Computational approaches to molecular recognition : from host-guest to protein-ligand binding

Joel José Montalvo Acosta

▶ **To cite this version:**

Joel José Montalvo Acosta. Computational approaches to molecular recognition : from host-guest to protein-ligand binding. Theoretical and/or physical chemistry. Université de Strasbourg, 2018. English. NNT : 2018STRAF051 . tel-02145764

HAL Id: tel-02145764
https://theses.hal.science/tel-02145764

Submitted on 3 Jun 2019

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

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

UNIVERSITÉ DE STRASBOURG

ÉCOLE DOCTORALE DES SCIENCES CHIMIQUES

Institut de Science et d'Ingénierie Supramoléculaires

Institut de Chimie

THÈSE présentée par :

Joel José Montalvo Acosta

soutenue le : **4 september 2018**

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

Discipline/ Spécialité : Chimie/ Chimie informatique et théorétique

Approches Computationnelles de la Reconnaissance Moléculaire: L'analyse de la Liaison Hôte-Invité et Protéine-Ligand

THÈSE dirigée par :

M. CECCHINI Marco	Maître de Conférences, Université de Strasbourg

RAPPORTEURS :

M. MICHEL Julien	Professeur, Université d'Édimbourg
M. SIMONSON Thomas	Professeur, École Polytechnique

AUTRES MEMBRES DU JURY :

M. ROGNAN Didier	Professeur, Université de Strasbourg
M. RENDINE Stefano	Chercheur, Syngenta AG

COMPUTATIONAL APPROACHES TO MOLECULAR RECOGNITION: FROM HOST-GUEST TO PROTEIN-LIGAND BINDING.

JOEL JOSÉ MONTALVO ACOSTA

Institut de Science et d'Ingénierie Supramoléculaires - UMR 7006
Institut de Chimie - UMR7177
École Doctorale des Sciences Chimiques
Université de Strasbourg

Thesis Director: Dr. Marco Cecchini
August 3, 2018

Dedicated to my parents
Aida and Joel

ACKNOWLEDGMENTS

First of all, I would like to give infinite thanks to my doctoral supervisor, Dr. Marco Cecchini, with his patience, friendship, dedication and motivation has greatly strengthened personal and scientific aspects. Since my first day in the laboratory I have felt at home since Marco allows personal scientific development, curiosity to find new ideas and always open to listen to new proposals to solve the problems under study. I am eternally grateful with Marco for giving me the honor of working with him my doctorate studies.

I would also like to thank the members of the Laboratoire d'Ingénierie des Fonctions Moléculaires (IFM), with whom I have shared for many years excellent moments at scientific and personal perspectives. All of them, my friends, members of IFM, I thank you so much.

I thank the University of Strasbourg for being the place where I could develop my doctoral studies and to la Fondation pour la Recherche Médicale for providing the financial support to realize this important step in my scientific career.

Finally, I am very grateful to my family, who from a distance, beyond crossing the Atlantic, have been supporting me with love and patience in this long, although beautiful, life of science.

PUBLICATIONS

Concepts, ideas and figures have appeared previously in the following publications:

J. J. Montalvo-Acosta and M. Cecchini. "Computational Approaches to the Chemical Equilibrium Constant in Protein−ligand Binding.". *Molecular informatics*, 35(11-12), pp. 555–567, **2016**.

J. J. Montalvo-Acosta, P. Pacak, , D. Barreto-Gomes and M. Cecchini. "A Linear Interaction Energy Model for Cavitand Host−Guest Binding Affinities". *The Journal of Physical Chemistry B*, 122(26), pp. 6810–6814, **2018**.

J. J. Montalvo-Acosta, M. Dryzhakov, E. Richmond, M. Cecchini and J. Moran. "A Supramolecular Model for the Co−Catalytic Role of Nitro Compounds in Brønsted Acid Catalyzed Reactions". *The Journal of Organic Chemistry*, Submitted, **2018**.

RESUMÉ

LE CONTEXTE

La reconnaissance moléculaire représente un événement central dans de nombreux processus chimiques et biologiques pertinents. Dans le domaine biologique, un exemple de reconnaissance moléculaire est la liaison d'un ligand (un drogue ou modulateur endogène) à une protéine cible. Comprendre comment les ligands se lient aux biomolécules est d'une importance fondamentale pour les disciplines fondamentales et appliquées. D'autre part, la reconnaissance moléculaire est largement observée dans les complexes hôte-invité, une zone d'intérêt accru au cours des dernières années pour les chimistes expérimentaux et computationnels. L'hôte est une petite molécule synthétique avec une cavité bien définie où un certain nombre de composés se lient avec une affinité remarquable. La formation de complexes hôte-invité en solution est conduite par les mêmes forces non-covalentes qui apparaissent dans la liaison protéine-ligand, ce qui fait d'eux des systèmes modèles appropriés pour explorer la liaison non-covalente plus complexe comme protéine-ligand. De plus, de nombreux hôtes synthétiques ont montré des applications technologiques intéressantes en tant que chimiosensors, contenants de réaction, biomimétiques, amplificateurs de solubilité ou transporteurs de médicaments. Les principaux facteurs qui régissent la reconnaissance moléculaire sont de nature thermodynamique, en particulier, la valeur de la constante d'équilibre de liaison (K_{eq}), qui est dictée par le changement d'énergie libre molaire standard sur la complexation (ΔG_b°) ou la différence absolue de potentiel de liaison chimique ($\Delta\mu_b^\circ$), est la quantité d'intérêt. La possibilité d'accéder à la constante de liaison de manière précise et à partir des premiers principes fournirait une compréhension chimique de la reconnaissance récepteur-ligand, décrivant ainsi des lignes directrices pour des médicaments de conception rationnelle ou des échafaudages pour des hôtes synthétiques. De plus, des prédictions fiables de l'affinité de liaison récepteur-ligand par le calcul réduiraient considérablement les coûts d'innovation et de RetD, par exemple au début du développement de médicaments et stimuleraient un développement plus efficace de nouveaux produits pharmaceutiques par les sociétés pharmaceutiques. Cependant, le calcul de la constante de liaison dans la liaison récepteur-ligand pose en soi un défi théorique et de calcul exceptionnel. En même temps, les méthodes actuelles d'évaluation de l'affinité de liaison présentent divers degrés d'efficacité en termes de temps de calcul et de temps personnel pour obtenir des résultats fiables. Ce fait augmente encore plus les difficultés pour sélectionner une méthodologie appropriée pour l'analyse de cas spécifiques de reconnaissance de liaison récepteur-ligand. Dans ce contexte, le besoin de nouvelles approches computationnelles pour évaluer l'affinité de liaison avec une précision et une efficacité élevées est actuellement très nécessaire tant pour le secteur industriel que pour la recherche fondamentale.

Un cadre de mécanique statistique pour évaluer numériquement les affinités de liaison protéine-ligand

Le calcul de l'affinité de liaison protéine-ligand est loin d'être trivial. Dans ce but, plusieurs approches de calcul ont été développées au cours des années, qui abordent le problème à divers degrés d'approximation. L'interprétation de la mécanique statistique présentée ici suggère qu'il existe deux approches générales de l'énergie libre de liaison standard (voir Figure 0.1). Une approche passe par l'évaluation (directe) des potentiels chimiques absolus pour tous les composants de la réaction de liaison (c'est-à-dire le ligand, la protéine et le complexe). L'autre traite la réaction de liaison comme un équilibre de partition du ligand entre les états lié et non lié, ce qui suppose que la plupart des contributions protéiques à la différence de potentiel chimique s'annulent effectivement. Au meilleur de nos connaissances, toutes les approches rigoureuses de la constante de liaison telles que la perturbation de l'énergie libre (FEP en anglais) ou celles basées sur les Potential Mean Force (PMF en anglais) tombent dans la seconde classe; voir la Figure 0.1.

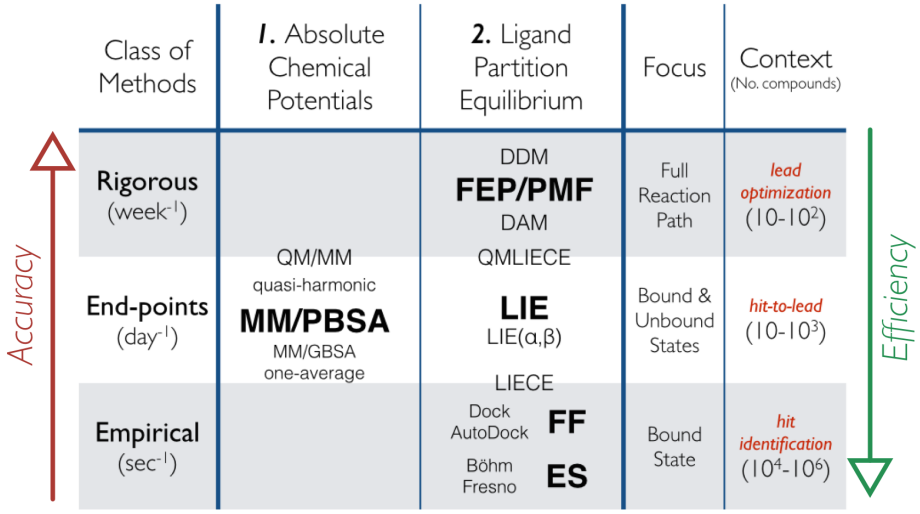


Figure 0.1: Classification des méthodes pour le calcul de l'affinité de liaison protéine-ligand.

Ces méthodes sont très intensives en termes de calculs et peuvent être utiles pour ne classer qu'un petit nombre de composés, typiquement moins d'une centaine, au stade de l'optimisation du plomb. La situation est différente pour les approches semi-rigoureuses ou terminales où MM/PBSA appartient à la première classe et le modèle d'énergie d'interaction linéaire (LIE en anglais) à la seconde classe; voir la Figure 0.1. Dans les deux cas, la constante de liaison est accessible en résolvant un cycle thermodynamique qui implique un transfert moléculaire vers la phase gazeuse. Cette stratégie transforme efficacement le calcul de l'énergie de liaison libre standard en une différence entre les énergies sans solvatation (approximatives), qui peuvent être évaluées avec beaucoup moins de calculs. En remplaçant la

représentation explicite du chemin de liaison par des estimations approximatives de l'énergie sans solvatation basées sur des modèles de continuum ou la théorie de la réponse linéaire, ces méthodes allègent la charge de calcul de façon significative. doit être évalué et classé. Bien sûr, la qualité des prédictions dépend de manière critique de la précision des calculs d'énergie sans solvatation, ce qui motive davantage d'efforts pour le développement de modèles de solvants implicites plus précis. L'analyse des approches empiriques rapides de la protéine-ligand (incluant certaines des fonctions les plus populaires pour l'amarrage) montre que ces méthodes décomposent le coût de calcul en se concentrant exclusivement sur l'état lié, c'est-à-dire sur la protéine ou le ligand en solution, et appartiennent donc à la seconde classe ; voir la Figure 0.1. Comme le taux de production moyen est d'une détermination de l'énergie libre par seconde, ces approches simplifiées sont appropriées pour le criblage de millions de composés et trouvent une utilisation répandue à l'étape d'identification des impacts. Néanmoins, l'accélération significative est obtenue en introduisant une série d'approximations théoriquement injustifiées, qui se traduisent par des erreurs systématiques importantes qui rendent les prédictions souvent peu fiables et/ou fortement dépendantes du système.

Le modèle d'énergie d'interaction linéaire pour les systèmes hôte-invité

Les complexes hôte-invité représentent un modèle intéressant pour tester de nouveaux développements de méthodes de calcul pour obtenir. Nous présentons un modèle d'énergie d'interaction linéaire (LIE en anglais) pour les systèmes cavit et hôte-invité. Dans LIE, ΔG°_{b} est calculé en évaluant l'énergie d'interaction du ligand avec son entourant à la fois dans les états liés (dans le complexe avec le récepteur) et non lié (libre dans la solution). Ces énergies d'interaction non-liantes sont normalement divisées en van der Waals et en contributions électrostatiques, pondérées par deux paramètres empiriques, α et β, obtenus par ajustement linéaire à partir de valeurs expérimentales. Ici, les paramètres LIE ont été générés en utilisant un ensemble d'entraînement de 14 complexes basés sur l'hôte cucurbit[7]uril (CB7) en ajustant linéairement les énergies d'interaction ligand / environnement calculées à partir des simulations de Dynamique Moléculaire avec le champ de force général ambre (GAFF) par rapport aux affinités de liaison expérimentales dans l'eau. Le caractère prédictif des paramètres LIE obtenus, $\alpha = 0.43$ et $\beta = 0.2$, a été évalué en utilisant un ensemble de test de 49 complexes de cavit et d'hôte-invité chimiquement divers. L'ensemble d'essai comprenait des complexes d'octa-acide (OAH), tétra-endométhyl-octa-acide (OAM), de β-cyclodextrine (BCD) et de CB7. Ces familles d'hôtes sont capables de lier un large spectre d'invités chimiques, de petites molécules rigides à des molécules plus flexibles. Les mesures statistiques utilisées pour évaluer l'exactitude de la méthode étaient l'erreur quadratique moyenne (RMSE) des expériences. Les résultats de la Figure 0.2 (à droit) montrent une corrélation frappante avec les expériences ($R = 0.81$) avec un RMSE calculé de 1.08 kcal/mol. Notez que cette erreur est plus faible que toute autre rapportée dans des études précédentes utilisant une variété de méthodes de calcul. Remarquablement, des prédictions précises ont été obtenues pour les hôtes OAH (RMSE = 0.66 kcal/mol), OAM (RMSE = 1.06 kcal/mol) et BCD (RMSE de 1.48 kcal/mol), qui ne faisaient pas partie de l'ensemble d'entraînement. Ces résultats indiquent que les paramètres

LIE ci-dessus sont transférables parmi des familles d'hôtes chimiquement diverses. Dans le cas des complexes CB7-guest, un RMSE de 1.17 kcal/mol a été obtenu, dépassant la précision obtenue par des méthodes plus rigoureuses basées sur des calculs quantiques coûteux (RMSE = 1.94 kcal/mol) et/ou des méthodes d'échantillonnage extensives basées sur MD (RMSE = 5.05 kcal/mol). La précision des prédictions ci-dessus indique qu'un modèle LIE simple est capable de capturer les détails modulant l'affinité de liaison hôte-invité dans la solution.

Pour évaluer l'impact du champ de force sur l'exactitude des prédictions, un nouvel ensemble de paramètres LIE a été dérivé en utilisant le champ de force général de la CHARMM (CGenFF). Le nouveau modèle LIE paramétré sur le même ensemble d'apprentissage présente $\alpha = 0.66$ et $\beta = 0.08$. Il est frappant de constater que malgré les valeurs de CGenFF LIE sensiblement différentes de la paramétrisation précédente, le RMSE calculé était de 0.92 kcal/mol, ce qui est cohérent avec la précision obtenue à l'aide de la GAFF. En outre, le modèle LIE basé sur CGenFF produit des résultats précis (Figure 0.2 à gauche) pour les hôtes individuels OAH (RMSE = 1.06 kcal/mol), OAM (RMSE = 0.97 kcal/mol), CB7 (RMSE = 0.88 kcal/mol) et BCD (RMSE = 0.54 kcal/mol). Ainsi, bien que les paramètres LIE dépendent du champ de force, la précision des prédictions d'affinité de liaison dans ces complexes hôte-invité ne l'est pas. Enfin, l'analyse de convergence des prédictions d'affinité de liaison en fonction de l'échantillonnage MD montre que des prédictions fiables peuvent être obtenues avec une simulation aussi faible que 1.1 ns dans les états liés et non liés. Pris ensemble, ces résultats supportent la conclusion que LIE fournit un accès précis et efficace à l'affinité de liaison de cavitand hôtes-invités; ce qui le rend approprié pour le criblage virtuel de grandes bibliothèques chimiques. En tant qu'application, le modèle LIE basé sur GAFF a été utilisé pour prédire l'affinité de liaison de 19 stéroïdes, qui se sont révélés se lier aux hôtes CB7 et CB8 avec des affinités nanomolaires dans l'eau. L'application directe de LIE a produit un RMSE de

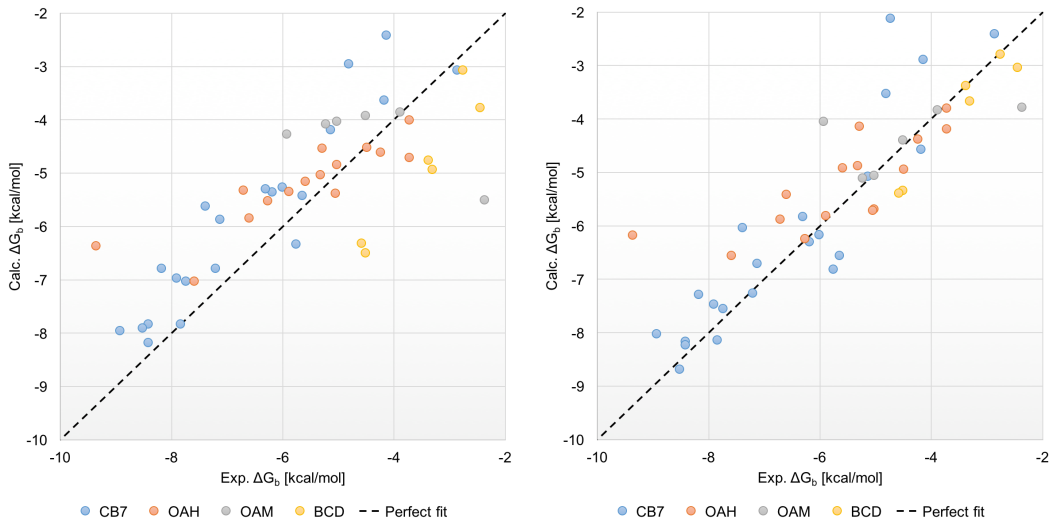


Figure 0.2: Valeurs d'énergie libre de liaison expérimentale vs calculée en solution aqueuse pour les systèmes hôte-invité de l'ensemble de test du modèle LIE basé sur GAFF (à droite) et CGenFF (à gauche).

2.45 kcal/mol à partir des expériences; voir la Figure 0.3 (points bleus). Nous avons observé que la déformation significative de l'hôte est introduite par les stéroïdes encombrants à l'état lié, en particulier sur le plus petit hôte CB7. Puisque la théorie de LIE suppose que le ligand est petit par rapport au récepteur et que la conformation de ce dernier est minimalement affectée lors de la complexation, il ne peut pas rendre compte de la souche du récepteur dans l'état lié. Dans les complexes stéroïde-cucurbituril, cependant, cette hypothèse est injustifiée car les hôtes stéroïdes ont une taille comparable à celle de l'hôte. Sur la base de ces considérations, nous avons développé un modèle LIE original qui tient compte de l'énergie de contrainte de l'hôte (ΔE_{str}) dans l'évaluation de l'énergie libre de liaison. Ainsi, dans la nouvelle formulation est la différence entre les énergies de champ de force pour les états liés et non liés au minimum, respectivement. Fait frappant, les résultats de la Figure 0.3 (points orange) montrent que l'inclusion de la contribution énergétique de la souche améliore considérablement les prédictions d'affinité de liaison avec un RMSE final de 0.81 kcal/mol et R^2 = 0.67.

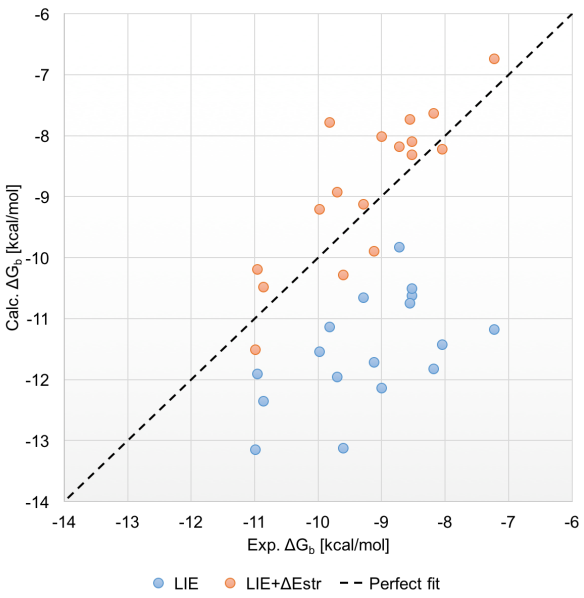


Figure 0.3: Valeurs de l'énergie libre reliée expérimentale vs calculée en solution aqueuse pour les complexes CB[7,8]-stéroïdes

L'effet co-catalytique des molécules "inertes" dans les réactions catalysées par l'acide de Brønsted

On a montré que la présence de composés nitrés modifiait les vitesses de réaction et la dépendance de la concentration cinétique des réactions catalysées par l'acide de Brønsted, y compris la déshydroazidation à l'alcool et l'hydrochloration des oléfines. Cependant, aucun modèle mécaniste n'existe pour rendre compte de ces observations. Ici, l'effet co-catalytique du composé nitro "inerte" dans la réaction d'azidation des alcools tertiaires catalysée par BCF, un acide de Bronsted organoborone, a été adressée par une approche combinée de la

modélisation moléculaire et des calculs DFT. Nous présentons un modèle supramoléculaire pour la forme catalytiquement active de l'acide généré par les calculs de DFT, constitué d'un agrégat lié à H de deux molécules d'acide de Brønsted et de deux molécules de composé nitro; voir Figure 0.4 (à droit). Les fréquences d'étirement O-H calculées pour l'agrégat sont d'excellents prédicteurs pour les vitesses de réaction expérimentales, contrairement aux bandes IR observées expérimentalement. En appliquant le modèle à un ensemble chimiquement divers de promoteurs potentiels, nous avons prédit et, de plus vérifié expérimentalement, que les esters de sulfate fournissent une alternative de travail aux composés nitrés; voir Figure 0.4 (à gauche). Ceci est le premier rapport de l'effet co-catalytique pour une famille chimique de composés non nitrés dans des réactions catalysées par l'acide de Bronsted.

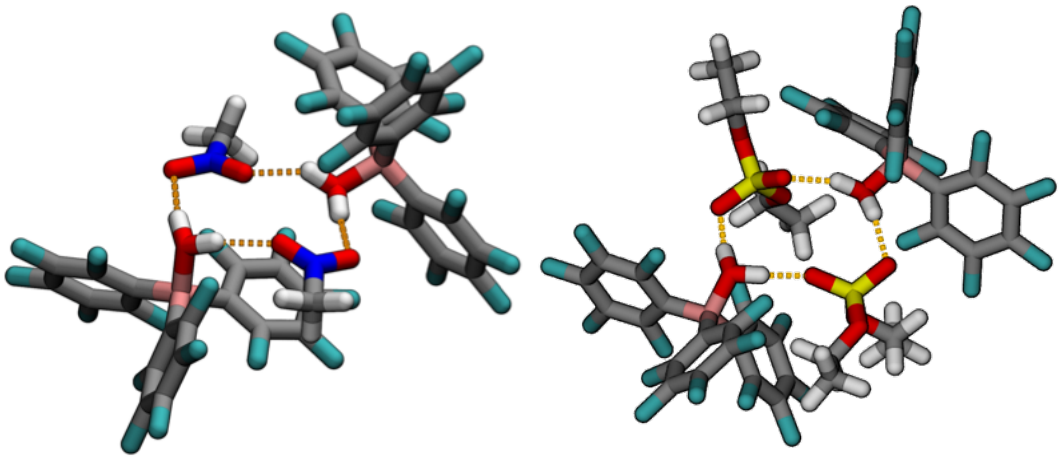


Figure 0.4: Structure optimisée par DFT pour l'auto-assemblage tétramérique 2:2 de BCF avec du nitrométhane (à droite) et du sulfate de diéthyle (à gauche).

CONCLUSIONS

Notre interprétation des approches computationnelles apparemment non reliées à la liaison récepteur-ligand dans le cadre commun de la mécanique statistique permet de repérer les approximations introduites pour accélérer les calculs, ce qui est utile pour rationaliser leur impact sur la précision de l'affinité de liaison prédictions. Notre analyse comparative met déjà en évidence des améliorations possibles d'approches semi-rigoureuses et empiriques bien établies et aide au développement de variantes avec un équilibre optimal entre précision et efficacité. En conséquence de ce cadre initial de mécanique statistique, nous avons présenté un modèle LIE pour les affinités de liaison cavit-hôte-hôte transférable entre des familles chimiquement diverses, précises et fiables, produisant des prédictions avec un RMSE <1.5 kcal/mol dans un grand test ensemble comprenant 49 invités et quatre hôtes différents. Notre modèle est efficace sur le plan des calculs et produit des résultats convergents en quelques nanosecondes de MD, ce qui ouvre la voie à des criblages informatiques à haut débit. On a montré que le caractère semi-empirique du modèle absorbait la plus grande

partie de l'erreur systématique du champ de force, rendant les prédictions essentiellement indépendantes du champ de force. Enfin, l'inclusion de l'énergie de contrainte de l'hôte dans le calcul de l'affinité de liaison, qui est absente dans la formulation LIE originale, s'est avérée améliorer sensiblement la qualité des prédictions, en particulier lorsque les hôtes et les invités ont des tailles similaires. L'utilité d'une formulation LIE pour la reconnaissance hôte-invité a été démontrée par la prédiction précise de la liaison stéroïdienne aux hôtes cucurbituril, qui sont technologiquement pertinentes pour le développement de chimiosensors. Enfin, un modèle de calcul a été généré pour tenir compte des changements dans l'échelle d'acidité d'un acide de Bronsted par l'inclusion d'une molécule apparemment inerte (composés nitrés) avec un effet co-catalytique. L'implication importante est que, pour être bien comprise, la catalyse acide de Brønsted doit être considérée dans une perspective supramoléculaire qui prend en compte non seulement le pKa de l'acide et les propriétés globales d'un solvant, mais aussi les interactions faibles entre toutes les molécules dans solution.

CONTENTS

1

INTRODUCTION

Nature is fascinating!!

It's remarkable the way how nature is able to carry out regulatory processes in a precise and controlled way. This phenomenon has molecular origin, through the interaction of two or more molecules, also called molecular recognition. In fact, it is essential in biology, chemistry and physics and it got highly attention from the 1980's with the emergence of the supramolecular chemistry by Lehn, Cram and Pedersen, Nobel prizes awared in 1987[1–3]. The essence of molecular recognition is the communication among molecules by non-covalent interactions and represents the heart of new research in chemistry aimed to design new chemical entities with technological applications. Thus, it is expected the creation of nano devices composed of agglomerations of molecules joined by non-covalent forces that can be controlled using external factors such as changes in temperature or concentration. A typical example is the use of self-assembly, a fascinating expression of molecular recognition in chemistry, for the fabrication of nano-scale electronic and photonic devices[4]. The importance of molecular recognition today is by far undeniable to scientists in many branches of basic and applied science.

Nevertheless, molecular recognition is, at least, little understood nowadays. For example, there are not (yet!) universal rules for the design of a small ligand that strongly and selectively binds to a specific binding site of a protein with known structural information, although this widely relevant as the fundamental principle for rational drug design. Similarly, the problem of "protein folding" remains almost intact after 50 years of research and it's not possible to predict the mechanism of folding for proteins (or agglomeration of them as in histons) knowing only the amino acid sequence.

Why such a relevant process still a mystery? a first possibility comes from the fact that we might not have a complete characterization of all non-covalent interactions present in complex molecular recognition events. With the birth of quantum mechanics since the 1920's, the chemical bond was practically solved[5] and much progress has been made in the elucidation of non-covalent interactions in simple systems such as homogeneous fluids or solid-liquid interfaces. Currently, we have very good knowledge of a list of non-covalent interactions such as electrostatic, dispersion or magnetic forces and we have characterized those of special relevance as the hydrogen bond[6]. However, the attraction experienced by anions and cations by the quadrupole of aromatic rings has only been revealed in the last 30 years but with little understood beyond simple models is present today as few applications exist exploiting this kind of interaction. Another special case is the halogen bond, a type of non-covalent attraction between an electrophilic region associated with a halogen atom and another nucleophilic region very similar to the hydrogen bridge, which was revealed on 1950's. A second and more definitive answer is related to the incomplete appreciation we have about the interconnection that exists among particular non-covalent forces in complex systems of molecular recognition.

That is, the mechanism by which individual non-covalent interactions reinforce or weaken/-compete with others in an event of molecular recognition with multiple and heterogeneous actors is still unclear. For example, in the field of rational drug design, one strategy for improving the binding affinity and specificity of a molecule is based on the addition of a polar group aimed to form a hydrogen bond with a specific residue in the binding site of the receptor. However, this extra polar group in the new ligand also increases its affinity for the aqueous medium (unbound state), as it also favors the hydrogen bond network in water and could end-up with a global decrease of the protein binding affinity with respect to the predecessor molecule. This case shows the difficulty to evaluate *a priori* how the introduction of a new chemical group in the ligand will increase its affinity for either the bound or the unbound states. It's hard to evaluate the contribution of this new interaction to reinforce the non-covalent interactions already present between the ligand and its two critical environments for binding affinity, inside the protein and free in solution, respectively.

This last point indicates a key aspect and is that today, in our opinion, we have a semi-quantitative understanding of molecular recognition. Excellent advances has been made oriented to characterize each individual non-covalent interaction participating in biological and chemical processes where molecular recognition is involved[6]. That is, from a qualitative point of view, we have certainty of what kind of non-covalent interactions are involved in most of the (bio)chemical process, for example the binding of a ligand to a protein or the self-assembly of monomers in solution. In fact, we know the relative strength which these interactions participating in a recognition event, the opening door for a more quantitative analysis. Since the last 40 years, great efforts have been made aimed at the quantification of molecular recognition using robust experiments based on isothermal calorimetry titration and nuclear magnetic resonance, at least in some systems with low and medium complexity as protein-ligand and protein-protein binding[7]. However, in more complex problems as catalysis in solution or 2D self-assembly, the insight and the estimation of affinity parameters is poor even by using the most advances experimental techniques (e.g., nuclear magnetic resonance or scanning electron microscopy) because a poor sensibility, low molecular resolution and/or other limitations of the experiments. On the other hand, notable theoretical advances on toy systems for quantitative molecular recognition has been performed since Hill's work[8] and others on 1950's[9,10]. These studies focused on simplified systems have provided the basis for the developing of computational methodologies to estimate parameters of binding affinities in more realistic systems, which is the base for the computer-aided molecular design. However, we lack a general method that allows us, from a quantitative point of view, to understand and control the recognition process of any molecular association. Moreover, we still do not have a universal methodology that allows us to predict events involving supramolecular systems of two or more molecules. Such predictions are the basis for the rational design of important chemical entities with social impact, such as new drugs, catalysts, drug carriers, biomimetics, among others. Although the computational power exponentially increase everyday, our state-of-the-art methodologies require a highly intense demand of computation not accessible nowadays, which forces to introduce drastic approximations confining the validity of the models to specific cases and loosing their generality. This lack of absolute predictive

power confirms that the phenomenon of molecular recognition remains not totally clear and we count with methods with an acceptable degree of success to make predictions and design just for specific cases of supramolecular recognition events.

This manuscript is focused on the study of molecular recognition from a quantitative approach, specifically in ligand-receptor binding systems such as host-guest and ligand-protein binding and catalysis in solution. This quantitative approach is possible thanks to the use of computational (numerical) methodologies based on statistical mechanics, a theory of atomistic resolution. With these methods it is possible to discern the conglomerate of non-covalent interactions present in these systems and provide a path to the understanding of chemical supramolecular events.

Part I

LES CONCEPTS

2

MOLECULAR RECOGNITION

2.1 CONCEPTS

The Nobel Prize in chemistry was awarded in 1987 to Cram, Lehn, and Pederson[1–3] for their work in the development and application of molecules that bind with high selectivity through structure-specific interactions. This gave birth to the supramolecular chemistry, a term given by Lehn[2], as a fundamental science encharged to study molecular complexes built by non-covalent interactions. Inside supramolecular chemistry emerged the concept of molecular recognition, another term given by Lehn[2], as the mechanism that decides the specific association among molecules. Thus, if a supramolecule is defined as "a molecular entity beyond a single molecule" bonded by reversible non-covalent interactions (e.g., hydrogen bonding, van der Waals forces, salt bridge, among others), the molecular recognition is the algorithm which dictates how this supramolecule is built. Molecular recognition covers many research areas of supramolecular chemistry in solution, as molecular (self)assembly and host-guest binding, and at interfaces, as the molecular imprinting[11,12].

Because we will focus on the analysis of molecular clusters ("supermolecules") bonded by intermolecular forces, we might first consider to define what is a molecule. For example, the diatomic specie Ar_2 formed by the binding of two argon atoms is not usually considered as a molecule. This is because the binding energy to form Ar_2 is low than kT (k is the Boltzmann constant) at room conditions and can be dissociated easily thorough atomic collisions. A contrast case is N_2, which has a binding energy of 226 kcal/mol at 298 K[13], one of the tightly bonded diatomic specie found in nature. Thus, we might define a molecule as a group of atoms with a binding energy larger than kT at room temperature and without losing its integrity when interacts with its environment. There are cases, as in proteins and other highly flexible molecules, where multiple and stable conformers of the molecule can be found after large conformational changes induced by the environment but they are accompanied by relative small changes in potential energy among the conformers.

