Theoretical Investigation on the Biomolecular Systems using Multiscale Modelling

Zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

von der KIT-Fakultät für Chemie und Biowissenschaften
des Karlsruher Instituts für Technologie (KIT)

genehmigte

Dissertation

von

M.Sc. Ziwei Pang

aus China

Tag der mündlichen Prüfung: 28. April 2022
Erster Gutachter: Prof. Dr. Marcus Elstner
Zweiter Gutachter: Prof. Dr. Andreas-Neil Unterreiner

Affidavit

Ich versichere hiermit wahrheitsgemäß, die Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt, die wörtlich oder inhaltlich übernommenen Stellen als solche kenntlich gemacht und die Satzung des Karlsruher Instituts für Technologie (KIT) zur Sicherung guter wissenschaftlicher Praxis in der jeweils gültigen Fassung beachtet zu haben.

Karlsruhe, 28.04.2022.

Acknowledgement

First of all, I would like to give my highest respect and gratitude to my supervisor Prof. Dr. Marcus Elstner. I would like to thank him for providing me these fantastic research projects. I would also like to thank him for his guidance, help and encouragement in my research, which has allowed me to face any difficulties with ease.

Special thanks to Dr. Monja Sokolov, for our perfect teamwork and outstanding achievements on the GGBP project and for every efficient online meeting during the special time of the corona pandemic. Also, thanks to her and Philipp Dohmen for their selfless help with the ML project at the last minute. Thanks to Dr. Tomáš Kubař for his outstanding technical support for over three years and for patiently answering my questions, even though some of them may seem naive now. Thanks to Dr. Daniel Holub, Dr. Weiwei Xie, Dr. Beatrix Bold and others for helping me at the beginning of my PhD so that I could start my research smoothly.

I would also like to thank Philipp Dohmen, Denis Maag and all my other colleagues, it has been an unforgettable experience for me to spend time with them and complete my PhD. Thanks to their selfless academic sharing, the relaxed group atmosphere and various impressive activities, especially the memorable lunchtime.

Thanks to Jochen Lauer and Dr. Sven Elbert from the Prof. Dr. Michael Masterlerz group, the collaborations on the Host-Guest projects were always enthusiastic and efficient.

Thanks to Dr. Violetta Schneider for the preparation of the GGBP project.

Thanks to all the supervisors and PhD students of GRK2039, every meeting and retreat together has been very rewarding.

Thanks to Sabine Holthoff for her help in solving various problems.

Thanks to the CSC scholarship for supporting me financially.

A special thanks to my flatmates, Xingchen Liu, Yicun Lu and Fangzhou Li, for taking care of me and making my life in Germany so enriching.

In addition, I would like to thank my friends in China and in Germany for their help and encouragement over the past three years.

Special thanks to my girlfriend Yuting Feng for your companionship and support during the long years. You are always there to support and inspire me when I am tough and cheer me on when I receive achievements. I love you.

Last but not least, thank you very much to all my family members who have always been my torch and my most substantial support on the road. Thank you all! I love you all!

Abstract

The study of protein-ligand interactions is crucial and challenging for biomolecular systems. In particular, traditional laboratory experiments often have difficulties explaining the mechanisms of the reactions, while classical theoretical computational methods have shortages in dealing with the system scale and time scale of biomolecular systems. In this work, enhanced sampling methods based on molecular dynamics (MD) simulations and artificial neural network (ANN) algorithms based on semi-empirical quantum mechanics (QM) approaches were applied to explore different biomolecular systems.

In the first part, host-guest chemistry of [4+4] as well as [2+3] imine cages was studied. In the study of [4+4] cages, the uptake process of ammonium ions with different sizes in cages with alternative volumes was simulated by well-tempered metadynamics (MetaD). Three plausible mechanisms are proposed to explain the guest uptake processes. In the [2+3] cages study, the nitrogen molecule transfer in three different cage crystals was calculated with funnel metadynamics (FM). The free energy surfaces obtained suggested the existence of two potential pathways to express the mechanism of nitrogen transfer between cages.

A novel glucose binding protein-based fluorescence probe was investigated in the second part. A detailed molecular understanding of changes in the glucose binding site due to mutations and their effect on glucose binding was achieved via MD simulations. The energetics of unbinding in these proteins were revealed and were consistent with the experimental results.

Last, a set of machines were trained by artificial neural networks (ANNs) to correct the misrepresentation of excited states by LC-DFTB when energy level crossing occurs. Most of the trained machines can accurately modify the excited state definition errors brought by LC-DFTB, whereas the machine trained in water is less accurate and required further training.

Key Words: *Protein-ligand interaction, Molecular dynamics simulation, Quantum mechanics, Metadynamics, Machine learning.*

Zusammenfassung

Die Untersuchung von Protein-Ligand-Wechselwirkungen ist für biomolekulare Systeme von entscheidender Bedeutung und eine Herausforderung. Insbesondere haben traditionelle Laborexperimente oft Schwierigkeiten, die Mechanismen der Reaktionen zu erklären, während klassische theoretische Berechnungsmethoden Defizite im Umgang mit der System- und Zeitskala biomolekularer Systeme aufweisen. In dieser Arbeit werden sogenannte *enhanced* Sampling-Methoden auf der Grundlage von Molekulardynamiksimulationen (MD) und Algorithmen für künstliche neuronale Netze (ANN), die auf semi-empirischen quantenmechanischen (QM) Ansätzen beruhen, zur Untersuchung verschiedener biomolekularer Systeme eingesetzt.

Im ersten Teil wurde die Wirt-Gast-Chemie von [4+4]- und [2+3]-Iminkäfigen untersucht. Bei der Untersuchung von [4+4]-Käfigen wurde der Aufnahmeprozess von unterschiedlich großen Ammoniumionen in Käfigen mit alternativen Volumina durch wohltemperierte Metadynamik (MetaD) simuliert. Es wurden drei mögliche Mechanismen vorgeschlagen, um die Gastaufnahmeprozesse zu erklären. Bei der Untersuchung von [2+3]-Käfigen wurde der Stickstoffmolekültransfer in drei verschiedenen Käfigkristallen mit Funnel-Metadynamik (FM) berechnet. Die erhaltenen freien Energieflächen deuten auf die Existenz von zwei möglichen Wegen hin, auf denen der Stickstofftransfer erfolgen kann.

Im zweiten Teil wurde eine neuartige Fluoreszenzsonde auf der Basis eines Glukose bindenden Proteins untersucht. Ein detailliertes molekulares Verständnis der Veränderungen an der Glukosebindestelle aufgrund von Mutationen und deren Auswirkungen auf die Glukosebindung wurde durch MD-Simulationen erreicht. Die Energetik der Dissoziation von Protein und Glukose wurde aufgedeckt und stimmte mit den experimentellen Ergebnissen überein.

Schließlich wurde eine Reihe von künstlichen neuronalen Netzen (ANNs) trainiert, um die falsche Darstellung von angeregten Zuständen durch LC-DFTB zu korrigieren, wenn Energieniveaus kreuzen. Die meisten der trainierten Maschinen sind in der Lage, die durch LC-DFTB verursachten Fehler bei der Beschreibung des angeregten Zustands zuverlässig zu korrigieren, während die für Farbstoffgeometrien in Wasser trainierte Maschine weniger genaue Ergebnisse liefert und weiteres Training erfordert.

Schlüsselwörter: *Protein-Ligand-Wechselwirkung, Molekulardynamiksimulation, Quantenmechanik, Metadynamik, Maschinelles Lernen*

Contents

Part I.

Introduction

1. Protein-Ligand Interactions

In nature, proteins perform their functions by interacting with other proteins or molecules and are involved in almost all biological processes. Therefore, an in-depth study of the structure, function, and interactions of proteins is essential in explaining various life activities phenomena.

Although some proteins can perform physiological functions independently, most protein molecules do not exist in isolation. They need to interact with other biological molecules to perform specific physiological functions, e.g. enzymes and substrates, hormones and receptors, antibodies and antigens and so on. These molecules that interact with the protein, such as biomolecules, metal ions, etc., are called *ligands*. Upon interaction with a ligand, the protein structure may change, ranging from the twisting of the amino acid side chains to protein domain structural changes. Likewise, ligands may undergo specific conformational changes when interacting with proteins. In addition, protein-ligand interactions require specific non-covalent bonding interactions, including hydrogen bonding, electrostatic forces, van der Waals (vdW) interactions, hydrophobic interactions, etc. For protein-ligand complexes, differences in the conformation and variations in interaction forces can lead to entirely different binding effects.

Therefore, a detailed study of protein-ligand interactions at the atomic level will help to reveal the effects of ligand binding on protein structure and deepen the understanding of many biological regulatory mechanisms. In addition, it also provides essential guidance for drug development and enzyme engineering.

To study protein-ligand interactions, it is necessary to explore both the kinetic and thermodynamic properties of the binding. Over the past century, three theoretical models have been developed to describe the protein-ligand interactions. In 1897, Fischer proposed the “Lock-Key” theory as the first theoretical model to describe the interaction between proteins and small molecules [1]. The model treats protein receptors as locks and ligands, such

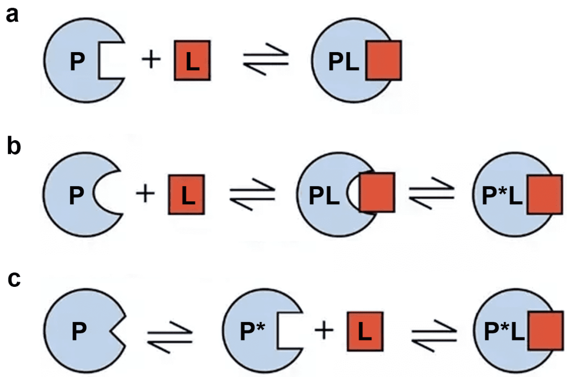


Figure 1.1.: Schematic view of protein-ligand binding mechanisms. **a).** The "Lock-Key" mechanism. **b).** The "Induced-fit" mechanism. **c).** The "Conformational Selection" mechanism.

as small molecules, as keys, both of which are rigid molecules that recognise and bind to each other by matching their spatial shapes. In 1958, the "Induced Fit model" was proposed by Koshland et al. [2]. This model states that when a small molecule binds to a protein, both conformations will change due to intermolecular interactions until an energetically stable binding pattern is formed. In contrast to "Lock-Key", this model compensates for the deficiencies of the first model by taking the structural changes of the bound molecules into account. The third model is the "Conformational Selection" model proposed by Straub in 1964 [3], which assumes that proteins have many conformations and only one or some of them can be bounded with ligands. To put it briefly, suppose a protein has only two states, A and B. State A can bind directly to a ligand molecule, but state B cannot, so if a protein in state B wants to bind to a ligand, its conformation needs to convert to the state A before binding. In recent years, with the development of biology, protein-ligand interactions have been found to co-exist with both "conformational selection" and "induced fit" mechanism reported by Csermely et al. [4].

The protein-ligand interaction can be described by the binding rate constant k_{on} and the dissociation rate constant k_{off}. The affinity of the ligand bound to the protein is determined by the equilibrium dissociation constant, where,

$$k_{\mathrm{D}} = \frac{k_{\mathrm{off}}}{k_{\mathrm{on}}} = \frac{[\mathrm{P}] \cdot [\mathrm{L}]}{[\mathrm{PL}]} \tag{1.1}$$

Here, [P] is the concentration of unbound-state protein, [L] the concentration of ligand molecules and [PL] the protein-ligand complexes concentration.

$$\mathrm{P} + \mathrm{L} \underset{\mathrm{k_{on}}}{\overset{\mathrm{k_{off}}}{\rightleftharpoons}} \mathrm{PL} \tag{1.2}$$

As shown in Equation 1.2, the protein-ligand binding and dissociation reactions are reversible, and hence are often subjected to standard equilibrium thermodynamic analysis. The affinity between protein and ligand is evaluated by the equilibrium dissociation constant, directly related to the standard Gibbs free energy (often referred to as the binding free energy in receptor-ligand interactions) as,

$$G^0_{\mathrm{binding}} = RT \ln k_{\mathrm{D}} = RT \ln \frac{k_{\mathrm{off}}}{k_{\mathrm{on}}}. \tag{1.3}$$

2. Fluorescence

Fluorescence is the emission of light by a substance that has been irradiated by the excitation light or other electromagnetic radiation. The earliest record of the fluorescence dates back to the 16th century when the Spanish botanist and physician N. Monardes discovered a piece of wood called "Lignum Nephriticura" whose aqueous solution is sky-blue. In 1852, when Sir G. Stokes studied quinine and chlorophyll solutions, they reported that the wavelength of the emitted light of these solutions was slightly longer than the wavelength of the incident light . Later, it was found that this phenomenon was caused by the material being able to re-emit light with different wavelengths after absorbing the excitation light and then such re-emitted light was first described as *fluorescence.*

In 1867, Goppelsröder realised the first fluorescence analysis in the determination of aluminium in aluminium-morin complexes [5]. With the rapid development of other related disciplines in recent years, the theory and application of fluorescence analysis methods have been greatly promoted and improved. Currently, the fluorescence analysis method has become a crucial spectroscopic analysis method for many fields of industrial production, daily life and scientific research due to its advantages of being sensitive, efficient, suitable for real-time, simple and in-situ detection [6–8].

2.1. Fluorescence Generation

When the external energy is encountered, orbital electrons of the fluorescent molecules absorb this energy and change ("hop") into higher electronical states (S_n). Photons are subsequently released by radiative decay, which returns the excited state molecule to the ground state, and fluorescence occurs. The specific process of fluorescence generation can be described by the Jablonski diagram [9] as follows.

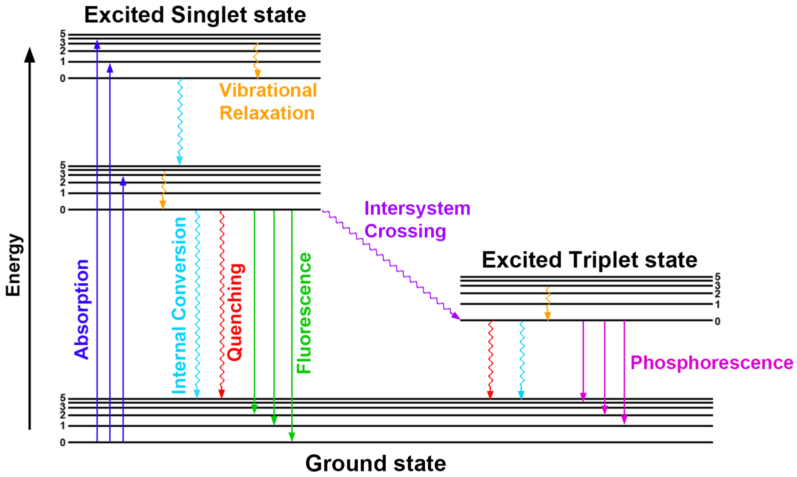


Figure 2.1.: Jablonski Energy Diagram.

When the number of electrons with spin up is equal to those with spin down, the molecule is in a singlet state (S); on the contrary, if the spin direction of an electron changes such that the total spin becomes one, then the molecule is in a triplet state (T). In ground state (S_0), fluorescent molecules have their electrons at the lowest energy level. When these molecules absorb the energy of the external light, they will transfer to different vibrational energy states, such as the first excited state (S_1) and the second excited state (S_2). Subsequently, the electrons at different vibrational energy states fall to the lowest vibrational energy level (S_1) through various non-radiative processes (Vibrational Relaxation and Internal Conversion). The electrons in the first excited state (S_1) will directly return to the ground state in the form of radiation, and the light emitted is called *fluorescence*. In some cases, in the first excited singlet state, an electron changes its spin direction and the molecule is trapped in the first excited triplet state (T_1) through intersystem crossing and then return to the ground state in the form of radiation. The light emitted is called *phosphorescence*.

As mentioned above, part of the energy absorbed by excited electrons will be dissipated through vibrational relaxation and internal conversion, and hence the energy emitted in the form of fluorescence is less than the energy absorbed when excited. Therefore, the spectrum shows that the fluorescence emission wavelength is longer than the absorption wavelength, and the difference between them is known as the *Stokes shift*. As another consequence of vibrational relaxation and internal conversion, the spectral shape of the fluorescence is not affected by the wavelength of the excitation light source. In general, fluorescent materials have only one emission peak. However, during energy absorption, the ground state electrons can jump to different excited states simultaneously. Therefore, a material can exhibit multiple absorption peaks at the same time.

2.2. Fluorescent Probe

Over the past decade, the rapid development of fluorescent probes has been witnessed. The low content of elements in the organism can be detected through the optimised fluorescent probe, which can respond specifically to a single element and eliminate the interference of other factors. At the same time, compared to the other detection approaches with complex measurement processes, its operation is significantly convenient. Most importantly, information on different temporal and spatial distributions can be revealed without damaging the biological sample, which is of great importance for the study of the physiological functions of important biological species [10, 11].

In general, fluorescent probes are composed of two main functional parts. The first part is the recognition group to identify specific groups and change the fluorescence signal when combined with the analyte; the second part is the chromophore as the signal group. The two parts of the probe can be attached directly, or the recognition group can be part of a chromophore. Besides, they can be connected by a linking group. For the signal group, small organic molecule dyes are the most commonly used chromophore of fluorescent probes, such as Fluorescein [12], BODIPY [13], Cyanine [14], and so on. In addition, upconverting nanomaterials [15], polymeric fluorescent materials [16] and fluorescent proteins [17] can also serve as the chromophore in the fluorescent probe.

Fluorescent probes can be classified according to different principles. On one hand, based on the type of fluorescence signal changes after interacting with the specific group, it can be divided into intensity-changing fluorescent probes (On-Off type and Off-On type) and ratio-type fluorescent probes. On the other hand, coordination fluorescent probes and reactive fluorescent probes (Chemodosimeter) can also be classified depending on the type of response between the recognition group and the specific group.

