charges need only be calculated for the variable part of the receptor. This

methodology saves a significant amount of time in the parameterisation of the receptors while ensuring the possibility of making structural modifications without restarting from scratch.

III. B. 5. C - MOLPRINT2D (MPD) FORMAT

MDP is an atom-environment fingerprint developed by Bender and al[144], used in QSAR studies and for measuring molecular similarity. Based on the molecular environment. At this step, we are not properly doing any QSAR analysis, but we wanted a way to consider and write every single atom of a molecule as a unique code that takes care of their neighbour.

In the case of a heavy atom, the generation of the MPD format starts from mol2 files. From this mol2 file, the TRIPOS atom type is considered, with the idea that every atom type can be defined as a number. The decomposition of the MPD code is presented in Figure 28:

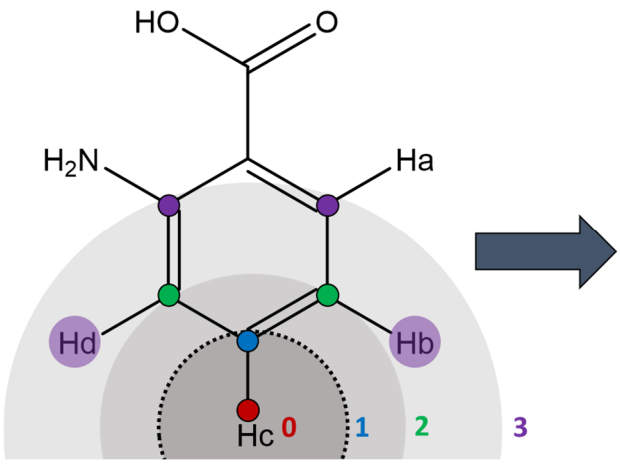


Layer	0	1	2
	C.ar	C.ar	C.ar
		C.ar	C.ar
		C.cooh	N.sp3
			O.sp2
			O.sp3
			Ha

Figure 28: MPD example: The aromatic carbon (in red) is described by his two neighbours

The MPD code for the carbon atom highlighted in red would be (i) the considered aromatic carbon atom for the layer 0 (in red), (ii) followed by three atoms in the first layer (in blue), two aromatic carbons and the carbon linked to the acidic function, and (iii) the second layer corresponding to the atoms linked to the three ones considered in the previous layer: two others aromatic carbons: a sp3-nitrogen, hydrogen and the two acidic oxygen: one sp2 and the other sp3.

In the specific case of hydrogen, two neighbours are not sufficient to efficiently describe the molecular environment. It makes sense because the first layer of the hydrogen represents only one atom leading to a simplification of the environment: almost all the hydrogens were considered equivalent. For that reason, for hydrogen, the third layer is considered, an example is presented in the following Figure 29:



Layer	0	1	2	3
	Hc	C.ar	C.ar	C.ar
			C.ar	C.ar
				Hb
				Hd

Figure 29: MPD example: the specific case of the hydrogens

In this figure, another layer is considered for the hydrogen, meaning that for hydrogen, the first layer will always be composed of a single atom corresponding to the atom that links the hydrogen, after that, the layers are developed as just presented.

III. B. 5. D - DATABASE BASED ON FRAGMENTS

The MPD database is built in two columns, the first one represents the MPD code, and the second one represents the associated RESP charge. It contains 480 samples that have been calculated during the thesis. For the parametrisation of a new host system containing atoms for which existing missing values in the database, a new ESP charge calculation is necessary.

The protocol is the following: at this step, we start from an optimised 3D molecule for which the metal centre and the substructure have been typed with gaff forcefield. The first step is to generate the MPD code of all the atoms in the molecule by following the procedure written in the following algorithm:

ALGORITHM 1: ***RESP charge generation***

Considering *n* atoms in a 3D molecule:

$\boldsymbol{for}\ i = 1\ to$ n, $\boldsymbol{do}$

Generate the MPD code.

$extract\ the\ MPD\ code\ of\ the\ atom\ i\ (MPD_i)$: $search\ MPD_i\ to\ the\ database$

$if\ MPD_i\ is\ existing\ in\ the\ database$; $extract\ the\ associated\ charge$

$if\ MPD_i\ is\ existing\ in\ the\ database;\ \ i = i + 1$

$\boldsymbol{else}; \boldsymbol{do}$

$Write\ a\ warning\ message\ for\ atom\ number\ i$

$Write\ the\ script\ for\ new\ ESP\ calculation\ to\ add\ new\ charge\ to\ the\ database$

$\boldsymbol{end}$ **for;**

$\boldsymbol{i = n}$

At this moment, we have two possibilities: In the first case, one or more MPD codes have not been found in the database, the extraction of the RESP charge is incomplete, and the parameterisation cannot be completed. In the second case, all atoms have their corresponding MPD code in the database, and a RESP charge for all atoms can be extracted. But as we are extracting the charge from a database, we can expect that the sum of partial charges will have a little deviation compared to the actual net charge of the molecule. In that case, all the charged-atom will be adjusted by distributing across all atoms the numerical difference between the calculated charge using the MPD-RESP protocol and the actual net charge. This difference is generally extremely low and is not expected to have any impact on the simulation.

Assuming the MPD-RESP procedure have been done without any problem, at this step, we already obtained the optimised 3D molecules typed with gaff atom type, with RESP charge, and extracted from the first module: the parameters of the metal centre and all the substructure that compose the molecule. The next step concerns the creation of the topological files to be able to launch the MD simulations.

III. B. 6 - GENERATION OF THE TOPOLOGICAL FILES

An overview of this process is presented in the following figure:

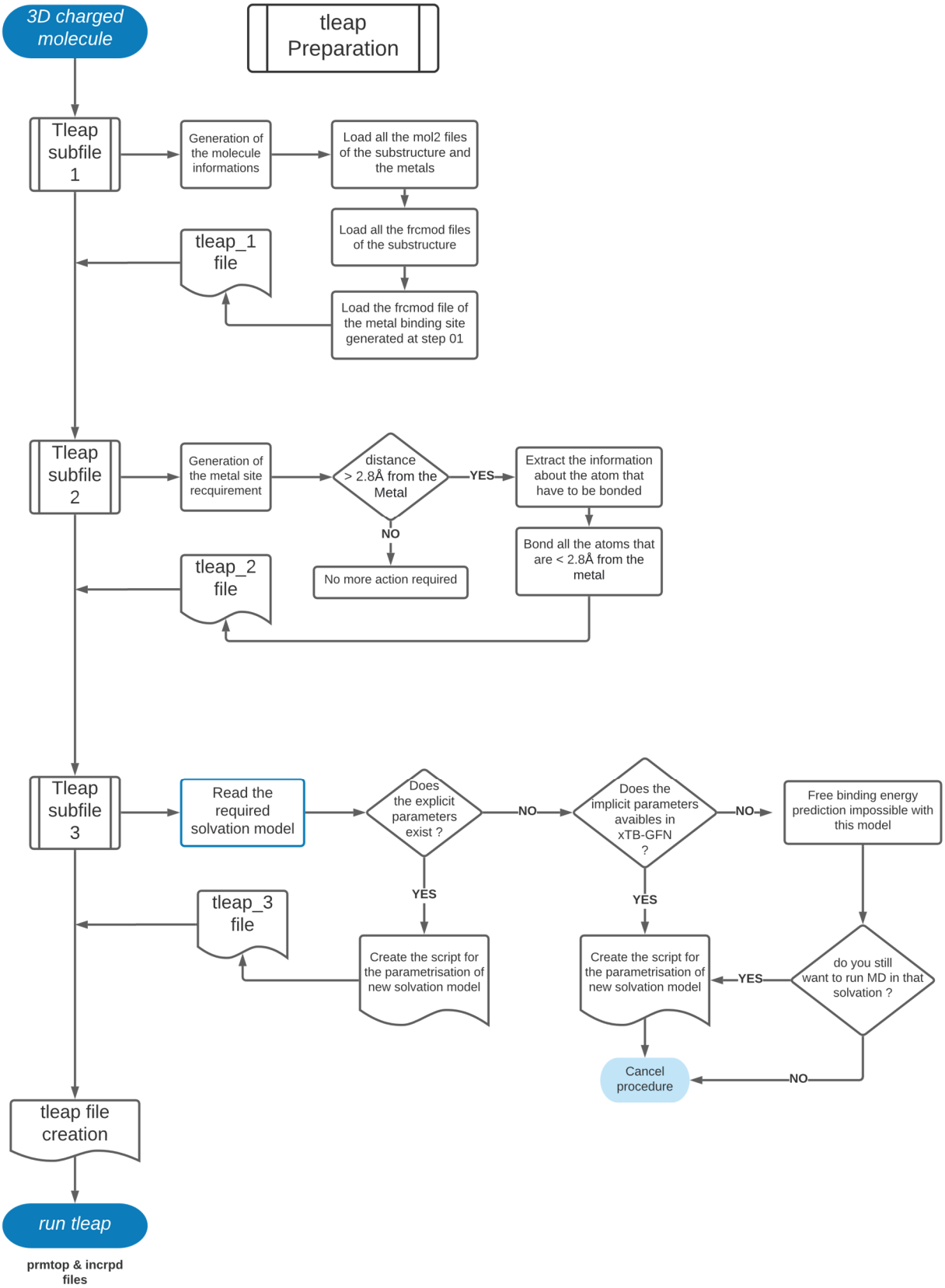


Figure 30: An overview of the generation of the topological files

The procedure to generate the topological files for the AMBER force field and the solvated model use the *tleap* helper program from the AMBERTOOLS suite. *tleap* will (i) read the previously generated parameters of the different substructures and the metal centre, (ii) bind the atoms close to the metal for the generation of the bonded model, and (iii) solvate and neutralize the solute with the requested solvated model.

The program will automatically check if the solvent specified in the input file is available or not. If the solvent is not parametrised, it is possible to generate the file to parametrize a new solvent model for the MD using *turbomole* software. But if the goal is to make binding free energy prediction, the solvent model has to be also available in the *xTB* software. A list of the available solvation model can be shown in the following Table 4. In both cases, if the solvent parameters are not available, the platform will stop with an error message.

Table 4: Available solvation models in the different software

Solvents	Parametrised for MD	Available in *xTB* software	
		GFN2(ALPB)	GFN2(ALPB)
Acetone	YES	YES	YES
Acetonitrile	YES	YES	YES
Aniline	NO	YES	NO
Benzaldehyde	NO	YES	NO
Benzene	NO	YES	YES
CH_2Cl_2	YES	YES	YES
$CHCl_3$	YES	YES	YES
CS_2	NO	YES	YES
Dioxane	NO	YES	NO
DMF	NO	YES	NO
DMSO	YES	YES	YES
Ether	NO	YES	YES
Ethylacetate	NO	YES	NO
Furane	NO	YES	NO
Hexandecane	NO	YES	NO
Hexane	NO	YES	YES
Methanol	NO	NO	YES
Nitromethane	NO	YES	NO
Octanol	NO	YES	NO
Phenol	NO	YES	NO
Toluene	NO	YES	YES
THF	NO	YES	YES

Water (H_2O)	YES	YES	YES

III. B. 7 - MINIMIZATION, EQUILIBRATION, AND PRODUCTION

The minimization, equilibration and production steps are presented in the following flowchart (Figure 31):

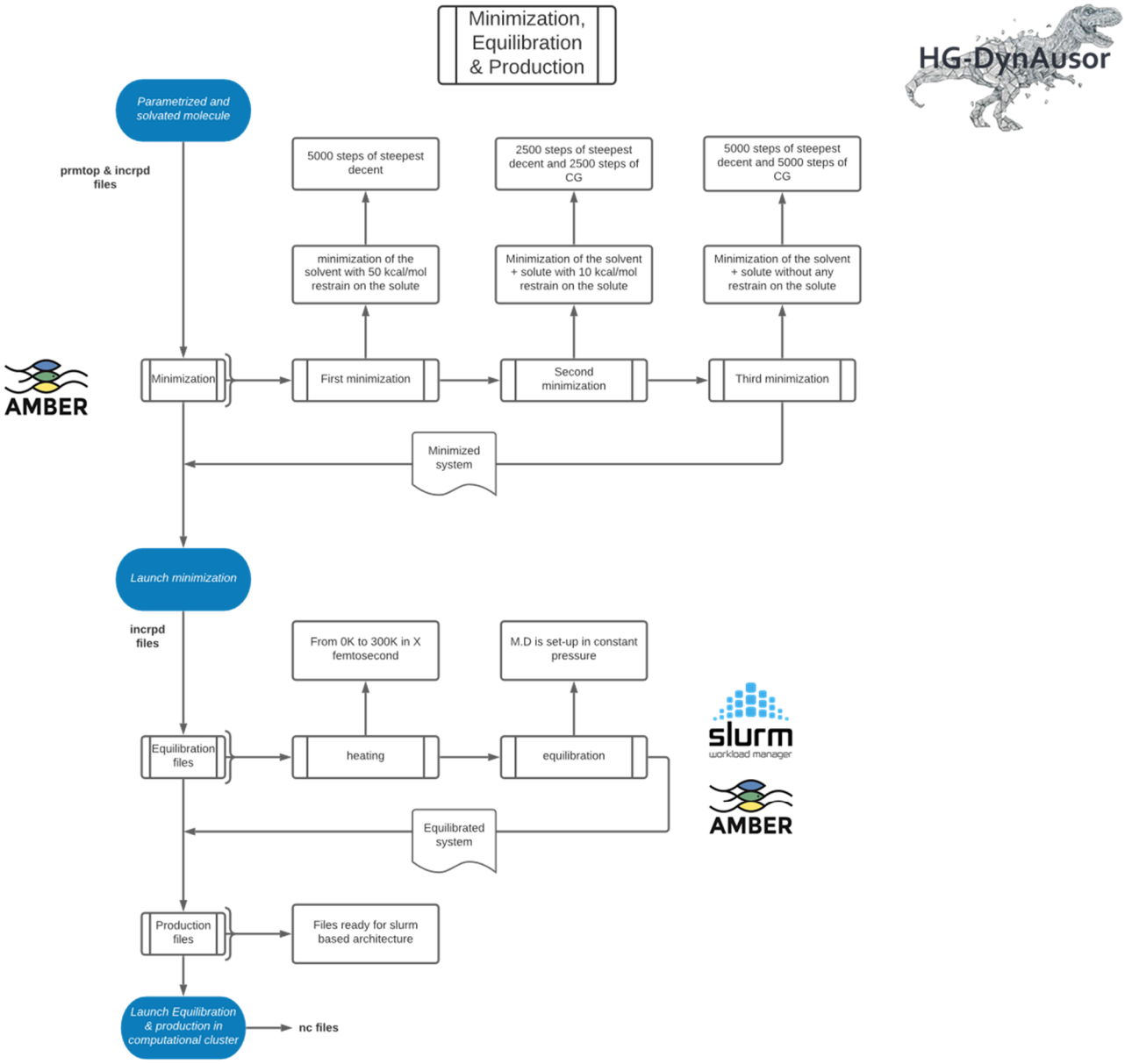


Figure 31: An overview of the last phases of module 01 of the HG-DYNAusor platform

Two different protocols for MD have been tried, the first on using the GROMACS software while the second one uses the AMBER software. In both cases, the topological files were generated using the AMBER force field. The AMBER topological files (prmtop and inpcrd) were converted into the GROMACS topological files (gro and top) using the *parmed* module of AMBERTOOLS. For simplification, only the details for the AMBER will be presented.

The system is minimized with the *sander* program of the AMBERTOOLS suite, using both the steepest descent algorithm and conjugate gradient algorithms in several steps (i) 5000 cycles of minimization are performed using the steepest-descent algorithm applying a very large force constant (500 kcal/mol/A²) to the host and the guest, (ii) 5 000 additional cycle of minimization using 2 500 steps of each algorithms applying a lower restraint on the solute (10 kcal/mol/A²), (iii) at the end, an additional 10 000 cycles of minimization are performed using 5 000 steps of each algorithm without any force restraint. The system is then slowly heated from 0 to 300 K during 500ps and equilibrated for 400ps with the leapfrog integrator in the NTP ensemble (constant pressure).

Both equilibration and production take too much time using the sander program available in the non-licenced version of AMBER. For that reason, a standard file for *Slurm Workload Manager* is created for the GPU usage inside a computing centre for the equilibration and production phases of the MD.

III. C - PARAMETRISATION OF THE GUEST SYSTEM AND BINDING MODE GENERATION (MODULE 02)

The second module of the HG-DYNAusor platform is dedicated to generating the binding mode of the host-guest complex. Different possibilities for binding mode generation are available in the platform and mainly depend on the host's accessible data and molecular structure. The possible options will be developed in this section.

In our application, the study of the binding mode is built as an additional module following the parameterisation of the host system. Three different approaches, already discussed in the introduction, can be considered: (i) Either the binding mode is known and can be directly extracted from a crystallographic structure, (ii) or the binding mode is not known but the host considered is in a conformation suitable for binding, (iii) the host has a very high intrinsic mobility and thus a spontaneous binding approach can be considered. An overview of module 02 of the HG-DYNAusor platform is presented in the following Figure 32.

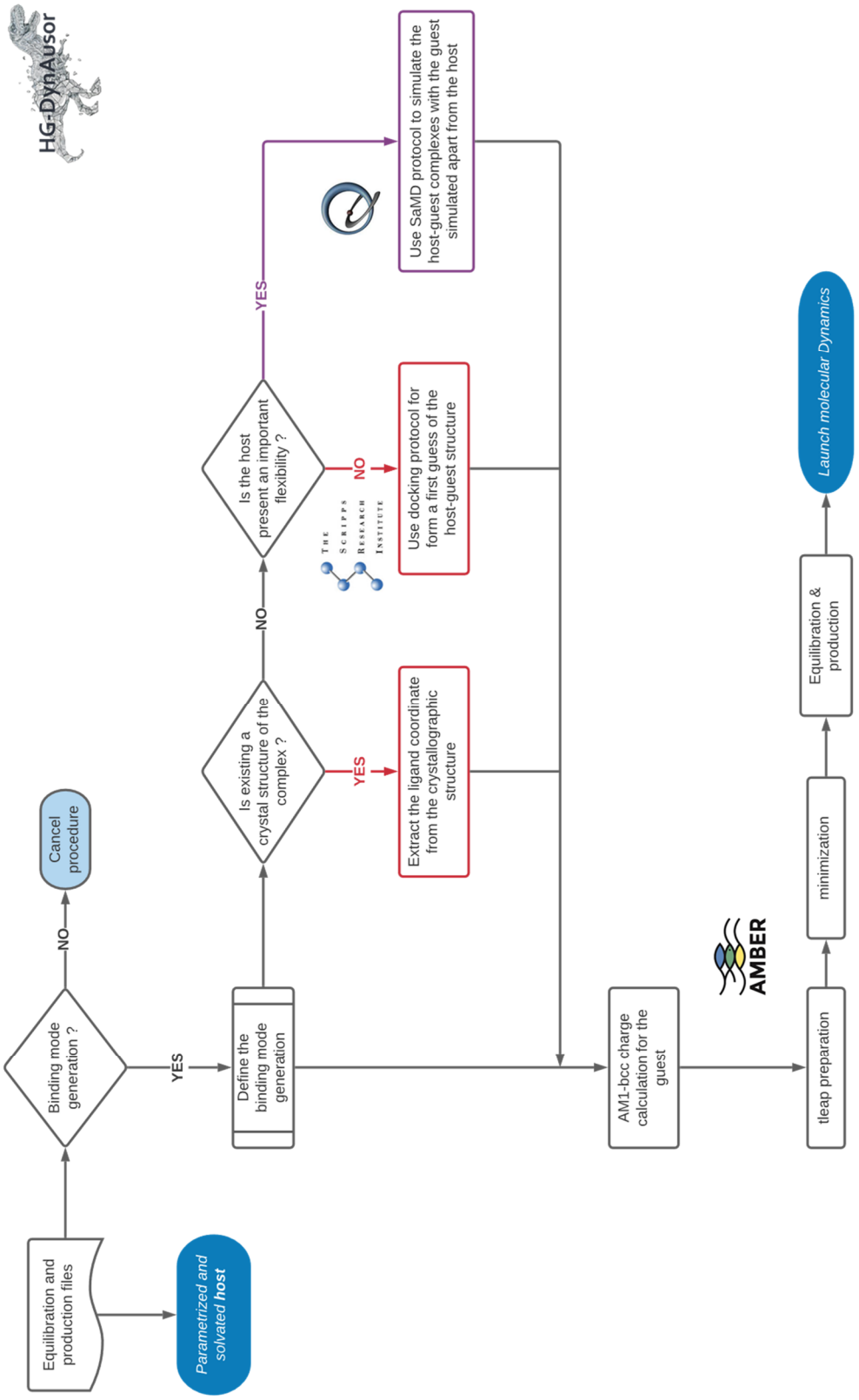


Figure 32: An overview of the second module of the HG-DYNAusor platform dedicated to the binding mode generation

In all the cases, the guest was parametrised apart using the AM1bcc charge model and gaff forcefield. Even though we said before that the RESP charge model is better for determining the molecular contact, if the guest is smaller enough, we can consider the difference between RESP and AM1-bcc charge as minimal. The platform will also generate both the files for the host without ligand and the files of the host-guest complexes apart, as the information about the systems without a guest is necessary for the binding free energy prediction.

The difference comes from the way the binding mode is considered:

- In the first case, as the structure come from a crystal, no further structural optimisation is necessary for the host, and then the coordinate remains unchanged. Thus, the coordinate of the ligand can be extracted and merged for the parametrisation of the host-guest system.

- In the second case, docking is used to generate a first guess of the Host-Guest structure. We chose ADV for our assessment for several reasons. It (i) is faster and generally performs better than AUTODOCK itself, (ii) is freely available and competitive with commercial tools. Docking is performed using AUTODOCK-VINA. Input comprises the host system, guest, and docking box, while output lists pose ranked by Δgbind, the predicted binding energy in kcal/mol ('score' = −Δgbind). To obtain the maximum number of poses, we set num_modes to 20. The top-scored solution is extracted. In some cases, the extraction was followed by the steepest descent and conjugated gradient minimization to correct any ligand distortion.

- In the last case, the best docking pose is modified to start from a dissociated configuration where the guest is > 8Å apart from the host.

Practically speaking, for the generation of new topological files, the pre-generated *tleap* file used for the parametrisation of the host is reused, and the information about the guest is added, leading to an easy generation of the topological files. Finally, the only difference between the host and the host-guest complex concerning the generation of the parameters consists of the addition of the guest's information during the step of generation of the topological parameters (*tleap* preparation).

III. D - THERMODYNAMIC BASED APPROACH FOR BINDING FREE ENERGY PREDICTION (MODULE 03)

Module 3 of the HG-DYNAusor platform is presented in the following Figure 33:

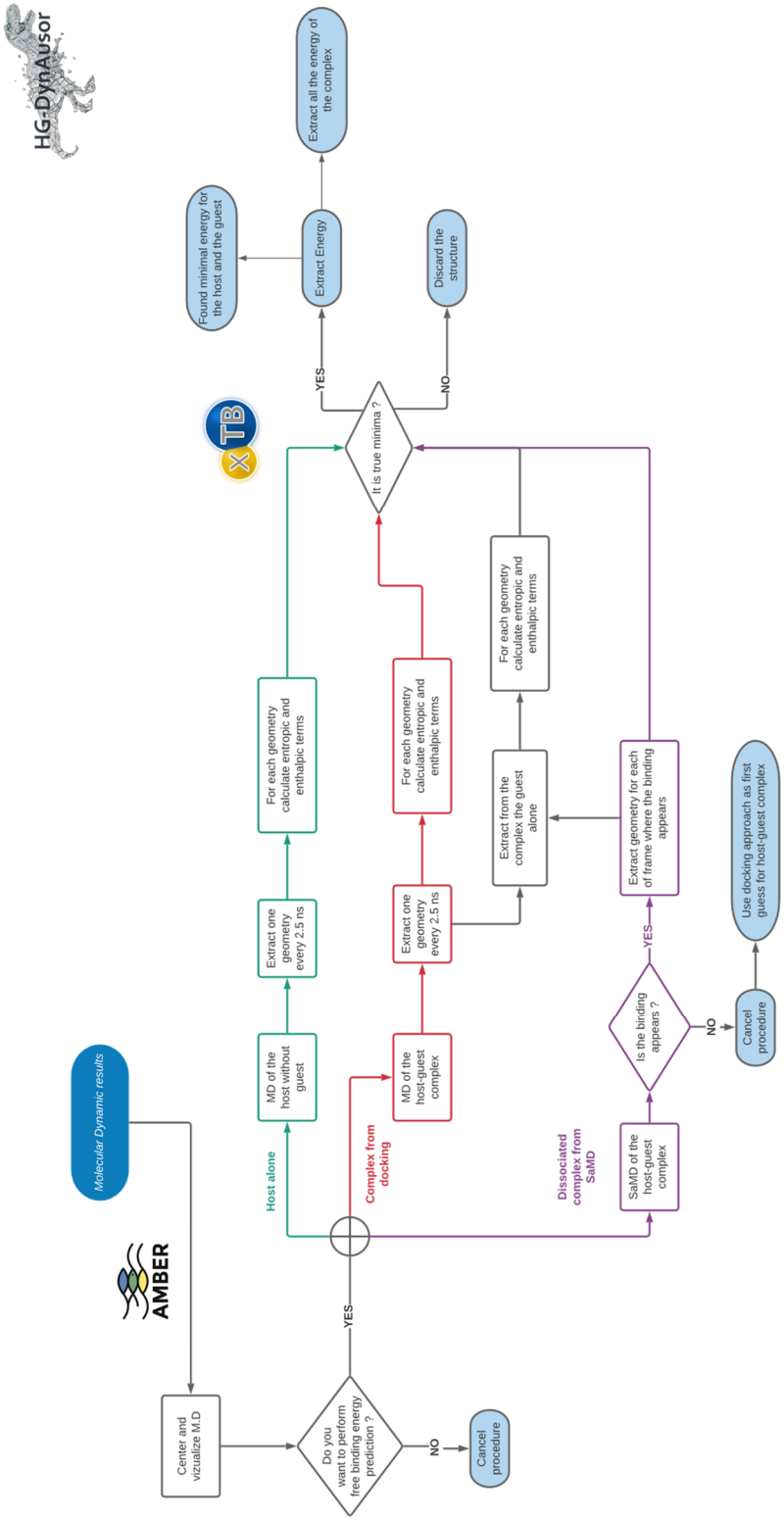


Figure 33: An overview of the third module of the HG-DYNAusor platform dedicated to the thermodynamic approach

The third module performs binding free energy predictions using a molecular simulation approach. It entirely depends on the three previous modules of the platform because it requires: an MD simulation of the host alone and an MD of the complex. If the users want to predict multiple compounds on the same receptor, only one host simulation is necessary. This module uses the *cpptraj* module of the AMBERTOOLS suite from the AMBER software to pre-process the MD.

After the alignment and the extraction of the representative geometries, a thermodynamic calculation is performed to determine both enthalpic and entropic properties of the host, the guest, and the host-guest system using the *xTB* program package (version 6.4.1). It uses the GFN-2B basis on an extended SQM tight-binding model, which has shown to be efficient for determining structures and noncovalent interaction energies for large molecular systems (in the order of 1000 atoms)[65]. Solvents effects are included through GBSA and ALPB models, and the convergence criteria thresholds were set as *extreme*. Optimisation, followed by a hessian calculation, are performed for all extracted geometries.

IV - BEHAVIOUR ANALYSIS

For the clustering of the systems studied in the thesis, non-supervised ML approaches are used to visualise the clusters. Generally speaking, for every MD simulation, some descriptors describing the macroscopic states are extracted for each step of the dynamic (every step represent a geometry): RMSD from the first frame (the equilibrated structure), the SASA, the Rg (representing the compactness of the studied object). To that macroscopic descriptors, multiple other geometrical descriptors can be considered depending on the structure: distance between twos atoms, angles, torsional angles, dihedrals…

Using all the possible descriptors, a dimensional reduction analysis is performed using our protocol, followed by a PCA to cluster the systems using MD. The idea of that clusterisation is to compare similar systems (using the same descriptors) or compare for one system the different geometries taken over time. With the idea that we are able, with these procedures, to study any system over time as long as the descriptors considered are likely to describe the structural variability. During the dimensional reduction phase, the descriptors that are not positively correlated to the variability of the MD will be discarded. In this way, it is possible to study whether the MD should be extended or the effects of solvents on the same system.

V - THE KNOWLEDGE-BASED APPROACH OF THE HG-DYNAUSOR PLATFORM (MODULE 04)

V. A - OVERVIEW OF THE KNOWLEDGE-BASED METHODS

A fourth and last module has been built in the HG-DYNAusor platform dedicated to predicting binding free energy using an ML approach. This last module is completely operational but still not released on the platform because, as we already stated, some steps need manual intervention and are not sufficiently automatized to be used by non-experts.

V. A. 1 - GRAPHICAL OVERVIEW OF THE KNOWLEDGE-BASED METHODS

This approach predicts the binding free energy as a combination of molecular descriptors. It only requires the molecular information of the host and the guest as input. Thus, it does not rely on the structure of the binary complex. The data used to train the ML module are extracted from the Binding Database (BindingDB) and are mainly dedicated to predicting binding free energy in water.

For the knowledge-based protocol using ML algorithms, we tried to decompose the binding mode of the host-guest systems by using an innovative analysis that merges the molecular descriptors of the guest with those of the host to make a model that learns both from the host and the guest structure to predict binding free energy. The initial idea behind this methodology is to predict different binding free energy for the same guest depending on which host it is binding. In addition, we tried to make a model that predicts the binding free energy of one guest based on the molecular consideration of the host, meaning we are predicting more the binding free energy depending on the molecular environment. In conclusion, our ML protocol is a possible way to predict the "unpredictable" because ML learns from the molecular environment of multiple host systems. We hypothesize that by using a new host for which no data are available in the literature but described by molecular descriptors, we should be able to predict the binding free energy with relatively low error (or in the case where we cannot trust numerical prediction, by predicting the ranking of multiple guests in a new host system).

A graphical overview of the protocol is presented in the following Figure 34:

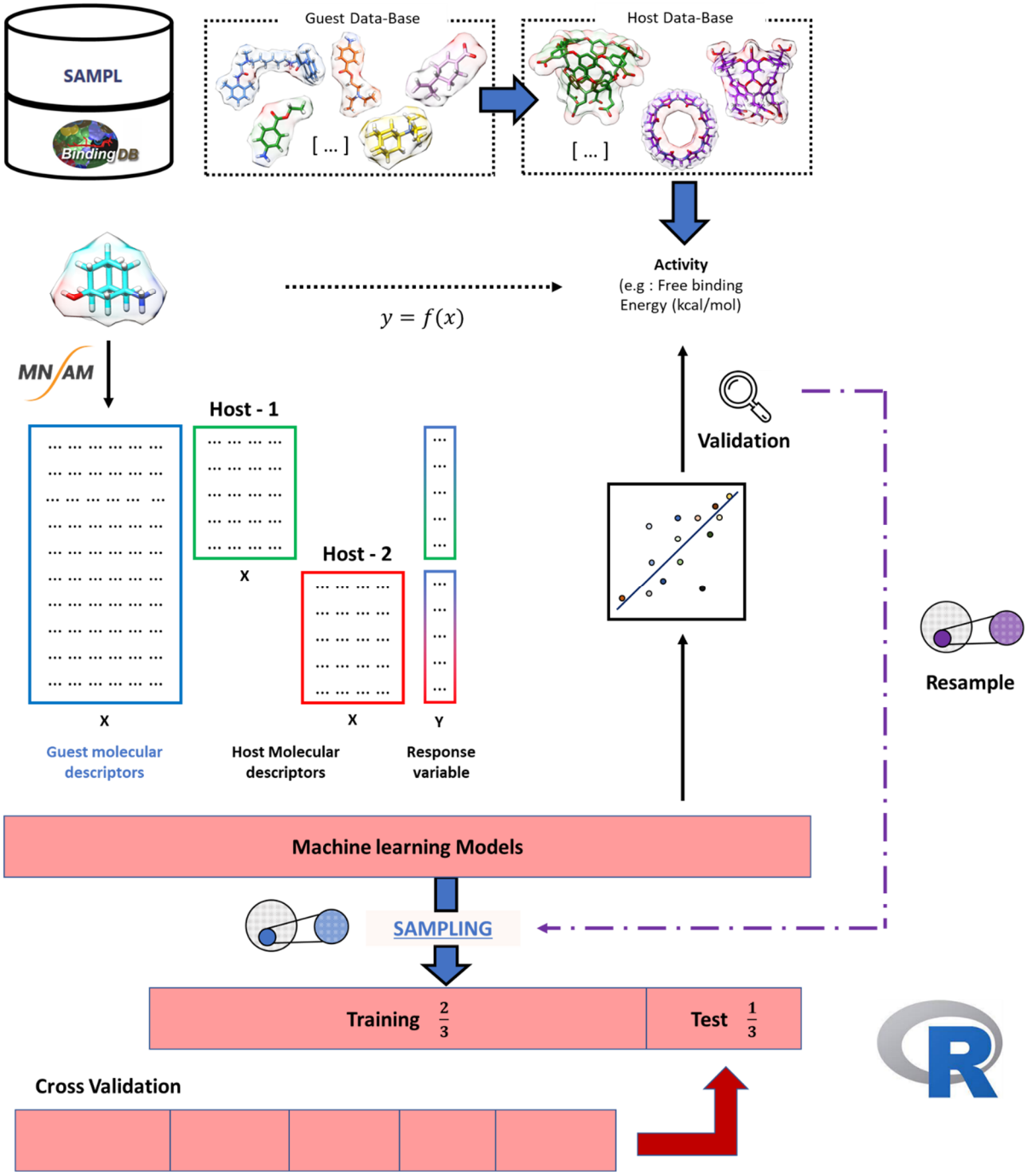


Figure 34: Graphical overview of the knowledge-based protocol

V. A. 2 - DATA PRE-PROCESSING

V. A. 2. A - PROCEDURE

All the binding data from the BindingDB have been extracted and parsed. All the guests involved in the BindingDB and SAMPLs challenge are reconstructed in two steps from SMILES using OPEN BABEL: (i) generating 150 3D conformers based on Genetic-Algorithm,

(ii) and selecting the lowest energy conformers. These conformers are then minimized at an SQM level using GFN2-xTB, giving us an optimised 3D structure. The guests from SAMPL6, SAMPL7, and SAMPL8 challenges are directly extracted in 3D from the GitHub repository and minimized at an SQM level using GFN2-xTB, giving us an optimised 3D structure.

The same methods are used to reconstruct the hosts. In total, > 25 different hosts are extracted and constructed from SMILES provided by the binding-DB following the same protocol. In some cases, we encountered some problems reconstructing the hosts: (i) for the host represented by the following BindingDB(id): BDBM197280, BDBM197287, BDBM197309, BDBM197310, BDBM36281 that mainly correspond to hosts that were used in the previous SAMPL challenges, SMILES reconstruction failed, and we had to extract the 3D structure from different SAMPL-repository followed by minimization at an SQM level using GFN2-xTB giving us an optimised 3D structure, (ii) For BDBM36250 as it was impossible to reconstruct from SMILES, the cyclodextrins were so extracted from 4J3U PDB code that was structurally close, and manually modified with the molecular builder, then minimized at an SQM level within same procedure as before.

V. A. 2. B - DISTRIBUTION OF THE BINDING FREE ENERGY IN THE MODEL

In Figure 35 and Figure 36, the distribution of the binding affinity extracted from the BindingDB is analysed into two different graphics. In the first one (Figure 35), we saw how the binding free energy of our samples are distributed in the database:

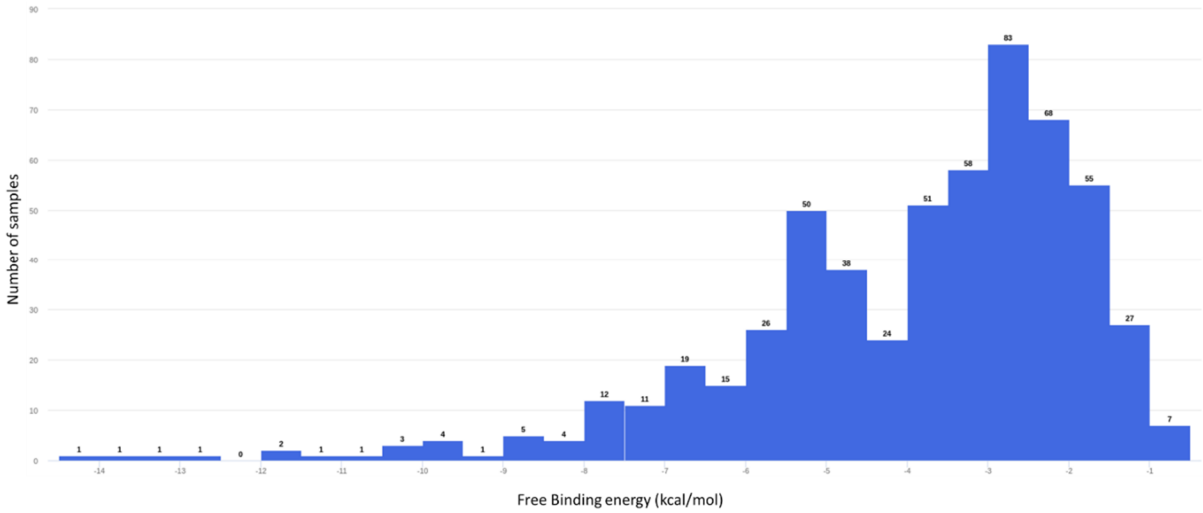


Figure 35: Barplot of the binding free energy distribution of the different systems extracted from the BindingDB

In contrast, Figure 36 shows the distribution of the energy along with all the considered systems. As the system does not have the same number of samples, they overlap. For example, we can see that the BDBM11 host has a much larger dataset than the others.

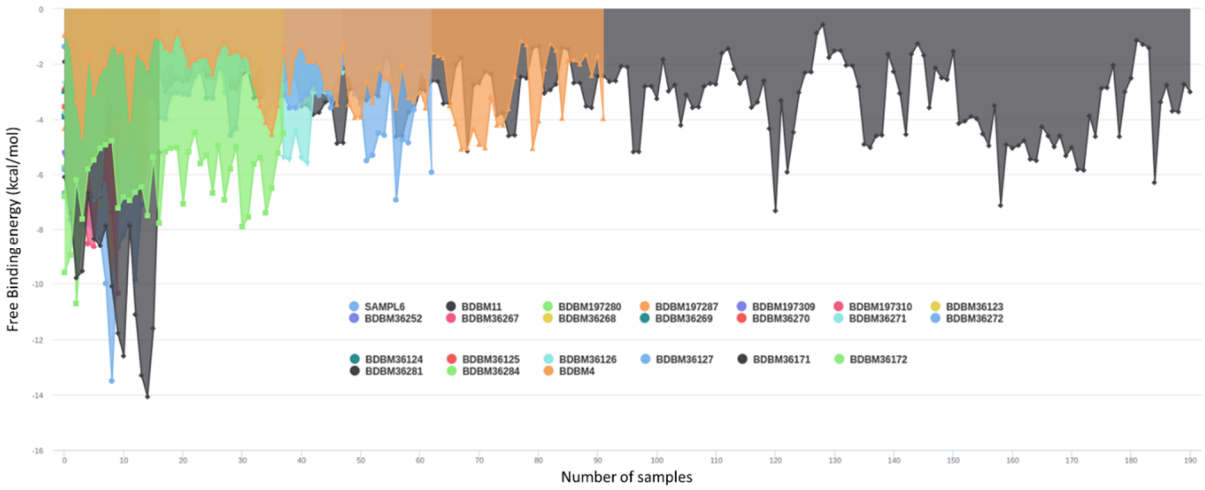


Figure 36: Density plot of the distributed binding free energy of the different systems extracted from the BindingDB

In both figures, we can see that the number of samples is much more important below < 7 kcal/mol.

The repartition of the samples for all the systems is the following:

Table 5: Composition of the data table for the binding free energy ML prediction model

BindingDB reference	Number of samples	Margins: (-kcal/mol)
SAMPL6	14	[-6.45 – 13.5]
BDBM11	191	[-0.57 – -7.32]
BDBM197280	7	[-4.17 – -10.71]
BDBM197287	15	[-3.72 – -9.36]
BDBM197309	6	[-2.37 – -5.94]
BDBM197310	10	[-2.50 – -10.34]
BDBM36123	4	[-2.78 – -2.94]
BDBM36124	4	[-2.84 – -3.15]
BDBM36125	4	[-2.77 – -3.03]
BDBM36126	49	[-1.43 – -5.64]
BDBM36127	63	[-1.36 – -6.93]
BDBM36171	4	[-4.69 – -5.76]
BDBM36172	4	[-4.26 – -5.02]
BDBM36250	2	[-3.30 – -4.34]
BDBM36252	2	[-5.10 – -5.29]
BDBM36267	8	[-3.53 – -6.19]
BDBM36268	8	[-3.81 – -6.35]
BDBM36269	8	[-3.90 – -6.85]
BDBM36270	8	[-3.54 – -6.76]

BDBM36271	8	[-3.80 – -6.98]
BDBM36272	4	[-5.67 – -6.76]
BDBM36281	17	[-5.20 – -14.08]
BDBM36284	38	[-4.77 – -9.59]
BDBM4	92	[-0.79 – -5.09]

V. A. 2. C - DESCRIPTORS CALCULATED WITH CORINA

Both host and Guest structures are passed into the CORINA web platform to compute 200 2D and 3D molecular descriptors for each host and guest. The descriptors of the Guest-dataset and the Host-dataset are reduced separately using the R software with different approaches: (i) deleting the descriptors that have a near-zero variance using Caret package; (ii) deleting the most correlated descriptors using Caret package; (iii) using PCA to combine descriptors that explain the most the variability. Host-dataset and Guest-dataset are merged to form the final dataset where each line corresponds to a guest interacting with a specific host. For creating data partition, for each numeric y, the sample is split into groups sections based on percentiles, and sampling is done within these subgroups. Several random seed numbers are used to generate a bunch of different partitions.

The description of the calculated descriptors before dimensional reduction is presented in the following tables (Table 6, Table 7, and Table 8). The descriptors can be separated into three different tables: (i) global molecular descriptors table, (ii) the atom pair properties for 2D autocorrelation, and (iii) the atom pair properties for 3D autocorrelation.

Table 6 represents the global molecular descriptors. It is composed of 16 different numerical molecular descriptors. These descriptors are very common and represent a chemical structure by a structural, chemical, or physicochemical feature or property of the molecule expressed by a numerical value. They are derived either from the growth formula, the 2D structure, or the 3D structure.

Table 6: Global molecular descriptors[145]

Descriptors	Description	Abbreviation
Molecular weight	Molecular weight in [g/mol] derived from the gross formula	Weight
Number of hydrogen bonding acceptors	Number of hydrogen bonding acceptors derived from the sum of nitrogen and oxygen atoms in the molecule	HAcc
Number of oxygen atom-based hydrogen bonding acceptors	Number of hydrogen bonding acceptors derived from the sum of oxygen atoms only in the molecule	HAcc_O
Number of nitrogen atom-based hydrogen bonding acceptors	Number of hydrogen bonding acceptors derived from the sum of nitrogen atoms only in the molecule	HAcc_N
Number of hydrogen bonding donors	Number of hydrogen bonding donors derived from the sum of N-H and O-H groups in the molecule	HDon_O
Number of oxygen atom-based hydrogen bonding donors	Number of hydrogen bonding donors derived from the sum of O-H groups only in the molecule	HDon
Number of nitrogen atom-based hydrogen bonding donors	Number of hydrogen bonding donors derived from the sum of N-H groups only in the molecule	HDon_N
Octanol/water partition coefficient (logP)	Octanol/water partition coefficient in [log units] of the molecule following the XlogP	XlogP
Topological polar surface area	Topological polar surface area in [Å2] of the molecule derived from polar 2D fragments	TPSA
Number of rotatable bonds	Number of open-chain, single rotatable bonds	NRotBond
Number of Rule of 5 violations	Number of violations of the Lipinski's rule of 5 (Weight > 500, XlogP > 5, HDon > 5, HAcc > 10)	NViolationsRo5
Number of extended Rule of 5 violations	Number of violations of the extended Lipinski's rule of 5 (additional rule: number of rotatable bonds > 10)	NViolationsExtRo5
Number of atoms	Number of all atoms in the molecule (including hydrogen atoms)	NAtoms
Number of tetrahedral stereo centres	Number of tetrahedral chiral centres in the molecule	NStereo
Molecular complexity	Molecular complexity according to the approach by J. Hendrickson	Complexity
Ring complexity	Ring complexity according to the approach by J. Gasteiger and C. Jochum	RComplexity

In Table 7, we have the Topological or 2D Property-Weighted Autocorrelation descriptors. These descriptors use the 2D structure of a molecule (i.e., the molecular graph as expressed by

the connection table) and atom pair properties as a basis to derive vectorial molecular descriptors. The products of atom pair properties are summed up for a certain topological distance: the number of bonds on the shortest path between two atoms. Thus, for each topological distance, a single value is obtained: one coefficient of the resulting 2D autocorrelation vector.

Table 7: Atom pair properties for 2D autocorrelation[145]

Atom Pair Property	Description	Abbreviation
Identity	2D autocorrelation weighted by atom identities, i.e., "1" for an atom	2DACorr_Ident
σ charge	2D autocorrelation weighted by σ atom charges	2DACorr_SigChg
π charge	2D autocorrelation weighted by π atom charges	2DACorr_PiChg
Total charge	2D autocorrelation weighted by total atom charges (sum of σ and π charges)	2DACorr_TotChg
σ electronegativity	2D autocorrelation weighted by σ atom electronegativities	2DACorr_SigEN
π electronegativity	2D autocorrelation weighted by π atom electronegativities	2DACorr_PiEN
Lone pair electronegativity	2D autocorrelation weighted by lone pair electronegativities	2DACorr_LpEN
Effective polarizability	2D autocorrelation weighted by effective atom polarizabilities	2DACorr_Polariz

The 2D autocorrelation vectors are calculated using the following parameters that have proven useful for most modelling:

- Hydrogen atoms are ignored, and only non-hydrogen (heavy) atoms are taken into account.
- The minimum topological distance is taken into account is 0, i.e., the first coefficient (element) of the 2D autocorrelation vector is the sum of the products of the properties of the atom pairs of each atom with itself.
- The maximum topological distance taken into account is 10, i.e., there can be up to ten intermediate bonds between a pair of atoms.
- Therefore, the dimensionality of a single 2D autocorrelation vector is "11".

In total, eight eleven-dimensional 2D autocorrelation vectors using eight different atom pair properties are calculated for a molecule.

A similar procedure can be considered for Table 8 and the spatial or 3D Property-Weighted Autocorrelation. The spatial descriptors are calculated using the 3D structure of a molecule (i.e., the Cartesian atomic coordinates) and atom pair properties as a basis to derive vectorial molecular descriptors. The products of atom pair properties are summed up for certain 3D distance intervals. Thus, for each 3D distance interval, a single value is obtained: one coefficient of the resulting 3D autocorrelation vector.

Table 8: Atom pair properties for 3D autocorrelation[145]

Descriptors	Description	Abbreviation
Identity	3D autocorrelation weighted by atom identities, i.e., "1" for an atom	3DACorr_Ident
σ Charge	3D autocorrelation weighted by σ atom charges	3DACorr_SigChg
π Charge	3D autocorrelation weighted by π atom charges	3DACorr_PiChg
Total charge	3D autocorrelation weighted by total atom charges (sum of σ and π charges)	3DACorr_TotChg
σ Electronegativity	3D autocorrelation weighted by σ atom electronegativities	3DACorr_SigEN
π Electronegativity	3D autocorrelation weighted by π atom electronegativities	3DACorr_PiEN
Lone pair electronegativity	3D autocorrelation weighted by lone pair electronegativities	3DACorr_LpEN
Effective polarizability	3D autocorrelation weighted by effective atom polarizabilities	3DACorr_Polariz

The 3D autocorrelation vectors are calculated using the following parameters that have proven useful for most modelling:

- Hydrogen atoms are ignored, and only non-hydrogen (heavy) atoms are considered.
- The minimum spatial distance taken into account is 1 Å, i.e., the first coefficient (element) of the 3D autocorrelation vector is the sum of the products of the properties of the pairs of atoms that are 1 to 2 Å away from each other.
- The maximum spatial distance considered is 13 Å, i.e., up to ten intermediate bonds between a pair of atoms.
- The number of equal 3D distance intervals is set to 12, i.e., the first interval sums the property products of the atom pairs from 1 to 2 Å. Therefore, the dimensionality of a single 3D autocorrelation vector is "12".

A total of eight 12-dimensional 3D autocorrelation vectors using eight different atom pair properties are calculated for a molecule.

Before calculating the global descriptors, hydrogens have to be added to all of the molecules. As the CORINA web platform uses 3D molecules, a strong focus has to be done on the constructions of the 3D molecules. For that reason, we used the GFN2-xTB SQM methods for geometrical optimisation on all the structures after their construction from the 2D SMILES code or after their extraction from the SAMPLs GitHub repository.

V. B - OPTIMISATION OF THE ML ALGORITHM

Several ML algorithms were used for this work and were described in chapter 2 (Methods). By modifying the tuning parameters of the four ML algorithms, multiple different models were able to be generated. The variation in the tuning parameters is presented in the following table:

Table 9: Variation of the tuning parameters of the NNET (in blue), the SVM (in green), the RF (in red), the Knn (in yellow).

Regression method used	Tuning Parameters	Explanation	Values	Final number of model generated
Neural network	Size	Number of units in the hidden layer.	From 1 to 10 by 1	$> 10^5$ per partition
	Decay	Parameter for weight decay	From 0.01 to 2 by 0.01	
	maxit	Maximum number of iterations.	500	
Support Vector Machines with Polynomial Kernel	Degree	Parameter needed for kernel of type Polynomial	From 1 to 5 by 1	$> 10^3$ per partition
	Scale	A logical vector indicating the variables to be scaled	0.01 and 0.1	
	Cost (c)	Cost of constraints violation : it is the constant of the regularization term in the Lagrange formulation.	From 10^{-4} to 10^4 by 10^1.	
Random Forest	mtry	Number of variables randomly sampled as candidates at each split	From 1 to 15, by 1	$> 10^3$ per partition
	ntree	Number of trees to grow	From 1000 to 2500, by 500	
k-Nearest Neighbour Classification	k	number of neighbours considered.	From 1 to 20 by 1	~ 600 per partition

In total, more than a million models are generated using the variation of the tuning parameters, the 3x repeated-cross validation using 10-fold, and the ten different data partition we initially created with ten different random seed numbers. These millions of models will be ranked using the respective values of the RMSE, MAE, and R^2 in the training-set. The best-ranked model of each regression model will be selected and compared by (i) their prediction of the test-set and (ii) the descriptors they used the most for the prediction.

The clustering methods, the thermodynamically based approach, and the knowledge-based methods have been used on multiple systems belonging to different chemical families. These systems and the associated results will be presented in the next chapters related to the application of the HG-DYNAusor platform.

V. C - DIMENSIONAL REDUCTION PROCEDURE:

The dimensional reduction procedure has been applied for all the systems studied with the knowledge-based methods and for constructing the global model, applied apart for the guest and the host. The process will only show once for practical reasons as the principle is the same for all the models.

V. C. 1 - PRE-PROCESSING:

To analyze the physical-chemical space, all the variables are analyzed together using non-supervised ML algorithms. The outcome of the PCA on the raw-dataset is shown in Figure 37:

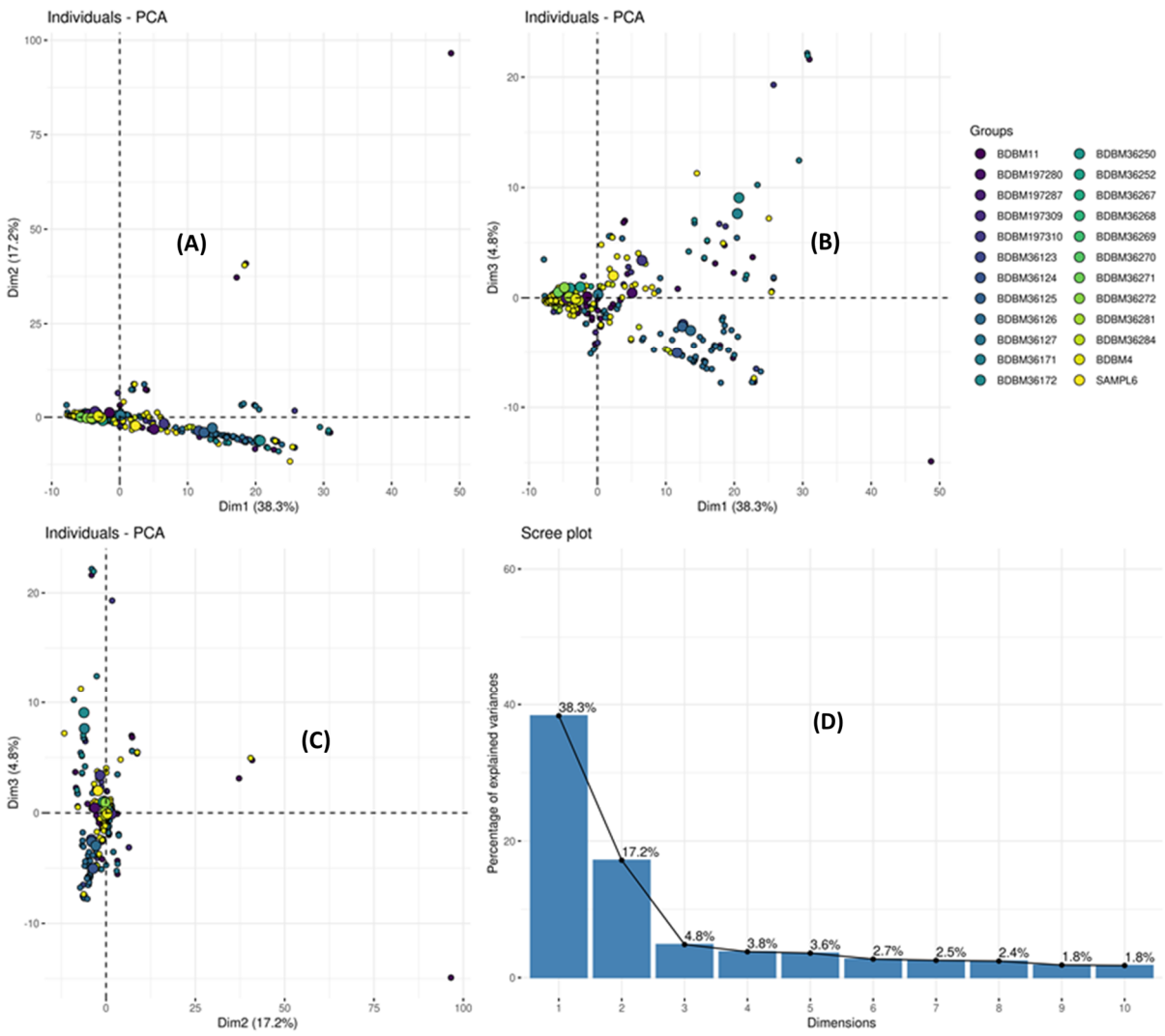


Figure 37: PCA of the Guest chemical space before any dimensional reduction described by a set of molecular descriptors generated with CORINA. In (A), the space formed by the combination of PC1 and PC2 explains respectively 38.3% and 17.2% of the variability. In (B), the space formed by the combination of PC1 and PC3 explains respectively 38.3% and 4.8% of the variability. In (C), the space formed by the combination of PC2 and PC3 explains respectively 17.2% and 4.8% of the variability. In (D), the scree-plot represents the variability of all the principal components of the analysis. The molecules

are coloured by the system there are supposed to interact with and, their size is a function of their binding free energy.

In Figure 37, the initial dataset explains in the three first components 60.3% of the total variability of the dataset. This value is acceptable for a raw dataset since some variables bring only noise at this step (imagine, for example, a variable that would accurately describe 10% of the dataset and give opposite information to the remaining 90%). One of the first pre-process we can do is a dimensional reduction using a zero-variance or near-zero-variance (*nzv*) protocol: a variable that presents almost no variability in the dataset is a useless variable that is incapable of giving interesting information for the discrimination of one molecule between one other. In most cases, this may cause the model to crash or the fit to be unstable. Similarly, predictors may have only a handful of unique values that occur at very low frequencies.

To identify these predictors, we are using the *caret* package from R. Two main metrics are used for the identification of the *nzv*-variables:

- The **frequency of the most prevalent** (*freqRatio*) represent the value over the second most frequent value (called the "frequency ratio'')
- The **percent of unique values** represent the number of unique values divided by the total number of samples

As an example, for our dataset, the application of the nzv-algorithms for the raw-data gives the following Table 10, where only the five first variables are shown:

Table 10: Output of the Near-Zero-Variance (*nzv*) protocol on the dataset:

Variables:	freqRatio	percentUnique	zeroVar	nzv
NViolationsRo5	27.35	0.35	false	true
X2DACorr_Ident_9	20.82	6.50	false	true
X2DACorr_Ident_11	33.38	4.92	false	true
X2DACorr_PiEN_11	38.42	8.08	false	true
X2DACorr_LpEN_2	130.75	3.69	false	true
…	…	…	…	…

In total, 30 descriptors were found to have non-zero variance at this step and were deleted from the dataset. The dataset presents then 170 numerical descriptors. Important to note that when you split the data set into a training set and a test set prior to prediction, the same protocol can be tested again to ensure that splitting the data into two groups has not resulted in new variables

with zero or near-zero variance. Another approach is to use dimensional reduction using correlated descriptors. The idea is that two highly correlated descriptors (99% or 95% depending on the dataset) will bring almost the same information to the models and thus should not be considered together. For our dataset, the correlated descriptors that are presenting a correlation up to 95% are deleted from the dataset, and at this step, the analysis of the chemical space of the model gives the following Figure 38, presenting at this step 142 descriptors:

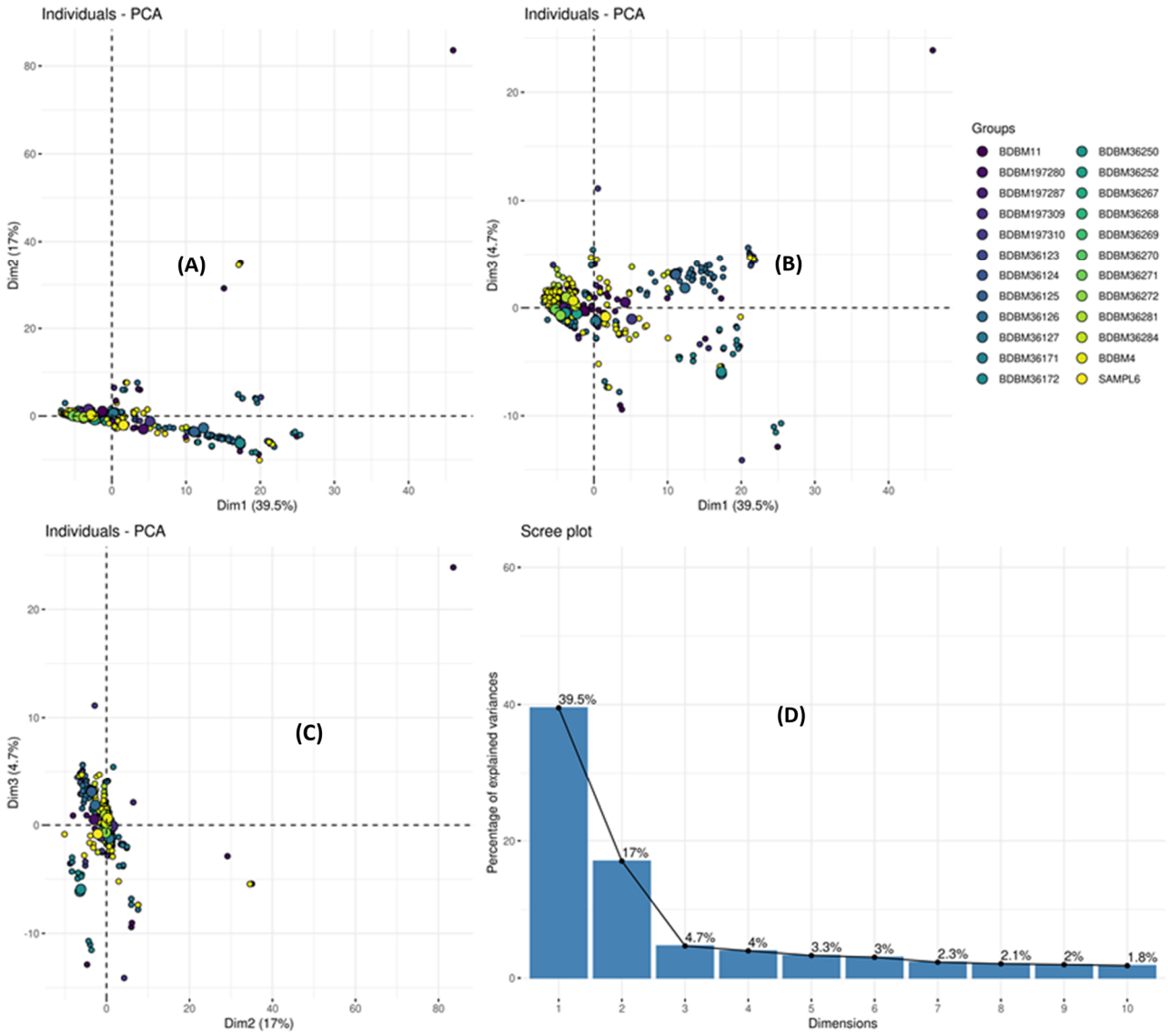


Figure 38: PCA of the Guest chemical space after a dimensional reduction using near-zero-variance and correlated approach described by a set of molecular descriptors generated with CORINA. In (A), the space formed by the combination of PC1 and PC2 explains respectively 39.5% and 17.0% of the variability. In (B), the space formed by the combination of PC1 and PC3 explains respectively 39.5% and 4.7% of the variability. In (C), the space formed by the combination of PC2 and PC3 explains respectively 17.0% and 4.7% of the variability. In (D), the scree-plot represents the variability of all the principal components of the analysis. The molecules are coloured by the system there are supposed to interact with, and their size is a function of their binding free energy.