Molecular recognition has been found among the most relevant events in chemical and biological systems, and provide a rich pool of chemical diversity and functions for both characterization and design of materials and assemblies[11]. Thus, we can classify many molecular recognition events according to the properties of the supramolecular assembly such as size, shape, non-covalent interactions involved, etc. Along this line, we might find molecular recognition at the molecular, meso- and nano- scales according to the system's complexity based on size and shape of the supramolecule[12]. From the development of the ionophores crown ethers and cryptand[2,3], a wide group of synthetic receptors as tweezers, clefts and cavitand hosts have been prepared[14]. They normally form 1:1 complexes with variable guests; ions, drug-like compounds, amino-acids, etc, and they are examples of molecular recognition at the molecular scale (Figure 2.1). Many approaches aimed to the design of receptors have

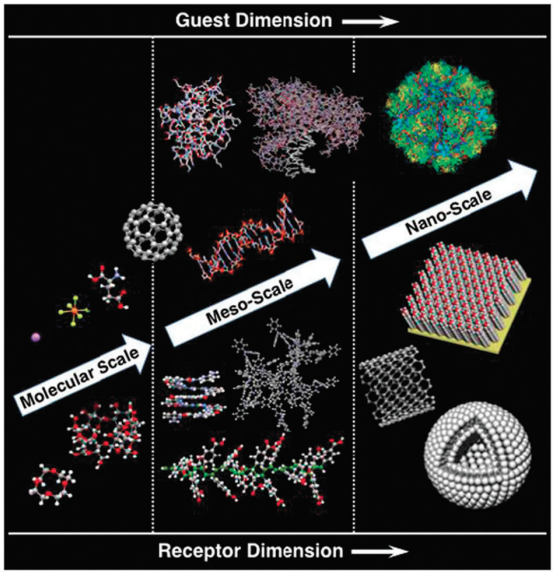


Figure 2.1: Different scales for molecular recognition from solution to material chemistry[12].

been emerged using molecular imprinting protocols, template reaction and computer-aided design at the molecular scale. Moreover, new research is point out to molecular recognition on the meso-scale using more specialized receptors (Figure 2.1). These families of host are designed to have larger guest cavities (>1 nm) to accept many types of guests with high diversity in shape and size and chemical groups. As examples on meso-scale recognition we have dendrimers and molecular capsules, normally used as drug delivery systems. When the molecular association forms a supramolecule with dimensions around hundreds nanometers we are in the regime of molecular recognition at nano-scale. Here, the ultimate goal is the fabrication of nanodevices with autonomous (smart) or semi-autonomous behavior, able to perform specific and efficient tasks by following of external instructions from the user[12] (Figure 2.1).

2.2 RELEVANCE OF MOLECULAR RECOGNITION

After 30 years, supramolecular chemistry and molecular recognition is being used in many branches of basic science and the technological revolution new millenium[7,14]. From the perspective of basic science, molecular recognition is at the heart of biology and biochemistry. Knowing the intrinsic mechanisms that underline the selective binding process of a modulator ligand to a protein will lead to clarify and control all metabolic pathways in a cell. Beyond, it opens the door to the field of rational design of receptors and/or ligands for technological advances. Among these technologies we find medical applications, energy storage, chemosensors in environment chemistry among others, where the relevance of non-covalent interactions among molecules is highlighted. Other advances done after the emergence of molecular recognition is at the nanotechnology level with the fabrication of nano- and micro-

devices based on molecular self-assembly [4]. Clearly the importance of study molecular recognition today is highlighted from several edges of basic and applied sciences.

As a case study we analyze the sugar-protein recognition, among many others process with biological relevance, where a deep insight and quantitative analysis of molecular recognition provide the basis for controlling the biochemistry of the cell. The sugar-protein recognition is involved in critical process such as cell-cell communication, energy production by their own catabolism, inmunological recognition among others. Sugars constitute the main and preferred energy source for cells. A complete network of regulatory gens is activated or induced by a specific sugar in order to produce proteins which are able to metabolize that particular sugar. In contrary cases when the inducer is absent, the entire genetic system is stopped or repressed. This beautiful circuit is carried out on the basis of a well-defined molecular recognition principle. In a first case, a repressor protein is synthesized when there is low concentration or absence of the sugar, which binds to DNA (operational region) and block the RNA polymerase movement. As a consequence, the transcription of genes needed for the catabolism of the sugar is stopped. In a second case, the activation process is based on the binding of the sugar with the repressor protein which inhibits the binding of the repressor with DNA sequence and allows the synthesis of catabolic enzymes for sugars. Here, the sugar-inducer/repressor mechanism is an unique example of molecular recognition. The whole regulatory circuit uses the substrate sugar as guarantee of the cell survival and it's an excellent target for drug design.

2.3 NONCOVALENT INTERACTIONS

In general, a molecule selectively recognizes its partner through various intermolecular interactions. These interactions are usually weak forces compared to chemical bonds in molecules and usually are called weak interactions. Several kind of non-covalent interactions are found with different features, although in essence, all of them comes ultimately from the electrostatic interaction among particles. Historically, these intermolecular interactions are classified in two main families: "long-range", with the interaction energy follows an in inverse power of the distance, and short-range, where the magnitude of the energy exponentially decreases with the distance (Table 2.1) .

The long-range interactions are: electrostatic, induction and dispersion. They are present when the interacting particles are separated at large distances. In general terms, the electrostatic effects are the most simplistic to follow. They can be quantified using the Coulomb law and they are pairwise additive with attractive or repulsive characters depending of the distance distribution of atoms.

Induction effects come from the changes in electronic structure for a particular molecule under the influence of a net electric field from all atoms in its environment. The induction interaction is strictly attractive but non-pairwise additive as the electric fields of molecules in the environment may intensify or cancel out each other.

Dispersion interactions are the less understood effects from a classical point of view because they arise from the highly fluctuating charge distribution as the electrons are in con-

Contribution	Additive?	Sign	Comment
Long-range $(U \sim R^{-n})$			
Electrostatic	Yes	$\pm$	Strongly dependence of orientation
Induction	No	-	Always present
Dispersion	Approx.	-	
Resonance	No	$\pm$	Degenerate states only
Magnetic	Yes	$\pm$	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	-	

Table 2.1: Classification of non-covalent interactions for molecular recognition [6].

stant motion in molecules. They have mostly quantum nature and are attractive interactions as the two molecules approach close each other. They come from the pre-defined correlation existing on the electron movements, which favors specific low-energy configurations for the two interacting molecules.

Two other interactions with long-range nature, the magnetic and resonance effects, are not deeply described as they usually don't participate in molecular recognition events at room temperature in most of the chemical and biological problems treated here. Firstly, the magnetic effect can occur involving electrons or nuclei. The electronic magnetic effects appear when both interacting molecules have unpaired spins, which is not so common in biological process, while the magnetic interactions involving nuclei are relatively several order of magnitude smaller in magnitude compared to other interactions as electrostatics or dispersion. Lastly, the resonance interactions arise when at least one interacting molecule is at a degenerate state (e.g., as an excited state) and they are not present in closed-shell molecules usually treated in molecular recognition problems at ground states (Table 2.1).

Interesting contributions to the total potential energy come from molecular interactions at short distances where electron exchange effects may be possible as the molecular wavefunctions might overlap significantly. The most significant effect at short-range is the exchange-repulsion (also known as just exchange), which is the result of two effects, one attractive and one repulsive. The former arises because in the proximity of the two interacting molecules, the electrons can freely flow in both molecules rather than one, which decreases the electronic momentum and energy as the uncertainty of the position is increased. The latter, repulsive effect comes from restrictions imposed by the Pauli antisymmetry principle where electrons with the same spin should not be found in the same orbitals which involves an energetic cost to pay. Normally the repulsive component is higher than the attractive one and the exchange-repulsion interaction has a net repulsive effect. Finally, other short-range interactions are exchange-induction, exchange-dispersion and charge transfer, which all of them arise from the overlap of the wavefunction (Table 2.1).

2.4 HOST-GUEST BINDING

Chemical hosts are defined as molecules with low molecular weight and a well-defined cavity used for binding other group of compounds, their chemical guests, to form stable complexes at experimental conditions. The host-guest binding is driven by conventional non-convalent interactions found in other association reactions as protein-ligand binding such as hydrogen bond, salt bridges, van der Waals interactions, etc. Revolutionary techniques in organic chemistry are allowed the synthesis of host families with divese applications in the field of chemosensors, biomimetics, reactions containers and chemo/drug transporters[15].

Hosts usually are constituted by specific monomers which provide the chemical identity to the whole host's family. These monomers can be present in the host in several units and localized in specific positions. A typical case is the cyclodextrin, where all members of this host family are constituted by glucose, a chiral and highly flexible small molecule, which transfers these properties to the cyclodextrin family[14].

Although host-guest complexes are relatively small, i.e., they present fewer degrees of freedom compared to protein-ligand systems, they still present critical issues specific of association reactions as flexibility of the host, desolvation and hydrophobic effects or changes in protonation states and tautomerism upon binding, which involve challenges for quantify their intrinsic molecular recognition[15]. All these features make the host-guest systems an interesting benchmark systems for evaluating binding affinities using computational methodologies because their relative small size allows to collect large statistics by run extensive sampling and remove any source of systematic error due to incomplete conformational sampling[16,17]. Additionally, experiments can be carried out in conditions with greater control where it is possible to eliminate uncertainties in the assignment of protonation states of host and guests. In this way, computational chemists can use these systems to evaluate other factors or source of errors in the binding affinity predictions[17].

In this spirit, around the latest 10 years a group of enthusiastic researchers have organized blind benchmarks or challenges to compare different computational methods for predict the same properties in specific systems in an objective manner. These tests become essential as they represent almost an unique way to assess the predictive power of current computational methodologies in a similiar way as the challenges presented in real life. Also, results from these challenges are the base for further refinement of methodologies with poor predictions. Along this line, the Statistical Assessment of the Modeling of Proteins and Ligands (SAMPL)[18] blind challenges and the Drug Design Data Resource (D3R)[19] grand challenges have become excellent projects to this aim.

First, D3R is centered on release blind challenges on predict protein-ligand binding both for capturing the binding mode of the ligand and its affinity for the receptor. The data sets for these tests are obtained from the pharmaceutical industry and the methods are tested in systems with direct pharmaceutical relevance. On the other hand, SAMPL focusses on predict basic physicalchemsitry properties of small molecules as pK_a, partition coefficients in aqueous and organic solutions and binding affinity for host-guest systems. Both the SAMPL and D3R challenges, roughly cover assess the prediction for interesting properties for drug

design/discovery by current computational methods[18,19]. Next, we review some families of hosts used in SAMPL challenges and/or technologically relevant applications. Most of those host-guest complexes have been used in the development and application of computational methods described in next chapters of this manuscript.

2.4.1 *Cucurbiturils*

The cucurbiturils (Figure 2.2) are cavitand hots formed by glycoluril monomers. From the discover of the first family member, cucurbit[6]uril, with six glycoluril units, many synthetic efforts have made possible to obtain members with five-, seven-eight- and 10-monomers in cucurbit[n]urils (CBn with n= 5, 6, 7, 8, 10). Surprisingly, the member CB9 has not been synthesized yet. As the number of monomers increases, the diameter of the cucurbituril cavity increases and it is therefore possible to accommodate larger and larger guests. Cucurbituril family mainly bind hydrophobic-core, neutral or cationic guests. The interior of the cavity is very hydrophobic while the carbonyl groups in surface of the hosts makes a hydrophilic portal where polar (mainly cationic) groups of the guests can interact. Among all members of the cucurbituril family, CB7 is the most widely used host since it is able to bind a large amount of different chemical guests with a 1:1 stoichiometry. Also, CB8 is being used in many applications as chemosensor or drug carriers as allow to bind guests with large size to form 1:1 or 1:2 host-guest complexes. CB7 is experimentally suitable because is very soluble in water, and computationally, because it is compact, rigid, without ionizable groups and can tightly bind many guests. For all these reasons, CB7 has been used in several SAMPL challenges as well as being used in the most recent HYDROPHOBE challenge[20] aimed to assess the quality of predictions by computational methods for purely hydrophobic guests.

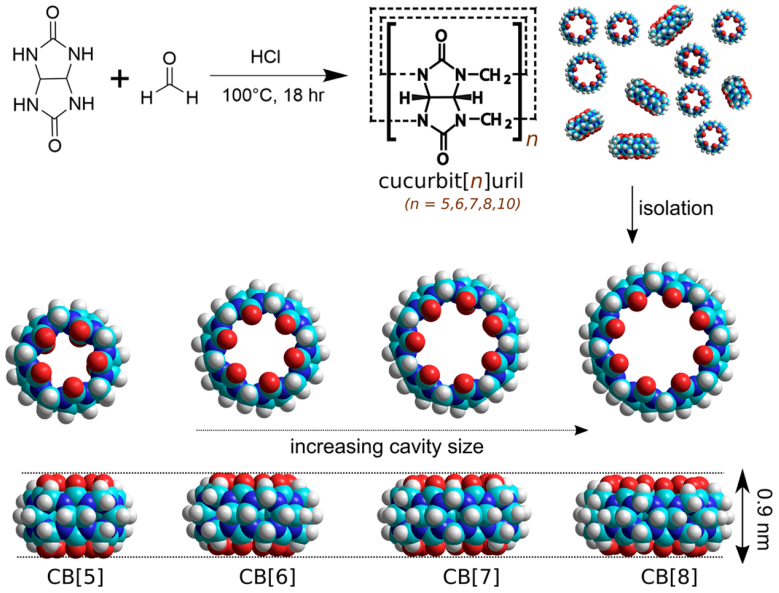


Figure 2.2: Chemical structures for members of the cucurbit[n]uril family[21].

2.4.2 *Octa acids*

The octa acid (OAH) and tetra-endomethyl octa acid (OAM) hosts (Figure 2.3) belong to a synthetic host family characterized by a basketshaped, hydrophobic and deep cavity[15]. These hosts have eight carboxylic acidic groups which make them highly soluble in water as they point out to the solvent. The cavity of OAM is much more sterically constrained than the cavity of OAH due to the presence of four methyl groups extra in the proximity of the host's portal. These extra methyl groups confer different binding properties in terms of selectivity and strength to OAM in comparison to OAH. Both OAH an OAM present a rigid cavity (compared to the cucurbituril family), which bind guests containing i) a hydrophobic core that fits inside the cavity of the hosts, and ii) a hydrophilic head that points out into the solvent. Thus, octa acids family are able to bind a diverse set of ligands as neutral, cationic and anionic organic compounds. In fact, complexes of OAH and OAM with cyclic carboxylic acids were included on previous versions of SAMPL challenges. These host-guest complexes provide a broad range of aqueous binding data suitable for exploring the performance for making predictions by diverse computational methods.

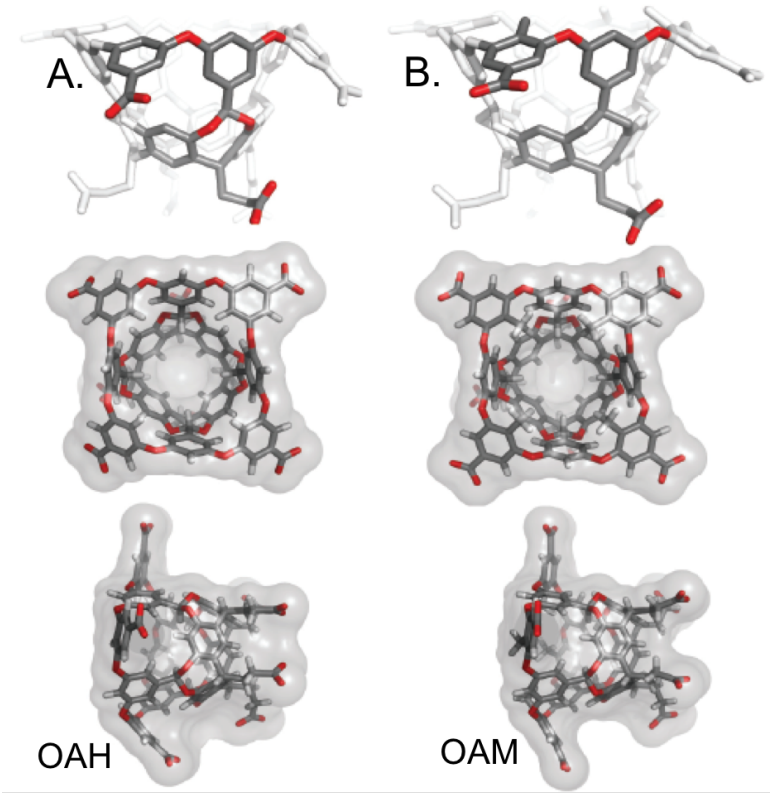


Figure 2.3: Chemical structures for octa acid (A) and tetra-endomethyl octa acid (OAM) hosts[15].

2.4.3 *Cyclodextrins*

The family of cyclodextrins (Figure 2.4) are composed by naturally ocurring hosts having a glucose molecule as the fundamental unit[22]. The most common cyclodextrins contain six, seven and eight units of glucose and are named α-, β- and γ- cyclodextrins (Figure 2.4). The monomers are organized in a vertically stand up position with the primary hydroxyl group located at the side of a narrow cavity while the secondary hydroxy groups are pointed out toward the reverse side in direction to a wide cavity. Thus, there are not hydroxyl groups on the borders of the wall, (i.e., the medium region) giving a hydrophobic character to the cavity. Thanks to this property, cyclodextrins can bind mostly hydrophobic guests in their cavities such as alkyl and aromatic derivatives. As in cucurbiturils, the size of the cyclodextrins defines the kind of guests to bind and it depends of the number of glucose units in the cyclodextrin host. Nevertheless, cyclodextrins are much more flexible than cucurbiturils because the flexibility of the glucose monomer in the former is higher than the glycouril monomer in the latter. Additionally, the glycosidic bounds which link all the glucose members gives more freedom for internal movements in cyclodextrins. Thus, they are also interesting systems for computational studies as the cavity of cyclodextrins remains a micro-hydrophobic medium for guests, analogous to the binding site of proteins. Some cyclodextrin derivatives obtained by chemical modifications of their hydroxyl groups have been widely used in the pharmaceutical and food industries as transporters of drugs and flavors. Those modified cyclodextrin present a major compatibility with biological fluids than natural cyclodextrins in terms of solubility, stability and side effects[14].

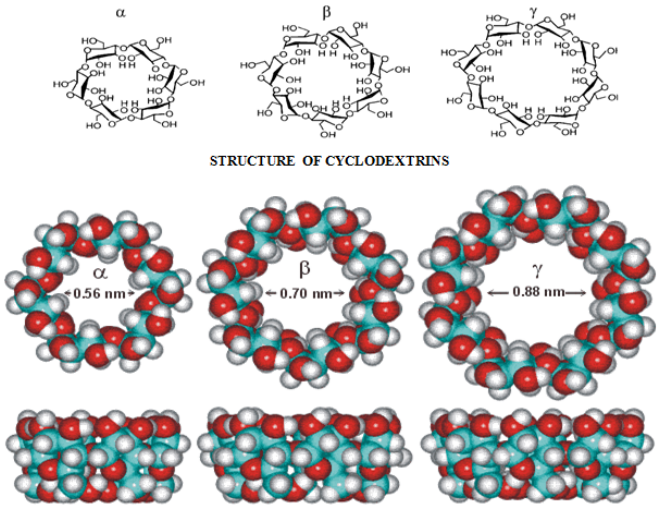


Figure 2.4: Chemical structures for natural products α-, β- and γ- cyclodextrins[22].

3

THERMODYNAMICS OF RECEPTOR-LIGAND BINDING

3.1 THERMODYNAMICS PARAMETERS OF BINDING REACTIONS

Although many kinetics effects can affect the binding of a receptor (as a host) by a ligand (as a guest), specially in heterogeneous medium as the interior of a cell, the molecular recognition is an equilibrium phenomena and follows the thermodynamics laws.

Considering a general form of an association reaction for a receptor (R) and a ligand (L) with a 1:1 stoichiometry as follows:

$$R + L \leftrightharpoons RL \tag{3.1}$$

This equilibrium reaction is characterized by thermodynamics parameters, being the standard molar Gibbs free energy difference of binding, the most important for quantitative analysis and design, as:

$$\Delta G_b^\circ = -RT \ln (K_{eq} C^\circ) = \Delta H_b^\circ - T\Delta S_b^\circ \tag{3.2}$$

where T is the experimental temperature and R is the gas constant. The equilibrium dissociation constant, K_{eq} is the ration between concentrations of the reacts and the products at specific equilibrium conditions at 1 M concentration (C°) and constant temperature and pressure. The other interesting thermodynamic quantities ΔH_b° and ΔS_b° are the standard enthalpy and entropy differences of binding, respectively.

Each thermodynamic parameters express the relationship between products and reacts. A negative value for ΔG_b° indicates that the binding reaction is favorable under standard conditions and the process is exergonic and spontaneous toward the formation of the complex. In contrary case, a positive value for ΔG_b° is described as endergonic change and the binding process is not spontaneous. Also, negative values for ΔH_b° and positive values for $T\Delta S_b^\circ$ are favorable for the complex formation. Thus, these two thermodynamic functions have the tendency to make oppose contributions to affinity, which is known as the enthalpy-entropy compensation. For example, a negative value of ΔH_b° indicates an increase of bonding terms, which reduces the disorder of the system, and consequently, decreases the magnitude for $T\Delta S_b^\circ$ is found. Negative values for ΔH_b° indicates that the binding reaction is an exothermic process, whereas a positive value corresponds to an endothermic process. On the other hand, the equilibrium constant K_{eq} can be understood as a measure of the free ligand concentration needed for saturate 50% the receptor, as smaller its value bigger will be the strenght of the ligand affinity by the receptor.

Another interesting thermodynamic parameter to consider is the change in heat capacity (or specific heat) upon binding at constant pressure, ΔC_p, as

$$\Delta C_p = \frac{d\Delta H_b^\circ}{dT} \tag{3.3}$$

A negative value of ΔC_p measures that the reacts have a higher heat capacity than the products and its macroscopic interpretation is hard since many factors induces a decrease of ΔC_p.

3.2 EXPERIMENTAL DETERMINATION OF BINDING THERMODYNAMICS

Many experimental techniques are useful for estimate thermodynamics of binding affinities, however we will focus on isothermal titration calorimetry (ITC) as the reference methodology for experimentally assess thermodynamic values for binding reactions. This is because ITC is a powerful tool for thermodynamic characterization of compounds which bind to a target receptor. It's widely used to gain a major understanding of (bio)molecular recognition for the design of improved ligands. Speficially, it's one of the most used techniques in drug design projects in the pharmaceutical industries. Also, the thermodynamics of host-guest chemistry are measured using this techniques. ITC allows to get directly, and in a single experiment, many of the thermodynamic parameters described above such as K_{eq}, ΔH_b° and the stoichiometry of the reaction. The others thermodynamic quantities as ΔG_b° and ΔS_b° are obtained using Eq. 3.1.

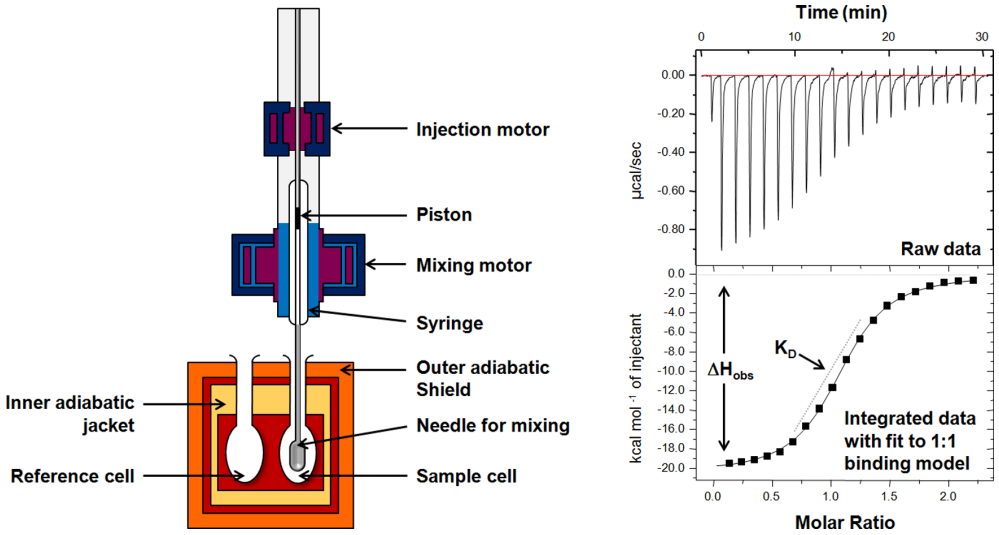


Figure 3.1: Schematic representation of an instrument for ITC experiments (left) and the typical titration curve obtained from an ITC experiment with the associated data analysis[23].

A typical instrument for ITC have two cells, one for the sample and one for the reference, which are inserted in an adiabatic jacket (Figure 3.1). The cells and jacket are connected to independent heaters with devices measuring the temperature difference between the two

cells and the jacket. In the course of an an experiment, the reference cell is heated by a very small constant power and a variable power is applied to the sample cell in order to maintain a fixed temperature difference relative to the reference. The voltage applied to the sample cell is the experimentally measured quantity which is related to the heat of the reaction. Exothermic reactions generate a negative signal, whereas endothermic reactions produce a positive applied voltage. Integration of this power with respect to time gives the amount of heat produced, or absorbed, by the reaction[24].

For an ITC run, the test compound is injected into a sample of target receptor. While the reaction takes places and Interaction between the compound and the target leads to release, or uptake, an small amount of heat (< 5 μcal in a modern instrument) whilst the mixture is kept at constant temperature. Monitoring of the relationship between dose of ligand and magnitude of heat change allows direct calculation of K_{eq} and ΔH_b° and stoichiometry (n) of the reaction (Figure 3.1). Additionally, ITC presents other advantages such as the straightforward analysis of data and the high precision of the experiments, make suitable for the widespread use in both academical as industrial environments. Few limitations can be found for ITC experiments. In a first case, the assays require a not negligible amount of the receptor, which can be an issue if this material is complicated to obtain (proteins, DNA, etc) although this technique does not destroy the sample. Also, ITC does not performance very well when there are cooperative interactions among multiple binding sites, as it is difficult to decompose the heat signal among the distinct binding modes.

3.3 THERMODYNAMIC PARAMETERS OF BINDING BY STATISTICAL MECHANICS

Thermodynamic parameters for characterizing binding reactions as involved in host-guest and protein-ligand complexes can be obtained from microscopic information of the system, in addition to isothermal titration calorimetry described before or other experimental techniques. In fact, there is a great interest into get binding affinities data without to perform experiments but using theoretical and computational (numerical) approaches. This is because in many cases experiments are difficult to perform as the compounds involved are unstable in the experimental conditions or the sensibility of the technique is low. Even more important, computational approaches based on fundamental laws of nature provide, in addition to the thermodynamic quantities of interests, an all-atomistic view of the binding reaction which allows a deeper understanding of the underline process. This knowledge is the fundamental base for the design of new ligands who binds a specific target in study. Finally, this computational approaches are cheaper and, with the exponential grow of the computer resources, they are eventually faster than actual experiments in a laboratory given as a consequence a better optimization of the resources involved in projects aimed to get binding affinities for large database of molecules.

Statistical mechanics is the set of laws which allows to obtain thermodynamic properties of a system using atomistic information, this is, the science that does a bridge and connects the macroscopic and the microscopic worlds. In the next sections, we will give a short summary of the most important statements in statistical mechanics that are involved for estimate ther-

modynamic quantities of binding reactions by computational approaches. A major review of the statistical mechanics concepts can be found elsewhere[9].

The formalism presented here aims to obtain the free energy of the system and the difference in free energy of two well-defined states of the system from atomistic information. Considering a system in the canonical ensemble where the temperature, the volume and the number of particles are constant, many important statistical thermodynamics variables can be obtained from numerical approaches. In all scenarios of statistical mechanics[9], the fundamental quantity is the configurational partition function,

$$Z_N = \frac{1}{N!h^{3N}} \int exp\left[-\beta E\left(\mathbf{x}^N\right)\right] d\mathbf{x}^N \tag{3.4}$$

where T is the absolute temperature, h is the Planck's constant, $\beta = (kT)^{-1}$ is the Boltzmann's factor and $E\left(\mathbf{x}^N\right)$ is the configurational energy and the integration is extended over all the configurational space $d\mathbf{x}^N$ for the N-particles system. From the configurational partition function, many thermodynamic quantities are obtained as the internal energy:

$$U = \left\langle E\left(\mathbf{x}^N\right)\right\rangle = \int E\left(\mathbf{x}^N\right) P\left(\mathbf{x}^N\right) d\mathbf{x}^N \tag{3.5}$$

where the $\langle\,\rangle$ implies an ensemble average obtained by molecular simulations and the Boltzmann's probability function, $P\left(\mathbf{x}^N\right)$, defined as:

$$P\left(\mathbf{x}^N\right) = \frac{exp\left[-\beta E\left(\mathbf{x}^N\right)\right]}{Z_N} \tag{3.6}$$

and the Helmholtz free energy of the system, i.e., the free energy at the canonical ensemble:

$$A = -kT \ln\left[Z_N\right] \tag{3.7}$$

On the other hand, the Gibbs free energy, the free energy at constant temperature and pressure, can be related to the Helmholtz free energy as:

$$G = A + PV = U + PV - TS \tag{3.8}$$

with P as the pressure of the system and the term $U + PV$ is the enthalpy, H. All thermodynamic quantities formulated in the canonical ensemble can be reformulated in the constant temperature and pressure ensemble, which is more compatible with experimental conditions and it could be used to compare directly computational results with experimental data[25].

Clearly, the computation of the configurational partition function, and subsequently, the absolute Helmholtz or Gibbs free energy of the system, is a challenge task because it requires the collection of statistics over all configurations of the systems, which is only accessible for systems in special conditions (e.g., molecules in gas phase treated in harmonic conditions) where the configurational integral has an analytical solution. The configurational partition function for most of the systems with chemical and biological relevance is not accessible in exact form as they count with an enormous number of degrees of freedom which difficult to get an analytical solution in their experimental conditions. Thus, The computation of A and G for these systems is only assessed numerically using a set of approximations to reduce the computational complexity[9,25].

To present the free energy problem in a mathematical context, let's express the Helmholtz free energy in terms of an ensemble average expression which can be numerically assessed by molecular simulations, in similar way as the internal energy was presented (Eq. 3.5). Inserting the equality:

$$1 = exp\left[\beta E\left(\mathbf{x}^N\right)\right] exp\left[-\beta E\left(\mathbf{x}^N\right)\right] \tag{3.9}$$

into the expression for the Helmholtz free energy (Eq. 3.7) gives:

$$A = kTln\int exp\left[\beta E\left(\mathbf{x}^N\right)\right] P\left(\mathbf{x}^N\right) d\mathbf{x}^N = kTln\left\langle exp\left[\beta E\left(\mathbf{x}^N\right)\right]\right\rangle \tag{3.10}$$

In general, the Eq. 3.10 gives a way to calculate the free energy in a single simulation of the system through the determination of an ensemble of configurations in consistency with the Boltzmann's probability and does an integration over all configurational space in the same spirit to obtain the average energy of the system. Nevertheless, $P\left(\mathbf{x}^N\right)$ is proportional to $exp\left[-\beta E\left(\mathbf{x}^N\right)\right]$ as indicated by Eq. 3.6 and because the fast increase of $exp\left[-\beta E\left(\mathbf{x}^N\right)\right]$ with the energy, there will be important configurations contributing to the integral with high energy. However, molecular simulations sample mainly regions of the configurational space with low energy and the barriers separating them from high-energy configurations are relatively high, which will require a much longer simulation time to collect all significant configurations for the ensemble average in the free energy expression.

Finally, for many relevant problems we are more interested on the difference in free energy, ΔA, for two well-defined states of the systems, than the absolute free energy for one state. For those states denoted as 0 and 1, ΔA is defined by ratio of the configurational partition function for the two states, Z_1 and Z_0, as

$$\Delta A = A_1 - A_0 = -kTln\frac{Z_1}{Z_0} \tag{3.11}$$

A direct approach to compute the free energy difference will require independent determinations of Z_1 and Z_0 through the energy functions E_0 and E_1. However, each computation of the partition function for both states will suffer of numerical difficulties as explained before. Another alternative for solving the problem is to connect both states through a coupling parameter, λ, which distinguishes both states.Thus, the potential energy, E, depends on this coupling parameter, and it smoothly passes from E_0 and E_1 when λ pass from 0 to 1, defining an equation for the Helmholtz free energy depending of λ.

$$A(\lambda) = -kTlnZ(\lambda) \tag{3.12}$$

Then, ΔA might be calculated through integration of the derivative of $A(\lambda)$ along λ as in the thermodynamic integration method. Also, it can be built a stratified path connecting the two end points of the coupling parameter and the computation of ΔA is done in a stepwise manner as in the case of the Free energy perturbation method.