2.3. Intramolecular Charge Transfer (ICT)

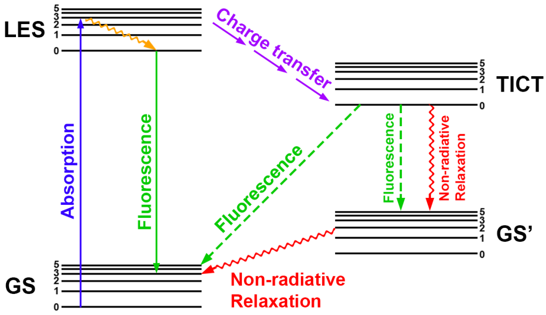


Figure 2.2.: Twisted Intramolecular Charge Transfer (TICT) dynamics.

Intramolecular charge transfer (ICT) refers to the transfer phenomenon of excited state electrons resulting in the separation of positive and negative charges, which forms the molecular charge transfer state [18]. Fluorescent probe molecules that use such a mechanism usually have electron-donor and electron-acceptor groups as the recognition group attaching to the chromophore. A strong "D-A" conjugate structure is generated through the electron transfer channels provided by the π bond. As the electron transfer occurs, the positive and negative charges within the molecule separate, and the dipole moment increases. Thus, the locally excited state (Franck-Condon state, LE) at the moment is no longer stable. As shown in Fig. 2.2, the energy of the CT state tends to be lower than that of the LE state, and as a result, the light emission peak of the ICT state is red-shifted, and the fluorescence intensity is reduced. In addition, as the environment polarity increases, the positive and negative charges tend to separate, which further exacerbates the decrease in energy of the ICT state and increases the red-shift of the fluorescence emission spectrum [19, 20].

Twisted intramolecular charge transfer (TICT) is a type of ICT mechanism in which some molecules with ICT properties twist or rotate themselves in the excited state when the molecule is in the TICT state [18, 21]. Fluorescent molecules with TICT properties are generally very sensitive to the polarity of their environment, producing short wavelength light in the LE state when the molecule is in a non-polar solvent environment and, as polarity increases, fluorescence is emitted in the long wavelength TICT state. Moreover, it can be observed that as the polarity increases, the intensity of the short wavelength LE fluorescence decreases and the fluorescence intensity of the long wavelength TICT state increases (Fig. 2.2). Therefore, based on this property of TICT, it is possible to design synthetic scaled fluorescent probes.

Part II.

Theoretical Background

3. Quantum Chemistry

Modern chemistry is not only a laboratory discipline but also requires extensive computer simulations. For instance, biochemical reactions often happen within only a few microseconds. However, it is impossible to track what happens in such a short time in the laboratory. Hence, the reaction mechanism cannot be fully explained.

In computational chemistry, if chemical molecules are considered to be composed of atoms and chemical bonds, many properties of molecules can be well described using the formulas of classical mechanics, such as stretching of chemical bonds, the opening and closing of bond angles, the rotation of dihedral angles and so on. Such a modelling approach is called *molecular mechanics* (MM) modelling. However, properties such as the breaking and generation of chemical bonds involve changes in the electronic structure. The MM model is unsuitable to describe the related properties because it treats atoms as individual particles. Therefore, it is necessary to consider the electrons when modelling, and such a model is called a *quantum mechanics* (QM) model. For the same system, the computational complexity of the QM model is much higher than that of the MM model, and the difference is often several orders of magnitude. As discussed above, the MM model regards atoms as classical particles, while the QM model needs to consider QM effects. Hence, the computation of energy and forces in quantum mechanics is far more complicated than the MM model. However, the QM model can not only simulate the breaking and generation of chemical bonds that cannot be achieved by the MM model but also obtain more accurate calculation results than the MM model.

As shown in Fig. 3.1, molecules can be calculated by different simulation approaches in computational chemistry, and the "accuracy" and "speed" of the simulation method are often a pair of contradictions. Specifically, finer methods can achieve higher computational accuracy but often pay higher computational costs and vice versa. There are not only QM models and

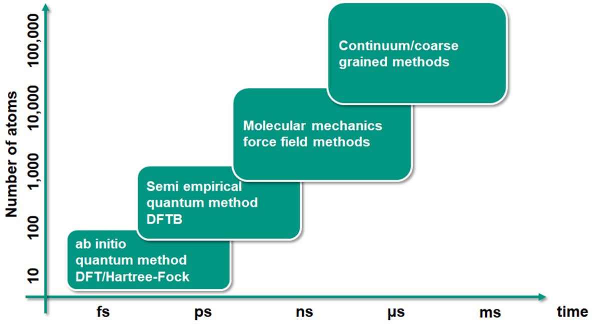

Number of atoms
10
100
1,000
10,000
100,000
Continuum/coarse grained methods
Molecular mechanics force field methods
Semi empirical quantum method DFTB
ab initio quantum method DFT/Hartree-Fock
fs
ps
ns
µs
ms
time

Figure 3.1.: Schematic representation of computational simulation methods according to the simulation time and the model size scale.

MM models in computational chemistry. When modelling, higher-precision calculation results can be obtained if we consider finer structures (such as quantum effects of atomic nuclei), but the calculation is even more expensive. Conversely, it is also possible to model in a coarser way (such as the coarse-grained method) to achieve a faster model but with lower accuracy.

Therefore, in practice, different simulation methods are often chosen according to the different research purposes. For instance, when unravelling the folding process of protein molecules, which contains thousands of atoms in the system and only involves the conformational changes of the molecules, a force field-based molecular mechanics approach or a coarse-grained method is sufficient. If the study focuses on the chemical reaction of small organic molecules, applying a QM method is necessary. However, in actual computational chemistry research, the system is always more complex and has more molecular properties that need to be investigated. For example, an important research direction of computational chemistry is the enzyme-catalysed reaction of proteins, which involves chemical reactions and conformational changes. Obviously, the MM model can only meet the requirements of conformational changes. If the QM method is applied to model the entire protein system, the calculation speed will be extremely slow. Since only a very small part of the protein is accounted for in the chemical reactions, we can perform QM calculations on the reaction-related protein zone and calculate the rest by using the MM model to achieve the chemical reaction while maintaining a high computational speed. This hybrid simulation method is called the Quantum Mechanics/Molecular Mechanics (QM/MM) model.

To sum up, in computational chemistry, modelling with different scales can achieve different effects. Usually, the finer scale, the higher the calculation accuracy, but the lower the calculation speed, and vice versa. For complex chemical systems, different parts of the system can be calculated with models of different scales as needed to achieve the required computational accuracy with the minimum computational cost. Various simulation approaches have emerged with the rapid development of computational chemistry in recent decades. This chapter will briefly introduce the simulation methods applied in this work.

3.1. Schrödinger Equation and Hamiltonian

In general, quantum chemistry is to apply the basic principles of quantum mechanics to solve the *Schrödinger* equation, and extract the potential energy surface structure and wave function of molecules to calculate the properties. The time-dependent form of the Schrödinger equation is

$$\left\{-\frac{\hbar^2}{2m}\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial z^2}\right)+V(\mathbf{r},t)\right\}\Psi(\mathbf{r},t)=i\hbar\frac{\partial\Psi(\mathbf{r},t)}{\partial t}. \tag{3.1}$$

In Equation 3.1, m is the mass of the single particle, $\hbar$ is the Planck's constant divided by 2π, i is the imaginary unit, V is the potential energy, $\Psi(\mathbf{r},t)$ is the wave function of a particle depended on the spatial coordinates $\mathbf{r}$ and time t, which describe the particle's motion. When the V is independent of time, the time-independent Schrödinger Equation can be written as

$$\left\{-\frac{\hbar^2}{2m}\nabla^2+V(\mathbf{r},t)\right\}\Psi(\mathbf{r})=E\Psi(\mathbf{r}). \tag{3.2}$$

Here, ∇^2 is the abbreviation for $\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial z^2}$ and E is the energy of the particle or the system. The left hand side of Equation 3.2 can be also written as $\hat{H}\Psi$. The $\hat{H}$ is called *Hamiltonian* operator, which contains the kinetic energy $-\frac{\hbar^2}{2m}\nabla^2$ and the potential energy V for the particle. This makes the time-independent Schrödinger Equation simplified as

$$\hat{H}\Psi(\mathbf{r}) = E\Psi(\mathbf{r}). \tag{3.3}$$

For a molecule containing n electrons and N nuclei, the $\hat{H}$ can be presented as the sum of kinetic energy and potential energy operators as,

$$\hat{H} = \hat{T}_e + \hat{T}_N + \hat{V}_{ee} + \hat{V}_{eN} + \hat{V}_{NN}, \tag{3.4}$$

where,

$$\begin{aligned}
\hat{T}_e &= -\sum_{i=1}^{n} \frac{\hbar^2}{2m_e}\nabla_i^2 \\
\hat{T}_N &= -\sum_{A=1}^{N} \frac{\hbar^2}{2M_N}\nabla_A^2 \\
\hat{V}_{ee} &= \sum_{i=1}^{n}\sum_{j>i}^{n} \frac{e^2}{|\mathbf{r_i} - \mathbf{r_j}|} \\
\hat{V}_{eN} &= -\sum_{i=1}^{n}\sum_{A=1}^{N} \frac{e^2 Z_A}{|\mathbf{r_A} - \mathbf{r_i}|} \\
\hat{V}_{NN} &= \sum_{A=1}^{N}\sum_{B>A}^{N} \frac{e^2 Z_A Z_B}{|\mathbf{r_A} - \mathbf{r_B}|}.
\end{aligned} \tag{3.5}$$

Here, $\hat{T}_e$ is the kinetic energy of the electrons, and $\hat{T}_N$ is the kinetic energy of the nuclei. $\hat{V}_{ee}$, $\hat{V}_{eN}$, $\hat{V}_{NN}$ stand for the repulsive interaction between electrons, the attractive interaction between electrons and the nuclei and repulsive interaction between nuclei, respectively. m_e is the mass of electrons, M_N the mass of nuclei, $\mathbf{r_i}$ and $\mathbf{r_j}$ represent the coordinates of electron i and j, $\mathbf{r_A}$ and $\mathbf{r_B}$ represent the coordinates of nuclei A and B, and Z_A and Z_B are the charges of nuclei A and B.

Since the Schrödinger equation is a partial differential equation, it is difficult to solve except for only a few cases, such as one-electron system as $H_2{}^+$. As the nucleus is much more massive than the electron and moves more slowly than the electron, the nucleus can be regarded as immobile. In such circumstances, the Hamiltonian operator can be written as

$$\hat{H} = -\sum_{i=1}^{n} \frac{\hbar^2}{2m_e}\nabla_i^2 - \sum_{i=1}^{n}\sum_{A=1}^{N} \frac{e^2 Z_A}{|\mathbf{r_A} - \mathbf{r_i}|} + \sum_{i=1}^{n}\sum_{j>i}^{n} \frac{e^2}{|\mathbf{r_i} - \mathbf{r_j}|} + \sum_{A=1}^{N}\sum_{B>A}^{N} \frac{e^2 Z_A Z_B}{|\mathbf{r_A} - \mathbf{r_B}|}. \tag{3.6}$$

The last term of Equation 3.6 is a constant, which indicates the repulsive potential between nuclei. Hence, the wave function only depends on the coordinates of electrons. In other words, the problem of the relative motion between nuclei and electrons is transformed into a problem of the motion of the electrons around immobile nuclei. This means that for any given arrangement of nuclei, the electrons have a corresponding state of motion, while the relative motion between the nuclei can be regarded as the average effect of the electron motion. This is the *Born–Oppenheimer approximation.* Under this approximation, the total wave function for the molecule can be described as follows:

$$\Psi_{tot}(\text{nuclei, electrons}) = \Psi(\text{electrons})\Psi(\text{nuclei}). \tag{3.7}$$

With the introduction of the Born–Oppenheimer approximation, solving the Schrödinger equation for a multi-electron molecular system is simplified. However, except for a few simplest molecule systems, the Schrödinger equation is still very difficult to be precisely solved. Therefore, the variational principle is often applied to optimise the solution of the Schrödinger equation.

$$\langle\Psi_{trial}|\hat{H}|\Psi_{trial}\rangle = E_{trial} \geq E_0 = \langle\Psi_0|\hat{H}|\Psi_0\rangle \tag{3.8}$$

In this principle, the true energy E_0 is lower than the energy E_{trial} corresponding to any guessed trail wave function Ψ_{trial}.

3.2. Hartree-Fock

Despite the introduction of the Born–Oppenheimer approximation, the electron's wave function has an exact solution only available for molecular ions containing one electron. In a many body system, the motion of the individual electrons is interconnected. Since the interaction energy between the electrons contained in the Hamiltonian cannot be found, further approximation methods must be applied. The simplest and most convenient approximation is the one-electron approximation proposed by Hartree, also known as the Hartree approximation [22]. That is, the effects of all the other electrons on each electron are replaced by a potential field. Thus, each electron in the system appears to be moving independently and has its own eigenvalue and eigenfunction.

The wave function of the total system could be assumed as the product of the state functions of the individual electrons. Hence, with a fixed nucleus position, the motion of a single electron depends only on the coordinates of itself so that the multi-electron wave function can be decomposed into the product of the single-electron wave functions as,

$$\Psi(\mathbf{r_1}, \mathbf{r_2}, ...\mathbf{r_n}) = \Phi(\mathbf{r_1})\Phi(\mathbf{r_2})...\Phi(\mathbf{r_n}). \tag{3.9}$$

The one-electron approximation inevitably leads to the emergence of a central concept of molecular orbital theory – the self-consistent field (SCF). The motion of each electron is influenced not only by nucleus, but also by the potential fields created by other electrons. Thus, in describing the nucleus-electron potential field, one must not only consider the state of the electron that is affected, but also its contribution to the potential field as the other electrons move, i.e. the self-consistent. Such a potential field is called a *self-consistent field.*

Since this representation of Equation 3.9 does not satisfy the antisymmetric principle, Fock improved the Hartree equation by rewriting the many-body wave function as a *Slater determinant* (Equation 3.10) of the single-electron wave function satisfying the exchange antisymmetry [23].

$$\Psi_{HF} = \frac{1}{\sqrt{N!}} \begin{bmatrix} \Phi_1(\mathbf{1}) & \Phi_1(2) & \cdots & \Phi_1(\mathrm{N}) \\ \Phi_2(\mathbf{1}) & \Phi_2(2) & \cdots & \Phi_2(\mathrm{N}) \\ \vdots & \vdots & \ddots & \vdots \\ \Phi_N(\mathbf{1}) & \Phi_N(2) & \cdots & \Phi_N(\mathrm{N}) \end{bmatrix} \tag{3.10}$$

Here, N is the total number of electrons, $\Phi_1(\mathbf{1})$ indicates a spin orbital containing position $\mathbf{r_1}$ and spin σ for the electron labelled as “1”. Applying the variational principle to the Slater determinant under the constraint of orthogonal normalisation yields the HF equation as [24],

$$\begin{aligned}&\left[-\frac{1}{2}\nabla_i^2-\sum_{A=1}^{M}\frac{Z_A}{\mathbf{r_{iA}}}\right]\Phi_i(\mathbf{1})+\sum_{j\neq i}\left[\int d\mathbf{r}_2\Phi_j^*(2)\Phi_j(2)\frac{1}{\mathbf{r_{12}}}\right]\Phi_i(\mathbf{1})\\&-\sum_{j\neq i}\left[\int d\mathbf{r}_2\Phi_j^*(2)\Phi_i(2)\frac{1}{\mathbf{r_{12}}}\right]\Phi_i(\mathbf{1})=\sum_j\varepsilon_{ij}\Phi_j(\mathbf{1}).\end{aligned}\tag{3.11}$$

On the left side of the Equation 3.11, the first term is the single-electron energy, which includes the kinetic energy and the attractive interaction to the nuclei. The second term is the Coulomb interaction between electrons, corresponding to the potential energy due to the average charge distribution of $N-1$ electrons in the orbitals Φ_j. The third term is the exchange interaction derived from Pauli's principle. In a "frozen" system, these three terms can be written by the core Hamiltonian operator $\hat{H}^{\text{core}}$, the Coulomb operator $\hat{J}_j$, and the exchange operator $\hat{K}_j$, respectively:

$$\begin{aligned}\hat{H}^{\text{core}}(\mathbf{1})&=-\frac{1}{2}\nabla_1^2-\sum_{A=1}^{M}\frac{Z_A}{\mathbf{r_{1A}}}\\\hat{J}_j(\mathbf{1})&=\int d\mathbf{r}_2\Phi_j^*(2)\frac{1}{\mathbf{r_{12}}}\Phi_j(2)\\\hat{K}_j(\mathbf{1})\Phi_i(\mathbf{1})&=\left[\int d\mathbf{r}_2\Phi_j^*(2)\frac{1}{\mathbf{r_{12}}}\Phi_i(2)\right]\Phi_j(\mathbf{1}).\end{aligned}\tag{3.12}$$

Equation 3.11 can be written as,

$$\hat{H}^{\text{core}}(\mathbf{1})\Phi_i(\mathbf{1})+\sum_{j\neq i}^{N}\hat{J}_j(\mathbf{1})\Phi_i(\mathbf{1})-\sum_{j\neq i}^{N}\hat{K}_j(\mathbf{1})\Phi_i(\mathbf{1})=\sum_j\varepsilon_{ij}\Phi_j(\mathbf{1}).\tag{3.13}$$

Introducing the Lagrangian multipliers, the Equation 3.13 can be further simplified as,

$$\hat{f}\Phi_i=\varepsilon_i\Phi_i.\tag{3.14}$$

Here, $\hat{f}$ is called the Fock operator, which can be described as follows:

$$\hat{f}_i(1)=\hat{H}^{\text{core}}(1)+\sum_{j=1}^{N}\left\{\hat{J}_j(1)-\hat{K}_j(1)\right\}.\tag{3.15}$$

For a closed-shell system, the Fock operator is written as,

$$\hat{f}_i(1) = \hat{H}^{\text{core}}(1) + \sum_{j=1}^{N/2} \left\{ 2\hat{J}_j(1) - \hat{K}_j(1) \right\}. \tag{3.16}$$

In conclusion, the HF method, which transforms the high-dimensional practically unsolvable multi-electron problem into a solvable problem of a single electron moving in an effective potential field, is the most fundamental method in quantum chemistry. However, this approach has some shortcomings.