At this point, there was little or no change in the explained variance: 61.2% compared to the previous 60.3%. It makes a lot of sense: at this point in the analysis, the deleted descriptors were uninformative descriptors that did not provide useful information about the variability of the data set, but that could disturb the prediction by the ML approach. At this step, we deleted the descriptors that are not important for the variability of the dataset using two different approaches. In contrast, now we will try to extract the descriptors that explained the most variability. For this, an unsupervised approach is used from a PCA result: a determined number of the most useful descriptors are extracted from each of the first three components to have a dataset that describes as much as possible the variability to improve the accuracy of the ML models. For this dataset, we extracted 90 descriptors (30 for each of the three first PCs), giving the following Figure 39.

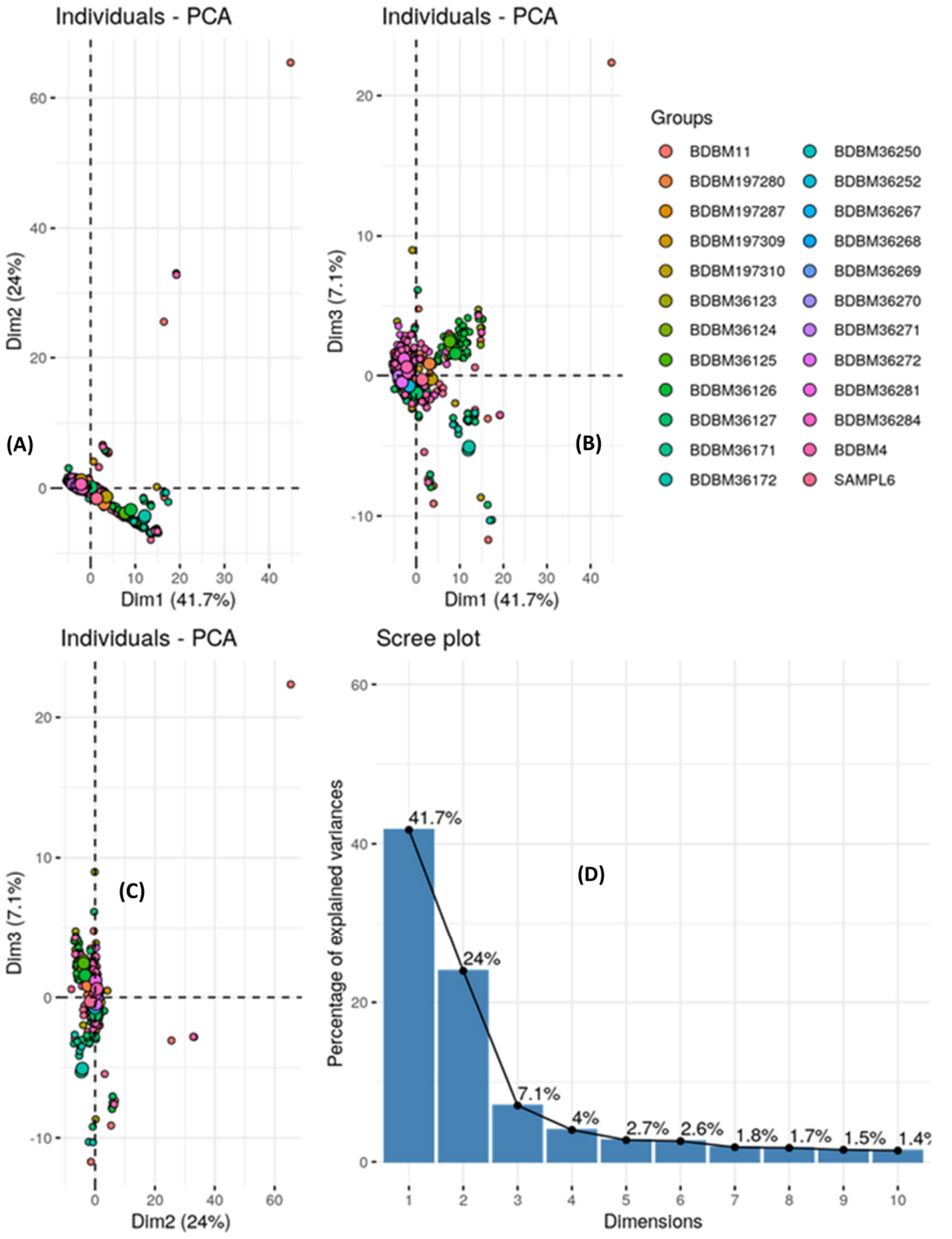


Figure 39: PCA of the Guest chemical space using a reduced set of molecular descriptors generated with CORINA. In (A), the space formed by the combination of PC1 and PC2 explains respectively 41.7% and 27% of the variability. In (B), the space formed by the combination of PC1 and PC3 explains respectively 41.7% and 7.1% of the variability. In (C), the space formed by the combination of PC2 and PC3 explains respectively 41.7% and 27% of the variability. In (D), the scree-plot represents the variability of all the principal components of the analysis. The molecules are coloured by the system there are supposed to interact with, and their size is a function of their binding free energy.

The PCA of the guest shows that the three first components explain 72.8% of the variability. The dataset for this PCA is composed of 570 ~ molecules, described by a set of 90 molecular descriptors calculated with CORINA web-platform and for which a numerical binding free energy has been extracted from the literature. In the terminology of the PCA, the dimensions are synonym with the PC(X), with X representing the considered dimension: as an example, the first dimension (Dim1) and the first component (PC1) represent the same space. As the other dimensions show a very low percentage (Figure 39D), we consider the sampling to be sufficient. As you can see, the guest model shows one outlier in the model (in the top-right of (A), (B), and (C)). This molecule can eventually lead to some computational noise in the prediction because it is far away from the other molecules in all the dimensions of the space: this molecule that is interacting with the BDM11 host is sampling its own conformational space. It is interesting to highlight that as the molecule are coloured by the system, there are interacting with, and as we said, there are molecules that interact with two different systems, thus some molecules completely overlap. In these cases, only one system is visible due to the overlap of the colours.

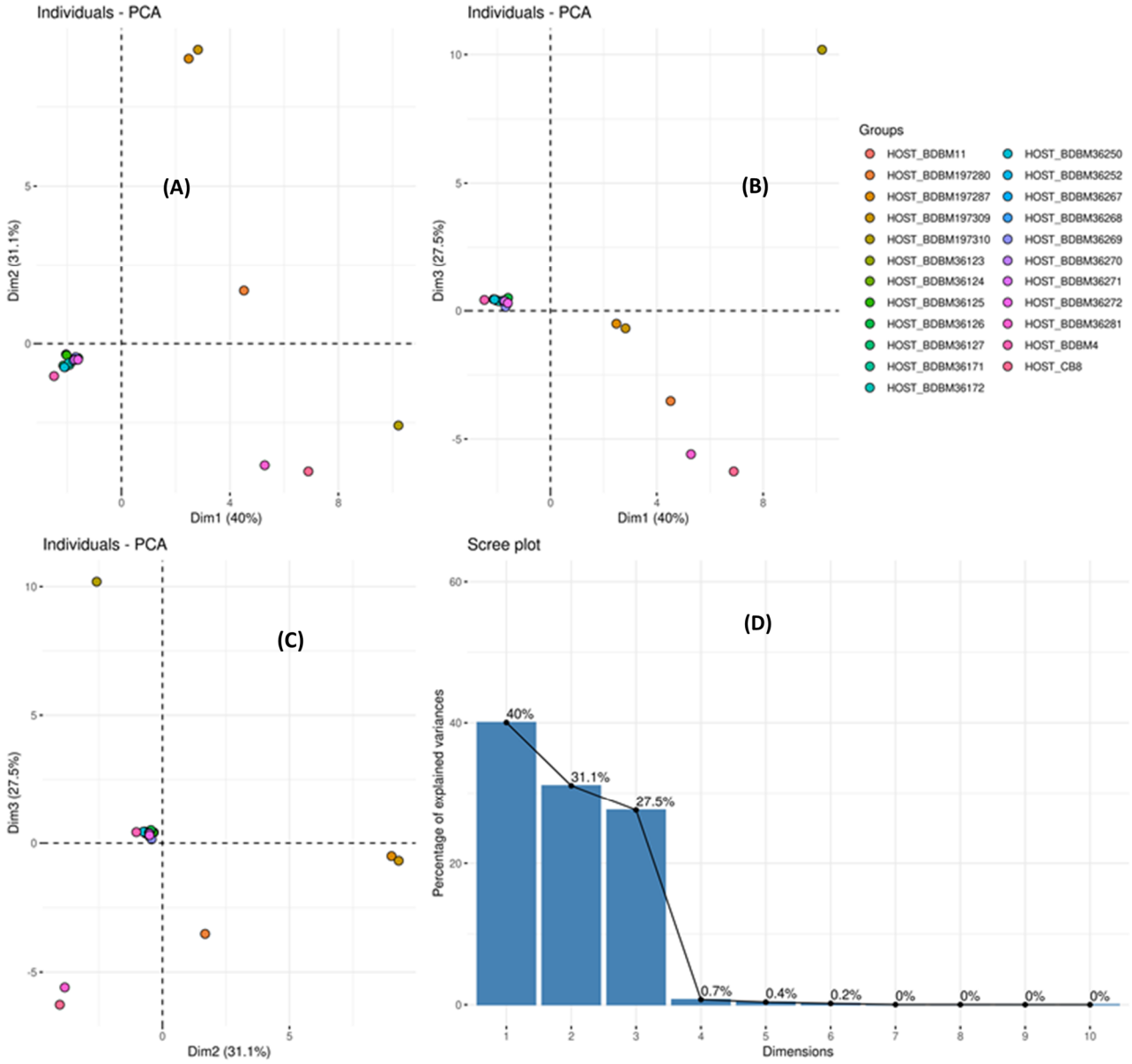


Figure 40: PCA of the Host chemical space using a reduced set of molecular descriptors generated with CORINA. In (A), the space formed by the combination of PC1 and PC2 explains respectively 40.0% and 31.1% of the variability. In (B), the space formed by the combination of PC1 and PC3 explains respectively 40.0% and 27.5% of the variability. In (C), the space formed by the combination of PC2 and PC3 explains respectively 31.1% and 27.5% of the variability. In (D), the scree-plot represents the variability of all the principal components of the analysis. The molecules are coloured by the system.

The three first components of the PCA of the host represent almost all the variability: 98.6% (Figure 40). The dataset for this PCA is composed of 24 different hosts, described by a set of 90 molecular descriptors calculated with CORINA web-platform and for which numerical binding free energies with a set of guests have been extracted from the literature. In that case, the hosts are sampling a relatively smaller space than the guest. It can be explained by the less

important diversity of the host systems compared to the guests. It is interesting to see that many hosts sample a similar conformational space, while others are quite far apart. However, it is very easy to differentiate between the different clusters because of the small number of samples. It is logical that the hosts belonging to the same chemical family's sample structurally close conformational spaces. However, due to the fact that in our method of extracting numerical variables that best explain the variability of the sample, possible that in some cases, two structurally close hosts do not sample the same conformational space because they do not have the same physicochemical properties. Unlike the previous graphic, each host corresponds to a unique colour, presenting the same size.

V. C. 4 - ANALYSIS OF THE HOST-GUEST MODEL

In Figure 41, we can see a PCA of the host-guest system composed by the two previous analyses on respectively host and guest. The dataset for the Host-Guest system is composed of a concatenation of the previous descriptors of the guest and the host to form the final dataset composed of 570 features, described by 90 variables. These variables are decomposed into 60 guest variables and 30 host variables. We wanted to weigh the analysis by giving more information about the guest than the host. The PCA only explains 62.7% of the variability after the three first PCs. Considering PC4, we are up to 70% of the variability, as we created a dataset with the host and the guest information inside, the values are acceptable for further analysis, with the idea that the inclusion of new systems will improve the quality of the dataset after several usages of the knowledge-based protocol.

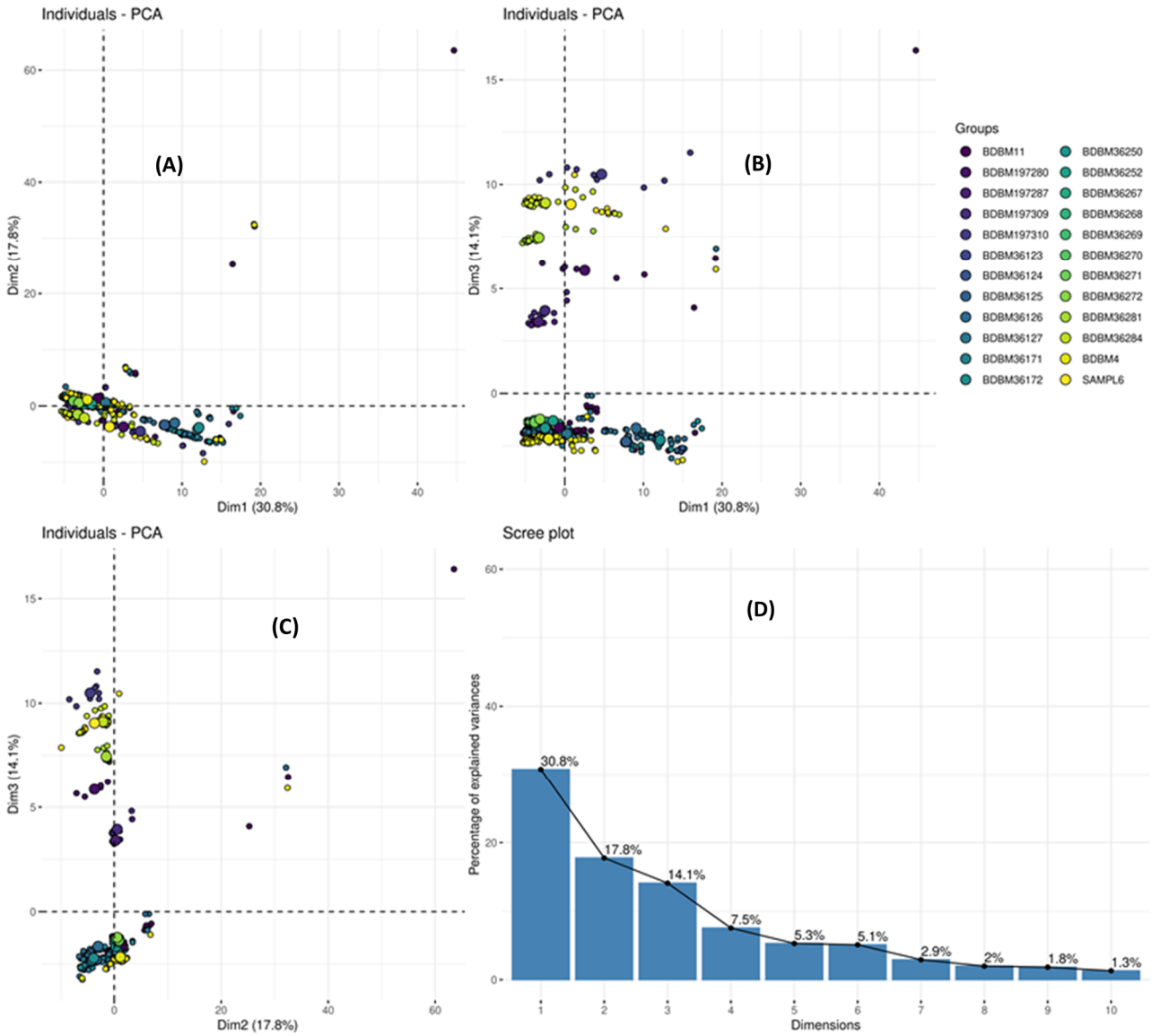


Figure 41: PCA of the Host-Guest chemical space using a reduced set of molecular descriptors generated with CORINA. In (A), the space formed by the combination of PC1 and PC2 explains respectively 30.8% and 17.8% of the variability. In (B), the space formed by the combination of PC1 and PC3 explains respectively 30.8% and 14.1% of the variability. In (C), the space formed by the combination of PC2 and PC3 explains respectively 17.8% and 14.1% of the variability. In (D), the scree-plot represents the variability of all the principal components of the analysis. The molecules are coloured by the system there are supposed to interact with, and their size is a function of their binding free energy.

An overview of the most useful variable and their respective contributions on each axis can be shown in the following Figure 42, Figure 43, and Figure 44.

You can see in these figures that the three first dimensions of the PCA are not described by the same descriptors-type. For the Dim-1 (or PC1), It is a majority of 2D and 3D-electronegativity descriptors, mainly related to the σ atoms.

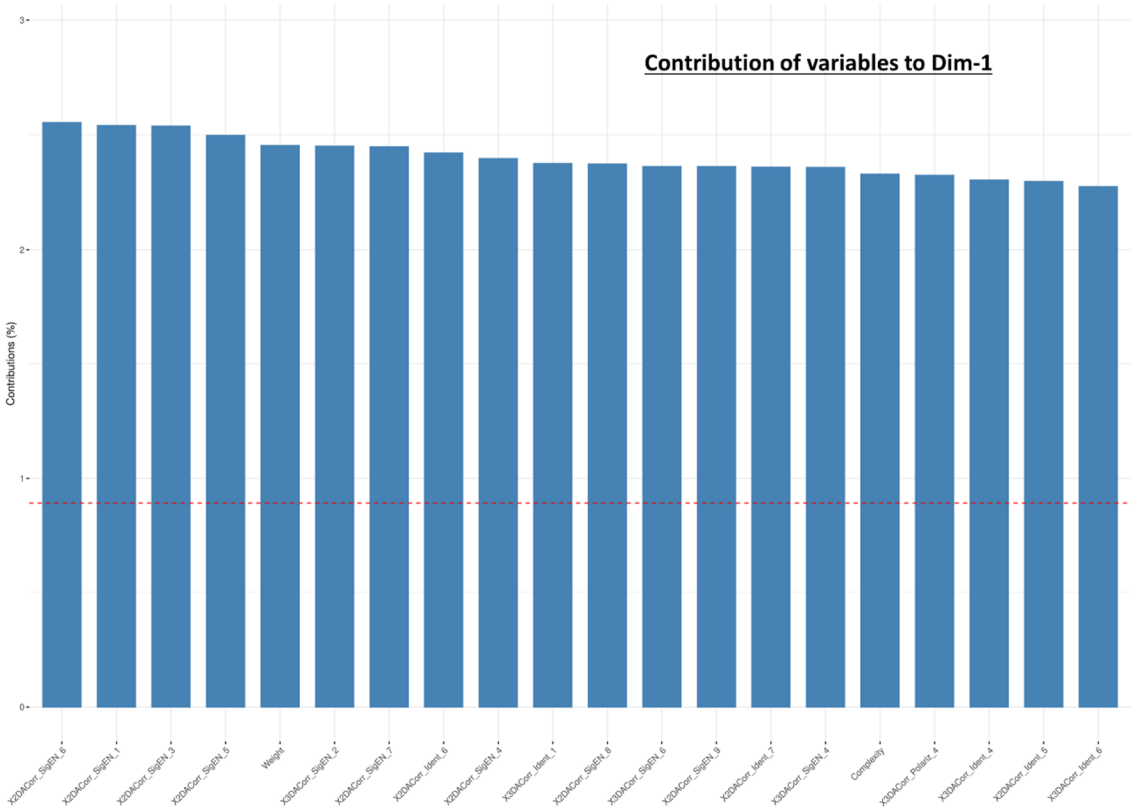


Figure 42: Top-contributions of variables to Dimension 1

For the Dim-2 (Figure 43), It is a majority of 2D and 3D-electronegativity descriptors, mainly related to the π atoms.

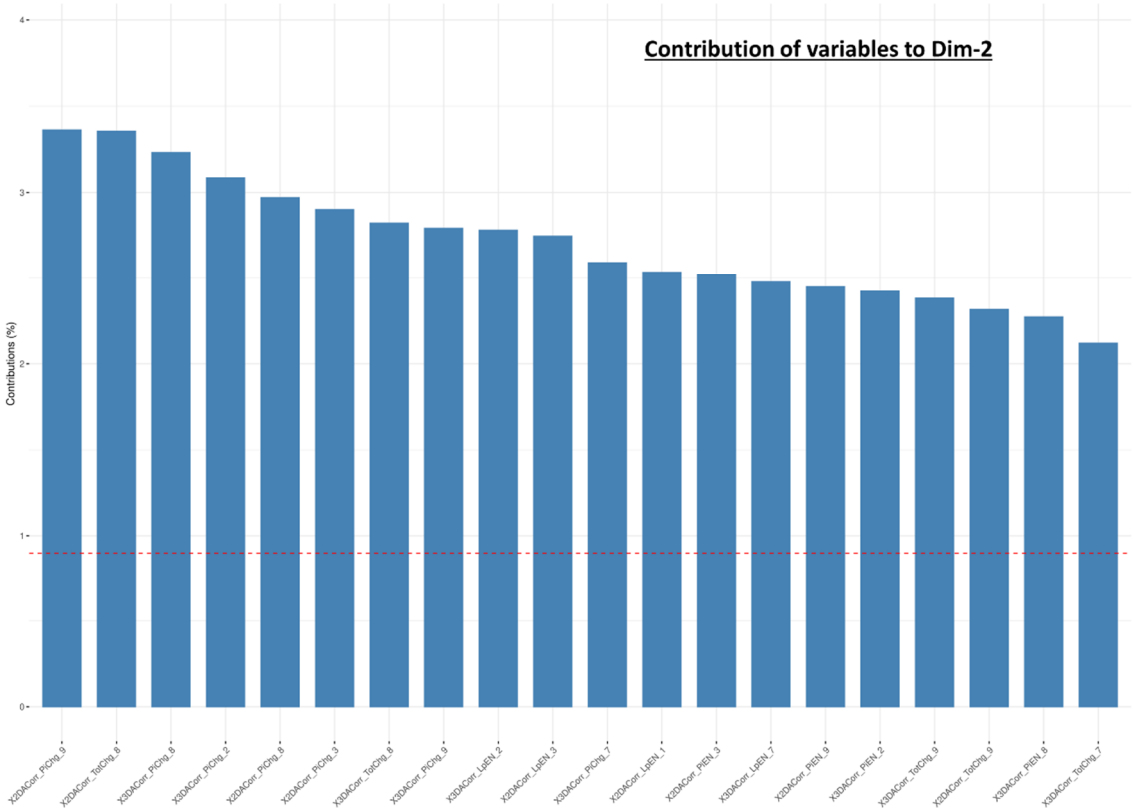


Figure 43: Top-contributions of variables to Dimension 2

For the Dim-3 (Figure 44), It is a majority of 2D and 3D-electronegativity descriptors, describing both total charges, π and, σ electrons atoms.

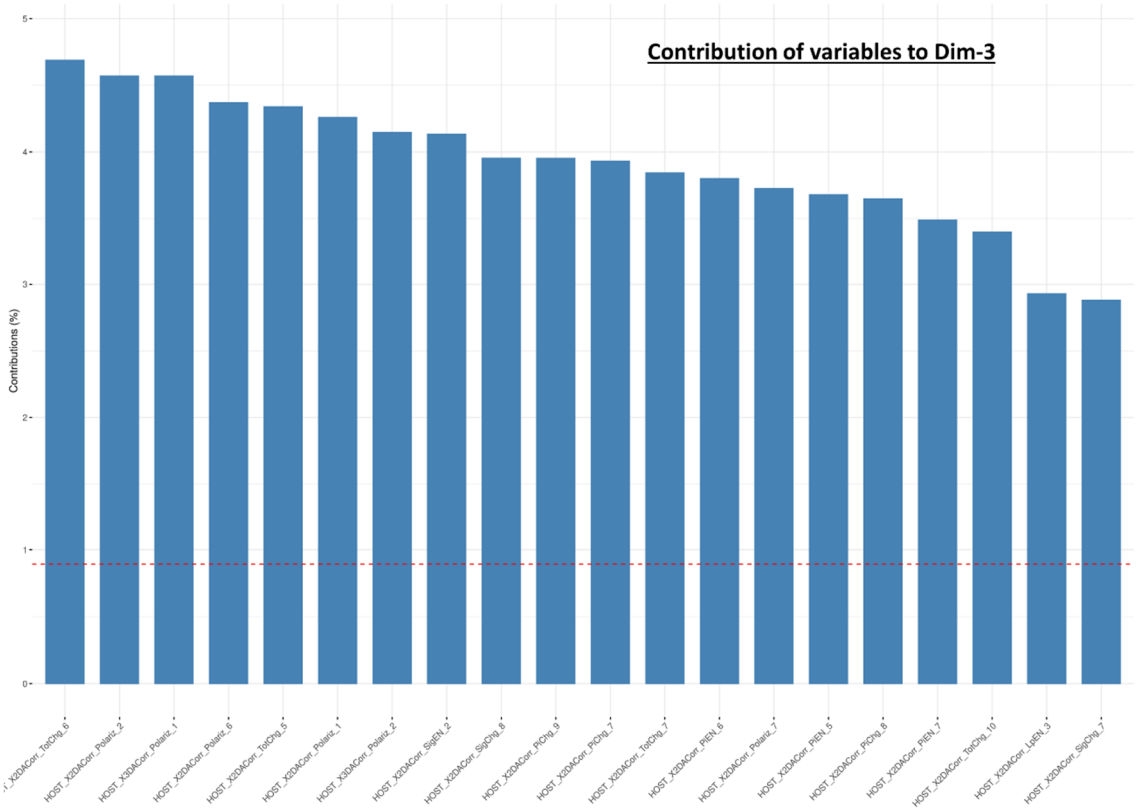


Figure 44: Top-contributions of variables to Dimension 3.

VI - HG-DYNAUSOR PLATFORM: FUTURE DIRECTIONS

VI. A - CLUSTERING METHODS

The clustering method using a non-supervised ML approach is difficult to automatize for any arbitrary system. At present, this method depends on general molecular descriptors that are easy to generate but do not describe the configurational variability of the molecules or how they interact. In order to improve accuracy, we could use additional molecular descriptors that depend on and describe the structure of the considered host (specific angles, distances, torsional, or dihedrals). For that reason, the automation of the clustering protocol was not possible in the time frame of the thesis. All the studied system has been clusterised manually depending on the chemical specificity of each other.

VI. B - THERMODYNAMIC BASED APPROACH

The thermodynamic-based approach is limited by the complexity of calculating the numerical values of the enthalpic and entropic terms. We knew and encountered several problems linked to the solvation modes and the protonation states.

We found that when the system is optimised at an SQM level, followed by a docking procedure, then the thermodynamic calculation on the resulting structure can lead to unrealistic results with a deformed structure. For that reason, sampling with explicit solvent is necessary, and the MD will be used as a sampling method for binding free energy prediction. We also encountered a problem concerning the thermodynamic prediction of hosts or guests that are multiple charged (positively or negatively), for which the binding free energy is mostly overestimated. This problem seems difficult to address with implicit solvent models. The use of a combined approach, where explicit solvent molecules are included in the first layers of solvation, should be investigated.

VI. C - KNOWLEDGE-BASED APPROACH

Within the limited timeframe we had during the thesis, we decided to use the BindingDB as a reference database for the ML protocol. But while the data coming from the SAMPL challenge are well standardized (the activity measurement is done in a very similar way by the same team for all the experiments), most of the activity data coming from the bindingDB originates in the scientific literature, and the experimental details are heterogeneous and often insufficiently described. There exist multiple limitations to the knowledge-based methods that are linked to

the methods and the data. As we are creating some logical function that learns from the molecular descriptors we generated with CORINA, but if the structure of the guest and the host used for the learning phase is completely different from the predicted ones, the prediction will be out of the scope of the model. Before any prediction, we should control that the space of the model overlaps with the space of the predicted molecules, thus avoiding making predictions when outside the scope of the model. Compared to other methods, the knowledge-based method is extremely fast. The moment the models are ready to be used, it takes only a few minutes to predict the binding free energy of a new guest inside a host.

In the next chapters, some systems on which the platform have been used will be presented.

4

APPLICATION OF THE HG-DYNAUSOR PLATFORM

I - INVESTIGATED SYSTEMS:

I. A - GIBB DEEP CAVITY CAVITAND (GDCC)

The Gibbs cavitand is an octa-acid macrocycle that we studied in the context of our participation in the SAMPL7 and SAMPL8 challenges. These macrocycles present a highly hydrophobic cavity and two preferential positions for chemical variations: endo and exo. The variations around these positions formed several different macrocycles investigated in this chapter. The initial structure (Figure 45) has very low solubility in water.

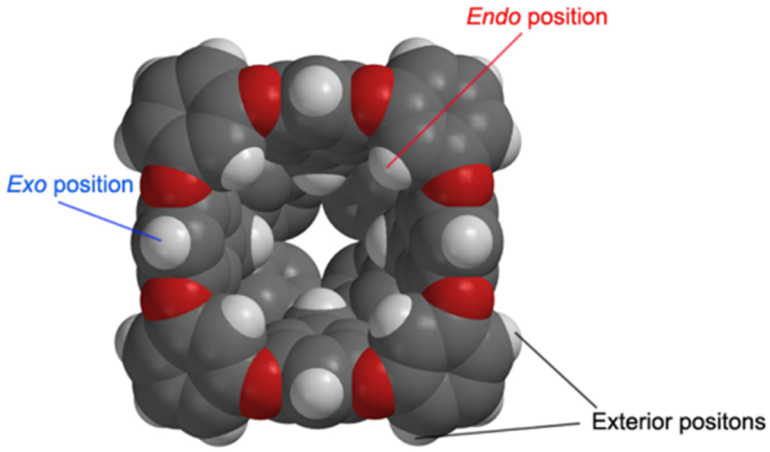


Figure 45: Presentation of the structure of the initial macrocycle[146]

The solubility issues were solved by the addition of several carboxylic acids (Figure 46), leading to the formation of the first GDCC-macrocycle: the octa-acid (OA). In addition, this macrocycle has a hydrophobic edge around the entrance to its inner hydrophobic pocket.

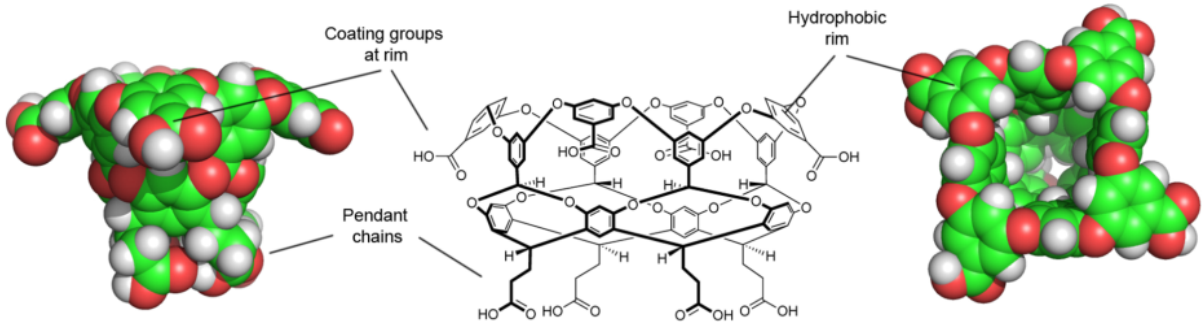


Figure 46: Presentation of the octa-acid host system: the exo position is not modified, but height carboxylic groups have been added to the structure: 4 at the end of the pendant chains and four at the external part of the endo position, linked to the hydrophobic rim[146]

From that octa-acid host, several molecular changes led to the formation of other macrocycles that have been studied in the thesis along the SAMPL7 and SAMPL8 challenges: the exo-Octa-Acid (exo-OA) with the carboxylic groups in the exo-position, the Tetra-endoMethyl Octa-

Acid (TeMOA) with an addition of a methyl group in the endo-position, and the Tetra-endoEthyl Octa-Acid (TeeTOA) with an addition of an ethyl group in the endo-position. These macrocycles are presented in the following Figure 47:

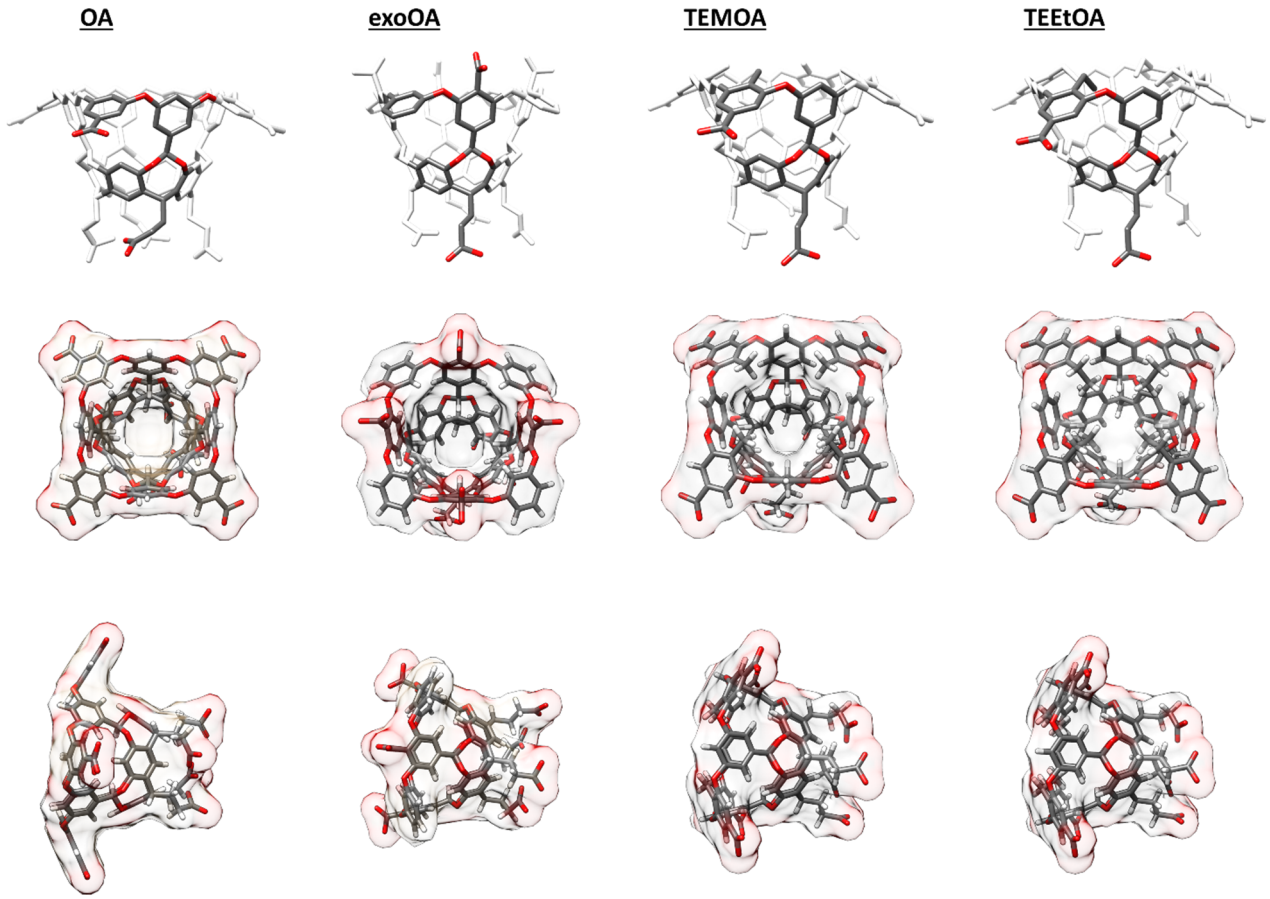


Figure 47: Presentation of the Gibbs cavitand used in the thesis: From the left to the right: the OA, the exoOA, the TEMOA, and TEETOA

I. B - CUCURBITURIL CB[8]

The cucubirt[n]uril (CB[n]) was investigated in the SAMPL8 challenge. The challenge focuses on the binding of CB[8] to nine guests, which are drugs of abuse (morphine, cocaine, hydromorphone…). The macrocycle was named cucurbituril due to its resemblance to a pumpkin. In terms of structure, the cucurbit[n]uril consists of a glycoluril unit bound together by two methylene bridges for each of them. The numbers of glycoluril units can differ, leading to several different cucurbit[n]urils differing by their cavity size (Figure 48).

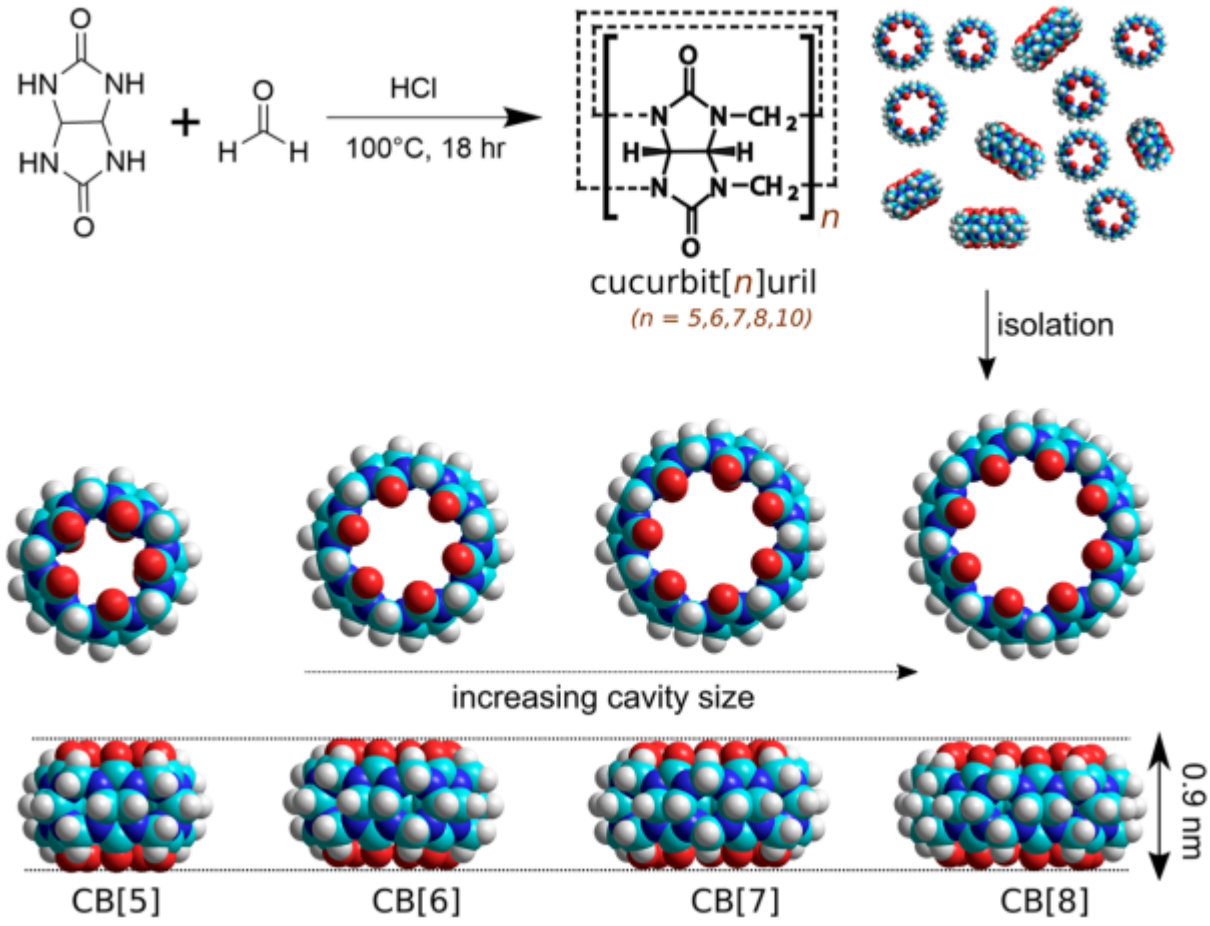


Figure 48: Geometries of the cucurbit[n]urils family[147]

The first member of this family was the CB[6] macrocycle (containing six units of glycoluril bound together with 12 methylene bridges). As the number of glycoluril units increases, the cavity size increases, allowing the new macrocycle to bind larger and larger guests inside their cavity. As the cavity size increase, the properties are also modified, especially the electrostatic potential, which becomes much more negative in the cavity from 8 units of glycoluril (Figure 49).

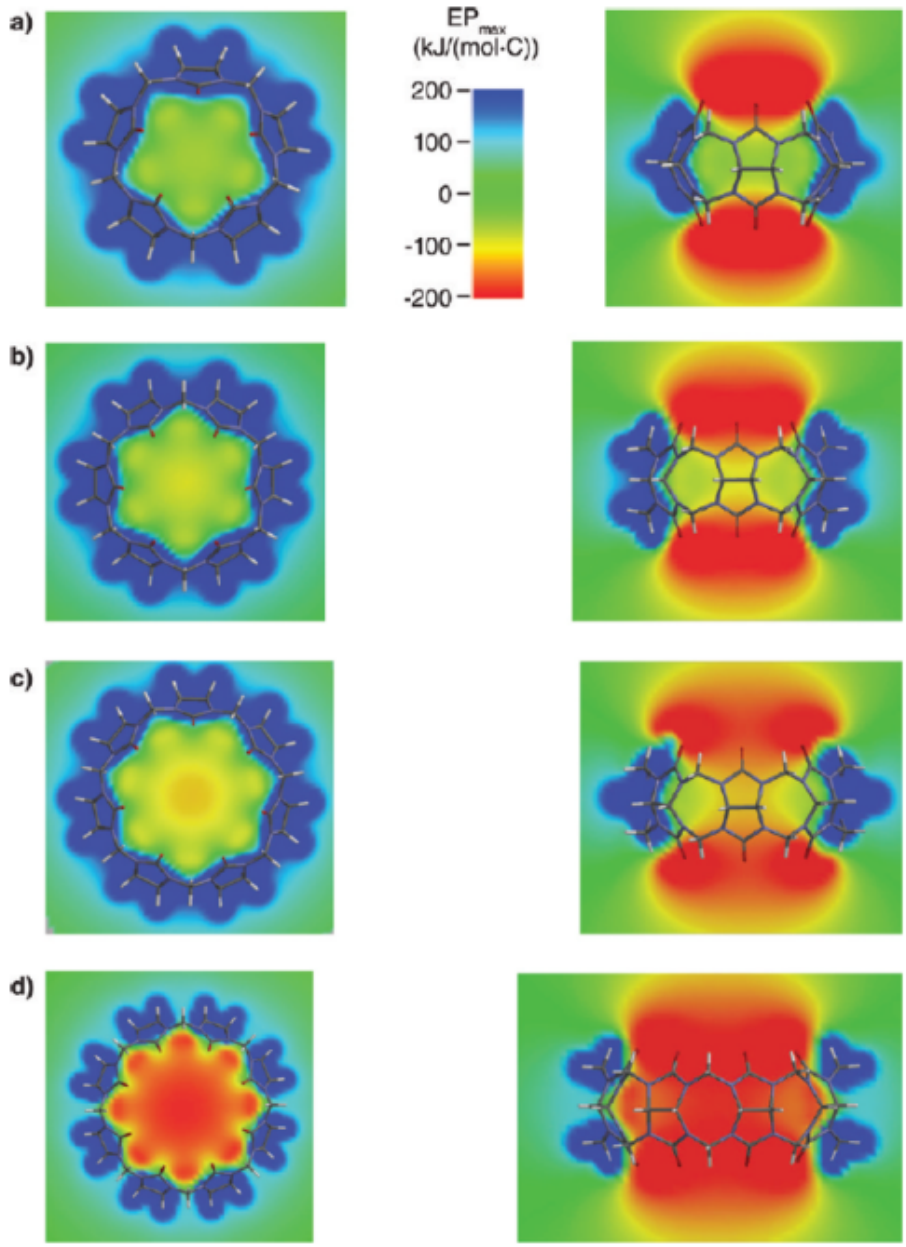


Figure 49: Calculated Electrostatic potential for several cucurbit[n]uril: CB[5] (a), CB[6] (b), CB[7] (c) and CB[8] (d)[147]

As shown, the glycoluril units' alignment results in a hydrophobic cavity with carbonyl-lined portals. The presence of the carbonyls makes the portal attractive for cation binding through the ion-dipole effect. Though the electrostatic potential is negative in all cases, the increase of cavity size increases the electrostatic potential. Unlike to rim, the inner part of the CB[n] family is remarkably hydrophobic and leads to a preferential encapsulation of hydrophobic compounds. From all the CB[n], the CB[7] is one of the most used cucurbiturils due to its capability of binding a diverse set of molecular structures. CB[8] is also well known and well used for biological application and drug encapsulation due to its large cavity size. CB[8] has been used for the SAMPL6 and SAMPL8 challenges.

An overview of the system is shown in Figure 50:

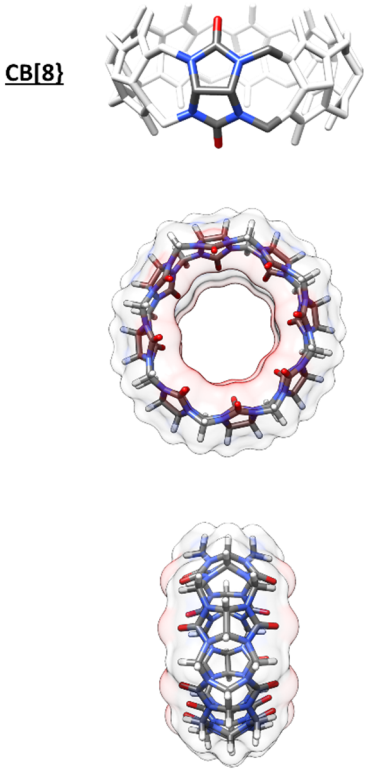


Figure 50: Presentation of the CB[8] structure used in the SAMPL8 challenge

The cavity volume of the CB[8] cavitand is around 480 $Å^3$, which is almost two times larger than the CB[7], but its binding capabilities are in many ways similar to the others CB[n] macrocycles. In the literature, CB[8] is described to have very strong affinities for bulky amphiphilic positively charged guests, but some encapsulation of very large macrocyclic guests is also described.[147]

Concerning the structural analysis of the CB[8] system, some NMR analysis suggests that the larger width of the cavity allows for a structural change compared to the other smaller CB[n] systems, and the macrocycle can present a U-shaped geometry that could allow some guests to be entirely encapsulated. While almost all the CB[n] macrocycles only allow only one guest inside the cavity, it appears that CB[8] system, due to its flexibility and larger cavity size, may bind two guests.

I. C - TRIMERTRIP

For the SAMPL7 challenge, the Isaacs group contributed a new host-system derived from the Cucurbituril macrocycle with associated binding data[148]. This CB-like clip is codenamed TrimerTrip in the challenge. In general, CB[n] are composed of *n* glycoluril unit linked by 2*n* methylene bridges. In this case, the TrimerTrip system is an acyclic receptor featuring a central glycoluril oligomer that is capped by aromatic sidewalls. Two different trimertrips have been studied in the thesis: the one extracted from the SAMPL3[39] challenge and the one extracted from the SAMPL7 challenge (Figure 51):

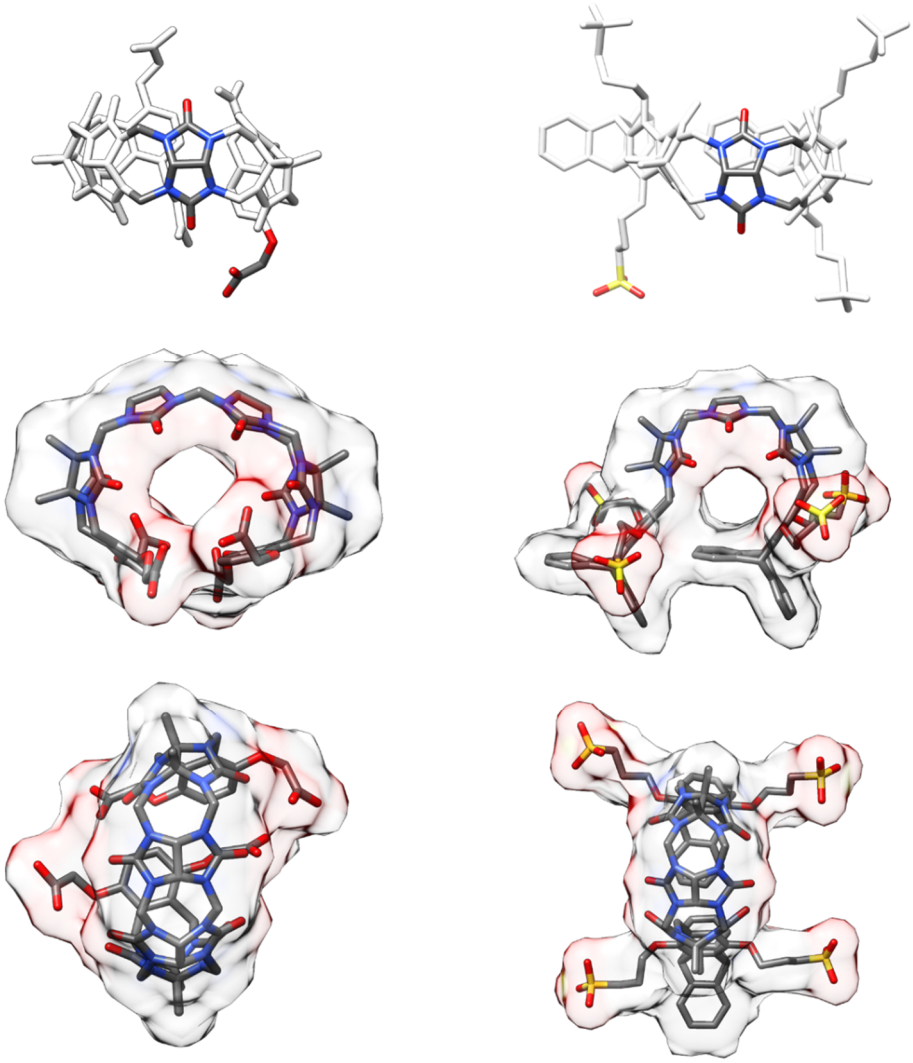


Figure 51: Presentation of the Trimertrip systems used in the SAMPL3 challenge (left) and in the SAMPL7 challenge (right)

The Trimertrip used for the SAMPL3 challenge features a central glycoluril tetramer, with o-xylylene sidewalls and carboxylate used to enhance water solubility. Due to their acyclic nature, these host systems are much more flexible than their non-cyclic counterparts.

I. D - CALIX[4]-PYRROLE

A molecular structure derived from calix[4]pyrrole was provided by one of the partners of the NOAH project, the ICIQ (Institut Català d'Investigació Química) to be studied.

Over the past few decades, calixpyrroles, particularly calix[4]-pyrroles, have become important members of supramolecular chemistry. The chemical structure of the calix[4]pyrrole is shown in Figure 52:

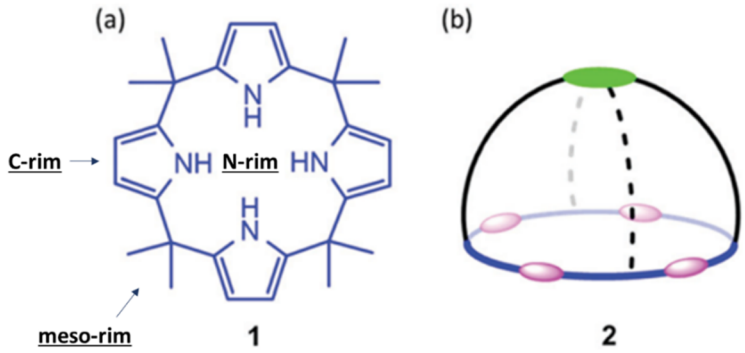


Figure 52: (a) Representation of the chemical structures of calix[4]pyrrole ; (b) schematic representation on how the calix[4]pyrrole assemble[149]

The calix[4]pyrroles are synthetic non-planar, non-aromatic tetrapyrrolic macrocyclic receptors capable of binding an array of anions as well as neutral substrates through N–H hydrogen bonding. The fundamental structure of calix[4]pyrroles consists of four pyrrole rings and is divided into three major parts: The C-rim, N-rim, and meso-rim.

These past decades, the calix[4]pyrroles were used in a very large variety of applications presented in the following Figure 53:

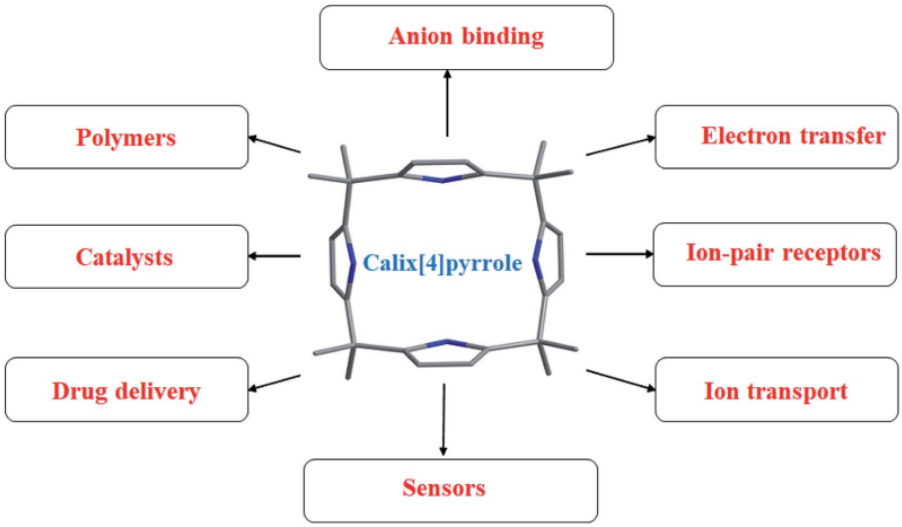


Figure 53: An overview of the possible applications for Calix[4]pyrrole[150]

II - SAMPL CHALLENGES AS A VALIDATION STEP FOR THE PLATFORM

II. A - THE SAMPL7 CHALLENGE

II. A. 1 - SAMPLING PROCEDURE

Considering the complexity of the conformational energy landscape of the complex and host molecule, we used multiple geometries of the unbound host system as starting points for minimization, thus increasing the probability of finding the absolute minimum. To do so, we extract approximately 15 structures from the classical molecular dynamics simulations to carry out a geometric optimization at a semi-empirical level, followed up by calculation of the hessian to confirm that the final energy is a true minimum (i.e., all vibrational frequencies are positive). The variation in free energy was as large as 10 kcal/mol for the different geometries, which confirmed the importance of conformational sampling. The overall lowest energy structure was defined as a reference for free energy calculation. Though the degrees of freedom of the guests are much reduced, we use a similar protocol for consistency.

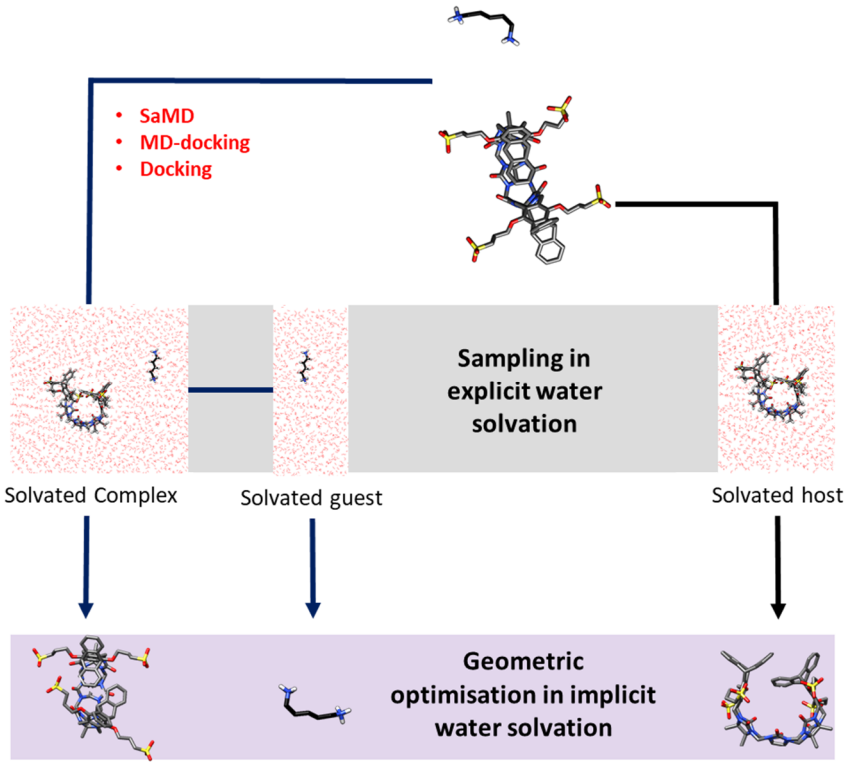


Figure 54: Protocol used to generate low-energy conformations of the apo host, the guest, and the host-guest systems. Three methods have been tested to generate initial models of the host-guest complex: SaMD, MD-Docking, and Docking. MD with explicit aqueous solvation is used to sample the conformational space. Then, for representative conformations, water is deleted, and the geometry is minimized with the GFN2B basis set in GBSA implicit water solvation.

II. A. 2 - RETROSPECTIVE ANALYSIS OF TRIMERTRIP SYSTEM

As a proof of concept for our methodology, we used the data from the trimertrip set in the SAMPL3 challenge. This host is similar but simpler than the one in SAMPL7. Docking with a large box (15 $Å^3$) produced complexes with negative binding energy (scoring), but the guest only formed surface interactions with the host. This led us to test two additional docking conditions where the docking space is progressively reduced. The resulting docking geometries have positive scores, indicative of conformational clashes, but in this case, the guest inserts into the host cavity. Three to five different binding modes were selected for each docking protocol. Minimization using *Chimera* allowed the system to relax before minimization and free energy calculation with GFN2B-xTB. Interestingly, the lowest-energy binding mode originated from the most restrictive docking protocol.