Part II

RÉSULTATS ET APPLICATIONS

4

COMPUTATIONAL APPROACHES FOR ASSESSING PROTEIN-LIGAND BINDING

4.1 INTRODUCTION

Protein-ligand binding is an example of a relevant biomolecular recognition event to many biological processes that take place in living systems. Some ligands inhibit protein function, some others promote it by stabilizing a large isomerization of their receptor, thereby controlling key cell-signaling pathways.[26] Understanding how ligands bind to biomolecules is of fundamental importance not only for the basic fields of biophysics and biochemistry, but also for applied disciplines such as medicinal chemistry and pharmacology.[27] Although kinetics may strongly affect the yield of binding in cellular and other non-equilibrium environments, the primary factors that govern molecular recognition are of thermodynamic nature. In particular, the value of the binding equilibrium constant, K_{eq}, which is dictated by the standard free energy change on complexation, ΔG_b°, is the primary quantity of interest[28,29]. Being able to access the binding constant accurately and from first principles would provide a chemical understanding of protein-ligand recognition, thus unraveling guidelines for drug design.[29]

Over the last decades, innovation costs in the Pharma industry have exceedingly increased and have recently approached 4 billion U.S. dollars per FDA-approved drug.[30] Lead optimization alone is estimated to involve about 150 million U.S. dollars per hit compound.[30,31] Reliable predictions of the protein-ligand binding affinity by computation would greatly reduce these costs and boost a more efficient development of new pharmaceuticals. However, the calculation of the binding constant in protein-ligand per se, poses an outstanding theoretical and computational challenge. For instance, there exists no method to solve this problem when binding of the ligand involves a global structural change of the receptor, which is crucially important in ligand-modulated allosteric equilibria.[26] When the protein response is more local, the calculation of the standard free energy of binding is possible and several computational approaches at various levels of sophistication have been developed, which are currently in use at different stages of the drug discovery pipeline based on a trade-off between accuracy and efficiency.[32]

Among the available methods, the so-called *rigorous* approaches evaluate the free energy of binding based on simplified descriptions of the reaction path, which typically involves a series of non-physical intermediates. In the alchemical route, which was first introduced by Jorgensen[33] and later improved by others,[34–36] the ligand is decoupled reversibly from its environment with the free energy of binding accessed by perturbation theory. Alternatively, the ligand can be physically separated from the receptor by forcing the unbinding along a one-dimensional reaction coordinate, and the free energy of binding measured by umbrella sampling.[37–39] In both cases, to ensure configurational overlap between consecutive steps of the microscopic transformation, these approaches involve a large number of

intermediate states between the end-points, which results in a large computational effort per free energy determination; see Figure 4.1 (on top). Moreover, to improve the efficiency of sampling, the transformation is typically performed in the presence of appropriately chosen restraints whose contribution to the binding affinity must be evaluated by additional computation. A rule of thumb, rigorous free energy approaches may grant an approximate output rate of one determination per week, which is clearly not suited for screening purposes almost independently of the accuracy of the predictions.

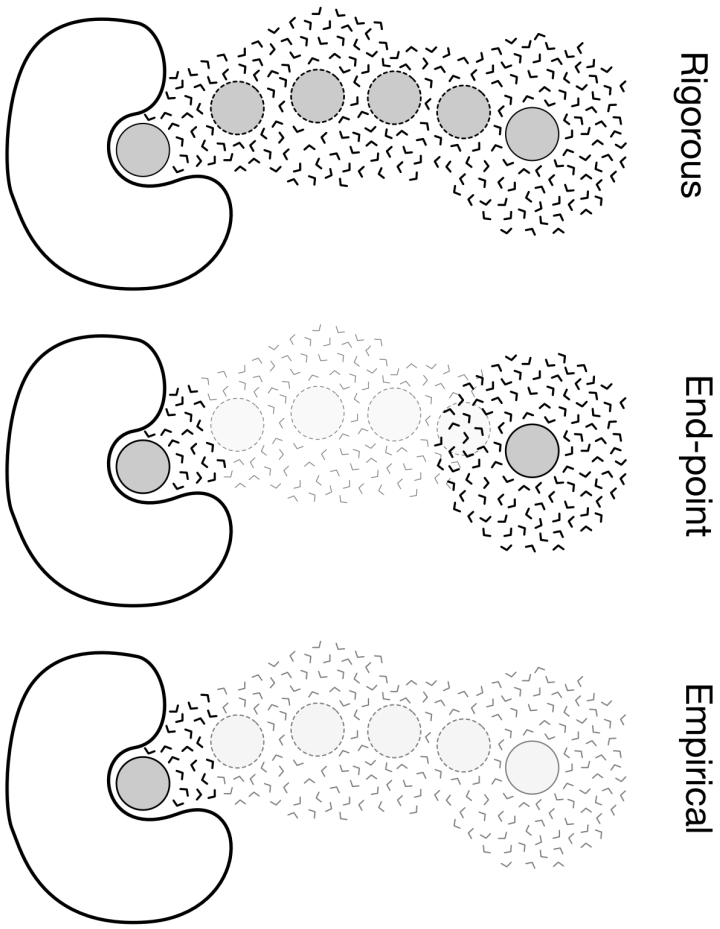


Figure 4.1: Different schemes to access the absolute free energy of binding in protein-ligand association.

To increase the efficiency of the calculation, more simplified computational approaches have been developed to focus only on the end-points of the binding reaction; see Figure 4.1 (middle). These methods reduce the computational burden by using approximated expressions for the solvation free energy, which can be efficiently evaluated by a continuum treatment of the solvent[40] (i.e. an implicit solvent model), or in the limit of the linear response approximation.[41] Prominent examples of *end-points* methods are MM/PBSA (molecular mechanics [MM] with Poisson-Boltzmann [PB] and surface area [SA]), which was originally developed by Kollman et al.,[42,43] and the linear interaction energy (LIE) approach of Åqvist.[44]

In MM/PBSA, the free energy of binding is estimated from the total change in the gas-phase internal energy, the solvation free energy and the configurational entropy upon protein-ligand association, with the solvation free energy accessed by solving the Poisson-Boltzmann equation plus a term accounting for the nonpolar contribution. In contrast, the LIE approach considers the solute/solvent interactions explicitly and estimates the binding free energy from changes in the electrostatic and van der Waals components of the ligand-surroundings interaction energy when the ligand is transferred from the solution bulk to the receptor binding site.[45] Both approaches rely on a detailed description of both the bound and the unbound states and explicitly include conformational effects by averaging over structural ensembles generated by e.g. Molecular Dynamics. These semi-rigorous approaches are clearly more efficient and grant an approximate output rate of one free-energy determination per day. For this reason, they have been fairly popular in the Pharma industry in both *hit-to-lead* and the *lead-optimization* phases.

Screening libraries or databases of small-molecule compounds require significantly more simplified schemes to allow for the efficient evaluation of thousands of millions of binding modes. Because of the strong computational restraints, *empirical* approaches based e.g. on molecular docking, which focus exclusively on the bound state (Figure 4.1, bottom), and generally neglecting the internal flexibility of the receptor,[46] have flourished. The introduction of rather drastic approximations results in output rates of one free energy determination per second,[47] which are suitable for the *hit identification* stage, although important aspects of the protein-ligand association reaction, such as translational and rotational entropy loss or desolvation of the ligand upon binding, are neglected.

Despite the existence of an inverse relationship between accuracy and required computer time per binding-affinity determination is well established,[45] less is known about the nature of the approximations that are introduced to reduce the computational effort and their actual impact on the accuracy of the predictions. In the following, we analyze some of the most popular approaches to the binding constant for protein-ligand and re-derive their fundamental equations in the common framework of statistical mechanics. In this chapter, we are able to pinpoint the approximations inherent to the various approaches (from most rigorous to most efficient), which represents a first step to devise variants with optimum accuracy/efficiency balance for each stage of the drug discovery pipeline. The final goal is to provide a self-consistent theoretical framework where apparently unrelated computational approaches to protein-ligand binding can be interpreted and compared.

4.2 STATISTICAL MECHANICS FRAMEWORK FOR PROTEIN-LIGAND BINDING

Let us consider the spontaneous association of a protein molecule (P) with a ligand (L) to form a non-covalent complex (PL) in aqueous solution

$$P + L \leftrightharpoons PL \tag{4.1}$$

At chemical equilibrium, the chemical potentials of the product and the reactants equalize so that

$$\Delta\mu_b = \mu_{PL} - (\mu_P + \mu_L) = 0 \tag{4.2}$$

By separating out the volume dependence of the chemical potentials in Eq. 4.2, which is customary done by introducing an arbitrary state of reference or *standard state*, as

$$\mu_i(V, T) = \mu_i^\circ(T) + kT \ln\left(\frac{C_i}{C^\circ}\right) \tag{4.3}$$

with T being the absolute temperature, C° the standard concentration, and μ_i° and C_i the standard chemical potential and the molar concentration of the i-th solute, respectively, and rearranging, it yields

$$\exp\left(-\frac{\Delta\mu_b^\circ}{kT}\right) = \frac{C_{PL}\,(C^\circ)}{C_P\,C_L} = K_{eq}C^\circ \tag{4.4}$$

which shows that the ratio between the equilibrium concentrations of the product over the reactants is volume independent (i.e. it is independent of the initial solute concentrations) and is therefore a chemical equilibrium constant. Importantly, the value of K_{eq}, or customarily its inverse K_d, which corresponds to the initial concentration of ligand for which the probability of binding at equilibrium is one-half (i.e. $C_P = C_{PL}$), sets an absolute scale of ligand-binding affinities. Hence, Eq. 4.4 quantifies the strength of the protein-ligand binding through the evaluation of $\Delta\mu_b^\circ$.

THE CANONICAL APPROACH. Straightforward access to the chemical potential difference in Eq. 4.4 is provided by a statistical mechanics treatment of the binding reaction in the canonical ensemble (N, V, T). In the limit of idealized solution behavior (i.e. the particle independent ansatz) and at constant temperature T and volume V, the chemical potential of the solute is

$$\mu_i(V, T) = -kT \ln \frac{q_i(V, T)}{N_i} \tag{4.5}$$

with q_i and N_i being the molecular partition function and the number of solute molecules. Introducing the rigid-rotor harmonic-oscillator (RRHO) approximation and in the limit of Born-Oppenheimer (BO) hypothesis, the molecular partition function in Eq. 4.5 can be further separated into translational, rotational, vibrational and electronic contributions, all having a closed form.[25] Also, by incorporating the net effect of the solvent in the electronic contribution through the evaluation of the potential mean force corresponding to the solvation free energy of the solute in its configuration at the minimum of the potential energy well ($\mathbf{X}_0$), Eq. 4.5 yields

$$\mu_i(V, T) = \mu_{i,v}(V, T) + W_{bulk}(\mathbf{X}_0) \tag{4.6}$$

which provides the chemical potential of the solute as a correction to the harmonic result in vacuum $(\mu_{i,v})$. By evaluating Eq. 4.6 for each component of the binding reaction at the standard $1M$ concentration (Figure 4.2) and introducing the results into Eq. 4.2, a rigorous estimate of $\Delta\mu_b^\circ$ is obtained, which provides numerical access to the binding constant in the limit of the RRHO approximation.

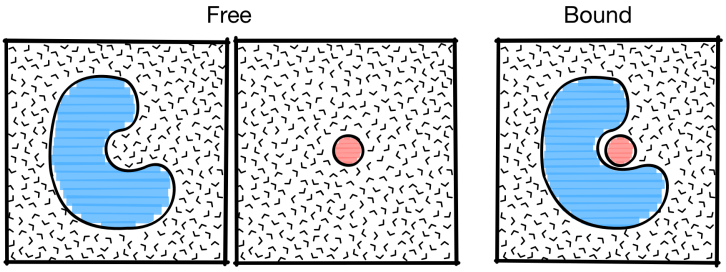


Figure 4.2: Schematic representation of the initial (free) and final (bound) states for the protein-ligand association used to compute $\Delta\mu_b^\circ$ in the canonical approach; see *Main Text*.

THE GRAND-CANONICAL APPROACH. A conceptually different approach to the binding constant in solution goes through a statistical mechanics treatment in the grand canonical ensemble (μ, V, T). This formalism[8] provides an alternative expression for the standard chemical potential of the solute

$$\mu_i^\circ(T) = -kT \ln\left(\frac{\mathcal{Q}_i(p,T)}{V}\right) \tag{4.7}$$

with $\mathcal{Q}_i$ being an effective partition function of the solute in the solvent at the constant pressure p. Introducing this result into Eq. 4.2 yields

$$\exp\left(-\frac{\Delta\mu_b^\circ}{kT}\right) = \frac{(\mathcal{Q}_{PL}/V_{PL})}{(\mathcal{Q}_P/V_P)\,(\mathcal{Q}_L/V_L)} = K_{eq} \tag{4.8}$$

which shows that the binding constant can be expressed by an effective partition function ratio. Because in the limit of infinite dilution each effective partition function can be approximated as

$$\mathcal{Q}_i(p,T) \approx \frac{Q_{N,1}}{Q_{N,0}} \tag{4.9}$$

with $Q_{N,1}$ and $Q_{N,0}$ being the canonical partition functions of a binary solution with N solvent molecules and one or no solute,[8] and the solution volume can be assumed as unchanged upon ligand binding (i.e. $V_P = V_{PL}$), Eq. 4.8 yields

$$K_{eq} = \frac{Q_{N,LP}/Q_{N,P}}{(Q_{N,L}/Q_{N,0})/V_L} \tag{4.10}$$

which shows that the value of the binding constant is related to the reversible work (or the free energy) to make the ligand disappear from the protein binding site (at the numerator) minus the work to make the ligand disappear from a box of solvent per unit of volume (at the denominator). Interestingly, Eq. 4.10 indicates that protein-ligand binding (Eq. 4.1) can be viewed as a transfer of ligand from the solution bulk to the binding site of the receptor (Figure 4.3), or as a partition equilibrium

$$L_{free} \leftrightharpoons L_{bound} \tag{4.11}$$

Finally, if the ligand is small relative to the receptor, the protein contributions in Eq. 4.10 almost cancel out and the numerator can be considered as an effective partition function of the ligand in the bound state, such that Eq.4.10 yields

$$K_{eq} = \frac{\mathcal{Q}_L(P)}{(\mathcal{Q}_L/V_L)} \tag{4.12}$$

with $\mathcal{Q}_L(P)$ being the volume-independent effective partition function of one ligand bound to the protein. As we shall see, this important result sets the ground to most statistical mechanics approaches to the binding constant.

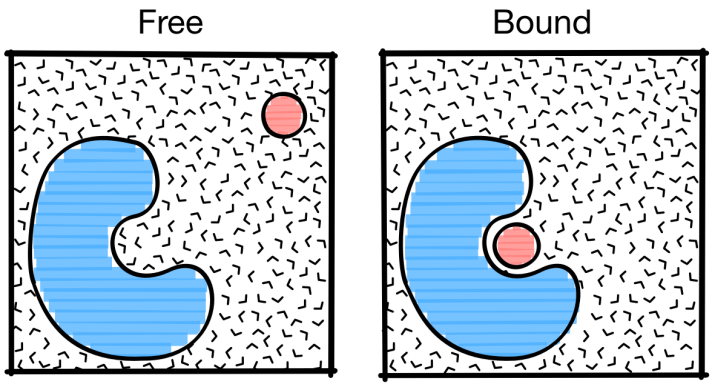


Figure 4.3: Schematic representation of the initial (free) and final (bound) states for the protein-ligand association used to compute $\Delta\mu^{\circ}_{b}$ in the grand canonical approach; see *Main Text*.

4.2.1 *Rigorous Statistical Mechanics approaches*

In the limit of highly dilute solutions, the results of Eq. 4.6 and Eq. 4.12 provide access to the binding equilibrium constant without too much approximation. In practice though, the use of Eq. 4.6 is strongly limited in protein-ligand problems by the evaluation of the solvation free energy of a large and flexible solute such as the protein alone or the complex, which is computationally challenging, and so far this approach has been successfully applied only to small peptide systems.[48] On the other hand, by factorizing out the kinetic energy contribution from the numerator and the denominator of Eq. 4.12, which cancel out as the total number of degrees of freedom is conserved, the binding constant can be expressed in terms of configurational integrals over the relevant portions of configurational space accessible to the ligand in the bound and the unbound states yielding

$$K_{eq} = \frac{\int_{site} d\mathbf{L} \int d\mathbf{X} \exp(-\beta U)}{\int_{bulk} d\mathbf{L}\, \delta(\mathbf{r}_L - \mathbf{r}^*) \int d\mathbf{X} \exp(-\beta U)} \tag{4.13}$$

with U being the total potential energy of the system, $\beta = 1/kT$, and $\mathbf{L}$ and $\mathbf{X}$ the coordinates of the ligand and the remaining (solvent and protein) atoms; note that the δ function at the denominator has been introduced to make the bulk configurational integral volume independent.[38] Importantly, Eq. 4.13 provides numerical access to the binding constant by computer

simulations and can be considered as the master equation for most *rigorous* approaches to protein-ligand binding. Two of them are briefly reviewed below.

4.2.1.1 *Alchemical free energy perturbation*

An effective strategy to compute protein-ligand binding free energies based on Eq. 4.13 was originally introduced by Jorgensen[33] and later improved by Gilson.[36] This approach, which is usually referred to as "double annihilation" or "double decoupling", solves Eq. 4.13 by making use of a thermodynamic cycle in which the ligand is transformed into a fictitious non-interacting body both in the bound and the unbound states. Such an *alchemical* transformation, termed annihilation, is achieved through the use a hybrid Hamiltonian with a coupling parameter λ of the form $U(\lambda) = (1-\lambda)U_1 + \lambda U_0$, with U_0 and U_1 being the total potential energy of the system with a non-interacting (decoupled) and a full-interacting (coupled) ligand.[49] By introducing an intermediate state in which the ligand is transferred to the gas phase, Eq. 4.13 yields

$$K_{eq} = \frac{\int_{site} d\mathbf{L} \int d\mathbf{X} \exp(-\beta U_1)}{\int_{bulk} d\mathbf{L}\, \delta(\mathbf{r}_L - \mathbf{r}^*) \int d\mathbf{X} \exp(-\beta U_0)} \times \frac{\int_{bulk} d\mathbf{L}\, \delta(\mathbf{r}_L - \mathbf{r}^*) \int d\mathbf{X} \exp(-\beta U_0)}{\int_{bulk} d\mathbf{L} \int d\mathbf{X} \exp(-\beta U_1)} \quad (4.14)$$

where the first term of the r.h.s is related to the reversible work for decoupling the ligand in the bound state and the second to the work for decoupling it in the solution bulk, which are both numerically accessible by multi-stage free energy perturbation molecular dynamics (FEP/MD) simulations.[50] Despite the elegance of the strategy, straightforward applications of Eq. 4.13 are often impractical as e.g. the ligand in the highly decoupled states becomes free to "wander" in the volume of the simulation box, which seriously hinders statistical convergence.[36] To improve sampling efficiency, more recent implementations of double decoupling make use of external restraints which reduce the configurational space accessible to the ligand. The sequential activation/deactivation of restraints on the position, the orientation, and the internal configuration of the ligand significantly improves statistical convergence but introduces a series of additional intermediates along the reaction path, which involve extra computation. Following one of the most recent implementations,[51] which denotes the har-

monic restraints on the translation, rotation, and the conformation of the ligand as u_t, u_o and u_c, Eq. 4.13 becomes

$$
\begin{aligned}
K_{eq} = & \frac{\int_{site} d\mathbf{L} \int d\mathbf{X} \exp\left(-\beta U_1\right)}{\int_{site} d\mathbf{L} \int d\mathbf{X} \exp\left[-\beta\left(U_1 + u_c\right)\right]} \times \\
& \frac{\int_{site} d\mathbf{L} \int d\mathbf{X} \exp\left[-\beta\left(U_1 + u_c\right)\right]}{\int_{site} d\mathbf{L} \int d\mathbf{X} \exp\left[-\beta\left(U_1 + u_c + u_o\right)\right]} \times \\
& \frac{\int_{site} d\mathbf{L} \int d\mathbf{X} \exp\left[-\beta\left(U_1 + u_c + u_o\right)\right]}{\int_{site} d\mathbf{L} \int d\mathbf{X} \exp\left[-\beta\left(U_1 + u_c + u_o + u_p\right)\right]} \times \\
& \frac{\int_{site} d\mathbf{L} \int d\mathbf{X} \exp\left[-\beta\left(U_1 + u_c + u_o + u_p\right)\right]}{\int_{bulk} d\mathbf{L}\, \delta\left(\mathbf{r}_L - \mathbf{r}^*\right) \int d\mathbf{X} \exp\left[-\beta\left(U_0 + u_c + u_o + u_p\right)\right]} \times \\
& \frac{\int_{bulk} d\mathbf{L}\, \delta\left(\mathbf{r}_L - \mathbf{r}^*\right) \int d\mathbf{X} \exp\left[-\beta\left(U_0 + u_c + u_o + u_p\right)\right]}{\int_{bulk} d\mathbf{L}\, \delta\left(\mathbf{r}_L - \mathbf{r}^*\right) \int d\mathbf{X} \exp\left[-\beta\left(U_0 + u_c + u_o\right)\right]} \times \\
& \frac{\int_{bulk} d\mathbf{L}\, \delta\left(\mathbf{r}_L - \mathbf{r}^*\right) \int d\mathbf{X} \exp\left[-\beta\left(U_0 + u_c + u_o\right)\right]}{\int_{bulk} d\mathbf{L}\, \delta\left(\mathbf{r}_L - \mathbf{r}^*\right) \int d\mathbf{X} \exp\left[-\beta\left(U_0 + u_c\right)\right]} \times \\
& \frac{\int_{bulk} d\mathbf{L}\, \delta\left(\mathbf{r}_L - \mathbf{r}^*\right) \int d\mathbf{X} \exp\left[-\beta\left(U_0 + u_c\right)\right]]}{\int_{bulk} d\mathbf{L}\, \delta\left(\mathbf{r}_L - \mathbf{r}^*\right) \int d\mathbf{X} \exp\left[-\beta\left(U_1 + u_c\right)\right]} \times \\
& \frac{\int_{bulk} d\mathbf{L}\, \delta\left(\mathbf{r}_L - \mathbf{r}^*\right) \int d\mathbf{X} \exp\left[-\beta\left(U_1 + u_c\right)\right]}{\int_{bulk} d\mathbf{L}\, \delta\left(\mathbf{r}_L - \mathbf{r}^*\right) \int d\mathbf{X} \exp\left[-\beta\left(U_1\right)\right]}
\end{aligned}
\tag{4.15}
$$

Eq. 4.15 shows that numerical access to the binding constant is provided by sequential confinement of the conformation, orientation and position of the ligand in the binding site; decoupling of the ligand from the protein environment; release of the positional and orientational restraints in the gas phase; re-coupling of the ligand in the bulk; and release of the conformational restraint in solution. In this approach, the standard free energy of binding is determined by summing up the reversible work associated with each of the eight steps of a complex microscopic transformation.

4.2.1.2 *Potential of mean force*

An alternative approach, which is also based on Eq. 4.13 consists on measuring the free-energy of binding/unbinding along a highly simplified representation of the reaction path by a potential of mean force (PMF) calculation. In this case, the ligand is physically separated from the receptor and the free energy of binding is obtained by umbrella sampling over one or more geometric reaction coordinates, typically the Euclidean distance between its initial position in the binding site and an arbitrary point in the bulk.[37] Because of the high-dimensionality of the *true* reaction coordinate for binding, the conformational freedom of the ligand that may differ substantially in the bound and unbound states, and the intrinsic difficulty of sampling the translational degrees of freedom in the unbound state, this approach is used in combination with restraints.[38,52] In analogy to the alchemical route, a series of intermediates corresponding to both orientational and conformational confinement/release of the

ligand in the bound and the unbound states are introduced, which provides the following expression for the binding constant

$$
\begin{aligned}
K_{eq} = & \frac{\int_{site} d\mathbf{L} \int d\mathbf{X} \exp\left(-\beta U\right)}{\int_{site} d\mathbf{L} \int d\mathbf{X} \exp\left[-\beta\left(U+u_c\right)\right]} \times \\
& \frac{\int_{site} d\mathbf{L} \int d\mathbf{X} \exp\left[-\beta\left(U+u_c\right)\right]}{\int_{site} d\mathbf{L} \int d\mathbf{X} \exp\left[-\beta\left(U+u_c+u_o\right)\right]} \times \\
& \frac{\int_{site} d\mathbf{L} \int d\mathbf{X} \exp\left[-\beta\left(U+u_c+u_o\right)\right]}{\int_{site} d\mathbf{L} \int d\mathbf{X} \exp\left[-\beta\left(U+u_c+u_o+u_a\right)\right]} \times \\
& \frac{\int_{site} d\mathbf{L} \int d\mathbf{X} \exp\left[-\beta\left(U+u_c+u_o+u_a\right)\right]}{\int_{bulk} d\mathbf{L}\, \delta\left(\mathbf{r}_L-\mathbf{r}^*\right) \int d\mathbf{X} \exp\left[-\beta\left(U+u_c+u_o\right)\right]} \times \\
& \frac{\int_{bulk} d\mathbf{L}\, \delta\left(\mathbf{r}_L-\mathbf{r}^*\right) \int d\mathbf{X} \exp\left[-\beta\left(U+u_c+u_o\right)\right]}{\int_{bulk} d\mathbf{L}\, \delta\left(\mathbf{r}_L-\mathbf{r}^*\right) \int d\mathbf{X} \exp\left[-\beta\left(U+u_c\right)\right]} \times \\
& \frac{\int_{bulk} d\mathbf{L}\, \delta\left(\mathbf{r}_L-\mathbf{r}^*\right) \int d\mathbf{X} \exp\left[-\beta\left(U+u_c\right)\right]}{\int_{bulk} d\mathbf{L}\, \delta\left(\mathbf{r}_L-\mathbf{r}^*\right) \int d\mathbf{X} \exp\left(-\beta\left(U\right)\right)}
\end{aligned}
\tag{4.16}
$$

Compared to Eq.4.15, the fourth term in the r.h.s. of the equation above is unique to the PMF approach and corresponds to the physical separation of the ligand from the protein in the presence of configurational and orientational restraints.

In the absence of large conformational changes of the protein, Eqs. 4.15 and 4.16 provide rigorous, first-principle access to the binding constant and have been successfully applied in a number of cases.[51,53,54] However, these approaches require thorough statistical sampling over a large number of intermediates and are computationally very demanding. Thus, despite their accuracy, they are not suited to handle large databases of ligands and have found, so far, little room in the Pharma industry, although this trend is likely to change in the near future.[55]

4.2.2 *Simplified end-points approaches*

To facilitate the computational task without compromising too much the quality of the results, simplified approaches based on Eq. 4.6 and Eq. 4.12 have been also developed. A prominent group of them explicitly consider the conformational dynamics of the protein-ligand complex and the ligand alone in solution and accesses the binding constant by focusing on the relevant initial and final states (end points) of the reaction (Figure 4.1). In this section, two amongst the most popular *end-points* strategies for protein-ligand binding, i.e. the MM/PBSA and the linear interaction energy (LIE) methods, are shortly reviewed. Based on the statistical mechanics framework introduced above, their fundamental equations are re-derived with emphasis on the approximations introduced.

4.2.2.1 *The MM/PBSA method*

The MM/PBSA approach aims at the binding constant from the numerical evaluation of the absolute chemical potentials of the ligand, the protein and the complex in isolation (Fig-

ure 4.2). Starting from Eq. 4.6 and separating out the enthalpy versus entropy contributions to the chemical potential in vacuum, it yields

$$\mu_i(V,T) = (3n-3)\,kT + U(\mathbf{X}_0) - TS_i(V) + W_{bulk}(\mathbf{X}_0) \tag{4.17}$$

with S_i being the configurational entropy *in vacuum* that includes contributions from the translational, rotational and vibrational degrees of freedom and W_{bulk} the solvation free energy of the solute, which accounts for all enthalpic and entropic contributions of the solvent. Assuming that W_{bulk} can be accessed by a continuum model of water such PB/SA, which evaluates the polar contribution by solving the Poisson-Boltzmann equation and the non-polar contribution as a linear function of the solvent accessible surface area (SASA), Eq. 4.17 yields

$$\mu_i(V,T) = (3n-3)\,kT + U(\mathbf{X}_0) + G_{PBSA}(\mathbf{X}_0) - TS_i(V) \tag{4.18}$$

where the configurational entropy can be evaluated in the harmonic limit by classical statistical mechanics in combination with normal-mode analysis[56] including quantum corrections.[57] To account for part of the anharmonicity, which can be significant in proteins,[58] the electronic energy of the solute as well as its solvation free energy are often accessed by ensemble averages based on sampling from room-temperature MD in a box of explicit water; note that in this case the electronic energy contribution is obtained by subtracting the temperature-dependent vibrational energy from the average potential energy of the solute in vacuum, as $-D_e = \langle U\rangle - \sum^{\kappa} 1/2kT$, with κ being the total number of internal degree of freedom of the solute. Introducing this result into Eq. 4.18, it yields

$$\mu_i(V,T) = \frac{3}{2}n_i kT + \langle U\rangle + \langle G_{PBSA}\rangle - TS_i(V) \tag{4.19}$$

with the configurational entropy corrected for anharmonicity by a quasi-harmonic vibrational analysis.[59] Using the notation in the original paper by Kollman,[42] i.e. decomposing the force field energy into bonded (E_{bond}), electrostatic (E_{elec}), and van der Waals (E_{vdW}) contributions and splitting the solvation free energy into polar (G_{pol}) and nonpolar (G_{np}) terms, Eq. 4.19 gives

$$\mu_i(V,T) = \frac{3}{2}n_i kT + \overline{E}_{bond} + \overline{E}_{elec} + \overline{E}_{vdW} + \overline{G}_{pol} + \overline{G}_{np} - TS_i(V) \tag{4.20}$$

Eq. 4.20 is the master equation for the MM/PBSA approach; note that the first term on the r.h.s. corresponds to the total kinetic energy of the solute, which was missing in the original MM/PBSA formulation as pointed out by Gohlke and Case.[60] Evaluating Eq. 4.20 at the standard 1M concentration for the ligand, the protein and the complex separately (Figure 2), the standard chemical potential difference and therefore the binding constant are straightforwardly accessed.

In practice, to solve Eq. 4.20 explicit-solvent MD simulations of the ligand, the protein and the complex are carried out and ensemble averages of the force field energy of the three solutes in the gas phase and the configurational-dependent solvation free energy are evaluated as arithmetic averages over a large series of MD snapshots. Alternatively, configurational

ensembles for the uncomplexed reactants can be generated from the trajectory of the complex only, by removing selectively the atoms of the protein or the ligand. Surprisingly, this simplified version of MM/PBSA, termed the one-average variant, was shown to yield more accurate results than the original strategy,[61] perhaps due to the exact cancellation of the bonded energy in Eq. 4.20. Great efforts to improve both the accuracy and the efficiency of MM/PBSA have been done in recent years by benchmarking and optimizing the calculation of the individual contributions in Eq. 4.20.[62–64] To this aim, various continuum-electrostatics models including e.g. the generalized Born (GB) model in the MM/GBSA variant[40] have been tested and benchmarked searching for improved estimates of the polar solvation term (G_{pol}).[65] Similarly, alternative charge schemes for computing the electrostatic contribution (E_{elec})[66] or different approaches to evaluate the configurational entropy of the solute e.g. by quasi-harmonic analyses[60,67,68] have been reported. In general, the accuracy of the predictions is found to be fairly system-dependent, which makes it difficult to draw conclusions on the performance of the individual variants. Finally, by replacing the MM terms by a quantum mechanical model, a significantly improved correlation with experimental binding affinity results was obtained.[69] In this case, a hybrid QM/MM approach with the atoms of the ligand assigned to the QM region and the rest treated as MM, was used to sample the configurational ensembles for evaluation of Eq. 4.20, which significantly deteriorated the performance of the calculation.

4.2.2.2 *The Linear Interaction Energy (LIE) method*

An alternative simplified approach to the binding constant, which treats the protein-ligand binding reaction as a partition equilibrium (Eq. 4.11), can be derived from Eq. 4.12. Introducing the RRHO approximation and incorporating the partition function contributions from the protein and the solvent in the electronic energy of the ligand in the bound and the unbound states, Eq. 4.12 yields

$$K_{eq} = \frac{q_{CM} q_{rock}\, e^{-\beta W_{site}}}{q_{tr}/V\, q_{rot}\, e^{-\beta W_{bulk}}} \tag{4.21}$$

with q_{CM} and q_{rock} corresponding to oscillations and rocking of the ligand in the bound state, q_{tr}/V and q_{rot} to translations (per unit of volume) and rotations of the free ligand in solution, and W_{site} and W_{bulk} to the reversible work for transferring the ligand from the gas phase to the protein binding site and the solution bulk, respectively. Note that Eq. 4.21 includes no contribution from the internal vibrations or the potential energy of the ligand *in vacuum*, which effectively cancel out in the limit of rigid ligands. By extracting the logarithm and multiplying by $-kT$, Eq. 4.21 yields

$$\Delta\mu_b^\circ = -kT \log \zeta + W_{site} - W_{bulk} \tag{4.22}$$

with ζ corresponding to the fraction of translational and rotational motion left to the ligand in the bound state relative to its free rotation in the unbound state in solution. Importantly, Eq. 4.22 indicates that the standard free energy of binding can be accessed from the difference in the "solvation" free energy of the ligand in the protein and the solvent environments plus a

contribution corresponding to the entropic confinement. Provided that accurate estimates of W_{site} and W_{bulk} can be determined e.g. by FEP, Eq. 4.21 is essentially equivalent to Eq. 4.13 and provides a quantitative estimate of the binding constant in the harmonic limit. However, this result would come with no computational advantage. To simplify the calculation of the binding constant, the idea developed by the linear interaction energy (LIE) approach, which follows from linear response theory,[70,71] is that both polar and non-polar contributions to the solvation free energies in Eq. 4.22 can be approximated by linear functions of the mean electrostatic and van der Waals interaction energies of the ligand with the surroundings. In fact, in the limit of a linear response of the solution to changes in the local electric field, e.g. the appearance of a charged solute, it can be formally shown[44] that the polar contribution to the solvation free energy equals one half of the mean solute-solvent electrostatic contribution as

$$W^{pol} = \frac{1}{2}\left\langle U_{l/s}^{elec}\right\rangle \tag{4.23}$$

where the brackets $\langle\rangle$ indicate a thermodynamic average of the ligand-surroundings (l/s) interaction energy. Furthermore, based on the observations that the experimental free energy of solvation for various hydrocarbons in water was approximately linear with the length of the carbon chain,[72] and that the corresponding solute-solvent van der Waals energies from computer simulations were also linear with the number of carbon atoms, Åqvist *et al* assumed that the non-polar contribution to the solvation free energy could be approximated as a linear function of the mean van der Waals interaction energy[44] as

$$W^{np} \approx \alpha\left\langle U_{l/s}^{vdW}\right\rangle \tag{4.24}$$

with α being an adjustable parameter subject to empirical calibration. If so, by expressing the solvation free energy of the ligand in the bound and the unbound states as a sum of polar (Eq. 4.23) and non-polar (Eq.4.24) contributions, Eq. 4.22 yields

$$\Delta\mu_b^\circ = \frac{1}{2}\left[\langle U_{l/s}^{elec}\rangle_{site} - \langle U_{l/s}^{elec}\rangle_{bulk}\right] + \alpha\left[\langle U_{l/s}^{vdw}\rangle_{site} - \langle U_{l/s}^{vdw}\rangle_{bulk}\right] + \gamma \tag{4.25}$$

which provides the master equation for the LIE approach.[44] In this expression, γ includes the entropic confinement contribution, which can be evaluated numerically in the limit of the RRHO model (Eq.4.21) or determined empirically by fitting on experimental binding data. To account for (minor) deviations of the polar term from the exact (linear response) scaling factor of 1/2, a more general expression of LIE may include an additional fitting parameter β, which was shown to improve the accuracy of the computational predictions.[73] Note that unlike the original derivation of LIE,[44] Eq. 4.25 was obtained in the limit of the harmonic approximation, which would restrain the validity of the approach to rigid ligands. Although this assumption is not strictly required in LIE, the treatment above provides an explicit expression for the entropic confinement contribution, which would be otherwise hidden in the empirical coefficient of the non-polar term.