Firstly, the electrons in the model do not interact with each other directly and instantaneously, but indirectly interact through the mean field, which neglects the dynamical correlation of the electrons. Secondly, when the system needs to be simplified, the wave function needs to be represented by a linear combination of multiple Slater determinants. However, the HF equation only employs a single Slater determinant to represent the wave function ignoring the static correlation of electrons. Therefore, the HF approximation is reasonable for molecular or cluster systems where exchange interactions are dominant. However, in periodic systems where the correlation effects between electrons cannot be ignored, the HF approximation does not give accurate results [24].

3.3. Density Functional Theory

ab initio electronic structure theories, such as the HF approach, are based on complex multi-electron wave functions. Specifically, in the HF approximation, for a system with N electrons, the wave functions depend on $3N$ spatial variables, making calculations increasingly difficult as the system gets larger and the number of electrons increases. In contrast, the density functional theory (DFT) considers the overall electron density distribution instead of whole wave functions in the calculation. Since the electron density is a function of only three spatial variables, the number of electronic density variables remains constant when the number of electrons increases. In other words, under the DFT theory, the degrees of freedom of the N-electron system are independent of the size of the system. This makes the DFT theory much simpler to deal with, both conceptually and practically than solving multi-electron wave functions.

3.3.1. Hohenberg-Kohn theorem

Hohenberg and Kohn proposed the Hohenberg-Kohn theorem in 1964, which is the fundamental basis of the DFT theory [25]. The Hohenberg-Kohn theorem contains two theorems. First, the external potential V_{ext} in any multi-electron system can be uniquely determined by the ground state charge density ρ of that system. Second, for any external potential V_{ext}, the true ground state charge density gives the global minimum of the energy functional and is equal to the ground state energy [24].

In the framework of the Hohenberg-Kohn theorem, the energy functional has the following form:

$$\begin{aligned} E[\rho] &= T[\rho] + E_{\text{ee}}[\rho] + \int V_{\text{ext}}(\mathbf{r})\rho(\mathbf{r})\mathrm{d}\mathbf{r} \\ &= F[\rho] + \int V_{\text{ext}}(\mathbf{r})\rho(\mathbf{r})\mathrm{d}\mathbf{r} \end{aligned} \tag{3.17}$$

Here, the $F[\rho]$ denotes the sum of the kinetic energy $T[\rho]$ and the inter-electronic energy $E_{\text{ee}}[\rho]$ of the system, and is a general functional that is independent of the external potential and depends only on the ground state electron density ρ.

The first theorem ensures that the ground state properties of the multi-electron system are all functionals of the electron density, while the second theorem provides a variational principle to calculate the total ground state energy and electron density of the system. Although in principle, the ground state properties of the system can be obtained from the ground state charge density, the exact form of the generalised functional is still unknown. In addition, the Hohenberg-Kohn theorem only deals with the ground state of the system. When studying the properties of excited states, corresponding extensions of the theory will be needed, such as the Time-dependent density-functional theory (TD-DFT).

3.3.2. Kohn-Sham equation

The Hohenberg-Kohn theorem does not provide a specific expression for the energy functional $F[\rho]$. Kohn and Sham proposed the Kohn-Sham equation in 1965 [26]. They assumed that there are equivalent non-interacting electron systems with the same charge densities at the ground state in any interacting multi-electron system. Although there is no rigorous theoretical proof of this assumption, the widespread use of the DFT theory based on the Kohn-Sham equation indirectly demonstrates its plausibility. Based on this assumption, the functional $E[\rho]$ in Equation 3.17 can be written as:

$$E[\rho] = T_\mathrm{s}[\rho] + E_\mathrm{H}[\rho] + E_\mathrm{xc}[\rho] + \int V_\mathrm{ext}(\mathbf{r})\rho(\mathbf{r})\mathrm{d}\mathbf{r}. \tag{3.18}$$

Here, ρ can be characterised by the orbital of the non-interacting electron system, $T_\mathrm{s}[\rho]$ is the kinetic energy, $E_\mathrm{H}[\rho]$ is the electron-electron Coulomb energy, $E_\mathrm{xc}[\rho]$ is the exchange-correlation energy and V_ext is the external potential for the interacting system, which are defined as:

$$\begin{aligned}
\rho(\mathbf{r}) &= \sum_{i=1}^{N} \psi_i^*(\mathbf{r})\psi_i(\mathbf{r}) \\
T_\mathrm{s}[\rho] &= \sum_{i=1}^{N} \int \psi_i^*(\mathbf{r}) \left(-\frac{\nabla^2}{2}\right) \psi_i(\mathbf{r})\mathrm{d}\mathbf{r} \\
E_\mathrm{H}[\rho] &= \frac{1}{2} \iint \frac{\rho(\mathbf{r}_1)\rho(\mathbf{r}_2)}{|\mathbf{r}_1 - \mathbf{r}_2|}\mathrm{d}\mathbf{r}_1\mathrm{d}\mathbf{r}_2 \\
E_\mathrm{xc}[\rho] &= (T[\rho] - T_\mathrm{s}[\rho]) + (E_\mathrm{ee}[\rho] - E_\mathrm{H}[\rho])
\end{aligned} \tag{3.19}$$

$E_\mathrm{xc}[\rho]$ corrects for unknown many-body interaction errors introduced by replacing the kinetic energy $T[\rho]$ and interaction energy $E_\mathrm{ee}[\rho]$ of the interacting electron system with the kinetic energy $T_\mathrm{s}[\rho]$ and Coulomb potential $E_\mathrm{H}[\rho]$ of the non-interacting electron system, which contains the full range of many-body effects beyond the HF approximation.

Applying the appropriate variational condition yields the one-electron Kohn-Sham equation as:

$$\left\{-\frac{\nabla_1^2}{2} - \left(\sum_{A=1}^{M} \frac{Z_A}{\mathbf{r}_{1\mathrm{A}}}\right) + \int \frac{\rho\left(\mathbf{r}_2\right)}{\mathbf{r}_{12}}\mathrm{d}\mathbf{r}_2 + V_\mathrm{xc}\left(\mathbf{r}_1\right)\right\} \psi_i\left(\mathbf{r}_1\right) = \varepsilon_i \psi_i\left(\mathbf{r}_1\right) \tag{3.20}$$

In Equation 3.20, the exchange-correlation functional is presented by

$$V_\mathrm{xc}(\mathbf{r}) = \frac{\delta E_\mathrm{xc}[\rho]}{\delta\rho(\mathbf{r})}. \tag{3.21}$$

The Kohn-Sham equations need to be solved iteratively and self-consistently due to the dependence between the Hamiltonian, the charge density and the orbital. The orbital ψ_i and the eigenvalue ε_i have no real physical meaning, and the calculation of the total energy of the interacting electron system requires a combination of Equations 3.18 and 3.20 [24, 26–28].

$$E = \sum_i \varepsilon_i - E_{\mathrm{H}}[\rho] + E_{\mathrm{xc}}[\rho] - \int V_{\mathrm{xc}}(\mathbf{r})\rho(\mathbf{r})\mathrm{d}\mathbf{r} \tag{3.22}$$

3.3.3. Exchange Correlation Functionals

Theoretically, the Kohn-Sham equation is strictly accurate for the multi-electron system, but the exact form of the exchange-correlation energy functional E_{xc} is still unknown. This functional can be so complex that it is difficult to find an exact expression, and only some parametric approximations can be made to it. Although the relevant theory of Kohn-Sham DFT theory was proposed in 1965, it did not gain popularity until the 1980s, when the exchange-correlation functional could be fitted by quantum Monte Carlo approaches. A major development in DFT theory has therefore been the search for suitable forms of exchange-correlation functionals, which are graphically classified by Jacob's ladder proposed by Perdew and Karla [29], as shown in Figure 3.2. This subsection gives a brief introduction to the commonly used exchange-correlation functionals.

3.3.3.1. Local-Density Approximation

The *Local Density Approximation* (LDA) was first proposed by Kohn and Sham [26] under the assumption of a homogeneous electron gas and is the simplest exchange-correlation functional approximation. This approximation suggests that when the spatial variation in electron density is sufficiently slow, the exchange correlation energy is only related to the local electron density and can be expressed as,

$$E_{\mathrm{xc}}^{\mathrm{LDA}}[\rho] = \int \epsilon_{\mathrm{xc}}^{\mathrm{LDA}}(\rho(\mathbf{r}))\rho(\mathbf{r})\mathrm{d}\mathbf{r} \tag{3.23}$$

Here ϵ_{xc} is the exchange-correlation potential for the homogeneous electron gas without parsed expression, and the parameters are generally fitted by

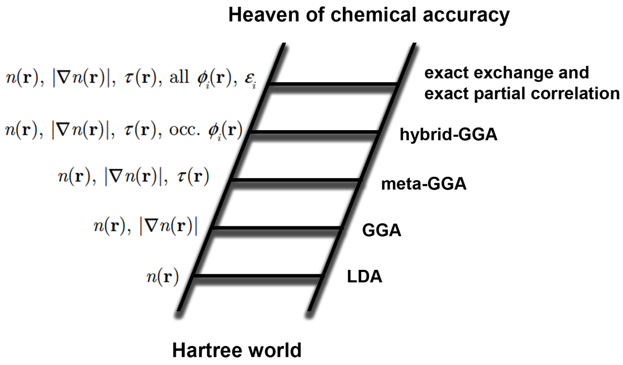

Heaven of chemical accuracy
$n(\mathbf{r})$, $|\nabla n(\mathbf{r})|$, $\tau(\mathbf{r})$, all $\phi_i(\mathbf{r})$, ε_i
exact exchange and exact partial correlation
$n(\mathbf{r})$, $|\nabla n(\mathbf{r})|$, $\tau(\mathbf{r})$, occ. $\phi_i(\mathbf{r})$
hybrid-GGA
$n(\mathbf{r})$, $|\nabla n(\mathbf{r})|$, $\tau(\mathbf{r})$
meta-GGA
$n(\mathbf{r})$, $|\nabla n(\mathbf{r})|$
GGA
$n(\mathbf{r})$
LDA
Hartree world

Figure 3.2.: Jacob's ladder diagram for the DFT theory. The higher order functionals contain additional new information on top of the lower order ones: the LDA takes into account the electron density, the GGA takes into account more density gradients, the meta-GGA takes into account second order derivatives of the density, the fourth level takes into account the exact exchange energy of the occupied state orbitals, while the fifth level takes into account the effect of the unoccupied states.

Monte Carlo simulations. LDA is widely used in material science due to its simple form and reasonable results. In general, LDA is relatively accurate for solid rather than molecules and generally gives good structural and elastic properties estimates. However, since the LDA treats the electron density as the same in the system, it tends to overestimate the correlation energy and underestimate the exchange energy. In addition, it has the following drawbacks: overestimation of binding energies, underestimation of reaction activation energies, excessive preference for high spin structures, misestimation of phase stability, etc.

3.3.3.2. Generalised Gradient Approximation

In general, LDA does not perform well in systems where the electron density changes rapidly, so the easier way to improve it is to take the electron density gradient into account. This expansion is referred to as *Generalised Gradient Approximation* (GGA), and the expression is:

$$E_{\mathrm{xc}}^{\mathrm{GGA}}[\rho] = \int \epsilon_{\mathrm{xc}}^{\mathrm{GGA}}(\rho(\mathbf{r}), \nabla\rho(\mathbf{r}))\rho(\mathbf{r})\mathrm{d}\mathbf{r}. \tag{3.24}$$

There are no analytic expressions for the exchange and correlation energy in GGA, and there are currently two main schools of thought on how to construct them: one school, represented by Becke, advocates the introduction of a large number of parameters to obtain a more accurate functional [30]; the other school, represented by Perdew, believes that the form of the functional should follow the basic physical laws and pay more attention to the scalar relationship and asymptotic behaviour[31]. Currently, the commonly used GGA functionals are PBE [32], BLYP [33] and so on.

Compared with the LDA, the GGA functional contains an extra portion of long-range correlations, which not only allows for a better description of non-uniform charge density systems, but also provides a good correction for the binding energy of the system. Therefore, the GGA functional generally achieves a more accurate energy and structure. However, the GGA results are not always more accurate than LDA and are even worse in some cases. For example, the GGA functionals usually give lower bond energies. Both LDA and GGA contain only electron density or density gradients instead of the Kohn-Sham orbitals explicitly, hence they are also known as pure density functionals[24, 34].

3.3.3.3. Hybrid Functionals

The exact exchange energy of the system is obtained with HF approximation. Therefore, in order to improve the deficiencies of the LDA as well as the GGA functional, the HF exchange energy has been mixed with the LDA/GGA in a certain ratio to obtain a new functional called the hybrid functional with the expression as:

$$E_{\mathrm{xc}} = \int_0^1 \mathrm{d}\lambda U_{\mathrm{xc}}^{\lambda} \tag{3.25}$$

In hybrid functionals, the HF exchange potential is incorporated into the exchange potential, and its calculation results are significantly improved compared to the calculation results of GGA. After the most important exchange-correlation hybrid functional B3LYP was proposed in 1994 [35, 36], the DFT method quickly became popular. The B3LYP became a general method for the computational study of various problems in physical chemistry. It employs three parameters for the mixing of exchange-correlation energies and can be expressed as,

$$E_{\mathrm{xc}}^{\mathrm{B3LYP}} = aE_{\mathrm{x}}^{\mathrm{HF}} + (1-a)E_{\mathrm{x}}^{\mathrm{LSDA}} + bE_{\mathrm{x}}^{\mathrm{B88}} + cE_{\mathrm{c}}^{\mathrm{LYP}} + (1-c)E_{\mathrm{c}}^{\mathrm{VWN}} \tag{3.26}$$

where $a = 0.2$, $b = 0.72$ and $c = 0.81$. $E_\mathrm{x}^\mathrm{B88}$ is the Becke 88 exchange functional with the generalised gradient approximation [30], $E_\mathrm{c}^\mathrm{LYP}$ is the Lee-Yang-Parr correlation functional with the gradient term [33] and $E_\mathrm{c}^\mathrm{VWN}$ is the local spin density approximation (LSDA, an unrestricted extension to LDA) to the correlation functional from Vosko, Wilk and Nusair [37].

B3LYP is a semi-empirical hybrid functional, and the coefficients a, b and c are obtained by fitting to a large number of data such as atomisation energies, electron and proton affinities, and ionisation energies. This functional is mainly used in chemistry and performs well in describing the ground state geometry and electronic structure of single molecules. However, it also has many shortcomings, such as a decrease in calculation accuracy as the molecular system increases, a significant underestimation of full reaction energies, bond dissociation energies and isomerisation energies, and an unsatisfactory estimation accuracy for the thermochemical properties of small and medium-sized systems. By adjusting the ratio of HF exchange energy, different forms of hybrid functionals can be obtained, such as PBE0 [38], HSE06 [39] and so on. In this work, all DFT calculations applied B3LYP as exchange-correlation functionals.

3.4. Density Functional Tight Binding

Compared to most wave function-based QM methods, DFT methods can produce accurate results in solving many chemical problems while at the same time significantly reducing computational costs. Nevertheless, the computational speed of the DFT method is still not satisfactory when calculating larger systems. To address this problem, semi-empirical and approximations have been applied to QM methods, and semi-empirical wave function-based QM methods, such as CNDO [40], MINDO [41] and PM3 [42] have been derived. At the same time, density-functional tight-binding (DFTB) based on DFT theory was also proposed, which can improve the computational speed by 2–3 orders of magnitude with only a slight loss of accuracy [43, 44]. The tight-binding model is from solid-state physics, which assumes that electrons are tightly bound to their nuclei and hence highly localised in space. The electron density $\rho[\mathbf{r}]$ can be defined by the combination of a reference density ρ^0 and its fluctuations as,

$$\rho(\mathbf{r}) = \rho^0(\mathbf{r}) + \delta\rho(\mathbf{r}). \tag{3.27}$$

Adding the approximated electron density in the DFT total energy (Equation 3.22) with the additional nuclei–nuclei repulsive energy E_{NN} yields the total energy for DFTB as,

$$
\begin{aligned}
&E[\rho^0(\mathbf{r}) + \delta\rho(\mathbf{r})] = \\
&\sum_i \left\langle \psi_i \middle| -\frac{\hbar^2}{2m}\nabla^2 + V_{\mathrm{eN}} + \int \frac{(\rho_1^0 + \delta\rho_1)(\rho_2^0 + \delta\rho_2)}{|\mathbf{r}_1 - \mathbf{r}_2|} \mathrm{d}\mathbf{r}_2 + V_{\mathrm{xc}}[\rho^0 + \delta\rho] \middle| \psi_i \right\rangle \\
&- \frac{1}{2} \iint \frac{(\rho_1^0 + \delta\rho_1)(\rho_2^0 + \delta\rho_2)}{|\mathbf{r}_1 - \mathbf{r}_2|} \mathrm{d}\mathbf{r}_1 \mathrm{d}\mathbf{r}_2 \\
&- \int V_{\mathrm{xc}}[\rho^0 + \delta\rho] \left(\rho^0 + \delta\rho\right) \mathrm{d}\mathbf{r} + E_{\mathrm{xc}}[\rho^0 + \delta\rho] + E_{\mathrm{NN}}.
\end{aligned}
\tag{3.28}
$$

Expanded in a Taylor series for the exchange-correlation energy, leading to the following expression for the total energy [45]:

$$
\begin{aligned}
E^{\mathrm{DFTB}} = &- \frac{1}{2} \iint \frac{\rho_1^0 \rho_2^0}{|r_1 - r_2|} \mathrm{d}\mathbf{r}_1 \mathrm{d}\mathbf{r}_2 - \int V_{\mathrm{xc}}[\rho^0]\rho^0 \mathrm{d}\mathbf{r} + E_{\mathrm{xc}}[\rho^0] + E_{\mathrm{NN}} \\
&+ \sum_i \left\langle \psi_i \middle| -\frac{\hbar^2}{2m}\nabla^2 + V_{\mathrm{eN}} + \int \frac{\rho_2^0}{|\mathbf{r}_1 - \mathbf{r}_2|} \mathrm{d}\mathbf{r}_2 + V_{\mathrm{xc}}[\rho^0] \middle| \psi_i \right\rangle \\
&+ \frac{1}{2} \iint \left(\frac{1}{|\mathbf{r}_1 - \mathbf{r}_2|} + \left. \frac{\partial^2 E_{\mathrm{xc}}[\rho]}{\partial\rho\partial\rho_2} \right|_{\rho_1^0, \rho_2^0} \right) \delta\rho_1 \delta\rho_2 \mathrm{d}\mathbf{r}_1 \mathrm{d}\mathbf{r}_2 \\
&+ \frac{1}{6} \iiint \left. \frac{\partial^3 E_{\mathrm{xc}}[\rho]}{\partial\rho_1\partial\rho_2\partial\rho_3} \right|_{\rho_1, \rho_2, \rho_3} \delta\rho_1 \delta\rho_2 \delta\rho_3 \mathrm{d}\mathbf{r}_1 \mathrm{d}\mathbf{r}_2 \mathrm{d}\mathbf{r}_3 + \ldots \\
&= E^0\left[\rho^0\right] + E^1\left[\rho^0, \partial\rho\right] + E^2\left[\rho^0, (\partial\rho)^2\right] + E^3\left[\rho^0, (\partial\rho)^3\right] + \ldots
\end{aligned}
\tag{3.29}
$$

When only the first order term of the Taylor expansion is considered, the standard DFTB is obtained, also known as the non-self-consistent DFTB. If the second order term is further included, the DFTB2 is obtained, also called Self-consistent charge density functional tight-binding (SCC–DFTB). If the third order term is further included, the DFTB3 is obtained.