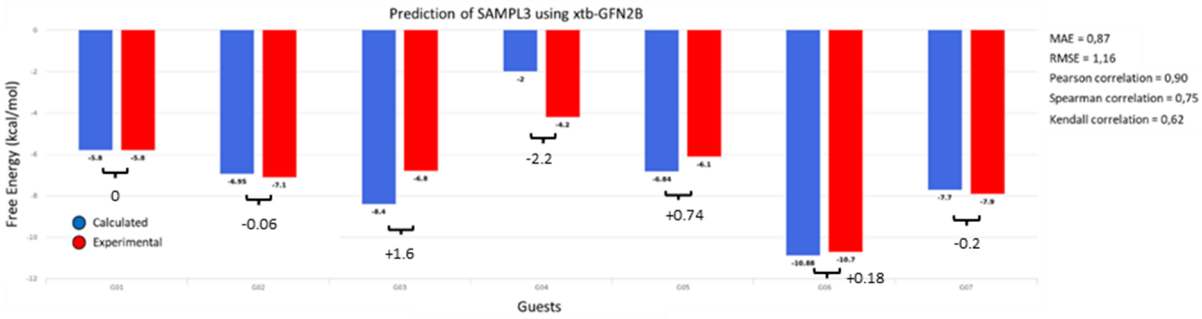


Figure 55: Results of the retrospective analysis of SAMPL3 Host-Guest complexes. Free energy predictions (blue bars) and experimental values (red bars) are in excellent agreement.

As shown in Figure 55, the predicted binding free energies are in excellent agreement with the experiment (RMSE = 1.16 kcal/mol; MAE = 0.87 kcal/mol; Pearson's correlation ® = 0.90; Spearman's rank correlation (ρ) = 0.75, Kendall's tau correlation = 0.62(τ)). In fact, in four out of the seven test cases, we obtain quantitative agreement. In one case, the error is below 1 kcal/mol, and in the two remaining cases, the errors are 1.6 kcal/mol and 2.2 kcal/mol. This led us to believe that, given the correct binding mode, the GFN2B-xTB semiempirical method could provide QM-level results at a small fraction of the computational cost (minimization plus calculation of the vibrational frequencies takes one to two hours per geometry on a desktop computer).

For that specific SAMPL3 dataset, retrospective analysis of the results shows very accurate results compared to the ones that have been published initially.

II. A. 3 - SAMPL7 TRIMER-TRIP BINDING MODE GENERATION

As in the test systems above, host-guest interactions were predicted by molecular docking considering different docking volumes in order to obtain a variety of binding modes, including some where the guest is fully inserted into the host. In the most restrained volume (which forces the guest to be located inside the host but yields positive score values), a molecular mechanics (MM) minimization of the docking solution is performed with MOE and *Chimera*, thus removing any potential clash between host and guest. For some particular systems (G08 and G10), the MM minimization was deemed insufficient to attain a relaxed complex. In those cases, docking was followed by 200 ns of MD simulations. Even then, it failed to generate any binding mode where the guest is embedded into the cavity of the cyclic host. Further adding to our problems, the sulfonate groups tended to form unrealistic interactions after minimization with *xTB*. In some cases, the sulfonates were even inserted into the host pocket, which is largely hydrophobic, instead of remaining solvent-exposed, as expected for a negatively charged group (Figure 56). This indicated that the implicit solvation model in *xTB* underestimates the desolvation cost of ionic groups.

Unlike what was observed with the trimer-trip host-guest systems of previous editions, we had to conclude that a better method was necessary to generate correct binding modes for the SAMPL7 test set. Our method should allow for host flexibility in order to allow guest embedding with reasonable geometries. On the other hand, it was clear that the implicit solvation model implemented in xTB was falling short for ionic systems, and explicit solvation would be necessary for the conformational sampling stage. Both requisites pointed to MD simulations as an optimal solution, which we proceeded to implement and test.

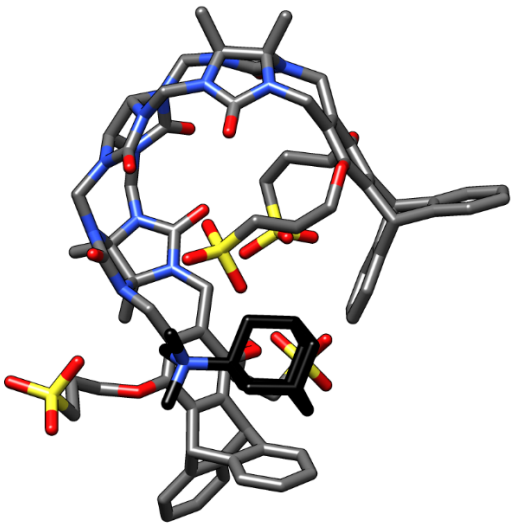


Figure 56: Binding mode of guest molecule G06 generated with docking and xTB. A sulfonate group enters the host pocket during geometric optimization, revealing an inadequate balance of solvation terms.

For the linear guests G01, G02, and G05, SaMD successfully completed the inclusion process, which proceeded in two steps: (i) rapid formation of surface contact between host and guest, leading to stable interactions; and (ii) a small opening of the host system, enabling the entry of the guest into the host cavity and formation of a stable complex (Figure 57). The second step is the bottleneck in the process. It occurs in a simulation time of 50 ns to 500 ns for the G01 compound, but for systems with longer alkene chains (more degrees of freedom) takes a much longer time. In G05, for instance, the simulation had to be extended to 1μs to observe a single association event (ca. 700ns). The application of the same methodology to the cyclic guest (i.e., G06, G07, G08, G09, G10, G11, G18, G19) failed to produce correct binding modes. While the compounds form stable surface interactions, they do not enter the host. This is in line with the above observation that host opening to admit the guest is the bottleneck in the association process. The bulkier nature of the cyclic guests implies that the host must (transitorily) adopt a wide-open conformation that is energetically unfavourable and cannot be sampled in the relatively short timescale of the MD simulations. To confirm this hypothesis, for the cyclic guest G07, we carried out an MD simulation starting from a fully open host system (generated by geometrical optimization in a vacuum). The guest rapidly proceeds to interact with the (now exposed) interior of the host, forming a stable but dynamic binding mode. After approximately 100ns, the host folds, trapping the guest in its interior (Figure 57). This result indicates that starting from metastable host conformations may be a general strategy to accelerate SaMD and generate valid host-guest geometries.

Notably, the binding mode of the guests inside the host is very dynamic, with fast rotations and frequent sliding movements that are only limited by the resistance of the charged group of the guest to enter the hydrophobic core of the host. As expected, the ionic groups rarely form direct contacts. Instead, they preserve their solvation shells. Overall, these results suggest that SaMD is an optimal and feasible strategy not only to obtain a bound conformation of the host-guest complex but also to capture the rich conformational diversity of the bound state. Unfortunately, between the setting up and testing of this protocol and the computational cost of the MD simulations, it was impossible to complete all these calculations by the challenge deadline. Posterior analysis confirms that correct identification of the binding mode through SaMD improves the quality of the binding free energy predictions (see next section).

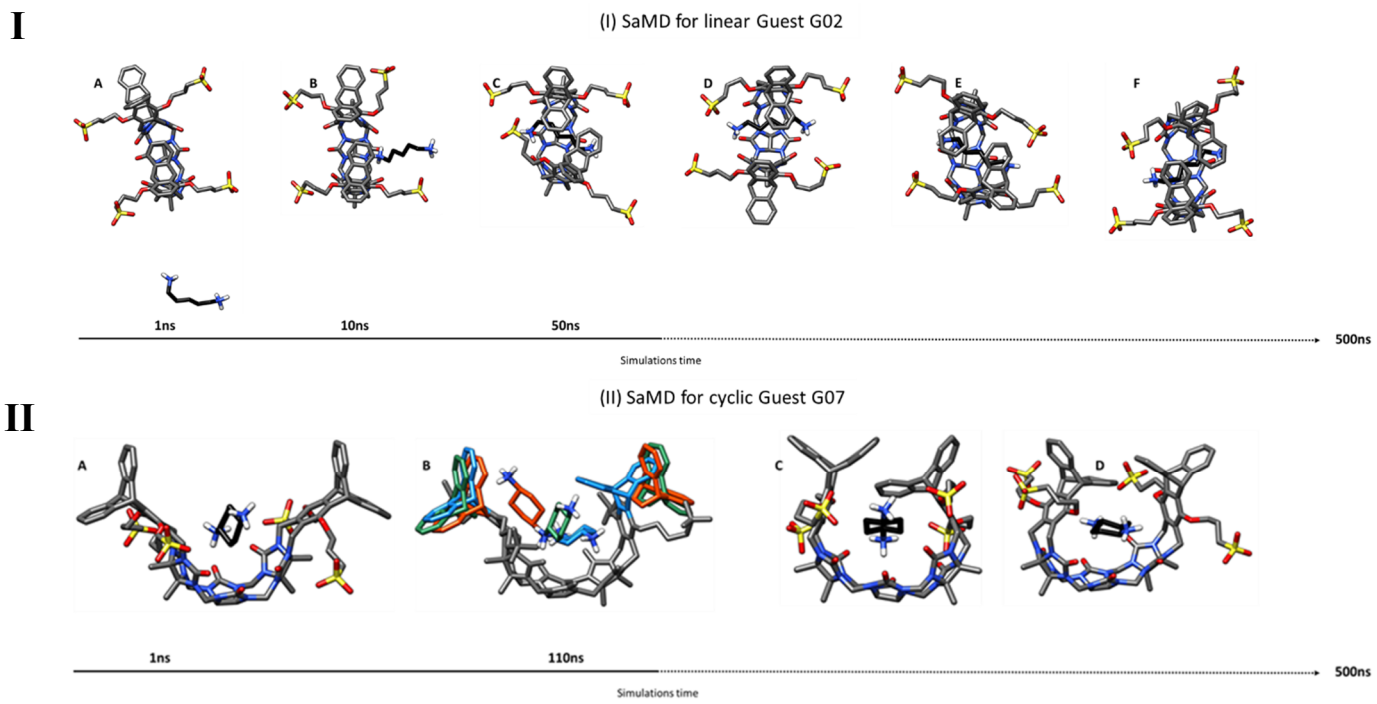


Figure 57: Inclusion process for trimer-trip dost-guest complexes observed with SaMD. (I) Linear guest G02 (A) starts from a fully dissociated state; (B) after ~ 10ns, surface interaction is formed between host and guest; (C) eventually, the host widens the cavity, and the guest molecule slides across to form a complex; (D-E-F) the complex remains stable but explores a variety of conformations for the remaining of the simulation. (II) Cyclic guest G07 (A) forms an encounter complex very early (~1ns); (B) and remains in contact with the host for over 100ns, until the host clicks into the closed geometry; (C-D) the complex remains stable but explores a variety of conformations for the remaining of the simulation.

II. A. 4 - SAMPL7 TRIMER-TRIP FREE ENERGY PREDICTION

For each complex, we extract 5 to 10 different binding modes generated with the above-described protocols. These geometries are then individually minimized at the GFN2B-xTB semi-empirical level, and only those yielding a true minimum (i.e., all vibrational frequencies are positive) are considered. The lowest energy complex is considered as the true minimum, except for a few cases where visual inspection identified issues with the corresponding geometry, always related to inadequate screening of charges by the implicit solvation method, such as those shown in Figure 56. Predictions for each system are shown in Table 11.

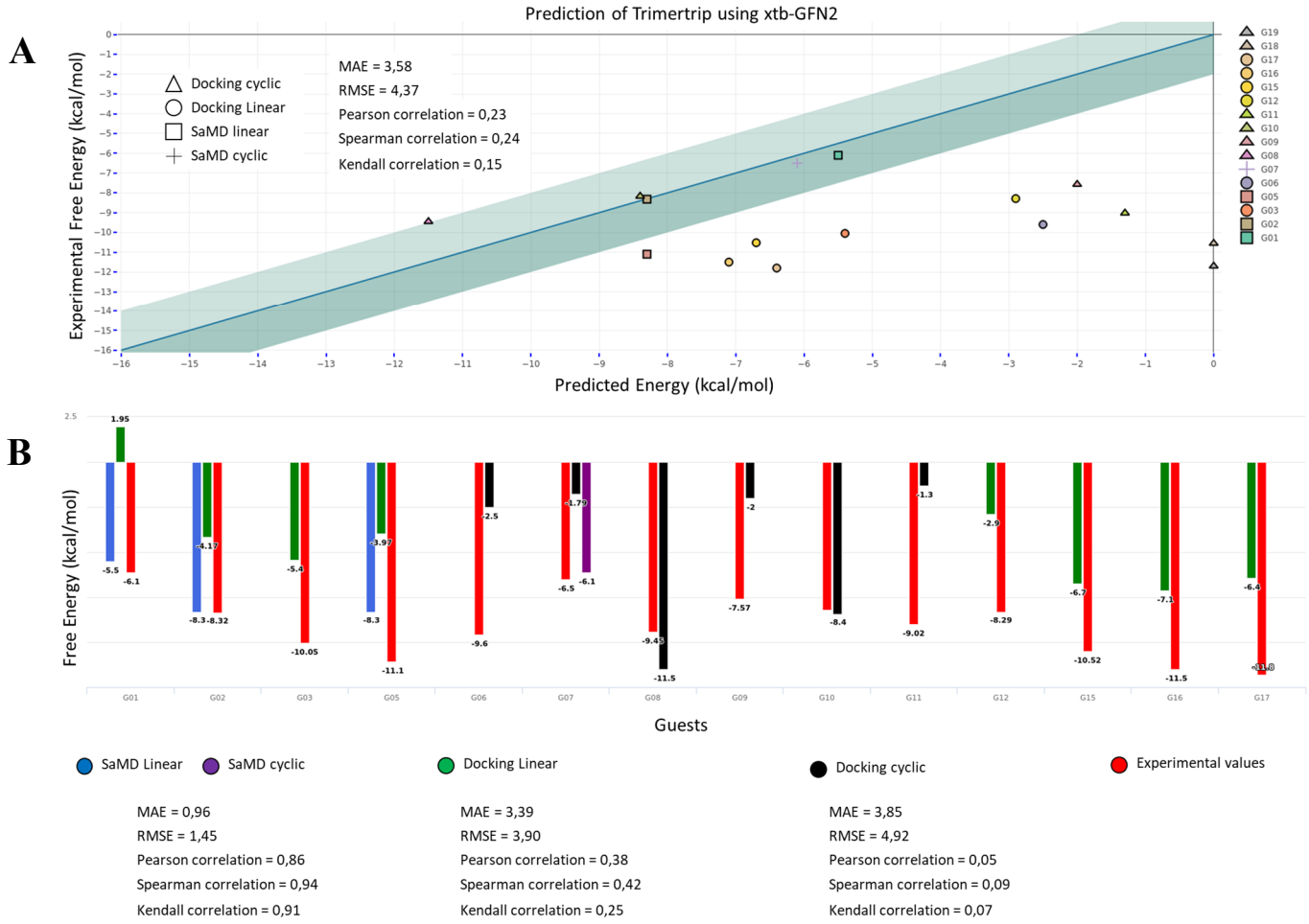


Figure 58: Comparison of experimental binding free energies with predicted values. (Top) correlation plot; the green-shaded area represents a threshold of +1/-1 kcal/mol from the experimental energy; the symbols indicate the nature of the guest and the method used for binding mode generation (triangle = docking for cyclic guest, circle = docking for linear guest, square = SaMD for linear guest, cross = SaMD for cyclic guest). (Bottom) histogram of binding free energy coloured by the method used for binding mode generation (black = docking for cyclic guest, green = docking for linear guest, blue = SaMD for Linear guest, purple = SaMD for cyclic guest). G18 and G19 guests are not shown or considered for statistical analysis because it was not possible to generate a plausible binding mode for them.

For guests G18 and G19, we could not find a correct binding mode SaMD, and the docking results gave positive binding energy. As both protocols failed for these two cyclic guests (presumably due to their large volumes), we desisted from making predictions for them.

We can see in Figure 58-A three different zones in the graphics: The first zone corresponds to the 5 Host-Guest systems that have been predicted well. Concerning these systems, G01, G02, and G07 are extracted from the SaMD protocol. At the same time, G08 and G10 are the two cyclic host from where interaction outside the cavity have been extracted from MD-docking. The second zone corresponds to the 5 Host-Guest systems, where our prediction was incorrect but still within the range from the experimental values (3 to 5 kcal/mol errors). These complex

(G03, G05, G15, G16, G17) are mainly linear, and the results originate from docking poses with the exception of G05, which originates from SaMD (result obtained after the submission deadline). The third zone corresponds to the 6 Host-Guest with large errors, including the G18 and G19 (for which no negative binding energy has been found). Most of them are cyclic, and the errors can be attributed to our inability to find reasonable binding modes in the timeline of the challenge.

In Figure 58-B, we show that for the complexes where SaMD delivers a correct binding mode, the binding free energy predictions are far superior to the results obtained from docking poses. In fact, most cases (G01, G02, G05, G07) are in quantitative agreement with the experiment (+/- 1kcal mol) and the overall performance statistics are excellent: for RMSE = 1.45 kcal/mol; MAE = 0.96 kcal/mol; Pearson's correlation (r) = 0.86; Spearman's rank correlation (ρ) = 0.94, Kendall's rank correlation = 0.91(τ)**.** Compared to SaMD, the results from docking underestimate the binding free energy, which suggests that lower-energy conformations of the Host-Guest complex can be sampled with MD but not with the MM protocols.

II. A. 5 - KNOWLEDGE-BASED APPROACH

For GDCC prediction, as there was an important amount of pre-existing data from previous challenges, we decided to try an orthogonal approach-based ML. The dataset includes 35 compounds in total, belonging to three classes of host systems that are similar in structure and chemical composition: OA, TEMOA, and exoOA (Figure 47). The binding free energy values range between -3.73 kcal/mol and -8.38 kcal/mol. The used model is an NNET, using 90 CORINA descriptors (60 describing the guest and 30 describing the host system). As expected, the predictions for the training set are very accurate, with RMSE = 0.92 kcal/mol and all the predicted values within a 1 kcal/mol range from the experimental values (Figure 59). For the test set, all the predicted values are close to the experimental one, with maximum and minimum errors of -1.49 kcal/mol and +0.22 kcal/mol, respectively.

The GDCC-7 dataset to be predicted this year consisted of 8 guest compounds (four charged and four non-charged) binding to two related host systems. After the model has been optimized, it takes only 10 seconds to calculate the binding free energy of the eight guests in the two hosts. With RMSE and MAE values of 1.67 kcal/mol and 1.21 kcal/mol, respectively, the overall performance is rather satisfactory, especially by comparison with the thermodynamic-based approach. Worth noting that the four negative guests are not predicting well, which can be explained by the limits of the model imposed by the composition of the training set: since the

least favourable binding free energy value is -3.73 kcal/mol, the model cannot predict more positive values. Even then, the hierarchy between the guest values is respected (G4 < G3 < G2). There is no experimental value for G1, so it has not been considered for this analysis. If we apply the same analysis to every subgroup (based on the positive or negative charge and the host they are interacting with), we obtain an almost perfect hierarchical prediction. The only exception is the OA-G7 complex, which was predicted lower than OA-G6 due to the fact that OA-G7 has been underestimated (-5,67 kcal/mol instead of -6.98 kcal/mol) while OA-G6 have been predicted very close to his experimental values (-5.92 for -5.83 experimental values). In fact, all systems, except for the four negative compounds interacting with exo-OA, are predicted within 1 kcal/mol of the experimental values (Figure 60). For the complexes involving the OA system, which features prominently in the training set, the predictions are better still, with MAE = 0.55 kcal/mol and RMSE = 0.85 kcal/mol.

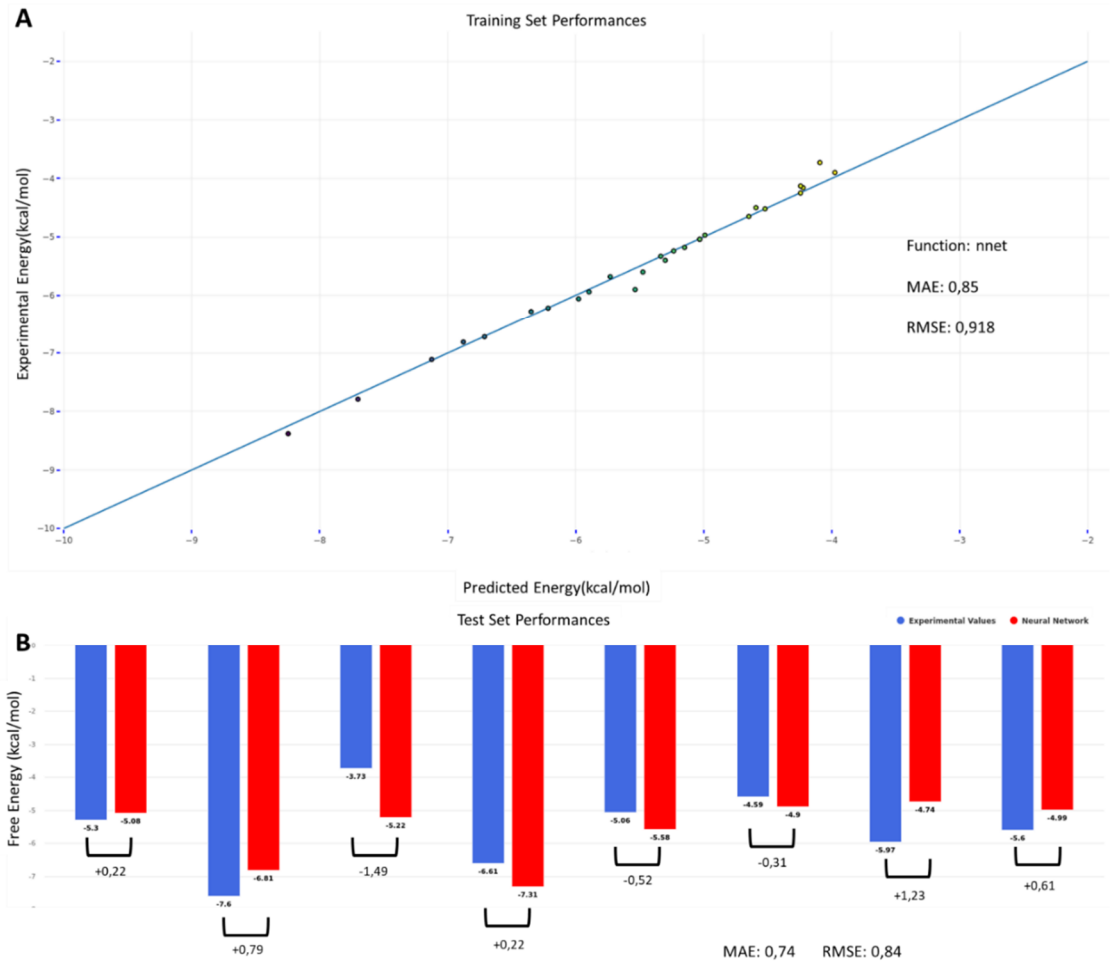


Figure 59: A performance of the training set including 27 different guests interacting with two different systems. (B) the test set includes eight guest molecules with free energy predicted using the training set.

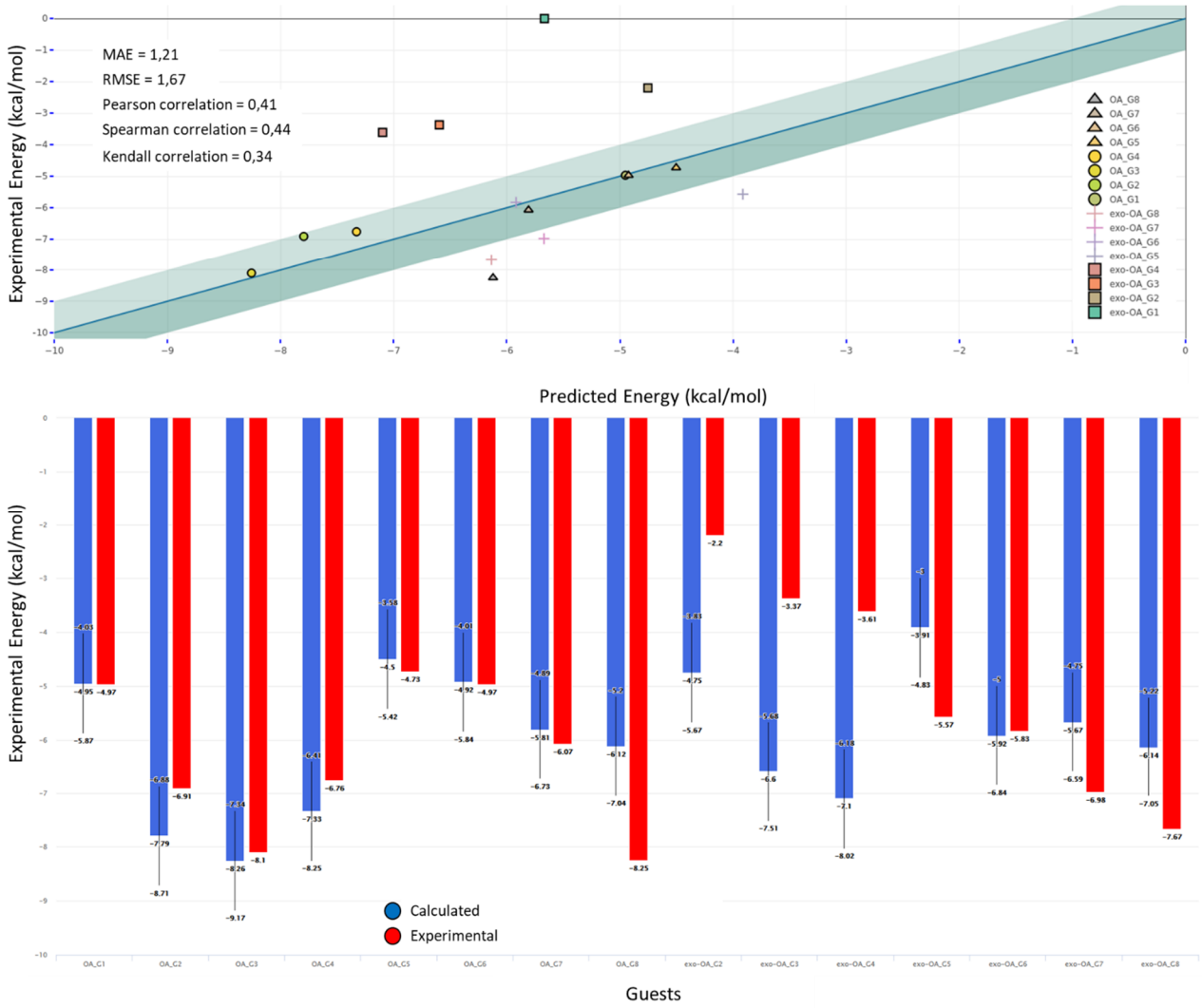


Figure 60: Comparison of experimental binding free energies with predicted values. (Top) correlation plot; The green-shaded area represents a threshold of ±1 kcal/mol from the experimental energy; the symbols indicate the nature of the guest, and each prediction has a different colour (triangle = positively charged guest interacting with OA system, circle = negatively charged guest interacting with OA system, square = negatively charged guest interacting with the exo-OA system, cross = positively charged guest interacting with the exo-OA system). (Bottom) histogram of binding free energy with calculated (blue) and experimental values (red). The error bars reflect the RMSE of the nnet model on the training set (0.918 kcal/mol).

II. A. 6 - CONCLUSION ON THE SAMPL7 CHALLENGE

The participation in SAMPL7 allowed us to test two orthogonal approaches to calculate host-guest binding free energies, identifying in each case strengths and limitations.

The thermodynamic-based approach is absolutely general and can be used, in principle, on any host-guest system. The use of an advanced semiempirical basis set (GFN2B) to calculate energies and thermostatistical corrections offers increased performance relative to MM approaches with a moderate computational cost (1-2h on a single CPU) and eliminates the

dependency on small-molecule force-fields, which are often inaccurate. However, we have identified critical aspects that can lead to incorrect predictions. The first one is a critical dependency on the structure of the host-guest complex used to generate the prediction (the binding mode). For systems with significant host flexibility, rigid receptor docking can be inappropriate, and host conformational sampling is necessary. Direct observation of the host-guest pair formation through molecular dynamics with explicit solvent is an optimal solution in terms of quality of the binding free energy predictions but can be unpractical due to the long simulations times, which increase with the number of degrees of freedom of the system. In the trimertrip case, we identified a slow transition between the closed and open conformation of the host as the bottleneck in the association process. For such cases, starting the SaMD simulations with open host conformations can yield excellent results at a fraction of the simulation cost. The second limitation of our approach is the implicit solvation method (GBSA) which can underestimate the desolvation cost of ionic species in aqueous solvation, leading to the formation of ionic pairs which contribution is overvalued. Other reports have observed a systematic bias with implicit solvation models. We do not observe such systematic bias, but the implicit solvation model remains one of the weaknesses of the approach. In any case, the explicit solvation in MD simulations is better suited to preserve the solvation shells around the solute's ionic groups. Thus, the use of MD snapshots as input geometries in GFN2B-xTB calculations seems to provide better results than exhaustive conformational sampling with implicit solvation.

The use of knowledge-based methods can be highly advantageous when there is sufficient pre-existing data. Unlike protein-ligand complexes, where a large body of data exists, host-guest systems cannot benefit from massive training sets. Thus, we were particularly interested in examining the suitability of machine learning approaches, with a particular concern on the risk of overfitting. The results obtained on the GDCC system are really encouraging and motivate us to build a database of host-guest systems with their corresponding binding free energies and train both general and host-specific models.

Overall, the participation in SAMPL7 has allowed us to design an automatic pipeline to compute binding free energies for any Host-Guest system. This automatic pipeline was the basis of the thesis, and the final version was presented in the previous chapter (HG-DYNAusor).

II. A. 6. A - OVERVIEW OF THE RESULTS

Table 11: Final results with experimental, calculated binding free energy and the error related.

Challenge system	Case name	Experimental ΔG_{bind} (kcal/mol)	Predicted ΔG_{bind} (kcal/mol)	Error (kcal/mol)
TRIMERTRIP	G01	-6,10	-5,50	-0,60
	G02	-8,32	-8,30	-0,02
	G03	-10,05	-5,40	-4,65
	G05	-11,10	-8,30	-2,80
	G06	-9,60	-2,50	-7,10
	G07	-6,50	-6,10	-0,40
	G08	-9,45	-11,50	2,05
	G09	-7,57	-2,00	-5,57
	G10	-8,17	-8,40	0,23
	G11	-9,02	-1,30	-7,72
	G12	-8,29	-2,90	-5,39
	G15	-10,52	-6,70	-3,82
	G16	-11,50	-7,10	-4,40
	G17	-11,80	-6,40	-5,40
	G18	-10,55	0,00	-10,55
	G19	-11,70	0,00	-11,70
GDCC	OA-G1	-4,97	-4,95	-0,02
	OA-G2	-6,91	-7,79	0,88
	OA-G3	-8,10	-8,26	0,16
	OA-G4	-6,76	-7,33	0,57
	OA-G5	-4,73	-4,50	-0,23
	OA-G6	-4,97	-4,92	-0,05
	OA-G7	-6,07	-5,81	-0,26
	OA-G8	-8,25	-6,12	-2,13
	ExoOA-G1	0,00	-5,67	5,67
	ExoOA-G2	-2,20	-4,75	2,55
	ExoOA-G3	-3,37	-6,60	3,23
	ExoOA-G4	-3,61	-7,10	3,49
	ExoOA-G5	-5,57	-3,91	-1,66
	ExoOA-G6	-5,83	-5,92	0,09
	ExoOA-G7	-6,98	-5,67	-1,31
	ExoOA-G8	-7,67	-6,14	-1,53

Table 12: Statistical analysis of SAMPL3 calculation, SAMPL7 TRIMERTRIP, and SAMPL7 GDCC prediction.

	MAE (kcal/mol)	RMSE (kcal/mol)	Pearson Correlation	Spearman correlation (rho)	Kendall correlation (tau)
SAMPL3					
S3_completedataset (n=7)	0,87	1,16	0,90	0,75	0,62
SAMPL7 TRIMERTRIP					
S7_complete dataset (n =14)	3,58	4,37	0,23	0,24	0,15
S7_Linear (n =8)	3,39	3,90	0,38	0,42	0,25
S7_cyclic (n =6)	3,85	4,92	0,05	0,09	0,07
S7_confident (n =4)	0.96	1.45	0,86	0,94	0,91
SAMPL7 GDCC					
S7_complete dataset (n=15)	1,21	1,67	0,41	0,44	0,34
S7_OA (n=8)	0,54	0,85	0,80	0,85	0,76
S7_exoOA (n=7)	1,98	2,27	-0,04	0,00	0,05
S7_positive_guest (n=7)	1,56	2,08	0,74	0,89	0,81
S7_negative_guest (n=8)	0,91	1,20	0,73	0,81	0,57
S7_Positive_OA (n=4)	0,41	0,53	0,97	1,00	1,00
S7_negative_OA (n=4)	0,67	1,08	0,90	1,00	1,00
S7_positive_exoOA (n=3)	3,09	3,12	1,00	1,00	1,00
S7_negative_exoOA (n=4)	1,15	1,31	0,68	0,80	0,67

II. B - SAMPL8: CB[8] DRUG ABUSE CHALLENGE

II. B. 1 - PRESENTATION OF THE CHALLENGE

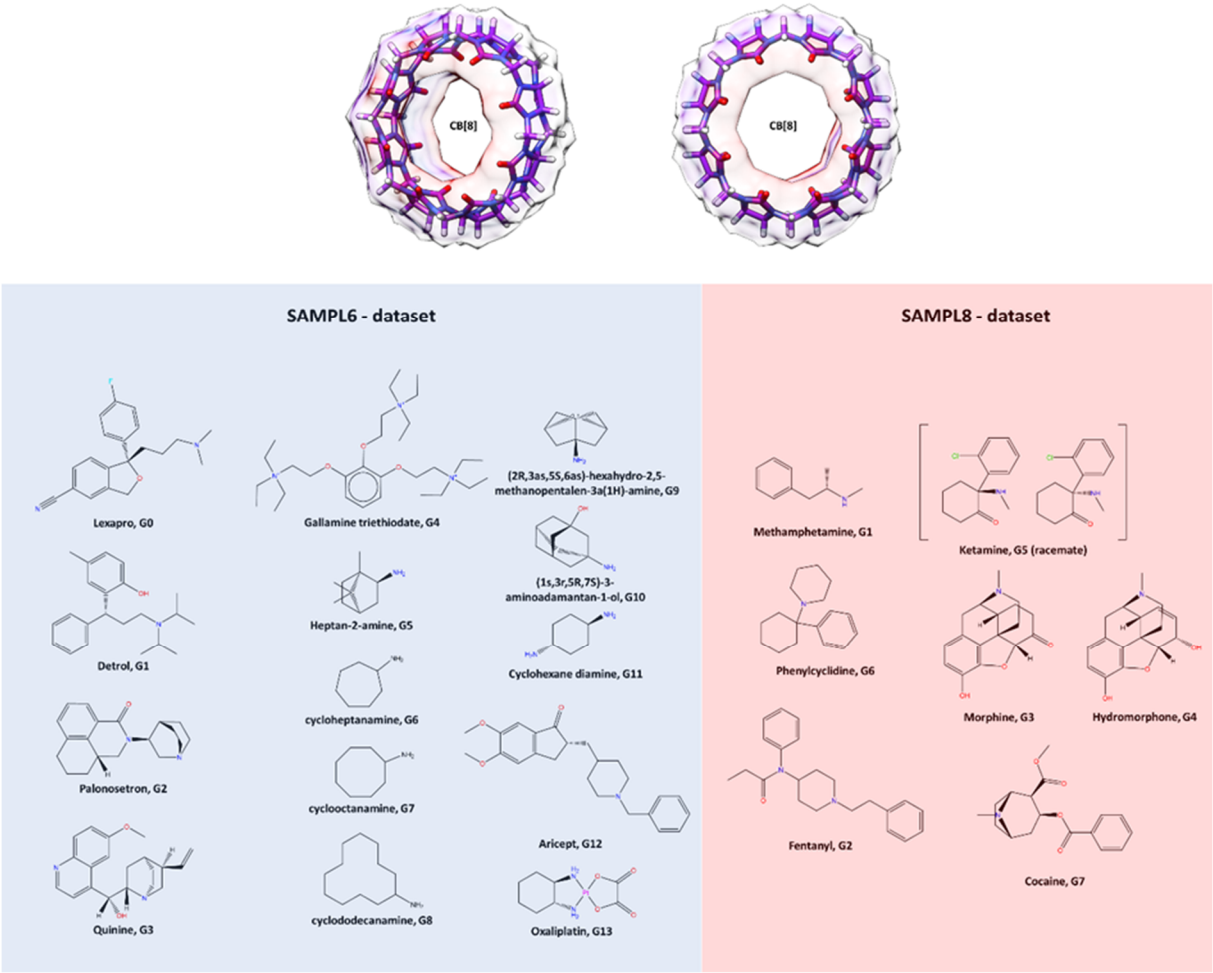


Figure 61: Molecules used for the SAMPL challenges; in blue, the SAMPL6 dataset used for the retrospective analysis (in blue); in red, the SAMPL8 dataset for which the binding free energy has to be predicted.

The CB8 "drugs of abuse" challenge focused on the binding of CB8 to seven guests, which are drugs of abuse, including morphine, hydromorphone, methamphetamine, cocaine, and others. Binding has been experimentally characterized and a provisional patent filed by the Isaacs group for potential biological application. As always with the SAMPL challenge, the prediction is blinded, and the users have a limited amount of time to provide a prediction. In the CB[8] SAMPL challenge, like in the previous SAMPL7, we have followed two different approaches: the thermodynamic based approach based on modules one to three of the HG-DYNAusor platform and the knowledge-based method based on module five of the said platform.

II. B. 2 - RETROSPECTIVE ANALYSIS

As in the SAMPL7 challenge, we have used some pre-existing data as a test set for our methodology. It consists of 14 molecules (Figure 61) binding the CB[8] host, with their corresponding binding free energies.

Compared to the SAMPL7 challenge, and inspired by it, the protocol evolves to become the actual protocol of the HG-DYNAusor platform:

- 500ns to 1µs of simulation is done for the CB[8] host alone in the water.
- Several MD simulations are launched in an attempt to observe spontaneous association (SaMD). Compared to the previous Trimertrip used in the SAMPL7 challenge, the CB[8] host does not display important intrinsic mobility, leading us to think that, in the course of the SaMD simulation, it will not open sufficiently to take up the ligands.
- For each of the guests, the binding is obtained by docking protocol on the circular form of the CB[8] system. Then a simulation of 250ns of the complex is realized, and geometry is extracted every 2.5ns. In some cases, several docking solutions were comparable and could not be discriminated: in that specific cases, two or three MD simulations are launched from different starting points.

The retrospective analysis (Figure 62) shows a very good agreement between the predicted binding free energy and the experimental values for most of the compounds. The binding mode is presented in Figure 63 to Figure 65. We can see three exceptions in the prediction for which the binding mode is coloured in yellow: G1, G2, and G11. G1 is underpredicted (with a very low binding free energy), while the G2 and G11 are overestimated.

Taking a look at the overestimated structure, we do not find any perturbation in the binding mode, the compound G4 is very large, and all the charged nitrogen are facing the solvent, while the G11 is fitting well in the cavity and realize two hydrogen bonds on each of the nitrogen atoms. In both cases, the compounds present multiple ionic centres, suggesting a problem in the calculation of the thermodynamic properties due to the limitations of the continuous solvation model and a larger error in the balance between interaction energy and desolvation cost. Compound G1 is clearly underestimated, but in the predicted binding mode, it does not fully interact in the pocket and has an intramolecular hydrogen bond that could contribute to the wrong prediction. Considering the outcome of the G1 structure, we reasoned that the error in the prediction was essentially due to our inability to find the proper binding mode. It would

have been interesting to investigate if it was possible to find a more plausible binding mode by launching MD simulations from various starting. However, due to the timing of the challenge, this was not feasible. Except for these errors, we considered the proof of concept a success, particularly as the guests to be predicted were charged entities, which are more challenging for the solvation methods.

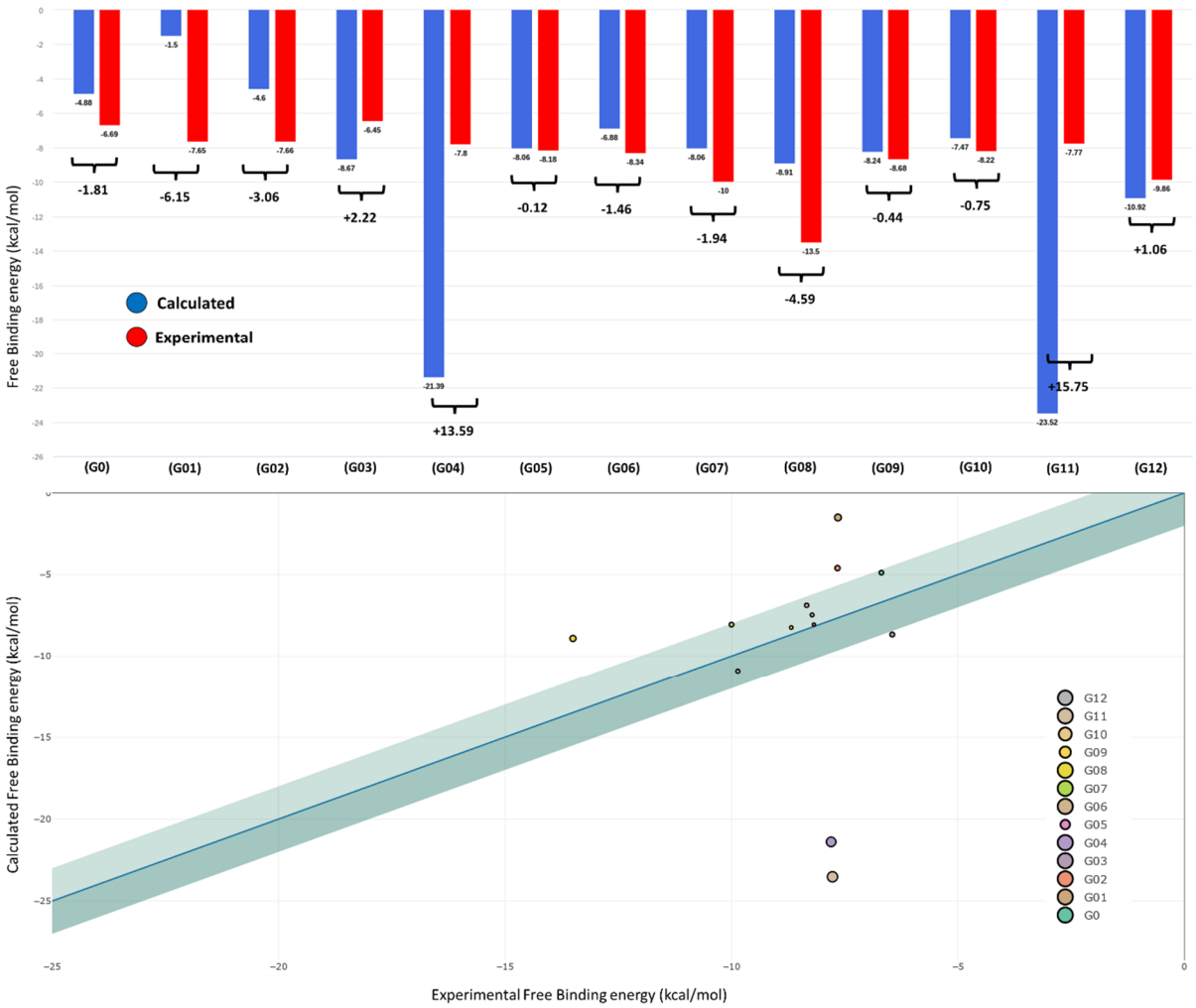


Figure 62: Overview of the results of the retrospective analysis done on the CB[8] system with data extracted from the SAMPL6 challenge. In the (Top) correlation plot, the green-shaded area represents a threshold of +2/-2 kcal/mol from the experimental energy. In the Bottom: histogram of binding free energy.

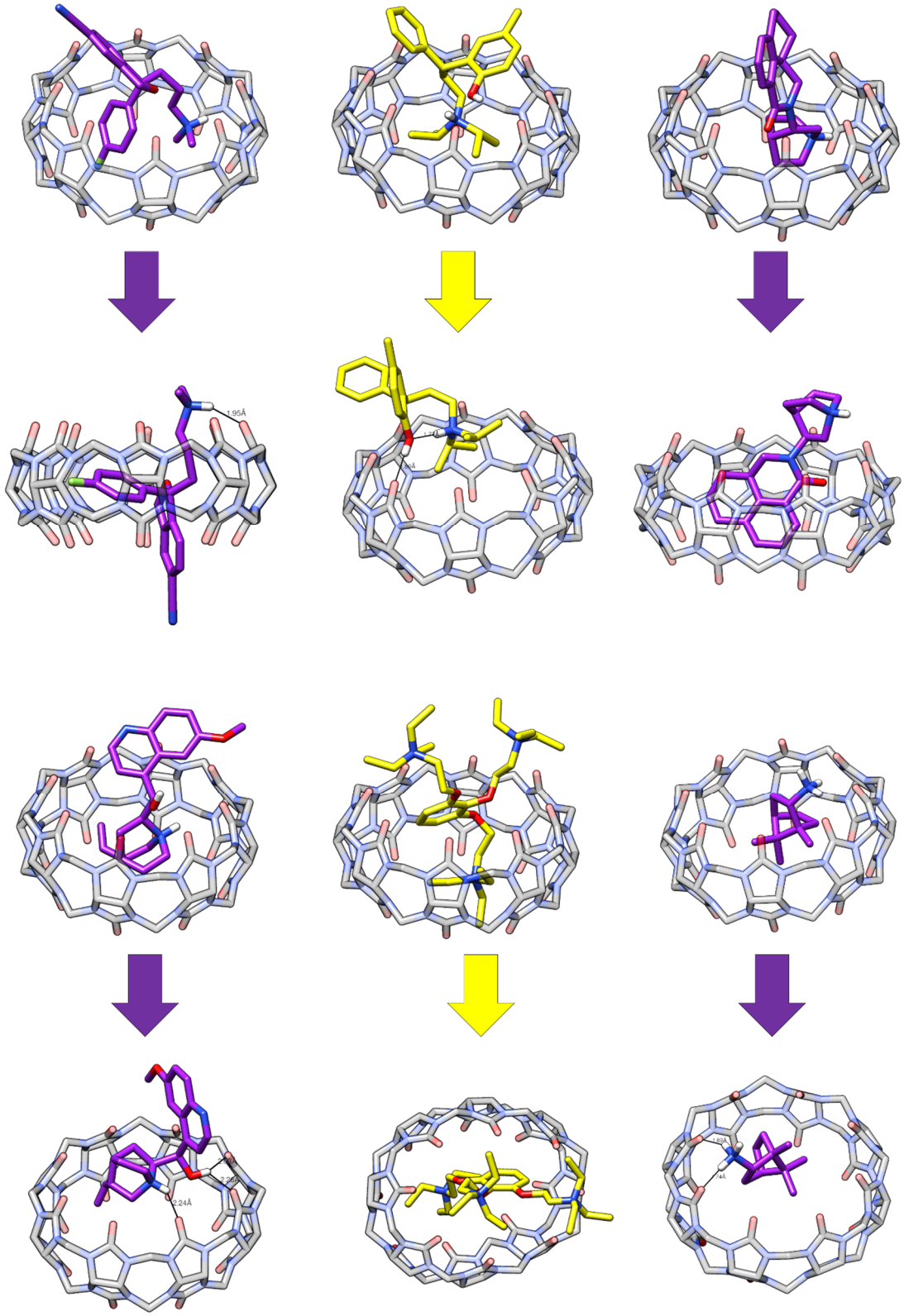


Figure 63: Overview of the outcome of the thermodynamic-based approach for the retrospective analysis: in the upper part from left to right: complex G0, G1, and G2 and in the bottom part from left to right: G3, G4, and G5. The arrows represent the transformation between the docking outcome and the minimal energy structure extracted from MD (. In purple, the results with good agreement from experimental and in the yellow the results with a bad agreement.

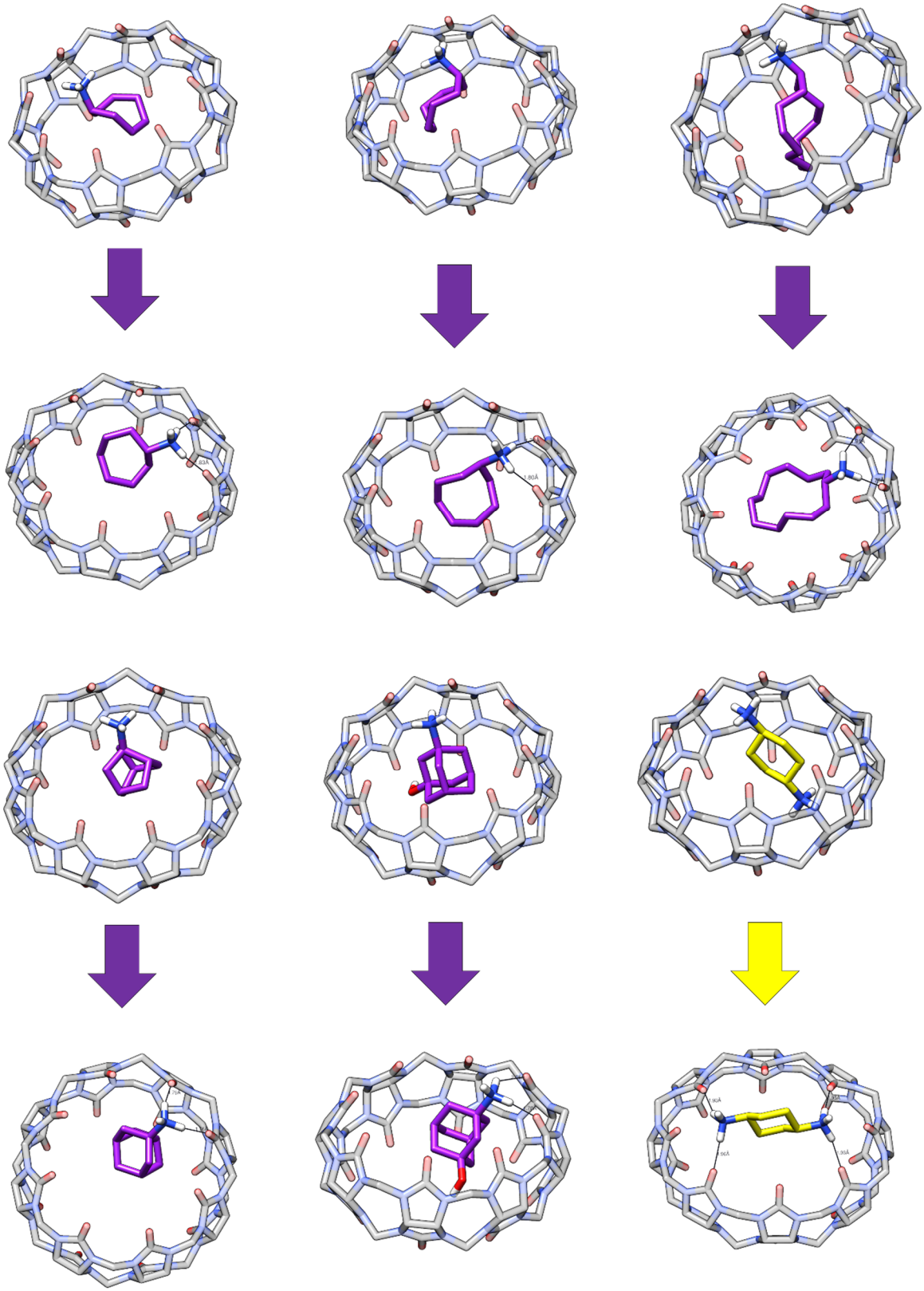


Figure 64: Overview of the outcome of the thermodynamic-based approach for the retrospective analysis: in the upper part from left to right: complex G6, G7 and, G8 and in the bottom part from left to right: G9, G10 and, G11. The arrows represent the transformation between the docking outcome and the minimal energy structure extracted from MD. In purple, the results with good agreement from experimental, and in yellow, the results with a bad agreement.

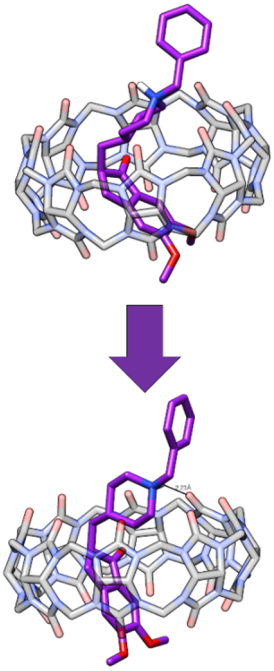


Figure 65: Overview of the outcome of the thermodynamic-based approach for the retrospective analysis: complex G12. The arrows represent the transformation between the docking outcome and the minimal energy structure extracted from MD.

II. B. 3 - THERMODYNAMIC BASED METHOD

II. B. 3. A - BINDING MODES

The CB[8] system presents mainly two different modes in a solvated environment (U-shaped and circular). We had a good agreement in the retrospective analysis with the circular geometry, and for this reason, it was selected as the conformation of choice for docking studies. Initial host-guest interactions were predicted by molecular docking, considering the same receptor structure for all the ligands. As the cavity of the CB[8] system is well defined, blind docking using a very large box leads to a binding mode without any distortion in the ligand, and the first scored solution is extracted for all the complexes. Even though no distortion was apparent, a molecular mechanics (MM) minimization of the docking solution was performed with *Chimera*, thus removing any potential clash between host and guest.

For some particular systems (G1 and G7), two different docking solutions could be extracted with a similar score, in those cases, as it was impossible to discriminate between them, the two solutions were extracted and simulated. All the docking solutions were followed by a 250ns simulation, from which one geometry is extracted every 2.5ns for a total of 100 geometries for each of the complexes. These geometries are then individually minimized at the GFN2B-xTB semi-empirical level, and only those yielding a true minimum (i.e., all vibrational frequencies

are positive) are considered. Two different approaches are then considered: (i) the lowest energy complex is considered as the true minimum (this option was submitted to the SAMPL8 CB[8] challenge), and (ii) a Boltzmann-weighted average is calculated for all the structures corresponding to true minima. The initial geometry and the optimized complex of the lowest energy configuration are shown in Figure 66 and Figure 67:

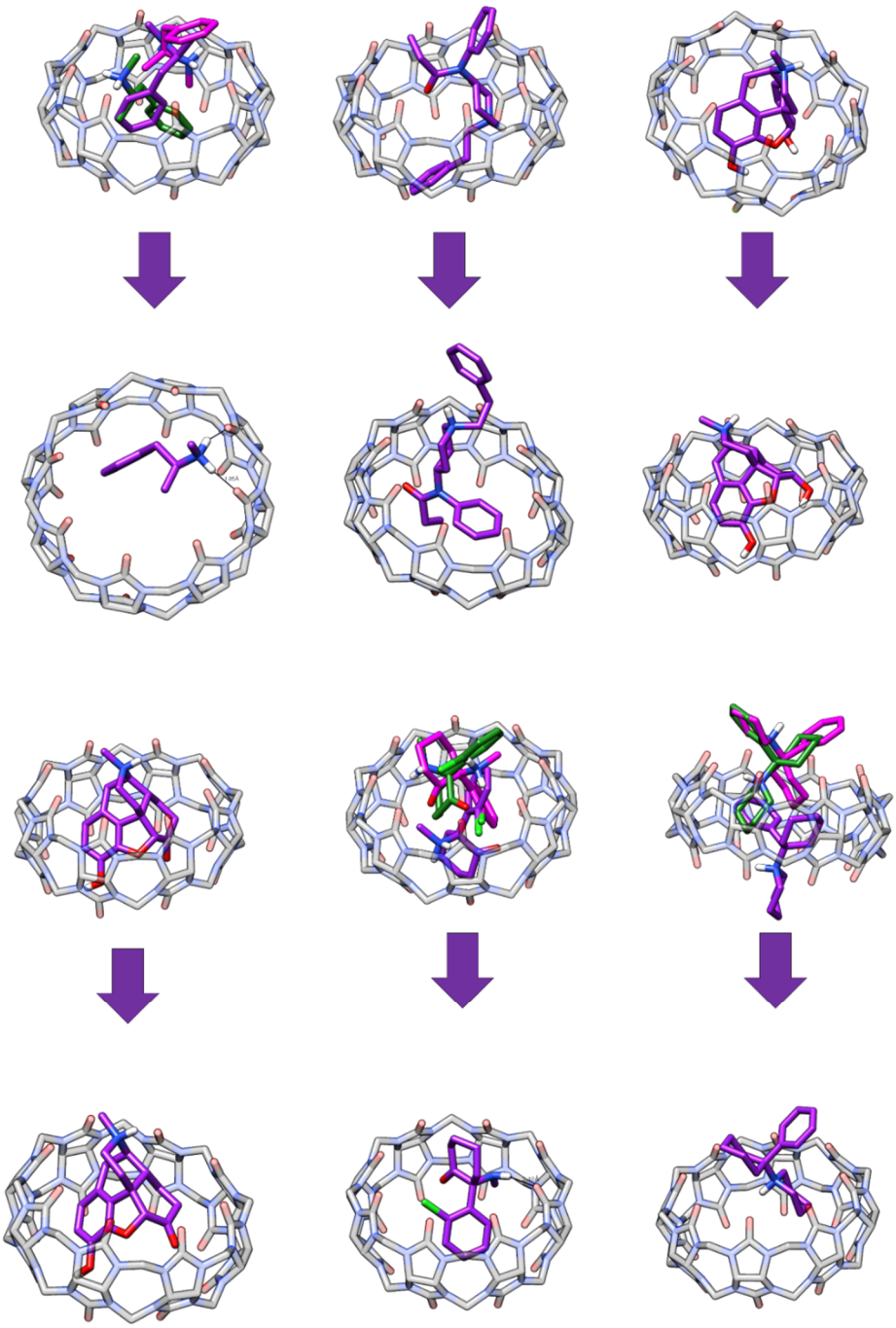


Figure 66: Overview of the outcome of the thermodynamic based approach for the SAMPL8 CB[8] challenge: in the upper part from left to right: complex G1, G2, and G3 and in the bottom part from left to right: G4, G5, and G6. The purple arrows represent the transformation between the docking outcome and the extracted structure from MD.