In the limit of the linear response theory, Eq. 4.25 solves the protein-ligand problem by measuring the mean ligand/surroundings interaction energy in the bound and the unbound states on a series of snapshots extracted from room-temperature MD simulations of the fully

solvated complex and the free ligand in solution. In this approach, the strongest approximations regard both the validity of the linear response assumption and the rather simplistic idea that the non-polar contribution to the binding free energy, which includes hydrophobic effects and both repulsive and dispersive solute/solvent interactions, can be extracted from the analysis of the non-electrostatic component of the ligand/surroundings interactions. Significant effort was made to validate the former hypothesis e.g. by comparing LIE results with rigorous FEP calculations, and the linear response approximation was found to be accurate for both the solvent and protein environments.[74,75] More difficult is the validation of the second assumption, which involves the determination of the parameter α, and possibly γ when absolute binding free energies are of interest. Because the physical nature of these parameters is unclear and their value is force-field, ligand and even protein dependent, their existence introduces a significant degree of empiricism, which has hindered the development of a "universal" and fully transferable LIE parameterization. Nonetheless, the fact that the intermolecular energies from simulations of the end-points are sufficient to predict absolute binding free energies with an accuracy of < 1 kcal/mol from experiments is absolutely remarkable and justifies the use of LIE in computer-aided drug discovery.

From a practical viewpoint, the implementation of LIE requires extensive configurational sampling of both the complex and the free ligand in solution typically by Molecular Dynamics or Monte Carlo simulations with an explicit treatment of the solvent, which makes this approach not suitable for high-throughput screening. To increase the computational performance, Huang and Caflisch developed the Linear Interaction Energy with Continuum Electrostatics (LIECE), where the MD sampling is replaced by energy minimization plus finite-difference Poisson calculations for a rigorous treatment of both the protein and the ligand desolvation energies.[76] When applied to β-secretase (BACE) and HIV-1 protease, a two-parameter LIECE model was shown to reach a predictive power of of < 1 kcal/mol relative to experiments, while being about two orders of magnitude faster than previous LIE. As such, the LIECE method can be effectively used to screen large libraries of compounds docked by automatic computational tools and has been successfully applied in virtual screening campaigns against important drug targets,[77] although the parameters developed for one target are generally not transferable. The strong dependence of the binding affinities on the electrostatic component of the protein-ligand interaction energy motivated the development of a LIECE variant (QMLIECE) in which the ligand-surrounding interactions are evaluated by a semiempirical quantum mechanical model to include e.g. polarization effects.[78] Interestingly, the QM variant was shown to be superior when dealing with formally charged compounds as the peptidic inhibitors of the West Nile serine protease.[79]

4.2.3 *Empirical approaches*

Although the endpoint approaches significantly reduce the computing time, they are still too expensive to estimate binding affinities for a large library or databases of compounds. Therefore, even more approximated approaches are required for structure-based virtual high-throughput screening, where hundreds of thousands of compounds for a specific target must

be evaluated and ranked. To speed up the calculations, the so-called *empirical* approaches simplify the description of the binding reaction one step further and focus exclusively on the bound state (Figure 4.1, bottom). Among them, molecular docking is with no doubt the most popular approach.[80] In this case, the binding affinity is usually estimated in the context of a rigid conformation of the receptor with no sampling procedure to account for its dynamics. The flexibility of the ligand is included through a systematic or stochastic search typically Monte Carlo or a genetic algorithm, and its fitness is quantified by a crude scoring function, which is used both for ranking the binding modes and prioritizing compounds extracted from the library. In most docking protocols, the atomistic details of the protein are replaced by a grid representation centered on the binding site, where each point stores the interaction energy of an atomic "probe" with the rest of the receptor. Also, solvent effects are usually neglected or efficiently accounted for by continuum solvation models. In general, three main classes of scoring functions are used in protein-ligand binding: empirical, force field-based and knowledge-based.[46,81] In the following, we will focus on the first two, which despite their crude nature can still be interpreted in the framework of statistical mechanics.

4.2.3.1 *Empirical scoring function*

A first approach to efficiently score a large number of docking poses is based on the evaluation of the binding free energy by a weighted sum of empirical descriptors as

$$\Delta\mu_b^\circ = \sum_i W_i \Delta\mu_i \tag{4.26}$$

with $\Delta\mu_i$ corresponding to independent contributions selected based on chemical intuition, e.g. the electrostatic and van der Waals components of the protein-ligand interaction energy, the number of H-bonding donors and acceptors, the number of rotable bonds, etc., each accounting for a critical interaction in protein-ligand binding. The weight of the contributions, i.e. the coefficients in Eq. 4.26 (W_i), are empirically determined by fitting on a training set of experimental binding affinities typically using multivariate linear regressions. The simplicity and flexibility of Eq. 4.26 grants for the required computational efficiency at the hit-identification stage. However, the accuracy of the predictions is strongly dependent of the quality of training set (both in size and composition), which makes the transferability of the model rather challenging. One of the earliest model for docking was introduced by Böhm[82]

$$\begin{aligned}\Delta\mu_b^\circ =& \Delta G_0 + \Delta G_{hb} \sum_{hbonds} f(\Delta R, \Delta\alpha) \; + \\ & \Delta G_{io} \sum_{io\ int.} f(\Delta R, \Delta\alpha) \; + \\ & \Delta G_{lipo} \sum_{lipo\ cont.} A_{lipo} \; + \\ & \Delta G_{aro} \sum_{aro\ int.} f(\Delta R) + \Delta G_{rot} \times N_{rot}\end{aligned} \tag{4.27}$$

In this equation, the polar contribution is accounted for by explicit H-bonding (ΔG_{hb}) and ionic interaction (ΔG_{io}) terms, which are both distant and angle dependent to penalize deviations from optimum geometries. The apolar contribution is accounted for by lipophilic

contacts (ΔG_{lipo}) and aromatic interactions (ΔG_{aro}), with the former being dependent on the contact surface $\left(A_{lipo}\right)$ and the latter on the distance. Finally, the flexibility of the ligand is indirectly included by counting the number of rotable bonds (N_{rot}), which approximates the entropy cost to confine the ligand in the bound state.

Another example of fast and simple empirical scoring is provided by Fresno, a model introduced by Rognan *et al*[83]

$$\Delta\mu_b^\circ = K + \alpha\,(HB) + \beta\,(LIPO) + \gamma\,(ROT) + \delta\,(BP) + \epsilon\,(DESOLV) \tag{4.28}$$

This empirical scoring function includes five terms corresponding to hydrogen bonds, lipophilic, rotational, buried-polar and ligand desolvation contributions. The first three are common to the Böhm model and are similarly evaluated. The buried-polar term (BP) accounts for unfavorable interactions or "bumps", which result from contacts between polar and nonpolar groups in the binding site. The last term accounts for the desolvation free energy of the ligand in the unbound state $(DESOLV)$, which is efficiently evaluated by a Poisson-Boltzmann calculation. Despite the empirical nature of Eq. 4.28, this scoring function can be connected with the theoretical framework above, particularly with the LIE approach. In fact, the hydrogen bond, lipophilic and buried-polar terms are clearly related to the electrostatic and van der Waals interaction energies of the ligand in the bound state, whereas rotational and desolvation contributions correspond to contributions related to the unbound state of the ligand, which account for the rotational entropy loss and ligand desolvation upon biding. However, the arbitrary selection of the energy components in such empirical schemes makes it difficult to rationalize the connection with first-principle statistical mechanics.

4.2.3.2 *Force field scoring function*

Another class of scoring functions aim at the evaluation of the binding affinity through the quantification of the physical protein-ligand interactions as quantified by molecular mechanics, where the nonbonding energy is typically evaluated as a sum of pairwise electrostatic and van der Waals atomic contributions. This two ingredients provide the minimalistic score used by the program Dock (v. 4.0)[84]

$$\Delta\mu_b^\circ = \sum_i^{prot}\sum_j^{lig}\left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} + 332.0\frac{q_i q_j}{\epsilon\left(r_{ij}\right) r_{ij}}\right) \tag{4.29}$$

with r_{ij} being the distance between the j-th atom of the ligand and the i-th atom of the protein, A_{ij} and B_{ij} the atomic van der Waals parameters, and q_j and q_i their partial charges. These empirical parameters are usually obtained from popular biomolecular force fields such as AMBER or CHARMM. Calculations of the nonbonding interactions in the bound state are typically performed *in vacuum*, sometimes using a distance-dependent dielectric constant $\epsilon\left(r_{ij}\right)$ to account for solvation effects. Alternatively, more realistic descriptions of the solvent are achieved through the use of more rigorous implicit solvent models as PB/SA or GB/SA, which require extra computation. Overall, the great advantage of these methods is that the binding affinity can be evaluated, in principle, for any non-covalent protein-ligand association as force-field parameters are developed to be transferable. Interestingly, Eq. 4.29 can be

derived from the fundamental equation of LIE (Eq. 4.25) introducing two additional assumptions: i. the binding affinity can be accessed from a single structure of the protein-ligand complex, which turns the ensemble averages in Eq. 4.25 into single-point energy evaluation; and ii. the desolvation of the ligand upon binding is negligible, i.e. the ligand/surrounding interaction energy in the unbound state can be set to zero. Thus, Eq. 4.25 yields

$$\Delta\mu_b^\circ = \left[\left(U_{l/s}^{elec}\right)_{site} - 0\right] + \left[\left(U_{l/s}^{vdw}\right)_{site} - 0\right] = \sum_i^{prot}\sum_j^{lig}\left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} + \frac{q_i q_j}{\epsilon\left(r_{ij}\right) r_{ij}}\right) \tag{4.30}$$

Finally, to effectively account for contributions of the unbound state, i.e. the strain energy of the ligand or the total entropy change upon binding, hybrid implementations of Eq. 4.29 have been developed. A prototypical example is implemented in AutoDock (v. 4.2)[85]

$$\begin{aligned}\Delta\mu_b^\circ =& W_{vdw}\sum_i^{prot}\sum_j^{lig}\left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}}\right) + \\ & W_{ele}\sum_i^{prot}\sum_j^{lig}\left(\frac{q_i q_j}{\epsilon\left(r_{ij}\right) r_{ij}}\right) + \\ & W_{hbond}\sum_i^{prot}\sum_j^{lig}\left(E\left(t\right)\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}}\right) + \\ & W_{sol}\sum_i^{prot}\sum_j^{lig}\left(S_i V_j + S_j V_i\right)\exp\left(\frac{-r_{ij}^2}{2\sigma^2}\right)\end{aligned} \tag{4.31}$$

where W_i are weighting constants calibrated on experimental binding data. In this case, the van der Waals and the electrostatic contributions, which are typical of a force field, are complemented by a specialized 10/12 Lennard-Jones potential for directional hydrogen bonding and a desolvation term that is related to the excluded volume for the ligand in the bound state. In general, the goal of a hybrid scheme is to introduce critical free energy contributions to binding (e.g. ligand desolvation) which are missing in the force field representation without compromising the computational efficiency.

4.3 DISCUSSION AND CONCLUSIONS

The calculation of the protein-ligand binding affinity is a fundamental problem that poses an outstanding theoretical and computational challenge. To this aim, several computational approaches have been developed over years, which tackle the problem at various degrees of approximation.

The statistical mechanics interpretation presented here that there are two general approaches to the standard free energy of binding. One approach goes through the (direct) evaluation of the absolute chemical potentials for all components of the binding reaction (i.e. the ligand, the protein and the complex). The other one treats the binding reaction as a partition equilibrium of the ligand between the bound and the unbound states, which assumes that most of the

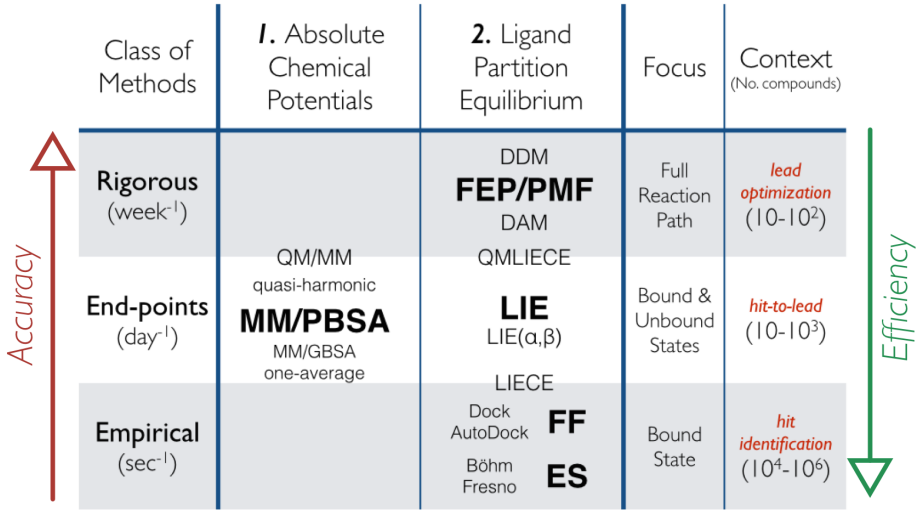


Figure 4.4: Classification of methods for the calculation of the protein-ligand binding affinity based on a unified statistical mechanics framework.

protein contributions to the chemical potential difference effectively cancel out. To the best of our knowledge, all rigorous approaches to the binding constant, such as FEP (Eq. 4.15) or PMF (Eq. 4.16) fall in the second class; see Figure 4.4. Surprisingly and perhaps due to the intrinsic challenge posed by the accurate evaluation of the solvation free energy of the protein with and without the ligand, there exists no rigorous approach belonging to the former class. Because the evaluation of the free energy of binding involves the detailed analysis of a large number of intermediate states along the reaction path, these methods are computationally very intensive and may be useful to rank only a small number of compounds, typically less than a hundred, at the *lead-optimization* stage.

The situation is different for the semi-rigorous or *end-points* approaches where MM/PBSA (Eq. 4.20) belongs to the first class, whereas LIE (Eq. 4.25) to the second; see Figure 4.4. In both cases, the binding constant is accessed by solving a thermodynamic cycle that involves molecular transfer to the gas phase. This strategy effectively transforms the calculation of the standard free energy of binding into a difference between (approximate) solvation free energies, which can be evaluated with much less computation. By replacing the explicit representation of the binding path with approximate solvation free energy estimates based on continuum models or the linear response theory, these methods alleviate the computational burden quite significantly, extending their scope to the *hit-to-lead* stage where thousands of compounds must be evaluated and ranked. Of course, the quality of the predictions critically depends on the accuracy of the solvation free energy calculations, which motivates further effort on the development of more accurate implicit solvent models. Also, the striking similarity with the strategy implemented in the (rigorous) alchemical route suggests that the use of restraints to control the configurational freedom of the ligand, particularly in the unbound state, could be beneficial to accelerate numerical convergence.

Analysis of the fast *empirical* approaches to protein-ligand (including some of the most popular scoring functions for docking) shows that these methods break down the computational cost by focusing exclusively on the bound state, i.e. forgetting about the protein or the ligand in solution, and therefore belong to the second class; see Figure 4.4. Because the average output rate is of one free-energy determination per second, these simplified approaches are suitable for screening millions of compounds and find widespread use at the *hit identification* stage. Nonetheless, the significant speed-up is achieved by introducing a series of theoretically unjustified approximations, which result in sizeable systematic errors that make the predictions often unreliable and/or highly system-dependent. Comparison of the force-field (FF) scoring functions with less-approximated approaches in the same class, i.e. LIE, demonstrates that the strongest approximations in the former are related to both the neglecting of entropic effects, which results from a rigid-body treatment of the receptor, and the deliberate exclusion of contributions from the unbound state, i.e. the strain energy upon binding and ligand desolvation. In light of this, the straightforward implementation of a statistical mechanics treatment of the vibrational entropy e.g. by normal-mode analysis and/or the explicit inclusion of ligand desolvation by fast implicit solvent models are expected to improve the quality of the docking predictions. Interestingly, these contributions are already included in some empirical scoring (ES) functions.[83].

Finally, our classification of methods (Figure 4.4) highlights different sources of systematic error in the evaluation of the protein-ligand binding affinity. Sampling of the configurational space accessible to the system in the bound and the unbound state is one of them, which explains, for instance, the observed increase in accuracy on moving from the fast empirical methods to the end-points strategies. An accurate treatment of the solvent is another important aspect, which is well exemplified by the comparison of the end-points strategies with the rigorous methods. In this case, when the computationally intensive evaluation of the solvation free energy in the latter is replaced by continuum models (MM/PBSA) or an ensemble average of the ligand/surroundings interactions (LIE), sizable inaccuracies may be introduced. Last but not least, a force-field representation of interactions which neglects polarization effects is another source of systematic error, which will affect the quality of the predictions independently of sampling. Simplified quantum-mechanical treatments such those introduced in QM-MM/PBSA[69] and QMLIECE[79] represent pioneering attempts to quantify this type of errors. The development of strategies in which errors due to undersampling or a continuum treatment of the solvent are roughly equal in size to those introduced by the force-field parameterization will be key for the development of optimal computational approaches to protein-ligand binding.

5

A LINEAR INTERACTION ENERGY MODEL FOR CAVITAND HOST-GUEST SYSTEMS

5.1 INTRODUCTION

Host-guest complexes have attracted significant interest in recent years both from experimental and computational chemists.[86–88] The host is typically a small synthetic molecule with a well-defined cavity or cleft, where a number of compounds (i.e. the guests) bind with remarkable affinity and/or selectivity.[15] The formation of host-guest complexes in solution is driven by the same non-covalent forces that steer protein-ligand binding such as hydrogen bonding, electrostatic and Van der Waals interactions, etc., which makes them suitable model systems to explore molecular recognition in solution.[16,17,89] In addition, a number of synthetic hosts were shown to be interesting targets for technological applications as chemosensors, biomimetics, solubility enhancers, reaction containers or drug carriers,[7,14,90,91] and the design of scaffolds that bind potently and selectively specific families of guests is currently an active research field.[92,93] In this context, the development of accurate and efficient numerical approaches to evaluate the binding constant in host-guests may open to the rational design of molecular function(s).

The accurate calculation of the free energy of binding in solution, ΔG°_{b}, remains a grand challenge in computational chemistry.[94] Despite a number of numerical strategies have been proposed, from fast and crude scoring functions to the most accurate and expensive quantum chemistry methods, there is no universal approach to the binding constant.[95] Among the pool of available methods, "end-point" approaches such as LIE[44,45,74] and MM/PBSA[42] provide efficient (though approximate) strategies to the free energy of binding and have been recently used in drug discovery.[32] The great advantage of these methods is that they sample only the configurational space of the initial and final states of the binding reaction, which drastically increases the efficiency of the calculations relative to more rigorous approaches. However, the quality of their predictions has been questioned, the accuracy of LIE being dependent on a set of empirical and typically non-transferable parameters,[76,78] while that of MM/PBSA being limited by the evaluation of the solvent contribution by continuum electrostatics.[96]

This chapter is centered in the development of a LIE model for cavitand host-guests with remarkable efficiency and predictive power. Many unexplored details of the LIE model have been addressed regarding to the quality of the predictions in dependence of the energy model and training set used. The usefulness and limitations of the model is illustrated by the numerical evaluation of the differential binding affinity of challenge host-guest systems based on the cucurbituril (CB[n]) hosts.

5.2 THEORY OF LIE

The theory of LIE[44,45] states that ΔG°_{b} can be obtained from the ensemble averages of the electrostatics and van der Waals contributions to the interaction energy of the ligand with the surroundings in the bound and the unbound states as

$$\Delta G^{\circ}_{b} = \beta \left[\left\langle U^{elec}_{L-s} \right\rangle_{b} - \left\langle U^{elec}_{L-s} \right\rangle_{ub} \right] + \alpha \left[\left\langle U^{vdw}_{L-s} \right\rangle_{b} - \left\langle U^{vdw}_{L-s} \right\rangle_{ub} \right] \tag{5.1}$$

where α and β are empirical parameters, the symbol $\langle\ \rangle$ indicates ensemble averages typically collected by Molecular Dynamics (MD) simulations, and the subscripts b and ub refer to the bound and unbound states, respectively.

5.3 RESULTS AND DISCUSSION

5.3.1 *Building the LIE model*

The LIE parameters for cavitand host-guests were generated using a training set of 14 complexes based on the cucurbit[7]uril (CB7) host for which experimental binding affinities in water were available;[16]. The ability of CB7 to bind a highly diverse set of ligands,[97] i.e. rigid/flexible, neutral/charged, and alkyl/aromatic compounds, makes it an ideal framework for training the model (Figure A.1). For each guest, classical MD simulations with an explicit treatment of the solvent were carried both in the bound and free state in solution using the General Amber Force Field (GAFF);[98]. The LIE parameters were then obtained by linear fitting the ligand/surrounding interaction energies versus experimental binding affinities using Eq.5.1. Linear fitting of the GAFF simulation results produced a good correlation with experiments (i.e. RMSE of 1.35 kcal/mol and R^2 of 0.62; Figure 5.1 and Table 5.1) and yielded $\alpha = 0.43$ and $\beta = 0.20$. We note that these values are significantly different from those that are typically used in protein-ligand binding ($\beta = 0.33 - 0.5$ and $\alpha = 0.18$),[45] with the coefficient for the electrostatic interactions β being approximately half the theoretical value of 0.5, and α being twice as large.

Table 5.1: Statistical metrics used to assess the accuracy of the computed $\Delta G^{\circ}_{b,LIE}$ for the *training set* using the LIE models based on GAFF and CGenFF LIE models.

Metric	GAFF	CGenFF
RMSE	1.35	1.70
MAE	1.25	1.48
R	0.79	0.63
R^2	0.62	0.39
Slope	0.68	0.34
Intercept	-4.30	-6.75

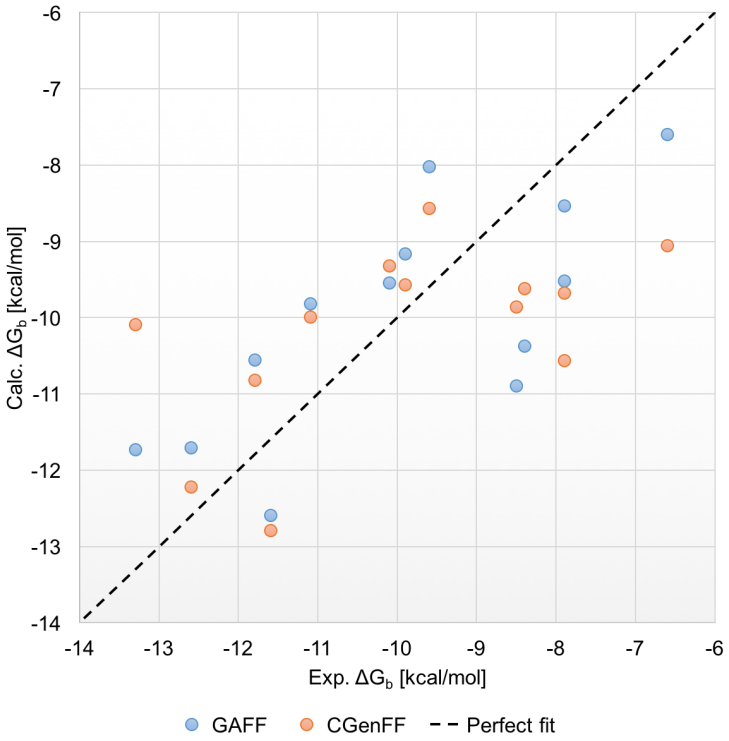


Figure 5.1: Experimental vs calculated binding free energy values in aqueous solution for the *training set* used to build the GAFF (R=0.79 and RMSE=1.35 kcal/mol) and CGenFF (R=0.63 and RMSE=1.70 kcal/mol) LIE models.

5.3.2 *Accuracy of the LIE model*

The predictive character of this LIE model was assessed using a test set of 49 chemically diverse cavitand host-guest complexes. The test set included 15 complexes of OAH, 6 complexes of the OAM, 22 complexes of CB7, and 6 complexes of BCD; see *Appendix A*. The OAH and OAM complexes were all part of SAMPL4 and SAMPL5 challenges[16,17] and provide a stringent benchmark for any computational approach. The 22 hydrocarbon guests in complex with CB7 were used in the HYDROPHOBE challenge, a recent experimental/computational benchmark of numerical approaches to the binding constant.[20] And, BCD is a flexible cavitand host that has been used as a solubility enhancer for drug formulation.[99]

The prediction strength of the LIE model was assessed by measuring the root mean square error (RMSE) and the mean absolute error (MAE) as metrics. The results in Figure 5.2 and Table 5.2 show a striking correlation with the experimental determinations (R=0.81) with a calculated RMSE of 1.08 kcal/mol . Note that this error is lower than any other reported in SAMPL4 and SAMPL5 using a variety of computational methods.[16,17] Remarkably, accurate predictions were obtained for the OAH (RMSE = 0.66 kcal/mol), OAM (RMSE = 1.06 kcal/mol) and BCD (RMSE of 1.48 kcal/mol) hosts individually, which were not part of the training set. Based on these results, we conclude that the LIE parameters above are transferable among chemically-diverse hosts families. In the case of the 22 CB7-hydrocarbons complexes, a RMSE

of 1.17 kcal/mol was obtained, surpassing the accuracy obtained by more rigorous methods based on expensive quantum calculations (RMSE = 1.94 kcal/mol) and/or extensive sampling based on MD (RMSE = 5.05 kcal/mol).[20] The accuracy of the predictions above indicate that a straightforward LIE model is able to capture the details modulating the host-guest binding affinity in solution.

Table 5.2: Statistical metrics used to assess the accuracy of the computed $\Delta G^{\circ}_{b,LIE}$ for the *test set* using the LIE models based on GAFF and CGenFF LIE models.

Force Field	Metric	Overall	OAH	OAM	BCD	CB7
GAFF	RMSE	1.08	0.66	1.06	1.48	1.17
	MAE	0.88	0.55	0.90	1.37	0.95
	R	0.81	0.88	0.98	0.95	0.92
	R^2	0.66	0.77	0.97	0.91	0.84
	Slope	0.72	0.56	0.20	1.47	1.05
	Intercept	-1.09	-2.10	-3.03	0.26	1.20
CGenFF	RMSE	0.92	1.06	0.97	0.54	0.88
	MAE	0.64	0.74	0.61	0.42	0.64
	R	0.87	0.77	0.55	0.97	0.91
	R^2	0.76	0.60	0.30	0.94	0.83
	Slope	0.80	0.44	0.26	1.26	1.05
	Intercept	-0.86	-2.75	-3.19	0.49	0.70

5.3.3 *Effect of the energy model*

To assess the impact of the force-field on the accuracy of the predictions, a new set of LIE parameters was derived using the CHARMM General Force Field (CGenFF).[100] The new LIE model parameterized on the same training set presents $\alpha = 0.66$ and $\beta = 0.08$; see Figure A.1 of the *Appendix A*. Despite the values of α and β are substantially different from the previous parameterization, the calculated RMSE was 0.92 kcal/mol (see Figure 5.3 and Table 5.2), which is consistent with the accuracy obtained using GAFF. In addition, the LIE model based on CGenFF produced accurate results for the individual host families; OAH (RMSE=1.06 kcal/mol), OAM (RMSE=0.97 kcal/mol), CB7 (RMSE=0.88 kcal/mol) and BCD (RMSE=0.54 kcal/mol). We conclude that although the LIE parameters are force-field dependent, the accuracy of the binding-affinity predictions in these host-guest complexes is not. Thus, provided a chemically diverse training set, the empirical parameterization of LIE is likely to absorb most of the systematic error of the force-field, making LIE even more accurate than the model of energetics in use.

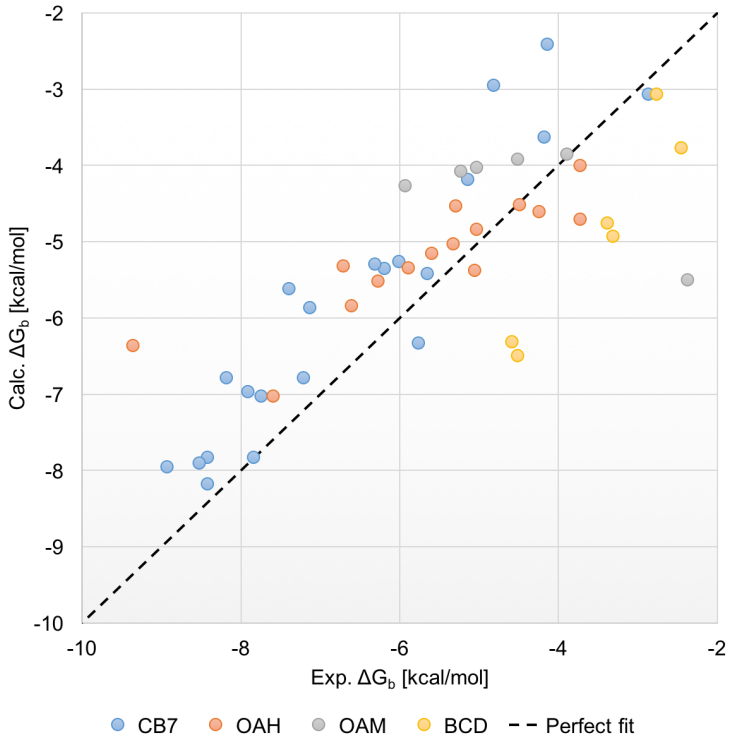


Figure 5.2: Experimental vs calculated binding free energy values in aqueous solution for host-guest systems of the *test set* from the GAFF LIE model.

5.3.4 *Efficiency and robustness of the LIE model*

In a first assay to test the robustness of our LIE model for cavitand host-guest systems with respect to the selection of the training set, two analyses were carried out. First, the entire data set of 61 host-guest complexes (training plus test) was randomly split in two new sets of 14 and 47 host-guest complexes, which were used as new training and test sets, respectively. New LIE parameters were obtained and the RMSE for the test set evaluated. By repeating the procedure 1×10^5 times, the frequency distributions for the LIE parameters α and β and the RMSE for the test set were obtained, which are presented in the Figures A.5 of the *Appendix A*. The results indicate that the most populated values of α and β are essentially equivalent to those used in the original model, which were based on an arbitrary selection of the training set. Moreover, the distribution of the RMSE shows that free energy results within 1.5 kcal/mol from experiments are obtained almost independently of the training set. Based on this analysis, we conclude that the empirical parameters of our LIE model for cavitand host-guests as well as the accuracy of the binding affinity predictions are essentially independent of the training set.

In a second analysis, the size of the initial training set, which was composed of n=14 CB7-guest complexes , was systematically reduced by random elimination of $k = 1, 2, 3 \ldots 11$ complexes. For each value of k, $\frac{n!}{k!(n-k)!}$ unique training sets of $n - k$ members were generated.