The first line of Equation 3.29 ($E^0[\rho^0]$) contributes to the short-range two-body repulsive energy E^{rep}, which contains electron interactions, exchange-correlation contribution and the nuclei-nuclei repulsive energy. In DFTB, this term depends on the reference density ρ_0 and is usually approximated as a sum of pairwise potentials,

$$E^0 \approx E^{\text{rep}} = \frac{1}{2} \sum_{ab,a\neq b} V_{ab}^{\text{rep}} \left[\rho_a^0, \rho_b^0, r_{ab}\right] \tag{3.30}$$

The second line of Equation 3.29 ($E^1[\rho^0, \delta\rho]$) only depends on the reference density and is obtained by an atomic orbital Hamiltonian. The total energy of the standard DFTB approach can be written as,

$$E^{\text{DFTB1}}[\rho(\mathbf{r})] = \sum_i \epsilon_i + E^{\text{rep}} \tag{3.31}$$

Since the standard DFTB method ignores the density fluctuation $\delta\rho(\mathbf{r})$, it cannot precisely describe systems with charge transfer. Therefore, it is often applied for solid state physics, such as unpolar crystals.

The third line of Equation 3.29 ($E^2[\rho^0, (\delta\rho)^2]$) takes the charge fluctuation into account. The energy of this term can be described by the sum of charge fluctuations as,

$$E^2 \approx E^{\gamma} = \frac{1}{2} \sum_{ab} \delta q_a \delta q_b \gamma_{ab} \tag{3.32}$$

Here, $\delta q_a = q_a - q_a^0$ indicates the net charge of atom a, defined as the Mulliken charge. γ contributes to the second order approximation, which describes the interaction between charge functionals.

Two cases need to be discussed for γ: first, when two atoms a and b have a long distance, i.e. $r_{ab} \to \infty$, γ_{ab} describes the long distance Coulomb interaction of partial charges δq_a and δq_b as $1/r_{ab}$; second, for $a = b$, i.e. $r_{ab} \to 0$, γ_{aa} indicates the on-site self-repulsion, and can be described by the Hubbard parameter U_a [45]. U_a is the twice hardness of an atom and related to the size of the atom. The total energy of the DFTB2 approach can be written as,

$$E^{\text{DFTB2}}[\rho(\mathbf{r})] = \sum_i \epsilon_i + E^{\text{rep}} + \frac{1}{2} \sum_{ab} \delta q_a \delta q_b \gamma_{ab}. \tag{3.33}$$

Since charge fluctuations have been introduced, DFTB2 is able to handle polar systems, such as biomolecules. However, such an approach still has a drawback due to the fixed atom size as well as the Hubbard parameter. In fact, the effective size of an atom varies as charge accumulates. To improve the reliability of the calculation, additional corrections for U_a are required.

The fourth line of Equation 3.29 ($E^3[\rho^0, (\delta\rho)^3]$) introduces the derivative of γ_{ab} with respect to the charge. Such Hubbard derivatives can be pre-calculated and increase the calculation's accuracy without paying additional time. The energy of this term can be written as,

$$E^3 = \frac{1}{3}\sum_{ab} (\delta q_a)^2 \, \delta q_b \Gamma_{ab} \tag{3.34}$$

where,

$$\begin{aligned} \text{if } a \neq b, \Gamma_{ab} &= \left(\frac{\partial \gamma_{ab}}{q_a}\right)_{q_a^0} = \left(\frac{\partial \gamma_{ab}}{U_a}\frac{\partial U_a}{q_a}\right)_{q_a^0} \\ \text{if } a \neq b, \Gamma_{ba} &= \left(\frac{\partial \gamma_{ab}}{q_b}\right)_{q_b^0} = \left(\frac{\partial \gamma_{ab}}{U_b}\frac{\partial U_b}{q_b}\right)_{q_b^0} \\ \text{if } a = b, \Gamma_{aa} &= \left(\frac{\partial \gamma_{aa}}{q_a}\right)_{q_a^0} = \left(\frac{\partial \gamma_{aa}}{U_a}\frac{\partial U_a}{q_a}\right)_{q_a^0} . \end{aligned} \tag{3.35}$$

The total energy of the DFTB3 approach is expressed as,

$$E^{\text{DFTB3}}[\rho(\mathbf{r})] = \sum_i \epsilon_i + E^{\text{rep}} + \frac{1}{2}\sum_{ab} \delta q_a \delta q_b \gamma_{ab} + \frac{1}{3}\sum_{ab} (\delta q_a)^2 \, \delta q_b \Gamma_{ab}. \tag{3.36}$$

With the introduction of the Hubbard derivatives term, DFTB3 now has the ability to describe highly polar systems.

4. Molecular Mechanics

After decades of development, the computational speed of quantum mechanics approaches has improved considerably, but it still cannot tackle many problems in molecular simulations, such as large biomolecular systems. In fact, for biomacromolecular systems, a reaction typically takes nanoseconds or even microseconds in real time, whereas existing QM methods can only afford to perform calculations within a few nanoseconds (only for semi-empirical methods). In addition, biomacromolecular systems often contain more than 100,000 atoms, which also makes the computation extremely time-consuming.

To compensate for the shortcomings of quantum mechanics, molecular mechanics (MM) methods applying empirical potential functions to describe the interactions between particles have been proposed. The molecular mechanics approach calculates the motion of individual particles in a system by the classical Newtonian mechanics equations. Although the molecular force field (FF) approach is not as accurate as the quantum mechanical approach, its application in biomacromolecular systems is beyond the reach of the QM approach.

4.1. Molecular Force Field

Molecular force fields are created by physical or chemical experiments, quantum chemical calculations, or both, and include both bonding and non-bonding interactions. Bonding interactions consist of bond length stretching potentials, bond angle bending potentials, dihedral angle twisting potentials and cross-interaction terms. The electrostatic interaction energy and van der Waals interaction energy make up the non-bonding interactions.

$$E_{\mathrm{total}} = E_{\mathrm{S}} + E_{\mathrm{B}} + E_{\mathrm{Tor}} + E_{\mathrm{cross}} + E_{\mathrm{vdw}} + E_{\mathrm{ele}} \tag{4.1}$$

4.1.1. Potential Functional Forms

The bond length stretching potential is described as energy changes due to the stretching motion of the chemical bonds within a molecule along the bond axis. It is usually described using the harmonic oscillator function (*Hooke's law*) as,

$$E_{\mathrm{S}} = \frac{1}{2} \sum_{ij} k_{ij}^{b} \left(r_{ij} - r_{ij}^{0} \right)^{2} \tag{4.2}$$

where k_{ij}^{b} indicates the bond stretching elasticity constant, r_{ij} is the distance between atoms i and j, and r_{ij}^{0} is the equilibrium distance. The harmonic oscillator function is the common functional form of the CHARMM [46] and AMBER [47] force fields. However, when the bond length deviates far from the equilibrium distance, the Morse potential function model is necessary. Although such Morse potential function model is more accurate, it is inevitably more time consuming, and hence not suitable for macromolecules. Nevertheless, the MM2 force field is developed based on such model, and its successors MM3 and MM4 force fields, can be used effectively to obtain reasonable results for organic small molecules [48–50].

The bond angles formed by three consecutive atoms in a molecule vibrate around the equilibrium bond angle and can be described by Hooke's law as,

$$E_{\mathrm{B}} = \frac{1}{2} \sum_{ijk} k_{ijk}^{a} \left(\theta_{ijk} - \theta_{ijk}^{0} \right)^{2} \tag{4.3}$$

where k_{ijk}^{a} indicates the bond angle bending constant, θ_{ijk} is the bond angle formed by atoms i, j and k, θ_{ijk}^{0} is the equilibrium angle. Here, when the bond angle deviates from the equilibrium angle by less than 10°, the harmonic oscillator model is reasonable. When the bond angle deviates beyond such value, higher-order correction terms need to be introduced. For example, the MM2 force field additionally contains a quartic term to improve the accuracy [48].

The dihedral angle is formed by four successively bonded atoms in a molecule. Dihedral angles are easily torsional and the Fourier series is commonly used to describe rotational potentials:

$$E_{\mathrm{Tor}} = \frac{1}{2} \sum_{n_{ijkl}}^{N} V_{n} \left[1 + \cos \left(n\omega_{ijkl} - \omega_{ijkl}^{0} \right) \right] \tag{4.4}$$

where V_n indicates torsional barrier height, n is the multiplicity, ω_{ijkl} is the dihedral angle, and the ω^0_{ijkl} is the phase factor which determines the position of the dihedral angle when it passes through its the minimum value.

In addition, when one of the three – bond length, bond angle and dihedral angle – is changed, accordingly, the other two are affected, i.e. the stretching potential energy of bond lengths, the bending potential energy of bond angles and the twisting potential energy of dihedral angles within a molecule are interconnected, and hence a so-called cross-interaction term needs to be considered. For instance, in the case of interaction between angle bending and dihedral twisting, the bend-torsion cross-interaction term is written as,

$$E^{\mathrm{BT}}_{\mathrm{cross}} = \frac{k_{\mathrm{BT}}}{2}(\theta - \theta_0)(1 - \cos 3\omega). \tag{4.5}$$

For consecutive atoms bonding atoms i,j,k, the change in bond length between i and j also affects the bond length between j and k, leading to the stretch-stretch cross-interaction term as,

$$E^{\mathrm{SS}}_{\mathrm{cross}} = \frac{k_{\mathrm{BT}}}{2}(r_{ij} - r_{ij,0})(r_{jk} - r_{jk,0}). \tag{4.6}$$

In fact, however, the introduction of cross-interaction terms is only necessary in a very small number of cases to reproduce precise structural properties. When dealing with macromolecular systems, these terms are usually ignored in order to save the computational costs.

The interactions discussed above are all bonding interactions, and we will now discuss non-bonding interactions, van der Waals (vdW) interactions and electrostatic interactions. Van der Waals interactions do not have a direct bonding connection and depend on the distance between two atoms. When the two atoms are close together, the vdW interaction between the two atoms is expressed as repulsion. When the two atoms are far apart, the vdW interaction is expressed as an attraction as a result of the instantaneous dipole moment or dispersion forces. The vdW interaction is usually described using the Lennard-Jones (LJ) potential function as,

$$E_{\mathrm{vdw}} = 4\varepsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right] \tag{4.7}$$

where ε refers to the potential well depth, σ is the distance at which the particle-particle potential energy is zero (i.e. the sum of the vdW radii of two atoms), and r is the distance between two interacting particles. The attractive interaction varies as r^{-6} and the repulsive interaction varies as r^{-12} in the LJ potential function. This 6-12 potential is commonly used in AMBER and CHARMM force field [46, 47], which are designed to deal with large systems.

Electrostatic interactions are mainly obtained by the dipole moment method and the point charge approach. In the dipole moment method, the electrostatic interactions are described by calculating dipole-dipole interactions as,

$$E_{\text{ele}} = \frac{\mu_{ij}\mu_{kl}}{Dr^3_{ij/kl}} \left(\cos\chi - 3\cos\alpha_{ij}\cos\alpha_{kl}\right). \tag{4.8}$$

Here, μ_{ij} and μ_{kl} represent the dipole moment of the bond between atoms i and j, and the bond between atoms k and l, respectively. D is the effective dielectric constant, $r_{ij/kl}$ is the distance between the centre of dipole moment μ_{ij} and μ_{kl}, χ and α are angles of the corresponding dipole moment.

In biomacromolecular systems, to save the computational costs, the point charge approach is usually applied to calculate the electrostatic interactions as,

$$E_{\text{coulomb}} = \sum_{i,j} \frac{q_i q_j}{Dr_{ij}} \tag{4.9}$$

where q_i and q_j are net charges of atoms i and j.

Both two non-bonding interactions weaken with increasing distance between the particles, such that the effect on the system can be ignored when a certain distance is exceeded. Therefore, in order to save the computational costs, vdW and electrostatic interactions are often set with a cut-off distance. And the interactions between two atoms are neglected when their distance is beyond the threshold.

4.1.2. Common Force Fields

The first molecular force fields for predicting molecular structure, vibrational spectra and enthalpies of isolated molecules originated in the 1960s [51]. These force fields were mainly used for small organic molecules, and some of them have continued to be developed and employed. The best examples are the MM2, MM3, and MM4 force fields developed by Allinger et al [48–

50]. With the development of molecular dynamics simulations, the scope of molecular force field research has shifted to dealing with more complex systems, and force fields have been developed to be more widely applicable. The common force fields in molecular dynamics simulations and in recent years are shown in Table 4.1.

Table 4.1.: Classical force fields and their applicable systems

Force Field	Applicable System	Ref.
AMBER	proteins, DNAs	[47, 52, 53]
CHARMM	small molecules, macromolecules	[46, 54]
CVFF	small molecules, macromolecules	[55, 56]
GROMOS	biomolecules	[57, 58]
MMX	small molecules	[48–50]
OPLS	liquid	[59, 60]

AMBER Force Field: The AMBER (Assisted model building with energy refinement) force field is proposed by Kollman et al., and is a widely used force field at present. This force field was initially only applicable for the study of protein and nucleic acid systems, but after many years of development, the AMBER force field, with the introduction of the GAFF (General Amber force field) force field, is now also applicable to certain small molecule and polymer systems. The parameters of this force field are derived from the comparison of calculated and experimental values. The functional form of the AMBER force field is,

$$V = \sum_{\text{bonds}} k_b \left(b - b_0\right)^2 + \sum_{\text{angles}} k_\theta \left(\theta - \theta_0\right)^2 + \sum_{\text{dihedrals}} \frac{1}{2} V_0 \left[1 + \cos\left(n\varphi - \phi_0\right)\right] \\ + \sum_{i<j} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}}\right) + \sum_{i<j} \frac{q_i q_j}{\varepsilon_{ij} r_{ij}} + \sum_{\text{H−bonds}} \left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}}\right). \tag{4.10}$$

CHARMM Force Field: The CHARMM (chemistry at Harvard macromolecular mechanism) force field is suitable for calculations and simulations of a wide range of properties, and it is supported by results from QM calculations of the interactions between model compounds and water.

CVFF Force Field: The CVFF (consistent valence force field) series force field was originally developed by Dauber et al. for biomolecular systems [55].

Later on, the CVFF force field was developed to be applicable also to large protein systems.

GROMOS Force Field: The GROMOS (Groningen Molecular Simulation) force field is mainly used for molecular dynamics simulations in the GROMACS software package. As a united-atom force field, it is suitable for a wide range of chemical substances, from n-alkanes to biopolymers.

MMX Force Field: Molecular mechanic force field is developed by Allinger et al. [48–50]. This series of force fields started out as the MM2 force field, and later developed into the MM3 and MM4 force fields. The MMX force field algorithm is complex (shown in Equation 4.11), as it subdivides some common atoms with different force field parameters. MMX can also be applied to simulate biomacromolecules, but it is time consuming.

$$\begin{aligned} V = &\sum E_{\mathrm{S}} + \sum E_{\mathrm{B}} + \sum E_{\mathrm{Tor}} + \sum E_{\mathrm{OOP}} + \sum E_{\mathrm{SB}} \\ &+ \sum E_{\mathrm{vdW}} + \sum E_{\mu} + \sum E_{\mathrm{ele}}. \end{aligned} \tag{4.11}$$

OPLS Force Field: The OPLS (Optimized Potentials for Liquid Simulations) force field is used to simulate the peptides and organic molecules, and contains two kinds of force fields: the OPLS-AA as an all-atom force field and the OPLS-UA as a united-atom force field. The bond-stretching and angle-bending parameters of the OPLS-AA, are mainly adjusted based on the AMBER force field and the torsional parameter is from *ab initio* calculations.

4.2. Molecular Dynamics

The essence of molecular dynamics (MD) simulations is to apply computer calculations to describe the movement and interaction of atoms and molecules in macromolecular systems by Newton's second law of motion. With the rapid development of computer science, the MD calculations are increasingly used in biological systems. Currently, the MD calculations are applied to study protein stability, protein folding, protein-ligand interactions, molecular recognition and so on.