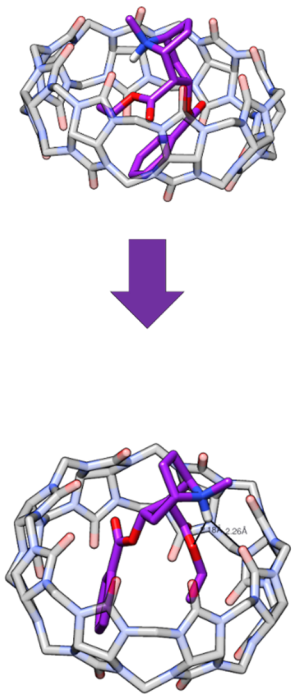


Figure 67: Overview of the outcome of the thermodynamic based approach for the SAMPL8 CB[8] challenge: complex G7

I. A. 11. A MINIMAL ENERGY STRUCTURES

The ranked solution using the minimal energy complex is presented in the following Figure 68:

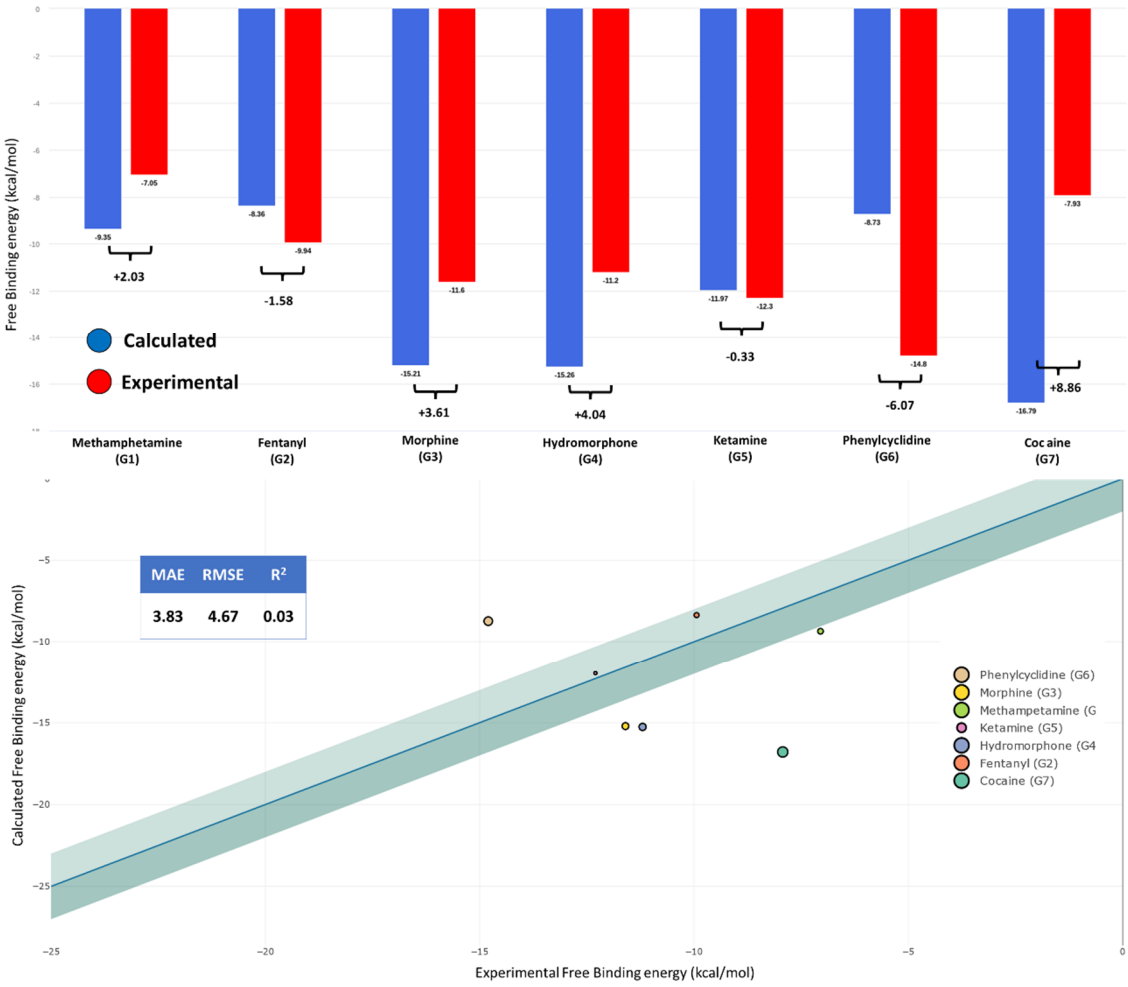


Figure 68: Comparison of experimental binding free energies with predicted values. (Top) histogram of binding free energy coloured by the origin of the data: in red the experimental data and in blue the calculated values. (Bottom) correlation plot; the green-shaded area represents a threshold of +2/-2 kcal/mol from the experimental energy. The statistical analysis is shown in the blue box: with MAE = 3.83, RMSE = 4.67, and R^2 = 0.03.

Based on the accuracy of the prediction (Figure 68), the results can be separated into three groups:

- (i) The first group represent the prediction with an excellent agreement with the experimental data (< 2 kcal mol) represented by the green shaded area. Concerning these systems, Ketamine (G5, -0.33 kcal/mol) is extracted from three different docking solutions, and Fentanyl (G2, -1.58 kcal/mol) is extracted from a single docking solution.

- (ii) The second group represent the incorrect prediction but are still in the range from the experimental values (~2 to 4 kcal/mol errors). These complexes: Morphine (G3, +3.61 kcal/mol), Hydromorphone (G4, +4.04 kcal/mol), are extracted from one docking pose, while Methamphetamine (G1, +2.3 kcal/mol) is extracted from three different docking solutions.

- (iii) The third group represent the prediction with large errors (> 4 kcal/mol). For these complexes: Phenylcyclidine (G6, -6.07 kcal/mol) is extracted from three docking solutions, while Cocaine (G7, +8.86 kcal/mol) is extracted from a single docking pose. For those compounds, our errors can be attributed to our inability to find reasonable binding modes in the timeline of the challenge.

Again, our results are mainly depending on the extracted binding mode. The binding mode of each of the ranked compounds (with the structure corresponding to the extracted minimum) is shown in Figure 66 and Figure 67. Additionally, for all of the overestimated structures (G3, G4, and G7), we found in our predictions binding free energy that corresponds to the experimental values, but it was not the best-scored energy, and as we wanted a fully automated application for the predictions, we followed the same protocol for all the predicted compound.

II. B. 3. B - BOLTZMANN AVERAGE

The solution corresponding to the Boltzmann average is presented in Figure 69:

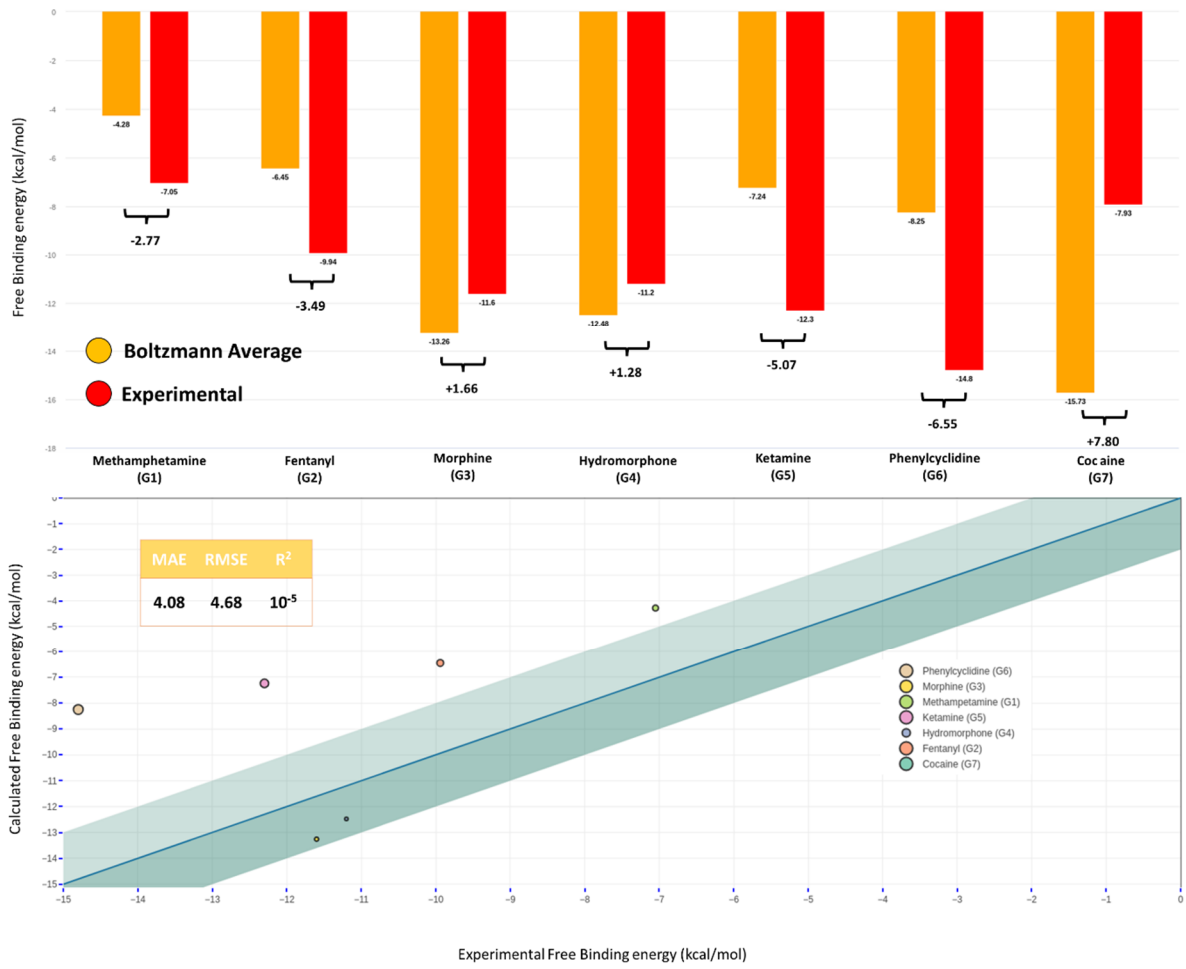


Figure 69: Comparison of experimental binding free energies with predicted values. (Top) histogram of binding free energy coloured by the origin of the data: in red the experimental data and in orange the Boltzmann average of the calculated values. (Bottom) correlation plot: the green-shaded area represents a threshold of +2/-2 kcal/mol from the experimental energy. The statistical analysis is shown in the orange box: with MAE = 4.08, RMSE = 4.68, and $R^2 = 10^{-5}$.

We can see in Figure 69 similar results as in the previous section. As we calculated a weighted average that gives more importance to the more negative results, in all the cases, the energy increases relative to those obtained using only the lowest-energy structure. We are closer for the predictions of morphine (G3, +1.66 kcal/mol) and hydromorphone (G4, +1.28 kcal/mol), but in comparison, Fentanyl (G2, -3.49 kcal/mol) and Ketamine (G5, -5.07 kcal/mol) for which the extracted minimal-energy corresponded to structure considered as a rare event in the simulations (only encountered few times) are not well predicted. For these compounds, the Boltzmann average approach underestimated the energy, while the prediction was good in the previous analysis. This result indicates that the set of structures obtained with our procedure

does not correspond to a physically meaningful conformational ensemble. In some instances, it may be able to correctly identify the true minima but does not provide a faithful representation of the free energy landscape.

II. B. 4 - KNOWLEDGE-BASED METHOD

II. B. 4. A - SCOPE OF THE MODELS

II. B. 4. A. (I) - THE NEURAL NETWORKS (NNET)

In order to define the best model possible in terms of performances, several tuning parameters are tested: (i) the number of hidden units, (ii) the number of hidden layers, and (iii) the weight decay.

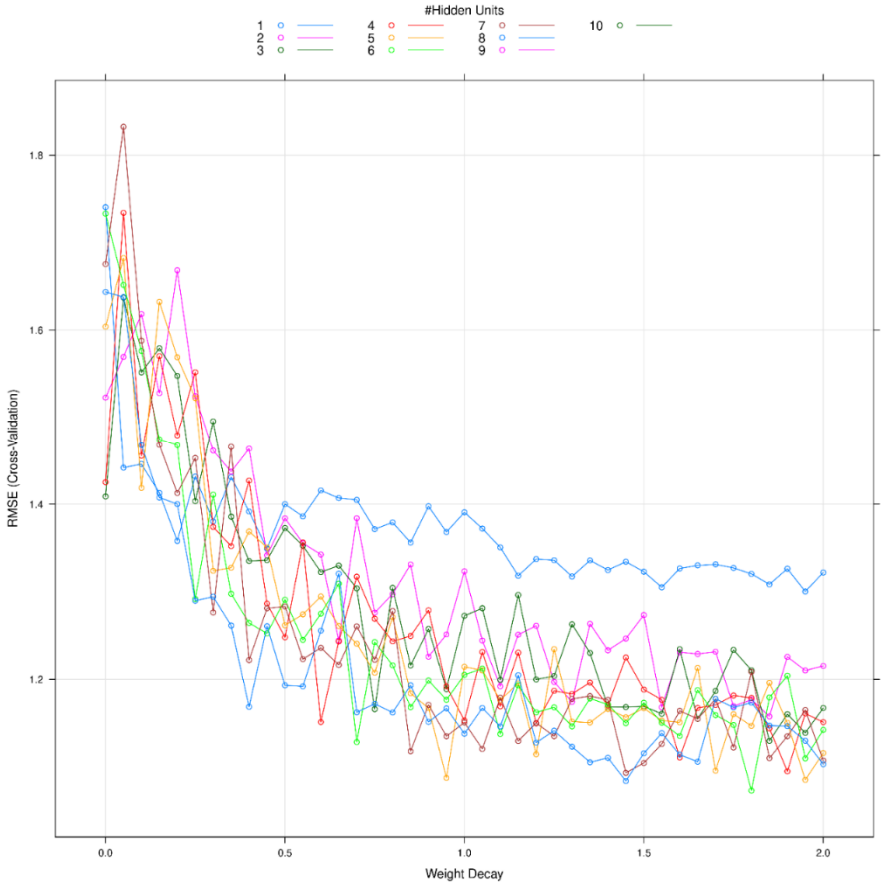


Figure 70: Performances of the NNET using a combination of hidden units and Weight decay to find the best performances for the model.

Theoretically speaking, the more hidden levels, the better and more the model will be able to capture nonlinear relationships in the data. But in practice, with more levels, the number of parameters increases, and finally, the best model corresponds to a combination of the specific values of weight decay and the hidden unit, corresponding to the lowest RMSE. In our case,

the results show that the optimum number of hidden units is six, and the decay parameter is 1.8 (Figure 70).

A list of the relative importance of the variables used by the NNET for the prediction can be visualized in the following Figure 71:

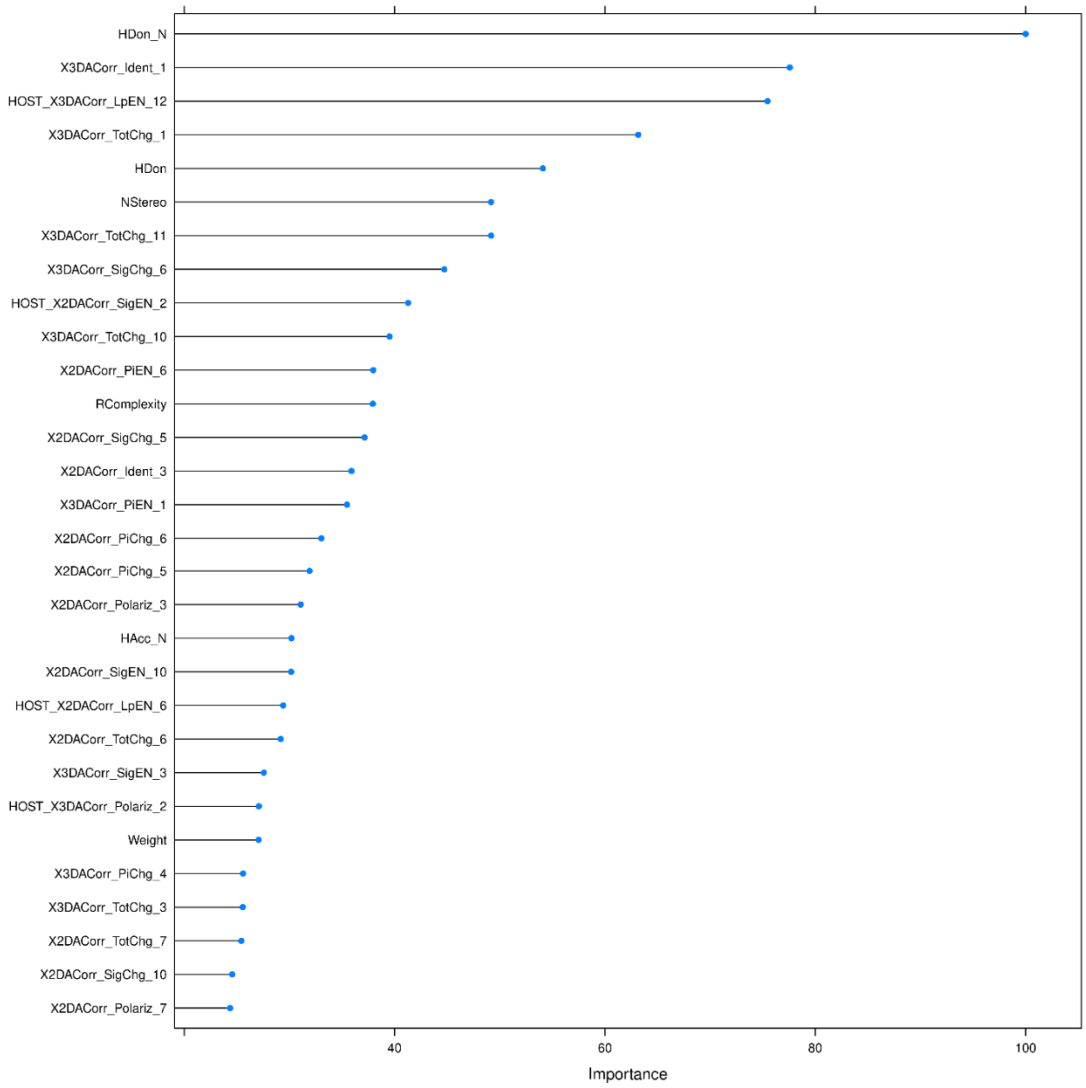


Figure 71: Most useful variable of the chosen NNET for the prediction of the training set for the *nnet* function from the *caret* package

For the NNET, the model uses both variables describing the guest and variables describing the host, and some of the most important variables are known to be correlated with the binding free energy (hydrogen bonds donor or acceptor). All measures of importance are scaled to have a maximum value of 100. The global performance of the training set and the test set of the NNET model can be visualized in the following Figure 72:

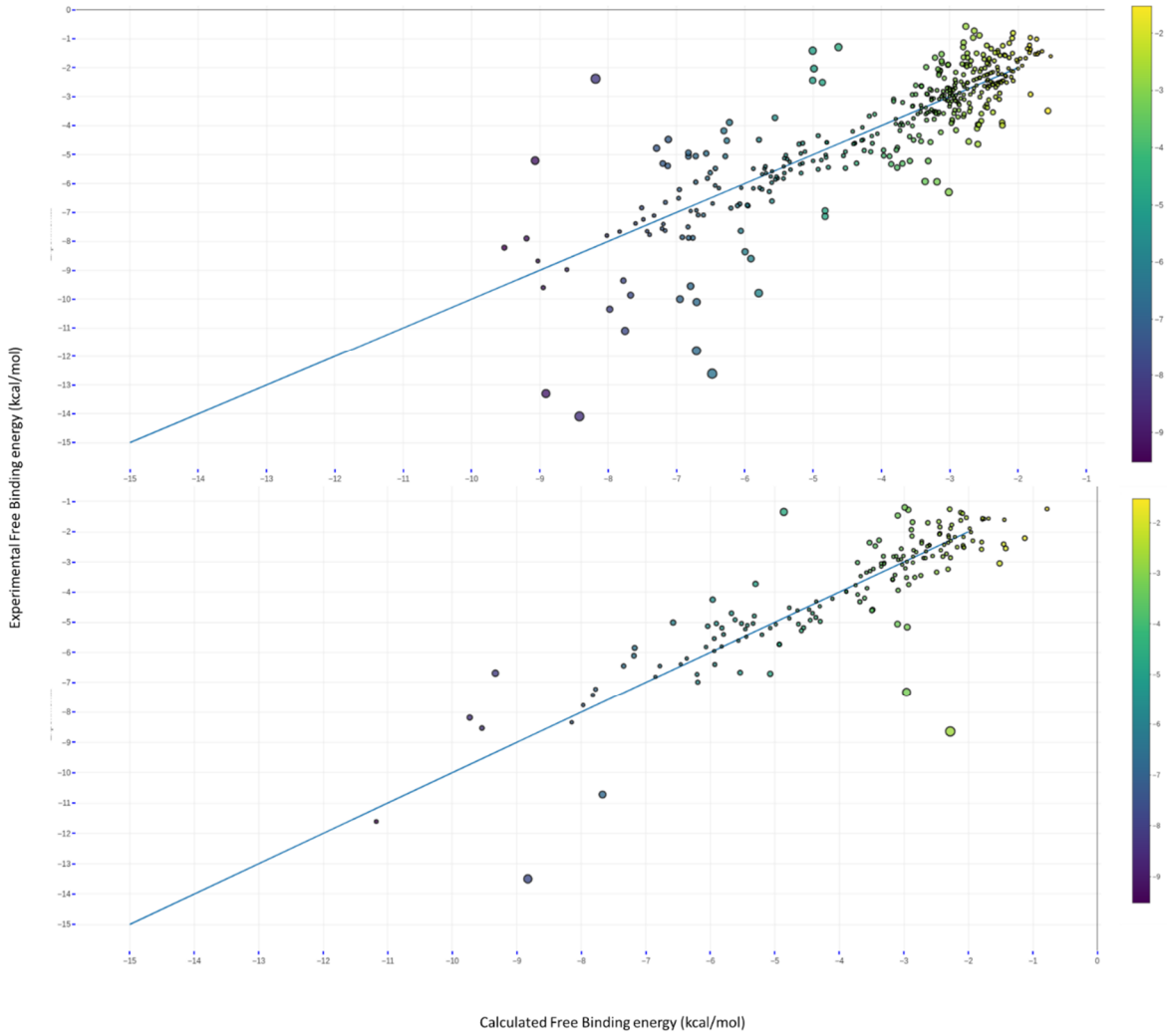


Figure 72:(Top) training-set prediction of the NNET; (Bottom) Testset prediction of the NNET. The points are coloured by the binding free energy, while the size is a function of the error of prediction (the smaller the points are and the smaller is the error).

The NNET shows very impressive performances: in the training set, some guest presents a more important error while we reach the most negative values of the binding free energy. It could be explained by the distribution of the score inside the model. As we show in the presentation of the machine learning model (chapter 03), the relative number of very negative binding free energy is low compared to the one presenting medium-range values. Explaining why the model may have some troubles predicting the compounds with binding free energy below -10kcal/mol. While for the others, they are predicted with an error that could be considered as important (+ 4kcal/mol), but in most of the cases, they belong to the most negative prediction for the model, suggesting that the NNET is capable of identifying the most negative compounds. The main problem here is there is no way to discriminate the well-predicted compound from the underpredicted compound based on the prediction of the machine

learning model. The only solution is to be cautious when the predicted values are outside the range of the training set (i.e.-9 kcal/mol).

II. B. 4. A. (II) - RANDOM FOREST (RF)

In a similar way, as we have done for the previous NNET model, we tried to improve the tunes parameters for the RF model as well. In order to define the best model possible in terms of performances (results & computational time), several tuning parameters are tested: (i) the number trees (*ntree*), (ii) the number of tested variables at each division (*mtry*). In our case, the results show that the optimum number of trees is 1000, and the *mtry* value is 10 (Figure 73).

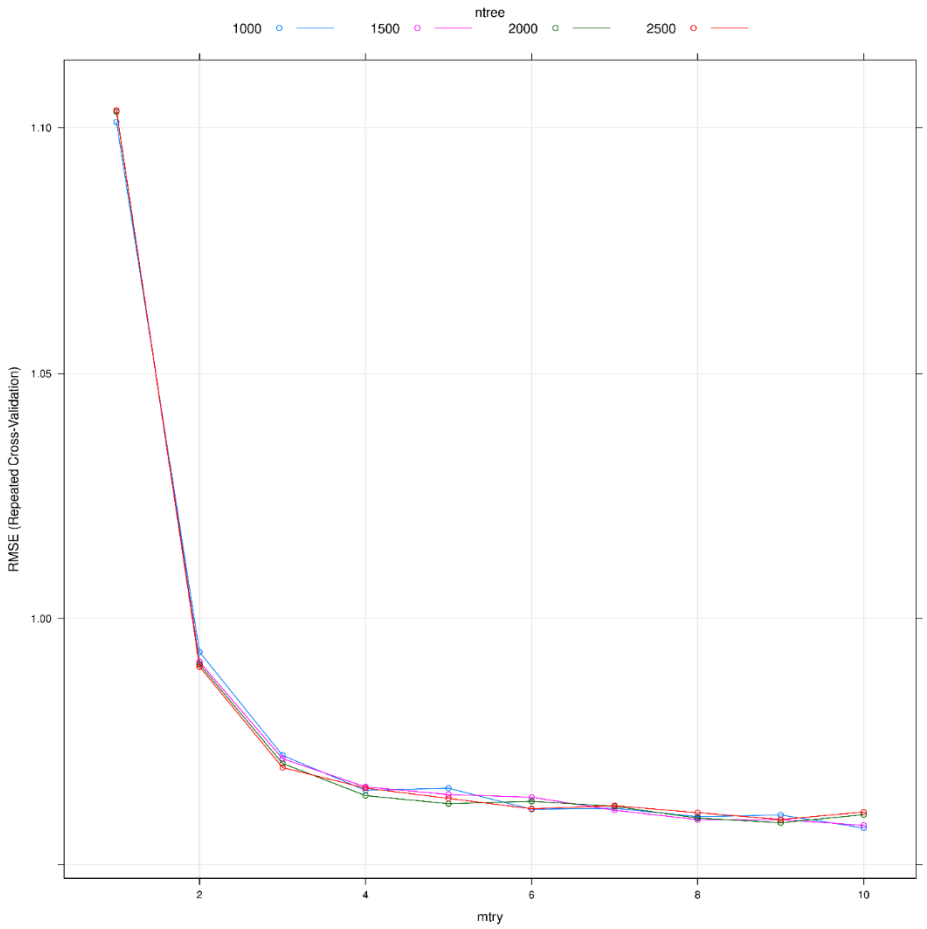


Figure 73: Performances of the RF using a combination of a number of trees (ntree) and a number of tested variables (mtry) to find the best performances for the model.

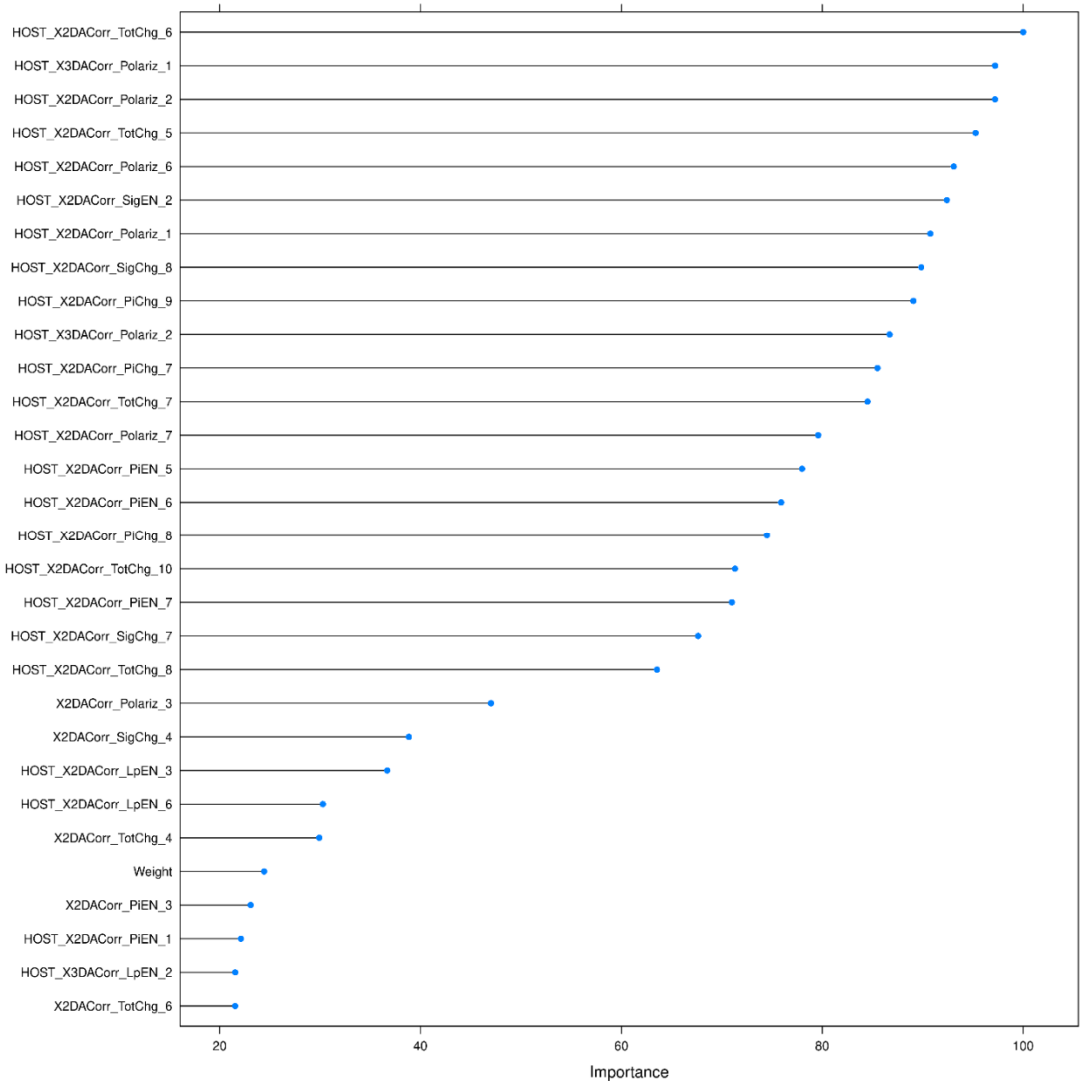


Figure 74: Most useful variables of the chosen RF for the prediction of the training set for the rf function from the caret package

Unlike the previous case, this time, almost all the most useful variables are from the host. In fact, out of the top 30 variables, only 6 (and not even the most important ones) describe the guest molecule. This was unexpected because, in theory, the host variables represent only 25% of the total number of variables. The variable importance algorithm for the RF is computed on the out-of-bag data for each tree, and then the same is computed after permuting a variable. The differences are then averaged and normalized by the standard error.

The global performance of the training set and the test-set of the RF model (using the *rf* function from *caret*) can be visualized in the following Figure 75.

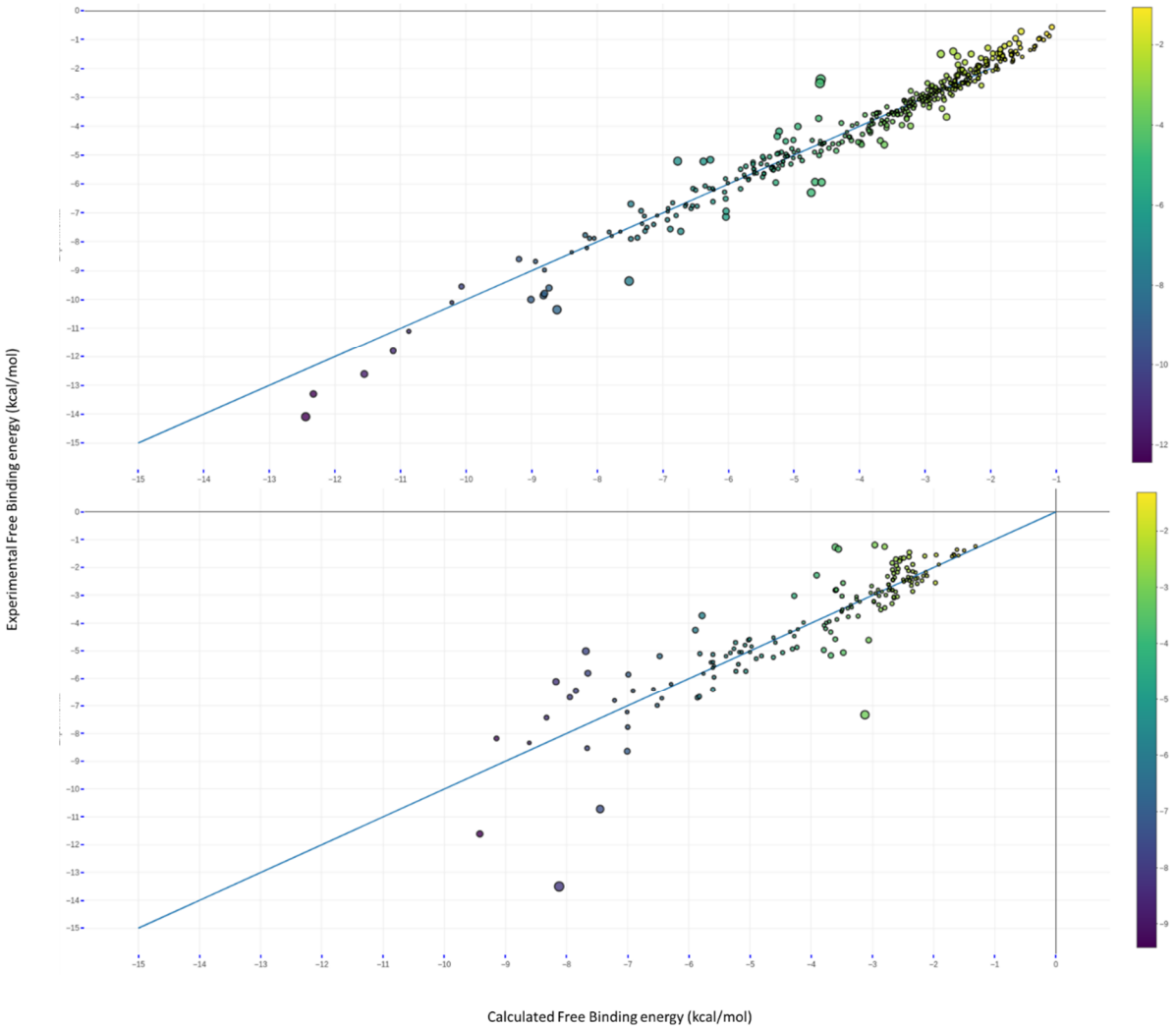


Figure 75: (Top) training-set prediction of the RF. (Bottom) Test-set prediction of the RF. The points are coloured by the binding free energy, while the size is a function of the error of prediction (the smaller the points are and the smaller is the error).

Compared to the NNET prediction, the prediction on the training set is impressive but shows some overtraining, depending on the features selections or the way the model is split into test-set and training-set. Still, we saw the same problem as stated before concerning the prediction of the most negative compound.

II. B. 4. A. (III) - POLYNOMIAL SVM

For Polynomial SVM, three tuning parameters are investigated: (i) the cost penalty (Cost), (ii) sigma value (polynomial degree), and (iii) the C value (scale).

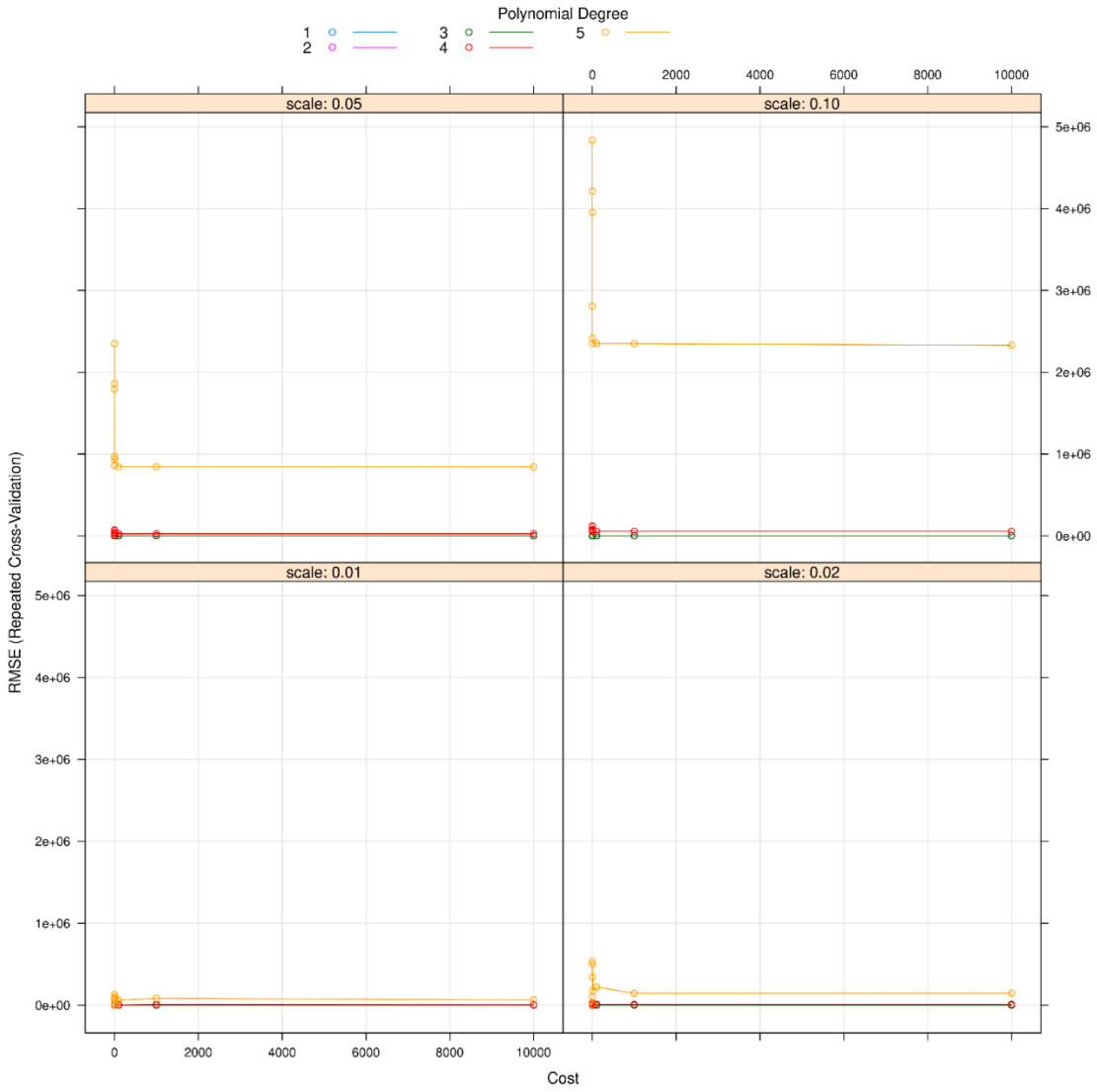


Figure 76: Performances of the *SVM* using a combination of internal parameters (C, sigma, and the cost) to find the best performances for the model.

The analysis of the training-set / test set (Figure 77) shows a more approximate prediction for both sets, indicating that the SVM algorithm performs worse than the other two algorithms. In that case, we found a similar problem regarding the prediction of very negative values, but this time, we have a verticalization of the prediction, obtaining bad results when the prediction approaches negative values, regardless of the molecular descriptors in use. On the lower end, the model is converging to a minimum value of ~ -9 kcal/mol both in the training and the test set.

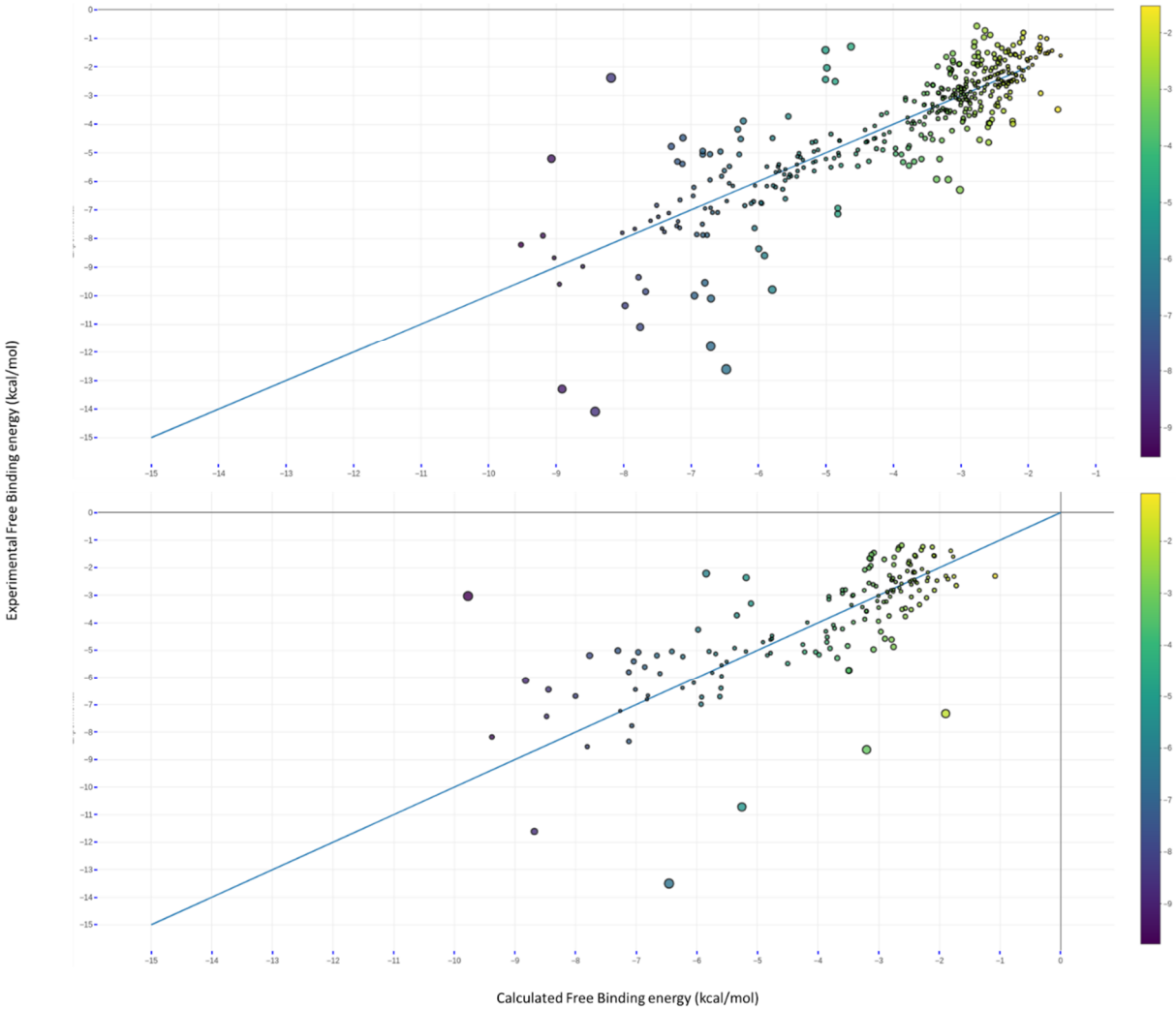


Figure 77: (Top) training-set prediction of the polynomial SVM. (Bottom) Test-set prediction of the SVM. The points are coloured by the binding free energy, while the size is a function of the error of prediction (the smaller the points are and the smaller is the error).

II. B. 4. A. (IV) - K-NEAREST NEIGHBOURS (KNN)

The last and final machine learning model we tried for the prediction of the binding free energy is the simplest one. Only one tune value is investigated for the improvement of the prediction: the number of neighbours. The optimal number of neighbours is four (Figure 78).

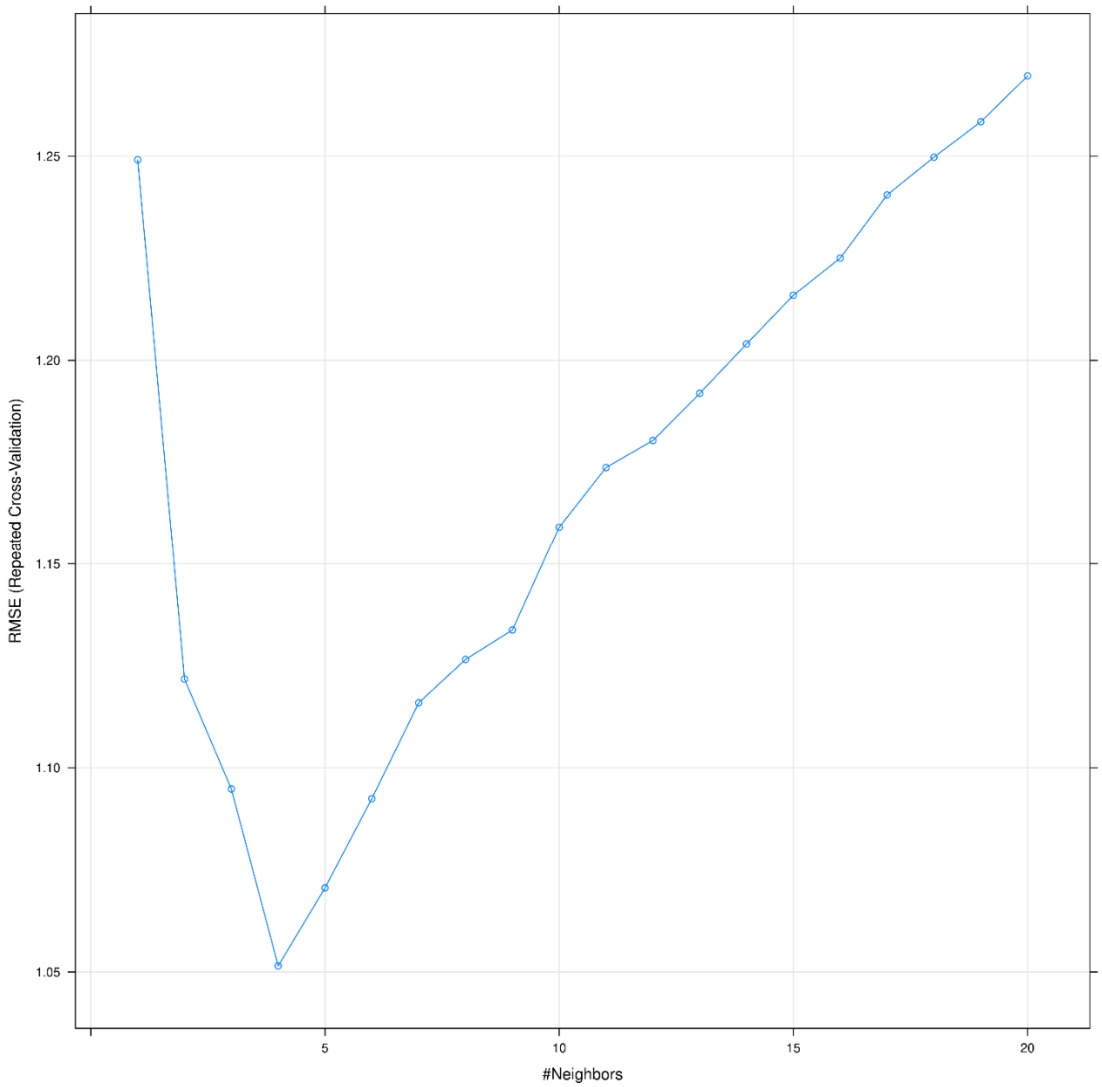


Figure 78: Performances of the Knn function using different neighbours number to find the best performances for the model.

The visualization of the training-set / test-set of the Knn model (Figure 79) shows similar results as the SVM model concerning the prediction of the top-scored compound with one difference: while the overall prediction on the training set is in good agreement with the experimental data (RMSE = 1.05 kcal/mol), the test-set presents a larger error than the training set (RMSE = 1.46 kcal/mol) associated with a very bad prediction of the top-scored guests, for which the error can be up to 7 kcal/mol.

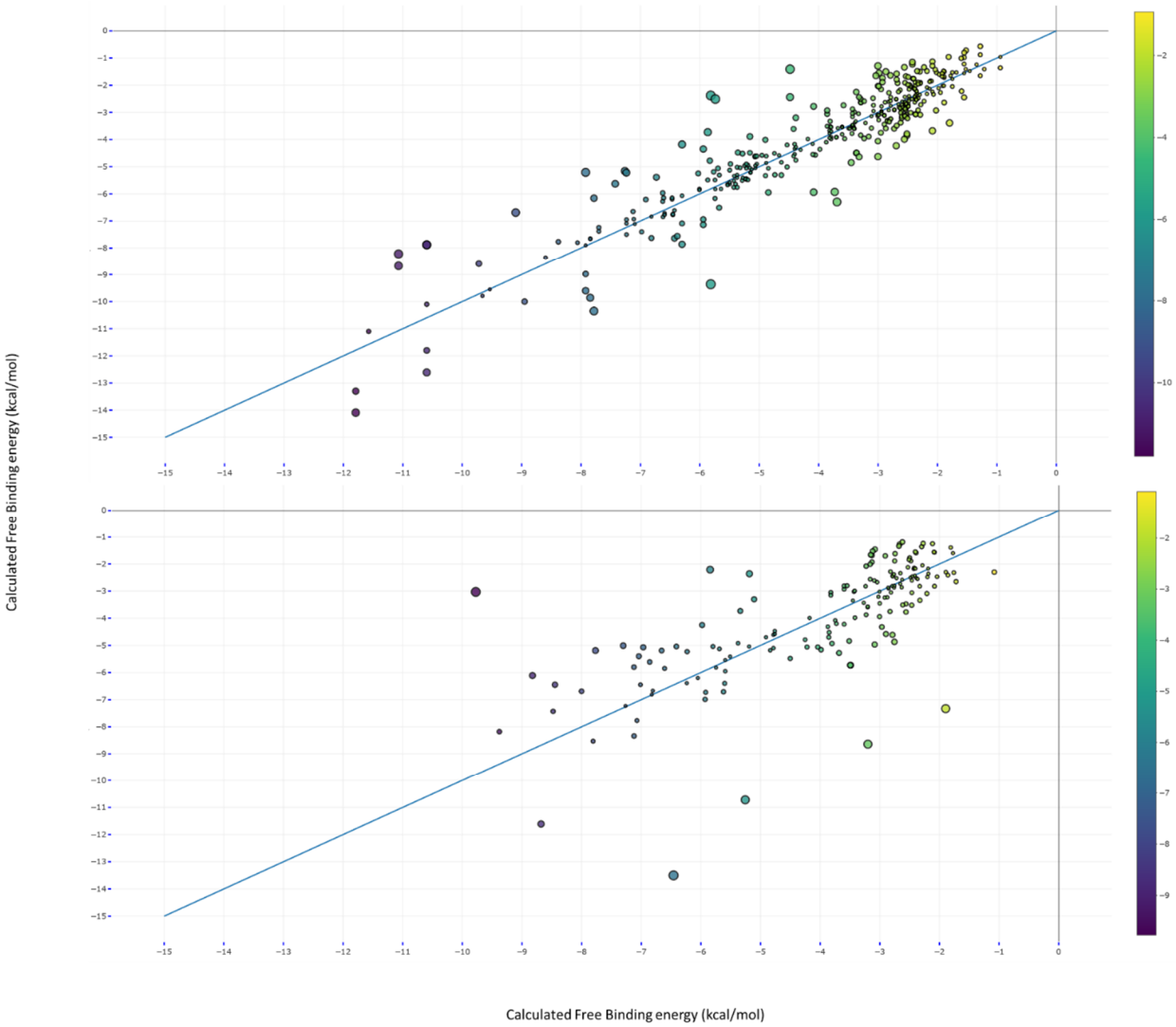


Figure 79: (Top) Training-set prediction of the Knn. (Bottom) Test-set prediction of the Knn. The points are coloured by the binding free energy, while the size is a function of the error of prediction (the smaller the points are and the smaller is the error).

The statistical analysis for the best machine learning models is presented in Table 13 and Table 14 and helps selecting the model that will be used for the prediction. The predictions for the test set are particularly relevant at this point.

Table 13: Statistical analysis of the Training-set for the best model of each of the machine learning algorithms

TrainRMSE	TrainRsquared	TrainMAE	Method
1.07	0.77	0.74	NNET
0.95	0.80	0.64	RF
1.59	0.57	1.02	SVM
1.05	0.77	0.73	Knn

Table 14: Statistical analysis of the Test-set for the best model of each of the machine learning algorithms

RMSE	Rsquare	MAE	Method
1.05	0.74	0.64	NNET
0.94	0.79	0.58	RF
1.46	0.52	0.94	SVM
1.46	0.52	0.94	Knn

Concerning the prediction with the *SVM*, this model represents the worst agreement with the experimental data looking at the training set. The values of the training set and the test set are similar, suggesting that the overtraining of the model is controlled. On the contrary, the Knn model presents a clear overtraining with a much lower RMSE and MAE in the training set compared to the test set, explaining the difference in terms of prediction for the negative values.

The two best models, the RF and the NNET are comparable with a very slight advantage of the *rf* model (with -0.09 and -0.06 respectively for the MAE and the RMSE of the test-set). But from the above-described analysis of the variables, the NNET function uses more variables dedicated to the guests than the hosts while the situation is reversed for the rf function. Considering the nature of the challenge, we reasoned that having more information about the guests should be beneficial, as it would help differentiate better different guests binding to the same host. For this reason, and as the difference in performance between the RF and the NNET functions was slight, we decided to use the NNET for the prediction of the SAMPL8-CB[8] dataset.

II. B. 5 - RESULTS:

II. B. 5. A - APPLICABILITY DOMAIN

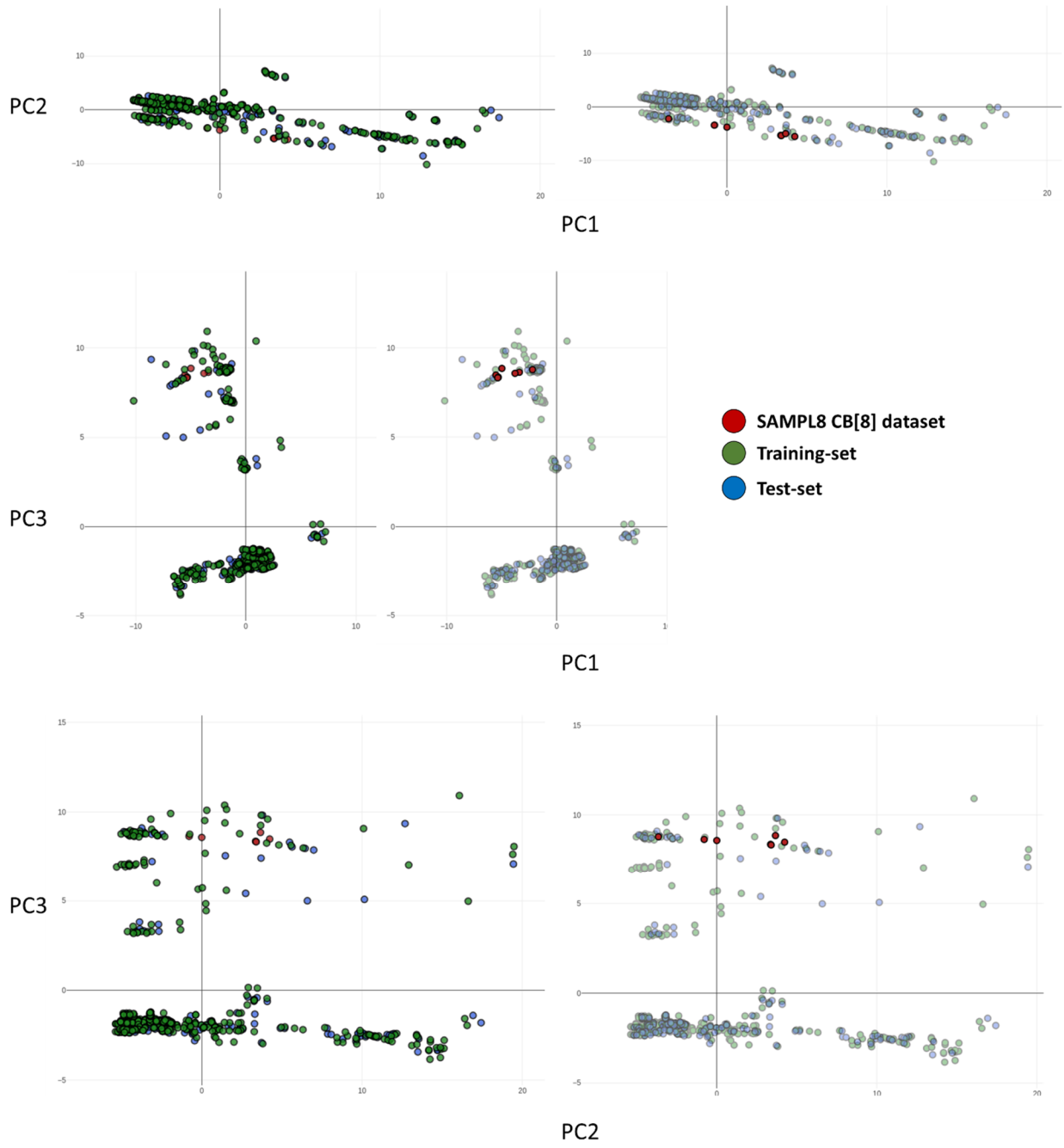
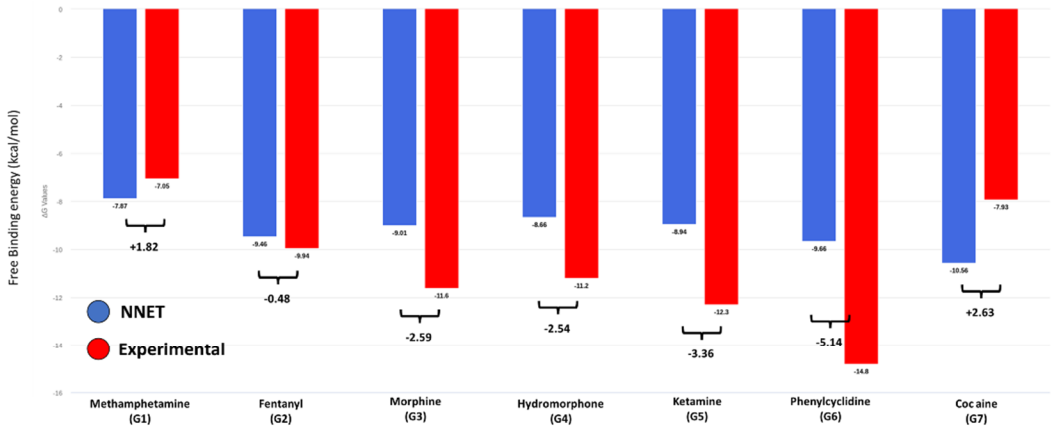


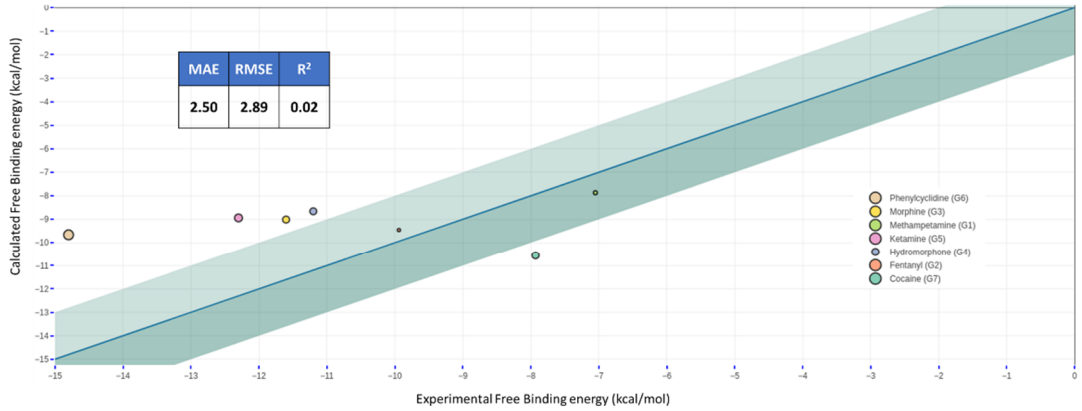
Figure 80: Visualization of the chemical space of the predicted set compared to the test set and the training set. If the new host-guest system descriptors overlap with the chemical space of the model, we can consider they are sampling a relatively similar space and fulfil the requirements for the prediction using our machine learning method.