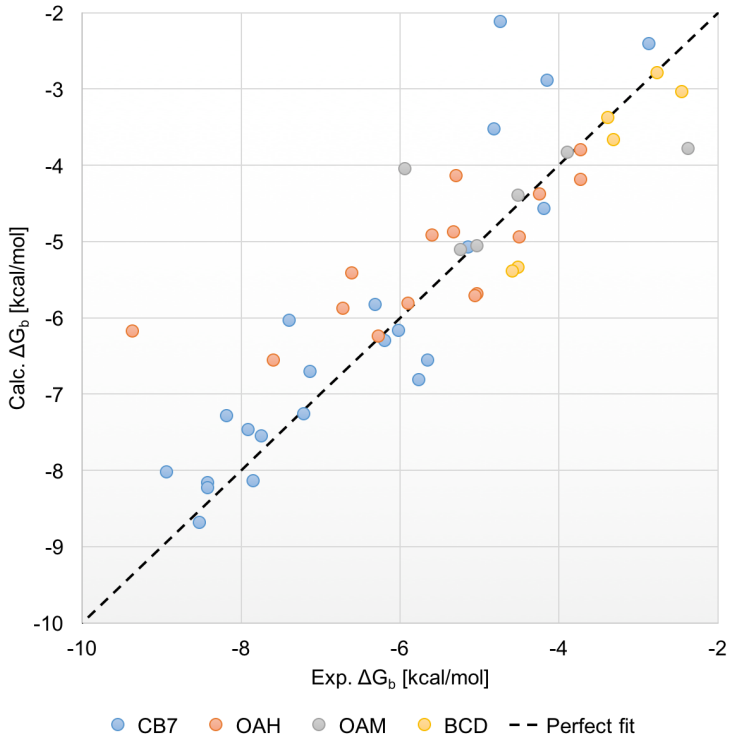


Figure 5.3: Experimental vs calculated binding free energy values in aqueous solution for the *test set* using the CGenFF LIE model (overall R=0.87 and RMSE=0.92 kcal/mol).

New LIE parameters were obtained and the RMSE for the test set evaluated; note that the k complexes removed from the training set were considered as part of the test set to preserve the size of the full data set (61 host-guest complexes). Average values and associated errors for α, β and RMSE for the test set as a function of the size of the training set are presented in Figures A.6 of the *Appendix A*. The results clearly show that reliable LIE models yielding RMSE <1.5 kcal/mol can be obtained using a training set including as little as 7 experimental determinations of the host-guest binding affinity.

Finally, the convergence analysis of the binding affinity predictions by LIE was performed to explore the efficiency of the methodology and its suitability for virtual screening campaigns. For this purpose, the simulation time (t_{min}) required to obtain predictions with a deviation of <0.5 kcal/mol from the ΔG_b° at full sampling (20 ns) was used as a convergence metric. The frequency distribution of t_{min} for all complexes of the test set was evaluated and fitted with an exponential function of the form $\exp(-t/\tau)$, whose characteristic time τ was of 1.1 ns. Based on this analysis, we conclude that the simulation time required to obtain converged binding affinity results in most complexes of the test set was <2 ns independently of the force field, which demonstrates the remarkable efficiency of our LIE model; see Figure A.7 (top) of the *Appendix A*. Based on these results we conclude that accurate binding-affinity predictions in host-guests can be obtained by LIE with a few nanosecond MD. Also, the rapid convergence of the LIE parameters (α and β) as a function of MD sampling was evaluated

as presented in Figure A.7 (bottom) of the *Appendix A*. Taken together, these results support the conclusion that LIE provides an accurate and efficient access to the binding affinity in cavitand host-guests; which makes it suitable for virtual screening of large chemical libraries.

5.3.5 *Applications of the LIE model*

As a first application, the LIE model based on GAFF was used to predict the standard binding free energy of 19 steroids to cucurbituril hosts, which were shown to bind CB7 and CB8 with nanomolar affinities in water;[101] their chemical structures are shown in Figure A.8 of the *Appendix A*.

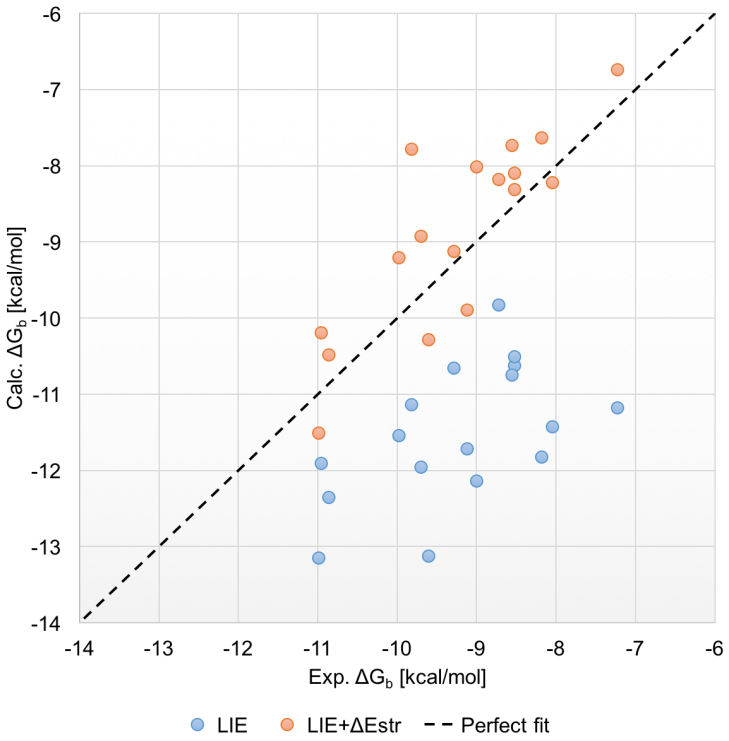


Figure 5.4: Experimental vs calculated binding free energy values in aqueous solution for CB[7/8]-steroids complexes in study.

Direct application of LIE produced a RMSE of 2.45 kcal/mol from experiments; see Figure 5.4 (blue points) and Table 5.3. This result was not totally satisfactory compared to previous RMSE values obtained for the test set. Visual inspection of the MD trajectories unveiled that a significant deformation of the host is introduced by the bulky steroids in the bound state, particularly on the smaller host CB7 (Figure 5.5).

Since the theory of LIE assumes that the ligand is small compared to the receptor and that the conformation of the latter is minimally affected upon complexation,[95] it cannot account for the receptor strain in the bound state. In steroid-cucurbituril complexes, however, this assumption is unjustified as steroid guests have a comparable size to the host. Based on these

Table 5.3: Statistical metrics used to assess the accuracy of the computed $\Delta G^{\circ}_{b,LIE}$ for the CB[7-8]/steroids complexes using the GAFF LIE model with and without ΔE_{str} values.

Metric	$\Delta G^{\circ}_{b,LIE}$	$\Delta G^{\circ}_{b,LIE} + \Delta E_{str}$
RMSE	2.45	0.81
MAE	2.27	0.67
R	0.55	0.82
R^2	0.30	0.67
Slope	0.46	0.77
Intercept	-7.24	-1.55

considerations, we developed an original LIE model that accounts for the strain energy of the host (ΔE_{str}) in the evaluation of the binding free energy. In the new formulation

$$\Delta G^{\circ}_{b} = \beta \left[\left\langle U^{elec}_{L-s} \right\rangle_{b} - \left\langle U^{elec}_{L-s} \right\rangle_{ub} \right] + \alpha \left[\left\langle U^{vdw}_{L-s} \right\rangle_{b} - \left\langle U^{vdw}_{L-s} \right\rangle_{ub} \right] + \Delta E_{str} \tag{5.2}$$

where $\Delta E_{str} = E^{host}_{b} - E^{host}_{ub}$, with E^{host}_{b} and E^{host}_{ub} are the force-field energies for the bound and unbound states at the minimum, respectively; see *Material and Methods* for details on the calculation of the strain energy correction. Strikingly, the results in Figure 5.4 (orange points) show that the inclusion of the strain energy contribution improves the binding-affinity predictions dramatically with a final RMSE of 0.81 kcal/mol and R^2 of 0.67 (Table 5.3).

The new implementation of LIE (Eq. 5.2) was used to re-evaluate the binding affinity of the 49 host-guests complexes of the test set. Since several guests are formally charged, the evaluation of the strain energy correction was modified and an implicit solvent model introduced to screen for the host-guest electrostatic interaction during geometry optimization of the host with and without the guest.The results (Table 5.4) show that the strain-energy contribution is small (< 1 kcal/mol) for all but three complexes, where inclusion of the correction improves the binding affinity predictions; i.e. R^2 increases from 0.66 to 0.71. It follows that when the guest is flexible and small compared to the host, the strain energy contribution is negligible

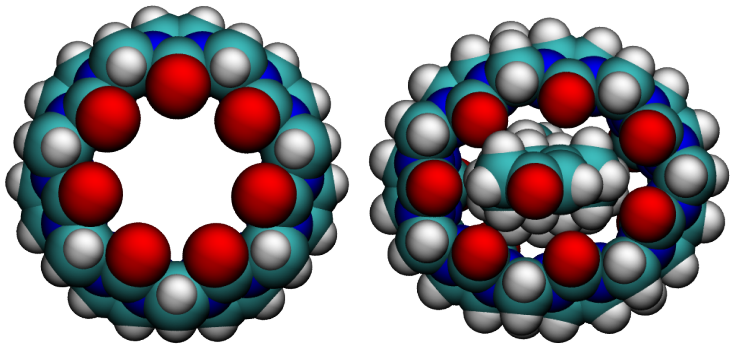


Figure 5.5: Geometry of the most populated cluster of CB7 in the unbound state (left) and the bound state to nandrolone (right) from MD simulations in explicit solvent. The structural deformation of the host in the bound state is striking.

and Eq. 5.2 is equivalent to Eq. 5.1. Hence, the strain-including formulation of LIE in Eq. 5.2 is a generalization of the original LIE, which accounts for the strain energy of the host when host and guest have comparable sizes.

Table 5.4: Statistical metrics used to assess the accuracy of the computed $\Delta G^{\circ}_{b,LIE}$ for the *test set* using the GAFF LIE model with ΔE_{str} values.

Metric	Overall	OAH	OAM	BCD	CB7
RMSE	1.14	0.74	1.28	0.90	1.35
MAE	0.96	0.65	1.21	0.75	1.17
R	0.84	0.86	0.95	0.78	0.90
R^2	0.71	0.74	0.89	0.61	0.82
Slope	0.72	0.56	0.43	0.83	0.97
Intercept	-0.87	-1.98	-1.61	-1.30	0.93

As a second application, the binding affinities of 28 guests for CB7 extracted from the recent work of Muddana *et al*[102] (Figure A.9 of the *Appendix A*) were analyzed. This additional test set includes highly charged and bulky compounds spanning a wide range of affinities from weak to ultra-tight; i.e. the ΔG°_{b} varies from -5.3 to -21.5 kcal/mol. We note that this dataset is challenging for numerical methods as demonstrated by the large RMSE of 4.8 and 10.2

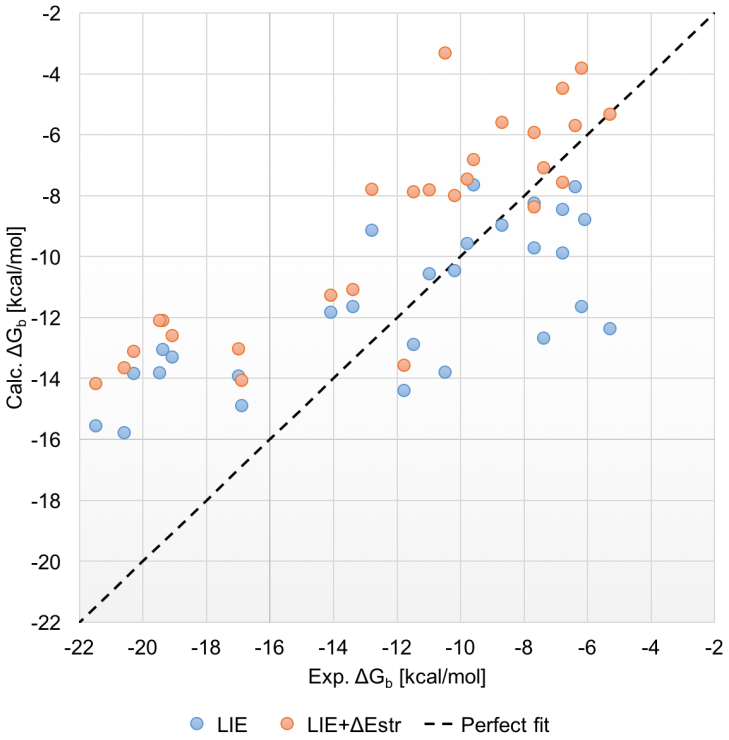


Figure 5.6: Experimental vs calculated binding free energy values in aqueous solution for the Muddana set in study.

kcal/mol produced by the mining minima method using force-field and quantum-chemical calculations, respectively, prior to fitting to the experimental data.[102] Predictions based on the GAFF/LIE model (Eq. 5.1) yielded a large RMSE of 3.47 kcal/mol from experiments with a poor R^2 of 0.3; see Figure 5.6 and Table 5.5. The predictions did not improve using the CGenFF/LIE model (RMSE of 4.26 kcal/mol) nor including the strain energy correction (RMSE of 4.40 kcal/mol); see Figure 5.6. A detailed analysis of the computational results highlighted the existence of two main sources of errors. First, our LIE model was unable to predict the affinity of the ultra-tight binders, i.e. guests with experimental $\Delta G^\circ_b < -16$ kcal/mol (e.g. M21-M28). Considering the numerical values of α and β in the GAFF/LIE model, a predicted binding affinity of -21.5 kcal/mol such that of M28 would require a vdW contribution to the LIE binding energy of -44 kcal/mol, which is ~ 11 kcal/mol larger than the maximum vdW component measured in this work. The situation is even worse when considering the electrostatic contribution as the β coefficient is half of the size of α. These data thus indicate that something may be fundamentally missing in the present LIE model, which would be key to capture the ultra-tight binding affinities to CB7. Perhaps, the use of polarizable force fields in combination with explicit water Molecular Dynamics will improve the performance, thus extending the scope of the model. Second, our LIE model overestimated the binding affinity of weak and bulky guests (e.g. M1, M2, M3, M7 and M14), whose binding to CB7 likely results in a strained conformation of the host. Upon inclusion of the strain-energy correction (Eq. 5.2), the correlation with the experiments did improve, i.e. R^2 increases from 0.53 to 0.76 and the slope of the regression goes from 0.35 to 0.61 (see Table 5.5), but the overall RMSE remains > 4 kcal/mol. Moreover, the results in Figure 5.4b show that inclusion of the strain energy of the host overshoots the experimental binding free energy systematically, thereby failing to correct the numerical predictions. Since the evaluation of the strain energy here required the use of an implicit-solvent model, we suspect that our numerical protocol is suboptimal (if not inadequate) with formally charged ligands. The development of better performing protocols for accurate strain-energy corrections on complexation is currently under investigation and will be reported elsewhere. Overall, the inaccurate predictions on the Muddana dataset highlight some of the shortcomings of the current LIE implementation and suggest future directions for improvement.

Table 5.5: Statistical metrics used to assess the accuracy of the computed $\Delta G^\circ_{b,LIE}$ for the *Muddana set* using the GAFF (with and without ΔE_{str} values) and the CGenFF LIE models.

Metric	GAFF		CGenFF
	$\Delta G^\circ_{b,LIE}$	$\Delta G^\circ_{b,LIE} + \Delta E_{str}$	$\Delta G^\circ_{b,LIE}$
RMSE	3.77	4.26	4.40
MAE	3.12	3.55	3.51
R	0.73	0.87	0.57
R^2	0.53	0.76	0.32
slope	0.35	0.61	0.23
intercept	-7.40	-1.38	-8.45

5.4 MATERIAL AND METHODS

5.4.1 *Computational details*

5.4.1.1 *Preparation of the systems*

The initial atomic coordinates for hosts in study; CB7, OAH and OAM and their guests, were obtained from publications of the blind challenges SAMPL4,[16] SAMPL5[17] and HYDROPHOBE.[20] The Initial geometries for both the BCD and CB8 hosts, which are not part of SAMPL4 and SAMPL5 challenges, were obtained from the Cambridge Structural Database[103] and their corresponding guests where built from SMILES. For all hosts and guests, the protonation states were assigned using the Marvin suite software[104] at the experimental pH. Here, the CB7, CB8, and BCD hosts were modeled as neutral, the OAH and OAM hosts had a net charge of -8. The chemical structures for all hosts were presented in the chapter 1, while the chemical structures and protonation states for all guests in study are presented in Figures A.1, A.2, A.3 and A.4 from the *Appendix A*. Initial atomic coordinates for the host-guest complexes were extracted from the top ranking binding mode predicted by docking using the CHEMPLP scoring function implemented in PLANTS.[105] Force field parameters for hosts and guests were generated using the General Amber Force Field (GAFF)[98] with AM1-bcc charges.[106] Also, to assess the impact of the force field on the binding affinity predictions, MD simulations were performed using the CHARMM general Force Field (CGenFF).[100]

5.4.1.2 *Protocol for Molecular dynamics simulations*

All Molecular Dynamics (MD) simulations were carried out using GROMACS 5.1.2[107] using periodic boundary conditions and a time step of 2 fs. Each molecular system was solvated in a cubic box with a minimum distance of 1.4 nm between the solute and the edge of the box. The TIP3P model was used to represent water molecules and counter-ions were added to grant neutrality of the simulation box. Electrostatic and van der Waals interactions were computed using particle mesh Ewald (PME) with a real-space cut-off of 1.2 nm, a grid spacing of 0.12 nm, a spline order of 4 and a relative tolerance of 1×10^{-6} for the reciprocal space. The LINCS algorithm was used to constrain all covalent bonds involving hydrogens. The simulation protocol started with an energy minimization of 10000 steps of steepest descent until a maximum force of $10\,\mathrm{kJ\,mol^{-1}\,nm^{-1}}$ was attained. The system was then slowly heated to the target temperature (i.e., 298 K) using a modified Berendsen thermostat[108] in 6 steps with increments of 50 K every 50 ps. The system was equilibrated for 1 ns at constant volume and for another 1 ns at the constant pressure of 1 bar. The first 500 ps of simulation were carried out using the Berendsen barostat,[109] the remaining 500 ps using the Parinello-Rahman barostat,[110] which grants correct sampling of the NPT canonical ensemble. In both cases, a barostat coupling parameter of 1 ps and a isothermal compressibility of $4.5 \times 10^{-5}\,\mathrm{bar^{-1}}$ were used. Finally, a production phase of 20 ns was performed at constant temperature and pressure. Molecular configurations were saved every 5 ps for further analysis.

5.4.1.3 *Practical aspects of LIE*

Since our simulations were performed using PME to treat both the electrostatic and van der Waals (LJ-PME[111]) interactions, which does not allow direct atomic pair-wise decomposition, a post-processing step was introduced to evaluate the ensemble averages in Equation 5.1. Starting with the bound state, the trajectory of the solvated complex was split in two trajectories, one containing the receptor and the solvent molecules (and also counter-ions if the system was not neutral), the other containing the ligand alone. From these two trajectories along with the one of the solvated complex, ensemble averages of the electrostatic and van der Waals (PME) energies for the complex in solution ($\langle U^{elec}_{RL+solv}\rangle_b$ and $\langle U^{vdw}_{RL+solv}\rangle_b$, respectively), the receptor in solution ($\langle U^{elec}_{R+solv}\rangle_b$ and $\langle U^{vdw}_{R+solv}\rangle_b$, respectively) and the ligand *in vacuum* ($\langle U^{elec}_{L}\rangle_b$ and $\langle U^{vdw}_{R+solv}\rangle_b$, respectively) were extracted. Then, the electrostatics and van der Waals contributions to the ligand interaction energy were evaluated as

$$\langle U^{elec}_{L/s}\rangle_b = \langle U^{elec}_{RL+solv}\rangle_b - \langle U^{elec}_{R+solv}\rangle_b - \langle U^{elec}_{L}\rangle_b \tag{5.3}$$

$$\langle U^{vdw}_{L/s}\rangle_b = \langle U^{vdw}_{RL+solv}\rangle_b - \langle U^{vdw}_{R+solv}\rangle_b - \langle U^{vdw}_{L}\rangle_b \tag{5.4}$$

Similarly, upon separating the MD trajectory of the ligand in three, i.e. free ligand in solution, free ligand in vacuum, and solvent alone, the electrostatics ($\langle U^{elec}_{L/s}\rangle_{ub}$) and the van der Waals ($\langle U^{vdw}_{L/s}\rangle_{ub}$) contributions to the ligand interaction energy in the unbound state were evaluated as

$$\langle U^{elec}_{L/s}\rangle_{ub} = \langle U^{elec}_{L+solv}\rangle_{ub} - \langle U^{elec}_{solv}\rangle_{ub} - \langle U^{elec}_{L}\rangle_{ub} \tag{5.5}$$

$$\langle U^{vdw}_{L/s}\rangle_{ub} = \langle U^{vdw}_{L+solv}\rangle_{ub} - \langle U^{vdw}_{solv}\rangle_{ub} - \langle U^{vdw}_{L}\rangle_{ub} \tag{5.6}$$

Finally, statistical errors associated with the numerical determination of ΔG°_{b} by LIE were estimated as in Baron *et al.*[112] For this purpose, the MD trajectories of the bound and unbound states were split in two chunks, named A and B, and ensemble averages of the electrostatic and van der Waals contributions to the ligand interaction energy were computed per chunk. Then, the statistical error associated with each ensemble average was estimated as

$$\langle E_{L-s}\rangle = \frac{1}{2}\left|\langle U^{A}_{L-s}\rangle - \langle U^{B}_{L-s}\rangle\right| \tag{5.7}$$

and that on ΔG°_{b} by a LIE-like equation as

$$Error_b = \beta\left[\left\langle E^{elec}_{L-s}\right\rangle_b + \left\langle E^{elec}_{L-s}\right\rangle_{ub}\right] + \alpha\left[\left\langle E^{vdw}_{L-s}\right\rangle_b + \left\langle E^{vdw}_{L-s}\right\rangle_{ub}\right] \tag{5.8}$$

5.4.1.4 *Computing the Strain Energy of the Host (ΔE_{str})*

Here, the computation of the strain energy of the host was done using `sander`[113] from the AmberTools17 suite package[114] as follows

1. Initially, a cluster analysis (using the GROMACS tool "`gmx cluster`") on the production trajectory of a host-guest complex was performed in order to isolate the structure of the most populated cluster. The clustering procedure was based on the "single linkage" method where one conformation of the complex is added to a particular cluster if its all-atom RMSD (after mass-weight fitting) was less than 0.25 nm.

2. The central conformation of the most populated cluster was submitted to energy minimization by performing 50000 steps of conjugate gradient, prior to four cycles of local minimization using steepest descent (`maxcyc=50000` and `ntmin=0` in `sander`) until the root-mean-square of the energy gradient was less than 1×10^{-5} kJ mol^{-1} Å^{-1} (`drms=`1×10^{-5} in `sander`). All geometry optimizations were done *in vacuum* for neutral system, whereas the GBSA implicit model by Hawkins *et al*[115] (`igb=1` in `sander`) was used for formally charged systems.

3. The atoms of the host were extracted from the optimized structure of the complex and its intramolecular energy evaluated, which yields the configurational energy of the host at the minimum in the bound state, E_b^{host}.

4. Then, the conformation of the host previously optimized in the presence of guest (i.e. the bound state) was energy minimized with the guest removed using the same procedure. Evaluation of the energy of the host yields the configurational energy of the host at the minimum in the unbound state, E_{ub}^{host}.

5. Finally, the strain energy of the host was determined as $\Delta E_{str} = E_{ub}^{host} - E_b^{host}$.

5.5 CONCLUSION

In conclusion, we have presented a LIE model for cavitand host-guest binding affinities that is transferable among chemically diverse families, accurate and reliable, producing predictions with a RMSE $<$ 1.5 kcal/mol in a large test set including 49 guests and four different hosts. Our model is computationally efficient, its performances are essentially independent of the training set, and produces converged results within a few nanoseconds of MD, which opens to high-throughput computational screenings. The semi-empirical character of the model was shown to absorb most of the systematic error of the force field, making the predictions essentially force-field independent; a considerable advantage over other physics-based approaches that cannot be more accurate than the model of energetics in use. Finally, the inclusion of the strain energy of the host in the calculation of the binding affinity, which is absent in the original LIE formulation, was shown to improve the quality of the predictions substantially, especially when hosts and guests have similar sizes. Nonetheless, the current formulation of LIE failed in predicting the binding affinity of ultra-tight (femto- to atto-molar) binders to CB7 and was shown to overestimate the strain energy of the host in complex with bulky and formally charged guests. The usefulness of a LIE formulation for host-guest recognition was demonstrated through the accurate prediction of steroid binding to cucurbituril hosts, which are technologically relevant for the development of chemosensors.

6

THE CO-CATALYTIC EFFECT OF "INERT" MOLECULES IN BRØNSTED ACID CATALYZED REACTIONS

6.1 INTRODUCTION

The proton provides the most common and important way to catalyze organic reactions, however it needs not always operate in isolation. Within a single molecular catalyst, such as in the active site of an enzyme, multiple hydrogen-bond donors (or Brønsted acidic sites) can work in cooperation to increase the overall acidity and hence catalytic activity.[116] Co-operative H-bonding and Brønsted acidity is also possible between molecules, when multiple H-bond donors or Brønsted acid molecules interact to generate an aggregate that is a more effective catalyst than either individual molecule.[117] However, what role the bystander molecules – those that are not H-bond donors or Brønsted acids – might play, if any, remains to be elucidated.[118,119]

Though poorly understood, a few papers indicate that interactions between Brønsted acids and seemingly innocuous solvents or additives, particularly nitro compounds, can have a profound accelerating effect on catalytic activity both in terms of reaction rate and the concentration dependence of the reaction. In the 1960s, Pocker and coworkers[120–122] found that the hydrochlorination of olefins was not only markedly faster when carried out in nitromethane compared to other solvents, but also displayed an atypical second order kinetic concentration dependence on HCl (Figure 6.1). Further investigations to uncover why this effect would be exclusive to nitromethane were not made, and some subsequent authors have attributed it to the increased polarity of the bulk solvent. Recently in 2015[123], it was observed that the presence of nitro compounds, either as solvent or co-catalyst, greatly accelerated the dehydroazidation of tertiary aliphatic alcohols catalyzed by $B(C_6F_5)_3 \bullet H_2O$ (BCF), a strong Brønsted acid of comparable strength to HCl. When carried out in benzene in the presence of catalytic quantities of nitro compound, the reaction was found to display a second order concentration dependence with respect to the nitro compound as well as a second order concentration dependence with respect to BCF. In contrast, when carried out in benzene in the absence of nitro compounds, the reaction was much slower and found to be first order with respect to BCF (Figure 6.1). No model currently exists to account for how the presence of a simple molecule like a nitro compound could change the concentration dependence and rate of a Brønsted acid catalyzed reaction.

Herein, it's presented a structural model for a higher order aggregate formed by nitro compounds and Brønsted acids based on DFT calculations. We examine the effect of aggregation on acidity of the Brønsted acid, show that the model can be used to predict reaction rates for a set of new nitro compounds, and even predict a new class of "template" molecules that can induce a similar co-catalytic effect. The study demonstrates that weak interactions between Brønsted acids and all molecules in solution must be taken into account to achieve a com-

prehensive understanding of Brønsted acid catalyzed reactions. This is an clear example of a quantitative analysis of molecular recognition in solution for catalysis.

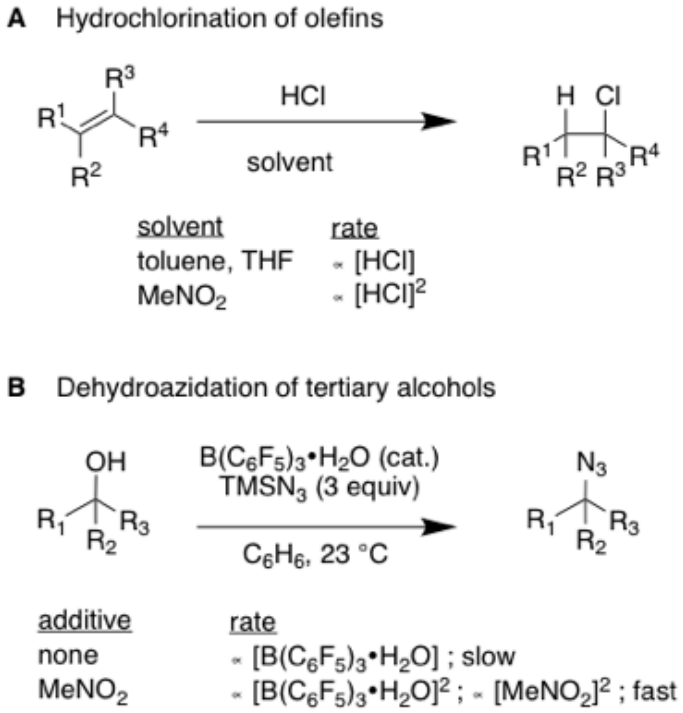


Figure 6.1: The effect of nitromethane on the kinetic concentration dependence of reactions mediated (A) or catalyzed (B) by Brønsted acids.

6.2 RESULTS

6.2.1 *Structural model for the BCF/nitro compound agregate*

Based on the observed second-order dependence of alcohol dehydroazidation on the concentration of BCF and nitro compound, as well as spectroscopic and catalytic evidence of hydrogen bonding to nitro compounds,[123] we postulated that the self-assembly of two BCF with two nitro-compound molecules is responsible for the co-catalytic effect. Structural models of the tetrameric assembly of BCF with nitromethane were generated by connecting the molecules via pairs of H-bonds and optimizing the geometry of the aggregate at the DFT level of theory using the ωB97X – D[124] functional with the 6-31G(d,p) basis set; see *Material and Methods* for details. The DFT optimization results in an almost flat, rectangular hydrogen-bonded network formed between the nitro groups of two nitromethanes and two boron hydrates (Figure 6.2). The tetrameric arrangement is stabilized by two sets of non-equivalent H-bonds, two shorter and two longer. Vibrational analysis of the DFT-optimized architecture

indicates that each pair of O-H groups in the network present two distinct stretching modes (i.e. one symmetric with a lower IR intensity signal and one anti-symmetric with higher intensity) with the symmetric stretching of the hydrogens involved in the shorter H-bonds characterized by the lowest vibrational frequency of 3501 cm-1; see Figure B.1 and B.4 of the *Appendix B*.

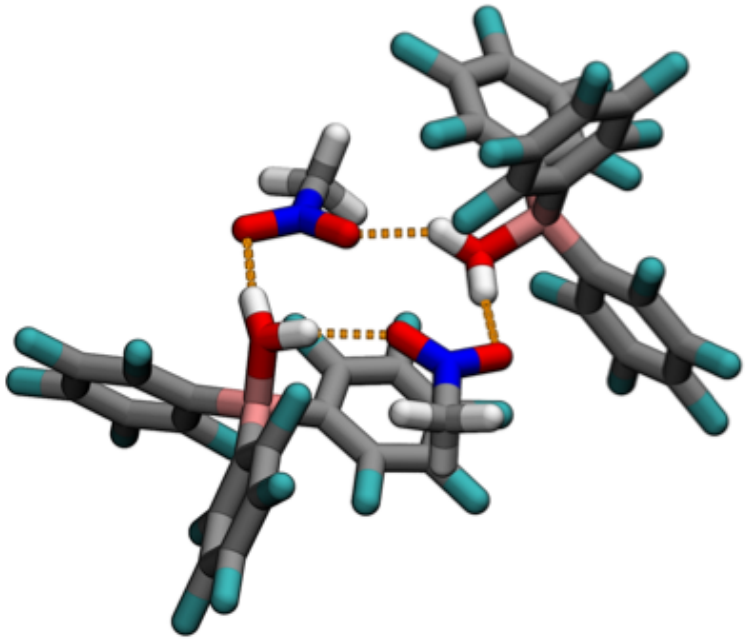


Figure 6.2: DFT-optimized structure for the 2:2 tetrameric self-assembly of BCF and nitromethane.

Thus, we conclude that the formation of a tetrameric adduct with nitromethane results in a red-shift of the O-H stretching frequency as large as the corresponding O-H stretching modes have vibrational frequencies of 3693 and 3670 cm^{-1} in the isolated BCF (i.e. the higher the frequency, the stronger the bond). Thus, the calculations suggest that BCF:nitromethane tetrameric self-assembly increases the acidity of BCF, in particular for the hydrogens involved in the shorter hydrogen bonds. The same vibrational analysis of 1:1 complex of BCF with nitromethane (see Figure B.2) shows that the dimeric form is not the specie responsable for the co-catalytic effect since a marginal red-shift of the O-H stretching frequency was detected in this 1:1 complex. These results imply that the interactions between the nitro compound and BCF in the 2:2 aggregate occurs at just the right angle and strength to have a net effect on the acidity.

Following the same procedure, tetrameric assemblies of BCF with a variety of nitro compounds including 4-nitrobenzotrifluoride, nitrobenzene, 4-nitroanisole, 2-methyl-2-nitropropane and 1-nitrohexane, which were shown to experimentally modulate the co-catalytic activity of BCF,[123] were generated. Upon geometry optimization at the DFT level of theory, the rectangular hydrogen-bonding network made of pairs of non-equivalent H-bonds is largely preserved in all complexes, with minor distortions in some cases; see Figure B.5 of the *Appendix B*.

Table 6.1: Computed and experimental OH stretching frequencies and Log(rate) values employed to build and validate correlations. [a]Outlier in the experimental Model, [b]in cm^{-1}, [c]Test set.