For a system, the total energy is the sum of the molecular kinetic and potential energies of the N particles that make up the system. Here, the total potential energy can be obtained by the position function $V(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n)$ of each particle inside the system.

The force on any atom i in the system is its potential energy gradient:

$$\vec{F}_i = -\nabla_i V = -\left(\vec{i}\frac{\partial}{\partial x_i} + \vec{j}\frac{\partial}{\partial y_i} + \vec{k}\frac{\partial}{\partial z_i}\right) \tag{4.12}$$

From force, the acceleration $\vec{a}_i$ of atom i can be determined by means of Newton's equation of motion and written as,

$$\vec{a}_i = \frac{\vec{F}_i}{m_i}. \tag{4.13}$$

After a certain time t, the velocity $\vec{v}_i$ and position $\vec{r}_i$ of atom i can be obtained by integrating over time t as,

$$\begin{aligned} &\frac{d^2}{dt^2}\vec{r}_i = \frac{d}{dt}\vec{v}_i = \vec{a}_i \\ &\vec{v}_i = \vec{v}_i^0 + \vec{a}_i t \\ &\vec{r}_i = \vec{r}_i^0 + \vec{v}_i^0 + \frac{1}{2}\vec{a}_i t^2. \end{aligned} \tag{4.14}$$

The superscript 0 represents the initial value of each physical quantity. By making $t = \delta t$, the positions and velocities of the individual atoms during the simulation time are obtained by integrating over small stages with fixed time δt.

Finite difference techniques are common ways to solve Newton's equations of motion in MD simulations. The *Verlet* algorithm is one of the most commonly used algorithms and the Taylor expansion of the particle positions is as follows [61]:

$$\begin{aligned} \mathbf{r}(t+\delta t) &= \mathbf{r}(t) + \delta t\mathbf{v}(t) + \frac{1}{2}\delta t^2\mathbf{a}(t) + \frac{1}{6}\delta t^3\mathbf{b}(t) + \frac{1}{24}\delta t^4\mathbf{c}(t) + \cdots \\ \mathbf{v}(t+\delta t) &= \mathbf{v}(t) + \delta t\mathbf{a}(t) + \frac{1}{2}\delta t^2\mathbf{b}(t) + \frac{1}{6}\delta t^3\mathbf{c}(t) + \\ \mathbf{a}(t+\delta t) &= \mathbf{a}(t) + \delta t\mathbf{b}(t) + \frac{1}{2}\delta t^2\mathbf{c}(t) + \cdots \\ \mathbf{b}(t+\delta t) &= \mathbf{b}(t) + \delta t\mathbf{c}(t) + \cdots \end{aligned} \tag{4.15}$$

where $\mathbf{v}$ indicates the velocity, $\mathbf{a}$ is the acceleration, $\mathbf{b}$ is the third derivative and so on. The Verlet algorithm can predict the position of an atom at the next moment $(t+\delta t)$ based on its position at the previous moment $(t-\delta t, t)$:

$$\mathbf{r}(t+\delta t) = \mathbf{r}(t) + \delta t \mathbf{v}(t) + \frac{1}{2}\delta t^2 \mathbf{a}(t) + \cdots \tag{4.16}$$

$$\mathbf{r}(t-\delta t) = \mathbf{r}(t) - \delta t \mathbf{v}(t) + \frac{1}{2}\delta t^2 \mathbf{a}(t) - \cdots \tag{4.17}$$

The new position can be predicted by adding Equation 4.16 to Equation 4.17 as,

$$\mathbf{r}(t+\delta t) = 2\mathbf{r}(t) - \mathbf{r}(t-\delta t) + \delta t^2 \mathbf{a}(t). \tag{4.18}$$

The current velocity can be determined by subtracting Equation 4.17 from Equation 4.16 as,

$$\mathbf{v}(t) = \frac{\mathbf{r}(t+\delta t) - \mathbf{r}(t-\delta t)}{2\delta t}. \tag{4.19}$$

Alternatively, the velocities can also be calculated at the half-step as,

$$\mathbf{v}\left(t+\frac{1}{2}\delta t\right) = \frac{[\mathbf{r}(t+\delta t) - \mathbf{r}(t)]}{\delta t}. \tag{4.20}$$

The Verlet algorithm only needs to calculate the forces once for each step of the motion and contains only two sets of coordinates, $\mathbf{r}(t)$ and $\mathbf{r}(t-\delta t)$, and their corresponding accelerations, which does not take up much memory resources. However, the Verlet algorithm lacks precision and does not include explicit velocities.

The *leap-frog* algorithm is another algorithm derived from the Verlet algorithm, which can be expressed as [62],

$$\begin{aligned} \mathbf{v}\left(t+\frac{1}{2}\delta t\right) &= \mathbf{v}\left(t-\frac{1}{2}\delta t\right) + \delta t \mathbf{a}(t) \\ \mathbf{r}(t+\delta t) &= \mathbf{r}(t) + \delta t \mathbf{v}\left(t+\frac{1}{2}\delta t\right) \end{aligned} \tag{4.21}$$

The *leap-frog* algorithm calculates the velocity $\mathbf{v}(t+\frac{\delta t}{2})$ from the velocity $\mathbf{v}(t-\frac{\delta t}{2})$ with an acceleration $\mathbf{a}(t)$, as shown the first line in Equation 4.21. Then, the position $\mathbf{r}(t+\delta t)$ is calculated by the position $\mathbf{r}(t)$ and the velocity $\mathbf{v}(t-\frac{\delta t}{2})$ obtained from the last step. The velocity at t reads as,

$$\mathbf{v}(t) = \frac{1}{2}\left[\mathbf{v}\left(t+\frac{1}{2}\delta t\right) + \mathbf{v}\left(t-\frac{1}{2}\delta t\right)\right] \tag{4.22}$$

Compared to the Verlet algorithm, the leap-frog algorithm includes explicit velocities. Besides, since only $t - \frac{\delta t}{2}$ and $\mathbf{r}(t)$ require to be recorded during calculations, the storage requirement is relatively low. In this work, the *leap-frog* algorithm is applied for most of the MD simulations.

It is crucial to select an appropriate *time step* is crucial for MD simulations. A suitable time step can ensure accuracy and save computational resources on both hands. In fact, there is no fixed value for the setting of time steps and we need to choose the appropriate value depending on the system and the simulation conditions. In general, time steps are usually set less than one tenth of the fastest period of motion in the system. For instance, if atoms in the simulation have a period of motion of 0.1 ps, a time step should be set in the femtosecond range. In the case of a water molecule, for example, which has a maximum vibration frequency of about 1.08×10^{14} Hz, the proper time step can be estimated as [24],

$$\delta t \leq \frac{1}{10} T = \frac{1}{10} \cdot \frac{1}{v} = 0.9 \times 10^{-15} \text{ s} = 0.9 \text{ fs} \tag{4.23}$$

Appropriate time steps for different types of systems are illustrated in Table 4.2 [24]:

Table 4.2.: Suggested time steps for various systems.

System	Type of motion	Time step (fs)
Atoms	Translation	10
Rigid molecules	Translation and rotation	5
Flexible molecules, rigid bonds	Translation, rotation, torsion	2
Flexible molecules, flexible bonds	Translation, rotation, torsion, vibration	1 or 0.5

In real life, each molecule can be considered as an individual unit in an infinite environment. However, with the current algorithms and techniques, only a limited number of molecules can be studied by computational chemistry, which would not be able to accurately reveal the properties of the whole system. To solve this problem, *periodic boundary conditions* (PBC) are taken into account in MD simulations.

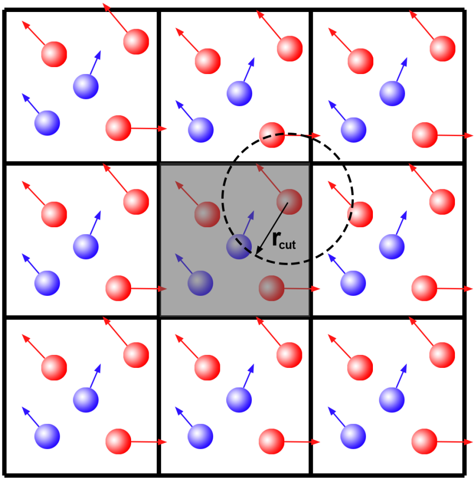


Figure 4.1.: 2D representation of periodic boundary conditions applied to a cubic simulation box. The velocity vector of each atom is indicated by an arrow and the simulation box is coloured grey. r_{cut} is the cut-off radius, which needs to be less than half the length of the simulation box.

Periodic boundary conditions employ the periodic replication of a small number of molecules to model the effect of the surrounding environment on the system. As shown in Figure 4.1, there are 8 periodic image cells around the simulation box in a two-dimensional system, and 26 in a three-dimensional system. The coordinates for replicated atoms in each image cells are taken from corresponding atoms in the simulation box. Therefore, when a particle leaves the simulation box and enters into an image cell, its replicated atom will come to the simulation box from the image cell on the other side with the same velocity.

The long-range electrostatic interactions need to be carefully considered when applying periodic boundary conditions. For a system containing N atoms, the total electrostatic potential under the PBC conditions reads as,

$$V = \frac{f}{2} \sum_{n_x} \sum_{n_y} \sum_{n_z} \sum_{i}^{N} \sum_{j}^{N} \frac{q_i q_j}{r_{ij,n}}. \tag{4.24}$$

To save computational cost while maintaining a certain accuracy, several approaches for the long-range electrostatic interactions have been proposed, such as the particle-Mesh-Ewald (PME) method used in this work [63].

In general, periodic boundary conditions come in a variety of shapes, such as cubes, hexagonal columns, truncated octahedra and so on. The cubic box is the simplest simulation box among these periodic boundary conditions. When choosing a periodic boundary condition, the appropriate shape needs to be considered depending on the type of simulation system.

For biomacromolecules, the computational cost for long-range interactions between particles will remarkably increase as the number of particles increases. To solve this issue and to enhance the calculation speed, the *cut-off* radius approach needs to be introduced. That is, when the distance between two particles exceeds the cut-off radius, van der Waals interactions as well as electrostatic interactions between them will be ignored. Note that, van der Waals interactions and electrostatic interactions have different cut-off radii, due to their different attenuation with distance. Van der Waals interactions are proportional to the r^{-6}, whereas electrostatic interactions are proportional to the r^{-1}. The decay of the electrostatic interaction with increasing distance is therefore weaker than the van der Waals interaction, and its cut-off radius is longer. The radius of cut-off for van der Waals interactions is typically 10-12 Å, while for electrostatic interactions, the radius of cut-off is typically more than 16 Å.

The concept of ensemble is introduced to describe the thermodynamic characteristics of MD simulation systems. In general, an ensemble is a collection of particles in systems with the same conditions, such as the number of particles N, the volume V, the total energy E, the temperature T, the pressure P and so on. To easily compared the simulation result and experimental result with a certain temperature and pressure, in MD simulations, commonly used ensembles include the canonical ensemble (NVT) as well as the isothermal–isobaric ensemble (NPT).

The NVT ensemble has a fixed temperature, system volume as well as number of particles. Here, the temperature is controlled by a thermostat to regulate velocities. The common used thermostats include Berendsen thermostat [64], Nosé–Hoover thermostat [65], V-rescale thermostat [66], etc. The NPT ensemble allows the change of the system volume during the simulation, while keeping the number of particles, pressure and temperature constant. The pressure is coupled by control barostats, such as the Parinello-Rahman barostat [67].

4.3. Free Energy Calculations

Free energy, one of the most important quantities in thermodynamics, is crucial in the study of receptor-ligand interactions. Thanks to the rapidly developed computer science, molecular dynamics simulations allow us to accurately predict the binding free energy, and hence to better understand the function and structure of proteins. The free energy can be expressed by the Helmholtz function, A, as well as the Gibbs function, G. The former one is suitable for the NVT system, while the latter one is appropriate for the NPT system. Since most laboratory experiments are carried out at constant temperature and pressure, it is more reasonable to compute the Gibbs free energy in MD simulations.

As introduced in section 4.2, in a MD simulation, the states of the system at each time step are calculated and can be recorded, resulting in a collection of states of the system relevant to the moment of the simulation, which is called as trajectory. Consider a simulation in which the molecule has "sufficiently" often visited two states A and B. The free energy difference between A and B can be obtained by counting the number of both states, $Q(A)$ and $Q(B)$, as [24],

$$\Delta G = G_B - G_A = -RT \ln \frac{Q(B)}{Q(A)} \tag{4.25}$$

where R is the gas constant, T is the absolute temperature.

However, this "sufficient" is almost impossible to achieve in the standard MD simulation. Despite the rapid advances in computer science, standard MD methods are currently still limited by timescales. For instance, for biomacromolecules, only simulations within a few microseconds are affordable. Since biomolecules usually involve large molecular weights, their interactions with ligands may happen on the order of milliseconds or even seconds. To obtain the full course of protein-ligand interactions on the atomic scale is computationally intensive.

To further extend the application of MD simulations and to solve the timescale problems, a large number of enhanced sampling methods have been proposed.

The common enhanced sampling methods fall into the following two main categories: those based on ensemble variables, such as umbrella sampling [68] and metadynamics [69], which sample and simulate with collective variables (CVs); and those that do not depend on reaction coordinates, such as replica-exchange molecular dynamics method [70]. In this work, metadynamics is employed and will be briefly introduced in the next section.

4.3.1. Metadynamics

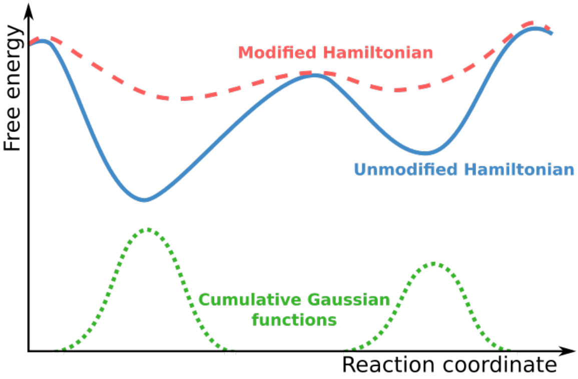


Figure 4.2.: Schematic view of of metadynamics with a single CV. The free energy landscape (blue solid curve) is calculated by the sum of deposited Gaussian kernels (green dot curve).

In metadynamics (MetaD), an external history-dependent bias potential is added along with the CVs ($\vec{s}(q)$) to fill the free energy surface (FES) of the system, thereby generally increasing the system Hamiltonian to achieve enhanced sampling. As shown in Fig. 4.2, when the local minima of the FES is filled by the bias potential, the system will cross the lowest free energy saddle point nearby and escape from the free energy “valleys”. This allows metadynamics not only to efficiently calculate free energies, but also to explore new reaction paths and accelerate rare events. During the simulation, the external bias potential is achieved by adding Gaussian shape kernels step by step, and the free energy can be obtained from the negative bias potential $V(\vec{s}, t)$ as,

$$V(\vec{s}, t) = \sum_{k\tau<t} W(k\tau) \exp\left(-\sum_{i=1}^{d} \frac{(s_i - s_i(q(k\tau)))^2}{2\sigma_i^2}\right) \tag{4.26}$$

Here, $\vec{s}$ is the CV, t is the simulation time, $W(k\tau)$ denotes the height of Gaussian kernels with the time of last bias deposition $(k\tau)$, s_i indicates the ith CV, and σ_i gives the width of Gaussian kernels. As convergence is reached after a long simulation time, the free energy along the CV space can be written as,

$$F(\vec{s}) = -V(\vec{s}, t \to \infty) + C \tag{4.27}$$

Since the height of Gaussian kernel is constant, the standard metadynamics faces the problem that a large Gaussian height, i.e. fast Gaussian deposition rate, will yield a large average error, which is proportional to the square root of the bias potential deposition rate, while a small Gaussian height will make the simulation take more time to fill the FES (Fig. 4.3). To solve this problem, the well-tempered metadynamics (wt-MetaD) was developed by Parrinello et al [71].

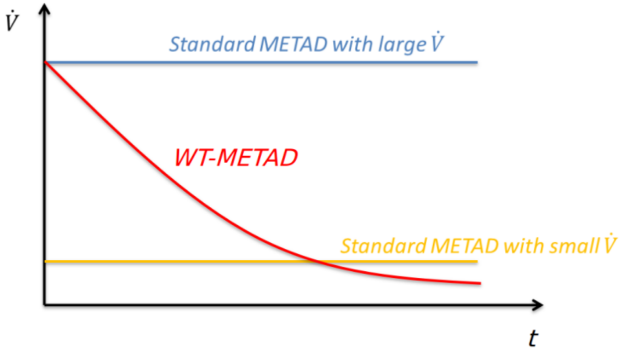


Figure 4.3.: Comparison of Gaussian bias deposition rate $\dot{V}$ between standard MetaD and wt-MetaD with time. The wt-MetaD deposits bias fast at the beginning and slow till the convergence, which can improve the accuracy while keeping the computation speed.

The well-tempered metadynamics employs the Gaussian kernel with variable height. The height of the Gaussian kernels decrease as the times of revisiting the same configuration increase as,

$$W(k\tau) = W_0 \exp\left(-\frac{V(\vec{s}(q(k\tau)), k\tau)}{k_B \Delta T}\right) \tag{4.28}$$

Here, W_0 is the initial Guassian height, ΔT is the energy related to the CV space exploration. The free energy along the CV space reads as,

$$F(\vec{s}) = -\frac{T + \Delta T}{\Delta T} V(\vec{s}, t \to \infty) + C \tag{4.29}$$

where T is the temperature of the system, $\frac{T+\Delta T}{\Delta T}$ indicates the bias factor γ, which needs to be carefully chosen based on relevant free-energy barriers to be crossed in the simulation. Note that, the $\Delta T = 0$ corresponds to standard MD simulation and $\Delta T \to \infty$ corresponds to standard metadynamics.