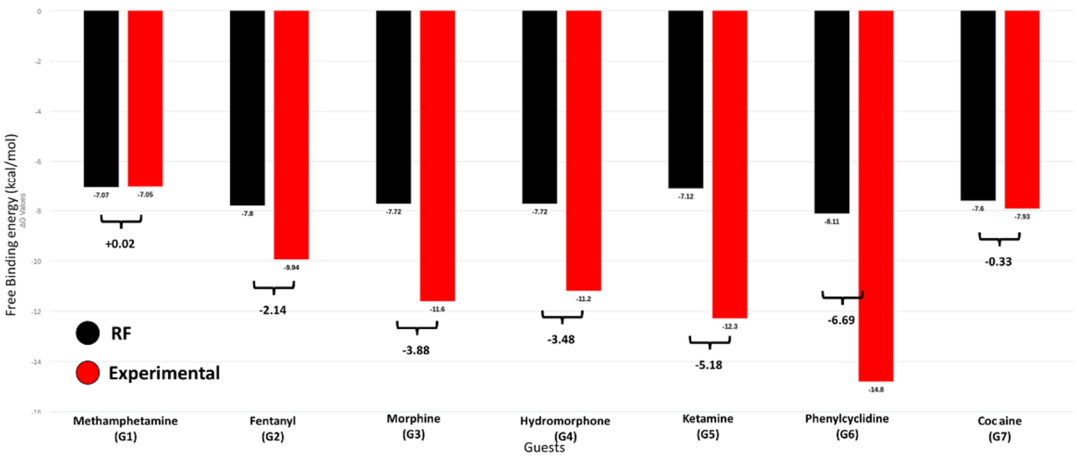
In order to verify if we are in the applicability domain of the machine learning model, we need to visualize the physicochemical space formed by the combination of the generated molecular

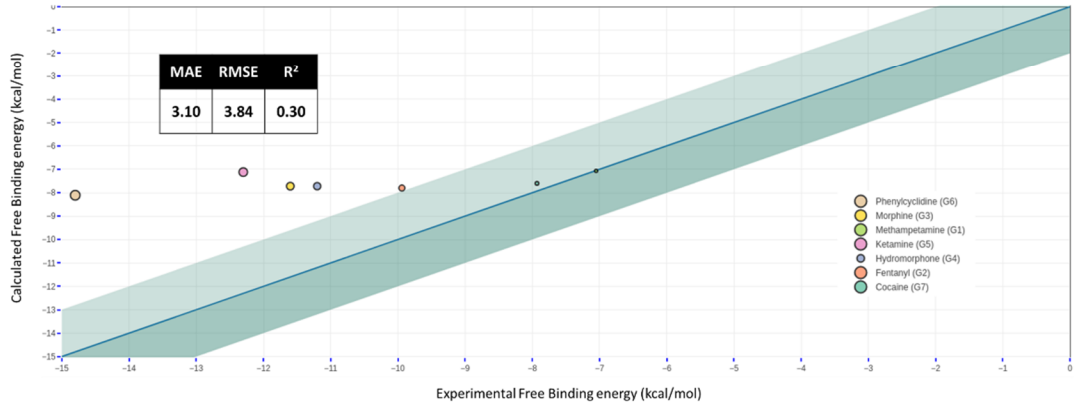
descriptors in a way to verify that the sampled space is similar between the model (split in training-set and test-set in Figure 80) and the molecules that have to be predicted. This figure is separated into three different parts corresponding for each of them to a specific part of the chemical space. Together, the three dimensions represent 63% of the variability of the dataset (30.6% for PC1, 18.09% for PC2, and 14.28% for PC3). Components with lower percentage variance (<10%) were visualized until 80% of the variability could be captured but are not shown for clarity. As the CB[8] dataset is composed of seven guests, there are seven red points that are represented in their space for each combination of PCs (on the left) and highlighted in the same space for visualization (on the right). As shown, the red points (representing the SAMPL8-CB[8] guests) overlap with the other samples in all the dimensions of the space, leading us to think that they are within the scope of the model. While this does not guarantee the quality of the prediction, the opposite is true (i.e. being outside the scope of the model guarantees unrealistic predictions). For that reason, the verification of the chemical space of the predictors is a minimal requirement. This procedure was done for all the molecules, even if only presented once.

II. B. 5. B - PERFORMANCES OF THE ML MODELS ON THE SAMPL8 CHALLENGE









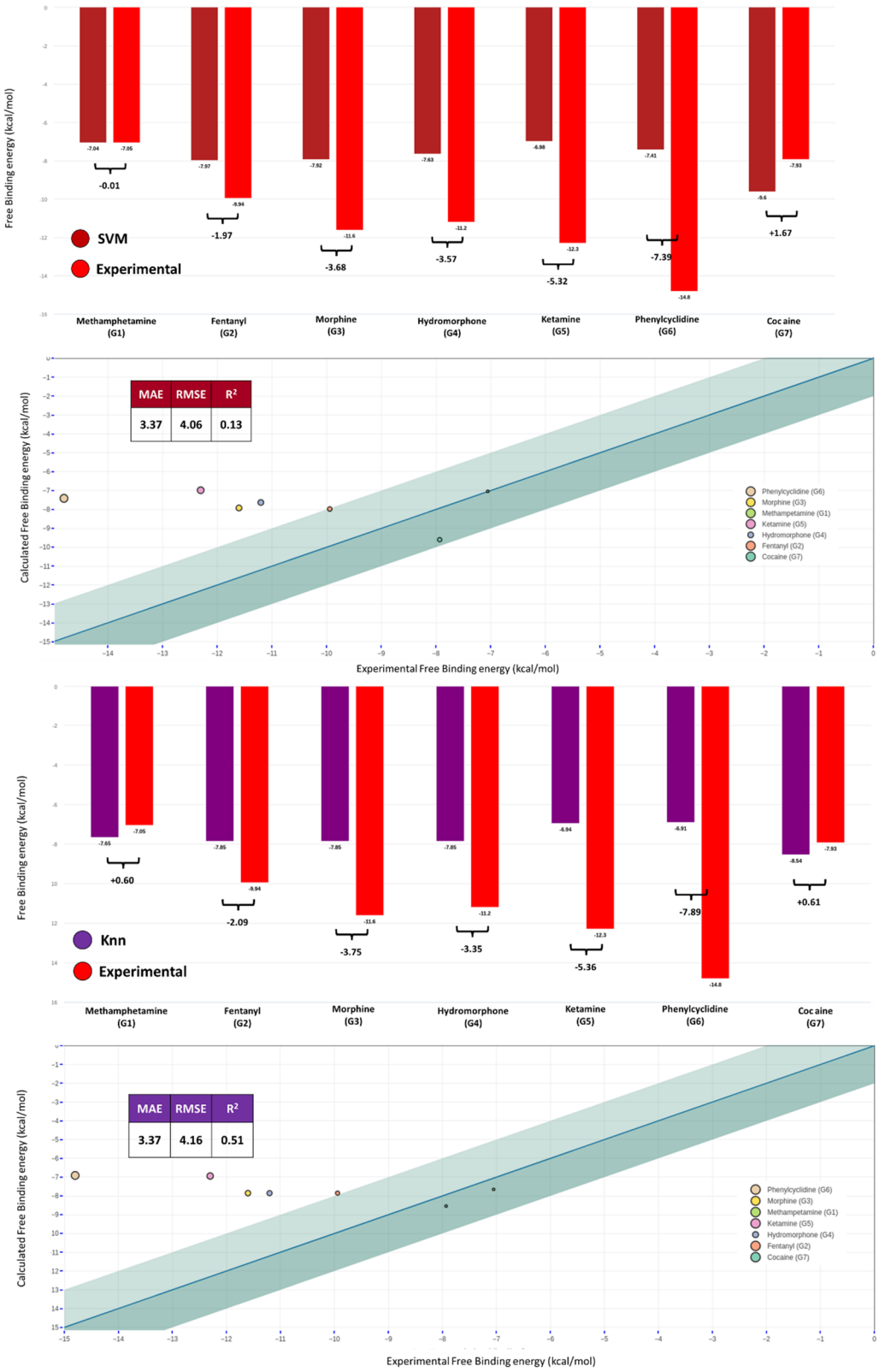


Figure 81: Comparison of experimental binding free energies and predicted values. Histograms of binding free energy coloured by the machine learning model used: in red the experimental data, and in blue the *nnet* prediction, in dark red the *SVM* prediction, in black the *RF* prediction, and in purple the Knn prediction. All of the histograms are associated with a correlation plot: the green-shaded area represents a threshold of +2/-2 kcal/mol from the experimental energy. The statistical analysis of the prediction compared to the experimental value is shown in the respective coloured box.

The prediction was run in the four different machine learning models analyzed in the previous part, but only the NNET was selected for submission considering the SAMPL8 challenge. All the machine learning models present a better overall performance than the thermodynamic-based method, and the NNET was the best performing one. The machine learning methods, compared to the thermodynamic-based methods, are underestimating most of the guest binding free energy. That was already discussed in the validation part, but considering the low amount of very negative binding free energy data (< -10 kcal/mol), it is not surprising that the model converged to [-9 – -10] kcal/mol and could not predict a -14.8 kcal/mol. Furthermore, most of the predicted guests were in the range where the prediction was uncertain.

Like in the thermodynamic based method, based on the accuracy of the prediction (Figure 81A), the results can be separated into three groups:

- (i) The first group represent the prediction with an excellent agreement with the experimental data (< 2 kcal mol) represented by the green shaded area. Two guests have been predicted with an excellent agreement: Methamphetamine (G1, +0.82 kcal/mol) and Fentanyl (G2, +0.48 kcal/mol).

- (ii) The second group represent the incorrect prediction but are still in the range from the experimental values (~2 to 4 kcal/mol errors). Four host-guest complexes have been predicted in the range of the experimental values, and from these four host-guest complexes: three of them were underestimated: Morphine (G3, -2.59 kcal/mol), Hydromorphone (G4, -2.54 kcal/mol), and Ketamine (G5, -3.36 kcal/mol), while Cocaine (G7 +2.63 kcal/mol) was overestimated. For almost all of them, the error is below 3 kcal/mol.

- (iii) The third group represent the prediction with large errors (> 4 kcal/mol). With the NNET, only one guest was presenting a bad prediction: the Phenylcyclidine (G6, -5.14 kcal/mol) presents an important underestimation. As explained, this can be rationalized on the basis the only one compound in the training set had a value of -14kcal/mol.

II. B. 5. C - PERFORMANCE OF THE CB[8] MODEL

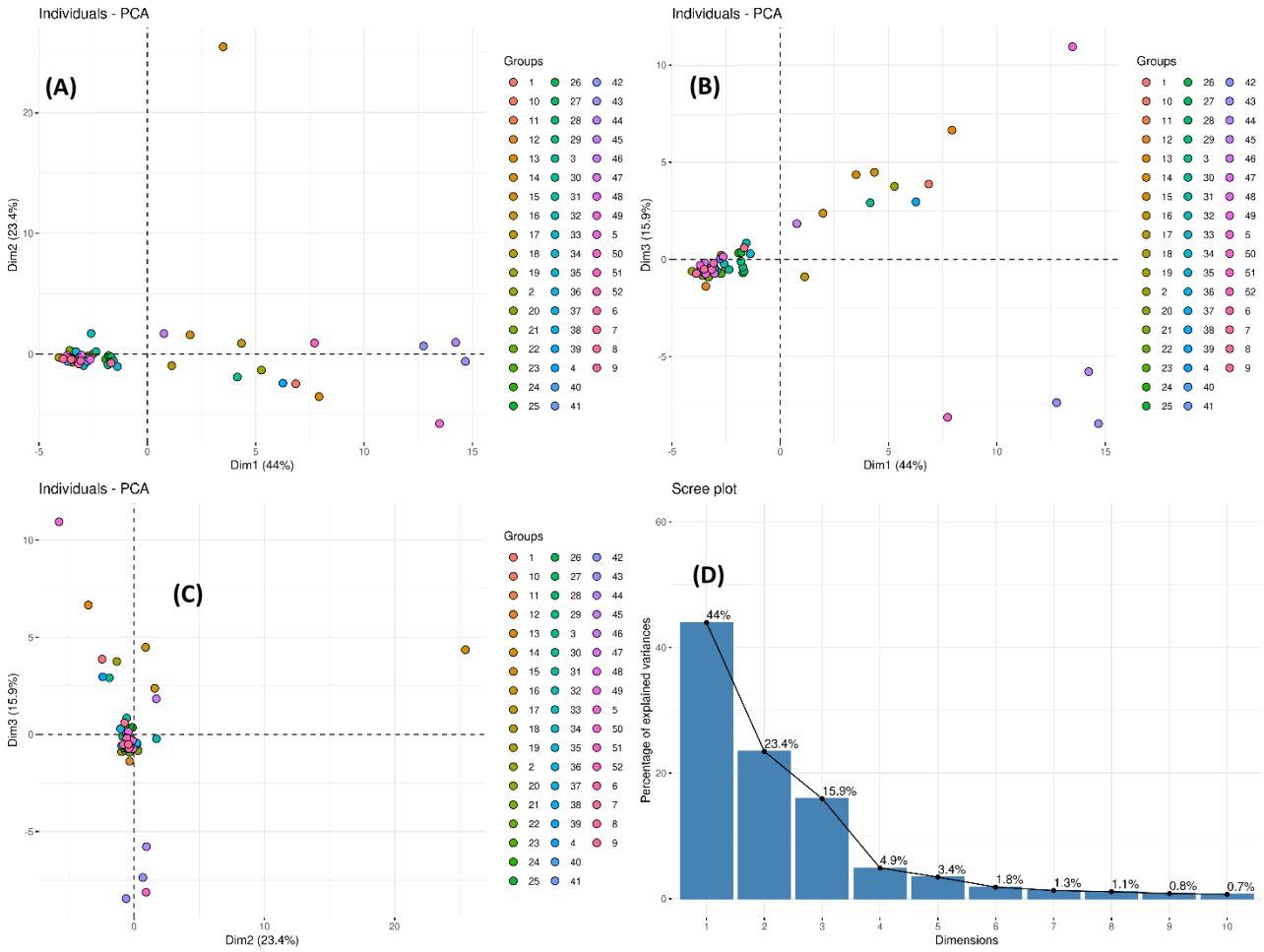


Figure 82: PCA of the Host-Guest chemical space using a reduced set of molecular descriptors generated with CORINA. In (A), the space formed by the combination of PC1 and PC2 explains respectively 30.8% and 17.8% of the variability. In (B), the space formed by the combination of PC1 and PC3 explains respectively 30.8% and 14.1% of the variability. In (C), the space formed by the combination of PC2 and PC3 explains respectively 17.8% and 14.1% of the variability. In (D), the scree-plot represents the variability of all the principal components of the analysis. The molecules are coloured by the system there are supposed to interact with and, their size is a function of their binding free energy.

The CB[8] machine learning model differs from the global approach by the training dataset. For this one, data corresponding to the CB[8] model were extracted from the global dataset, and an independent model was constructed using the same methodology. The CB[8] dataset is composed of 52 binding free energy of guest molecules interacting with the CB[8] host, 34 came from the BindingDB with the code "BDBM36284", and 14 came from the SAMPL6 challenge. For the CB[8] model, the host-descriptors are not required since the model is trained and expected to predict binding data on the same host. However, this is a local model, only

suitable to predict compounds interacting with the same host. Although, in theory, this approach is expected to generate more accurate results for a host for which data is available in the literature, but the amount of data can be a limiting factor in prediction, as we saw in the previous challenge (SAMPL7).

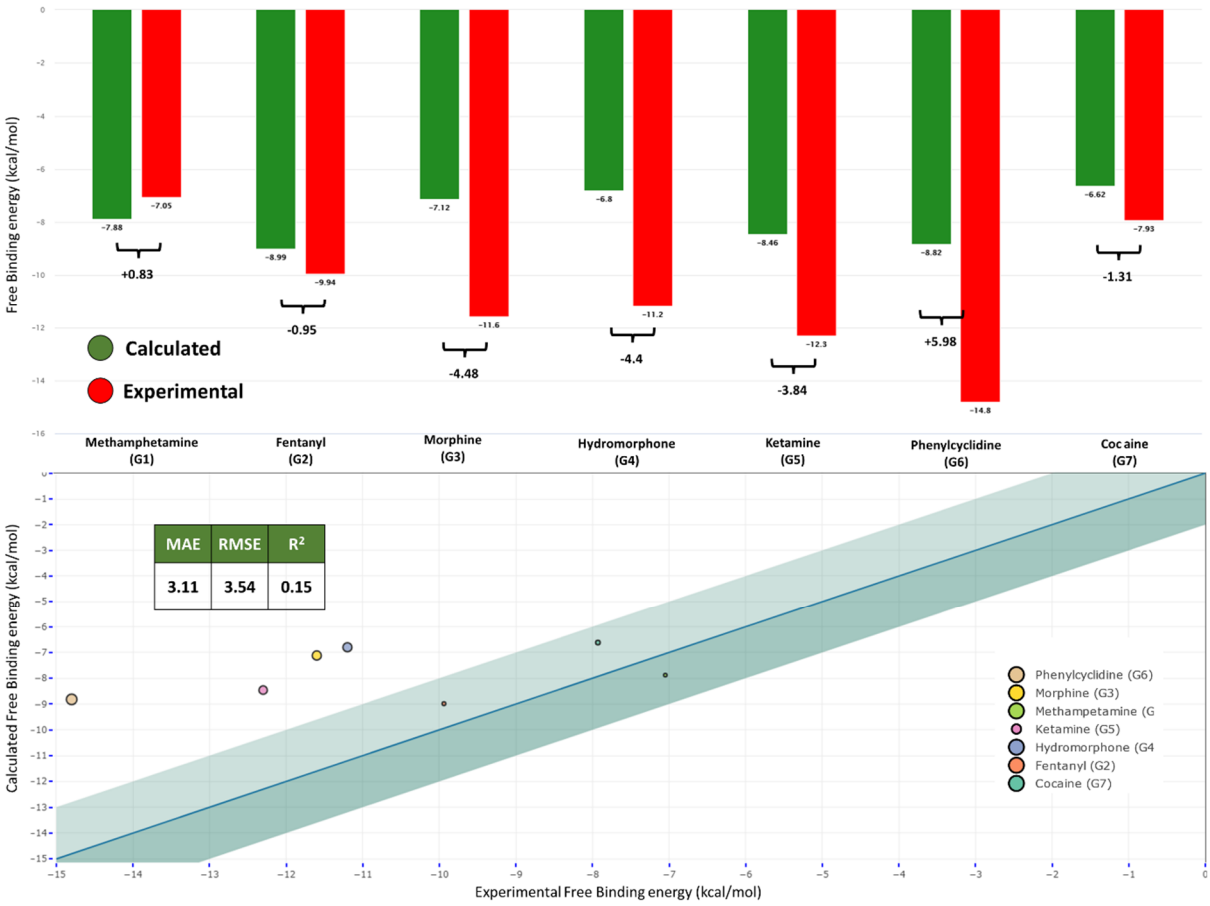


Figure 83: Comparison of experimental binding free energies with predicted values. (Top) histogram of binding free energy coloured by the origin of the data: in red the experimental data and in green the predicted binding free energy using *nnet* machine learning on the CB[8] dataset. (Bottom) correlation plot: the green-shaded area represents a threshold of +2/-2 kcal/mol from the experimental energy. The statistical analysis is shown in the green box: with MAE = 3.11, RMSE = 3.54, and R^2 = 0.15.

Overall, the CB8 model performs worst than the global model. This can be explained by the difficulty of the model to predict very active compounds. The range of the experimental values in the model was [-4.77 -13.5], but only a few compounds present energy below -10 kcal/mol. Concerning the predictions, it is interesting to note that despite the globally worse prediction (considering RMSE, MAE values) of the CB8 model, the two molecules Methamphetamine (G1) and Fentanyl (G2) were also correctly predicted. Compared to the global model, we find a more important underestimation of the group composed of Morphine (G3), Hydromorphone (G4), Ketamine(G5), and Phenylcyclidine (G6), but in this subgroup, these are ranked

correctly. Cocaine (G7), which was overestimated in the two previous approaches, is correctly predicted here (-1.31).

And finally, the prediction can be separated into two groups:

- (i) the first group is composed of the three well-predicted compounds: Methamphetamine (G1, +0.83 kcal/mol), Fentanyl (G2, -0.95 kcal/mol), and Cocaine (G7, -1.31 kcal/mol).

- (ii) for the second group, despite the fact that the ranking between these compounds in the subgroup is correct, it is composed of four guests, all predicted with large error (> ~ 4 kcal/mol): Morphine (G3, -4.48 kcal/mol), Hydromorphone (G4, -4.4 kcal/mol), Ketamine (G5, -3.84 kcal/mol), and Phenylcyclidine (G6, -5.98 kcal/mol).

II. C - SAMPL8 GDCC CHALLENGE

II. C. 1 - METHODS USED

The SAMPL8-GDCC challenge is a challenge that uses several Gibbs cavitand with a dataset of five guests for which the binding free energy must be predicted in association with two different hosts: the TEMOA host and the TEETOA host (Figure 47).

We studied the Gibbs cavitand during the SAMPL7 challenge, obtaining good results with the first approach of our global machine learning model. In addition, we had some difficulties with the thermodynamic approach concerning this family of host-guest complexes. The complexity of the modelling of the cavitand using the GFN2B-xTB may be due to the charges of the host system (-8), as we encounter problems with charged entities in the retrospective analysis of the SAMPL6 challenge concerning the CB[8] host (Figure 62). For these reasons, and because at the moment of the project, in relation to the timeframe of the challenge, it was not feasible to realize all the calculations needed for the thermodynamic methods, we decided to make a submission applying only the global machine learning approach.

In a similar way to the SAMPL7 challenge, the TEETOA is shown for the first time in the challenge, and there is no pre-existing data in the literature for this specific system. But it belongs to the family of Gibbs cavitand, for which we have lots of data. Considering their similar scaffold, we made the assumption that our Host-Guest machine learning model will understand the chemical specificity of these host-guest complexes and will be able to discriminate the TEETOA system from the TEMOA system.

For the prediction, the dataset for the global machine learning approach has been expanded, and the results from the previous SAMPL7 and SAMPL8-CB[8] were added. The dataset is now composed of ~600 features with 27 different host systems. As we already explained, when new molecules are added to the dataset, all the feature selection processes have to be redone.

II. C. 2 - RESULTS

The outcome of the machine learning results compared to the experimental value is presented in Figure 84:

Figure 84: Comparison of experimental binding free energies with predicted values. (Upper) Histogram of binding free energy coloured by the system: in red the experimental data, in purple the prediction on the TEMOA system, in blue the prediction of the TEETOA system. (Bottom) All of the histograms are associated with a correlation plot: the green-shaded area represents a threshold of +2/-2 kcal/mol from the experimental energy. The statistical analysis of the prediction compared to the experimental value is shown in the respective colored box: in purple the prediction concerning the *TEOMA* system (MAE =1.27, RMSE =1.64, R^2 = 0.001), in blue the prediction concerning the *TEETOA* system (MAE =2.74, RMSE =2.83, R^2 = 0.003), and in black the global prediction (MAE =1.92, RMSE =2.25, R^2 = 0.54)

Unlike the two previous cases, the range of the model included the experimental values. Concerning the prediction, there is a difference between the two systems: for the TEMOA

system for which some data are available, 4 of the 5 molecules are predicted with less than 2 kcal/mol error. While for the TEETOA system, only one molecule is predicted with less than 2kcal/mol of error. The global model also differentiates between the TEMOA and TEETOA systems: We can see in Figure 84 that the TEMOA system is in general underestimated while the TEETOA system is overestimated, meaning the ML model is treating both systems differently.

The prediction can be separated into groups:

- (i) The first group represent the prediction with an excellent agreement with the experimental data (< 2 kcal mol) represented by the green shaded area. Four guests have been predicted with an excellent agreement: TEMOA-G1 (-0.75 kcal/mol), TEMOA-G3 (+0.03 kcal/mol), TEMOA-G4 (-1.06 kcal/mol), TEMOA-G5 (-1.34 kcal/mol)

- The second group represent the incorrect prediction but are still in the range from the experimental values (~2 to 4 kcal/mol errors). Five host-guest complexes have been predicted in the range of the experimental values: TEETOA-G1 (+2.74 kcal/mol), TEETOA-G2 (+1.72 kcal/mol), TEETOA-G4 (+2.8 kcal/mol), TEMOA-G2 (-3.18 kcal/mol) and TEETOA-G5 (+3.72 kcal/mol).

In conclusion, for the SAMPL8-GDCC prediction, we were able to use the global machine learning approach to predict 4/9 molecules with an excellent agreement with experimental data while 5/9 molecules are predicted with larger errors, but perhaps still useful. Additionally, none of them was predicted with a large error (> 4 kcal/mol).

III - SOLVENT EXCHANGE ANALYSIS IN CALIX[4]PYRROL CAPSULE:

III. A - PRESENTATION OF THE HOST SYSTEM

For this analysis, two host complexes with solvents were provided by one of the partners of the NOAH project: the ICIQ (Institut Català d'Investigació Química). They are derivate from calix[4]pyrrole, and the structure are presented in Figure 85:

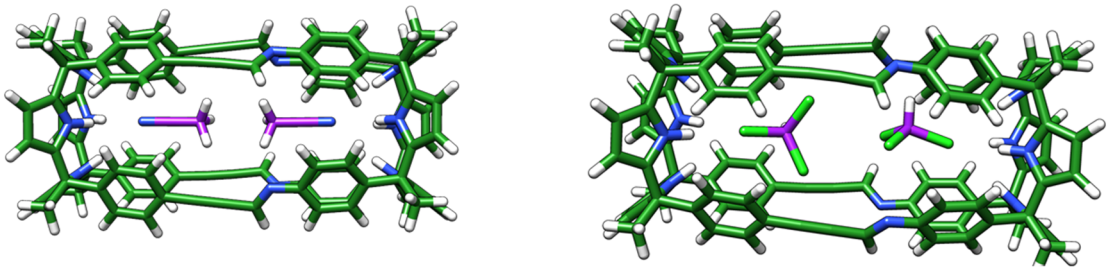


Figure 85: Presentation of the initial 3D structure that was used for solvent exchange analysis: (Left) the molecule in complex with two molecules of acetonitrile; (Right) the molecule in complex with two molecules of chloroform.

The capsules were synthesized from two calix[4]pyrrole joined by four dynamic covalent bonds (i.e., imines). The idea was to study the behaviour of these molecular cages under different stimuli in order to determine whether:

- In chloroform, if the capsule is likely to contain two molecules inside its cavity, and investigate the inclusion process.
- In a chloroform/acetonitrile (9/1) mixture, analyze the competition between the solvent molecules, and determine which chloroform or acetonitrile is more likely to be present in the host cavity. Particular attention was paid to the possible interactions of acetonitrile with the NH of pyrroles inside the cavity.

In order to simulate the cage in two different solvated environments, the system is prepared following the HG-DYNAusor platform using modules one and two. As there is no guest to be considered the modules three and four are ignored. The simulations on the optimized structure are done at two different levels: classical molecular dynamic simulations (i) in explicit chloroform solvation, (ii) in a mixture of chloroform/acetonitrile (9/1) and using semi-empirical molecular dynamic simulations: (iii) in implicit chloroform solvation starting from the initial configuration with two chloroform inside the cavity, and (iv) in explicit solvation

considering a mixture of chloroform/acetonitrile (9/1) in an implicit chloroform environment. The results of the simulations are presented in the next part.

III. B - SIMULATIONS OF THE HOSTS:

III. B. 1 - CLASSICAL MOLECULAR DYNAMIC SIMULATIONS

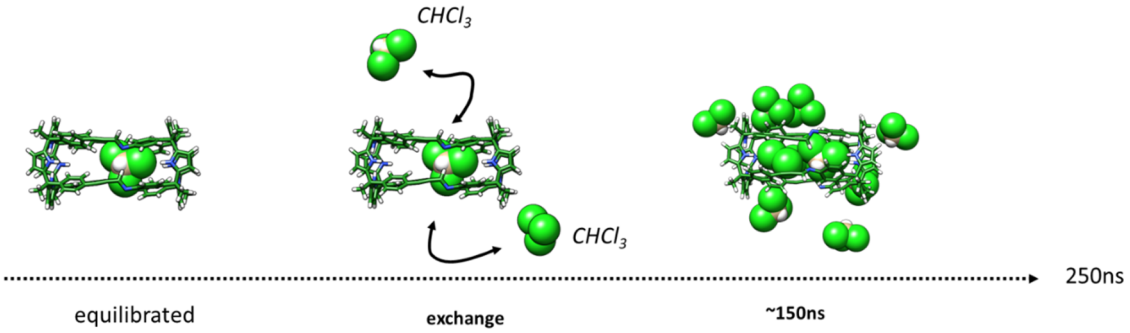


Figure 86: An overview of the molecular dynamics of the system in the explicit chloroform ($CHCl_3$) model. The graphics are separated into three parts: (i) in the left, we have the equilibrated system representing the first frame of the dynamic, (ii) in the middle, after a few ns of simulations, the first exchange between two molecules of solvents inside the cavity of the host appears, and (iii) after 150ns another chloroform molecule enter in the cavity.

As shown in Figure 86, the simulations of the host starting without any solvent molecules inside present right after the equilibration of one chloroform molecule inside the cavity. The chloroform molecules present a free rotation inside the cavity during the simulations, and we expect the rotation to have an impact on the exchanges processes.

III. B. 2 - SIMULATION IN CHLOROFORM/ACETONITRILE (9/1) MIXTURE

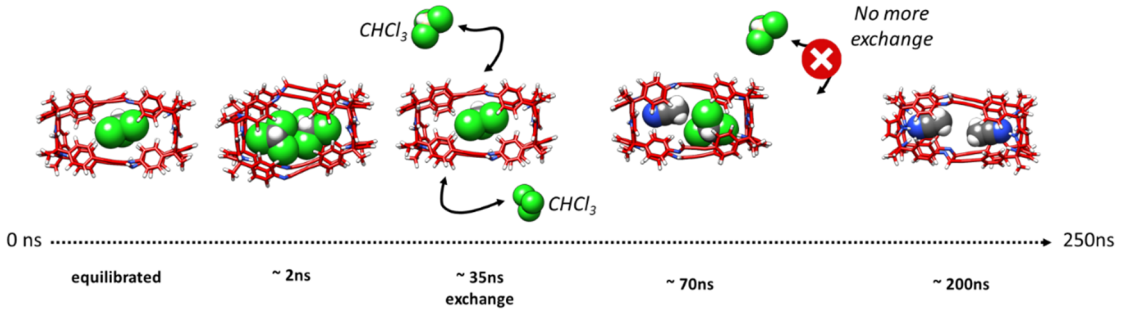


Figure 87: An overview of the molecular dynamics of the system in explicit chloroform ($CHCl_3$) / Acetonitrile (ACN) model. The graphics are separated into five parts: from left to right: (i) we have the equilibrated system representing the first frame of the dynamic, (ii) after 2ns of simulations, the second molecule of chloroform enter the cavity, (iii) multiple exchanges remain between the chloroforms molecule, (iv) after ~70ns of simulation, a molecule of acetonitrile is replacing one molecule of chloroform. At this moment, the exchange stopped until the second molecule of chloroform gets replaced after ~130ns (v).

For the simulations using a mixture of chloroform/acetonitrile (Figure 87), as in the previous simulations, the system presents a chloroform molecule inside the cavity after the equilibration phase. It was not the case after the minimizations. The initial results are equivalent to the one presented in the previous figure (Figure 86), but in this case, after 70ns of simulations, an acetonitrile molecule is replacing one molecule of chloroform. At this moment, the Acetonitrile in the cavity orienting itself for interacting with the four nitrogen of the calix[4]pyrrole unit and stabilizing the ensemble: at this moment, there is no more replacement of the chloroform molecule inside the cavity. The chloroform is less mobile in the cavity. At 200ns of simulation, a replacement occurs, and the last chloroform remaining in the cavity is exchanged with another acetonitrile from the solvated environment. At this moment, the capsule is stabilized around that two acetonitriles that both perform hydrogen interactions with the N-rim of the calix[4]pyrrole.

III. B. 3 - SEMI-EMPIRICAL MOLECULAR DYNAMIC:

In the previous simulations, we studied the dynamic process of the inclusion/release of the solvent's molecules in the capsule's cavity. From what we saw, the chloroform is susceptible to be replaced while the acetonitrile remains stable in the cavity during the time frame of the simulation. Two different simulations have been launched at SQM level in order to study with better accuracy the equilibrium between the solvent's molecules and the host.

In the first case, we studied the capsule with two molecules of chloroform inside the cavity in order to show if, at any moment of the simulations, one of the chloroforms molecules is getting spontaneously out of the cavity. For that, the simulation is launched starting from an initial configuration with two chloroform molecules in the cavity of the capsule. The simulation was launched for a total of 2ns in implicit chloroform solvation. The results of the simulated capsule are presented in Figure 88:

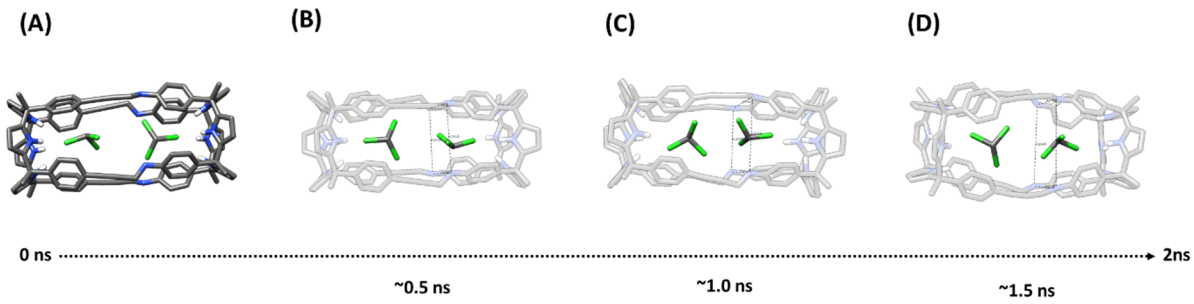


Figure 88: An overview of the semi-empirical molecular dynamics of the system in implicit chloroform starting from a configuration with two chloroforms explicitly defined inside the cavity.

We can see that the structure remains stable during all the simulations processes. There is a free rotation of each chloroform's molecules inside the cavity, but unlike to the other simulations done at the classical level, the capsule is not presenting any "open conformation". The square formed by the distance of each nitrogen at the middle of the cavity remains mostly stable during simulation. It looks like the distance is led by the orientation of the hydrogen of the N-rim ring from the calix[4]pyrrol unit: when the pyrrole hydrogens are making contacts with the solvents molecules (Figure 88A, B and D), the distance is almost the same in all the part of the square while the distance is higher in the Figure 88C when only half of the hydrogens of the N-rim are positioned for making contact. A zoom of the structures B and D, highlighting this process, is presented in Figure 89.

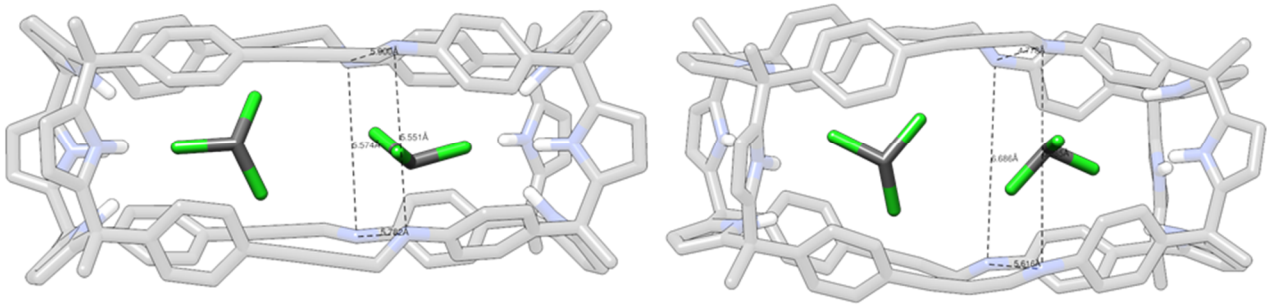


Figure 89: Zoom of the N-N distances of the capsules

Considering the outcome of this semi-empirical simulation, we tried to make another simulation, this time considering a mixture of explicitly defined solvent molecules in an implicit chloroform environment. The explicit solvation at the semi-empirical level consists of the addition around the cavity of 36 molecules of chloroforms and four molecules of acetonitrile respecting the experimental proportion of 9/1. These molecules are confined in the simulation in a little sphere, which keeps the molecules from escaping too far away from the cavity. For that, a *logfermi* potential is used through the *xTB* software. The outcome of the simulation is presented in the following Figure 90:

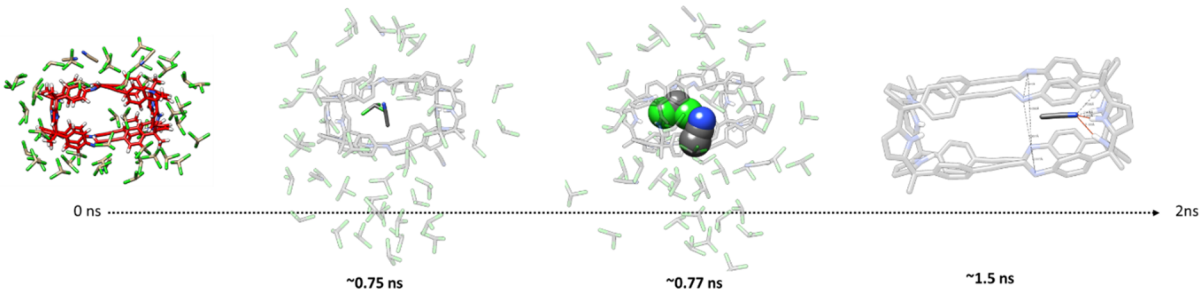


Figure 90: An overview of the semi-empirical molecular dynamics of the system in implicit chloroform starting from a configuration with an empty cavity and a mixture of chloroform and acetonitrile molecules explicitly defined around the cavity.

Unlike to the previous simulations, the empty-capsule adopts an open conformation associated with an antisymmetric orientation of the hydrogens of the pyrroles. Multiple times, in the early process, the chloroforms molecules half enter the cavity leading to an important opening of the capsule. After 0.75ns of simulation, this "half-entering" process occurs while an acetonitrile molecule is close to the cavity occurring the entry of the acetonitrile after 0.03 ns. At the moment the acetonitrile molecule enters the capsule, positioned face to face with the pyrrole's hydrogens, leading to multiple hydrogen bonds interactions. The acetonitrile is stabilizing the capsule, and we find similar values in the square formed by the nitrogens distances (see Figure 89) compared to the closed-conformation of the previous simulation (the semi-empiric one) but also from the end of the classical simulation with the two acetonitrile molecules inside the cavity (Figure 87). In the timeframe of the thesis, we do not extend the simulations, but we can expect chloroform molecules to be able to enter the cavity, and as we saw it multiple times, the chloroform molecule is rotating all the time in the cavity while the acetonitrile remains stabilized by the hydrogen bonds he made with the calix[4]pyrrole unit.

In conclusion, using MD simulations, we were able to analyse the behaviour of a calix[4]pyrrole cage under different stimuli (i.e. the use of different solvents). And we have put forward a hypothesis that will be investigated and compared to the experimental conditions in the ICIQ.

A novel Zn(II)-a porphyrin-acridinium receptor will be computationally explored in the next chapter, followed by synthesis and characterisation.

5

COMPUTATIONAL ANALYSIS, SYNTHESIS, AND CHARACTERISATION OF NOVEL Zn(II)-PORPHYRIN-ACRIDINIUM RECEPTORS

I - Zn(II) BISPORPHYRIN-ACRIDINIUM SCAFFOLD

I. A - PORPHYRIN RECEPTOR

Porphyrins are an important class of molecules present in many life processes. Chemically, porphyrins are highly coloured cyclic aromatic molecules formed by four modified pyrrole units held together by four sp2 hybridised carbon bridges, called *meso* carbons. Porphyrins are well-studied units in supramolecular chemistry, they are easily functionalised by metalation, and thus a wide variety of metal ions can be incorporated into the porphyrin ring. In this thesis, only the metallated porphyrin with a Zn(II) cation has been considered. The physicochemical properties of porphyrins can also be modified by functionalisation of the *meso* and β-pyrrolic positions. Both free base and metallated porphyrins show a characteristic intense band (the Soret band) with a maximum between 380 and 420 nm in their UV-vis spectrum. This band corresponds to a transition from the ground state to the second excited singlet state. For porphyrins, up to four additional bands between 480 and 700 nm (called Q bands), corresponding to the transitions between the ground state and the first excited singlet state, can be observed. These bands (Q bands and Soret band) are generally used for as an additional characterisation of porphyrin complexes.[130] A general overview of the application of the porphyrinoids system is presented in Figure 91.

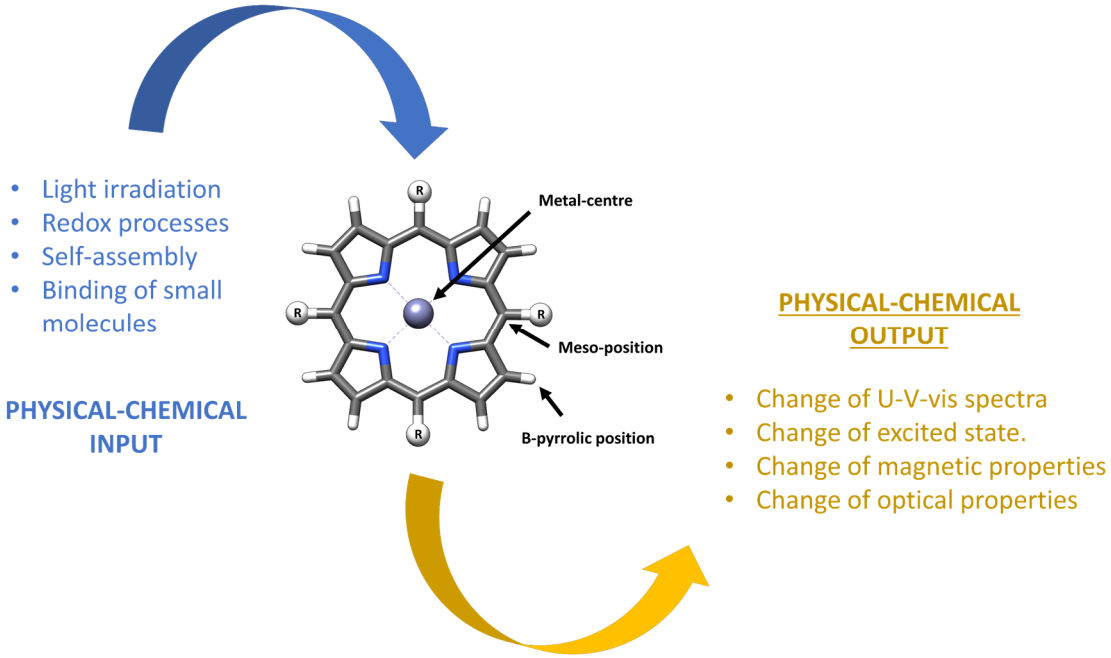


Figure 91: Overview of the application of porphyrinoids, whose physical-chemical output depend on the applied stimuli. The porphyrin core is also defined and described by the metal-centre and the main functionalisation sites (*meso* and *β*-positions)[151]

I. B - ALLOSTERISM

I. B. 1 - PRINCIPLE

Molecular systems presenting a cavity are particularly involved in the supramolecular recognition processes through non-covalent interactions, leading to the formation of supramolecular host-guest complexes. For these supramolecular systems capable of molecular recognition, allosterism offers the capabilities of controlling the interactions between the host and the guests. Allosterism can be defined as the ability of a receptor to change its conformation due to the first interaction with an effector (a guest molecule), resulting in a modification of its structural and physical properties, which modifies its ability to bind a second guest molecule at a different binding site. Add a reference These conformational changes can be cooperative (activation) when binding the effector enhances the binding affinity for the second guest or antagonistic (inhibition) when binding the effector decreases the binding affinity for the second guest. Additionally, the effector and the guest can be homotropic (the same chemical species) or heterotropic (different chemical species). Allosterism is a mechanism described mainly for proteins, an adapted illustration of this mechanism is shown in Figure 92.

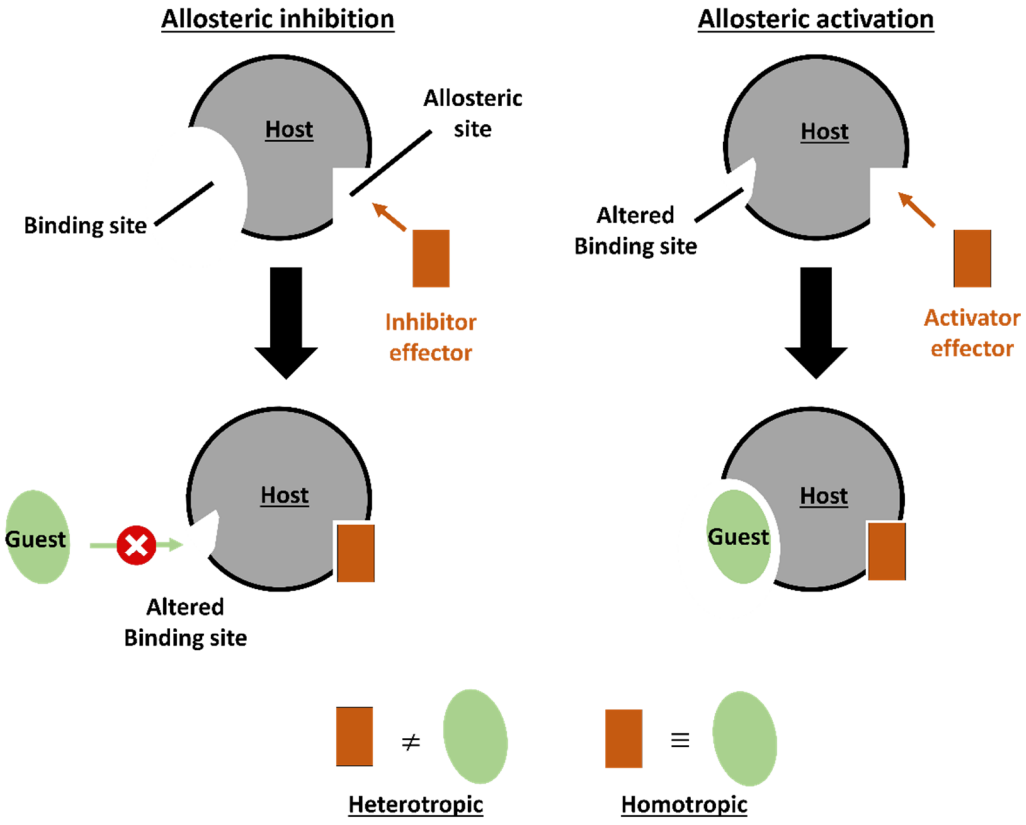


Figure 92: Illustration of the general principle of allosteric control adapted for host-guest systems. Left: allosteric inhibition. Right: allosteric activation

The first artificial allosteric system was reported in 1979 by the J. Rebek Jr group[152] as a macrocyclic polyether composed of a 2,2'-bipyridine ligand with an ether crown (Figure 93).

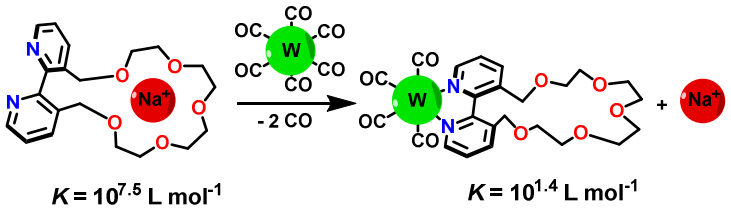


Figure 93: One of the first allosteric systems described in 1979[152]

This system can change its geometry upon chelation of metal in the bipyridine site, becoming coplanar. More precisely, the binding of certain cations (Na^+, K^+, Li^+) in the ether-crown cavity is affected by the chelation of a metal (Pd^{2+}, W^0) in the bipyridine site in a deuterated chloroform solvated environment. This coplanarity leads to a significant decrease in the capacity of the host to accept a cationic guest. And thus, the formation of the host-guest complex is not favoured. The system presents heterotropic allosteric inhibition illustrated by the diminution of the binding constant (Ka) of the Na^+ cation while the $W(CO)_6$ molecule binds the bipyridine site (from $10^{7.5}$ to $10^{1.4}$ $L.mol^{-1}$ in $CDCl_3$)

I. B. 2 - ALLOSTERIC RECEPTOR PRESENTING A PORPHYRIN SCAFFOLD

In some cases, porphyrin-receptors are also capable of doing allosterism: one of the relevant systems was described in 2009 by the group of W.-D. Jang (Figure 94). This system is a molecular tweezer composed of two different recognition motifs: (i) two Zn(II) metallated porphyrins capable of interacting with a 1,4-diazabicyclo[2.2.2]octane (DABCO) acting as a ditopic ligand and (ii) a 2,2'-bisindole acting as a spacer and capable of interacting with anionic molecules such as chloride. In the absence of guests, the indole bonds rotate freely, and the receptor alternates between the open and closed forms. However, the addition of a ditopic ligand (DABCO) coordinating the two Zn(II)-porphyrins leads to the stabilisation of the receptor and immobilising the porphyrins in a cofacial orientation. This new geometry of the receptor leads to a cis orientation of the bisindole unit, favouring the interaction of the receptor with anions chlorides. This change is associated with a significant improvement of the binding constants of (i) the chloride anions in the presence of DABCO (from 4.93×10^4 to 7.10×10^6 in THF solvated environment) and (ii) the association constant of the DABCO is also improved by the presence of chloride anions (from 2.02×10^6 to 2.48×10^7 in THF solvated environment). In that case, the bisporphyrin molecular tweezer shows an important heterotropic

allosteric effect where both the DABCO and the chloride anion play the roles of the effector for each other.

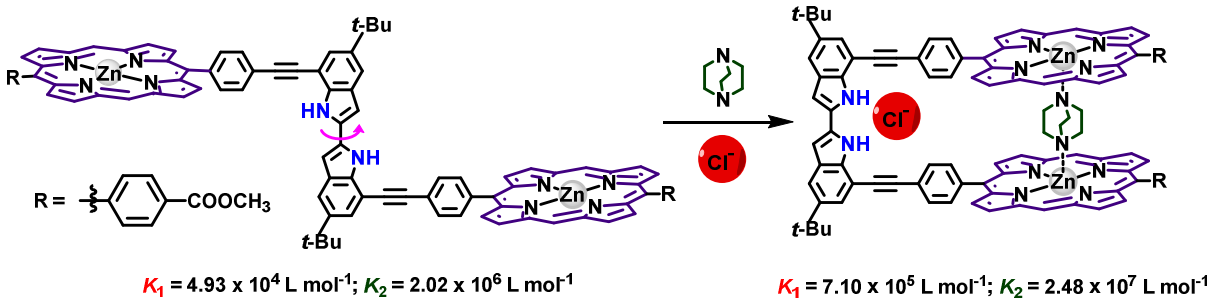


Figure 94: Allosteric receptor presenting Zn(II)porphyrin scaffold[153]

I. B. 3 - PREVIOUSLY SYNTHESISED PORPHYRINS-RECEPTOR WITH ALLOSTERIC PROPERTIES IN THE HOSTING INSTITUTION

This specific topic has been studied these past years by one of the hosting institutions of this thesis: the LSAMM laboratory partner of the NOAH project. In the following Figure 95, a three-dimensional receptor reported by the group is represented[154]:

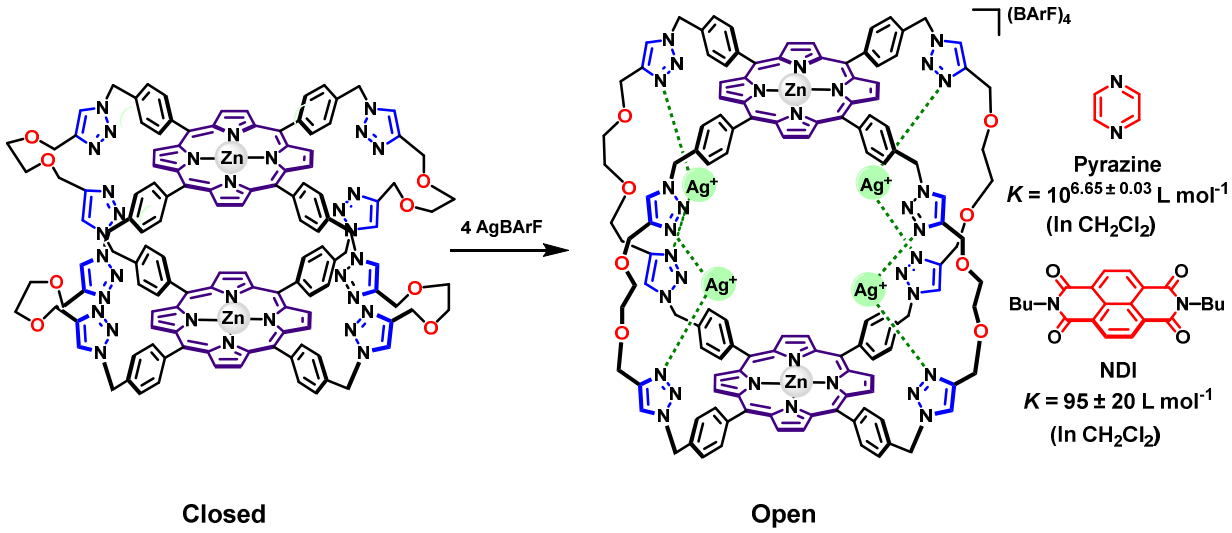


Figure 95: Zn(II)-porphyrin receptor presenting allosteric control designed at LSAMM[154]

This receptor is composed of two Zn(II)-porphyrin bonded with four different flexible covalent spacers each incorporating two triazole units. The initial state of the cage (Figure 95-left) has a closed conformation resulting from an intramolecular interaction between its porphyrins. The cavity of this system is not accessible to a ditopic ligand such as pyrazine or π-acceptor guest such as Naphthalenediimide (NDI). The triazole groups in the periphery of the receptor can coordinate Ag(I) ions (acting as the effectors) that induce the opening of the cavity (Figure 95-right). In its open form, the cage can accommodate any of the previously mentioned ditopic ligands. The association constant of these guests has been determined in dichloromethane using

respectively UV-visible titration (for pyrazine) and ^{1}H NMR (NDI) and are equal to $10^{6.65}$ $L.mol^{-1}$ (Pyrazine) and 95 $L.mol^{-1}$ (NDI).

II - PRESENTATION OF THE Zn(II)-PORPHYRIN RECEPTOR

II. A - GENERAL STRUCTURE

As part of my PhD project, we simulated receptor scaffolds with the general structure depicted in Figure 96. This receptor, designed by Amy Edo-Osagie, one of the E.S.R. of the NOAH project in the LSAMM group, presents two different binding cavities and is susceptible to acting as an allosteric receptor. The general structure consists of four different units: (i) the Zn(II)-porphyrin, (ii) a 9-phenyl-acridinium unit at one of the *meso*-position, (iii) the presence of solubilising groups at the remaining *meso*-positions, and (iv) a flexible alkene chain linking both acridiniums units together to form the expected tweezer.

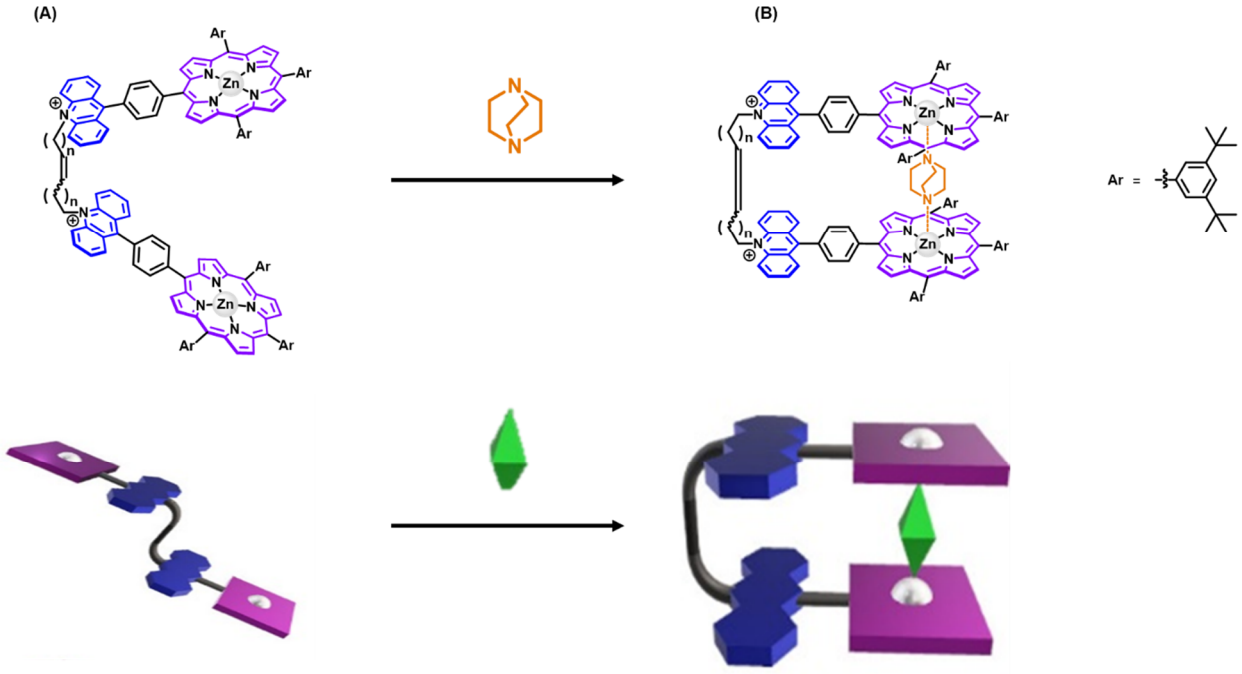


Figure 96: (A) Chemical structure (top) and schematic representation (bottom) of the tweezer in the unbound state. (B) Chemical structure (top) and schematic representation (bottom) of the DABCO-coordinated receptor, presenting an accessible binding cavity between the two acridinium units.

The receptor is susceptible to presenting several states: in its open conformation, the receptor is highly flexible and able to bind a ditopic ligand to coordinate the two Zn(II)-porphyrin (Figure 96A). In a previous example (Figure 94), we saw that DABCO is a good ditopic ligand for these systems, and when the DABCO is coordinating the Zn(II)-porphyrin, the receptor is stabilised in a closed conformation where both acridiniums are at a distance of 7Å, forming a binding cavity susceptible of accepting polyaromatic guests through π-π interactions (Figure

96B). To improve the solubility of the receptor in organic solvents, 1,3-di-tert-butylbenzene groups were used.

The 9-phenyl-acridiniums are an interesting motif since they are known to have a multi-responsive nature: chemiochromic, photochromic, and redox switching.[155,156] Acridinium presents interesting photochromic properties (Figure 97), leading to the reversible transformation into acridane within the influence of irradiation in a polar solvent. Acridinium also has the ability to bind electron-rich polyaromatic guests via π-π interactions[132], allowing for the release of substrates: indeed, switching the acridinium to acridane enables to control the catch an release of potential guests. These features remain novelties within the field of artificial allosteric systems.[131,157–159]

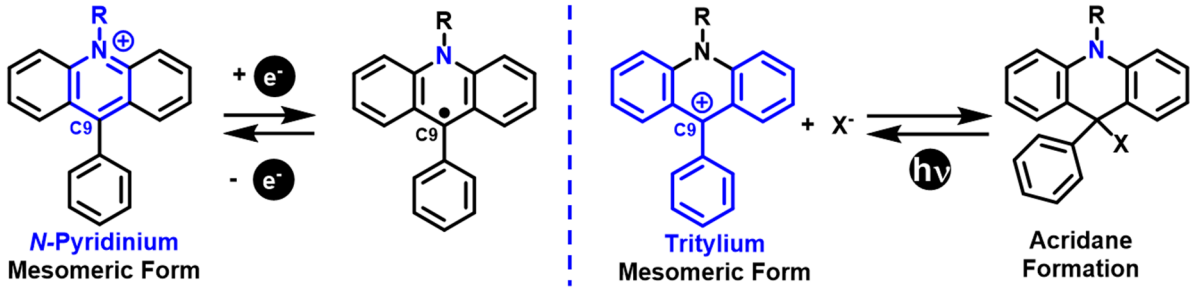


Figure 97: The multi-responsive properties of the 9-phenyl-acridinium moiety

Several alkene chain lengths were investigated. The initial idea was to use an eight-carbon chain that would correspond to the Zn-Zn distance when coordinated with DABCO (~7Å). Unfortunately, due to synthetic feasibility, the receptor was synthesised using a C18-carbon spacer with three 1,3-di-tert-butylbenzene as solubilising groups for organic solvents. Coupling the phenyl-acridinium groups with the porphyrins core, a new family of receptors can be accessed. Finally, using both the porphyrin and the multi-responsive acridinium properties, two different receptors were synthesised and characterised in the LSAMM labs by Amy Edo-Osagie (Figure 98).