Comp.	Name	Exp. Log(rate)	DFT model [b]IR Freq.	DFT model Pred. Log(rate)	IR model [b]IR Freq.	IR model Pred. Log(rate)
1	4-nitrobenzotrifluoride	-5.13	3601.31	-4.99	3523	-5.07
2	nitrobenzene	-4.58	3506.99	-4.65	3480	-4.63
3	4-nitroanisole	-4.34	3454.68	-4.46	3465	-4.47
4[a]	nitromethane	-4.62	3530.63	-4.74	-	-
5	1-nitrohexane	-4.94	3561.62	-4.85	3500	-4.83
6	2-methyl-2-nitropropane	-4.83	3502.86	-4.64	3488	-4.71
7	Background	-5.2	3692.52	-5.32	3547	-5.32
8[c]	1-nitropropane	-4.75	3531.32	-4.74	3510	-4.93
9[c]	2-nitropropane	-4.77	3528.85	-4.73	3510	-4.93
10[c]	nitroethane	-4.77	3525.08	-4.72	3510	-4.93
11[c]	2-nitroanisole	-4.39	3447.33	-4.44	3522	-5.06
12[c]	1-nitronaphthalene	-4.81	3615.81	-5.04	3468	-4.5

Vibrational analysis of the model complexes shows the same characteristic red shift in the O-H stretching frequency with a magnitude that is dependent on the chemical nature of the nitro compound; see Table 6.1. Most importantly, the calculated red shift in the anti-symmetric stretching of the short H-bonds is found to be strongly correlated with the experimental Log(rate) with a determination coefficient (R^2) of 0.81; see Figure 6.3 (red filled circles). This observation suggests that the O-H stretching frequencies predicted by DFT in the 2:2 aggregate can be used as predictors for the kinetic rate of dehydroazidation catalyzed by BCF. Intriguingly, the frequency of the symmetric stretching of the short H-bonds or those of the longer H-bonds do not correlate with the experimental Log(rate); see Table B.1 of the *Appendix B*.

To evaluate the statistical significance of the correlation in Figure 6.3, two validation schemes were followed. First, a cross (internal) validation based on a leave-one-out (LOO) analysis was carried out. Second, an external validation was performed using a new set of five nitro compounds (1-nitropropane, 2-nitropropane, nitroethane, 2-nitroanisole and 1-nitronaphtalene) whose influence on reaction rate was determined by GC-MS. These experimental measurements were provided by the Moran's Group at the University of Strasbourg.

The results of the internal validation indicate that the DFT model presents a positive predictive profile with a cross-validated squared correlation coefficient (q^2) greater than 0.46 (Table 6.2). Those of the external validation show that kinetic rates predicted by the DFT model are strongly correlated with the experimental rates with a determination coefficient (R^2) of 0.70. Also, the statistical parameters of the model appear to be robust to randomization of the experimental kinetic rates in alternative training/test sets, which is consistent with the absence of random correlations; see Table B.2. The predictive character of the DFT model in

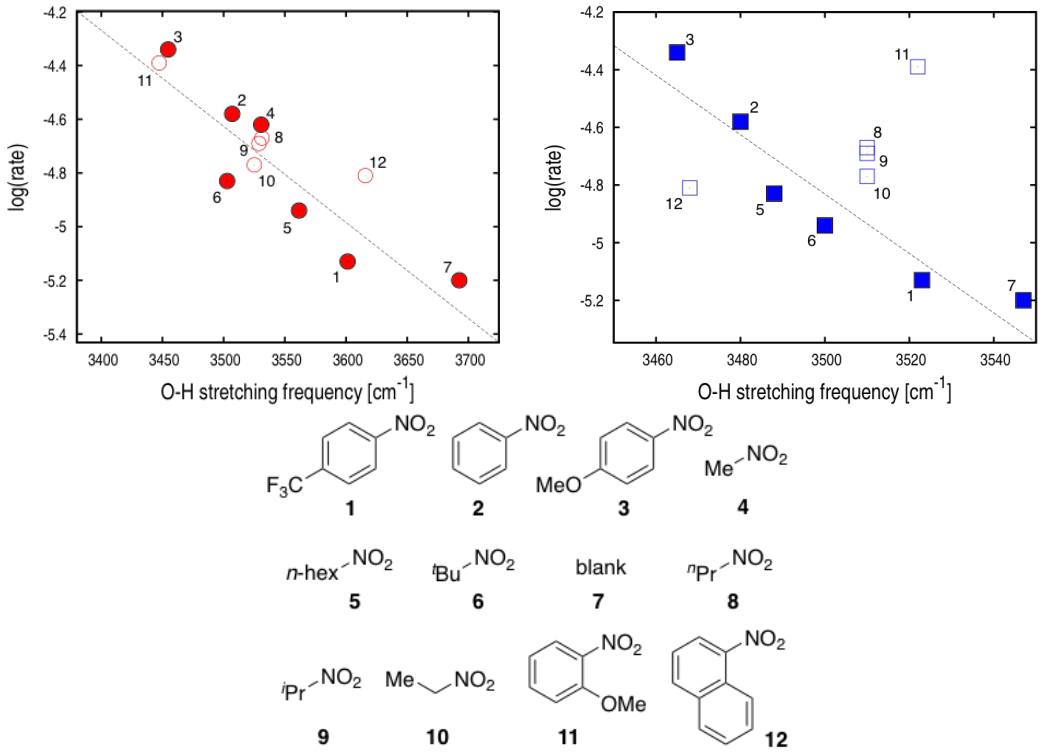


Figure 6.3: Correlations of the experimental kinetic rate with the DFT calculated frequency of the anti-symmetric stretching of the short H-bonds in the 2:2 aggregate (top left) and experimentally observed IR stretching frequencies (top right). Filled and empty data points correspond to the training (six + blank) and test (five) sets of nitro compounds investigated in this study. The numbering is the same as in Table 6.1.

Figure 6.3 supports the assertion that tetrameric self-assembly of BCF with nitro compounds is a critical feature of the catalytic mechanism in the azidation of tertiary alcohols.

A similar model for the Log(rate) was constructed using experimentally observed IR stretches. These experimental IR stretching frequencies were obtained from the Moran's group at the University of Strasbourg. Although this IR model was predictive for the initial set of nitro compounds, it is not for the test set (Table 6.2 and Figure 6.3). In particular, the accelerations produced by 2-nitroanisole and 1-nitronapthalene (entries 11-12 in Table 6.2) are reversed by the IR model, which produces an anti-correlation between predicted and observed Log(rate) in the external validation; see Figure 6.3. The more limited predictive power of the IR model compared to DFT can be explained on the basis that the intense OH-stretching bands derive from the overlap of multiple stretching modes, which limits the resolution of the experimental determination. Despite the inaccuracies associated with the DFT functional/basis set and the absence of solvent in the calculations, the computational vibrational analysis uniquely allows to quantify the red shift in the anti-symmetric stretching of the short H-bonds, which is predictive for the Log(rate).

Table 6.2: Statistical parameters for internal and external validations of the computational and experimental models. Capital and un-capital parameters refer to the internal and the external validation, respectively.

Parameter	DFT model	IR model
r^2	0.81	0.88
m	-0.00358	-0.010
b	7.891	31.229
q^2	0.46	0.6
SDEP	0.08	0.19
R^2	0.7	0.34
R_0^2	0.7	-1.68
$(R^2 - R_0^2)/R^2$	0.0005	5.97
k	1.01	1.04

6.2.2 *New promotors of the co-catalytic effect in the*

Lastly, the DFT model was used to explore new chemotypes capable of playing the same co-catalytic role as nitro compounds. For this purpose, six representative compounds were selected on the basis of their ability to weakly accept hydrogen bonds on two different atoms. For each compound, the tetrameric assembly with BCF was modeled and the relevant O-H stretching frequencies calculated by DFT (Table 6.3). In four out of six cases, the assembly preserved a rectangular-shape and a planar hydrogen-bond network after geometry optimization. Of these four, the Log(rate) predicted by the DFT model was in good correspondence with the observed reactivity after 24 h under standard reaction conditions in three cases (entries 1-3). These kinetics experiments were performed by the Moran's group at the University of Strasbourg. Also, the two compounds that did not result in stable tetrameric complexes in the calculations did not show any accelerating effect on the reaction (entries 5-6). Most importantly, as predicted by the DFT model, diethylsulfate (entry 1) as additive showed a similar accelerating effect to nitromethane. The computed aggregate for diethylsulfate is shown in Figure 6.4. Indeed, extra kinetics experiments shown a kinetic order dependence of 2.55 on diethylsulfate, consistent with higher order aggregates being involved in catalysis mediated by nitro compounds. To the best of our knowledge, this is the first time that an additive other than a nitro compound has been shown to induce changes in the kinetic order dependence and an accelerating effect in the azidation reaction. The incorrect prediction for Log(rate) in entry 4 illustrates that above a certain threshold of H-bond accepting ability, the buffering effect of the additive dominates any positive effects arising from aggregation.

6.2.3 *Limitations of the DFT model*

The apparent lack of predictivity with dimethyl-sulfone can be ascribed to both the level of theory used for the geometry optimization and/or the vibrational analysis, or the ab-

Table 6.3: Calculated frequency analysis for tetrameric self-assembly of BCF with various promoters and comparison to the observed reactivity in the dehydroazidation of alcohols. [a]A complex was considered stable if it preserves a rectangular-shape and planar hydrogen bond network as obtained in tetramer self-assembly of BCF with nitro compounds. [b] Difference between the OH-stretching frequency of BCF in the complex and alone (3692.52 cm^{-1}). [c] Reaction monitored by GCMS relative to dodecane as internal standard.

Entry	Promotor (Stable agregate?)[a]	OH stretch (cm^{-1})	$\Delta\mu$ (cm^{-1})[b]	Pred. Log(rate)	Conv. at 24 h (%)[c]
1	Diethylsulfate (yes)	3510.47	182.05	-4.66	>99%
2	Dimethyloxalate (yes)	3561.43	131.09	-4.85	60%
3	Dimethylmalonate (yes)	3562.81	129.71	-4.85	58%
4	Dimethylsulfone (yes)	3449.12	243.4	-4.44	<5%
5	Sulfolane (no)	-	-	-	<5%
6	Dichloroethane (no)	-	-	-	<5%

sence of solvent in the calculations. In addition, our modeling approach for the Log(rate) relies on single-point vibrational analyses, which is incorrect for flexible chemical entities. In this respect, modeling the self-assembly with dimethyl oxalate and dimethyl malonate indicates that multiple structures (i.e. energy minima) may contribute to the OH-stretching frequency, which would require more extensive sampling of the potential energy surface. Finally, the identification of short versus long H-bonds, which correspond to different acidities (see above), is not straightforward in irregular and/or non-rectangular H-bonded networks and may introduce a systematic error in the determination of the O-H stretching frequencies. A generalization of the current modeling approach to fix some of the shortcomings above is left for future studies.

6.3 MATERIAL AND METHODS

6.3.1 *Computational details*

In all cases, the initial geometries of the monomers were modeled using Avogadro[125]. Initial coordinates for the tetrameric assemblies were generated manually, optimized using the HF-3c semi-empirical method[126] in Orca 2.9[127], and finally refined at the DFT level of theory using the ωB97X − D/6-31G(d,p) functional in Gaussian09 [128]. Fully optimized structures for the supramolecular complex and monomers were used to compute IR frequencies (not scaled) by Hessian diagonalization in internal coordinates. In some cases, conver-

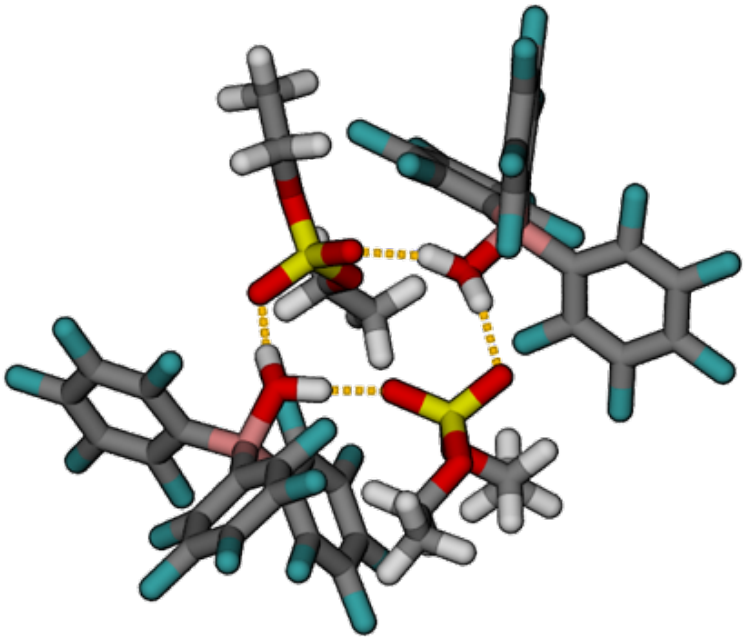


Figure 6.4: DFT-optimized structure for the 2:2 tetrameric self-assembly of BCF and diethylsulfate.

gence of the DFT optimization of the 2:2 aggregate was hindered by the flexibility of the nitrocompound, particularly the nitroalkanes (i.e. 1-nitrohexane, nitroethane and 2-methyl-2-nitropropane). Here, several optimization cycles using different starting structures were performed in order to identify a minimum in the potential energy surface.

6.3.2 *Modeling the self-assembly of BCF with nitromethane*

Quantum Chemistry methods were used to explore the co-catalytic role of nitro compounds in the acceleration of the azidation of tertiary alcohols by BCF. Considering the second-order dependence of the rate constant on the concentration of both BCF and nitro compound [123], we postulated that the self-assembly of two BCF and two nitro-compound molecules is responsible for the co-catalytic effect. Based on this assumption, atomistic models of the supramolecular adduct formed by the tetrameric association of BCF with nitromethane, i.e. the smallest nitro compound in *Dryzhakov et. al.*, 2015 [123], were built by connecting the water molecule hydrating the boron center in BCF with the nitro group of nitromethane via pairs of planar hydrogen bonds. The resulting square-shaped supramolecular adduct was submitted to geometry optimization at the DFT level of theory using the ωB97X $-$ D functional [124] with the 6-31G(d,p) basis set. However, due to bulky nature of the substituents in BCF and the large number of translational and rotational degrees of freedom to be optimized, this procedure resulted in poor convergence of the calculation. Thus, a step-wise approach was devised to obtain a fully optimized geometry for the BCF:nitromethane tetrameric complex. First, an initial guess of the tetrameric self-assembly was built using monohydrated tris(trifluomethyl)

boron (BMF) instead of BCF. The initial guess was optimized at the HF-3c semi-empirical level of theory[126] and then refined by ωB97X − D/6-31G(d,p) DFT optimization. The use of the sterically less hindrant BMF simplified the optimization of the hydrogen-bonding network, which converged successfully; i.e. no imaginary frequencies were found in the subsequent vibrational analysis. In addition, because BMF and BCF share similar boron-oxygen distances and O-H stretching frequencies in the calculated IR spectrum (data not shown), the optimized BMF:nitromethane (tetrameric) complex was a good starting point for modeling BCF self-assembly with nitromethane. By replacing BMF with BCF, a fully optimized 2:2 BCF/nitromethane complex was obtained at the ωB97X − D/6-31G(d,p) DFT level of theory *in vacuum*. The DFT-optimized geometry illustrates the formation of an almost flat, rectangular hydrogen-bonding network that is stabilized by two pairs of non-equivalent hydrogen bonds (Figure 6.2). One pair (hb1 = 1.77 Å and hb3 = 1.81 Å) is characterized by shorter distances between the BCF protons and the corresponding acceptor oxygens, while the other (hb2 = 1.98 Å and hb4 = 1.99 Å) by longer distances. In addition, DFT optimizations of BCF in isolation and the 1:1 complex with nitromethane were performed to obtain atomistic models of the monomeric state and the dimeric form. For the latter, two distinct minima were identified depending on the initial atomic coordinates of BCF and nitromethane. The most stable arrangement (by 0.78 kcal/mol) features a double H-bonding interaction between BCF and nitromethane (Figure B.2 of the *Appendix B*), the other involves a single hydrogen bond corresponding to a slightly lower OH-stretching frequency in the calculated IR spectrum (data not shown).

6.3.3 *Modeling the 2:2 self-assembly of BCF with nitro compounds*

To explore how the electron-rich character of the nitro compound affects the OH-stretching frequency in BCF, the 2:2 tetrameric assembly of BCF was modeled with 4-nitrobenzotrifluoride, nitrobenzene, 4-nitroanisole, 2-methyl-2-nitropropane, and 1-nitrohexane, which were previously shown to modulate the co-catalytic activity in azidation reactions[123]. DFT-optimized models were generated following the same procedure for nitromethane (see above). Remarkably, the tetrameric arrangement characteristic of the 2:2 self-assembly with nitromethane with pairs of non-equivalent hydrogen bonds was preserved in all complexes. In some cases, structural distortion of the network was observed, which is likely due to the presence of bulky substituents on BCF and/or the nitro compound.

6.3.4 *Vibrational analysis of BCF self-assembly with nitro compounds*

A DFT-based vibrational analysis of 2:2 self-assembly of BCF with nitro compounds was used to develop a model for the co-catalytic activity of the latter. DFT frequencies in vacuum indicate that the two sets of short and long H-bonds are characterized by distinct OH stretching signals in the calculated IR spectrum. In particular, the group that vibrates at a higher frequency corresponds to the stretching of the hydrogens involved in the longer H-bonds, whereas the one that vibrates at a lower frequency corresponds to the stretching of the hy-

drogens involved in the shorter H-bonds; see Figure B.1 for an illustration with nitromethane. In addition, each pair of H-bonds presents two distinct stretching modes: a symmetrical stretching where opposite protons move synchronously but in opposite direction (in-phase stretching), and an anti-symmetrical stretching where opposite protons move synchronously in the same direction (out-of-phase stretching). Notably, the antisymmetrical stretching of the short H-bonds corresponds to the highest IR intensity signal. The distinction of these four OH-stretching modes in the 2:2 self-assembly of BCF with nitro compounds is actually quite relevant. In fact, by using the experimental kinetic rates for the azidation of tertiary alcohols in the presence of nitrocompounds[123], we have found that the calculated frequency (DFT) of the anti-symmetrical stretching of the short H-bonds in the 2:2 aggregate is strongly correlated with the experimental Log(rate); see Figure 2 in the Main Text. In sharp contrast and for reasons that are not fully understood, the vibrational frequencies of the symmetrical stretching of the short H-bonds (hb1 and hb3) and the stretching frequencies of the longer H-bonds (hb2 and hb4) do not correlate with the experimental Log(rate) of azydation; confront the values of the determination coefficient in Table B.1 of the *Appendix B*. Also, vibrational analyses of the DFT optimized structures of BCF in isolation (Figure B.3 of the *Appendix B*) and in the 1:1 complex with nitromethane *in vacuum* (Figure B.2 of the *Appendix B*) indicate that dimeric association corresponds to a marginal red shift in the OH stretching frequency (i.e. of 23 cm-1), which is significantly smaller than the one predicted in the tetrameric network (i.e. of 192 cm-1). Based on these results we conclude that the 1:1 aggregate is unlikely to be the catalytic active specie in the azydation of tertiary alcohols.

6.3.5 *Validation of the DFT model to predict the Log(rate)*

The correlation between the frequency of the anti-symmetrical stretching of the short H-bonds and the experimental Log(rate) presented before (Figure 6.3) illustrates how the chemical nature of the nitro compound modulates the catalytic activity of BCF in the azidation of tertiary alcohols. In conjunction with modeling based on DFT, this correlation provides means to predict the co-catalytic power of the nitro compound. Here, we demonstrate the statistical robustness of the computational model for predicting the Log(rate) using both internal and external validation schemes.

6.3.5.1 *Internal validation*

To assess the validity of the DFT model presented in Figure 6.3, a leave-one-out (LOO) cross validation was carried out. In LOO cross-validation, each data-point is removed from the sample, linear regression is applied to the remaining data-points, and the resulting model is used to predict the Log(rate) of the left-out. Typically In LOO the statistical significance of the correlation is evaluated using the cross-validated squared correlation coefficient (q^2) and the standard deviation of prediction (SDEP). For validation purposes, a model is considered fully predictive if the value of q^2 is larger than 0.5[129]. The results in Table 6.2 indicate that the DFT and the experimental models present a positive statistical profile from an internal validation.

6.3.5.2 *External validation*

Good statistical parameters from LOO cross-validation is necessary but not sufficient to obtain reliable predictions[129]. To challenge the statistical robustness of the DFT model, an external validation was carried out. This more stringent test is based on the prediction of the Log(rate) for five additional nitro compounds, which were not part of the initial set used to build up the DFT model. The new set includes: 1-nitropropane, 2-nitropropane, nitroethane, 2-nitroanisole, and 1-nitronaphtalene. For these compounds the Log(rate) was predicted using the frequency of the anti-symmetrical stretching of the short H-bonds in the 2:2 aggregate optimized by DFT and compared with the experimental rates obtained by GC-MS from the Moran's Group at the University of Strasbourg. The results of the external validation show that kinetic rates predicted by the DFT model are strongly correlated with experiments with a determination coefficient. In addition to high q^2 in the LOO cross-validation, a model is deemed predictive if $R^2 > 0.6$, $(R^2 - R_0^2)/R^2 < 0.1$ and $0.85 < k < 1.15$; being k, R^2 and R_0^2 the slope and the determination coefficients of the linear regression with and without crossing the origin for the test set, respectively.

Finally, the predictive power of the DFT-model was challenged using a more rigorous validation scheme based on the randomization of the experimental kinetic rates for the twelve data-points in Table 6.1. For this purpose, three new pairs of training/test sets including eight and four nitro compounds, respectively, were generated as follows. After sorting compounds according to the experimental Log(rate) in descending order (i.e. from the slowest to the fastest), three new "test" sets were obtained by picking up one compound every three starting from number one, number two, or number three of the list and sending the rest to the "training" set. For each training set, a new model was generated and used to predict the Log(rate) for the corresponding test set. Statistical parameters of the three resulting models (named 1C-3C) are given in Table B.2 of the *Appendix B*. This analysis shows that in all cases the correlations between predicted and observed Log(rate) present positive statistical profiles independently of the training/test set, which demonstrates the absence of random correlations.

6.3.6 *Experimental IR model for Log(rate)*

In addition to the DFT model, an experimental model for predicting the Log(rate) was built by correlating the kinetic rate of azidation with the OH-stretching frequency estimated from the experimental IR spectra. These frequencies were obtained using the maximum peak in the OH stretching band of the IR spectrum recorded from a solution containing BCF and nitro compounds at room temperature. The predictive power of the IR model was assessed using the same internal and external validation schemes (Table 6.1). The experimental IR model presents good statistical parameters for the internal validation (training set) as shown in Table 6.2. However, its predictivity for the five additional nitro compounds (test set) is questionable, as indicated by the striking anti-correlation between predicted and observed Log(rate) for these compounds Table 6.2. The same tendency is also observed upon scrambling training/test sets to produce models 1E – 3E, which show poor statistical profiles; see

Table B.3 of the *Appendix B*. Therefore, we conclude that the experimental IR model is not suitable for predicting the Log(rate) of the azidation by BCF in the presence of nitro compounds. The comparison with the computational results (DFT model) suggests that the limited predictive power of the IR model may be attributed to the lower resolution of the experimental IR spectrum, whose broad bands result from the superimposition of many overlapping modes. Since, the DFT results show that the experimental Log(rate) is strongly correlated with the frequency of the anti-symmetrical stretching of the short H-bonds in the 2:2 aggregate, the lack of resolution in the experimental IR spectrum severely limits the predictive power of the model. Interestingly, the DFT vibrational analyses of the 2:2 aggregates indicate that the anti-symmetrical stretching of the short H-bonds, which is predictive for the Log(rate), has the highest intensity signal in the presence of nitro aromatic compounds (3/5 in the training set) but not with nitro aliphatic derivatives (3/5 in test set), which may explain the predictive character of the IR model for the training set but not the test set.

6.3.7 *Modeling the self-assembly of BCF with non-nitro compounds*

The striking correlation between the calculated OH-stretching frequency in the 2:2 aggregate and the experimental rate of azidation (Figure 6.3) provides computational evidence that the accelerating effect on the catalytic activity of the Bronsted acid BCF is related to the formation of 2:2 supramolecular aggregates. Based on this conclusion, we explored the possibility to identify new co-catalysts by searching for chemical entities able to mimic the nitrocompounds in the formation of a H-bonded tetrameric networks with BCF. To this aim, we modeled the 2:2 self-assembly of BCF with six chemically distinct compounds, i.e. dichloroethane, dimethyl sulfone, diethyl sulfate, sulfolane, dimethyl oxalate, and dimethyl malonate, using the procedure described for nitromethane (see above). At first, the stability of the tetrameric network with BCF was used as a criterion to evaluate the suitability of the compound as a promoter. Then, the co-catalytic power was quantified by isolating the vibrational frequency of the anti-symmetrical stretching of the short H-bonds and introducing it into the DFT model for the Log(rate) developed for nitro compounds. The DFT results show that the tetrameric networks of BCF in complex with dichloroethane and sulfolane are not stable and quickly disassemble; self-assembly was considered as stable if the main features of the 2:2 aggregate (i.e. the rectangular shape of the network with four H-bonds characterized by distinct OH-stretching frequencies) were preserved. In sharp contrast, the tetrameric networks formed by BCF in complex with diethyl sulfate and dimethyl sulfone were stable; see Figures 6.3 and B.6, respectively.

Finally, the self-assembly with dimethyl oxalate and dimethyl malonate require more attention. In fact, although the DFT-optimized tetrameric networks correspond to a minimum of the potential energy surface (Figure B.6), the association of BCF with these compounds produced hexagonal rather than rectangular arrangements, which makes the identification of the short H-bonds more challenging. Based on the stability of the DFT-optimized 2:2 aggregates and corresponding vibrational analyses, we predict: (1) significant co-catalytic activity for diethyl-sulfate (comparable to nitro-methane) and dimethyl-sulfone (comparable to 4-

nitroanisole); (2) marginal co- catalytic activity for dimethyl oxalate and dimethyl malonate (comparable to 1-nitrohexane); and (3) no activity for dichloroethane and sulfolane. Strikingly, the predictions for diethyl-sulfate, dimethyl-sulfone and dimethyl malonate are in qualitative agreement with experiments; see Table 1. The incorrect prediction for dimethyl-sulfone points to possible limitations of the model, which is unable to distinguish between sulphates and sulfones.

6.4 CONCLUSION

In summary, DFT modeling of 2:2 H-bonded aggregates of nitro compounds and BCF returns computed O-H stretching frequencies that are predictive of the experimentally observed rates in reactions that use the corresponding nitro compound as promoter. In contrast, a similar model constructed from experimental IR data was found to be not predictive for the reaction rate, which demonstrates the need for modeling. Also, the structural model of the aggregate was used to identify sulfate esters as a new class of promoters, which was verified experimentally a posteriori. The strong correlation between experiments and calculations suggests that such aggregates couldindeed be involved in the reaction mechanism, a conclusion that is consistent with prior studies on kinetic concentration dependence of Brønsted acid catalyzed reactions.8 More broadly, the important implication of this study is that a deep understanding of Brønsted acid catalysis requires consideration of not only the molecular structure and pKa of the Brønsted acid, but also the nature of the supramolecular environment that takes into account the weak interactions between all molecules in the reaction mixture.

Part III

CLÔTURE ET PERSPECTIVES FUTURES

7

CONCLUSIONS AND PERSPECTIVES

With the fast growth of technological tools for being used in science nowadays, we are able to study natural phenomena in a much precise and efficient way. These technological advances allow the design of experiments with an increased resolution and highly accuracy in the measurements with a drastic reduction of random uncertaints by noise in the equipment. Among these revolutionary techniques we have as example the Single molecule fluorescence resonance energy transfer(smFRET), scanning electron and cryo-electron microscopies, x-ray crystallography, nuclear magnetic resonance (NMR) and supercomputers, among others. All these techniques aim to study matter at increasingly smaller grade, ranging from the micro to the nano scales. The fundamental principle behind this fact states that a better understanding of matter at the atomistic level, as well as the laws governing at that scale, will give a deeper insight for solve current questions in natural sciences. In consequence, much more knowledge will be provided for designing new technologies and the development of better applications. Thus, the greater understanding of phenomena at the atomic level, the greater the progress of natural sciences.

Special attention must be paid to the advances in the development of supercomputers that have boosted the use of molecular simulations to study natural phenomena. Nowadays it's possible to simulate realistic systems of hundreds nanometers (as virus) at the atomistic level in the microsecond scale with relative facility. This has made possible thanks to the use of recent and inexpensive graphical processor units (GPUs), although they were originally designed for other purposes. Also, with the birth of ANTON, a supercomputer optimized to develop molecular dynamics (MD) simulations, it has been possible to reach the scale of milliseconds in simulation time. Despite the theory behind MD simulations is known since 18th (i.e., the fundamental laws of motion by Newton), these calculations could not even be conceived at the beginning of the 19th century for complex and relevant systems.

Along this line, molecular recognition, a fundamental problem in biology, chemistry and physics, has been one of the most benefited from the use of new and revolutionary technologies. Many open questions on molecular recognition can be treated today thanks to the use of atomic-scale techniques as smFRET, x-ray crystallography and molecular simulations. In fact, very little advances in the field of molecular recognition could be done without these advanced techniques. The exponential progress done in the understanding of this phenomena is evidenced from the middle of the 20th century since these advanced biophysical methods have emerged. Molecular recognition is not fully understood today, in fact, we could say that we have a semi-quantitative perspective of the problem. Then, studies aimed to get a more quantitative picture of molecular recognition are highly in demand.

In this sense, the present manuscript provided results to address, in a quantitative fashion, some molecular recognition events with chemical and biological relevance as binding reactions of host-guest and protein-ligand systems, and catalysis in solution. The methodolo-

gies presented here were not exclusively limited to compute or reproduce some experimental quantities for the studies systems as a validation scheme but also they were used for making reliable predictions on new and challenge systems, which were not experimentally tested before. In fact, they were used for guiding new experiments and gain insights of the particular phenomena in study. Furthermore, the quantitative methods presented here significantly might boost the field on molecular design, specially in the development of novel chemical entities as drugs, synthetic hosts/guests and catalysts.

In the first scenario, we have performed a theoretical analysis of apparently unrelated computational approaches to protein-ligand binding in the common framework of statistical mechanics. Our comparative approach allowed to pinpoint the approximations that are introduced to speed up the calculations, which is useful to rationalize their impact on the accuracy of the binding affinity predictions. As a perspective, our comparative analysis highlights possible improvements to well established semi-rigorous and empirical scoring strategies and will hopefully help in the development of variants with an optimum balance between accuracy and efficiency at each stage of the drug-discovery pipeline. Also, our analysis invites to the development of more reliable methods for computing solvation free energy, being the main source of systematic errors in many wel-known binding approaches as MM/PBSA.

In our second case, we built a LIE model for predict binding affinities in cavitand host-guest systems, a less explored computational approach for these relevant systems. This model turned out to be accurate (RMSE $<$ 1.5 kcal/mol), efficient and transferable among chemically diverse families of host-guest complexes. Also, it has been shown that the predictions given by LIE are independent of the training and test sets used for building and validate the model, respectively. Furthermore, the accuracy of the model does not dependent of the energy model (force field) used but it's parametrization does. As a remarkable note, we have presented a new LIE model which consider the strain energy of the host for predicting binding affinities, which is missing in the classical LIE formulation. The success for this new LIE model was evidenced in the the reliable prediction (RMSE $<$ 1.0 kcal/mol) of binding affinities for the challenge cucurbituril-steroid complexes, where the hosts undergo conformational changes upon the ligands binding. However, the LIE model also reported fails predictions for some other difficult cases, specially for those guests with ultra-high binding affinity for their hosts. As a perspective, we have proposed to build a LIE model using a polarizable force fields as a correction approach to these faults.

In the third case, the co-catalytic effect of the "inhert" nitro compound in the azidation reaction of tertiary alcohols catalyzed by BCF, an organoborone Bronsted acid, was analyzed by a computational approach based on DFT calculations. Our DFT model revels a correlation between the red-shift of the anti-symmetric O-H stretching frequencies for BCF and the experimental kinetic rate for the reaction in presence of nitro compounds. The DFT model proved to be predictive, passing several statistical validation tests, in contrast to a similar built based on experimental IR frequencies with low resolution. This DFT model provided a quantitative insight for explore other chemical entities with the same co-catalytic effect as nitro compounds in the azidation reaction. Thus, the DFT model predicted that sulfate esters will be a new family of promotors for the azidation reaction under study, which was experimen-

tally confirmed further. Nevertheless, our DFT model was not able to do good predictions in all tested cases, giving room for further improvements. In fact, the introduction of solvation effects in the modeling could give a more robust model. As a perspective, and complementing our DFT model, we will study the full reaction mechanism using other computational methodologies In order to gain a better understanding of the catalytic/co-catalytic role.

In conclusion, we have presented robust models with tested accuracy and pragmatic spirit, which can be easily applied to study other problems of molecular recognition which involve receptor-ligand binding and catalysis in solution. The results obtained in this manuscript revel significant contributions toward the definitive solution of the molecular recognition. Of course, advances in other fields of science such as the increase computational power or improvements of the energy models for molecular simulations, will also contribute to the final solution. There is still a long way to go towards the full understanding of molecular recognition. This is still a challenge but a very beautiful one. This further confirms that...

Nature is fascinating!!

.