Unfortunately, wt-MetaD still fails to deal with the ligand binding process in most cases of biomacromolecular systems. In fact, if there are no restrains for the ligand coordinates, once it leaves the binding pocket, it will sample all the possible solvated states. The solvated states contain a vast part of the configuration space that cannot be sampled within a limited simulation time, and hence the ligand will barely come back leading to a failure of sufficiently sampling the important binding states.

To overcome these issues arising in the binding of protein ligands, the funnel metadynamics (FM) approach was further proposed based on the metadynamics [72, 73]. In the FM, as shown in Fig. 4.4, the ligand position is restrained by a funnel shaped restraint potential. As the ligand is close to the binding site, it will be fully sampled within the conical part restraint. When the ligand is at a unbound state, it is constrained within the cylinderic part with smaller space, resulting in enhanced sampling. There is no repulsive bias when the ligand explores configurations inside the funnel restraint, whereas a repulsive bias will be applied to the system if the ligand moves to the edge of the restraint, which discourages the ligand from exploring the area outside. With the introduced repulsive bias potential, the binding constant K_b can be expressed as,

$$K_\mathrm{b} = C^0 \pi R_\mathrm{cyl}^2 \int_\mathrm{site} dz e^{-\beta[W(z) - W(\mathrm{ref})]} \tag{4.30}$$

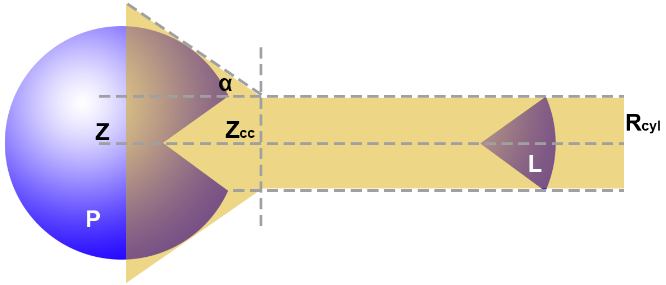


Figure 4.4.: Schematic view of protein-ligand binding process and the funnel-shaped restraint in funnel metadynamics.

where $C^0 = 1/1,660$ Å^{-3} is the standard concentration, R_{cyl} is the radius of the cylinder restraint, β is the reverse of Boltzmann constant k_{B} times system temperature T, $W(z)$ is the potential along the z-axis and $W(\mathrm{ref})$ is the potential in the unbound state derived from the potential of mean force (PMF).

4.3.2. Alchemical Calculation

In enhanced sampling approaches, the simulations are performed with a fixed charge model. Therefore, in the case of protein-ligand binding process, computing the free energy with a fixed charge model may lead to inaccurate results if the polarity of the binding pocket differs significantly from that of the solvent. To overcome this issue, alchemical free energy calculations need to be applied. In alchemical calculations, several non-physical intermediate states ($\lambda_i, i = 1, 2, 3\ldots$) are inserted between the known initial state A ($\lambda = 0$) and end state B ($\lambda = 1$). The chemical properties of parts of the system in these intermediate states are modified by the potential, leading to the change of interactions between the selected part and its environment [74]. The system Hamiltonian varies from H_A to H_B as the coupling parameter λ increases from 0 to 1, and each term of the force field for the intermediate state λ_i can be written as [24],

$$X(\lambda_i) = \lambda_i X(\mathrm{B}) + (1 - \lambda_i) X(\mathrm{A}) \tag{4.31}$$

Here, X represents k_{ij}^{b} and r_{ij}^{0} in Equation 4.2, k_{ijk}^{a} and θ_{ijl}^{0} in Equation 4.3, V_n in Equation 4.4, ϵ and σ in Equation 4.7 and q_i in Equation 4.9.

During the calculation, for each intermediate state λ_i, the system is firstly equilibrated with corresponding FF parameters from Equation 4.31. Afterwards, the free energy difference ΔG from intermediate state i to state $i+1$ is counted by

$$\begin{aligned} \Delta G(\lambda_i \longrightarrow \lambda_{i+1}) &= -k_{\mathrm{B}} T \ln \left\langle \exp\left(-\frac{\Delta H_i}{k_{\mathrm{B}} T}\right)\right\rangle \\ \Delta H_i &= H_{i+1} - H_i \end{aligned} \tag{4.32}$$

The total free energy between state A and state B is the sum of these free energy changes and such protocol is known as free energy perturbation (FEP) method.

4.4. Quantum Mechanics/Molecular Mechanics Simulation

In the previous section, both quantum mechanics and molecular mechanics have been briefly introduced and discussed. However, for biomacromolecular system, both methods are unlikely to reveal the chemical reactions because the QM approach can not afford for such large systems and the MM method does not consider changes in the electronic structure. Therefore, a combination method containing both QM and MM approaches has been proposed, which is known as Quantum Mechanics/Molecular Mechanics (QM/MM) method.

The combined QM/MM method is a method for calculating chemical reactions in the liquid phase. The reaction part of the system is calculated by quantum mechanical methods and the rest by force field-based simulation. The total energy of the system reads as,

$$E_{\mathrm{TOT}} = E_{\mathrm{QM}} + E_{\mathrm{MM}} + E_{\mathrm{QM/MM}} \tag{4.33}$$

where E_{QM} is the energy obtained by QM methods, E_{MM} is the energy from MM approach. $E_{\text{QM/MM}}$ is the interaction energy between the QM and MM part of the system. In the case of no chemical bonds between QM and MM parts, $E_{\text{QM/MM}}$ includes only non-bonding interactions between QM atoms and MM atoms. The Hamiltonian $H_{\text{QM/MM}}$ can be expressed as,

$$H_{\text{QM/MM}} = -\sum_{i}\sum_{M}\frac{q_M}{r_{i,M}} + \sum_{\alpha}\sum_{M}\frac{Z_\alpha q_M}{R_{i,M}} + \sum_{\alpha}\sum_{M}\left(\frac{A_{\alpha,M}}{R_{\alpha,M}^{12}} - \frac{C_{\alpha,M}}{R_{\alpha,M}^{6}}\right) \tag{4.34}$$

Here, i represents a QM electron and α is a QM nucleus. M refers to a MM nucleus and q indicates its net atomic charge. Hence, the interaction between QM and MM region contains electrostatic interactions between QM electrons and MM nuclei, electrostatic interactions between QM nuclei and MM nuclei and vdW interactions between QM atoms and MM atoms. The last two terms of Equation 4.34 do not involve the electron coordinates, and thus can be directly calculated. In the first term, QM calculations need to be included by adding one-electron integrals to the one-electron matrix as,

$$\int \phi_\mu(1)\frac{1}{r_{1,M}}\phi_v(1)dv(1) \tag{4.35}$$

In proteins, chemical bonds existing between the QM and MM regions and boundary bonds need to be carefully considered. Cutting the QM/MM bond results in the formation of unpaired electrons in the QM region, and the most common solution to the loss of electrons in the QM region due to truncation is to introduce an single bonded H atom in the QM region. This bonded H atom is called as cap atom and is only added to the QM calculation. Note that, it is important to choose a proper boundaries in the simulation, as a too large QM region will reduce the efficiency of the calculation, while a small QM region will lead to artificial risks being introduced into the calculation.

5. Machine Learning

Machine learning, in a broad sense, is making machines have the capacity to learn like humans, and then be more intelligent to complete tasks that are impossible to do with direct programming. In practical production terms, machine learning is about training based on big data, building models, and then using these models to predict new data. With the development of machine learning algorithms and improvements in hardware, machine learning has been integrated into quantum chemistry research in recent years [75–78]. Under certain conditions (e.g. good training data and machine learning algorithms), since there is no need to solve the Schrödinger equation, machine learning can predict the desired quantum chemical properties without loss of accuracy and at a computational speed comparable to that of molecular mechanics methods.

5.1. Supervised learning

Machine learning techniques can be divided into three types: supervised learning, unsupervised learning and reinforcement learning. In supervised learning, it can be classified into the classification process and regression process. The former is performed by labelling and training existing data samples, feature analysis to identify certain types of objects, and then classification. The main difference between classification and regression is the type of output data, usually called labels. If the type of label is qualitative, the process is called classification, which is also known as discrete variable prediction. While if the type of label is quantitative, the process is called regression, which is also known as continuous variable prediction.

Unsupervised learning refers to methods where the input is only sample data without labels and the network model actively explores the information contained in the data during the training process. Reinforcement learning is the process of feeding unlabelled data into a network and then continuously adjusting the direction of learning by feeding back the results of training to the network model, repeating the process until the network model converges. In general, supervised learning is the main focus in machine learning, with unsupervised learning as a supplement.

5.2. Overfitting and Underfitting

An important topic in machine learning is the generalisation ability of the model. The stronger the generalisation ability of the model, the better the model. If a trained model performs poorly in the training set, it will also perform poorly in the test set, which may be due to underfitting; if the model performs very well in the training set, but poorly in the test set, this is a result of overfitting.

Overfitting and underfitting can be explained in terms of *bias* and *variance*, with underfitting leading to high bias and overfitting leading to high variance, hence the need for the model to make a trade-off between bias and variance. Specifically, when a simple model is used to fit complex data, it can be difficult to fit the model to the true distribution of the data, and underfitting occurs. In this case, there is a large *bias*, which indicates the difference between the expected output of the model and its true output. In some cases, the model is over-training to get a more accurate model, or the noise in the training data is fitted as the model is too complex, resulting in a model that works very well on the training set but performs poorly in test set, in which case overfitting happens. At this point, there is a large *variance*, which portrays the difference between the output of the models obtained from different training sets and the expected output of these models.

In fact, as the model training proceeds, the complexity of the model increases and the training error of the model on the training data set gradually decreases. However, at a certain level of model complexity, the error of the model on the validation set increases as the complexity of the model increases. To prevent overfitting, several methods are used, such as Early Stopping [79], Data augmentation [80], Dropout [81], and so on. In this work, the Early Stopping is employed and will be briefly introduced in the following.

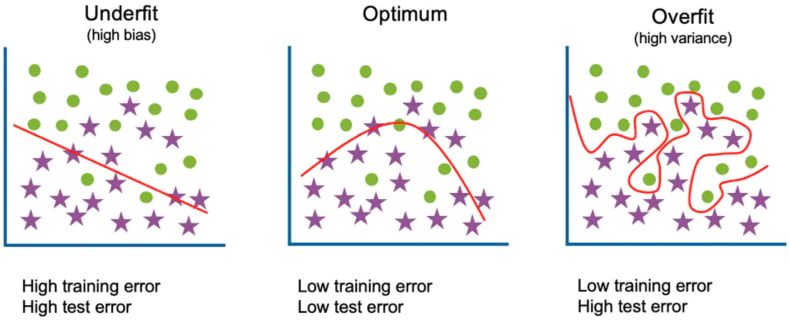


Figure 5.1.: Schematic view of overfitting, underfitting and proper fitting.

The process of training the model is the process of learning and updating the parameters of this model, and this parameter learning process often applies some iterative methods, such as the Gradient descent learning algorithm. Early stopping is a method of truncating the number of iterations to prevent overfitting, i.e. stopping the model before it converges on the training data set. The early stopping approach is to calculate the accuracy of the validation data at the end of each Epoch (an Epoch is a round of traversal of all the training data) and to stop training when the accuracy is no longer improving. Since accuracy does not decrease continuously in practice, it is not possible to judge that it is no longer improving based on one or two consecutive decreases. The general approach is to record the best validation accuracy to date during the training process, and to assume that accuracy is no longer improving when the best accuracy is not achieved for 10 consecutive Epochs (or more). At this point the iteration can be stopped (Early Stopping).

5.3. Neural Networks

Artificial neural networks (ANNs) are mathematical models that simulate the neuronal structure of a living creature's brain, using the structure of synaptic connections in biological neural networks to perform calculations on the input data [82]. In addition, ANNs can analyse and identify the acquired patterns through feature extraction and function fitting, and are used to solve problems such as classification or regression. ANNs work by extracting features from a large amount of data to analyse the patterns in the images and ultimately make decisions about new data based on the patterns acquired. This process of analysing and learning patterns is known as "training", after which the artificial neural network is capable of making autonomous judgements and decisions.

For a single neuron (called as *perceptron* in ANN), a set of training samples $(x^{(i)}, y^{(i)})$ provided, where x represents the input, y represents the corresponding label value and i represents the i-th sample. The neural network algorithm can be represented as a non-linear complex function $h_{W,b}(x)$ to fit a functional relationship between x and y, where W and b are the parameters of the functional model.

$$h_{W,b}(x) = f(W^T x) = f(\sum_{i=1}^{n} W_i x_i + b) \tag{5.1}$$

Here, f is called as activation function, which traditionally includes the Tanh or Sigmoid functions.

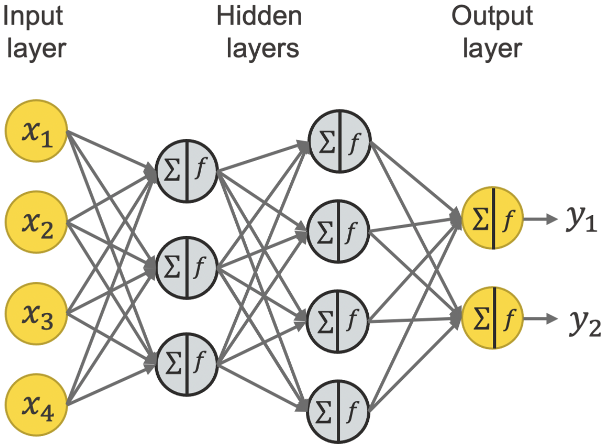


Figure 5.2.: Schematic view of ANN structure with two hidden layers.

For larger ANNs, as shown in Fig. 5.2, it contains one input layer, one output layer and one or more hidden layers. For two adjacent hidden layers, the outputs of neurons from the layer closer to the input layer give the inputs for the layer closer to the output layer, where the former hidden layer is also referred to as lower layer and the latter as upper layer. The layers are arranged in order of signal transmission, with neurons in layer i only receiving signals given by neurons in layer $(i - 1)$, without feedback between neurons.

Artificial neural networks reduce the error loss by forward pass and back propagation algorithms, and the process is repeated several times until the network converges. Specifically, when the signal is propagated forward, the input samples are passed in from the input layer, processed layer by layer by the hidden layer, and then passed to the output layer. If the actual output of the output layer does not match the desired output, it will move to the back-propagation stage of the error. Here the output error is back propagated in some form through the hidden layer to the input layer, and the error is apportioned to all neurons in each layer, so that the error signal of each neuron is obtained, and this error signal is taken to adjust the weights of each neuron. This process is carried out circumferentially. The constant adjustment of the weights is the learning and training process of the network.

Part III.

Results and Discussion

6. Unravelling the mechanism of Host-Guest Chemistry

Supramolecular chemistry is the chemistry of molecular aggregates based on non-covalent interactions between molecules. It focuses on the weak interactions between molecules with non-covalent bonds such as hydrogen bonds, ligand bonds, hydrophilic/hydrophobic interactions and the assembly, structure and function of molecular aggregates resulting from their synergistic interactions. The understanding of supramolecules dates back to the middle of the 20th century, in particular to the synthesis of macrocyclic molecules (crown ethers, cavity ligands, etc.) by Pedersen and his coworkers, which were able to selectively bind specific ions and small organic molecules based on non-covalent bonding interactions as well as the ring size [83]. This innovation in host and guest chemistry was awarded the Nobel Prize in Chemistry in 1987 [84]. Later, Lehn et al. designed three-dimensional congeners of crown-ethers, the macrobicyclic cryptands, which have larger association constants corresponding to a higher selectivity to the alkaline metals [85]. Based on these findings, in the past few decades, the rapid development of larger host molecules or supramolecular capsules has been witnessed, which can accommodate larger guests or molecular cations and hence reveal biochemical recognition events [86].

6.1. Truncated Tetrahedral [4+4] Imine Cages with Ammonium Ions

In parts reprinted with permission from Lauer, Jochen C. and Pang, Ziwei and Janßen, Paul and Rominger, Frank and Kirschbaum, Tobias and Elstner, Marcus and Mastalerz, Michael.
Host-Guest Chemistry of Truncated Tetrahedral Imine Cages with Ammonium Ions. ChemistryOpen, 9(2), pp.183-190.
DOI: https://doi.org/10.1002/open.201900357

6.1.1. Introduction

In this work, Lauer et al. experimentally investigated the host-guest binding of ammonium ions, NEt_4^+ and NMe_4^+, by shape-persistent [4+4] imine cages (Fig. 6.1) with a truncated tetrahedral geometry [87], and we subsequently revealed the mechanism of ammonium uptake through standard molecular dynamics simulations as well as metadynamics. The structure of the three [4+4] imine cages share the same main frames but have different window sizes due to various long substituents on the 1,3,5-triformylbenzene [87].

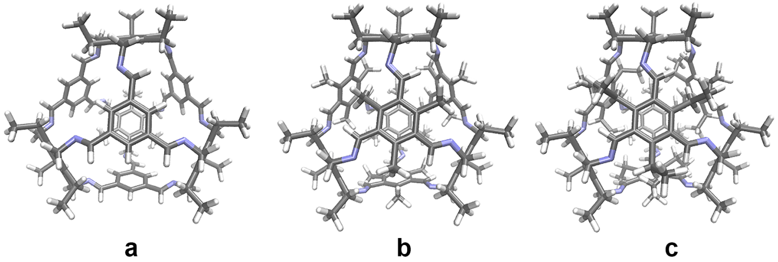


Figure 6.1.: Schematic view of **a).** 3-H, **b).** 3-Me and **c).** 3-Et imine cages and from standard MD simulations. The positions 2, 4 and 6 on 1,3,5-triformylbenzene are substituted with methyl and ethyl for 3-Me and 3-Et, respectively. The carbon atoms are coloured grey, nitrogen atoms ice blue and hydrogen atoms white.