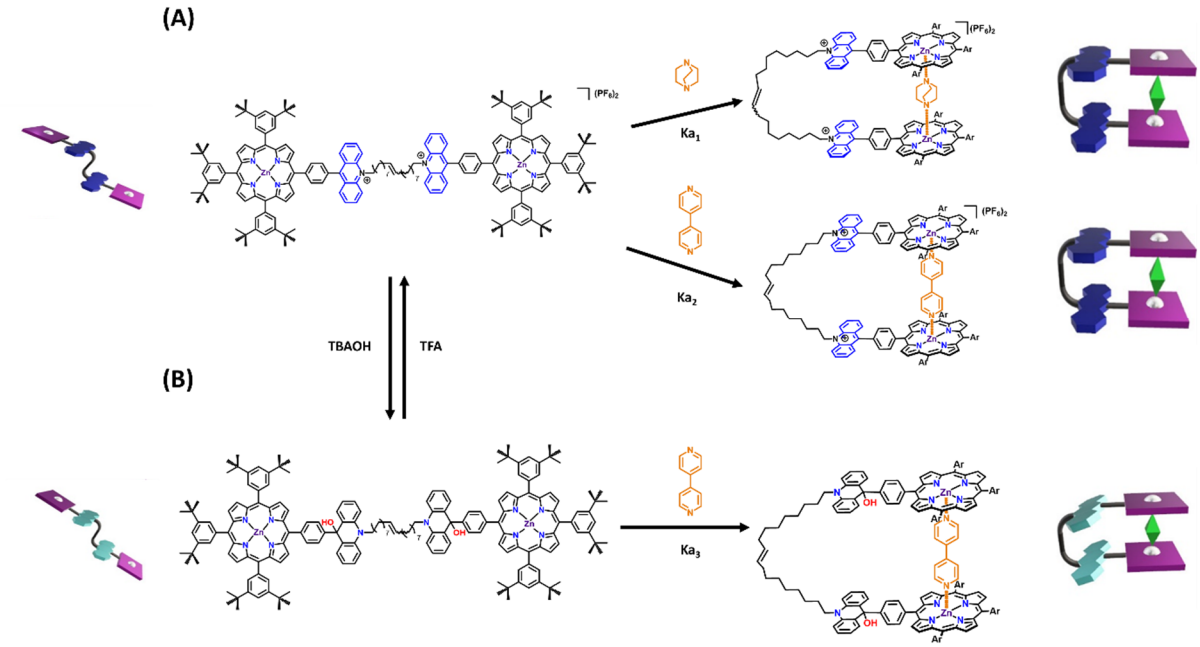


Figure 98: Synthesised receptors using a C18 carbon spacer: (A) a Zn(II)-porphyrin-acridinium receptor with two ditopic ligands and their respective association constant (DABCO, Ka_1) and (Bipyridine, Ka_2); (B) a Zn(II)-porphyrin-acridane receptor with a ditopic ligand and its respective association constant (Bipyridine, Ka_3)

These two receptors differ by the presence of the acridinium or the acridane unit binding the spacer, using the multi-responsive properties of the acridinium. With the binding of the ditopic ligands, three host-guest complexes are formed: DABCO with the Zn(II)-porphyrin-acridinium complex (Figure 98A) and the 4,4'-bipyridine to both the Zn(II)-porphyrin-acridinium and Zn(II)-porphyrin-acridane complexes (Figure 98A and Figure 98B).

As there is no crystallographic data of these hosts systems, computational analysis was performed to (i) get information about the behaviour of the receptor with and without ditopic ligands and (ii) measure the binding free energy of the ditopic ligands.

III - GENERATION OF PARAMETERS FOR Zn(II)-PORPHYRIN RECEPTOR

To study the receptors and the various complexes formed by the receptors in association with the ditopic ligands, we used classical molecular dynamics for binding free energy prediction and to analyse the behaviour of the receptor in a solvated environment.

This process was done in several steps, all connected between them using the procedure presented in chapter 3 of the thesis. First, the parameters for the host system, including the metal, are generated using the two first modules of the HG-DYNAusor platform. At this step, the receptor is optimised at a semi-empirical level using dichloromethane implicit solvation, and the parameters of the metal centre are generated from that optimised structure. The receptor is then charged using the previously explained MPD-RESP charge procedure, and the bonded model is explicitly solvated in dichloromethane. As the ditopic ligand binds the metal atoms, we are in a specific case where the complex has to be manually formed prior to the optimisation and parametrisation of the metal centre using the MBCPY.py module of the HG-DYNAusor platform (see Chapter 3). Thus, the complexes are formed manually by positioning the nitrogen atoms of the ditopic ligand at an optimal distance (<2.8Å) from the zinc atoms, followed by a complex optimisation at a semi-empirical level.

Five molecular dynamic simulations were launched corresponding to the different systems considered (Figure 98): (i) The Zn(II)-acridinium-DABCO complex, (ii) the Zn(II)-acridinium-bipyridine complex, (iii) the Zn(II)-acridinium host without ditopic ligand, (iv) the Zn(II)-acridane-Bipyridine complex and (v) the Zn(II)-acridane host without ditopic ligand. 2.500 geometries representing 250 nanoseconds of the simulation were then extracted for analysis. These geometries are considered for conformational analysis. In contrast, for the binding free energy prediction, due to the computational cost of calculating the thermodynamic properties, only one geometry is extracted every 2.5 ns for a total of 100 structures for which the binding free energy was predicted.

IV - COMPUTATIONAL ANALYSIS OF DITOPIC LIGAND BINDING

IV. A - CONFORMATIONAL ANALYSIS OF Zn(II)-PORPHYRIN RECEPTORS

Analysis of the MD simulations was based on the atomic RMSD relative to the initial conformation and the Rg. The Rg of the object can be calculated as the root-mean-square distance between each point in the object and its centre of mass.[160] The scatterplot representing the geometries described by RMSD and Rg are shown in Figure 99-A. In this graphic, the two receptors without ditopic-ligand (in orange and light-blue respectively in Figure 99-A) deviates more than the others and logically sample a larger space than the systems with ditopic ligand. It could be easily explained since, without ditopic ligands, the receptors are exploring extended geometries. The three other systems (in pink, blue and green in the Figure 99-A) represent receptors with a ditopic ligand and sample a relatively constrained space in the graphic. To identify representative geometries, we performed a Knn clustering analysis in each independent MD simulation. Four clusters are created for each of the systems (see technical details below). The centroid of the most representative cluster is shown in Figure 99 B-F. For the systems without ditopic ligand (Figure 99B and C), the representative structures are respectively present a very-open (in green) or semi-open conformation (in red), while for the systems with the ditopic ligand (in blue), the dynamic of the system is mainly related to the alkene chain and the mobility of the porphyrin cores around the axis formed by the ditopic ligand.

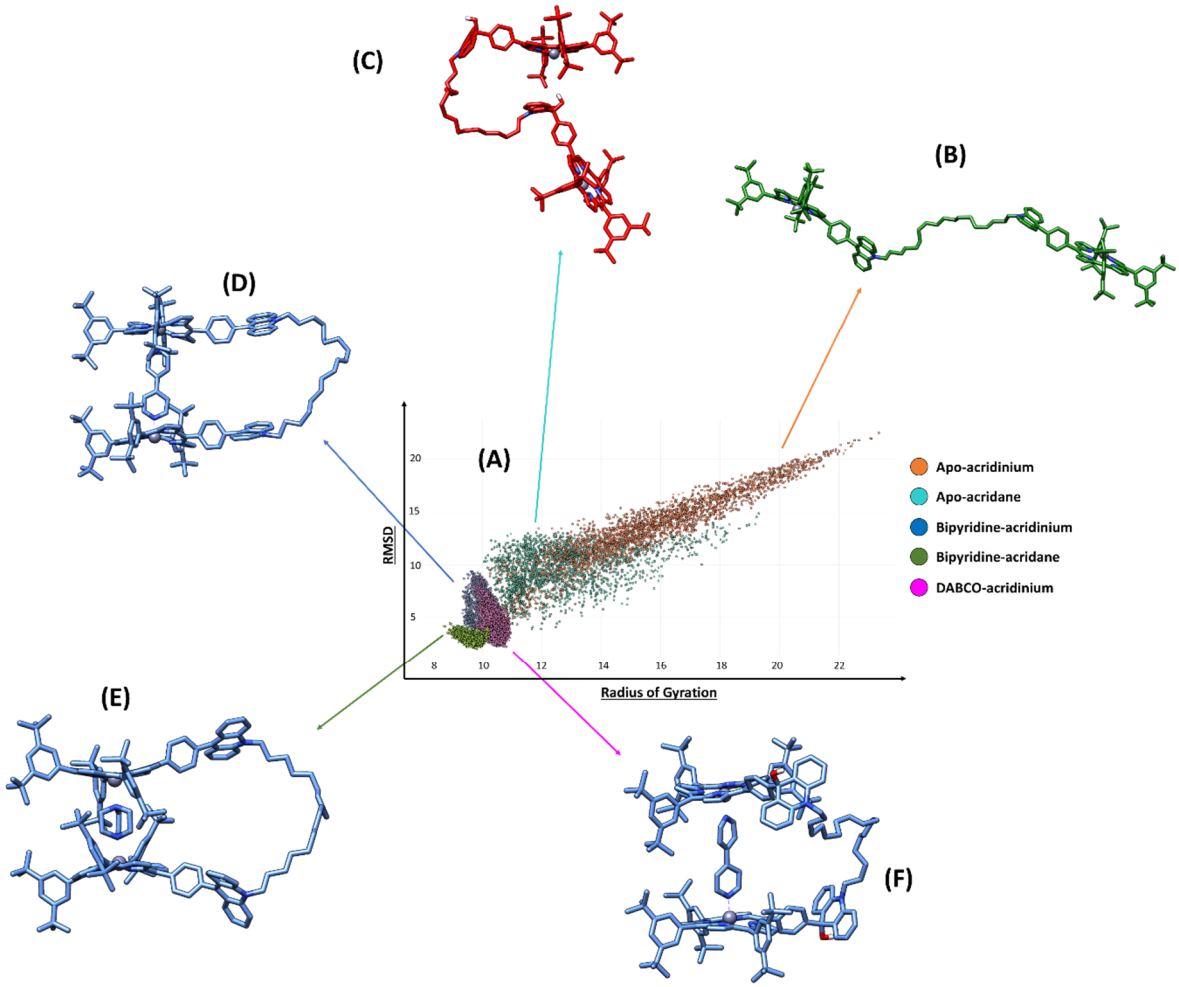


Figure 99: Clustering analysis of the porphyrinoids receptors; (A) scatterplot representing the geometries described by RMSD and radius-of-gyration ; kmeans clustering of (B) Apo-acridinium (orange) receptor ; (C) Apo-acridane (light-blue) receptor ; (D) Bipyridine-acridinium (blue) receptor ; (E) Bipyridine-acridane (green) receptor ; and (F) DABCO-acridinium (pink) receptor

Various metrics of the Knn clustering are displayed in Table 15 and Table 16.

Table 15: Performance of the Knn clustering

	DBI	pSF	SSR/SST
Apo-acridinium	1.90	431.86	0.34
Apo-acridane	1.87	388.16	0.31
Bipyridine-acridinium	1.99	570.36	0.40
Bipyridine-acridane	1.49	1358.11	0.62
DABCO-acridinium	2.80	196.41	0.19

In Table 15: the Davies-Bouldin Index (DBI) and pseudo-F statistic (pSF) values are metrics of clustering quality; low values of D.B.I. and high values of pSF indicate better results. DBI measures the sum over all clusters of the dispersion within the cluster versus the separation

between clusters; the smaller the DBI, the better. The pSF is another measure of clustering quality, intended to capture the "tightness" of clusters and is essentially a ratio of the average sum of squares between groups to the average sum of squares within the group. High values are best. The R-squared (SSR/SST) value represents the percentage of variance explained by the data. The overall performance of the Knn clustering (Table 11) is correct, even if for most systems the explained variance can be considered low. This can be explained by the high dynamics of the system and the rapid alternation of the different conformations extracted by Knn clustering, and thus the predominance of transient states and the absence of a stabilised structure over time. The exception is the bipyridine-acridane system, for which the open conformation (in green) corresponds to a stable conformation over time.

Table 16: Composition of the clusters of the kmeans clustering

Cluster	Frames	Frac	AvgDist	Stdev	Centroid	AvgCDist
Apo-acridinium						
0	849	0.340	8.388	1.773	129	6.306
1	721	0.288	7.862	1.672	1573	6.843
2	581	0.232	7.273	1.514	1991	9.031
3	349	0.140	7.788	2.004	68	9.140
Apo-acridane						
0	936	0.374	7.621	1.796	595	6.838
1	645	0.258	7.940	1.974	1226	6.786
2	514	0.206	7.957	1.780	1421	6.630
3	405	0.162	8.112	1.789	1656	6.588
Bipyridine-acridinium						
0	1027	0.411	3.123	0.449	1952	2.988
1	874	0.350	3.121	0.431	776	3.017
2	339	0.136	3.355	0.568	2468	4.203
3	260	0.104	3.197	0.525	1158	4.442
Bipyridine-acridane						
0	900	0.360	3.317	0.595	1154	6.185
1	809	0.324	3.543	0.625	983	4.553
2	499	0.200	3.677	0.730	30	4.546
3	292	0.117	3.752	0.850	1281	6.654
DABCO-acridinium						
0	943	0.377	2.614	0.290	660	1.508
1	692	0.277	2.730	0.366	1879	1.659
2	557	0.223	2.645	0.318	257	1.493
3	308	0.123	3.075	0.475	1759	1.773

Table 16 contains all the practical information about Knn clustering. Taking the first line as an example, we can say that the most populated cluster (#0) is composed of 849 frames which represent 34% of the processed frames. The average distance between the points in the cluster is 8.4 ± 1.8Å, frame 129 represent the centroid structure which has the lowest cumulative distance to every other point and the average distance of the cluster #0 to every other cluster is 6.3Å. Comparing the systems with each other, we can highlight the fact that the systems without ditopic ligand have a much larger variation (~8Å) between the clusters than the other systems (~3Å). This is logical considering that, without the ditopic ligand, the system explores a much broader conformational space.

The result of the Knn-clustering are presented in Figure 100 to Figure 104. Concerning the clustering of the apo-acridinium receptor (Figure 100), as the structure of the centroid is similar for clusters 2 and 3 (they symmetry-related), the two clusters were manually merged to correspond to the open conformation of the receptor (in green representing then 52.1% of the total processed frames, becoming the most representative structure).

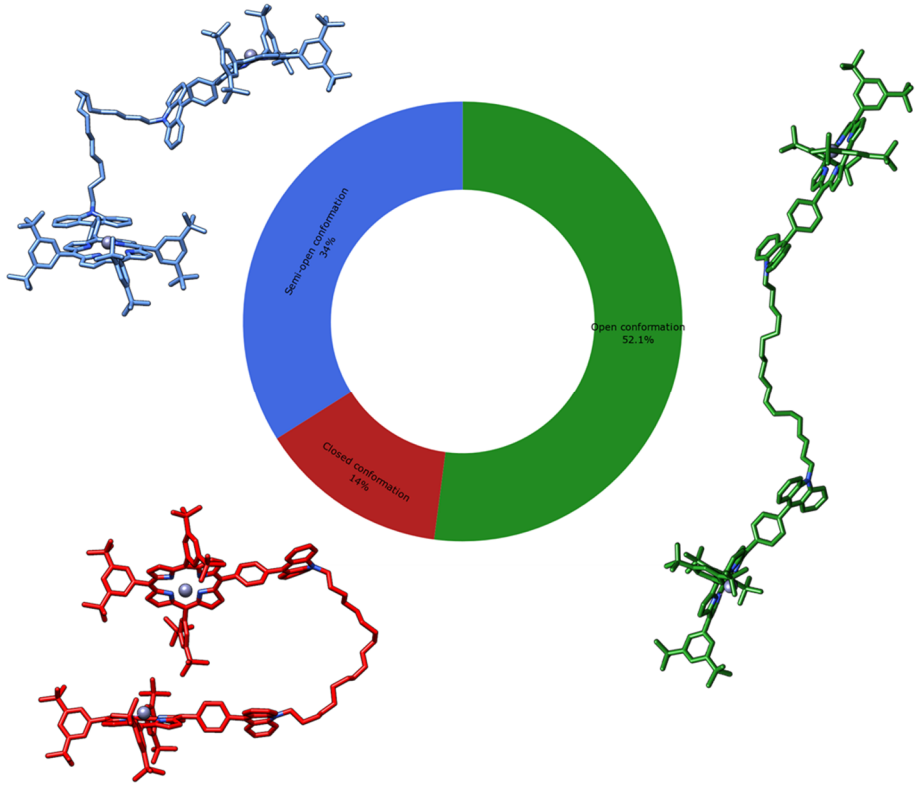


Figure 100: Knn clustering of the apo-acridinium receptor, coloured by clusters: (In green) the open-conformation, (In blue) the semi-open conformation, and (In red) the closed-conformation

The two remaining clusters can be defined as the closed conformation of the receptor (in red), representing a 14% of the population, and the semi-open conformation (in blue) corresponding to a transitional structure between the extended and closed states, which represents 34% of the total processed frames.

Concerning the clustering of the apo-acridane receptor, four clusters were generated and named based on the respective geometry of the centroid structure (Figure 101). This time, as every centroid presented a different structure, none of them were merged. The most populated cluster corresponds to the closed conformation (in red, 37.4%), which corresponds to the geometries not accessible for binding a ditopic ligand, with a perpendicular orientation of the porphyrins between each other. The second one (in green, 25.8%) represents the open conformation. Even if the two porphyrins are not facing each-others in a conformation close to the binding conformation, both of them were accessible, while the structure was stabilised by hydrogen bonding between the two hydroxyl groups of the acridane core.

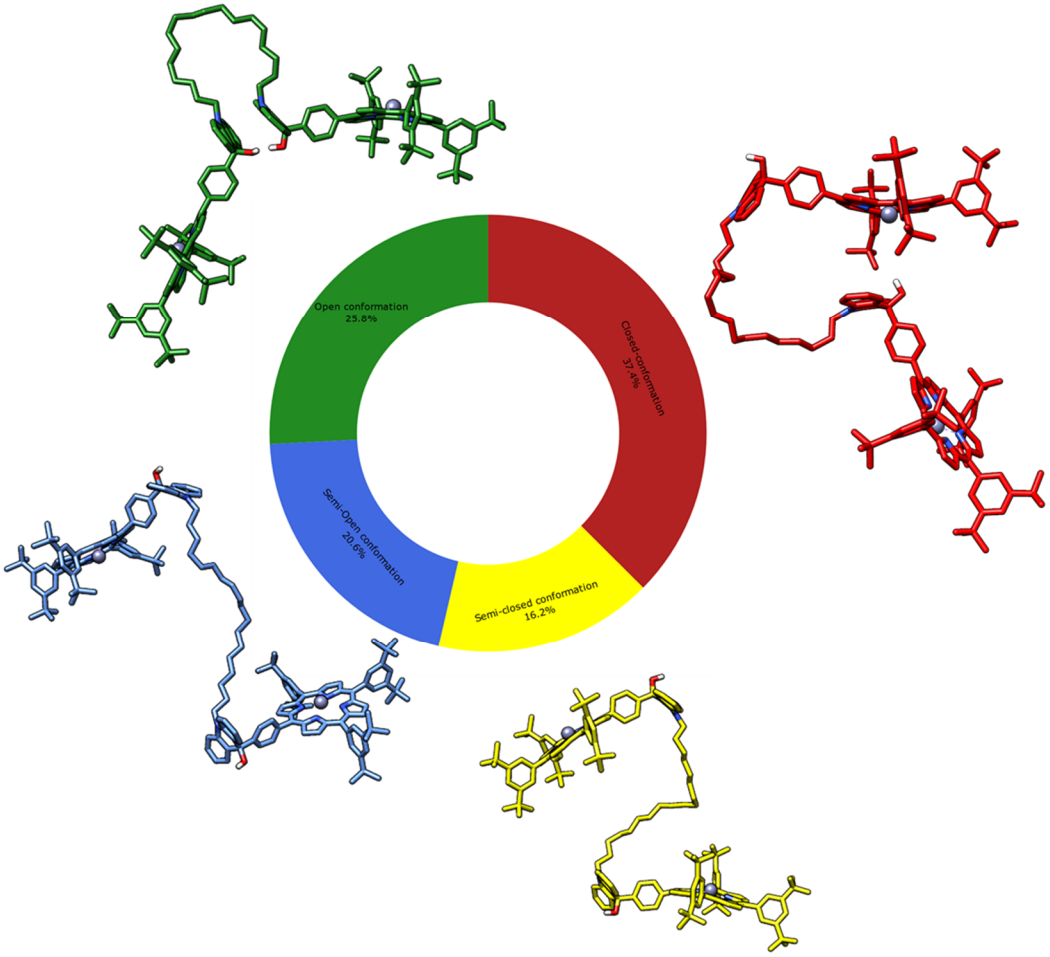


Figure 101: Knn clustering of the apo-acridane receptor, coloured by clusters: (In green) the open-conformation, (In blue) the semi-open conformation, in yellow the semi-closed conformation and (In red) the closed-conformation

The last two clusters are closed in terms of structure and can be defined as transitional states between the open and the closed conformation. They represent 20.6% (in blue) and 16.2% (in yellow) of the total frames of the dynamics, respectively. They were named semi-closed and semi-open due to the relative orientation of the porphyrin's cores.

For the clustering of the bipyridine receptor (Figure 102 and Figure 103), due to the rigidification of the porphyrin backbone related to the bipyridine binding at the so-called ditopic binding site, the main deviations in the geometry are related to the flexibility of the alkene chain and the induced rotation of the porphyrin core. For that reason, in all the cases, the four defined clusters were named in relation to the cluster order. Concerning the clustering of the bipyridine-acridinium receptor (Figure 102), the two most populated clusters (in blue and green) represent respectively 41.1% and 35% of the total processed frames. In these clusters, the receptor was in a conformation where the binding site formed between the two acridinium was open and accessible, suggesting a favourable orientation of the acridinium for the binding of guests that can perform π-π stacking interactions.

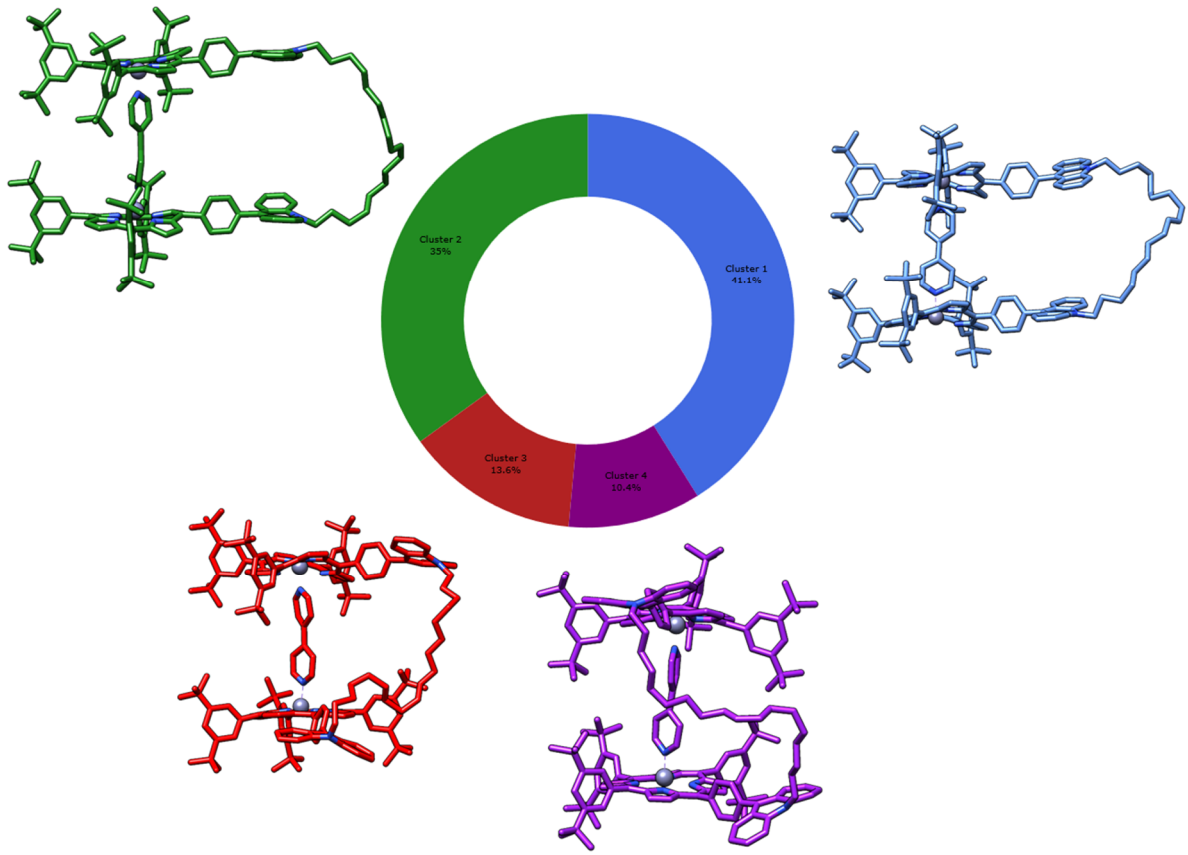


Figure 102: Knn clustering of the Bipyridine-acridinium receptor, coloured by clusters: from the most populated cluster to the less populated cluster: (In blue) the cluster 1, (In green) the cluster 2, (In red) the cluster 3 and (In purple) the cluster 4

In the two less populated clusters in red and purple representing respectively 13.6% and 10.4% of the total processed frames, the receptor is in a conformation where the binding-site formed between the two acridiniums is not accessible, linked to the flexibility of the alkene chain and the orientation of the porphyrins cores: in both cases, we can observe a 45° rotation of the porphyrins, which leads to inaccessibility of the binding site formed between the acridiniums.

For the clustering of the bipyridine-acridane receptor (Figure 103), it was very difficult to extract a trend from the clustering, and the four centroids represent structures that diverge in the orientation of the acridanes and the alkene chain. Although in some cases, the cavity formed between the two acridanes appears to be accessible (in red), the structural modification (between the acridinium and the acridane) makes ligand binding difficult due to the loss of the capabilities of making π-π interactions.

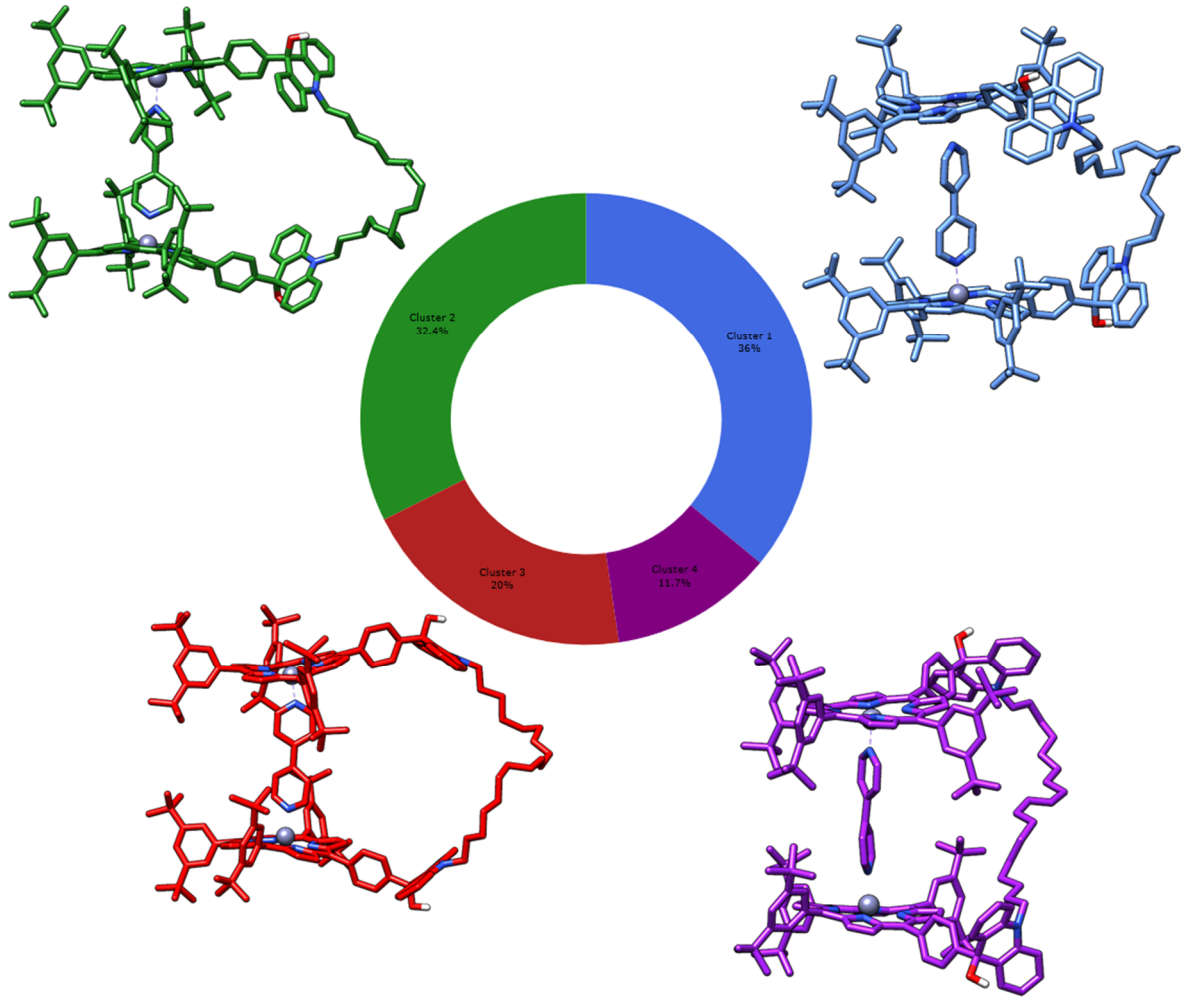


Figure 103: Knn clustering of the Bipyridine-acridane receptor, coloured by clusters: from the most populated cluster to the less populated cluster: (In blue) the cluster 1, (In green) the cluster 2, (In red) the cluster 3 and (In purple) the cluster 4

The DABCO-acridinium (Figure 104) is distinct, and even if the dynamic of the system is also mediated by the alkene chain and the porphyrins rotations, the short Zn-Zn distance imposed by DABCO lead to some distortion of the receptor. Unlike to the bipyridine-acridinium (Figure 102), the porphyrins cannot have an orientation that allows alignment on either side of the acridinium to form an accessible binding site due to the presence of a large steric clash between the 1,3-di-tert-butylbenzene groups that are located on the meso positions of the porphyrins.

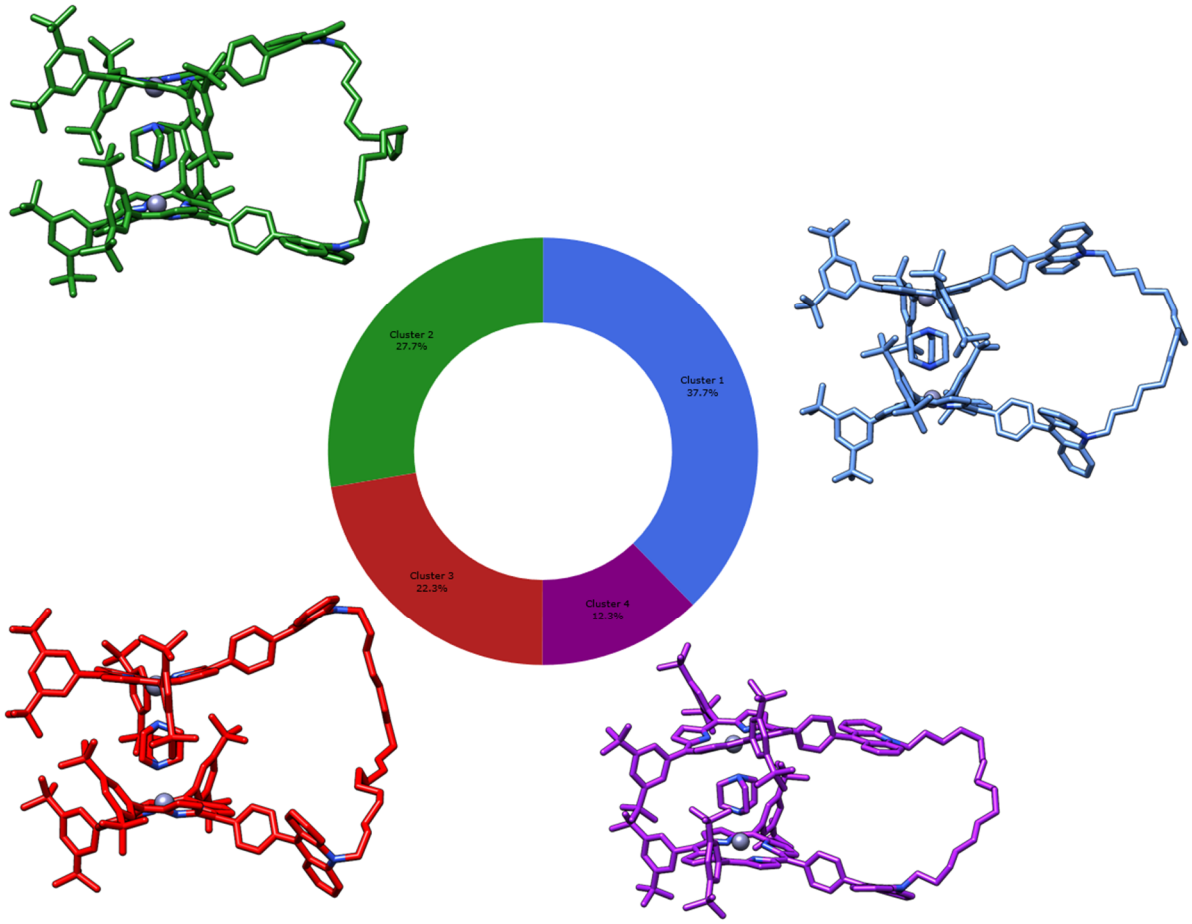


Figure 104: Knn clustering of the DABCO-acridinium receptor, coloured by clusters: from the most populated cluster to the less populated cluster: (In blue) the cluster 1, (In green) the cluster 2, (In red) the cluster 3 and (In purple) the cluster 4

Binding of the ditopic ligands DABCO and bipyridine to their respective systems was computed using a thermodynamic-based approach and experimentally measured by Amy Edo-Osagie in the LSAMM labs.

IV. B - BINDING FREE ENERGY PREDICTION OF DITOPIC LIGANDS

The computational prediction of numerical binding free energy is a very challenging problem. In our protocol, the molecular dynamics simulations are used as conformational sampling in explicit CH_2Cl_2 solvation. The semi-empirical method GFN2B-xTB is used to solve the thermodynamic properties (both the enthalpic and entropic terms are considered) in implicit solvation. We used dichloromethane as a solvent to be as close as possible to the experimental conditions in both cases. An overview of the results is presented in Figure 105:

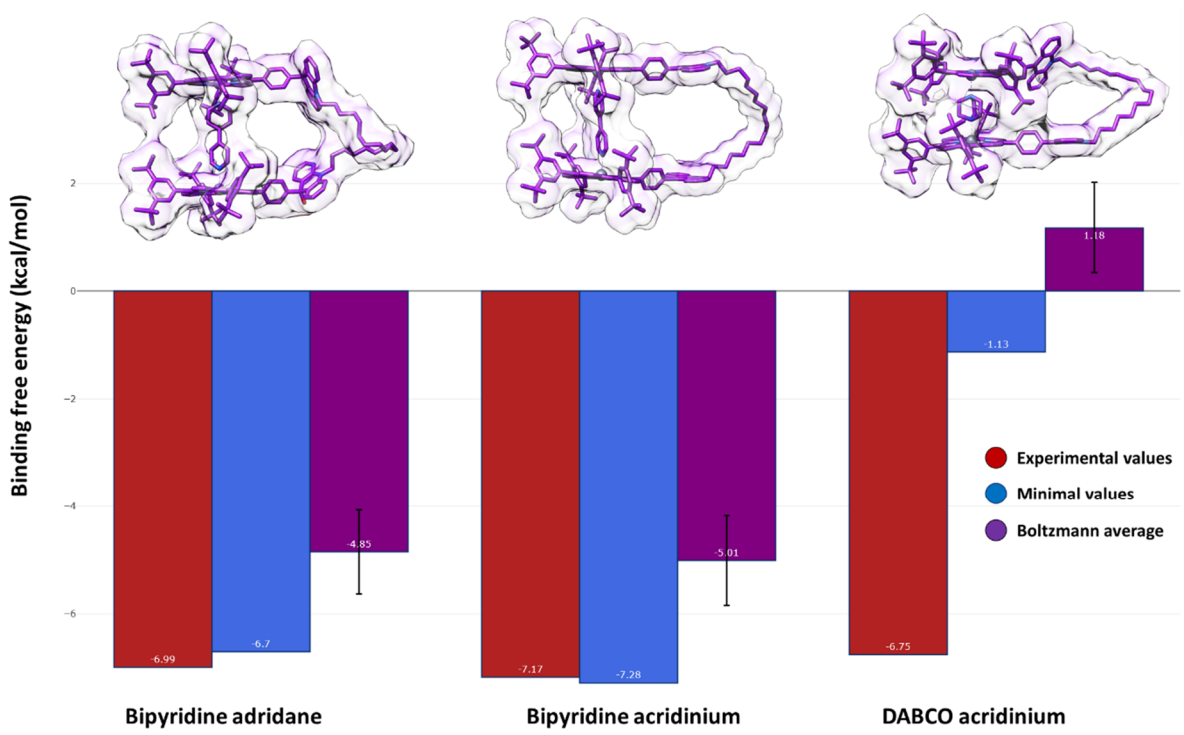


Figure 105: Binding free energy prediction of Zn(II)-porphyrin receptor considering three different receptors: (Left) the Zn(II)-porphyrin-acridane with bipirydine as a ligand ; (Middle) The Zn(II)-porphyrin-acridinium with bipyridine as a ligand ; (Right) The Zn(II)-porphyrin-acridinium with DABCO as a ligand

Predictions for both bipyridine systems are very accurate. We are close to the experimental binding free energy. with minimal errors (+0.19 kcal/mol for the bipyridine-acridane and -0.11 kcal/mol for bipyridine-acridinium). The binding free energy of the ditopic ligands using the minimal energy calculation gives better results than the Boltzmann-average ones. The limits of the Boltzmann average were already mentioned in the previous chapter. Probably the set of structures obtained with our procedure does not correspond to a physically meaningful conformational ensemble. In some instances, it may identify the true minima correctly but does not provide a faithful representation of the free energy landscape.

In the case of the DABCO-acridinium, the predicted energy is far from the experimental value. As in our protocol, the apo-structure (without ditopic ligand) is the same for the DABCO-acridinium and the bipyridine acridinium, and as we are good in the predictions of the bipyridine systems, we can assume the problem came from the complex with DABCO. Two explanations can be given:

- (i) The molecular dynamics simulations failed to find the binding conformation.

- (ii) Considering the size of the 1,3-di-tert-butylbenzene group and associated with the fact that with the DABCO coordinating the two Zn(II) metal centres, there is an important steric clash between the 1,3-di-tert-butylbenzene groups forcing the receptor to operate a rotation to avoid a face-to-face orientation of these groups. Considering the steric clash, the enthalpic cost of the deformation may be overestimated by the semi-empirical method, leading to an underestimated binding free energy.

In conclusion, we were able to cluster the different systems and calculate the binding free energy of the bipyridine systems with excellent accuracy. Next, we want to investigate a new type of receptor, where the 1,3-di-tert-butylbenzene groups are replaced by smaller (less sterically hindered) groups, to avoid deformation of the porphyrin ring and allow the face-to-face orientation of the acridinium cores, thus creating a suitable binding site for polyaromatic guest between two acridiniums. This new receptor will be synthesised in the LSAMM group, and computational analysis will be done to extract from databases a set of interesting guests able to bind the receptor in the presence of the ditopic ligand pre-organising the complex, comparing computational prediction with experimental measurement.

V - SYNTHESIS AND CHARACTERISATION OF A NEW Zn(II)-PORPHYRIN ACRIDINIUM RECEPTOR

V. A - GENERALITY

On the previous tweezer, we observed that the 1,3-di-tert-butylbenzene groups present a large steric hindrance, leading to a deformation of the porphyrin core to reduce the clashes. These constraints on the structure were measured previously with (i) the molecular dynamics of the DABCO-tweezer where the deformation was able to be shown and (ii) when we measured the binding free energy of the DABCO with the Zn(II)-porphyrin-acridinium. For those reasons, we decided to reduce the steric hindrance by replacing the 1,3-di-tert-butylbenzene from the previous system with less sterically hindered groups.

The new-tweezer $(\mathbf{1} \cdot (\boldsymbol{PF_6})_2)$ was designed based on the previously presented structure with two zinc porphyrins connected to phenyl acridinium and a long alkene chain of 18 carbons connecting the bis(acridinium-porphyrin) units to ensure enough flexibility to the system for its preorganisation as a receptor upon coordination ditopic guest. The introduction of six 4-methoxybenzene groups was considered on the porphyrin moieties to decrease the steric clashes of the 1,3-di-tert-butylbenzene and conserved sufficient solubility. A general figure of the scaffold is presented in Figure 106.

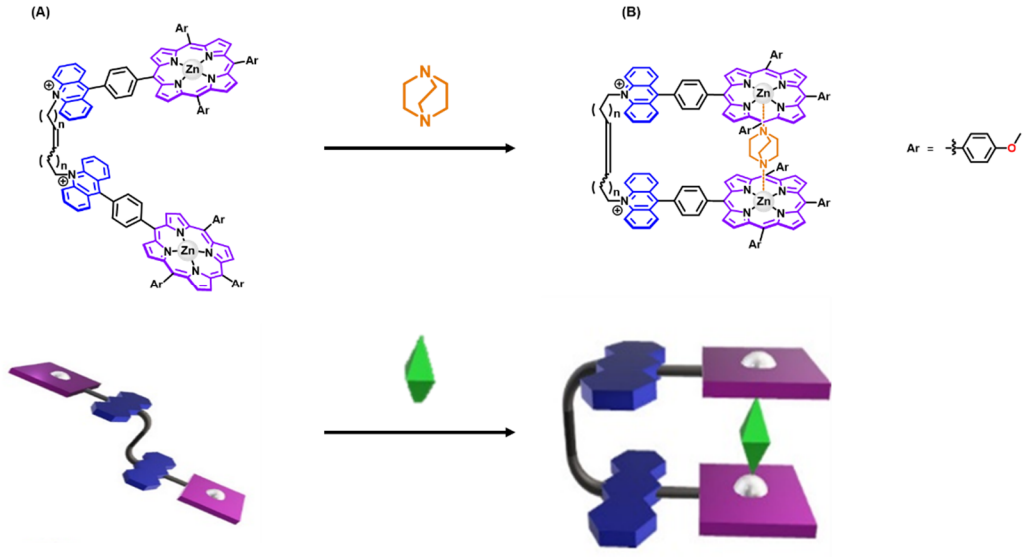


Figure 106: (A) Chemical structure of the tweezer; (B) Receptor formed by the DABCO coordinating the two Zn(II), forming an available binding cavity between the two acridiniums allowing the binding of polyaromatic guest interacting with π-π interactions.

We aim for the complex to bind aromatic guests such as perylene by π-π stacking with the acridiniums while avoiding the steric shown with 1,3-di-tert-butylbenzene groups.

V. B - SYNTHESIS AND CHARACTERISATION

V. B. 1 - CHEMICAL PATHWAY

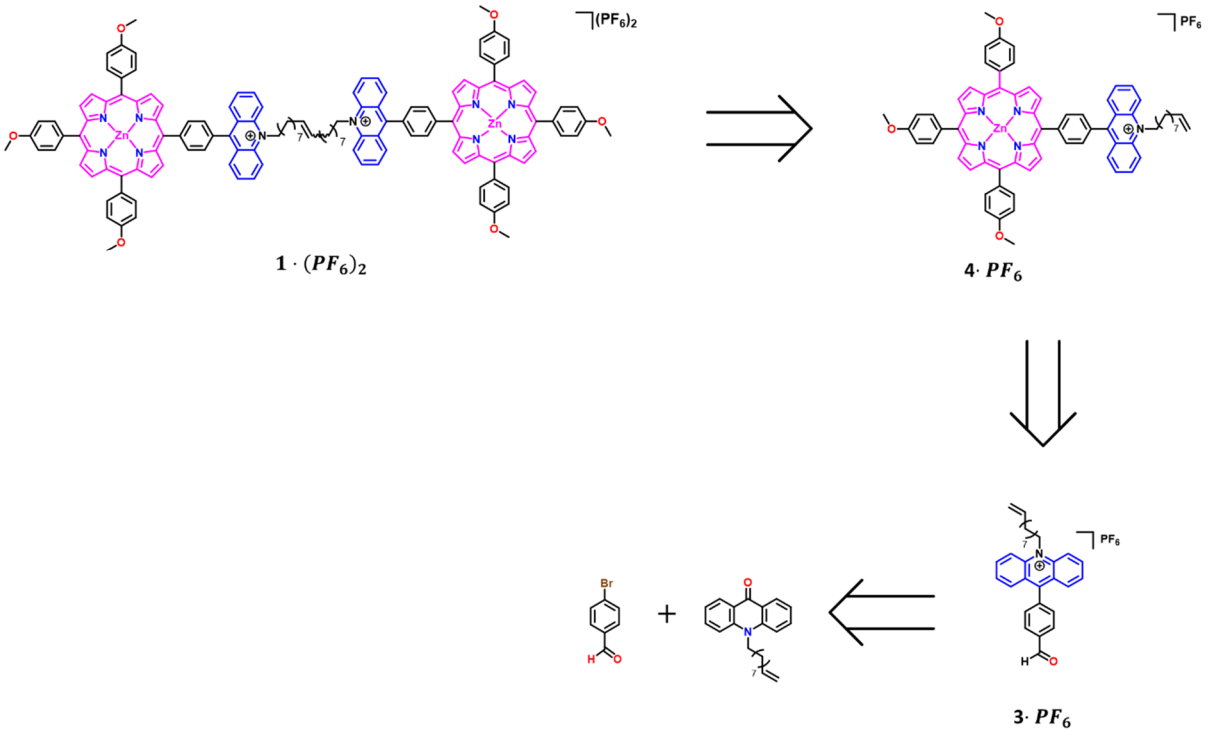


Figure 107: Retrosynthesis study of the porphyrin-acridinium conjugate: $1 \cdot (PF_6)_2$

The synthesis of the Zn(II)-porphyrin-acridinium tweezer using an 18 carbon spacer was inspired by Fukuzumi and coworkers[161] and was performed in four synthetic steps (Figure 107). A convergent approach is used for the formation of $\mathbf{1 \cdot (PF_6)_2}$. Compound $\mathbf{3 \cdot PF_6}$ is formed after two synthetic steps starting from a commercially available compound. The synthon $\mathbf{3 \cdot PF_6}$ is then used for the formation of an asymmetric porphyrin ($\mathbf{4 \cdot PF_6}$).

V. B. 2 - DESIGN AND SYNTHESIS

The synthesis started from commercially available 4-bromobenzaldehyde (Figure 108). The aldehyde functional group was first protected in the presence of 1.1 equivalent of pinacol in toluene using a dean-stark allowing to shift of the equilibrium in favour of the formation of protected 4-bromobenzaldehyde by extraction of the water formed from the reaction medium. The resulting acetal (**2**) was obtained with 90% yield after purification by column chromatography.[162]

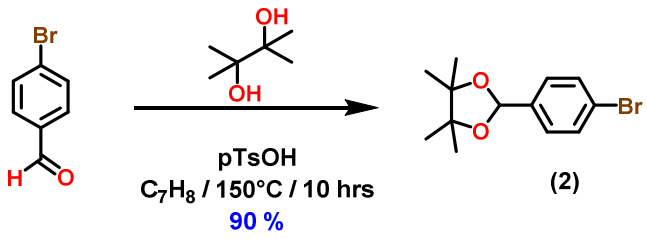


Figure 108: Synthesis of the 2-(4-bromophenyl)-4,4,5,5-tetramethyl-1,3-dioxolane (2)

The resulting 2-(4-bromophenyl)-4,4,5,5-tetramethyl-1,3-dioxolane (**2**) (Figure 109) was reacted with 1 equivalent of nBuLi in THF at −78°C, followed by the addition of 1 equivalent of dec-9-en-1-yl-acridin-9(10H)-one. The dec-9-en-1-yl-acridin-9(10H)-one (**2**) was obtained in a single step from commercially available 9(10H)-acridone in 70% yield. After acidification of the reaction mixture using HCl (37 wt%), the newly formed 10-allyl-9-(4-formylphenyl)acridin-10-ium chloride (**3·Cl**) was converted to the corresponding hexafluorophosphate salt (**3** · $\boldsymbol{PF_6}$) by anionic metathesis and isolated in 68% yield.

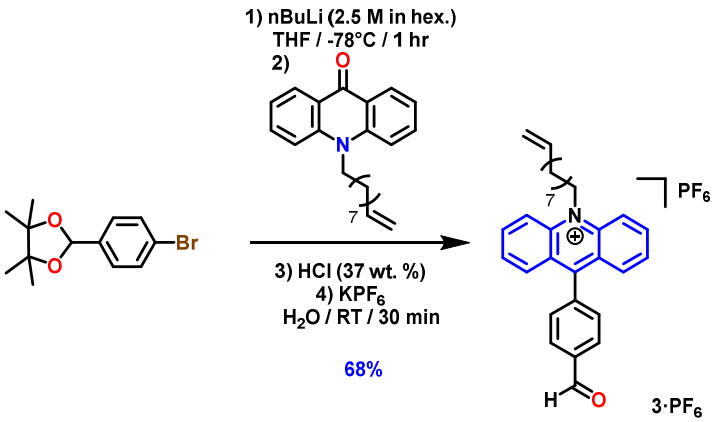


Figure 109: Synthesis of the 10-allyl-9-(4-formylphenyl)acridin-10-ium (3 · $\boldsymbol{PF_6}$)

The key intermediate **3** · $\boldsymbol{PF_6}$, was then reacted under Lindsey conditions[163] (Figure 110) in the presence of four equivalents of pyrrole, three equivalents of 4-methoxybenzaldehyde, and six equivalents of trifluoroacetic acid (TFA) in CH_2Cl_2. After aromatisation of the porphyrinogen using three equivalents of 2,3-dichloro-5,6-dicyano-1,4-benzoquinone (D.D.Q.), the reaction was followed by the metalation of the free base porphyrins using one equivalent of $Zn(OAc)_2{\cdot}2H_2O$. After the purification, the porphyrin-acridinium conjugate (**4** · $\boldsymbol{PF_6}$) was isolated with 11% yield.

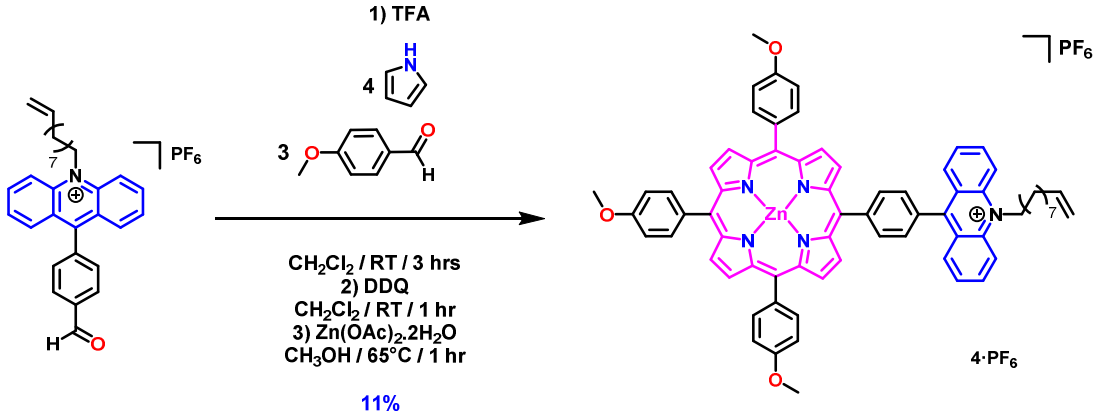


Figure 110: Chemical synthesis of the porphyrin-acridinium conjugate: $4 \cdot PF_6$

Finally, the bis(acridinium-porphyrin) tweezer ($\mathbf{1} \cdot (\boldsymbol{PF_6})_2$) was formed from two molecules of ($\mathbf{4} \cdot \boldsymbol{PF_6}$) under olefin metathesis conditions (Figure 111) using Grubbs I catalyst (10 mol%)[164] The targeted ($\mathbf{1} \cdot (\boldsymbol{PF_6})_2$) tweezer was obtained after purification as a purple-red crystalline solid in 21% yield.

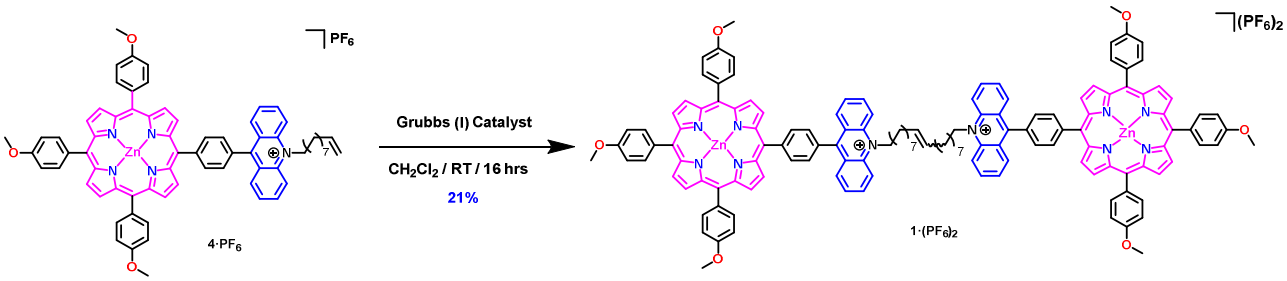


Figure 111: Synthesis of the porphyrin-acridinium conjugate: $\mathbf{1} \cdot (PF_6)_2$

Full characterisation of the bis(acridinium-porphyrin) tweezer ($\mathbf{1} \cdot (\boldsymbol{PF_6})_2$) was performed by NMR and U.V./Vis spectroscopies. At this stage of the thesis, the obtained ($\mathbf{1} \cdot (\boldsymbol{PF_6})_2$) was not isolated as a pure compound (~5-10% of impurities), but the overall purity seems acceptable to at least make the proof of concept for the binding of (i) a ditopic guest inside the Zn(II) binding-cavity and (ii) a polyaromatic system that will bind the acridinium binding cavity.

Nitrogenous ligands are known to bind Zn(II)-porphyrins.[130] Thanks to the flexibility of the linker, the two remote Zn(II) porphyrins of the tweezer can adopt a face-to-face arrangement upon binding a ditopic ligand such as DABCO. The DABCO was chosen for practical reasons as it fits the cavity of the previously studied system (Figure 98).

To do so, the obtained ($\mathbf{1} \cdot (\boldsymbol{PF_6})_2$) was reacted with a solution of DABCO, leading to the formation of the ($\mathbf{1} \cdot (\boldsymbol{PF_6})_2 \boldsymbol{\cdot DABCO}$) presented Figure 112.

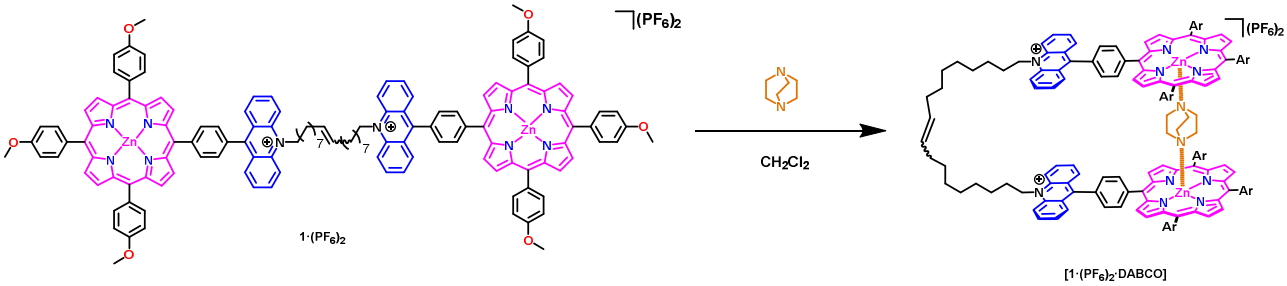


Figure 112: Chemical synthesis of the porphyrin-acridinium conjugate in complex with DABCO: [1 · $(PF_6)_2$•*DABCO*]

To the previously formed $\mathbf{1} \cdot (PF_6)_2$•***DABCO***, a solution of perylene was added. After manually stirring the mixture, the formation of the porphyrin-acridinium conjugate in complex with DABCO and Perylene [$\mathbf{1} \cdot (PF_6)_2$ •***DABCO***] ⊃ ***Perylene*** was observed (Figure 113).