Part IV

APPENDIX

APPENDIX A

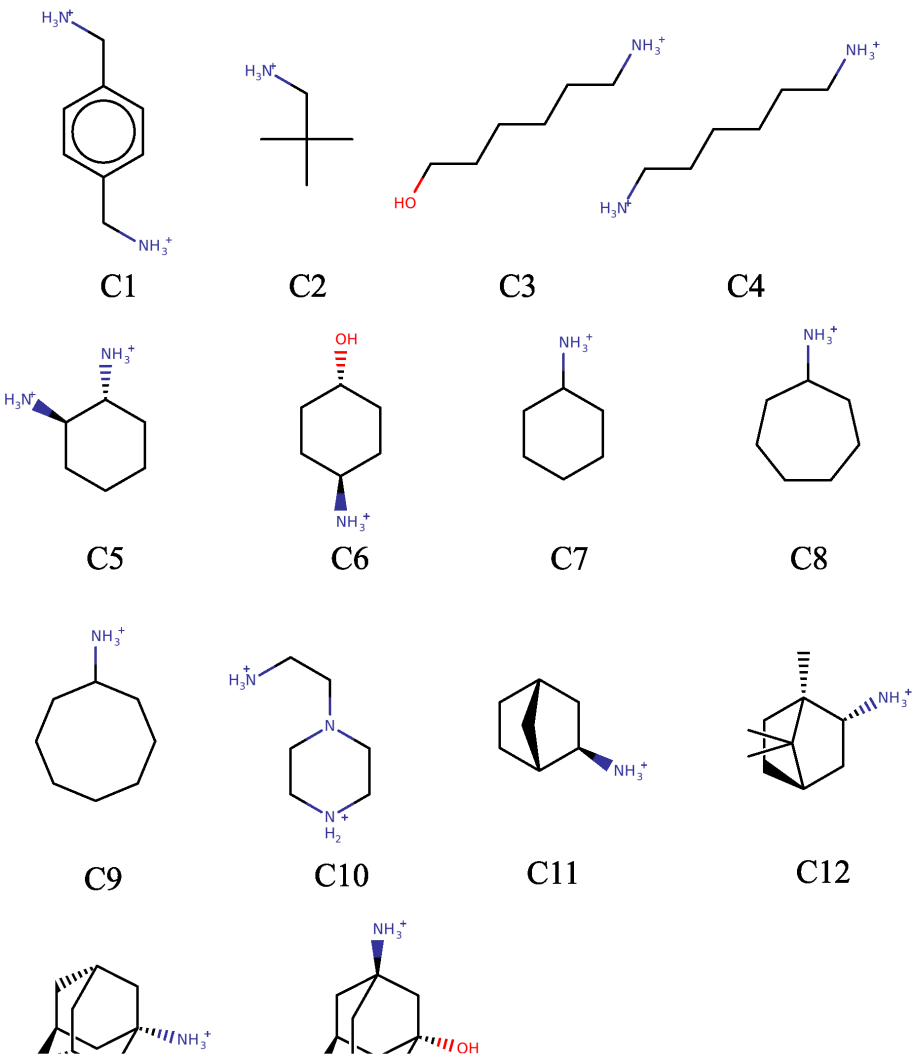


Figure A.1: Chemical structures of CB7 guests used as a *training* set, shown in their protonated states used in the computations to produce the LIE models.

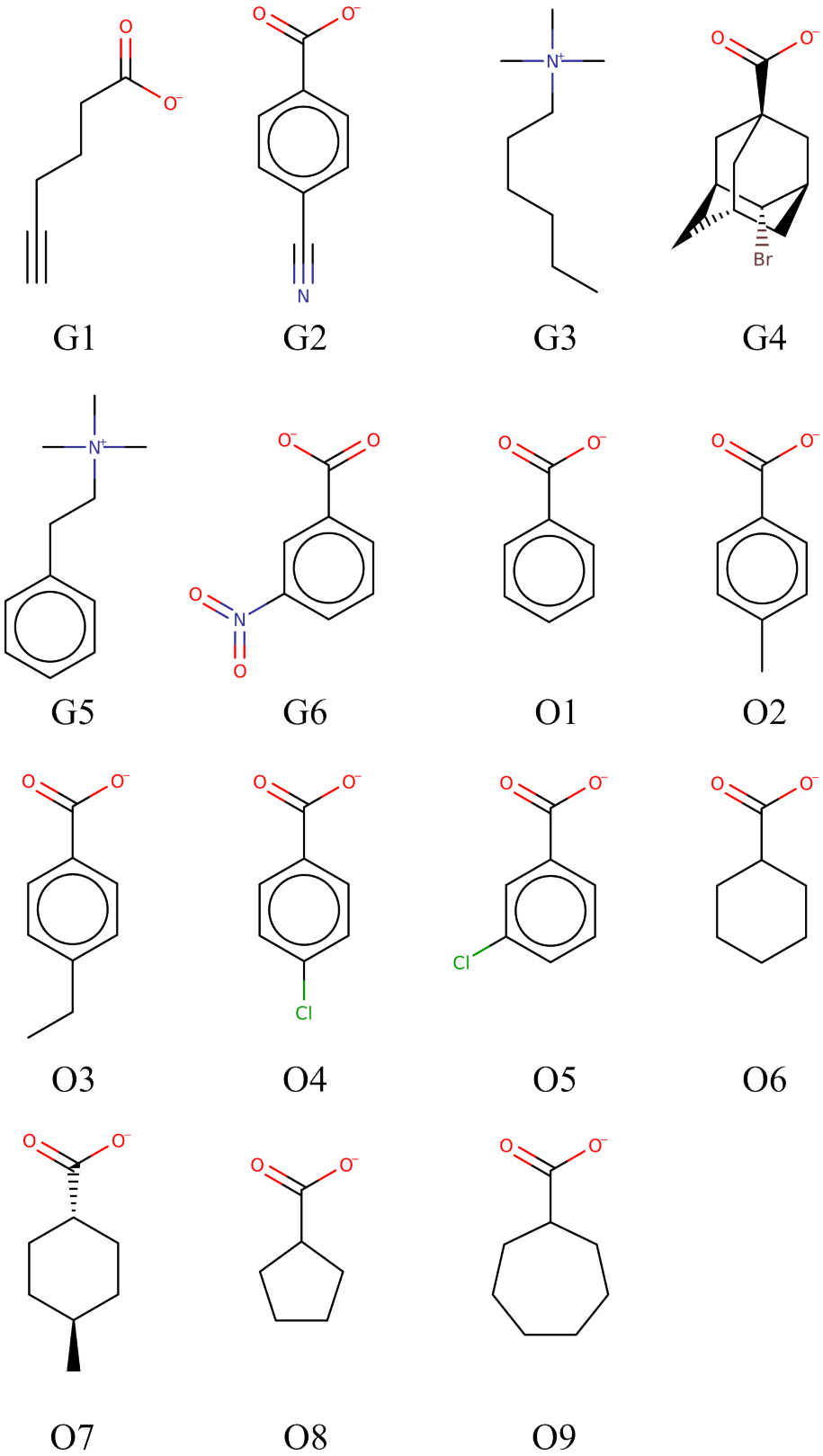
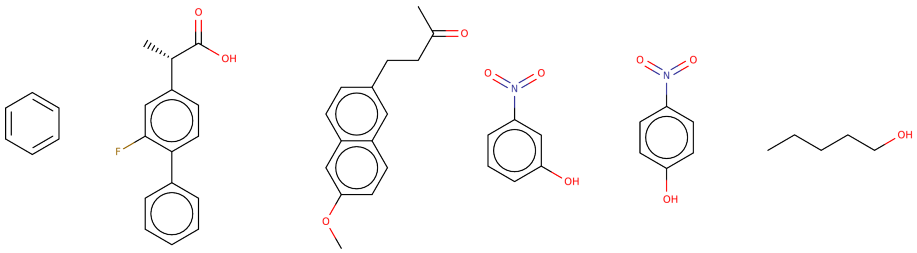


Figure A.2: Chemical structures of the 15 ligands for octa-acid, OAH, (Gxx and Oxx ligand families reported in references Yin *et al*[17] and Muddana *et al*,[16] respectively) and the 6 ligands for the tetramethylated octa-acid, OAM, (Gxx[17] ligand family), which were part of the *test* set. The guests are shown in the protonation form used in the calculations.



B1 B2 B3 B4 B5 B6

Figure A.3: Chemical structures of the six β-cyclodextrin guests that were part of the **test** set. The guests are shown in the protonation form used in the calculations.

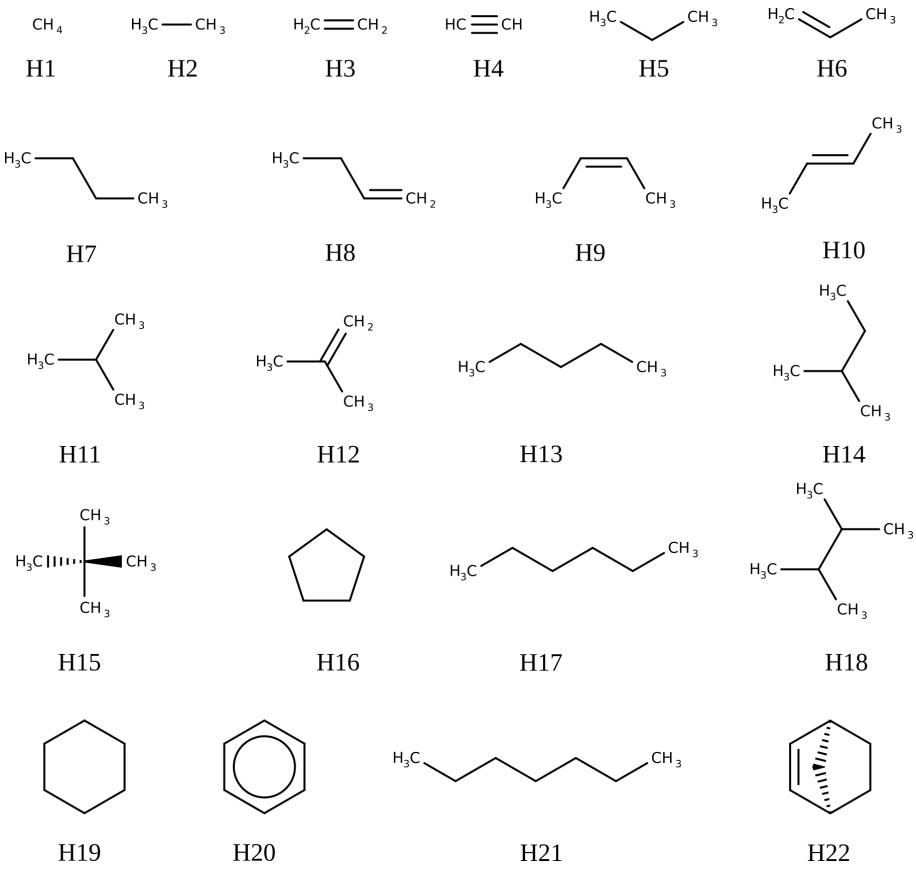


Figure A.4: Chemical structures the 22 cucurbit-7-uril (CB7) guests that were part of the **test** set. All these guests are neutral and belong to the HYDROPHOBE challenge.[20]

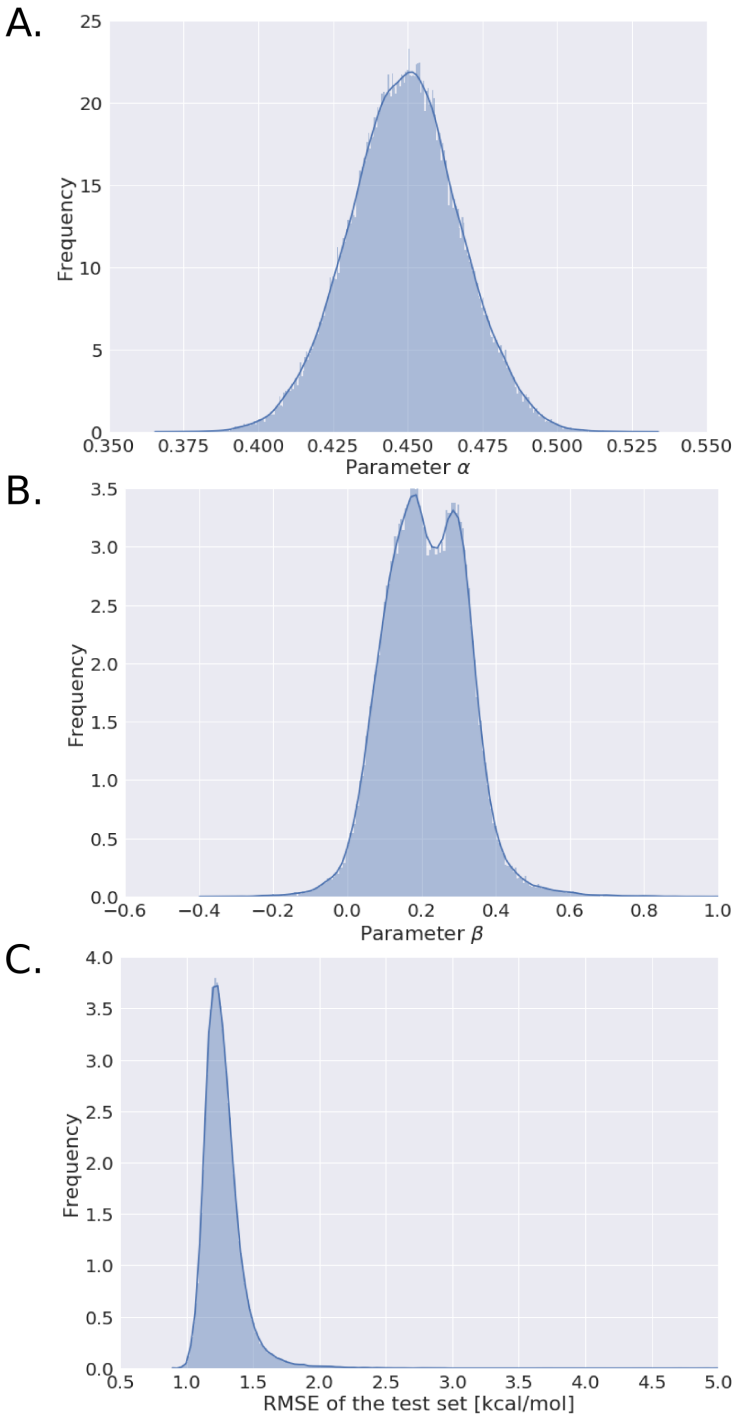


Figure A.5: Frequency distributions of the GAFF/LIE parameters α (A), β (B), and the RMSE for the test set (C) upon splitting of the full data set into 1×10^5 random training/test sets.

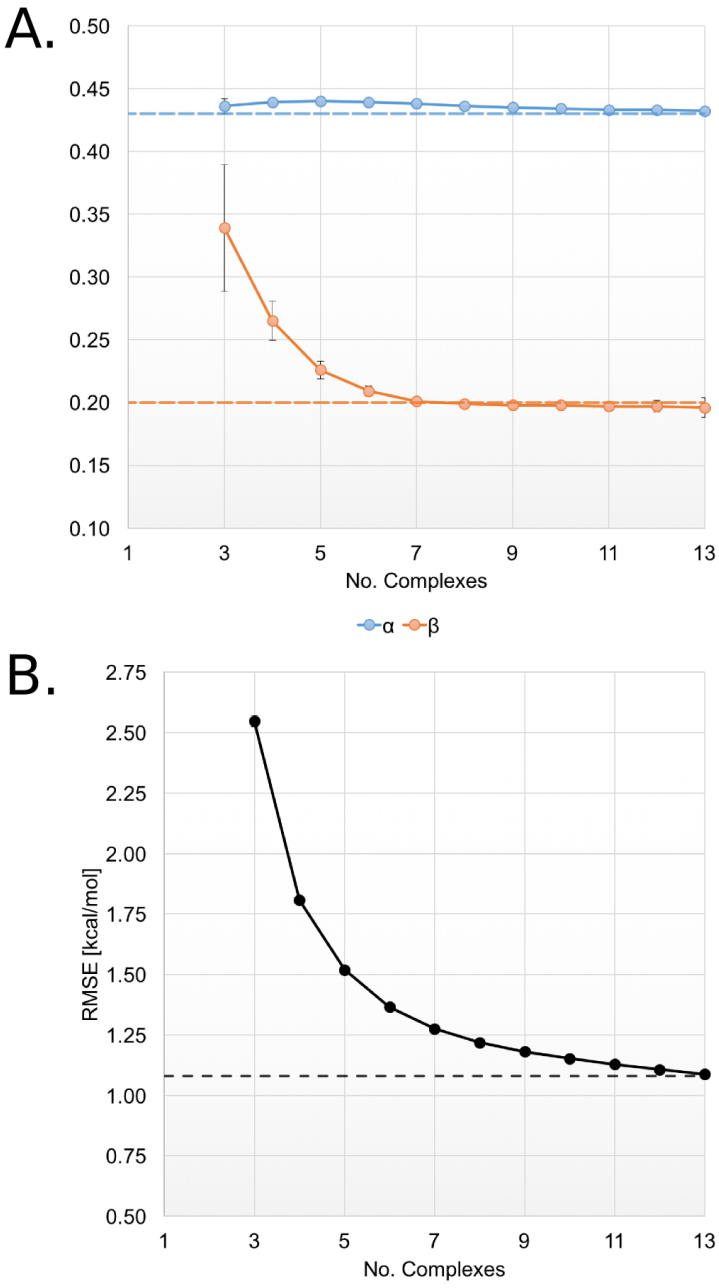


Figure A.6: Average values of the LIE parameters (A) and the RMSE for the test set (B) after removing $k = 1, 2, 3 \ldots 11$ members from the training set. The dashed lines represent the values for α, β and RMSE for the test set obtained using the original (n=14) training set.

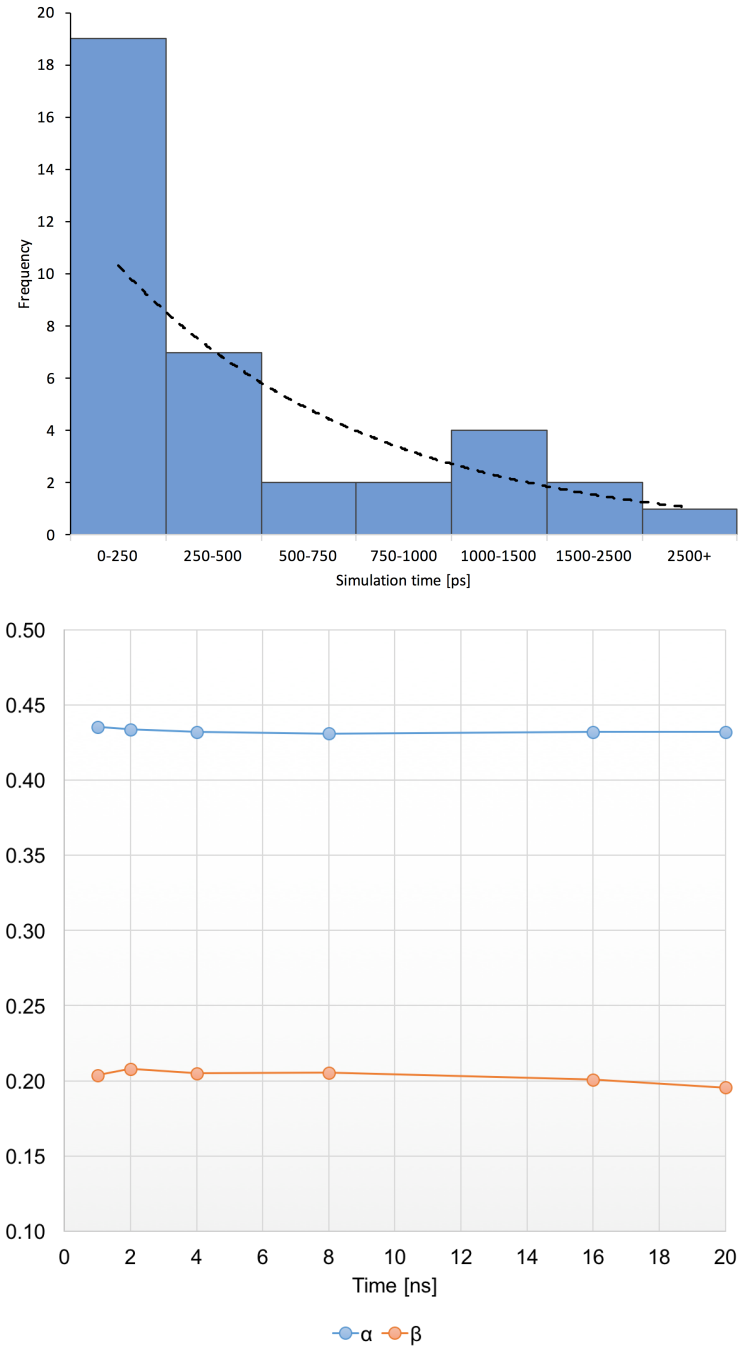


Figure A.7: On top, the frequency distribution for t_{min} values computed for the **test** set using GAFF. On bottom, convergence analysis for parameters of the GAFF/LIE model.

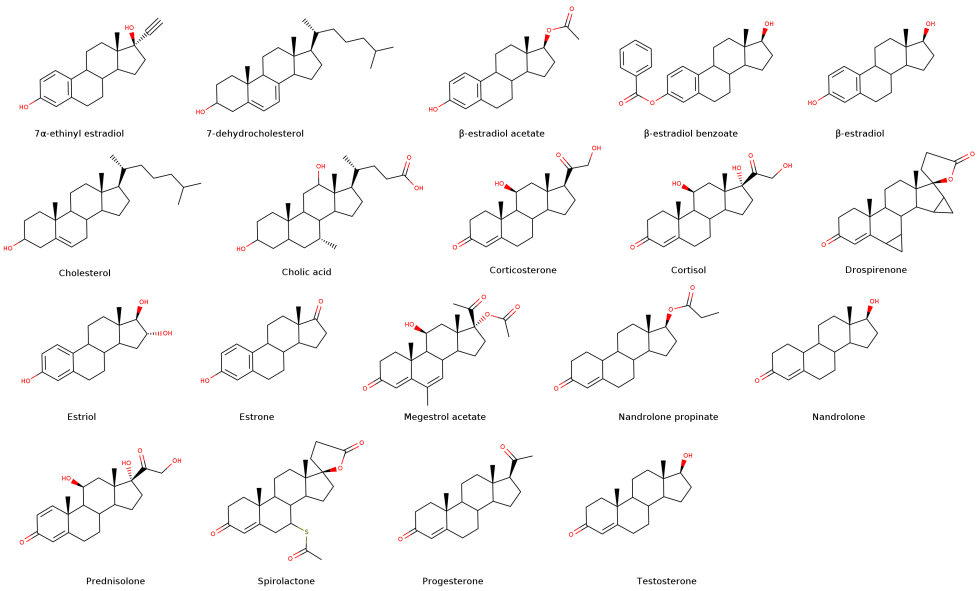


Figure A.8: Chemical structures of *steroid compounds* that bind to CB[7/8] used in this study.

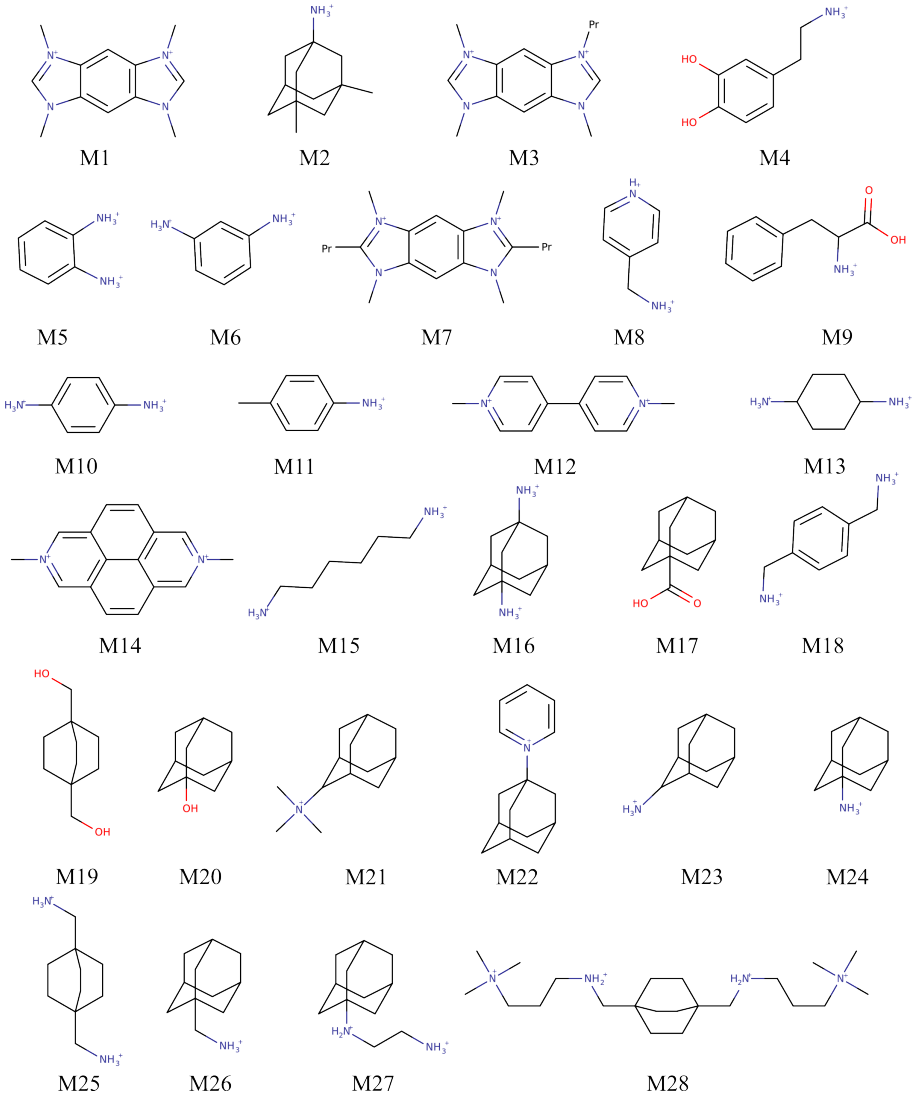


Figure A.9: Chemical structures of the CB7 guests from the *Muddana set*, shown in their protonated states used in the computations.

B

APPENDIX B

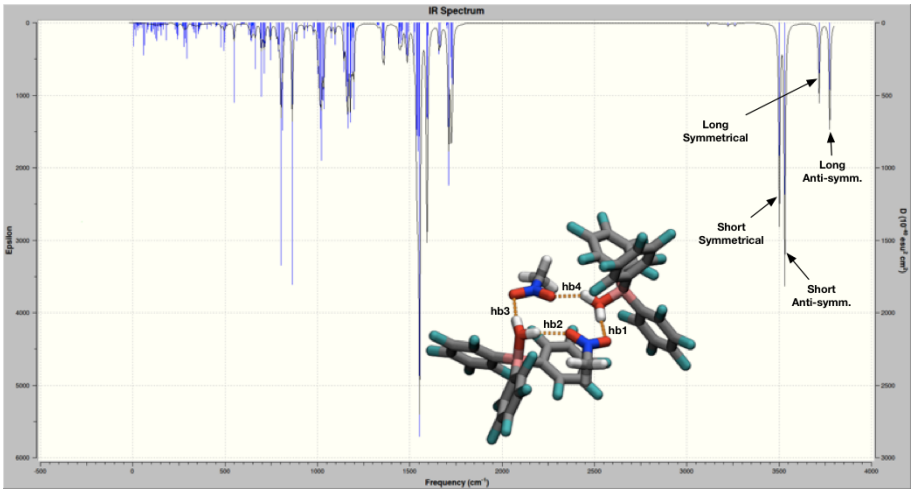


Figure B.1: Calculated IR spectrum for the tetrameric complex BCF:nitromethane. Labels and arrows indicate the O-H stretching modes for each group of hydrogen bonds. The DFT results show that the short H-bonds (hb1 and hb3) vibrate at 3501 cm-1 (symmetrical) and 3531 cm-1 (anti-symmetrical), whereas the longer H-bonds at 3717 cm-1 (symmetrical) and 3774 cm-1 (anti-symmetrical).

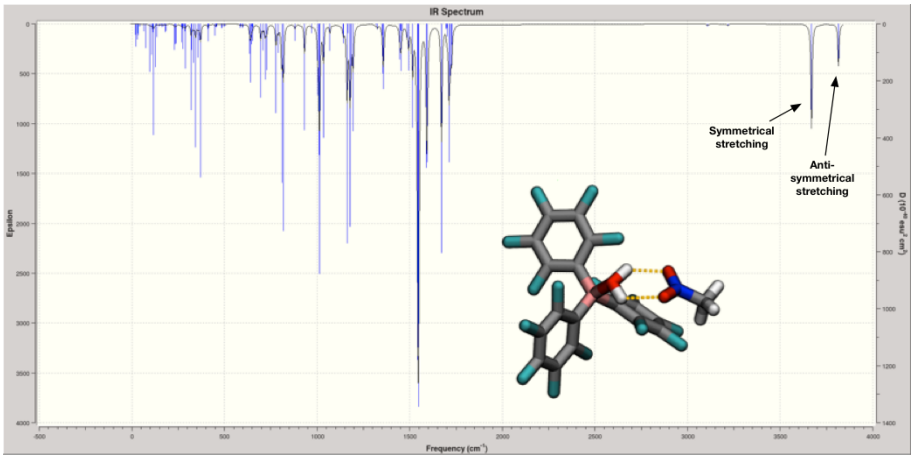


Figure B.2: Calculated IR spectrum for the most stable dimeric complex BCF:Nitromethane. Labels and arrows indicate the hydrogen stretching modes for each group of hydrogen bond.

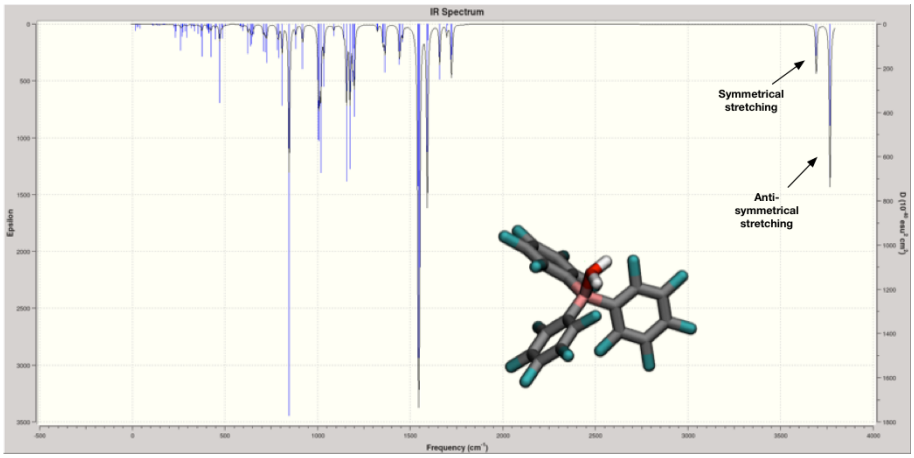


Figure B.3: Calculated IR spectrum for the BCF isolated. Labels and arrows indicate the hydrogen stretching modes.

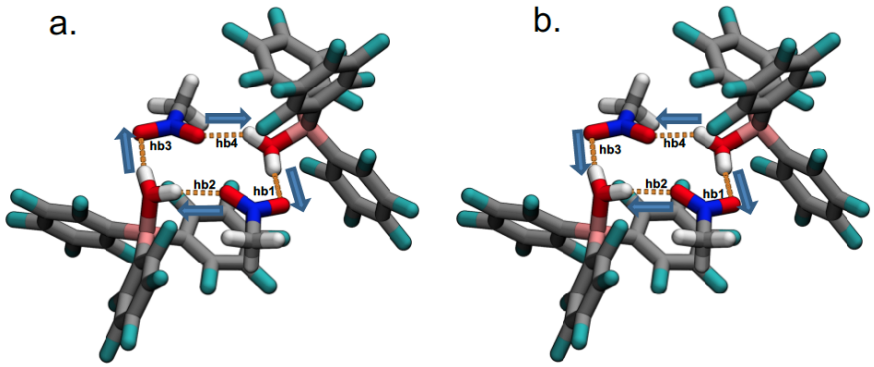


Figure B.4: Definition of the type of vibrational modes for the OH-stretching frequencies using the BCF:nitromethane complex as model. Vector displacements (blue arrows) for the symmetrical (a) and the anti-symmetrical OH-stretching modes (b).

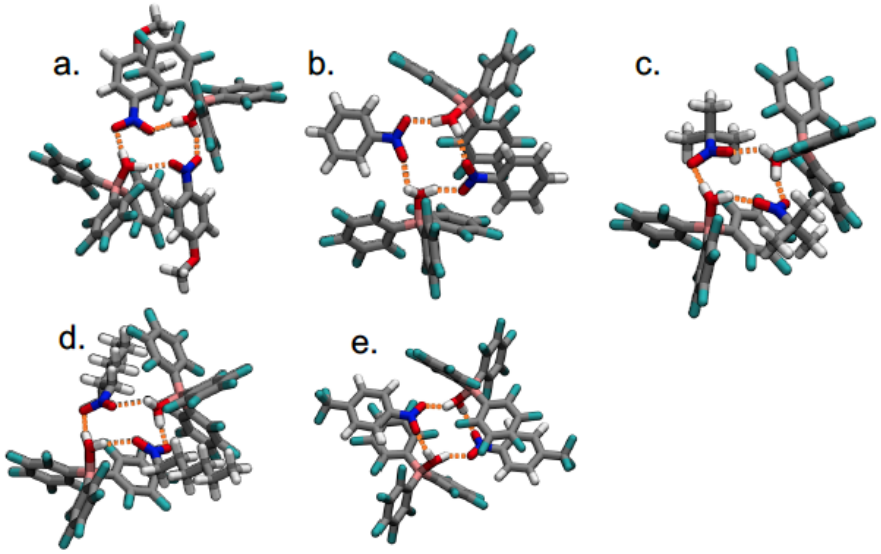


Figure B.5: Optimized structures for the 2:2 self-assembly of BCF with: (a) 4-nitroanisole; (b) nitrobenzene; (c) 2-methyl-2-nitropropane; (d) nitrohexane; and (e) nitrobenzotrifluoride.