As shown in Fig. 6.2, since there is no substituent on the 1,3,5-triformylbenzene of 3-H cage, the 3-H [4+4] imine cage has the largest window size of 49.7 Å^2 and cavity volume size of 337 Å^3. With methyl substituents, the window size of 3-Me cages significantly reduces to 28.4 Å^2 and volume to 253 Å^3. The 3-Et cage has the smallest window size of 24.1 Å^2 and volume of 218 Å^3. The volume of ammonium ion NEt_4^+ is 163 Å^3 and NMe_4^+ is 95 Å^3 [88]. To investigate ammonium ions uptake process, the activation energy of guest dissociation were firstly estimated based on nuclear magnetic resonance (NMR) results. However, underestimated results were produced according to the reaction time during experiments. For instance, for 3-H cages, it took a day for the guests to leave the cage completely, indicating a energy barrier over 80 kJ/mol based on the Arrhenius Equation, whereas the NMR results gave only 34 kJ/mol. Furthermore, the process of the guest leaving from the cage cavity could not be observed from laboratory experiments. Hence, in this work, we performed a set of MD simulations as well as metadynamics calculations to reveal the full extent of the guest's uptake process.

6.1.2. Computational details

6.1.2.1. Initial Structural Model

Models of all cage structures, as well as cationic ligands NEt_4^+ and NMe_4^+, were constructed with xLeap tool from the AmberTools package [89] employing the general Amber force field (GAFF) [47, 90]. The atomic charge of models was calculated by the Hartree-Fock method using GAUSSIAN09 with 6-31G* basis set [91]. The force field parameters of the solvent molecule dichloromethane and the Cl^- counter ions were also selected from the GAFF. Different cage structures with corresponding cationic ligands were separately settled in a 5×5×5 Å^3 simulation box with periodic boundary conditions. Afterwards, dichloromethane molecules were inserted into the box, and one of those solvent molecules were subsequently replaced by the Cl^- to achieve electroneutrality. A total of 6 systems were set up and each of them comprised ca. 4,000 atoms, which contained one cage molecule, one cationic ligand molecule, ca. 750 dichloromethane molecules and one counter ions.

6.1.2.2. All-atom MD Simulations

All the systems first underwent steepest descents energy minimisation and 2 ns NVT simulation at 298 K in sequence. Afterwards, a 2 ns NPT simulation was performed at 298 K and 1 bar with the Berendsen barostat [64] and was

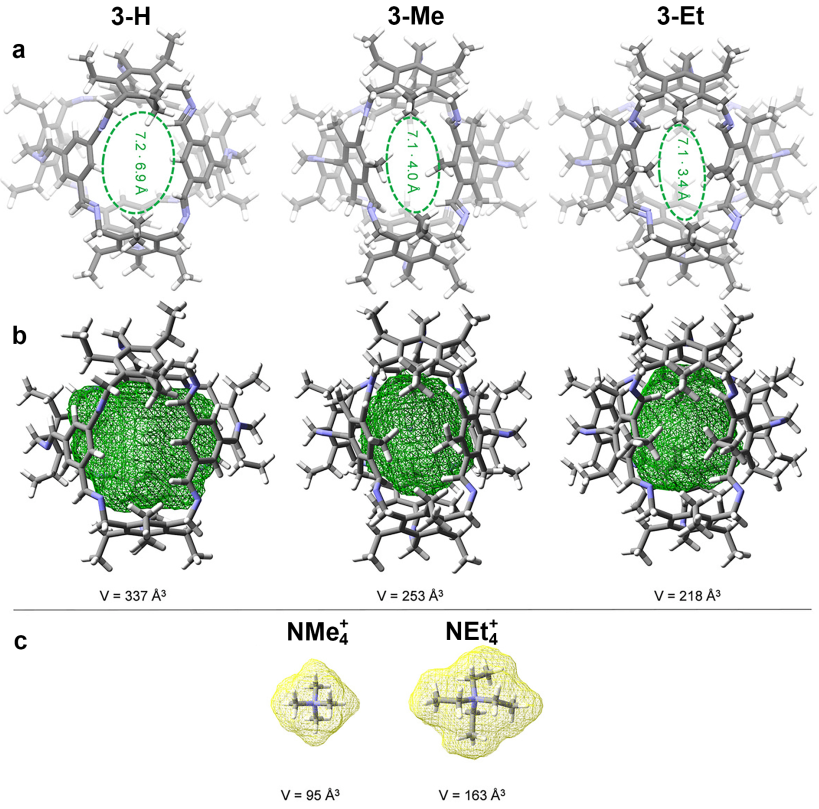


Figure 6.2.: a). Window size of [4+4] imine cages. The size was estimated by horizontal distance between the atom centres of two closest carbon atoms multiplied by vertical distance between the atom centres of two closest carbon atoms. **b).** Cavity volume of imine cages. **c).** Volume of NEt_4^+ and NMe_4^+. The volume was calculated via SwissPDBViewer by Dr. Jochen Lauer.

followed by 2 ns further NPT simulation with Parrinello-Rahman barostat [67] under the same thermodynamic conditions. Then, to further stabilise the system and to collect optimised structure, we performed 100 ns MD simulations for each system. In order to increase the accuracy, two groups (one comprised the cages molecule, the other one contained the ligand, solvent and counter ions) were set for temperature coupling in all simulations.

6.1.2.3. Free Energy Simulations

The free energies of the centre-of-mass distance between the cage and its corresponding cationic ligands were calculated with classical well-tempered metadynamics simulations [71]. The initial structure was taken from the 100 ns MD simulations for each system. The Gaussian bias potential was initially set with height = 0.5 kJ/mol, width $\sigma = 0.2$ Å, and was deposited each picosecond. Since the higher energy barriers to be crossed for the cage with the NEt_4^+, for those who have NEt_4^+, the bias factor was 100, and for systems with NMe_4^+, the bias factor was 50. To enhance the sampling efficiency, based on the estimation of the cage radius (r $\approx$ 4.5 Å), the ligand was restricted to move within 7 Å from the mass centre of the cage. According to the experiments, we never saw the counter ions BF_4^- enter into the cage, the counter ions Cl^- was restrained from moving 5 Å away from the cage centre. The initial systems were set up with GROMACS in version 4.6.7, and the following all-atom simulations were performed with GROMACS in version 2018.3. [92–94] The well-tempered metadynamics simulations were performed and analysed via GROMACS 2018.3 interfaced with the Plumed 2.5.1 package [95]. Molecular structures were visualised with VMD 1.9.2 [96].

6.1.3. Results and Discussion

There are two different mechanisms for the uptake process of the ammonium ions [97]. One possibility is a door-opening mechanism, whereby one or more imine bonds undergo reversible bond cleavage to “open the lid” of the cage, allowing the guest ion to be encapsulated without or with low energy barriers, followed by the reformation of the imine bonds to close the cage. This mechanism has been proven for an imine based hemicarcerand by Ro et al. [98]. The second possibility is a squeezing mechanism where the cage remains intact, and the guest is squeezed through the window into the cavity without bond breaking and reforming. It is proposed to be one of the most likely mechanisms for absorbing guest ions by tetrahedral metalcatecholate cages. Note that, under the squeezing mechanism, even excessively large guest ions, such as $CoCp_2^{*+}$, which need to cross the energy barrier of 251 kJ/mol, seem to be able to enter the host cavity [99]. As the kinetic uptake of the same ammonium ions was found to vary considerably in the experiments depending on the size of the cage window, it can be assumed that the squeezing mechanism is more likely than the door-opening mechanism.

Activation energies $\Delta G^{\ddagger}$ from NMR estimations as well as MetaD simulations are illustrated in Table 6.1. The MetaD results show that the larger the guest volume, the higher the activation energy for the same imine cage. In contrast, this trend could not be demonstrated in the NMR results for two reasons: firstly, no binding was found for both $NEt_4^+ \subset$3-Et and $NMe_4^+ \subset$3-Et; second, the larger activation energy for smaller guest was presented in the 3-Me cage, which is contrary to the pattern we found in the MetaD simulations. Besides, as discussed in section 6.1.1, underestimated activation energies were obtained via NMR estimation. Hence, it seems like activation energies calculated from NMR is not suitable in our case.

Table 6.1.: Activation energy $\Delta G^{\ddagger}$ (kJ/mol) from NMR and MetaD

Host	Guest	NMR*	MetaD
3-H	NMe_4^+	12	61
3-H	NEt_4^+	34	141
3-Me	NMe_4^+	50	123
3-Me	NEt_4^+	42	357
3-Et	NMe_4^+	–	91
3-Et	NEt_4^+	–	359

*NMR resuls are from Jochen Lauer

In addition to the relationship identified between activation energies and guest volume, compared to the 3-H cage, higher activation energies were found for both 3-Et and 3-Me cages, which have smaller windows. In other words, the energy barriers significantly increase as the window size decreases. For NMe_4^+ ammonium ions, the activation energy doubles when changing from 3-H cages with an aperture of 49.7 Å^2 to 3-Me or 3-Et cages with a window size of 28.4 Å^2 and 24.1 Å^2, respectively. Similarly, for NEt_4^+ guest, the activation energy nearly triples when changing from 3-H cages to 3-Et or 3-Me cages. Most interestingly, compared to the cage with medium size windows (3-Me), for NEt_4^+ guests, the activation energy of the cage with the smallest window size (3-Et) drops from $\Delta G^{\ddagger}$ =123 kJ/mol to $\Delta G^{\ddagger}$ =91 kJ/mol.

The free energy surfaces for all six complexes are presented in Fig. 6.3a to Fig. 6.8a, respectively. It is noticeable that when the cage has a smaller cavity volume ($NMe_4^+ \subset$3-Et in Fig. 6.8a) or a larger guest present ($NEt_4^+ \subset$3-Me in Fig. 6.4a) or both ($NEt_4^+ \subset$3-Et in Fig. 6.5a), a local minimum can be accessed at the state A, meaning that the guest ion prefers to stay in the centre of the cage. Conversely, the guest ion tends to be at the edge of the cage, and a local minimum is obtained at state B.

In the thermodynamically most stable host-guest complexes, different numbers of solvent molecules appear as co-guests in the cavity, which may explain different favourite positions of guests in our three cages. In those complexes where guests prefer to be located in the centre of the cavity (state A), as shown in Fig. 6.4A, 6.5A and 6.8A, no solvent molecules appear in the cage cavity. A possible reason is the small size of the cage cavity or the large size of the guests. Furthermore, for these complexes, even if the guest has deviated from the centre of the cage, the excess space in the cavity is still not sufficient to allow full access to solvent molecules. In Fig. 6.5B and 6.8B, there is no solvent molecule in the cavity, while in Fig. 6.4B, only half of a solvent molecule is squeezed into the interior of the cage.

In those complexes where the guests favour to stay close to the cage wall, as shown in Fig. 6.3B, 6.6B and 6.7B, some solvent molecules were ultimately squeezed into the cage cavity, thus forcing the guest away from the centre. Note that substable states are observed in state A in these complexes, suggesting that the guest can still be stable in the middle of the cage under certain coincidental conditions. Spatially, the six windows of the cage are symmetric based on the centre of the cage. Thus, for a complex with a large cavity volume as well as a small guest volume ($NEt_4^+ \subset$3-H and $NMe_4^+ \subset$3-H), there may be a situation where all six windows of the cage are occupied by a single solvent molecule, resulting in the guest ion being sandwiched by six solvent molecules in the centre of the cage (Fig. 6.3A and 6.6A). Interestingly, for the complex $NMe_4^+ \subset$3-Me, we found that at state A, where solvent molecules occupy only one symmetric pair of cage windows, the guest ion also remains stable at the centre (Fig. 6.7A). Nevertheless, this coincidental equilibrium can not be maintained for long in the system, and once the equilibrium is broken, the solvent molecules occupying the windows will squeeze into the excess space of the cage, thus reaching state B.

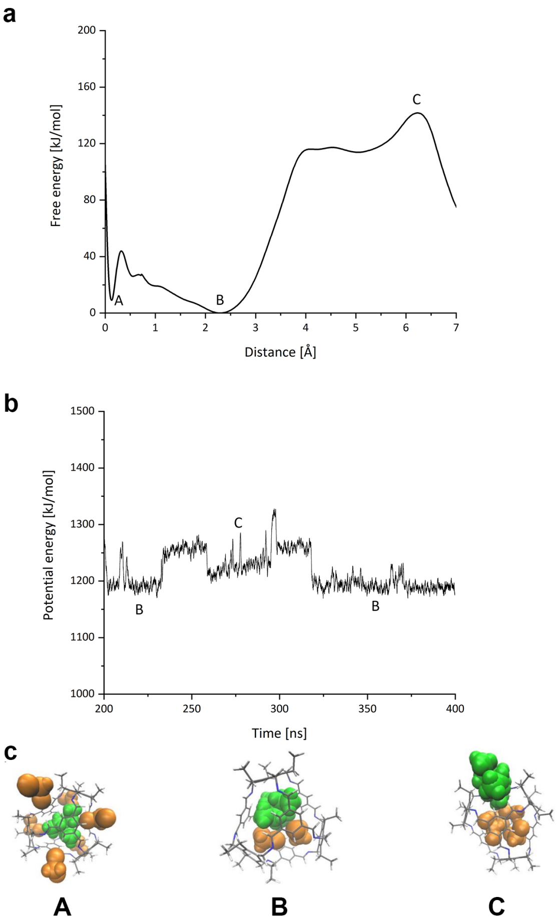


Figure 6.3.: Computed decapsulation process for 3-H cage with NEt_4^+: **a).** Free energy surface with activation energy $\Delta G^{\ddagger}$= 141 kJ/mol. **b).** Host potential energy changes when guest leaves and comes back once. **c).** Different conformations at state A, B and C. Due to solvation effects, the guest is not favoured at the centre of the cage (A), the most stable position is at the cage walls (B).

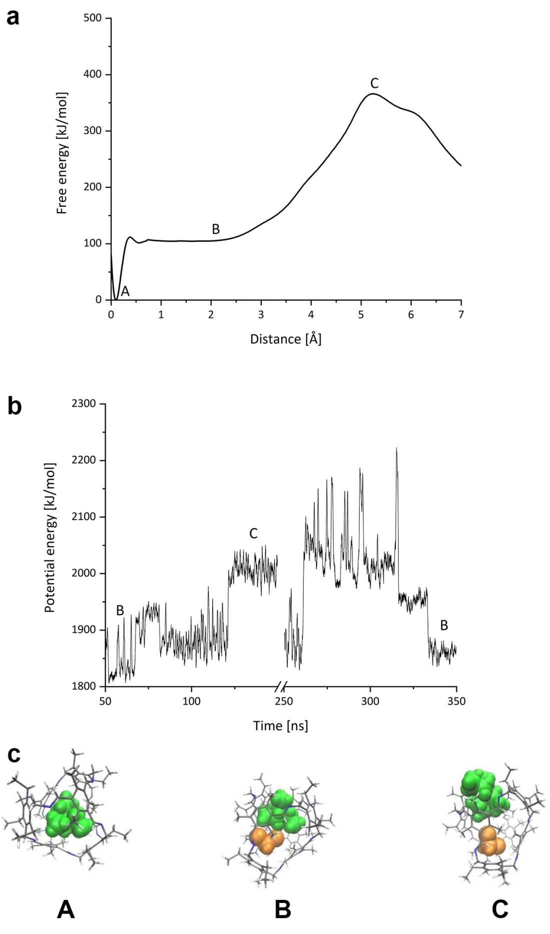


Figure 6.4.: Computed decapsulation process for 3-Me cage with NEt_4^+: **a).** Free energy surface with activation energy $\Delta G^{\ddagger}$= 357 kJ/mol. **b).** Host potential energy changes when guest leaves and comes back once. **c).** Different conformations at state A, B and C.

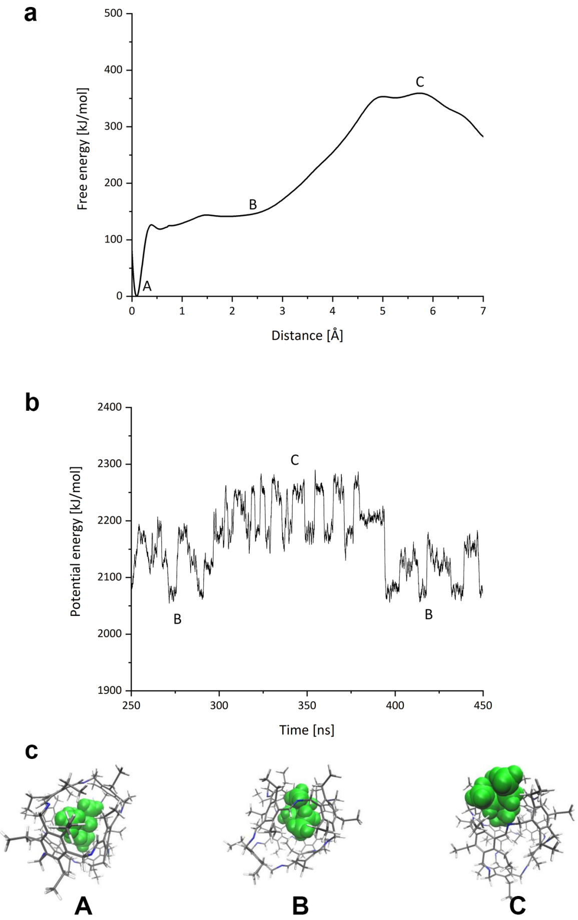


Figure 6.5.: Computed decapsulation process for 3-Et cage with NEt_4^+: **a).** Free energy surface with activation energy $\Delta G^{\ddagger}$= 359 kJ/mol. **b).** Host potential energy changes when guest leaves and comes back once. **c).** Different conformations at state A, B and C.