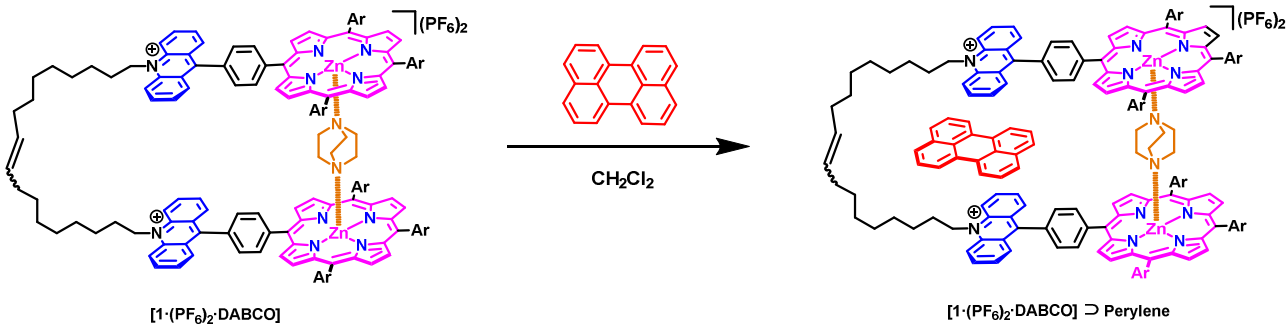


Figure 113: Chemical synthesis of the porphyrin-acridinium conjugate in complex with DABCO and Perylene: [1 · $(PF_6)_2$•*DABCO*] ⊃ *Perylene*

V. C - CHARACTERISATION

V. C. 1 - CHARACTERISATION OF $\mathbf{1} \cdot (\boldsymbol{PF_6})_2$

V. C. 1. A - NMR ANALYSIS OF THE $(\mathbf{4} \cdot \boldsymbol{PF_6})$ AND THE $(\mathbf{1} \cdot (\boldsymbol{PF_6})_2)$

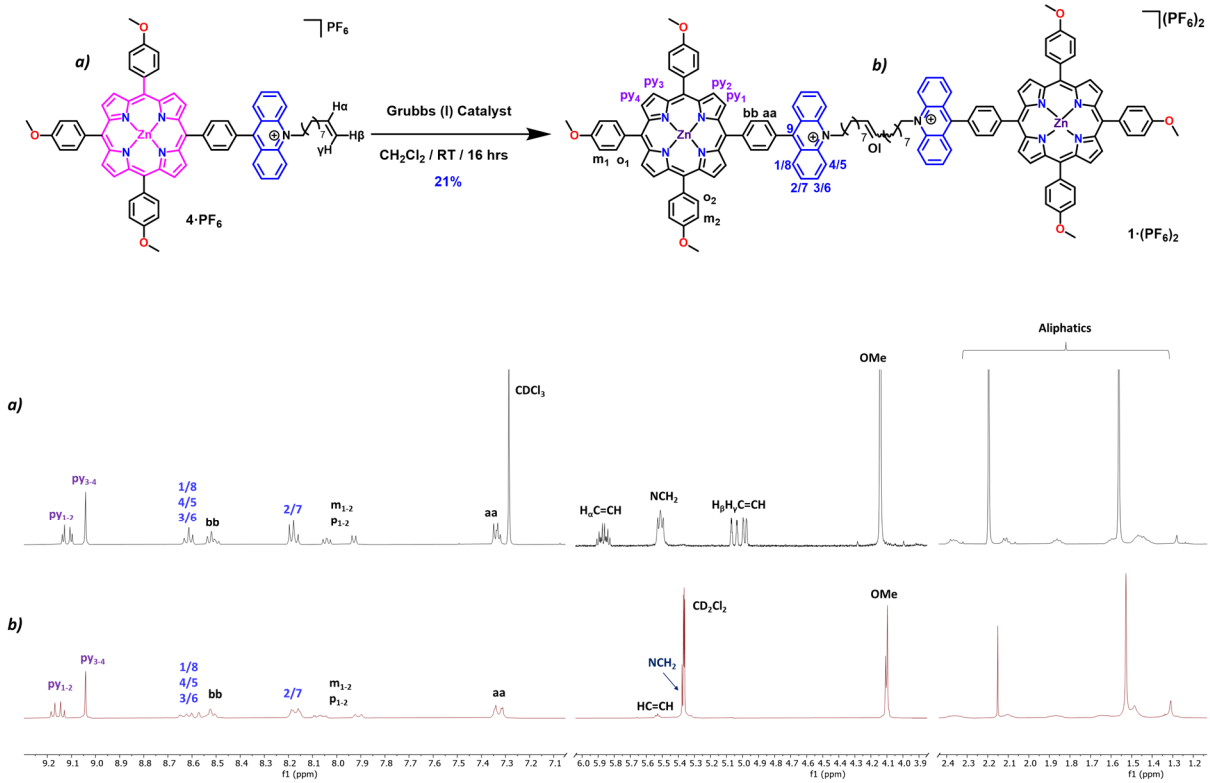


Figure 114: Stacked ^{1}H NMR spectra of a) 4·$(PF_6)_2$ ($CDCl_3$, 298K, 500MHz), b) 1·$(PF_6)_2$ (CD_2Cl_2, 298 K, 500MHz)

The porphyrin-acridinium conjugate $(\mathbf{1} \cdot (\boldsymbol{PF_6})_2)$ was analysed by ^{1}H NMR in Figure 114, we can see the important changes in the spectra of the $\mathbf{1} \cdot \boldsymbol{PF_6}$, with the dimerization, the pics corresponding to the protons of the alkene chain disappeared, associated with a shift of the NCH_2 pic ($\Delta\delta H_{NCH2}$) = –0.3 ppm). As we mentioned before, the ^{1}H NMR attested to the formation of the $\mathbf{1} \cdot (\boldsymbol{PF_6})_2$ porphyrin acridinium conjugate. The U.V./Visible spectrum of the molecule will be presented in the next section.

V. C. 1. B - UV/VISIBLE SPECTRUM OF $(\mathbf{1} \cdot (\boldsymbol{PF_6})_2)$

The formation of the named $\mathbf{1} \cdot (\boldsymbol{PF_6})_2$ porphyrin acridinium conjugate was also confirmed using U.V./visible spectroscopy. The spectrum (Figure 115) shows the corresponding Soret band at 427nm and the two Q-bands at 554 nm and 598 nm, respectively, which are characteristic of the formation of metalled porphyrins (without metals, we would have four Q-

bands). In addition, we can see a band at 364 nm representing the π-π* transition centred on the acridinium fragment.[132]

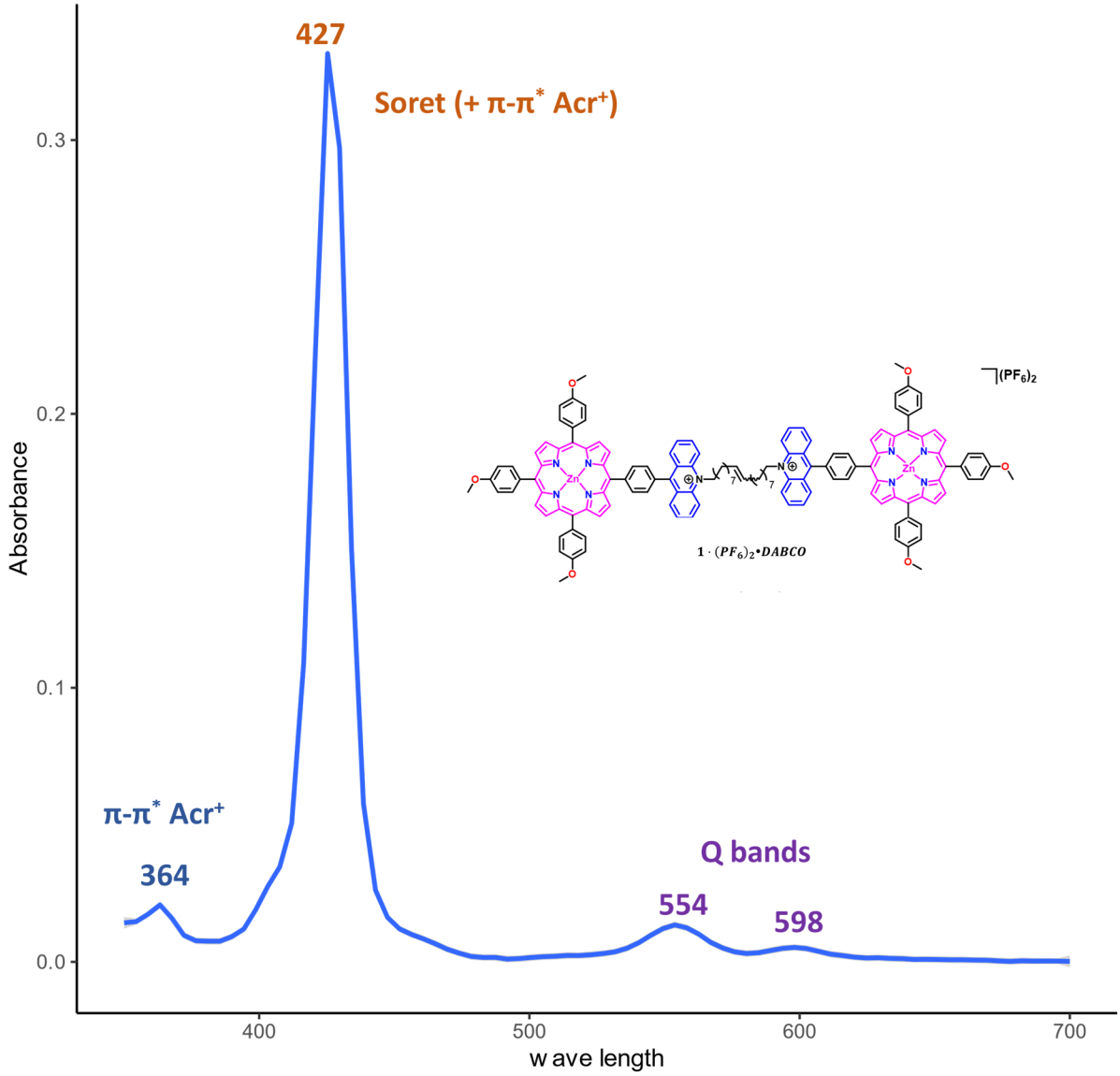


Figure 115: UV/Visible spectrum (CH_2Cl_2, 298 K, l = 1cm, c = $4.37.10^{-6}$ M) of the tweezer $(\mathbf{1} \cdot (\boldsymbol{PF_6})_2)$

Even if the $\mathbf{1} \cdot (\boldsymbol{PF_6})_2$ was not isolated as a pure compound (~5-10% of impurities), and correlated with the computational analysis coming in the next section, we decided to make a proof of concept of the ability for the porphyrin acridinium to bind to a ditopic ligand and polyaromatic molecules. As everything depends on our capacity to synthesise the given product, we wanted to try before investing time in the purification process and scale up to obtain sufficient amount of the product.

V. C. 2 - NMR ANALYSIS OF THE $(\mathbf{1}\cdot(\boldsymbol{PF_6})_2\bullet\boldsymbol{DABCO})$

In order to confirm the capability of $\mathbf{1}\cdot(\boldsymbol{PF_6})_2$ to complex **DABCO** and form a macrocyclic receptor, one equivalent of **DABCO** was added to a solution of $\mathbf{1}\cdot(\boldsymbol{PF_6})_2$ in CD_2Cl_2, and a ^{1}H-NMR recorded (Figure 116).

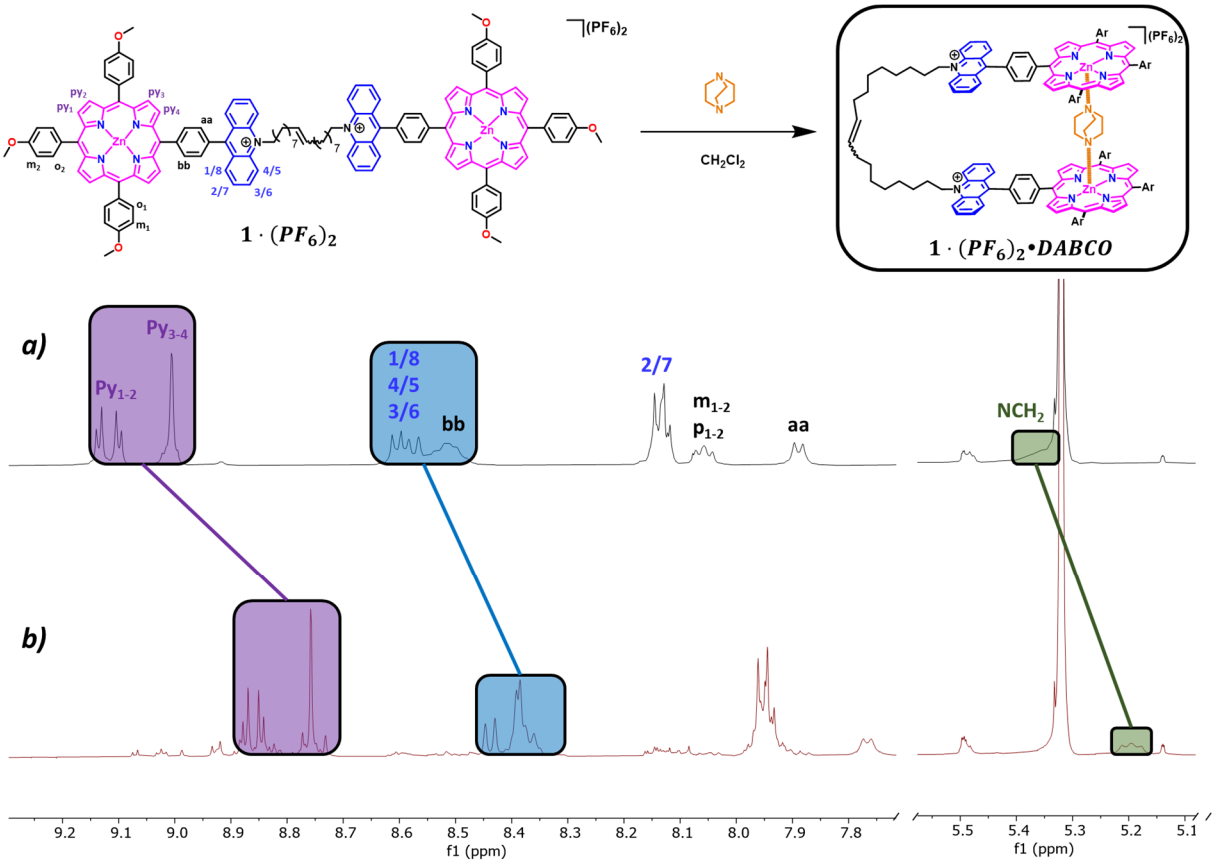


Figure 116: Stacked ^{1}H NMR spectra ($CDCl_3$, 298 K, 500MHz) of a) $\mathbf{1}\cdot(\boldsymbol{PF_6})_2$ b) $\mathbf{1}\cdot(\boldsymbol{PF_6})_2\bullet\boldsymbol{DABCO}$

Even an excess of DABCO can be shown, a notable change in the spectrum can be evidenced with the addition of DABCO, with upfield shifts of the pyrrolic protons ($\Delta\delta(Py_{1\text{-}2}) = \Delta\delta(Py_{3\text{-}4})$ = –0.25 ppm, upfield shifts of the smaller magnitude of the acridinium peaks ($\Delta\delta(H_{1/8}) = \Delta\delta(H_{2/7}) = \Delta\delta(H_{3/6}) = \Delta\delta(H_{4/5})$ = –0.1 ppm), and upfield shift of the NCH_2 ($\Delta\delta(NCH_2$ = –0.15ppm). This NMR evidenced binding of the DABCO by the porphyrin and the formation of the macrocyclic receptor $\mathbf{1}\cdot(\boldsymbol{PF_6})_2\bullet\boldsymbol{DABCO}$.

V. C. 3 - NMR ANALYSIS OF ($[\mathbf{1}\cdot(PF_6)_2\bullet DABCO]\supset Perylene$)

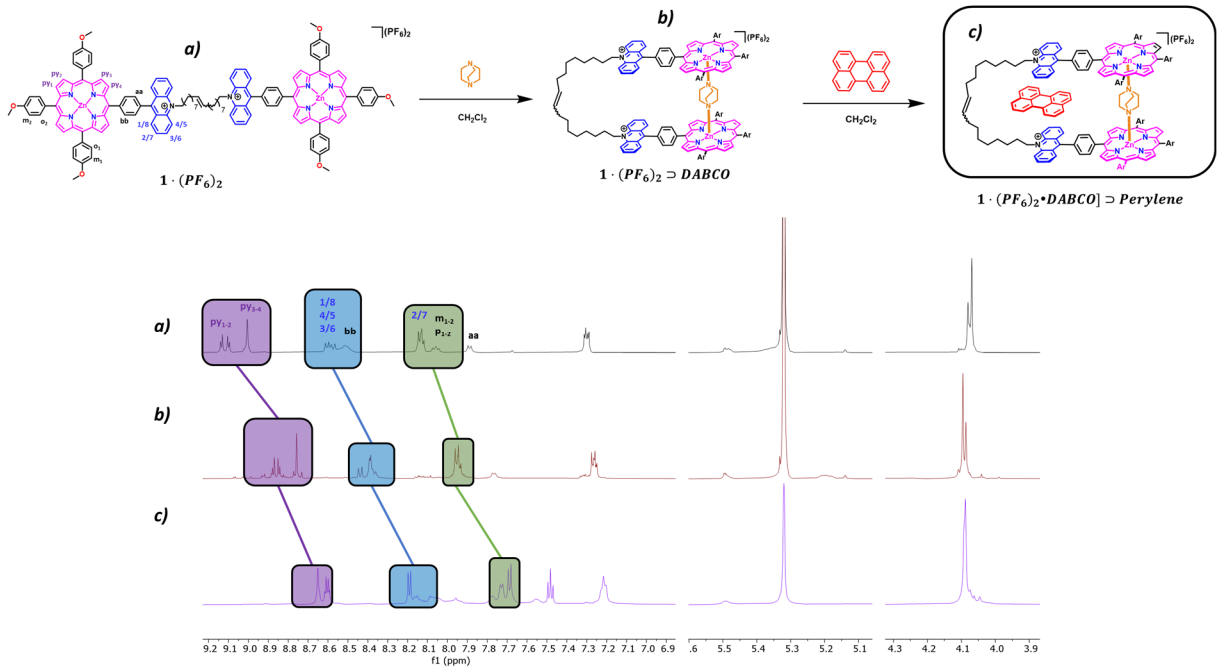


Figure 117: Stacked ^{1}H NMR spectra ($CDCl_3$, 298 K, 500MHz) of a) $\mathbf{1}\cdot(PF_6)_2$ b) $\mathbf{1}\cdot(PF_6)_2\bullet DABCO$ and c) $[\mathbf{1}\cdot(PF_6)_2\bullet DABCO]\supset Perylene$

With the previous analysis suggesting confirmation of the binding of the DABCO, the ability to bind polyaromatic guests in an allosteric manner was studied by ^{1}H-NMR. For that, one equivalent of perylene is added to the previous $\mathbf{1}\cdot(PF_6)_2\bullet DABCO$ in CD_2Cl_2 (Figure 117). The acridinium peaks present a slight shift, suggesting a possible binding of the perylene inside the binding cavity. ($\Delta\delta$(Py$_{1\text{-}2}$) = $\Delta\delta$(Py=$_{3\text{-}4}$) = –0.2 ppm, ($\Delta\delta$(H$_{1/8}$) = $\Delta\delta$(H$_{2/7}$) = $\Delta\delta$(H$_{3/6}$) = $\Delta\delta$(H$_{4/5}$) = –0.08 ppm). Additionally, sub-peaks can be shown representing a possible pre-existing excess of DABCO or $\mathbf{1}\cdot(PF_6)_2$ due to the difficulties of measuring precisely one equivalent of each product in the previous steps. This suggests a possible increase in the electronic shielding of the acridinium protons upon the addition of pyrene, consistent with an acridinium-pyrene π-π interaction.

Along with the synthesis process, a computational exploration was initiated to identify possible binders for the new receptor.

VI - IDENTIFICATION OF POTENTIAL BINDERS FOR THE NEW Zn(II)-PORPHYRIN ACRIDINIUM: A FUTURE PERSPECTIVE

VI. A - PROTOCOL

VI. A. 1 - PHARMACOPHORIC SEARCH AND VIRTUAL SCREENING

Virtual screening is a computational technique used in drug discovery to search libraries of small molecules to identify structures most likely to bind to a drug target, in our case, the Zn(II)-porphyrin-acridinium receptor. Historically, virtual screening improves hit rates and reduces costs by generating small, highly enriched subsets of compound libraries that could be physically screened. Pharmacophore search is an established and effective strategy for virtual screening.[165] Our thesis uses two different databases and molecular docking to rank molecules that will be analysed through the HG-DYNAusor platform.

A pharmacophore describes the structural arrangement of essential interaction features. Common pharmacophore features include several steric and electronic features describing the molecular interactions (hydrophobic, hydrogen bonds, charged, …). A pharmacophore query is defined by the spatial arrangement of features and a search radius around each feature. The pharmacophore of a molecule describes the essential characteristic describing the interactions.[166]

VI. A. 2 - DATABASES

VI. A. 2. A - DRUG BANK

The latest release of the DrugBank dataset (released 2021-01-03) contains 14575 drug entries. These drugs are separated into the following groups: 2700 approved small molecule drugs, 1496 approved biologics (proteins, peptides, vaccines, and allergenics), 132 nutraceuticals, and over 6652 experimental (discovery-phase) drugs.[167–171]

VI. A. 2. B - T3DB

The Toxin and Toxin Target Database (*T3DB*) or, is a unique bioinformatics resource that combines toxin data with comprehensive toxin target information. The database currently contains 3,678 toxins described by 41,602 synonyms, including pollutants, pesticides, some drugs, and food toxins, for which the toxicity record can be shown. This information has been

extracted from over 18,143 sources, including other databases, government documents, books, and scientific literature.[172,173]

VI. B - BEHAVIOUR OF NEW Zn(II)-PORPHYRIN ACRIDINIUM HOST

For the Zn(II)-porphyrin-acridinium with para-anisyl solubilising groups, three different molecular dynamics are launched from three different starting points. As we did before, a set of molecular descriptors describing the deviation (RMSD), the Rg and the SASA of the receptor are calculated. In addition, other descriptors related to the geometry of the receptors in more detail are added to the analysis. These descriptors are separated into three types: distance descriptors, angles descriptors, and dihedrals angles descriptors, which represent the dynamics of the alkene chain and the acridinium. In total, 33 additional descriptors are added. As the dataset contains much more information than in previous times, the variability is more diluted in the different PCs. Before the analysis, a dimensional reduction process is done, and a set of 12 variables with largest contributions are selected for further analysis.

The results can be shown in Figure 118: ~80% of the variability is explained by the four first PCs. Concerning the "MD2" and the "MD3" respectively coloured in green and red in the graphic, they sample a very similar conformational space. "MD1" presents some variation: although the largest part of the dynamics overlaps with the two others, some geometries are only sampled by this system, which likely represents a particular conformation associated with a rare event. Some specific geometries are highlighted in the bottom part of Figure 118. In blue, the rare conformation highlighted with *1 in the graphic represents a conformation of the receptor where the alkene chain enters the cavity, generating steric hindrance and preventing the binding of a potential guest.

Most of the geometries present an accessible binding cavity where the receptor is in an open configuration allowing binding within the cavity between the two acridiniums (*2 in blue and *1 and *2 in orange in Figure 118). This shape is predominant in all simulations, so the points are mostly located in the same area and overlapping. For MD3 (green), two specific geometries can be extracted: the first one (green *1) shows the closed conformation of the receptor with the two acridiniums interacting with each other. In contrast, the second represents the very open conformation, where the two acridiniums have rotated by 90° and are no longer facing each other. The alkene chain seems to be flexible enough to allow this conformational change in the structure.

From the observations of molecular dynamics, the predominant open conformation can take several orientations over time. From what we have seen in the simulations, the open conformation can be considered as an equilibrium point from which the geometry is likely to diverge to another conformation (closed, semi-closed, rotated...). We expect that the geometry with the alkene chain entering the cavity will have a lower absolute energy than the configuration presenting a binding cavity between the two acridiniums, which is generally the one used to bind a guest.

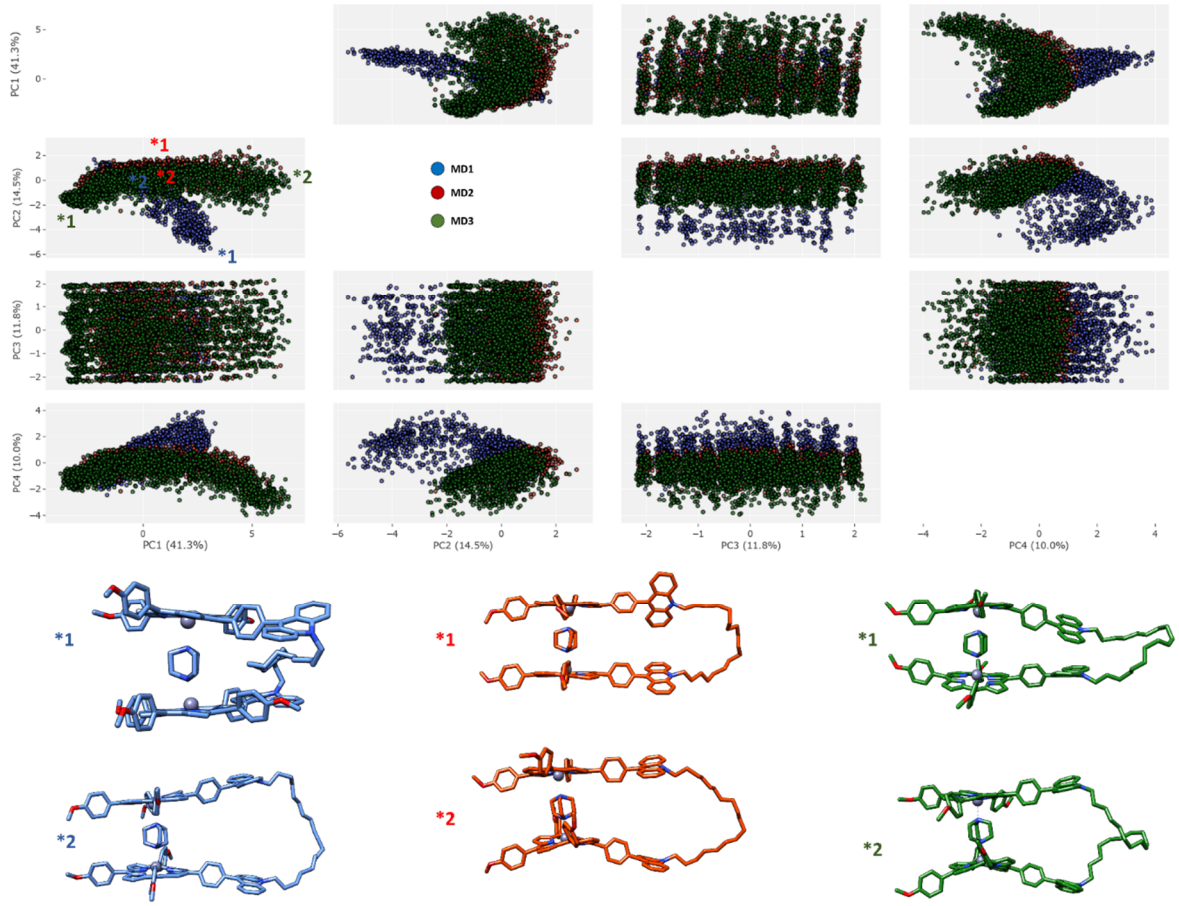


Figure 118: PCA of Zn(II)-porphyrin receptors chemical space described by a set of molecular descriptors derived from the simulations. The space formed by the combination of PC1, PC2, PC3, and PC4 explaining respectively, ~80% of the variability. Each point represents a geometry, and the point is coloured by the simulations: MD1 (blue), MD2 (red), MD3 (green). (upper part): the combination of each P.C.s (bottom part): an overview of some specific geometry.

VI. C - VIRTUAL SCREENING

VI. C. 1 - PHARMACOPHORIC SEARCH

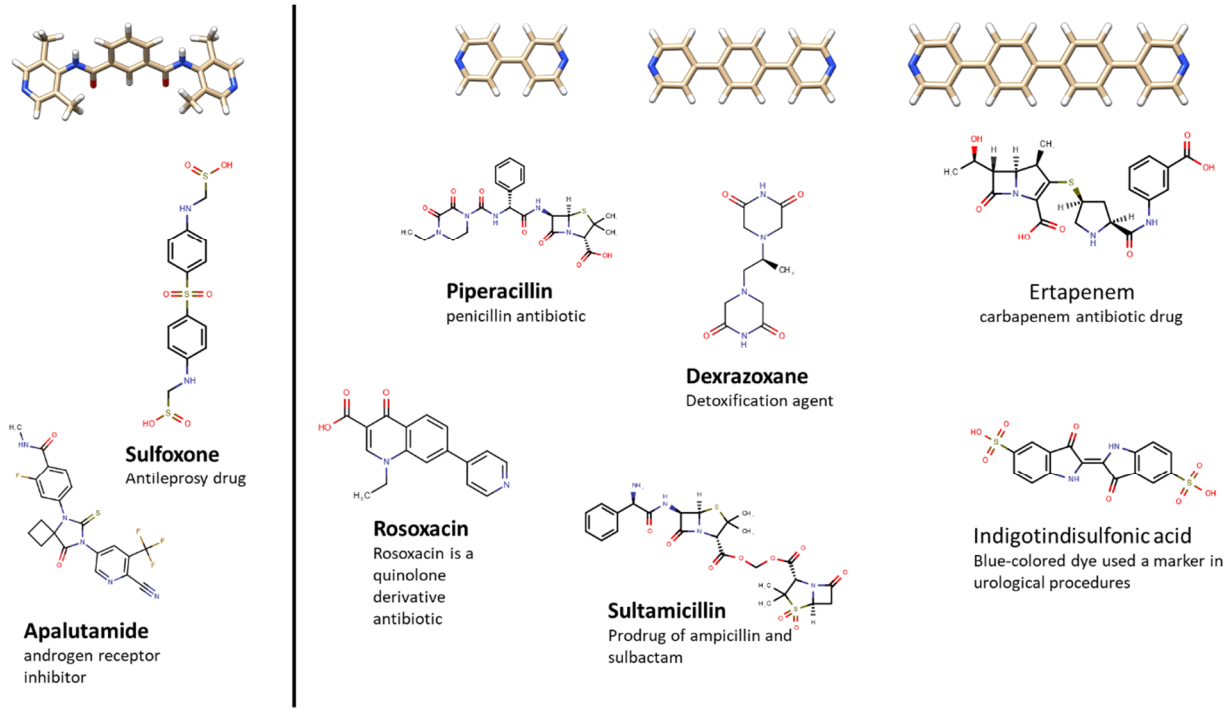


Figure 119: An overview of the results of the first pharmacophoric search

The pharmacophoric search is done with MOE[174] software using the constructed 3D databases. Four different pharmacophores are investigated (upper part of the Figure 119). Only the distance between the nitrogen and the donor or acceptor character is considered for each of them. The drug bank is used, and the different filters are applied. An overview of some results is shown in the Figure 119. The idea is to select molecules that would coordinate the two zinc molecules coming from the porphyrins. Unfortunately, most of the extracted molecules present steric groups near the nitrogen that are likely to reduce the affinity of this ditopic guest to the Zn(II), making their use complicated in this context. In conclusion, it would be necessary to generate more detailed filters to avoid the steric clashes.

VI. C. 2 - FILTERING

The number of molecules to be analysed was reduced using a filtering procedure. This filtering procedure can be summarised as follows:

1) The potential guests that contain metals are removed from the databases.
2) Mono-anionic and mono-cationic species are removed from the databases.
3) The molecules presenting a counter-ion are removed from the databases.
4) Peptides are removed from the databases.

5) The Solvent Accessible Surface Area (SASA) were measured for all remaining molecules. A threshold was applied from the SASA measurement, and all the guests that may not fit into the cavity were removed.
6) Molecules that present errors during 3D constructions were removed from the databases.

Using the filtering approach, both databases are reduced, and finally, the DrugBank is reduced to 5881 molecules and the T3DB to 1943 molecules.

VI. C. 3 - DOCKING RESULTS

In Table 17 and Table 18, the ten best results of the docking procedure are presented with the associated score. Each database has been considered separately, and the docking is realised once for each of them in receptor extracted from the previous MDs and presenting an open binding cavity. The docking box is sufficiently large to allow any type of external or internal interaction. On the top twenty results, all molecules are inside the binding cavity.

The ten best molecules extracted from the T3DB database are mainly pollutants. All of them can be classified as polycyclic aromatic hydrocarbons (PAH), susceptible to interact with a good affinity with the molecular tweezer using π-π interactions. Considering the score of the T3DB compounds, they are globally high, suggesting that these molecules are good binders for the receptor. Concerning the DrugBank, only the best result is a PAH molecule, but most of the ranked molecules can interact with the receptor via π-π interactions. Very interesting to highlight that from the ten best molecules, four of them present a steroid scaffold (Figure 120). That could suggest an interesting particularity of the receptor to bind this kind of molecules.

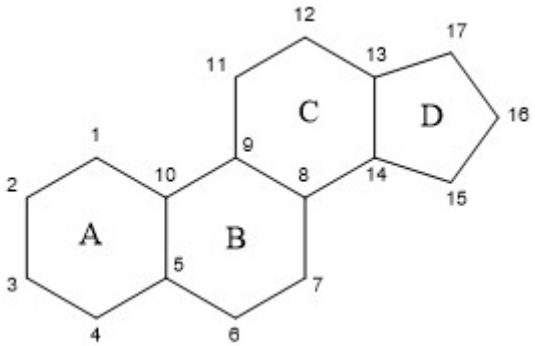


Figure 120: Representation of the steroid scaffold

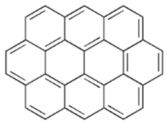
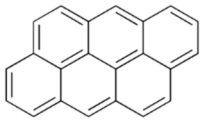
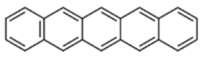
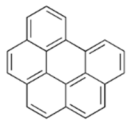
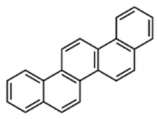
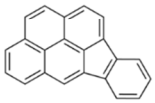
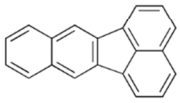
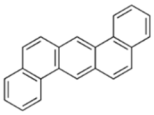
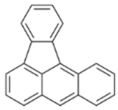
One of the main limitations of the docking process is the initial geometry of the receptor. Unlike the guest molecules, the host cannot be considered fully flexible in the docking algorithm. The initial structure was extracted from the previous simulation of molecular dynamics, and its

internal energy was not considered, but could be important to improve the results[175]. However, considering the low amount of time for running the docking, it is still possible to consider running an ensemble docking based on several different geometries presenting low energy.

Generally speaking, the docking results are consistent with the literature and the best ligands extracted are almost all molecules capable of π-π interactions with the receptor. Binding can be studied on these ten molecules by molecular dynamics following the protocol described in Chapter 3 concerning using the HG-DYNAusor platform. In conclusion, the top-five molecules of each database will be extracted, and the binding free energy will be measured for these ten molecules.

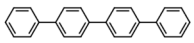
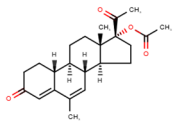
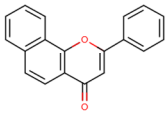
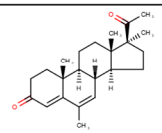
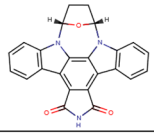
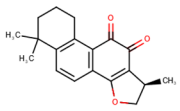
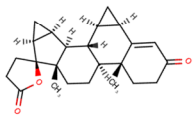
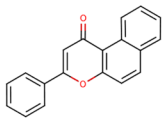
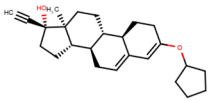
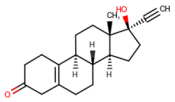
VI. C. 3. A - T3DB

Table 17: An overview of the results of the docking on the T3DB after the filtering procedure

Score	Name	structure	Score	Name	structure
-14.9	Ovalene		-11.9	Anthanthrene	
-12.3	Pentacene		-11.9	Benzo[ghi]perylene	
-12.3	Coronene		-11.8	Picene	
-12.2	Indeno(1,2,3-cd)pyrene		-11.7	Benzo[k]fluoranthene	
-12.1	Dibenzo[a,h]anthracene		-11.5	Benzofluoranthene	

VI. C. 3. B - DRUG-BANK

Table 18: An overview of the results of the docking on the T3DB after the filtering procedure

Score	Name	structure	Score	Name	structure
-11.3	pQuaterphenyl **Investigational**		-10.6	Nomegestrol **Approved**	
-11.0	alphaNaphthoflavone **Experimental**		-10.6	Medrogestone **Approved**	
-10.9	DB08683 **Experimental**		-10.6	Cryptotanshinone **Experimental**	
-10.8	Drospirenone **Approved**		-10.5	betaNaphthoflavone **Experimental**	
-10.8	Quingestanol **Experimental**		-10.5	Norethynodrel **Approved**	

VI. D - PERSPECTIVE

At the time of writing, this part is still in progress. As we already explained, the first try of the binding of $\mathbf{1} \cdot (\boldsymbol{PF_6})_2$ with DABCO and perylene shows sufficiently interesting results to investigate other possible binders. In parallel to the scale-up for the obtention of $\mathbf{1} \cdot (\boldsymbol{PF_6})_2$ with sufficient purity, the computational studies have been launched.

We aim to:

1) Extract several low-energy conformations of the $\mathbf{1} \cdot (\boldsymbol{PF_6})_2$ -DABCO system from the clustering analysis done on the three molecular dynamics simulations. If the extracted conformations are very different, the docking procedure can be repeated.
2) Perform MD simulations of the docking top-scoring molecules and predict the binding free energy for each of them using the HG-DYNAusor platform. Select the best ones for experimental testing.
3) Re-synthesise $\mathbf{1} \cdot (\boldsymbol{PF_6})_2$, investing time in the synthetic and purification process to obtain the tweezer with sufficient purity.
4) Experimentally determine the binding affinities of the selected ligands $\mathbf{1} \cdot (\boldsymbol{PF_6})_2$.

EXPERIMENTAL PART

I - SYNTHESIS OF 2-(4-BROMOPHENYL)-4,4,5,5-TETRAMETHYL-1,3-DIOXOLANE

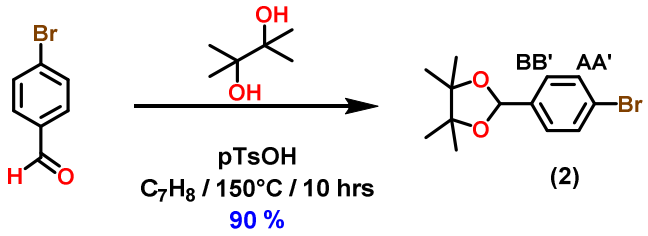


Figure 121: Chemical synthesis of the 2-(4-bromophenyl)-4,4,5,5-tetramethyl-1,3-dioxolane

To a solution of 4-bromobenzaldehyde (2.004 g, 10.83 mmol, 1 eq.) and tetramethylethylene glycol (1.42 g, 12.01 mmol, 1.11 eq.) in toluene (100 mL), para-tosylic acid (0.55 g) was added. After Dean-Stark extraction for ~3 hours, the reaction mixture was washed with a saturated aqueous solution of Na_2CO_3 (2 × 100mL), and the aqueous layer was extracted using ethyl acetate (2 × 100 mL).

The organic layers were combined and dried ($MgSO_4$). After evaporation of the solvents, the crude product was purified by column chromatography (SiO2, petroleum ether/EtOAc, 5% then 7%). Three different fractions have been extracted and submitted to the NMR. Due to the purity of the three fractions, they are merged in Ch_2Cl_2 and evaporated. The desired product was obtained as colourless crystals in 90% yield (2.762 g).

^{1}H NMR (500 MHz, $CDCl_3$, 298 K) δ (ppm) = 7.52 – 7.45 (m, 2H, $H_{AA'}$), 7.39 – 7.33 (m, 2H, $H_{BB'}$), 5.93 (s, 1H, O_2CH), 1.31 (s, 6H, CH_3), 1.24 (s, 6H, CH_3).

^{13}C NMR (126 MHz, $CDCl_3$, 298 K) δ (ppm) = 139.1($CC_{BB'}$), 131.5($C_{AA'}$), 128.1($C_{BB'}$), 122.7($CC_{AA'}$), 99.3(O_2CH), 83.0((CH_3)2C), 24.4(CH_3), 22.3(CH_3).

II - SYNTHESIS OF 10-(BUT-EN-1-YL)ACRIDIN-9(10H)-ONE

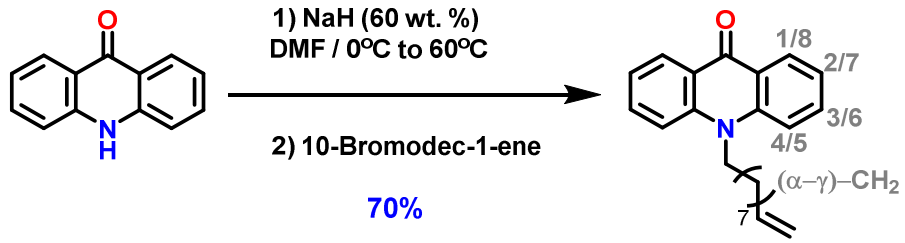


Figure 122: Chemical synthesis of the 10-(but-en-1-yl)acridin-9(10H)-one

To a solution of 9(10H)-Acridone (780 mg, 4 mmol, 1.0 eq.) in anhydrous DMF (40 mL), was added NaH 60 wt. % dispersion in mineral oil (240 mg, 6 mmol, 1.5 eq.) at 0°C. The solution was stirred for 30 minutes before adding 1-(10)Bromohex-1-ene (1.20 mL, 6 mmol, 1.5 eq.). The mixture was then heated to 60°C overnight. After the addition of H_2O (500 mL), the aqueous layer was extracted using Et_2O (3 × 100 mL). After evaporation of the solvents, the resulting oil was precipitated from H_2O (100 mL). After filtration, the crude product was purified by column chromatography (SiO_2, CH_2Cl_2/acetone, 100:0 to 99:1). The desired product 2 was obtained as a pale-yellow solid in 70% yield (783 mg).

^{1}H NMR (500 MHz, $CDCl_3$, 298 K) δ (ppm) = 8.60 (dd, J = 8.0, 1.8 Hz, 2H, $H_{1/8}$), 7.74 (ddd, J = 8.7, 7.0, 1.8 Hz, 2H, $H_{3/6}$), 7.51 (d, J = 8.7 Hz, 2H, $H_{4/5}$), 7.30 (ddd, J = 7.9, 7.0, 0.9 Hz, 2H, $H_{2/7}$), 5.82 (ddt, J = 17.0, 10.2, 6.7 Hz, 1H, H_2C=C*H*), 5.01 (dq, J = 17.0, 2.2 Hz, 1H, (trans)H*H*C=CH), 4.95 (ddt, J = 10.2, 2.2, 1.3 Hz, 1H, (cis)*H*HC=CH), 4.38 – 4.31 (m, 2H, NCH_2), 2.06 (tdd, J = 6.7, 5.3, 1.3 Hz, 2H, γ-CH_2), 2.00 – 1.90 (m, 2H, α-CH_2), 1.58 – 1.52 (m, 2H, β-CH_2), 1.50 – 1.42 (m, 2H, δ-CH_2), 1.41 – 1.30 (m, 4H, (χ&ϕ)-CH_2).

^{13}C NMR (126 MHz, $CDCl_3$, 298 K) δ (ppm) = 178.2 (*C*O), 141.9 ($C_{4/5}$*C*), 139.2 (H_2C=*C*), 134.0 ($C_{3/6}$), 128.2 ($C_{1/8}$), 122.6 ($C_{1/8}$*C*), 121.4 ($C_{2/7}$), 114.7 ($C_{4/5}$), 114.4 (H_2*C*=C), 46.4 (N*C*H_2), 33.9 (γ-*C*H_2), 29.6 (χ-*C*H_2), 29.5 (δ-*C*H_2), 29.2 (ε-*C*H_2), 29.0 (ϕ-*C*H_2), 27.3 (α-*C*H_2), 27.1 (β-*C*H_2).

III - SYNTHESIS OF THE 10-ALLYL-9-(4-FORMYLPHENYL)ACRIDIN-10-IUM

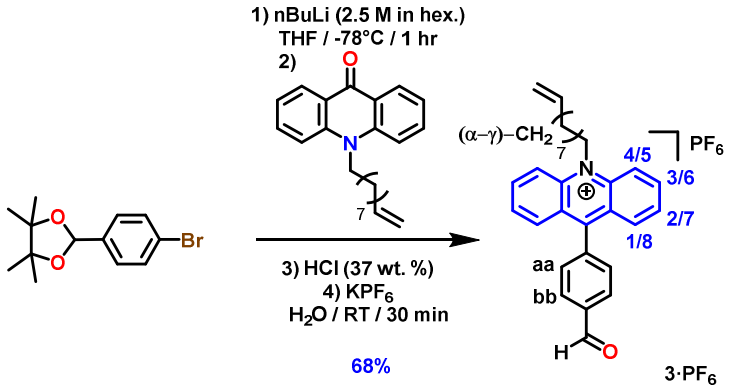


Figure 123: Chemical synthesis of the synthesis of the 10-allyl-9-(4-formylphenyl)acridin-10-ium ($3 \cdot PF_6$)

To a solution of 2-(4-bromophenyl)-4,4,5,5-tetramethyl-1,3-dioxolane (**2**) (770 mg, 2.70 mmol, 1.2 eq.) in anhydrous THF (30 mL), was added a 2.5M solution of nBuLi in hexanes (0.9 mL, 2.25 mmol, 1.0 eq.) at –78°C. After 1 hours, 10-(dec-9-en-1-yl)acridin-9(10H)-one (2) (750 mg, 2.25 mmol) was added. The resulting mixture was allowed to room temperature overnight. A 37 wt.% aqueous solution of HCl (30 mL) was added, and after 30 min, the acidified solution was poured slowly into an aqueous solution of KPF_6 (6g, 150 mL). The resulting oil was extracted with CH_2Cl_2 (3 × 100 mL), and the combined organic layers were washed with H_2O (2 × 100 mL). After evaporation of the solvents, the crude product was precipitated using ACN / diethyl ether mixture, giving the desired product $\mathbf{3 \cdot PF_6}$ was obtained as a yellow solid in 62% yield.

^{1}H NMR (500 MHz, $CDCl_3$, 298 K) δ (ppm) = 10.25 (s, 1H, C*H*O), 8.48 (d, *J* = 9.3 Hz, 2H, $H_{4/5}$), 8.42 (ddd, *J* = 9.3, 6.6, 1.5 Hz, 2H, $H_{3/6}$), 8.27 – 8.22 (m, 2H, H_{bb}), 7.98 (dd, *J* = 8.7, 1.5 Hz, 2H, $H_{1/8}$), 7.87 (ddd, *J* = 8.7, 6.6, 1.0 Hz, 2H, $H_{2/7}$), 7.73 – 7.67 (m, 2H, H_{aa}), 5.84 (ddt, *J* = 17.0, 10.2, 6.7 Hz, 1H, H_2C=C*H*), 5.38 – 5.31 (m, 2H, NCH_2), 5.01 (dq, *J* = 17.0, 1.7 Hz, 1H, (trans)H*H*C=CH), 4.93 (ddt, *J* = 10.2, 2.3, 1.2 Hz, 1H, (cis)*H*HC=CH), 2.33 – 2.22 (m, 2H, α-CH_2), 2.11 – 2.03 (m, 2H, γ-CH_2), 1.78 (p, *J* = 7.6 Hz, 2H, β-CH_2), 1.57 – 1.49 (m, 2H, χ-CH_2), 1.46 – 1.35 (m, 6H, (δ - ϕ)-CH_2).

^{13}C NMR (126 MHz, $CDCl_3$, 298 K) δ (ppm) = 191.7 (*C*HO), 160.6 (C_{aa}*C*), 141.0 ($C_{4/5}$*C*), 140.0 ($C_{3/6}$), 139.6 (C_{aa}*C*), 138.9 (H_2C=*C*H), 138.0 (*C*CHO), 131.0 (C_{bb}), 130.8 ($C_{1/8}$), 130.3 (C_{aa}), 128.7 ($C_{2/7}$), 126.3 ($C_{1/8}$*C*), 118.2 ($C_{4/5}$), 114.3 (H_2*C*=CH), 51.7 (NCH_2), 34.1 (γ-*C*H_2), 29.7 (δ-*C*H_2), 29.6 (χ-*C*H_2), 29.3 (ϕ-*C*H_2), 29.3 (ε-*C*H_2), 29.2 (α-*C*H_2), 27.1 (β-*C*H_2).

31P NMR (121 MHz, CD_2Cl_2, 298 K) δ (ppm) = -144.78 (sept, J = 711.0 Hz).

19F NMR (282 MHz, CD_2Cl_2, 298 K) δ (ppm) = -73.58 (d, J = 710.9 Hz).

IV - SYNTHESIS OF PORPHYRIN-ACRIDINIUM CONJUGATE (4 · PF_6)

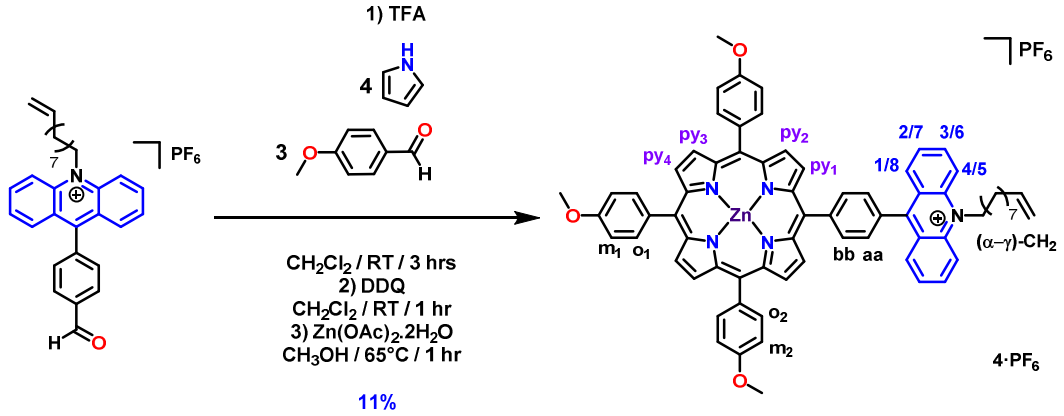


Figure 124: Chemical synthesis of the porphyrin-acridinium conjugate (4 · PF_6)

To a solution of 10-(dec-9-en-1-yl)-9-(4-formylphenyl)acridin-10-ium (**4**) (1 g, 1.76 mmol, 1.0 eq.), pyrrole (0.489 mL, 7.04 mmol, 4 eq.), and 4-methoxybenzaldehyde (6.642 mL, 5.28 mmol, 3 eq.) in degassed CH_2Cl_2 (450 mL), was added TFA (0.866 mL, 3 mmol, 6 eq.). After three hours of stirring in the dark, DDQ (1.2 g, 5.28 mmol, 3 eq.) was added. The reaction mixture was stirred at room temperature for an additional hour, and the solution was neutralised using triethylamine (12 mL). After evaporating the solvents, the resultant solid was dissolved in CH_2Cl_2 and filtered through a plug of SiO_2. After evaporation of the solvents, a solution of $Zn(OAc)_2{\cdot}2H_2O$ (386 mg, 1.76 mmol, 1 eq.) in CH_3OH (330 mL) was then added. The mixture was heated to reflux for one hour with completion of the reaction assessed via UV-Vis. After evaporating the solvents, the crude product was purified by column chromatography (SiO_2, $CHCl_3$/ACN, 100:0 to 95:5) followed by precipitation from ACN/Petroleum ether (1:10) and finally recrystallised in toluene. The desired product **4 · PF_6** was obtained as a dark-blue solid in 11% yield (234 mg).

^{1}H NMR (500 MHz, $CDCl_3$, 298 K) δ (ppm) = 9.13 (d, J = 4.6 Hz, 2H, , H_{py1}), 9.10 (d, J = 4.6 Hz, 2H, , H_{Py2}), 9.04 (s, 4H, $Py_{3\text{-}4}$), 8.61 (t, J = 8.8 Hz, 4H, $H_{bb\text{-}1/8}$), 8.55 – 8.48 (m, 4H, $H_{3/6\text{-}4/5}$), 8.18 (t, J = 9.1 Hz, 6H, $H_{m\text{-}m'}$), 8.04 (dd, J = 8.6, 6.6 Hz, 6H, $H_{p\text{-}p'}$), 7.93 (d, J = 7.9 Hz, 2H, $H_{2/7}$), 7.37 – 7.31 (m, 6H, H_{aa}), 5.87 (ddt, J = 16.9, 10.2, 6.7 Hz, 1H, H_2C=C*H*), 5.51 (t, J = 8.7 Hz, 2H, NCH_2), 5.09 – 4.95 (m, 2H, (HHC=CH), 4.14 (s, 6H, OCH_3), 2.54 – 2.25 (m, 2H, , α- γCH_2), 2.16 – 1.99 (m, 2H, , α- γCH_2), 1.93 – 1.75 (m, 2H, α- γCH_2), 1.47 – 1.43 (m, 10H, , α- γCH_2)

^{13}C NMR (126 MHz, CD_2Cl_2, 298 K) δ (ppm) = 207.13 (s, C_{OMe}), 140.69 (s, C_9), 139.21(s, $C_{q\text{-}OMe}$), 139.08 (s, $C_{q\text{-}porph}$), 130.70 (s, $C_{q\text{-}porph}$), 128.02 (s, $C_{q\text{-}porph}$), 118.06, 114.22, 77.23(s, $C_{q\text{-}OMe}$), 76.97(s, $C_{q\text{-}Ac^+}$), 76.72(s, $C_{4/5\ \&\ 3/6}$), 55.51 (s, $C_{N\text{-}CH2}$), 33.71(s, $C_{aliphatic}$), 30.90(s, $C_{aliphatic}$), 29.65(s, $C_{aliphatic}$), 29.30(s, $C_{aliphatic}$), 29.26(s, $C_{aliphatic}$), 29.07(s, $C_{aliphatic}$), 28.96(s, $C_{aliphatic}$), 28.80(s, $C_{aliphatic}$), 26.70(s, $C_{aliphatic}$), 25.37(s, $C_{aliphatic}$), 22.65(s, $C_{aliphatic}$).

31P NMR (203 MHz, CD_2Cl_2, 298 K) δ (ppm) = -137.49 – -151.17 (sept, J = 712.0 Hz),

19F NMR (471 MHz, CD_2Cl_2, 298 K) δ (ppm) = -74.59 (d, J = 712.0 Hz).

HRMS (ESI-TOF): for $C_{70}H_{60}N_5O_3Zn$, m/z_{calc} = 1082.3988, m/z_{found} = 1082.3982 (100%, $[M]^+$).

V - SYNTHESIS OF $1 \cdot (PF_6)_2$

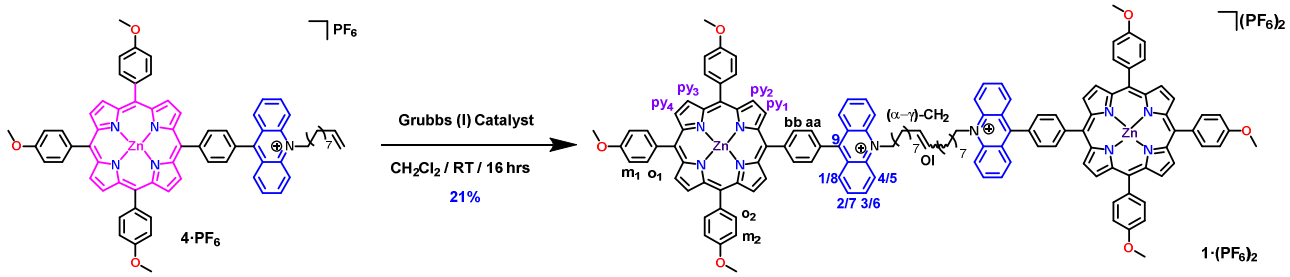


Figure 125: Chemical synthesis of the porphyrin-acridinium tweezer ($1 \cdot (PF_6)_2$)

To a solution of Zn(II)porphyrin $\mathbf{4 \cdot PF_6}$ (50 mg, 1eq.) in degassed anhydrous CH_2Cl_2 (4mL), was added Grubbs (I) catalyst (3.6 mg, 10 mol%). After 48 hours in the dark, the solution was washed with brine (10 mL) and H_2O (3 × 10 mL). After evaporating the solvents, the crude product was purified by column chromatography (SiO_2, $CHCl_3/CH_3CN$, 98:2 to 97:3) followed by crystallisation from acetonitrile/petroleum ether (1:3). The desired product $\mathbf{1 \cdot (PF_6)_2}$ was obtained as a purple crystalline solid in 25% yield.

^{1}H NMR (500 MHz, CD_2Cl_2, 298 K) δ (ppm) = 9.13 (d, ^3J = 4.6 Hz, 4H, H_{py1}), 9.09 (d, ^3J = 4.6 Hz, 4H, H_{Py2}), 8.99 (s, 8H, $Py_{3\text{-}4}$), 8.57 (d, ^3J = 7.7 Hz, 4H, H_b), 8.47 (d, ^3J = 8.3 Hz, 4H, $H_{1/8}$), 8.39 – 8.36 (m, 8H, $H_{3/6\text{-}4/5}$), 8.11 (d, ^3J = 8.5 Hz, 12H, $H_{m\text{-}m'}$), 7.95 (ddd, J = 8.6, 5.1, 2.3 Hz, 4H, $H_{2/7}$), 7.81 (d, J = 7.7 Hz, 4H, H_{aa}), 7.34 – 7.16 (m, 12H, $H_{o1\text{-}o2}$, 5.48 (t, ^4J = 3.6 Hz, 2H, H_{ol}), 5.19 – 5.17 (m, 4H, NCH_2), 4.02 (s, 6H, OCH_3), 4.01 (s, 12H, OCH_3), 2.28-2.23 (m, 4H, CH_2), 1.81 – 1.75 (m, 4H, CH_2), 1.58 – 1.55 (m, 4H, CH_2), 1.47 – 1.42 (m, 20H, CH_2).

^{13}C{^{1}H} NMR (126 MHz, CD_2Cl_2, 298 K) δ (ppm) = 206.89 (s, C_{OMe}), 162.48 (s, C_9), 159.80 (s, $C_{q\text{-}OMe}$), 151.18(s, $C_{q\text{-}porph}$), 151.02(s, $C_{q\text{-}porph}$), 150.91(s, $C_{q\text{-}porph}$), 149.99 (s), 146.03 (s, $C_{q\text{-}Ac}^+$), 140.96 (s), 139.81(s, $C_{4/5,\ 3/6}$), 135.88 (s), 135.85 (s), 135.38 (s), 135.17 (s, C_b), 132.75 (s, C_{py2}), 132.56 (s, $C_{py3/4}$),, 132.44 (s, $C_{12/13}$),, 132.26 (s, C_{py1}), 131.67 (s), 131.39 (s, $C_{1/8}$), 130.82(s, C_{Ol}), 128.77 (s, C_{aa}), 128.51 (s, $C_{2/7}$), 126.70 (s), 121.87 (s), 121.52 (s, $C_{q\text{-}OMe}$), 119.00 (s, $C_{q\text{-}Ac}^+$), 118.02 (s, $C_{4/5\ \&\ 3/6}$), 112.51 (s), 55.92 (s, $C_{N\text{-}CH2}$), 32.94 (s, $C_{aliphatic}$), 30.94 (s, $C_{aliphatic}$), 30.09 (s, $C_{aliphatic}$), 29.99 (s, $C_{aliphatic}$), 29.79 (s, $C_{aliphatic}$), 29.62 (s, $C_{aliphatic}$), 29.37 (s, $C_{aliphatic}$), 29.25 (s, $C_{aliphatic}$), 27.17 (s, $C_{aliphatic}$).

31P NMR (121 MHz, CD_2Cl_2, 298 K) δ (ppm) = -133.85 – -154.92 (sept, , J = 711.0 Hz).

19F NMR (282 MHz, CD_2Cl_2, 298 K) δ (ppm) = -73.31 (d, J = 712.0 Hz).

HRMS (ESI-TOF): for $C_{138}H_{116}N_{10}O_6Zn_2$, m/z_{calc} = 1068.3830, m/z_{found} = 1068.3825 (100%, $[M]^{2+}$).

UV-Vis (CH_2Cl_2, 298 K) λmax (nm) (ε (L mol^{-1} cm^{-1}) = 364 (4980), 427 (78600), 554 (3200), 598 (1300),

VI - SYNTHESIS OF $\mathbf{1 \cdot (PF_6)_2 \bullet DABCO}$

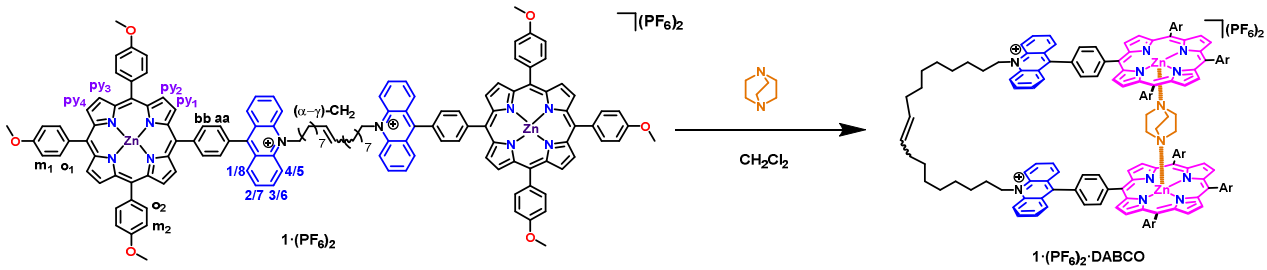


Figure 126: Synthesis of $\mathbf{1 \cdot (PF_6)_2 \bullet DABCO}$

To a solution of $\mathbf{1 \cdot (PF_6)_2}$ (2 mg, 0.87 µmol) in CD_2Cl_2 (2 mL) was added a solution of 1,4-diazabicyclo[2.2.2]octane (**DABCO**) (1 eq) in CD_2Cl_2 (0.1 mL). The mixture was manually stirred, and formation of the $\mathbf{1 \cdot (PF_6)_2 \bullet DABCO}$ complex assessed via ^{1}H NMR spectroscopy.