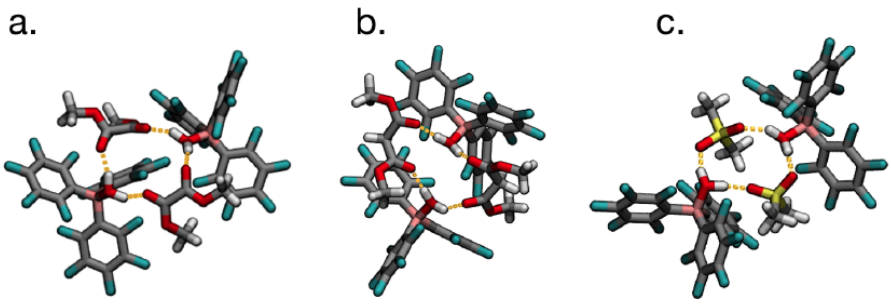


Figure B.6: DFT-optimized structures for the 2:2 self-assembly of BCF with: (a) dimethyl oxalate; (b) dimethyl malonate; and (c) dimethyl sulfone. These compounds correspond to entries 2-4 in Table 6.1.

Table B.1: Statistical parameters for the correlation between the experimental Log(rate) and the DFT vibrational frequency of each of the four OH-stretching modes predicted in the BCF:nitrocompound tetrameric self-assembly *in vacuum*.

Parameter	DFT OH-stretching frequencies			
	Short symmetrical	Short anti-symmetrical	Long symmetrical	Long anti-symmetrical
r^2	0.00	0.81	0.13	0.09
m	-0.00002	-0.00358	0.00199	-0.00124
b	-4.724	7.891	-12.204	-0.174

Table B.2: Statistical parameters for both internal and external validation of three DFT models (1C – 3C) generated by randomizing training/test sets in the list of eleven nitro compounds.

Parameter	Model 1C	Model 2C	Model 3C
r^2	0.78	0.79	0.76
m	-0.00386	-0.00328	-0.00295
b	8.9298	6.8527	5.6747
q^2	0.53	0.38	0.61
SDEP	0.06	0.07	0.05
R^2	0.83	0.74	0.91
R_0^2	0.8	0.63	-0.01
$(R^2 - R_0^2)/R^2$	0.04	0.16	1.01
k	0.99	1	1.01

Table B.3: Statistical parameters for both internal and external validation of three IR models (1E – 3E) generated by randomizing training/test sets in the list of eleven nitro compounds.

Parameter	Model 1E	Model 2E	Model 3E
r^2	0.54	0.17	0.09
m	-0.00813	-0.00478	-0.00336
b	23.6387	11.97059	7.04388
q^2	-0.06	-1.23	-1.12
SDEP	0.09	0.13	0.14
R^2	0.14	0.56	0.83
R_0^2	-1.62	-1.79	-2.43
$(R^2 - R_0^2)/R^2$	12.25	4.2	3.93
k	1.03	0.99	0.97

BIBLIOGRAPHY

[1] D. J. Cram. "The design of molecular hosts, guests, and their complexes (nobel lecture)". *Angewandte Chemie International Edition in English*, 27(8), pp. 1009–1020, 1988.

[2] J.-M. Lehn. "Supramolecular chemistry—scope and perspectives molecules, supermolecules, and molecular devices (nobel lecture)". *Angewandte Chemie International Edition in English*, 27(1), pp. 89–112, 1988.

[3] C. J. Pedersen. "The discovery of crown ethers (noble lecture)". *Angewandte Chemie International Edition in English*, 27(8), pp. 1021–1027, 1988.

[4] B. A. Parviz, D. Ryan, and G. M. Whitesides. "Using self-assembly for the fabrication of nano-scale electronic and photonic devices". *IEEE transactions on advanced packaging*, 26(3), pp. 233–241, 2003.

[5] P. A. M. Dirac. *The principles of quantum mechanics*. 27. Oxford university press, 1981.

[6] A. Stone. *The theory of intermolecular forces*. OUP Oxford, 2013.

[7] J.-M. Lehn. "From supramolecular chemistry towards constitutional dynamic chemistry and adaptive chemistry". *Chemical Society Reviews*, 36(2), pp. 151–160, 2007.

[8] T. L. Hill. *Cooperativity theory in biochemistry: steady-state and equilibrium systems*. Springer Verlag, 1985.

[9] T. L. Hill. *An introduction to statistical thermodynamics*. Courier Corporation, 2012.

[10] A. Levitzki and D. Koshland. "Negative cooperativity in regulatory enzymes". *Proceedings of the National Academy of Sciences*, 62(4), pp. 1121–1128, 1969.

[11] D. Chatterji. *Basics of molecular recognition*. CRC Press, 2016.

[12] K. Ariga, H. Ito, J. P. Hill, and H. Tsukube. "Molecular recognition: from solution science to nano/-materials technology". *Chemical Society Reviews*, 41(17), pp. 5800–5835, 2012.

[13] S. J. Blanksby and G. B. Ellison. "Bond dissociation energies of organic molecules". *Accounts of chemical research*, 36(4), pp. 255–263, 2003.

[14] J. W. Steed, J. L. Atwood, and P. A. Gale. *Definition and emergence of supramolecular chemistry*. Wiley Online Library, 2012.

[15] D. L. Mobley and M. K. Gilson. "Predicting binding free energies: Frontiers and benchmarks". *Annual Review of Biophysics*, 46, pp. 531–558, 2017.

[16] H. S. Muddana, A. T. Fenley, D. L. Mobley, and M. K. Gilson. "The sampl4 host–guest blind prediction challenge: an overview". *Journal of computer-aided molecular design*, 28(4), pp. 305–317, 2014.

[17] J. Yin, N. M. Henriksen, D. R. Slochower, M. R. Shirts, M. W. Chiu, D. L. Mobley, and M. K. Gilson. "Overview of the sampl5 host–guest challenge: Are we doing better?" *Journal of computer-aided molecular design*, 31(1), pp. 1–19, 2017.

[18] A. Rizzi, S. Murkli, J. N. McNeill, W. Yao, M. Sullivan, M. K. Gilson, M. W. Chiu, L. Isaacs, B. C. Gibb, D. L. Mobley, *et al.* "Overview of the sampl6 host-guest binding affinity prediction challenge". *bioRxiv*, p. 371724, 2018.

[19] Z. Gaieb, S. Liu, S. Gathiaka, M. Chiu, H. Yang, C. Shao, V. A. Feher, W. P. Walters, B. Kuhn, M. G. Rudolph, *et al.* "D3r grand challenge 2: blind prediction of protein–ligand poses, affinity rankings, and relative binding free energies". *Journal of computer-aided molecular design*, 32(1), pp. 1–20, 2018.

[20] K. I. Assaf, M. Florea, J. Antony, N. M. Henriksen, J. Yin, A. Hansen, Z.-w. Qu, R. Sure, D. Klapstein, M. K. Gilson, S. Grimme, and W. M. Nau. "Hydrophobe challenge: A joint experimental and computational study on the host–guest binding of hydrocarbons to cucurbiturils, allowing explicit evaluation of guest hydration free-energy contributions". *The Journal of Physical Chemistry B*, 121(49), pp. 11144–11162, 2017.

[21] S. J. Barrow, S. Kasera, M. J. Rowland, J. del Barrio, and O. A. Scherman. "Cucurbituril-based molecular recognition". 2015.

[22] M. V. Rekharsky and Y. Inoue. "Complexation thermodynamics of cyclodextrins". *Chemical reviews*, 98(5), pp. 1875–1918, 1998.

[23] S. Geschwindner, J. Ulander, and P. Johansson. "Ligand binding thermodynamics in drug discovery: still a hot tip?" *Journal of medicinal chemistry*, 58(16), pp. 6321–6335, 2015.

[24] W. H. Ward and G. A. Holdgate. "Isothermal titration calorimetry in drug discovery". In "Progress in medicinal chemistry", vol. 38, pp. 309–376. Elsevier, 2001.

[25] D. McQuarrie. *Statistical Mechanics*. New York: Harper and Row, 1976.

[26] J.-P. Changeux and S. J. Edelstein. "Allosteric mechanisms of signal transduction". *Science*, 308(5727), pp. 1424–1428, 2005.

[27] F. Feixas, S. Lindert, W. Sinko, and J. A. McCammon. "Exploring the role of receptor flexibility in structure-based drug discovery". *Biophysical chemistry*, 186, pp. 31–45, 2014.

[28] C. Chipot. "Frontiers in free-energy calculations of biological systems". *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 4(1), pp. 71–89, 2014.

[29] P. Kollman. "Free energy calculations: applications to chemical and biochemical phenomena". *Chemical reviews*, 93(7), pp. 2395–2417, 1993.

[30] S. M. Paul, D. S. Mytelka, C. T. Dunwiddie, C. C. Persinger, B. H. Munos, S. R. Lindborg, and A. L.

Schacht. "How to improve r&d productivity: the pharmaceutical industry's grand challenge". *Nature reviews Drug discovery*, 9(3), pp. 203–214, 2010.

[31] S. Morgan, P. Grootendorst, J. Lexchin, C. Cunningham, and D. Greyson. "The cost of drug development: a systematic review". *Health Policy*, 100(1), pp. 4–17, 2011.

[32] N. Homeyer, F. Stoll, A. Hillisch, and H. Gohlke. "Binding free energy calculations for lead optimization: assessment of their accuracy in an industrial drug design context". *Journal of chemical theory and computation*, 10(8), pp. 3331–3344, 2014.

[33] W. L. Jorgensen, J. K. Buckner, S. Boudon, and J. Tirado-Rives. "Efficient computation of absolute free energies of binding by computer simulations. application to the methane dimer in water". *The Journal of chemical physics*, 89(6), pp. 3742–3746, 1988.

[34] J. Hermans and L. Wang. "Inclusion of loss of translational and rotational freedom in theoretical estimates of free energies of binding. application to a complex of benzene and mutant t4 lysozyme". *Journal of the American Chemical Society*, 119(11), pp. 2707–2714, 1997.

[35] S. Boresch, F. Tettinger, M. Leitgeb, and M. Karplus. "Absolute binding free energies: a quantitative approach for their calculation". *The Journal of Physical Chemistry B*, 107(35), pp. 9535–9551, 2003.

[36] M. K. Gilson, J. A. Given, B. L. Bush, and J. A. McCammon. "The statistical-thermodynamic basis for computation of binding affinities: a critical review." *Biophysical journal*, 72(3), p. 1047, 1997.

[37] W. L. Jorgensen. "Free energy calculations: a breakthrough for modeling organic chemistry in solution". *Accounts of Chemical Research*, 22(5), pp. 184–189, 1989.

[38] H.-J. Woo and B. Roux. "Calculation of absolute protein–ligand binding free energy from computer simulations". *Proceedings of the National Academy of Sciences of the United States of America*, 102(19), pp. 6825–6830, 2005.

[39] M. S. Lee and M. A. Olson. "Calculation of absolute protein-ligand binding affinity using path and endpoint approaches". *Biophysical journal*, 90(3), pp. 864–877, 2006.

[40] G. Rastelli, A. D. Rio, G. Degliesposti, and M. Sgobba. "Fast and accurate predictions of binding free energies using MM-PBSA and MM-GBSA". *Journal of computational chemistry*, 31(4), pp. 797–810, 2010.

[41] M. Ikeguchi, J. Ueno, M. Sato, and A. Kidera. "Protein structural change upon ligand binding: linear response theory". *Physical review letters*, 94(7), p. 078102, 2005.

[42] P. A. Kollman, I. Massova, C. Reyes, B. Kuhn, S. Huo, L. Chong, M. Lee, T. Lee, Y. Duan, W. Wang, O. Donini, P. Cieplak, J. Srinivasan, D. A. Case, and T. E. Cheatham. "Calculating structures and free energies of complex molecules: Combining molecular mechanics and continuum models". *Accounts of Chemical Research*, 33(12), pp. 889–897, 2000.

[43] J. Srinivasan, T. E. Cheatham, P. Cieplak, P. A. Kollman, and D. A. Case. "Continuum solvent studies of the stability of dna, rna, and phosphoramidate-dna helices". *Journal of the American Chemical Society*, 120(37), pp. 9401–9409, 1998.

[44] J. Åqvist, C. Medina, and J.-E. Samuelsson. "A new method for predicting binding affinity in computer-aided drug design". *Protein engineering*, 7(3), pp. 385–391, 1994.

[45] H. Gutiérrez-de Terán and J. Åqvist. "Linear interaction energy: method and applications in drug design". *Computational Drug Discovery and Design*, pp. 305–323, 2012.

[46] D. B. Kitchen, H. Decornez, J. R. Furr, and J. Bajorath. "Docking and scoring in virtual screening for drug discovery: methods and applications". *Nature reviews Drug discovery*, 3(11), pp. 935–949, 2004.

[47] H.-Y. Liu, I. D. Kuntz, and X. Zou. "Pairwise GB/SA scoring function for structure-based drug design". *The Journal of Physical Chemistry B*, 108(17), pp. 5453–5462, 2004.

[48] J. Esque and M. Cecchini. "Accurate calculation of conformational free energy differences in explicit water: The confinement–solvation free energy approach". *The Journal of Physical Chemistry B*, 119(16), pp. 5194–5207, 2015.

[49] T. Simonson, G. Archontis, and M. Karplus. "Free energy simulations come of age: protein-ligand recognition". *Accounts of chemical research*, 35(6), pp. 430–437, 2002.

[50] A. Pohorille, C. Jarzynski, and C. Chipot. "Good practices in free-energy calculations". *The Journal of Physical Chemistry B*, 114(32), pp. 10235–10253, 2010.

[51] J. C. Gumbart, B. Roux, and C. Chipot. "Standard binding free energies from computer simulations: What is the best strategy?" *Journal of chemical theory and computation*, 9(1), pp. 794–802, 2012.

[52] J. Wang, Y. Deng, and B. Roux. "Absolute binding free energy calculations using molecular dynamics simulations with restraining potentials". *Biophysical journal*, 91(8), pp. 2798–2814, 2006.

[53] A. Y. Lau and B. Roux. "The hidden energetics of ligand binding and activation in a glutamate receptor". *Nature structural & molecular biology*, 18(3), pp. 283–287, 2011.

[54] J. C. Gumbart, B. Roux, and C. Chipot. "Efficient determination of protein–protein standard binding free energies from first principles". *Journal of chemical theory and computation*, 9(8), pp. 3789–3798, 2013.

[55] L. Wang, Y. Wu, Y. Deng, B. Kim, L. Pierce, G. Krilov, D. Lupyan, S. Robinson, M. K. Dahlgren, J. Greenwood, *et al.* "Accurate and reliable prediction of relative ligand binding potency in prospective drug discovery by way of a modern free-energy calculation protocol and force field". *Journal of the American Chemical Society*, 137(7), pp. 2695–2703, 2015.

[56] B. Brooks and M. Karplus. "Harmonic dynamics of proteins: normal modes and fluctuations in bovine pancreatic trypsin inhibitor". *Proceedings of the National Academy of Sciences*, 80(21), pp. 6571–

6575, 1983.

[57] M. Cecchini. "Quantum corrections to the free energy difference between peptides and proteins conformers". *Journal of chemical theory and computation*, 11(9), pp. 4011–4022, 2015.

[58] R. M. Levy, M. Karplus, J. Kushick, and D. Perahia. "Evaluation of the configurational entropy for proteins: application to molecular dynamics simulations of an α-helix". *Macromolecules*, 17(7), pp. 1370–1374, 1984.

[59] M. Karplus and J. Kushick. "Method for estimating the configurational entropy of macromolecules". *Macromolecules*, 14(2), pp. 325–332, 1981.

[60] H. Gohlke and D. A. Case. "Converging free energy estimates: MM-PB(GB)SA studies on the protein–protein complex ras–raf". *Journal of computational chemistry*, 25(2), pp. 238–250, 2004.

[61] S. Genheden and U. Ryde. "Comparison of end-point continuum-solvation methods for the calculation of protein–ligand binding free energies". *Proteins: Structure, Function, and Bioinformatics*, 80(5), pp. 1326–1342, 2012.

[62] T. Hou, J. Wang, Y. Li, and W. Wang. "Assessing the performance of the MM/PBSA and MM/GBSA methods. 1. the accuracy of binding free energy calculations based on molecular dynamics simulations". *Journal of chemical information and modeling*, 51(1), pp. 69–82, 2010.

[63] T. Hou, J. Wang, Y. Li, and W. Wang. "Assessing the performance of the MM/PBSA and MM/GBSA methods: Ii. the accuracy of ranking poses generated from docking". *Journal of computational chemistry*, 32(5), p. 866, 2011.

[64] H. Sun, Y. Li, S. Tian, L. Xu, and T. Hou. "Assessing the performance of MM/PBSA and MM/GBSA methods. 4. accuracies of MM/PBSA and MM/GBSA methodologies evaluated by various simulation protocols using pdbbind data set". *Physical Chemistry Chemical Physics*, 16(31), pp. 16719–16729, 2014.

[65] S. Genheden, T. Luchko, S. Gusarov, A. Kovalenko, and U. Ryde. "An MM/3D-RISM approach for ligand binding affinities". *The Journal of Physical Chemistry B*, 114(25), pp. 8505–8516, 2010.

[66] L. Xu, H. Sun, Y. Li, J. Wang, and T. Hou. "Assessing the performance of MM/PBSA and MM/GBSA methods. 3. the impact of force fields and ligand charge models". *The Journal of Physical Chemistry B*, 117(28), pp. 8408–8421, 2013.

[67] C. Gao, M.-S. Park, and H. A. Stern. "Accounting for ligand conformational restriction in calculations of protein-ligand binding affinities". *Biophysical journal*, 98(5), pp. 901–910, 2010.

[68] S. Genheden, O. Kuhn, P. Mikulskis, D. Hoffmann, and U. Ryde. "The normal-mode entropy in the MM/GBSA method: effect of system truncation, buffer region, and dielectric constant". *Journal of chemical information and modeling*, 52(8), pp. 2079–2088, 2012.

[69] M. Kaukonen, P. Soderhjelm, J. Heimdal, and U. Ryde. "QM/MM-PBSA method to estimate free energies for reactions in proteins". *The Journal of Physical Chemistry B*, 112(39), pp. 12537–12548, 2008.

[70] F. S. Lee, Z.-T. Chu, M. B. Bolger, and A. Warshel. "Calculations of antibody-antigen interactions: microscopic and semi-microscopic evaluation of the free energies of binding of phosphorylcholine analogs to mcpc603". *Protein Engineering*, 5(3), pp. 215–228, 1992.

[71] R. Marcus. "Chemical and electrochemical electron-transfer theory". *Annual Review of Physical Chemistry*, 15(1), pp. 155–196, 1964.

[72] A. Ben-Naim and Y. Marcus. "Solvation thermodynamics of nonionic solutes". *The Journal of chemical physics*, 81(4), pp. 2016–2027, 1984.

[73] M. Almlöf, J. Carlsson, and J. Åqvist. "Improving the accuracy of the linear interaction energy method for solvation free energies". *Journal of chemical theory and computation*, 3(6), pp. 2162–2175, 2007.

[74] J. Åqvist and J. Marelius. "The linear interaction energy method for predicting ligand binding free energies". *Combinatorial chemistry & high throughput screening*, 4(8), pp. 613–626, 2001.

[75] J. Åqvist and T. Hansson. "On the validity of electrostatic linear response in polar solvents". *The Journal of Physical Chemistry*, 100(22), pp. 9512–9521, 1996.

[76] D. Huang and A. Caflisch. "Efficient evaluation of binding free energy using continuum electrostatics solvation". *Journal of medicinal chemistry*, 47(23), pp. 5791–5797, 2004.

[77] D. Huang and A. Caflisch. "Library screening by fragment-based docking". *Journal of Molecular Recognition*, 23(2), pp. 183–193, 2010.

[78] T. Zhou, D. Huang, and A. Caflisch. "Is quantum mechanics necessary for predicting binding free energy?" *Journal of medicinal chemistry*, 51(14), pp. 4280–4288, 2008.

[79] T. Zhou, D. Huang, and A. Caflisch. "Quantum mechanical methods for drug design". *Current topics in medicinal chemistry*, 10(1), pp. 33–45, 2010.

[80] A. Breda, L. A. Basso, D. S. Santos, J. De Azevedo, and F. Walter. "Virtual screening of drugs: score functions, docking, and drug design". *Current Computer-Aided Drug Design*, 4(4), pp. 265–272, 2008.

[81] S.-Y. Huang, S. Z. Grinter, and X. Zou. "Scoring functions and their evaluation methods for protein–ligand docking: recent advances and future directions". *Physical Chemistry Chemical Physics*, 12(40), pp. 12899–12908, 2010.

[82] H.-J. Böhm. "Prediction of binding constants of protein ligands: a fast method for the prioritization of hits obtained from de novo design or 3d database search programs". *Journal of computer-aided molecular design*, 12(4), pp. 309–309, 1998.

[83] D. Rognan, S. L. Lauemøller, A. Holm, S. Buus, and V. Tschinke. "Predicting binding affinities of protein ligands from three-dimensional models: application to peptide binding to class i major histocompatibility proteins". *Journal of medicinal chemistry*, 42(22), pp. 4650–4658, 1999.

[84] E. C. Meng, B. K. Shoichet, and I. D. Kuntz. "Automated docking with grid-based energy evaluation". *Journal of computational chemistry*, 13(4), pp. 505–524, 1992.

[85] G. M. Morris, R. Huey, W. Lindstrom, M. F. Sanner, R. K. Belew, D. S. Goodsell, and A. J. Olson. "Autodock4 and autodocktools4: Automated docking with selective receptor flexibility". *Journal of computational chemistry*, 30(16), pp. 2785–2791, 2009.

[86] J. Stoddart. ". host–guest chemistry". *Annual Reports Section" B"(Organic Chemistry)*, 85, pp. 353–386, 1988.

[87] J. W. Lee, S. Samal, N. Selvapalam, H.-J. Kim, and K. Kim. "Cucurbituril homologues and derivatives: new opportunities in supramolecular chemistry". *Accounts of chemical research*, 36(8), pp. 621–630, 2003.

[88] S. Grimme. "Supramolecular binding thermodynamics by dispersion-corrected density functional theory". *Chemistry-A European Journal*, 18(32), pp. 9955–9964, 2012.

[89] F. Biedermann, V. D. Uzunova, O. A. Scherman, W. M. Nau, and A. De Simone. "Release of high-energy water as an essential driving force for the high-affinity binding of cucurbit [n] urils". *Journal of the American Chemical Society*, 134(37), pp. 15318–15323, 2012.

[90] J.-M. Lehn. "Perspectives in supramolecular chemistry—from molecular recognition towards molecular information processing and self-organization". *Angewandte Chemie International Edition in English*, 29(11), pp. 1304–1319, 1990.

[91] Y. Furusho, I. M. Rahman, H. Hasegawa, and N. E. Izatt. "Application of molecular recognition technology to green chemistry: Analytical determinations of metals in metallurgical, environmental, waste, and radiochemical samples". *Metal Sustainability: Global Challenges, Consequences, and Prospects*, p. 271, 2016.

[92] W. Liu, S. K. Samanta, B. D. Smith, and L. Isaacs. "Synthetic mimics of biotin/(strept) avidin". *Chemical Society Reviews*, 46(9), pp. 2391–2403, 2017.

[93] T. Ogoshi, T.-a. Yamagishi, and Y. Nakamoto. "Pillar-shaped macrocyclic hosts pillar [n] arenes: new key players for supramolecular chemistry". *Chem. Rev*, 116(14), pp. 7937–8002, 2016.

[94] J. H. Jensen. "Predicting accurate absolute binding energies in aqueous solution: thermodynamic considerations for electronic structure methods". *Physical Chemistry Chemical Physics*, 17(19), pp. 12441–12451, 2015.

[95] J. J. Montalvo-Acosta and M. Cecchini. "Computational approaches to the chemical equilibrium constant in proteinligand binding". *Molecular Informatics*, 35(11-12), pp. 555–567, 2016.

[96] B. Roux and T. Simonson. "Implicit solvent models". *Biophysical chemistry*, 78(1-2), pp. 1–20, 1999.

[97] K. I. Assaf and W. M. Nau. "Cucurbiturils: from synthesis to high-affinity binding and catalysis". *Chemical Society Reviews*, 44(2), pp. 394–418, 2015.

[98] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case. "Development and testing of a general amber force field". *Journal of computational chemistry*, 25(9), pp. 1157–1174, 2004.

[99] T. Loftsson and D. Duchene. "Cyclodextrins and their pharmaceutical applications". *International journal of pharmaceutics*, 329(1), pp. 1–11, 2007.

[100] K. Vanommeslaeghe, E. Hatcher, C. Acharya, S. Kundu, S. Zhong, J. Shim, E. Darian, O. Guvench, P. Lopes, I. Vorobyov, and A. D. Mackerell. "Charmm general force field: A force field for drug-like molecules compatible with the charmm all-atom additive biological force fields". *Journal of Computational Chemistry*, 31(4), pp. 671–690, 2010.

[101] A. I. Lazar, F. Biedermann, K. R. Mustafina, K. I. Assaf, A. Hennig, and W. M. Nau. "Nanomolar binding of steroids to cucurbit [n] urils: selectivity and applications". *Journal of the American Chemical Society*, 138(39), pp. 13022–13029, 2016.

[102] H. S. Muddana and M. K. Gilson. "Calculation of host–guest binding affinities using a quantum-mechanical energy model". *Journal of chemical theory and computation*, 8(6), pp. 2023–2033, 2012.

[103] C. R. Groom, I. J. Bruno, M. P. Lightfoot, and S. C. Ward. "The cambridge structural database". *Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials*, 72(2), pp. 171–179, 2016.

[104] ChemAxon. "Marvinsketch", 2014. Accessed Feb 8, 2017.

[105] O. Korb, T. Stützle, and T. E. Exner. "Plants: Application of ant colony optimization to structure-based drug design". In "International Workshop on Ant Colony Optimization and Swarm Intelligence", pp. 247–258. Springer, 2006.

[106] A. Jakalian, B. L. Bush, D. B. Jack, and C. I. Bayly. "Fast, efficient generation of high-quality atomic charges. am1-bcc model: I. method". *Journal of Computational Chemistry*, 21(2), pp. 132–146, 2000.

[107] M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, and E. Lindahl. "Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers". *SoftwareX*, 1, pp. 19–25, 2015.

[108] G. Bussi, D. Donadio, and M. Parrinello. "Canonical sampling through velocity rescaling". *The Journal of chemical physics*, 126(1), p. 014101, 2007.

[109] G. J. Martyna, D. J. Tobias, and M. L. Klein. "Constant pressure molecular dynamics algorithms". *The Journal of Chemical Physics*, 101(5), pp. 4177–4189, 1994.

[110] M. Parrinello and A. Rahman. "Strain fluctuations and elastic constants". *The Journal of Chemical Physics*, 76(5), pp. 2662–2666, 1982.

[111] C. L. Wennberg, T. Murtola, S. Pall, M. J. Abraham, B. Hess, and E. Lindahl. "Direct-space corrections enable fast and accurate lorentz–berthelot combination rule lennard-jones lattice summation". *Journal of chemical theory and computation*, 11(12), pp. 5737–5746, 2015.

[112] R. Baron. *Computational drug discovery and design*. Humana Press, 2012.

[113] D. A. Case, T. E. Cheatham, T. Darden, H. Gohlke, R. Luo, K. M. Merz, A. Onufriev, C. Simmerling, B. Wang, and R. J. Woods. "The amber biomolecu-

lar simulation programs". *Journal of computational chemistry*, 26(16), pp. 1668–1688, 2005.

[114] D. Case, D. Cerutti, T. Cheatham, and T. Darden. "Amber 2017", 2017.

[115] G. D. Hawkins, C. J. Cramer, and D. G. Truhlar. "Parametrized models of aqueous free energies of solvation based on pairwise descreening of solute atomic charges from a dielectric medium". *The Journal of Physical Chemistry*, 100(51), pp. 19824–19839, 1996.

[116] P. M. Pihko. *Hydrogen bonding in organic synthesis*. John Wiley & Sons, 2009.

[117] M. Eigen. "Proton transfer, acid-base catalysis, and enzymatic hydrolysis. part i: Elementary processes". *Angewandte Chemie International Edition in English*, 3(1), pp. 1–19, 1964.

[118] H. Yamamoto and K. Futatsugi. ""designer acids": combined acid catalysis for asymmetric synthesis". *Angewandte Chemie International Edition*, 44(13), pp. 1924–1942, 2005.

[119] Z. Tian, A. Fattahi, L. Lis, and S. R. Kass. "Single-centered hydrogen-bonded enhanced acidity (shea) acids: a new class of brønsted acids". *Journal of the American Chemical Society*, 131(46), pp. 16984–16988, 2009.

[120] Y. Pocker. "Kinetics and mechanisms of addition of acids to olefins. part i. the addition of hydrogen chloride to isobutene in nitromethane". *Journal of the Chemical Society (Resumed)*, pp. 1292–1297, 1960.

[121] Y. Pocker, K. D. Stevens, and J. Champoux. "Kinetics and mechanism of addition of acids to olefins. iii. addition of hydrogen chloride to 2-methyl-1-butene, 2-methyl-2-butene, and isoprene in nitromethane". *Journal of the American Chemical Society*, 91(15), pp. 4199–4205, 1969.

[122] Y. Pocker and K. D. Stevens. "Kinetics and mechanism of addition of acids to olefins. iv. addition of hydrogen and deuterium chloride to 3-methyl-1-butene, 3, 3-dimethyl-1-butene, 1-methylcyclopentene, and 1-methylcyclopentene-2, 5, 5-d3". *Journal of the American Chemical Society*, 91(15), pp. 4205–4210, 1969.

[123] M. Dryzhakov, M. Hellal, E. Wolf, F. C. Falk, and J. Moran. "Nitro-assisted brønsted acid catalysis: Application to a challenging catalytic azidation". *Journal of the American Chemical Society*, 137(30), pp. 9555–9558, 2015.

[124] J.-D. Chai and M. Head-Gordon. "Long-range corrected hybrid density functionals with damped atom–atom dispersion corrections". *Physical Chemistry Chemical Physics*, 10(44), pp. 6615–6620, 2008.

[125] M. D. Hanwell, D. E. Curtis, D. C. Lonie, T. Vandermeersch, E. Zurek, and G. R. Hutchison. "Avogadro: an advanced semantic chemical editor, visualization, and analysis platform". *Journal of cheminformatics*, 4(1), p. 17, 2012.

[126] J. G. Brandenburg, M. Hochheim, T. Bredow, and S. Grimme. "Low-cost quantum chemical methods for noncovalent interactions". *The journal of physical chemistry letters*, 5(24), pp. 4275–4284, 2014.

[127] F. Neese. "The orca program system". *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 2(1), pp. 73–78, 2012.

[128] M. Frisch, G. Trucks, H. B. Schlegel, G. Scuseria, M. Robb, J. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. Petersson, *et al.* "Gaussian 09, revision d. 01", 2009.

[129] P. Gramatica. "Principles of qsar models validation: internal and external". *QSAR & combinatorial science*, 26(5), pp. 694–701, 2007.

UNIVERSITÉ DE STRASBOURG

Joel José MONTALVO ACOSTA

Approches Computationnelles de la Reconnaissance Moléculaire: L'analyse de la Liaison Hôte-Invité et Protéine-Ligand

Résumé

La reconnaissance moléculaire est un problème très intéressant et surtout un défi actuel pour la chimie biophysique. Avoir des prévisions fiables sur la reconnaissance spécifique entre les molécules est hautement prioritaire, car il fournira un aperçu des problèmes fondamentaux et suscitera des applications technologiques pertinentes. La thèse présentée ici est centrée sur une analyse quantitatif de la reconnaissance moléculaire en solution pour la liaison l'hôte-invité, la liaison protéine-ligand et la catalyse. Le cadre de la mécanique statistique utilisé pour décrire l'état de la technique de liaison récepteur-ligand est un point d'inflexion pour le développement de nouvelles méthodes améliorées. En fait, un modèle très performant et précis a été obtenu pour l'analyse de la liaison hôte-invité. Enfin, les modèles présentés ont été utilisés comme outils prédictifs fiables pour la découverte de nouvelles entités chimiques destinées à améliorer la catalyse en solution.

Mots-clés : Reconnaissance moléculaire, liaison d'énergie libre, effet co-catalytique, mécanique statistique

Résumé en anglais

Molecular recognition is a very interesting problem, and foremost, a current challenge for biophysical chemistry. Having reliable predictions on the specific recognition between molecules is highly priority as it will provide an insight of fundamental problems and will raise relevant technological applications. The dissertation presented here is centered on a quantitative analysis of molecular recognition in solution for host-guest, protein-ligand binding and catalysis. The statistical mechanics framework used to describe the state-of-the-art for receptor-ligand binding is an inflection point for the developing of new improved and methods. In fact, a highly performanced and accurate model was obtained for the analysis of host-guest binding. Finally, the presented models were used as a reliable predictive tools for discovering new chemical entities for enhance catalysis in solution.

Keywords : Molecular recognition, binding free energy, co-catalytic effect, statistical mechanics