VII - SYNTHESIS OF $[\mathbf{1} \cdot (PF_6)_2 \bullet DABCO] \supset Perylene$

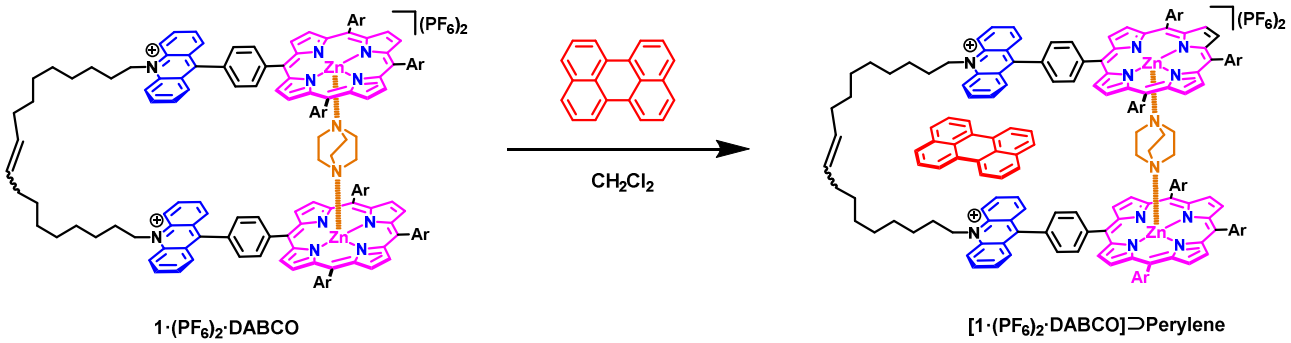


Figure 127: Synthesis of $[\mathbf{1} \cdot (PF_6)_2 \bullet DABCO] \supset Perylene$

To a solution of $\mathbf{1} \cdot (PF_6)_2 \bullet DABCO$ (0.87 µmol) in CD_2Cl_2 (2 mL) was added a solution of perylene (1eq, 0.87 µmol) in CD_2Cl_2 (0.1 mL). The mixture was manually stirred, and the formation of the $[\mathbf{1} \cdot (PF_6)_2 \bullet DABCO] \supset Perylene$ complex assessed via ^1H NMR spectroscopy.

LIST OF THE FIGURES

Figure 1: Classification of the supramolecular complex[14] 7

Figure 2: Flowchart of the molecular modelling approaches 16

Figure 3: Artificial intelligence history and description inspired by nvidia[75] 18

Figure 4: General aim of the project: multiple approaches we used for the exploration of the host-guest complexes 19

Figure 5: Overview of the GFN family of methods with main components and classification of the most important terms. Dark grey shaded areas denote a quantum mechanical description, while light grey parts indicate a classical or semi-classical description[94] 28

Figure 6: A schematic one-dimensional energy surface: starting from three different geometries, the minimization methods move downhill to the nearest minimum[95] 33

Figure 7: Schematic view of the periodic boundary conditions inspired by ISAAC program[104] 40

Figure 8: The schematic transformation of the Ewald summation method 42

Figure 9: Representation of a water molecule inside the TIP3P model with the values extracted from idc technical reference[106] 45

Figure 10: Implementation of implicit solvation on a solute 46

Figure 11: Geometric interpretation of PCA as the search for the best representation subspace of the considered sample[110] 50

Figure 12: Individuals analysis using the iris dataset provided by R and the factoextra package[111] 53

Figure 13: Variables analysis using the decathlon dataset provided by R and the factoextra package[111] 54

Figure 14: Screeplot using the decathlon dataset provided by R and the factoextra package[111] 54

Figure 15: The SVM residuals versus the predicted values[118] 59

Figure 16: Schematic representation of an artificial neuron where $= \boldsymbol{wj}, \boldsymbol{x} + \boldsymbol{netj}$ *..... 65*

Figure 17: Representation of an NNET 66

Figure 18: Illustration of a genetic algorithm[128] ... 70

Figure 19: An overview of the modules of the HG-DYNAusor platform; the three first modules are part of the thermodynamic based approach: (in red) module 1 dedicated to the parametrisation, (in purple) the module 2 dedicated to the binding mode generation for host-guest complexes, (in blue) the module 3 dedicated to the binding free energy calculation and in green the knowledge-based approach with the module 4 dedicated to the binding free energy prediction. 73

Figure 20: An overview of the docking results using AUTODOCKVina, starting from the crystal structure (A), for the docking of the crystal, one of the tweezer will be the host and considered as rigid while the other one considered as the guest will be allowed to rotate on some bonds (B), two different docking poses representing two different complexes can be extracted from the docking results (C), and one tweezer can be extracted, and his geometry optimised at DFT level (D), both using SaMD protocol and docking the DFT-optimised structure in the previous host lead to the formation of the similar complex (E). 76

Figure 21: Clustering of the MD of the tweezer-monomer, (A) Clustering of the closed-tweezer in orange, (B) Clustering of the semi-open or semi-closed form in yellow, and (C) clustering of the open-tweezer in green, each point represent an individual geometry, the points are colored by the frames of the simulations 78

Figure 22: Clustering of the MD of the tweezer-dimer, (A-B) Clustering of the closed-semi-open tweezer, (B-B) Clustering of the semi-open-semi-closed dimer, and (A-C) clustering of the open-closed dimer, each point represents an individual geometry, the points are colored by the frames of the simulations 79

Figure 23: Geometrical comparison between QM level (B3LYP) and two SQM methods (PM6 and GFN2-xTB), the number represents the frame of the MD from which the representative structure is extracted..... 80

Figure 24: Energetical analysis of the molecular tweezer 81

Figure 25: Overview of the first module of the HG-DYNAusor platform..... 85

Figure 26: Example of the substructure determination..... 86

Figure 27: Overview of the MD-RESP protocol for the generation of the atomic partial charges..... 91

Figure 28: MPD example: The aromatic carbon (in red) is described by his two neighbours..... 94

Figure 29: MPD example: the specific case of the hydrogens 95

Figure 30: An overview of the generation of the topological files..... 97

Figure 31: An overview of the last phases of module 01 of the HG-DYNAusor platform 99

Figure 32: An overview of the second module of the HG-DYNAusor platform dedicated to the binding mode generation 101

Figure 33: An overview of the third module of the HG-DYNAusor platform dedicated to the thermodynamic approach..... 103

Figure 34: Graphical overview of the knowledge-based protocol..... 107

Figure 35: Barplot of the binding free energy distribution of the different systems extracted from the BindingDB 108

Figure 36: Density plot of the distributed binding free energy of the different systems extracted from the BindingDB 109

Figure 37: PCA of the Guest chemical space before any dimensional reduction described by a set of molecular descriptors generated with CORINA. In (A), the space formed by the combination of PC1 and PC2 explains respectively 38.3% and 17.2% of the variability. In (B), the space formed by the combination of PC1 and PC3 explains respectively 38.3% and 4.8% of the variability. In (C), the space formed by the combination of PC2 and PC3 explains respectively 17.2% and 4.8% of the variability. In (D), the scree-plot represents the variability of all the principal components of the analysis. The molecules are coloured by the system there are supposed to interact with and, their size is a function of their binding free energy. 115

Figure 38: PCA of the Guest chemical space after a dimensional reduction using near-zero-variance and correlated approach described by a set of molecular descriptors generated with CORINA. In (A), the space formed by the combination of PC1 and PC2 explains respectively 39.5% and 17.0% of the variability. In (B), the space formed by the combination of PC1 and PC3 explains respectively 39.5% and 4.7% of the variability. In (C), the space formed by the combination of PC2 and PC3 explains respectively 17.0% and 4.7% of the variability. In (D), the scree-plot represents the variability of all

the principal components of the analysis. The molecules are coloured by the system there are supposed to interact with, and their size is a function of their binding free energy. 117

Figure 39: PCA of the Guest chemical space using a reduced set of molecular descriptors generated with CORINA. In (A), the space formed by the combination of PC1 and PC2 explains respectively 41.7% and 27% of the variability. In (B), the space formed by the combination of PC1 and PC3 explains respectively 41.7% and 7.1% of the variability. In (C), the space formed by the combination of PC2 and PC3 explains respectively 41.7% and 27% of the variability. In (D), the scree-plot represents the variability of all the principal components of the analysis. The molecules are coloured by the system there are supposed to interact with, and their size is a function of their binding free energy. 119

Figure 40: PCA of the Host chemical space using a reduced set of molecular descriptors generated with CORINA. In (A), the space formed by the combination of PC1 and PC2 explains respectively 40.0% and 31.1% of the variability. In (B), the space formed by the combination of PC1 and PC3 explains respectively 40.0% and 27.5% of the variability. In (C), the space formed by the combination of PC2 and PC3 explains respectively 31.1% and 27.5% of the variability. In (D), the scree-plot represents the variability of all the principal components of the analysis. The molecules are coloured by the system. 121

Figure 41: PCA of the Host-Guest chemical space using a reduced set of molecular descriptors generated with CORINA. In (A), the space formed by the combination of PC1 and PC2 explains respectively 30.8% and 17.8% of the variability. In (B), the space formed by the combination of PC1 and PC3 explains respectively 30.8% and 14.1% of the variability. In (C), the space formed by the combination of PC2 and PC3 explains respectively 17.8% and 14.1% of the variability. In (D), the scree-plot represents the variability of all the principal components of the analysis. The molecules are coloured by the system there are supposed to interact with, and their size is a function of their binding free energy. 123

Figure 42: Top-contributions of variables to Dimension 1..... 124

Figure 43: Top-contributions of variables to Dimension 2..... 124

Figure 44: Top-contributions of variables to Dimension 3..... 125

Figure 45: Presentation of the structure of the initial macrocycle[144] 129

Figure 46: Presentation of the octa-acid host system: the exo position is not modified, but height carboxylic groups have been added to the structure: 4 at the end of the pendant chains and four at the external part of the endo position, linked to the hydrophobic rim[144] 129

Figure 47: Presentation of the Gibbs cavitand used in the thesis: From the left to the right: the OA, the exoOA, the TEMOA, and TEETOA 130

Figure 48: Geometries of the cucurbit[n]urils family[145] 131

Figure 49: Calculated Electrostatic potential for several cucurbit[n]uril: CB[5] (a), CB[6] (b), CB[7] (c) and CB[8] (d)[145] 132

Figure 50: Presentation of the CB[8] structure used in the SAMPL8 challenge..... 133

Figure 51: Presentation of the Trimertrip systems used in the SAMPL3 challenge (left) and in the SAMPL7 challenge (right)..... 134

Figure 52: (a) Representation of the chemical structures of calix[4]pyrrole ; (b) schematic

representation on how the calix[4]pyrrole assemble[147] 135
Figure 53: An overview of the possible applications for Calix[4]pyrrole[148] 135
Figure 54: Protocol used to generate low-energy conformations of the apo host, the guest, and the host-guest systems. Three methods have been tested to generate initial models of the host-guest complex: SaMD, MD-Docking, and Docking. MD with explicit aqueous solvation is used to sample the conformational space. Then, for representative conformations, water is deleted, and the geometry is minimized with the GFN2B basis set in GBSA implicit water solvation. 136
Figure 55: Results of the retrospective analysis of SAMPL3 Host-Guest complexes. Free energy predictions (blue bars) and experimental values (red bars) are in excellent agreement. 137
Figure 56: Binding mode of guest molecule G06 generated with docking and xTB. A sulfonate group enters the host pocket during geometric optimization, revealing an inadequate balance of solvation terms. 138
Figure 57: Inclusion process for trimer-trip dost-guest complexes observed with SaMD. (I) Linear guest G02 (A) starts from a fully dissociated state; (B) after ~ 10ns, surface interaction is formed between host and guest; (C) eventually, the host widens the cavity, and the guest molecule slides across to form a complex; (D-E-F) the complex remains stable but explores a variety of conformations for the remaining of the simulation. (II) Cyclic guest G07 (A) forms an encounter complex very early (~1ns); (B) and remains in contact with the host for over 100ns, until the host clicks into the closed geometry; (C-D) the complex remains stable but explores a variety of conformations for the remaining of the simulation. 140
Figure 58: Comparison of experimental binding free energies with predicted values. (Top) correlation plot; the green-shaded area represents a threshold of +1/-1 kcal/mol from the experimental energy; the symbols indicate the nature of the guest and the method used for binding mode generation (triangle = docking for cyclic guest, circle = docking for linear guest, square = SaMD for linear guest, cross = SaMD for cyclic guest). (Bottom) histogram of binding free energy coloured by the method used for binding mode generation (black = docking for cyclic guest, green = docking for linear guest, blue = SaMD for Linear guest, purple = SaMD for cyclic guest). G18 and G19 guests are not shown or considered for statistical analysis because it was not possible to generate a plausible binding mode for them. 141
Figure 59: A performance of the training set including 27 different guests interacting with two different systems. (B) the test set includes eight guest molecules with free energy predicted using the training set. 143
Figure 60: Comparison of experimental binding free energies with predicted values. (Top) correlation plot; The green-shaded area represents a threshold of ±1 kcal/mol from the experimental energy; the symbols indicate the nature of the guest, and each prediction has a different colour (triangle = positively charged guest interacting with OA system, circle = negatively charged guest interacting with OA system, square = negatively charged guest interacting with the exo-OA system, cross = positively charged guest interacting with the exo-OA system). (Bottom) histogram of binding free energy with calculated (blue) and experimental values (red). The error bars reflect the RMSE of the nnet model on the training set (0.918 kcal/mol). 144

Figure 61: Molecules used for the SAMPL challenges; in blue, the SAMPL6 dataset used for the retrospective analysis (in blue); in red, the SAMPL8 dataset for which the binding free energy has to be predicted. *148*

Figure 62: Overview of the results of the retrospective analysis done on the CB[8] system with data extracted from the SAMPL6 challenge. In the (Top) correlation plot, the green-shaded area represents a threshold of +2/-2 kcal/mol from the experimental energy. In the Bottom: histogram of binding free energy. *150*

Figure 63: Overview of the outcome of the thermodynamic-based approach for the retrospective analysis: in the upper part from left to right: complex G0, G1, and G2 and in the bottom part from left to right: G3, G4, and G5. The arrows represent the transformation between the docking outcome and the minimal energy structure extracted from MD (. In purple, the results with good agreement from experimental and in the yellow the results with a bad agreement. *151*

Figure 64: Overview of the outcome of the thermodynamic-based approach for the retrospective analysis: in the upper part from left to right: complex G6, G7 and, G8 and in the bottom part from left to right: G9, G10 and, G11. The arrows represent the transformation between the docking outcome and the minimal energy structure extracted from MD. In purple, the results with good agreement from experimental, and in yellow, the results with a bad agreement. *152*

Figure 65: Overview of the outcome of the thermodynamic-based approach for the retrospective analysis: complex G12. The arrows represent the transformation between the docking outcome and the minimal energy structure extracted from MD. *153*

Figure 66: Overview of the outcome of the thermodynamic based approach for the SAMPL8 CB[8] challenge: in the upper part from left to right: complex G1, G2, and G3 and in the bottom part from left to right: G4, G5, and G6. The purple arrows represent the transformation between the docking outcome and the extracted structure from MD. *154*

Figure 67: Overview of the outcome of the thermodynamic based approach for the SAMPL8 CB[8] challenge: complex G7. ..... *155*

Figure 68: Comparison of experimental binding free energies with predicted values. (Top) histogram of binding free energy coloured by the origin of the data: in red the experimental data and in blue the calculated values. (Bottom) correlation plot; the green-shaded area represents a threshold of +2/-2 kcal/mol from the experimental energy. The statistical analysis is shown in the blue box: with MAE = 3.83, RMSE = 4.67, and R^2 *= 0.03.* *156*

Figure 69: Comparison of experimental binding free energies with predicted values. (Top) histogram of binding free energy coloured by the origin of the data: in red the experimental data and in orange the Boltzmann average of the calculated values. (Bottom) correlation plot: the green-shaded area represents a threshold of +2/-2 kcal/mol from the experimental energy. The statistical analysis is shown in the orange box: with MAE = 4.08, RMSE = 4.68, and R^2 *=* 10^{-5}*.* *158*

Figure 70: Performances of the NNET using a combination of hidden units and Weight decay to find the best performances for the model. *159*

Figure 71: Most useful variable of the chosen NNET for the prediction of the training set for the nnet function from the caret package *160*

Figure 72:(Top) training-set prediction of the NNET; (Bottom) Testset prediction of the NNET. The points are coloured by the binding free energy, while the

size is a function of the error of prediction (the smaller the points are and the smaller is the error). 161

Figure 73: Performances of the RF using a combination of a number of trees (ntree) and a number of tested variables (mtry) to find the best performances for the model. 162

Figure 74: Most useful variables of the chosen RF for the prediction of the training set for the rf function from the caret package 163

Figure 75: (Top) training-set prediction of the RF. (Bottom) Test-set prediction of the RF. The points are coloured by the binding free energy, while the size is a function of the error of prediction (the smaller the points are and the smaller is the error). 164

Figure 76: Performances of the SVM using a combination of internal parameters (C, sigma, and the cost) to find the best performances for the model. 165

Figure 77: (Top) training-set prediction of the polynomial SVM. (Bottom) Test-set prediction of the SVM. The points are coloured by the binding free energy, while the size is a function of the error of prediction (the smaller the points are and the smaller is the error). 166

Figure 78: Performances of the Knn function using different neighbours number to find the best performances for the model. 167

Figure 79: (Top) Training-set prediction of the Knn. (Bottom) Test-set prediction of the Knn. The points are coloured by the binding free energy, while the size is a function of the error of prediction (the smaller the points are and the smaller is the error). 168

Figure 80: Visualization of the chemical space of the predicted set compared to the test set and the training set. If the new host-guest system descriptors overlap with the chemical space of the model, we can consider they are sampling a relatively similar space and fulfil the requirements for the prediction using our machine learning method. 170

Figure 81: Comparison of experimental binding free energies and predicted values. Histograms of binding free energy coloured by the machine learning model used: in red the experimental data, and in blue the nnet prediction, in dark red the SVM prediction, in black the RF prediction, and in purple the Knn prediction. All of the histograms are associated with a correlation plot: the green-shaded area represents a threshold of +2/-2 kcal/mol from the experimental energy. The statistical analysis of the prediction compared to the experimental value is shown in the respective coloured box. 173

Figure 82: PCA of the Host-Guest chemical space using a reduced set of molecular descriptors generated with CORINA. In (A), the space formed by the combination of PC1 and PC2 explains respectively 30.8% and 17.8% of the variability. In (B), the space formed by the combination of PC1 and PC3 explains respectively 30.8% and 14.1% of the variability. In (C), the space formed by the combination of PC2 and PC3 explains respectively 17.8% and 14.1% of the variability. In (D), the scree-plot represents the variability of all the principal components of the analysis. The molecules are coloured by the system there are supposed to interact with and, their size is a function of their binding free energy. 175

Figure 83: Comparison of experimental binding free energies with predicted values. (Top) histogram of binding free energy coloured by the origin of the data: in red the experimental data and in green the predicted binding free energy using nnet machine learning on the CB[8] dataset. (Bottom) correlation

plot: the green-shaded area represents a threshold of +2/-2 kcal/mol from the experimental energy. The statistical analysis is shown in the green box: with MAE = 3.11, RMSE = 3.54, and R^2 = 0.15. .. 176

Figure 84: Comparison of experimental binding free energies with predicted values. (Upper) Histogram of binding free energy coloured by the system: in red the experimental data, in purple the prediction on the TEMOA system, in blue the prediction of the TEETOA system. (Bottom) All of the histograms are associated with a correlation plot: the green-shaded area represents a threshold of +2/-2 kcal/mol from the experimental energy. The statistical analysis of the prediction compared to the experimental value is shown in the respective colored box: in purple the prediction concerning the TEOMA system (MAE =1.27, RMSE =1.64, R^2 = 0.001), in blue the prediction concerning the TEETOA system (MAE =2.74, RMSE =2.83, R^2 = 0.003), and in black the global prediction (MAE =1.92, RMSE =2.25, R^2 = 0.54)..... 178

Figure 85: Presentation of the initial 3D structure that was used for solvent exchange analysis: (Left) the molecule in complex with two molecules of acetonitrile; (Right) the molecule in complex with two molecules of chloroform. 180

Figure 86: An overview of the molecular dynamics of the system in the explicit chloroform ($CHCl_3$) model. The graphics are separated into three parts: (i) in the left, we have the equilibrated system representing the first frame of the dynamic, (ii) in the middle, after a few ns of simulations, the first exchange between two molecules of solvents inside the cavity of the host appears, and (iii) after 150ns another chloroform molecule enter in the cavity. 181

Figure 87: An overview of the molecular dynamics of the system in explicit chloroform ($CHCl_3$) / Acetonitrile (ACN) model. The graphics are separated into five parts: from left to right: (i) we have the equilibrated system representing the first frame of the dynamic, (ii) after 2ns of simulations, the second molecule of chloroform enter the cavity, (iii) multiple exchanges remain between the chloroforms molecule, (iv) after ~70ns of simulation, a molecule of acetonitrile is replacing one molecule of chloroform. At this moment, the exchange stopped until the second molecule of chloroform gets replaced after ~130ns (v)..... 181

Figure 88: An overview of the semi-empirical molecular dynamics of the system in implicit chloroform starting from a configuration with two chloroforms explicitly defined inside the cavity. 182

Figure 89: Zoom of the N-N distances of the capsules 183

Figure 90: An overview of the semi-empirical molecular dynamics of the system in implicit chloroform starting from a configuration with an empty cavity and a mixture of chloroform and acetonitrile molecules explicitly defined around the cavity. 184

Figure 91: Overview of the application of porphyrinoids, whose physical-chemical output depend on the applied stimuli. The porphyrin core is also defined and described by the metal-centre and the main functionalisation sites (meso and β-positions)[149] 186

Figure 92: Illustration of the general principle of allosteric control adapted for host-guest systems. Left: allosteric inhibition. Right: allosteric activation 187

Figure 93: One of the first allosteric systems described in 1979[150] 188

Figure 94: Allosteric receptor presenting Zn(II)porphyrin scaffold[151] 189

Figure 95: Zn(II)-porphyrin receptor presenting allosteric control designed at LSAMM[152] 189
Figure 96: (A) Chemical structure (top) and schematic representation (bottom) of the tweezer in the unbound state. (B) Chemical structure (top) and schematic representation (bottom) of the DABCO-coordinated receptor, presenting an accessible binding cavity between the two acridinium units. 191
Figure 97: The multi-responsive properties of the 9-phenyl-acridinium moiety 192
Figure 98: Synthesised receptors using a C18 carbon spacer: (A) a Zn(II)-porphyrin-acridinium receptor with two ditopic ligands and their respective association constant (DABCO, Ka_1) and (Bipyridine, Ka_2); (B) a Zn(II)-porphyrin-acridane receptor with a ditopic ligand and its respective association constant (Bipyridine, Ka_3) 193
Figure 99: Clustering analysis of the porphyrinoids receptors; (A) scatterplot representing the geometries described by RMSD and radius-of-gyration ; kmeans clustering of (B) Apo-acridinium (orange) receptor ; (C) Apo-acridane (light-blue) receptor ; (D) Bipyridine-acridinium (blue) receptor ; (E) Bipyridine-acridane (green) receptor ; and (F) DABCO-acridinium (pink) receptor..... 196
Figure 100: Knn clustering of the apo-acridinium receptor, coloured by clusters: (In green) the open-conformation, (In blue) the semi-open conformation, and (In red) the closed-conformation 198
Figure 101: Knn clustering of the apo-acridane receptor, coloured by clusters: (In green) the open-conformation, (In blue) the semi-open conformation, in yellow the semi-closed conformation and (In red) the closed-conformation 199
Figure 102: Knn clustering of the Bipyridine-acridinium receptor, coloured by clusters: from the most populated cluster to the less populated cluster: (In blue) the cluster 1, (In green) the cluster 2, (In red) the cluster 3 and (In purple) the cluster 4 200
Figure 103: Knn clustering of the Bipyridine-acridane receptor, coloured by clusters: from the most populated cluster to the less populated cluster: (In blue) the cluster 1, (In green) the cluster 2, (In red) the cluster 3 and (In purple) the cluster 4 201
Figure 104: Knn clustering of the DABCO-acridinium receptor, coloured by clusters: from the most populated cluster to the less populated cluster: (In blue) the cluster 1, (In green) the cluster 2, (In red) the cluster 3 and (In purple) the cluster 4 202
Figure 105: Binding free energy prediction of Zn(II)-porphyrin receptor considering three different receptors: (Left) the Zn(II)-porphyrin-acridane with bipirydine as a ligand ; (Middle) The Zn(II)-porphyrin-acridinium with bipyridine as a ligand ; (Right) The Zn(II)-porphyrin-acridinium with DABCO as a ligand 203
Figure 106: (A) Chemical structure of the tweezer; (B) Receptor formed by the DABCO coordinating the two Zn(II), forming an available binding cavity between the two acridiniums allowing the binding of polyaromatic guest interacting with π-π interactions. 205
Figure 107: Retrosynthesis study of the porphyrin-acridinium conjugate: $\mathbf{1} \cdot (\boldsymbol{PF_6})_2$ 206
Figure 108: Synthesis of the 2-(4-bromophenyl)-4,4,5,5-tetramethyl-1,3-dioxolane (2)..... 207
Figure 109: Synthesis of the 10-allyl-9-(4-formylphenyl)acridin-10-ium $(\mathbf{3} \cdot \boldsymbol{PF_6})$ 207
Figure 110: Chemical synthesis of the porphyrin-acridinium conjugate: $\mathbf{4} \cdot \boldsymbol{PF_6}$ 208

Figure 111: Synthesis of the porphyrin-acridinium conjugate: $\mathbf{1} \cdot (\boldsymbol{PF_6})_2$ *208*
Figure 112: Chemical synthesis of the porphyrin-acridinium conjugate in complex with DABCO: $[\mathbf{1} \cdot (\boldsymbol{PF_6})_2 \bullet \boldsymbol{DABCO}]$ *209*
Figure 113: Chemical synthesis of the porphyrin-acridinium conjugate in complex with DABCO and Perylene: $[\mathbf{1} \cdot (\boldsymbol{PF_6})_2 \bullet \boldsymbol{DABCO}] \supset \boldsymbol{Perylene}$ *209*
Figure 114: Stacked ^{1}H NMR spectra of a) 4·($\boldsymbol{PF_6}$)$_2$ ($CDCl_3$, 298K, 500MHz), b) 1·($\boldsymbol{PF6}$)$_2$ (CD_2Cl_2, 298 K, 500MHz) *210*
Figure 115: UV/Visible spectrum (CH_2Cl_2, 298 K, l = 1cm, c = 4.37.10^{-6} M) of the tweezer $(\mathbf{1} \cdot (\boldsymbol{PF_6})_2)$ *211*
Figure 116: Stacked ^{1}H NMR spectra ($CDCl_3$, 298 K, 500MHz) of a) $\mathbf{1} \cdot (\boldsymbol{PF_6})_2$ *b)* $\mathbf{1} \cdot (\boldsymbol{PF_6})_2 \bullet \boldsymbol{DABCO}$ *212*
Figure 117: Stacked ^{1}H NMR spectra ($CDCl_3$, 298 K, 500MHz) of a) $\mathbf{1} \cdot (\boldsymbol{PF_6})_2$ *b)* $\mathbf{1} \cdot (\boldsymbol{PF_6})_2 \bullet \boldsymbol{DABCO}$ *and c)* $[\mathbf{1} \cdot (\boldsymbol{PF_6})_2 \bullet \boldsymbol{DABCO}] \supset \boldsymbol{Perylene}$ *213*
Figure 118: PCA of Zn(II)-porphyrin receptors chemical space described by a set of molecular descriptors derived from the simulations. The space formed by the combination of PC1, PC2, PC3, and PC4 explaining respectively, ~80% of the variability. Each point represents a geometry, and the point is coloured by the simulations: MD1 (blue), MD2 (red), MD3 (green). (upper part): the combination of each P.C.s (bottom part): an overview of some specific geometry. *216*
Figure 119: An overview of the results of the first pharmacophoric search *217*
Figure 120: Representation of the steroid scaffold *218*
Figure 121: Chemical synthesis of the 2-(4-bromophenyl)-4,4,5,5-tetramethyl-1,3-dioxolane *222*
Figure 122: Chemical synthesis of the 10-(but-en-1-yl)acridin-9(10H)-one *223*
Figure 123: Chemical synthesis of the synthesis of the 10-allyl-9-(4-formylphenyl)acridin-10-ium $(\mathbf{3} \cdot \boldsymbol{PF_6})$ *224*
Figure 124: Chemical synthesis of the porphyrin-acridinium conjugate ($\mathbf{4} \cdot \boldsymbol{PF_6}$*)* *225*
Figure 125: Chemical synthesis of the porphyrin-acridinium tweezer ($\mathbf{1} \cdot (\boldsymbol{PF_6})_2$*)* *227*
Figure 126: Synthesis of $\mathbf{1} \cdot (\boldsymbol{PF_6})_2 \bullet \boldsymbol{DABCO}$ ... *228*
Figure 127: Synthesis of $[\mathbf{1} \cdot (\boldsymbol{PF_6})_2 \bullet \boldsymbol{DABCO}] \supset \boldsymbol{Perylene}$ *229*

BIBLIOGRAPHY

1. Tramontano, A. The role of molecular modelling in biomedical research. *FEBS Letters* **580**, 2928–2934 (2006).
2. Tanaka, F. Theoretical Study of Molecular Association and Thermoreversible Gelation in Polymers. *Polym J* **34**, 479–509 (2002).
3. Young, D. C. *Computational Chemistry*. (John Wiley & Sons, Inc., 2001). doi:10.1002/0471220655.
4. *Reviews in Computational Chemistry*. (John Wiley & Sons, Inc., 1990). doi:10.1002/9780470125786.
5. Shakerzadeh, E. Theoretical investigations of interactions between boron nitride nanotubes and drugs. in *Boron Nitride Nanotubes in Nanomedicine* 59–77 (Elsevier, 2016). doi:10.1016/B978-0-323-38945-7.00004-3.
6. Mobley, D. L. & Gilson, M. K. Predicting Binding Free Energies: Frontiers and Benchmarks. *Annu. Rev. Biophys.* **46**, 531–558 (2017).
7. Cramer, C. J. *Essentials of computational chemistry: theories and models*. (Wiley, 2004).
8. Lehn, J. M. Cryptates: inclusion complexes of macropolycyclic receptor molecules. *Pure and Applied Chemistry* **50**, 871–892 (1978).
9. Lehn, J.-M. *Supramolecular chemistry concepts and perspectives*. (VCH, 2006).
10. *Supramolecular catalysis*. (Wiley-VCH, 2008).
11. Chrisstoffels, L., de Jong F, F. & Reinhoudt, D. Facilitated transport of

salts by neutral anion carriers. *Chemistry* **6**, 1376–1385 (2000).

12. Lehn, J.-M. Supramolecular Chemistry—Scope and Perspectives Molecules, Supermolecules, and Molecular Devices(Nobel Lecture). *Angew. Chem. Int. Ed. Engl.* **27**, 89–112 (1988).
13. Chatterji, D. *Basics of molecular recognition*. (CRC Press, Taylor & Francis Group, 2016).
14. Ariga, K., Ito, H., Hill, J. P. & Tsukube, H. Molecular recognition: from solution science to nano/materials technology. *Chem. Soc. Rev.* **41**, 5800 (2012).
15. Aguilar, Z. P. *Nanomaterials for medical applications*. (Elsevier, 2013).
16. Therrien, B. Drug Delivery by Water-Soluble Organometallic Cages. in *Chemistry of Nanocontainers* (eds. Albrecht, M. & Hahn, E.) vol. 319 35–55 (Springer Berlin Heidelberg, 2011).
17. Wenz, G. An Overview of Host-Guest Chemistry and its Application to Nonsteroidal Anti-Inflammatory Drugs: *Clinical Drug Investigation* **19**, 21–25 (2000).
18. Yi, J. W. *et al.* Delivery of Floxuridine Derivatives to Cancer Cells by Water-Soluble Organometallic Cages. *Bioconjugate Chem.* **23**, 461–471 (2012).
19. Yu, G. & Chen, X. Host-Guest Chemistry in Supramolecular Theranostics. *Theranostics* **9**, 3041–3074 (2019).
20. Zhang, J. & Ma, P. X. Host–guest interactions mediated nano-assemblies using cyclodextrin-containing hydrophilic polymers and their biomedical applications. *Nano Today* **5**, 337–350 (2010).
21. Cram, D. J. The design of molecular hosts, guests, and their complexes. *Journal of Inclusion Phenomena* **6**, 397–413 (1988).

22. Sauvage, J.-P. From Chemical Topology to Molecular Machines (Nobel Lecture). *Angew. Chem. Int. Ed.* **56**, 11080–11093 (2017).
23. Stoddart, J. F. Mechanically Interlocked Molecules (MIMs)-Molecular Shuttles, Switches, and Machines (Nobel Lecture). *Angew. Chem. Int. Ed.* **56**, 11094–11125 (2017).
24. Feringa, B. L. The Art of Building Small: From Molecular Switches to Motors (Nobel Lecture). *Angew. Chem. Int. Ed.* **56**, 11060–11078 (2017).
25. Fischer, E. & Fourneau, E. Ueber einige Derivate des Glykocolls. *Ber. Dtsch. Chem. Ges.* **34**, 2868–2877 (1901).
26. Müller-Dethlefs, K. & Hobza, P. Noncovalent Interactions: A Challenge for Experiment and Theory. *Chem. Rev.* **100**, 143–168 (2000).
27. Lehn, J.-M. Supramolecular Chemistry: Receptors, Catalysts, and Carriers. *Science* **227**, 849–856 (1985).
28. Stone, A. J. *The theory of intermolecular forces*. (Oxford University Press, 2013).
29. Steed, J. W. & Atwood, J. L. *Supramolecular chemistry*. (Wiley, 2009).
30. *Molecular cell biology*. (W.H. Freeman, 2008).
31. Ariga, K., Ariga, K. & Kunitake, T. *Supramolecular chemistry: fundamentals and applications advanced textbook*. (Springer, 2006).
32. Wang, L., Berne, B. J. & Friesner, R. A. On achieving high accuracy and reliability in the calculation of relative protein-ligand binding affinities. *Proceedings of the National Academy of Sciences* **109**, 1937–1942 (2012).
33. Pohorille, A. & Chipot, C. *Free energy calculations: theory and applications in chemistry and biology*. (Springer, 2007).

34. Geballe, M. T., Skillman, A. G., Nicholls, A., Guthrie, J. P. & Taylor, P. J. The SAMPL2 blind prediction challenge: introduction and overview. *J Comput Aided Mol Des* **24**, 259–279 (2010).
35. Guthrie, J. P. A Blind Challenge for Computational Solvation Free Energies: Introduction and Overview. *J. Phys. Chem. B* **113**, 4501–4507 (2009).
36. Gaieb, Z. *et al.* D3R Grand Challenge 2: blind prediction of protein–ligand poses, affinity rankings, and relative binding free energies. *J Comput Aided Mol Des* **32**, 1–20 (2018).
37. SAMPL challenges. *https://www.samplchallenges.org/*.
38. Muddana, H. S. *et al.* Blind prediction of host–guest binding affinities: a new SAMPL3 challenge. *J Comput Aided Mol Des* **26**, 475–487 (2012).
39. Skillman, A. G. SAMPL3: blinded prediction of host–guest binding affinities, hydration free energies, and trypsin inhibitors. *J Comput Aided Mol Des* **26**, 473–474 (2012).
40. Muddana, H. S., Fenley, A. T., Mobley, D. L. & Gilson, M. K. The SAMPL4 host–guest blind prediction challenge: an overview. *J Comput Aided Mol Des* **28**, 305–317 (2014).
41. Yin, J. Overview of the SAMPL5 host–guest challenge: Are we doing better? *J Comput Aided Mol Des* 19.
42. Rizzi, A. *et al.* Overview of the SAMPL6 host–guest binding affinity prediction challenge. *J Comput Aided Mol Des* **32**, 937–963 (2018).
43. Genheden, S. & Ryde, U. The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opinion on Drug Discovery* **10**, 449–461 (2015).
44. Kästner, J. Umbrella sampling: Umbrella sampling. *WIREs Comput Mol Sci* **1**, 932–942 (2011).
45. Zwanzig, R. W. High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. *The*

Journal of Chemical Physics **22**, 1420–1426 (1954).

46. Lemkul, J. GROMACS Tutorial - Umbrella Sampling. *http://www.mdtutorials.com/gmx/umbrella/index.html* http://www.mdtutorials.com/gmx/umbrella/index.html.
47. You, W., Tang, Z. & Chang, C. A. Potential Mean Force from Umbrella Sampling Simulations: What Can We Learn and What Is Missed? *J. Chem. Theory Comput.* **15**, 2433–2443 (2019).
48. *Quantum mechanics in drug discovery*. (Humana Press, 2020).
49. Wang, J., Wang, W., Kollman, P. A. & Case, D. A. Automatic atom type and bond type perception in molecular mechanical calculations. *Journal of Molecular Graphics and Modelling* **25**, 247–260 (2006).
50. Vanommeslaeghe, K. *et al.* CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J. Comput. Chem.* NA-NA (2009) doi:10.1002/jcc.21367.
51. Robertson, M. J., Tirado-Rives, J. & Jorgensen, W. L. Improved Peptide and Protein Torsional Energetics with the OPLS-AA Force Field. *J. Chem. Theory Comput.* **11**, 3499–3509 (2015).
52. Ponder, J. W. *et al.* Current Status of the AMOEBA Polarizable Force Field. *J. Phys. Chem. B* **114**, 2549–2564 (2010).
53. Gasteiger, J. & Marsili, M. A new model for calculating atomic charges in molecules. *Tetrahedron Letters* **19**, 3181–3184 (1978).
54. Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A. & Case, D. A. Development and testing of a general amber force field. *J. Comput. Chem.* **25**, 1157–1174 (2004).
55. D.A. Case, K. Belfon, I.Y. Ben-Shalom, S.R. Brozell, D.S. Cerutti,

T.E. Cheatham, III, V.W.D. Cruzeiro,T.A. Darden, R.E. Duke, G. Giambasu, M.K. Gilson, H. Gohlke, A.W. Goetz,R Harris, S. Izadi, S.A. Iz-mailov, K. Kasavajhala, A. Kovalenko, R. Krasny, T. Kurtzman, T.S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu,T. Luchko, R. Luo, V. Man, K.M. Merz, Y. Miao, O. Mikhailovskii, G. Monard, H. Nguyen, A. Onufriev, F.Pan, S. Pantano, R. Qi, D.R. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, J. Shen, C.L. Simmerling, N.R.Skrynnikov, J. Smith, J. Swails, R.C. Walker, J. Wang, L. Wilson, R.M. Wolf, X. Wu, Y. Xiong, Y. Xue,D.M. York and P.A. Kollman (2020), AMBER 2020, University of California, San Francisco.

56. Cornell, W. D., Cieplak, P., Bayly, C. I. & Kollman, P. A. Application of RESP charges to calculate conformational energies, hydrogen bond energies, and free energies of solvation. *J. Am. Chem. Soc.* **115**, 9620–9631 (1993).

57. 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.* **23**, 1623–1641 (2002).

58. Vanommeslaeghe, K. *et al.* CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J. Comput. Chem.* NA-NA (2009) doi:10.1002/jcc.21367.

59. Kuntz, I. D., Blaney, J. M., Oatley, S. J., Langridge, R. & Ferrin, T. E. A geometric approach to macromolecule-ligand interactions. *Journal of Molecular Biology* **161**, 269–288 (1982).

60. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water.

The Journal of Chemical Physics **79**, 926–935 (1983).

61. Mark, P. & Nilsson, L. Structure and Dynamics of the TIP3P, SPC, and SPC/E Water Models at 298 K. *J. Phys. Chem. A* **105**, 9954–9960 (2001).
62. Klamt, A. Conductor-like Screening Model for Real Solvents: A New Approach to the Quantitative Calculation of Solvation Phenomena. *J. Phys. Chem.* **99**, 2224–2235 (1995).
63. Marenich, A. V., Cramer, C. J. & Truhlar, D. G. Universal Solvation Model Based on Solute Electron Density and on a Continuum Model of the Solvent Defined by the Bulk Dielectric Constant and Atomic Surface Tensions. *J. Phys. Chem. B* **113**, 6378–6396 (2009).
64. Zheng, Z., Wang, T., Li, P. & Merz, K. M. KECSA-Movable Type Implicit Solvation Model (KMTISM). *J. Chem. Theory Comput.* **11**, 667–682 (2015).
65. Bannwarth, C., Ehlert, S. & Grimme, S. GFN2-xTB—An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions. *J. Chem. Theory Comput.* **15**, 1652–1671 (2019).
66. *Host-guest-systems based on nanoporous crystals*. (Wiley-VCH, 2003).
67. Brooijmans, N. & Kuntz, I. D. Molecular Recognition and Docking Algorithms. *Annu. Rev. Biophys. Biomol. Struct.* **32**, 335–373 (2003).
68. Poole, D. L., Mackworth, A. K. & Goebel, R. *Computational intelligence: a logical approach*. (Oxford University Press, 1998).
69. Turing, A. M. Computating machinery and intelligence. *Mind* **LIX**, 433–460 (1950).
70. Legg, S. & Hutter, M. A Collection of Definitions of Intelligence. *arXiv:0706.3639 [cs]* (2007).

71. Qu'est-ce que l'intelligence artificielle ? *https://www.netapp.com/fr/artificial-intelligence/what-is-artificial-intelligence/*.
72. Nilsson, N. J. Introduction to Machine Learning.
73. Alpaydin, E. *Introduction to machine learning*. (MIT Press, 2010).
74. Marblestone, A. H., Wayne, G. & Kording, K. P. Toward an Integration of Deep Learning and Neuroscience. *Front Comput Neurosci* **10**, 94 (2016).
75. What's the Difference Between Artificial Intelligence, Machine Learning and Deep Learning? *https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/*.
76. King, E., Aitchison, E., Li, H. & Luo, R. Recent Developments in Free Energy Calculations for Drug Discovery. *Front. Mol. Biosci.* **8**, 712085 (2021).
77. Aprahamian, I. The Future of Molecular Machines. *ACS Cent. Sci.* **6**, 347–358 (2020).
78. Born, M. & Oppenheimer, R. Zur Quantentheorie der Molekeln. *Ann. Phys.* **389**, 457–484 (1927).
79. Hohenberg, P. & Kohn, W. Inhomogeneous Electron Gas. *Phys. Rev.* **136**, B864–B871 (1964).
80. Kohn, W. & Sham, L. J. Self-Consistent Equations Including Exchange and Correlation Effects. *Phys. Rev.* **140**, A1133–A1138 (1965).
81. Becke, A. D. Density-functional thermochemistry. III. The role of exact exchange. *The Journal of Chemical Physics* **98**, 5648–5652 (1993).
82. Thiel, W. Semiempirical quantum–chemical methods. *WIREs Comput Mol Sci* **4**, 145–157 (2014).
83. Dewar, M. J. S., Zoebisch, E. G., Healy, E. F. & Stewart, J. J. P. Development and use of quantum

mechanical molecular models. 76. AM1: a new general purpose quantum mechanical molecular model. *J. Am. Chem. Soc.* **107**, 3902–3909 (1985).

84. Stewart, J. J. P. Optimization of parameters for semiempirical methods V: Modification of NDDO approximations and application to 70 elements. *J Mol Model* **13**, 1173–1213 (2007).
85. Beveridge, D. L. Approximate Molecular Orbital Theory of Nuclear and Electron Magnetic Resonance Parameters. in *Semiempirical Methods of Electronic Structure Calculation* (ed. Segal, G. A.) 163–214 (Springer US, 1977). doi:10.1007/978-1-4684-2559-8_5.
86. Jensen, F. *Introduction to computational chemistry*. (Wiley, 1999).
87. Elstner, M. *et al.* Self-consistent-charge density-functional tight-binding method for simulations of complex materials properties. *Phys. Rev. B* **58**, 7260–7268 (1998).
88. Niehaus, T. A. *et al.* Tight-binding approach to time-dependent density-functional response theory. *Phys. Rev. B* **63**, 085108 (2001).
89. Yang, Y., Yu, H., York, D., Cui, Q. & Elstner, M. Extension of the Self-Consistent-Charge Density-Functional Tight-Binding Method: Third-Order Expansion of the Density Functional Theory Total Energy and Introduction of a Modified Effective Coulomb Interaction. *J. Phys. Chem. A* **111**, 10861–10873 (2007).
90. Cui, Q., Elstner, M., Kaxiras, E., Frauenheim, T. & Karplus, M. A QM/MM Implementation of the Self-Consistent Charge Density Functional Tight Binding (SCC-DFTB) Method. *J. Phys. Chem. B* **105**, 569–585 (2001).
91. Frauenheim, T. *et al.* Atomistic simulations of complex materials: ground-state and excited-state

properties. *J. Phys.: Condens. Matter* **14**, 3015–3047 (2002).

92. Pracht, P., Caldeweyher, E., Ehlert, S. & Grimme, S. A Robust Non-Self-Consistent Tight-Binding Quan- tum Chemistry Method for large Molecules. 19.
93. Gaus, M., Cui, Q. & Elstner, M. DFTB3: Extension of the Self-Consistent-Charge Density-Functional Tight-Binding Method (SCC-DFTB). *J. Chem. Theory Comput.* **7**, 931–948 (2011).
94. Bannwarth, C. *et al.* Extended tight-binding quantum chemistry methods. 49.
95. Leach, A. R. *Molecular modelling: principles and applications*. (Prentice Hall, 2001).
96. Haug, E. J., Arora, J. S. & Matsui, K. A steepest-descent method for optimization of mechanical systems. *J Optim Theory Appl* **19**, 401–424 (1976).
97. Shewchuk, J. R. (1994). An introduction to the conjugate gradient method without the agonizing pain.
98. Levesque, D. & Verlet, L. Molecular dynamics and time reversibility. *J Stat Phys* **72**, 519–537 (1993).
99. Van Gunsteren, W. F. & Berendsen, H. J. C. A Leap-frog Algorithm for Stochastic Dynamics. *Molecular Simulation* **1**, 173–185 (1988).
100. Hünenberger, P. H. Thermostat Algorithms for Molecular Dynamics Simulations. in *Advanced Computer Simulation* (eds. Dr. Holm, C. & Prof. Dr. Kremer, K.) vol. 173 105–149 (Springer Berlin Heidelberg, 2005).
101. Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A. & Haak, J. R. Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics* **81**, 3684–3690 (1984).
102. Evans, D. J. & Holian, B. L. The Nose–Hoover thermostat. *The Journal*

of Chemical Physics **83**, 4069–4074 (1985).

103. Loncharich, R. J., Brooks, B. R. & Pastor, R. W. Langevin dynamics of peptides: The frictional dependence of isomerization rates ofN-acetylalanyl-N?-methylamide. *Biopolymers* **32**, 523–535 (1992).
104. Le Roux, S. & Petkov, V. ISAACS, Model Box Periodic Boundary Conditions - P.B.C.
105. *Compendium of chemical terminology: IUPAC recommendations*. (Blackwell Science, 1997).
106. Water Models. *https://www.idc-online.com/technical_references/pdfs/chemical_engineering/Water_models.pdf*.
107. Mark, P. & Nilsson, L. Structure and Dynamics of the TIP3P, SPC, and SPC/E Water Models at 298 K. *J. Phys. Chem. A* **105**, 9954–9960 (2001).
108. Bannwarth, C. *et al.* Extended TIGHT-BINDING quantum chemistry methods. *WIREs Comput Mol Sci* **11**, (2021).
109. *Principal Component Analysis*. (Springer-Verlag, 2002). doi:10.1007/b98835.
110. Besse, P. PCA stability and choice of dimensionality. *Statistics & Probability Letters* **13**, 405–410 (1992).
111. Dunn, K. Process Improvement Using Data. 421.
112. Alboukadel Kassambara and Fabian Mundt (2017). factoextra: Extract and Visualize the Results of Multivariate Data Analyses. R package version 1.0.5. https://CRAN.R-project.org/package=factoextra.
113. Roy, K., Kar, S. & Das, R. N. *A Primer on QSAR/QSPR Modeling: Fundamental Concepts*. (Springer International Publishing : Imprint: Springer, 2015). doi:10.1007/978-3-319-17281-1.

114. *Handbook of computational chemistry*. (Springer, 2017). doi:10.1007/978-3-319-27282-5.
115. Todeschini, R. & Consonni, V. *Molecular descriptors for chemoinformatics*. (Wiley-VCH, 2009).
116. Chemoinformatics ProgramPackage CORINA Symphony, developed and distributed by Molecular Networks GmbH, Nuremberg, Germany and Altamira LLC, Columbus, OH, USA (www.mn-am.com).
117. Altman, N. S. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *The American Statistician* **46**, 175–185 (1992).
118. Cortes, C. & Vapnik, V. Support-vector networks. *Mach Learn* **20**, 273–297 (1995).
119. Kuhn, M. & Johnson, K. *Applied Predictive Modeling*. (Springer New York, 2013). doi:10.1007/978-1-4614-6849-3.
120. Tin Kam Ho. Random decision forests. in *Proceedings of 3rd International Conference on Document Analysis and Recognition* vol. 1 278–282 (IEEE Comput. Soc. Press, 1995).
121. Breiman, L. Bagging predictors. *Mach Learn* **24**, 123–140 (1996).
122. Breiman, L. Random Forests. *Machine Learning* **45**, 5–32 (2001).
123. Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning*. (MIT Press, 2016).
124. Bishop, C. M. *Neural networks for pattern recognition*. (Clarendon Press ; Oxford University Press, 1995).
125. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning*. (Springer New York, 2009). doi:10.1007/978-0-387-84858-7.
126. Hornik, K., Stinchcombe, M. & White, H. Multilayer feedforward networks are universal approximators. *Neural Networks* **2**, 359–366 (1989).

127. Hornik, K. Approximation capabilities of multilayer feedforward networks. *Neural Networks* **4**, 251–257 (1991).
128. Trott, O. & Olson, A. J. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* NA-NA (2009) doi:10.1002/jcc.21334.
129. Empereur-Mot, C. Développement d'outils statistiques d'évaluation de méthodes de criblage virtuel : courbes de prédictivité & Screening Explorer. 211.
130. Valderrey, V., Aragay, G. & Ballester, P. Porphyrin tweezer receptors: Binding studies, conformational properties and applications. *Coordination Chemistry Reviews* **258–259**, 137–156 (2014).
131. Jacquot de Rouville, H.-P., Gourlaouen, C. & Heitz, V. Self-complementary and narcissistic self-sorting of bis-acridinium tweezers. *Dalton Trans.* **48**, 8725–8730 (2019).
132. Jacquot de Rouville, H., Hu, J. & Heitz, V. N-Substituted Acridinium as a Multi-Responsive Recognition Unit in Supramolecular Chemistry. *ChemPlusChem* **86**, 110–129 (2021).
133. *open-babel software*.
134. *UCSF Chimera & UCSF ChimeraX*.
135. *Amber software*.
136. *Ambertools software*.
137. *vmd software*.
138. *xtb sofware*.
139. *R software*.
140. Grimme, group. *xtb reference manual*.
141. *Amber reference manual*.
142. Bayly, C. I., Cieplak, P., Cornell, W. & Kollman, P. A. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *J. Phys. Chem.* **97**, 10269–10280 (1993).

143. AMBER case study: Setting up an Advanced System (including basic charge derivation).
144. Bender, A., Mussa, H. Y., Glen, R. C. & Reiling, S. Similarity searching of chemical databases using atom environment descriptors (MOLPRINT 2D): evaluation of performance. *J Chem Inf Comput Sci* **44**, 1708–1718 (2004).
145. Chemoinformatics ProgramPackage CORINA Symphony, developed and distributed by Molecular Networks GmbH, Nuremberg, Germany and Altamira LLC, Columbus, OH, USA (www.mn-am.com).
146. Gibb, B. C. From steroids to aqueous supramolecular chemistry: an autobiographical career review. *Beilstein J. Org. Chem.* **12**, 684–701 (2016).
147. Barrow, S. J., Kasera, S., Rowland, M. J., del Barrio, J. & Scherman, O. A. Cucurbituril-Based Molecular Recognition. *Chem. Rev.* **115**, 12320–12406 (2015).
148. Ma, D., Zavalij, P. Y. & Isaacs, L. Acyclic Cucurbit[n]uril Congeners Are High Affinity Hosts. *J. Org. Chem.* **75**, 4786–4795 (2010).
149. Peng, S. *et al.* Strapped calix[4]pyrroles: from syntheses to applications. *Chem. Soc. Rev.* **49**, 865–907 (2020).
150. Rather, I. A., Wagay, S. A., Hasnain, M. S. & Ali, R. New dimensions in calix[4]pyrrole: the land of opportunity in supramolecular chemistry. *RSC Adv.* **9**, 38309–38344 (2019).
151. Martynov, A. G., Safonova, E. A., Tsivadze, A. Yu. & Gorbunova, Y. G. Functional molecular switches involving tetrapyrrolic macrocycles. *Coordination Chemistry Reviews* **387**, 325–347 (2019).
152. Rebek, J., Trend, J. E., Wattley, R. V. & Chakravorti, S. Allosteric effects in organic chemistry. Site-specific

binding. *J. Am. Chem. Soc.* **101**, 4333–4337 (1979).

153. Lee, C.-H., Yoon, H. & Jang, W.-D. Biindole-Bridged Porphyrin Dimer as Allosteric Molecular Tweezers. *Chem. Eur. J.* **15**, 9972–9976 (2009).
154. Djemili, R. *et al.* Positive Allosteric Control of Guests Encapsulation by Metal Binding to Covalent Porphyrin Cages. *Chem. Eur. J.* chem.201805498 (2018) doi:10.1002/chem.201805498.
155. Koper, N. W., Jonker, S. A., Verhoeven, J. W. & van Dijk, C. Electrochemistry of the 9-phenyl-10-methyl-acridan/acridinium redox system; a high-potential NADH/NAD $^{+}$ analogue. *Recl. Trav. Chim. Pays-Bas* **104**, 296–302 (1985).
156. Ackmann, A. J. & Fréchet, J. M. J. The generation of hydroxide and methoxide ions by photo-irradiation: use of aromatization to stabilize ionic photo-products from acridine derivatives. *Chem. Commun.* 605–606 (1996) doi:10.1039/CC9960000605.
157. Gosset, A. *et al.* A chemically-responsive bis-acridinium receptor. *New J. Chem.* **42**, 4728–4734 (2018).
158. Jacquot de Rouville, H.-P., Zorn, N., Leize-Wagner, E. & Heitz, V. Entwined dimer formation from self-complementary bis-acridiniums. *Chem. Commun.* **54**, 10966–10969 (2018).
159. Hu, J. *et al.* A Bis-Acridinium Macrocycle as Multi-Responsive Receptor and Selective Phase-Transfer Agent of Perylene. *Angew. Chem. Int. Ed.* **59**, 23206–23212 (2020).
160. Sneha, P. & George Priya Doss, C. Molecular Dynamics. in *Advances in Protein Chemistry and Structural Biology* vol. 102 181–224 (Elsevier, 2016).
161. Tanaka, M., Yukimoto, K., Ohkubo, K. & Fukuzumi, S. Intermolecular vs. intramolecular photoinduced electron

transfer from nucleotides in DNA to acridinium ion derivatives in relation with DNA cleavage. *Journal of Photochemistry and Photobiology A: Chemistry* **197**, 206–212 (2008).

162. Yi, H. *et al.* Visible-Light-Induced Acetalization of Aldehydes with Alcohols. *Org. Lett.* **19**, 122–125 (2017).
163. Lindsey, J. S., Hsu, H. C. & Schreiman, I. C. Synthesis of tetraphenylporphyrins under very mild conditions. *Tetrahedron Letters* **27**, 4969–4970 (1986).
164. Schwab, P., Grubbs, R. H. & Ziller, J. W. Synthesis and Applications of $RuCl_2(CHR')(PR_3)_2$: The Influence of the Alkylidene Moiety on Metathesis Activity. *J. Am. Chem. Soc.* **118**, 100–110 (1996).
165. Leach, A. R., Gillet, V. J., Lewis, R. A. & Taylor, R. Three-Dimensional Pharmacophore Methods in Drug Discovery. *J. Med. Chem.* **53**, 539–558 (2010).
166. Koes, D. R. & Camacho, C. J. Pharmer: Efficient and Exact Pharmacophore Search. *J. Chem. Inf. Model.* **51**, 1307–1314 (2011).
167. Wishart, D. S. *et al.* DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Research* **46**, D1074–D1082 (2018).
168. Law, V. *et al.* DrugBank 4.0: shedding new light on drug metabolism. *Nucl. Acids Res.* **42**, D1091–D1097 (2014).
169. Knox, C. *et al.* DrugBank 3.0: a comprehensive resource for 'Omics' research on drugs. *Nucleic Acids Research* **39**, D1035–D1041 (2011).
170. Wishart, D. S. *et al.* DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Research* **36**, D901–D906 (2008).
171. Wishart, D. S. *et al.* DrugBank: a comprehensive resource for in silico drug discovery and exploration. 5.
172. Wishart, D. *et al.* T3DB: the toxic exposome database. *Nucleic Acids Research* **43**, D928–D934 (2015).

173. Lim, E. *et al.* T3DB: a comprehensively annotated database of common toxins and their targets. *Nucleic Acids Research* **38**, D781–D786 (2010).

174. Molecular Operating Environment (MOE), 2019.01; Chemical Computing Group ULC, 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, 2021.

175. Kamenik, A. S. *et al.* Energy penalties enhance flexible receptor docking in a model cavity. *Proc Natl Acad Sci U S A* **118**, e2106195118 (2021).

Dylan SERILLON

Computational exploration of host-guest complexes

Résumé

Ces dernières années, la chimie supramoléculaire a connu un énorme essor. Les processus supramoléculaires et, en particulier, les interactions hôte-invité sont étudiées pour la variété des applications possibles (des processus industriels au domaine médical). Actuellement, les découvertes dans le domaine de la chimie supramoléculaire hôte-invité sont entravées par la complexité de la caractérisation thermodynamique et cinétique des processus d'inclusion/libération, ce qui rend difficile la génération de prédictions utiles pour l'encapsulation moléculaire.

Dans ce contexte, ce projet de thèse s'est concentré sur le développement d'une plateforme de calcul pour la prédiction de l'énergie libre de Gibbs de complexes hôte-invité en utilisant deux approches différentes : la première basée sur la prédiction des paramètres thermodynamiques et la seconde basée sur les connaissances. L'objectif est non seulement d'améliorer les connaissances globales dans le domaine de la chimie supramoléculaire, mais également de fournir de nouvelles opportunités et applications pour les conteneurs existants de manière à aider au développement de ces derniers.

Mots-clés : Prédiction d'énergie libre de Gibbs, Méthode quantique semi-empirique, Méthode d'apprentissage automatique, plateforme automatisée

Abstract

Supramolecular chemistry has experienced enormous growth in recent years. Supramolecular processes and, in particular, host-guest interactions are studied for the variety of their potential applications (from industrial processes to medical field application). At the moment, breakthrough discoveries in molecular host-guest chemistry are hampered by the complexity of the thermodynamic and kinetic characterization of the inclusion/release processes, which make it difficult to generate useful predictions about molecular encapsulation.

In this context, this thesis project focused on the development of a computational platform for binding free energy prediction of host-guest complexes using both thermodynamic-based and knowledge-based approaches. The aim is not only to improve the overall knowledge in the field of supramolecular chemistry but also to provide new opportunities and applications for existing containers and provide direction in their rational development.

Key-words: Binding free energy prediction, semi-empirical quantum mechanics, Machine learning, automated platform