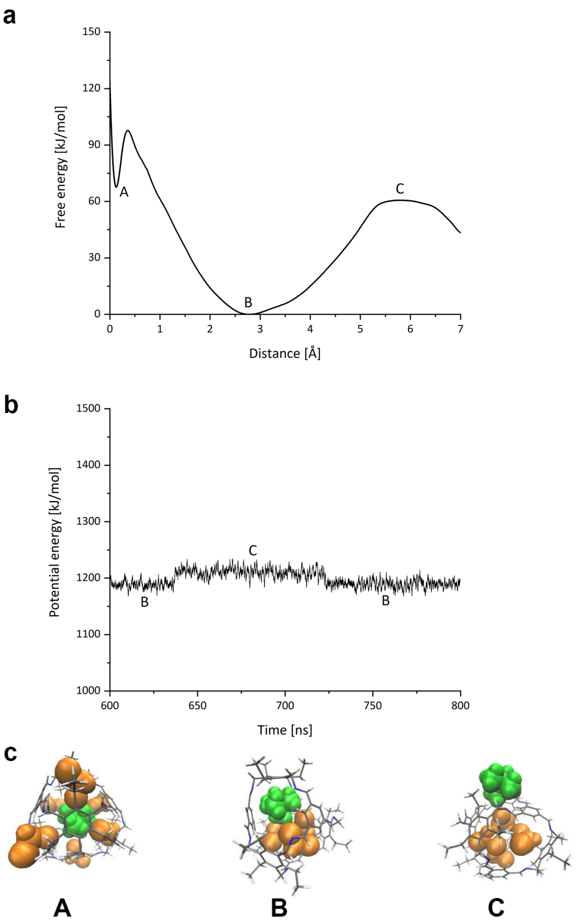


Figure 6.6.: Computed decapsulation process for 3-H cage with $NMe_4{}^+$: **a).** Free energy surface with activation energy $\Delta G^{\ddagger}$= 61 kJ/mol. **b).** Host potential energy changes when guest leaves and comes back once. **c).** Different conformations at state A, B and C. Due to solvation effects, the guest is not favoured at the centre of the cage (A), the most stable position is at the cage walls (B).

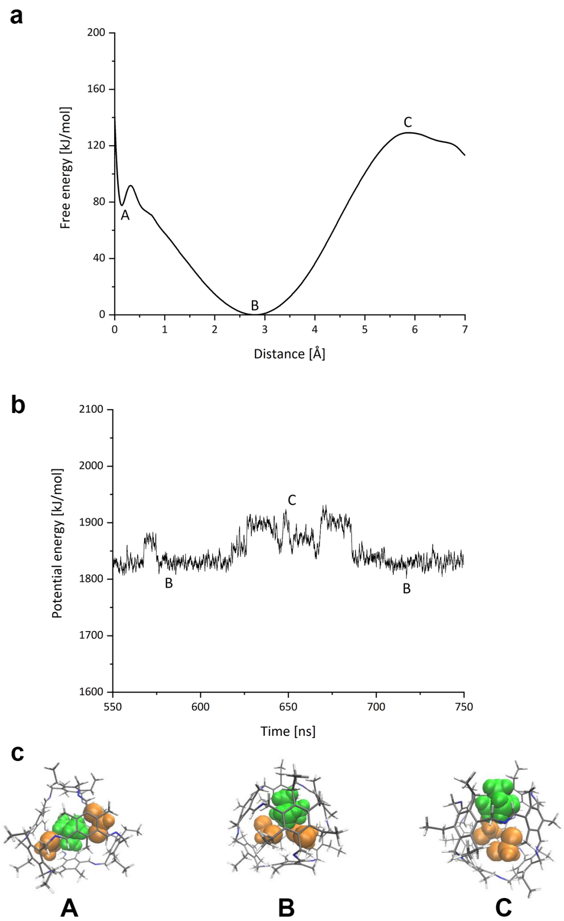


Figure 6.7.: Computed decapsulation process for 3-Me cage with NMe_4^+: **a).** Free energy surface with activation energy $\Delta G^{\ddagger}$= 123 kJ/mol. **b).** Host potential energy changes when guest leaves and comes back once. **c).** Different conformations at state A, B and C. Due to solvation effects, the guest is not favoured at the centre of the cage (A), the most stable position is at the cage walls (B).

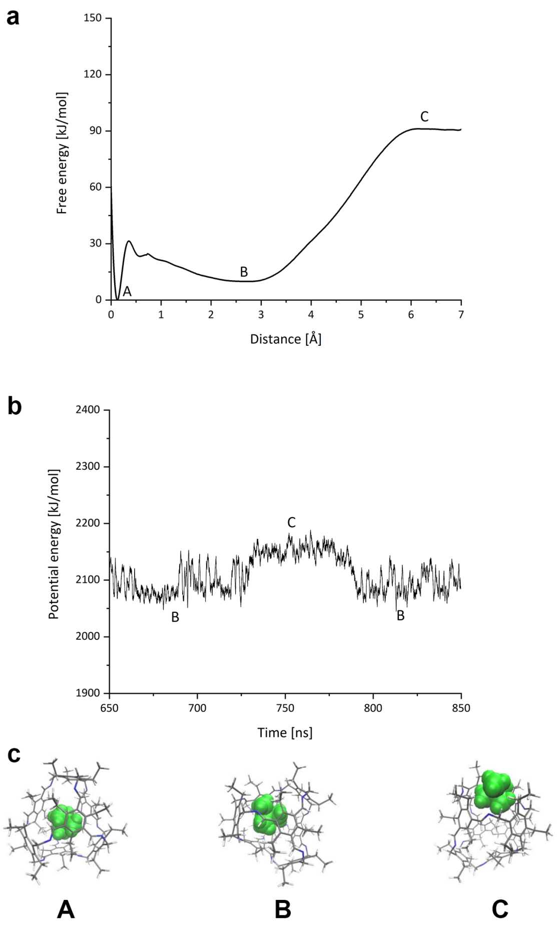


Figure 6.8.: Computed decapsulation process for 3-Et cage with NMe_4^+: **a).** Free energy surface with activation energy $\Delta G^{\ddagger}$= 91 kJ/mol. **b).** Host potential energy changes when guest leaves and comes back once. **c).** Different conformations at state A, B and C.

The kinetics of ammonium complexation is undoubtedly impacted by these co-complexed solvent molecules. For those complexes with large cage cavity, like $NEt_4^+ \subset$3-H and $NMe_4^+ \subset$3-H, as shown in Table 6.2, more solvent molecules enter the cage when the guest is located at the aperture and is leaving the cage (state C). However, for the $NEt_4^+ \subset$3-Me and $NMe_4^+ \subset$3-Me complexes, when the guest is on the way to leave the cage, no additional solvent molecules enter the cage due to the smaller cavity volume. It is worth noting that when the guest is at state C and multiple solvent molecules have entered the cage, these solvent molecules accumulate symmetrically inside the cage, centred on the central position of the cage, as shown in Fig.6.3C, 6.6C and 6.7C. If there is only one solvent molecule inside the cage (complex $NEt_4^+ \subset$3-Me as shown in Fig. 6.4C), it will stand directly in the centre of the cage.

Table 6.2.: Solvation effects of NMe_4^+ and NEt_4^+ in different cage cavities.

Host	Guest	Number of solvent molecules		
		A	B	C
3-H	NMe_4^+	6*	3	4
3-H	NEt_4^+	6*	2	3
3-Me	NMe_4^+	2*	2	2
3-Me	NEt_4^+	0	1*	1
3-Et	NMe_4^+	0	0	0
3-Et	NEt_4^+	0	0	0

*Solvent molecules are half inside the cavity.

Based on the above findings, we propose the following speculations on the mechanism of guest de-encapsulation. Since all three cages are approximately spatially symmetrical structures, they can be regarded as a non-polar sphere. When only the guest ion is present in the cavity, since both NEt_4^+ and NMe_4^+ are polarised molecules, homogeneous repulsions interact between the host and the guest due to the solvation effect whereby the guest is located in the centre of the host. However, when solvent molecules are squeezed into the host, the repulsive effects between the solvent molecules and the non-polar cage sphere also force the solvent molecules to prefer the central position of the cavity due to the strong polarity of the CH_2Cl_2 solvent molecule. Note that the dipole moment of CH_2Cl_2 is 1.67 D, which is larger than the NEt_4^+ of 1.27 D and the NMe_4^+ of 0.37 D. Once the repulsive effects between the cage and solvent molecules are strong enough, it will break the original equilibrium between the guest and the host so that the guest is squeezed out from the centre and tends to occupy a position close to the cage's wall. As soon as the solvent molecules occupy entirely the centre of the cage or more solvent molecules enter the cavity, the guest ion will be pushed out of the host. However, such "solvent competition" mechanism can not act on complexes with small windows, such as 3-H cages, because solvent molecules are too large to enter the cavity despite the guest being about to leave the cage as shown in Fig. 6.5C and 6.8C. In fact, whilst the guest ion is in the host, it is not only affected by the cage but also the polar solvent outside the cage. Therefore, for 3-H cages, the solvation effects between inside guest ions and outside solvent molecules may play the leading role in the guest de-encapsulation mechanism.

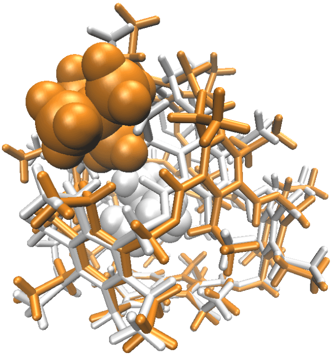


Figure 6.9.: Schematic view of the mine cage conformational change during the de-encapsulation of ammonium ion (example for 3-Et with NEt_4^+). White: Guest in the middle of the cage; Orange: Guest is leaving the cage.

Finally, for the guest uptake process in all six complexes, the deformations of cages were found as shown in Fig. 6.9, suggesting that such deformations also have effects on the de-encapsulation of ammonium ions. Potential energies changes as guest leaves and enters the host are illustrated in Fig. 6.3b to Fig. 6.8b for all complexes, respectively. In general, the smaller the aperture of the cage and the larger the volume of the guest, the greater the change in potential energy and the more significant the deformation of the host. For the $NMe_4^+ \subset$3-H complex (Fig. 6.6b), which has the largest window size and the smallest guest ions, no significant change in potential energy is observed between state B and state C, indicating only slight cage deformation happens when the guest ion leaves the host and therefore has a weak effect on the guest uptake process. However, in cases of $NEt_4^+ \subset$3-Me and $NEt_4^+ \subset$3-Et complex (Fig. 6.4b and 6.5b), the change in potential energy is up to 300 kJ/mol for the former and 200 kJ/mol for the latter, demonstrating a significant change in the structure of the cage due to its smaller window size, which may explain why both have activation energies of around 350 kJ/mol, despite the presence of the "solvent competition" mechanism in the $NEt_4^+ \subset$3-Me complex.

To sum up, the uptake process of the object is influenced by three main aspects. First, the "solvent competition" mechanism, in which the more polar CH_2Cl_2 solvent molecules enter the cage cavity and occupy the position of the guest, thus contributing to the guest uptake process. This mechanism plays a leading role when the complex has a larger cage cavity with a smaller guest. Second, the solvation effects between the guest inside the host and solvent molecules outside, which also drives the guest uptake process and plays a dominant role when the solvent can not enter the cage, especially for those hosts with small cavity volumes as well as small window sizes. The third is the cage deformation mechanism, in which the cage partly structural changes when a large guest ion crosses a smaller cage window, preventing the guest from the uptake process. This mechanism acts significantly in complexes that possess smaller cage apertures as well as larger guests. Overall, the sum of all these energy contributions may explain the mechanism of ammonium ions' de-encapsulation in our [4+4] imine cages. However, further investigations are required to understand the extent to which these mechanisms contribute to the guest uptake process.

6.2. Nitrogen Transfer within [2+3] Imine Cages

In parts collaborated with Dr. Sven M. Elbert from the Group of Prof. Mastalerz Michael at Heidelberg University.

6.2.1. Introduction

In this section, Elbert et al. experimentally synthesised [2+3] imine cages, and we then computationally investigated the nitrogen transfer dynamics within self–assembled cage crystals by funnel MetaD approach. As shown in Fig. 6.10, the structure of the three [2+3] imine cages are mainly structured by two triptycenes and three 1,4-diphenylbenzenes. The only difference is substituents at position 2,5 of the middle benzene ring in the three 1,4-diphenylbenzenes. Furthermore, the length between the two external triptycenes bridgehead carbons is 1.85 nm, and the length between the two internal triptycenes bridgehead carbons is 1.35 nm.

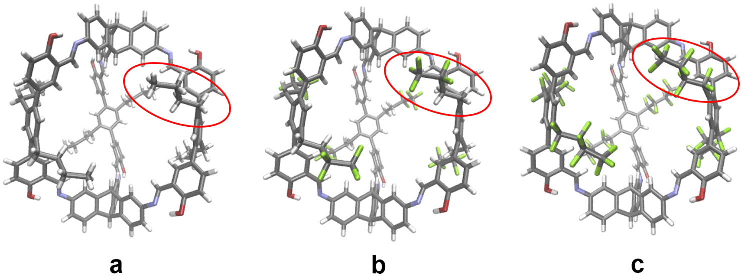


Figure 6.10.: Schematic view of **a).** F-cage, **b).** HF-cage and **c).** H-cage from standard MD simulations. The positions 2 and 5 on the middle benzene ring of 1,4-diphenylbenzene are substituted with butyl containing 9 fluorine atoms, butyl containing 5 fluorine atoms and butyl without fluorine atom, respectively (red circle marked). The carbon atoms are coloured grey, nitrogen atoms ice blue, hydrogen atoms white, oxygen atoms red and fluorine atoms green.

As shown in Fig. 6.11, the cages self-assemble by π–π stacking and form dense hexagonal crystals. As in hexagonal stacking, a void exists between any six adjacent cages in space. In this void, six butyl side chains from surrounding cages constructed a new “cage” (hereafter called a void cage). Therefore, for the nitrogen transfer between cages that are standing at diagonal positions,

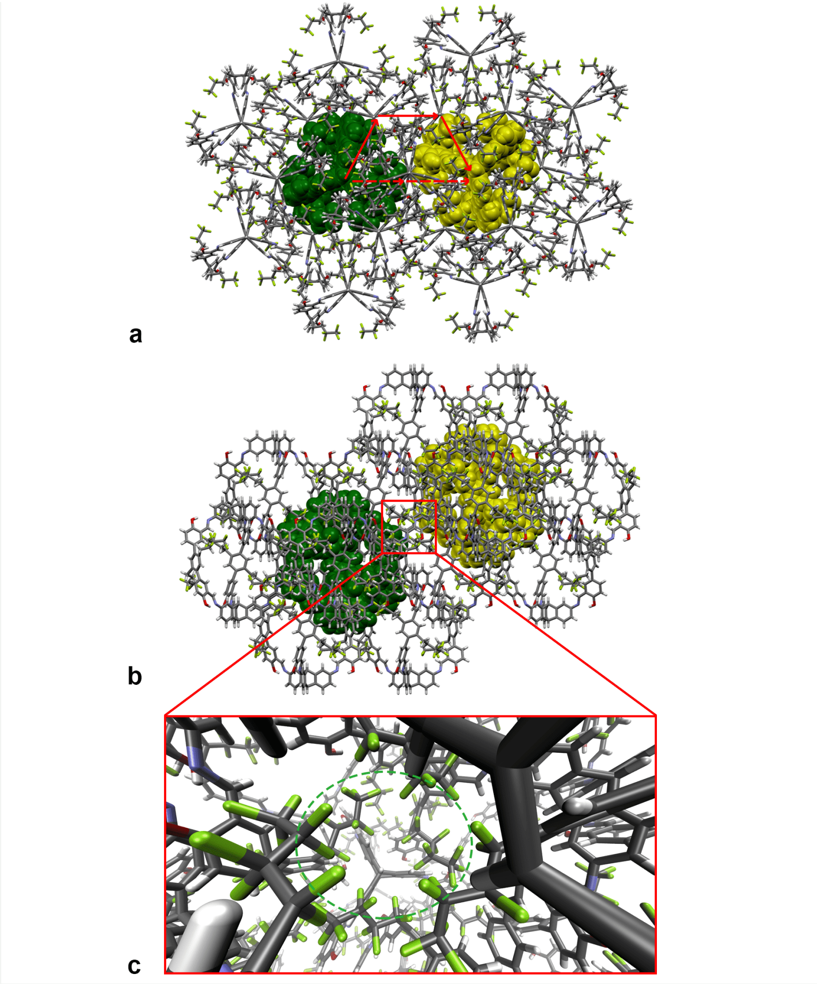


Figure 6.11.: Schematic view HF-cage crystal. **a).** View along the crystal c-axis. The solid arrows represent the “neighbour cage” three-steps pathway, the dashed arrows represent the “void cage” two-steps pathway. **b).** View along the crystal a-axis. **c).** The void cage constructed by six butyl side chains from surrounding HF-cages.

for instance, from the green cage to yellow cage in Fig. 6.11a, two possible pathways need to be considered. First, the guest will leave the green cage and enter directly into the adjacent cage (grey cage). Then, after passing two neighbouring cages, the guest will go into the yellow cage. In the other pathway, the guest will enter the void cage between green and yellow cages and subsequently move into the yellow cage. As laboratory experiments are not able to reveal which mechanism the nitrogen molecules prefer to transfer between cages, molecular dynamics simulations with an enhanced sampling approach were performed to explain the nitrogen transfer process.

6.2.2. Computational details

6.2.2.1. Initial set up

The atomic charge of cages was calculated by DFT/6-31G* with B3LYP functionals in GAUSSIAN09 [91]. Then, all cages were set up with the xLeap tool from the AmberTools package [89] employing the general Amber force field (GAFF) [47, 90]. Each crystal was settled with 22 cages in a 8×8×8 Å^3 simulation box with periodic boundary conditions. The host nitrogen molecule was placed in the middle of the cage that closed to the centre of the crystal. In total, three systems were established, and each of them contained ca. 5500 atoms. All the systems first underwent steepest descents energy minimisation and 1 ns NVT simulation at 77 K in sequence. Afterwards, a 10 ns NPT simulation was performed at 77 K and 1 bar with the Berendsen barostat [64], followed by 50 ns MD simulations to stabilise the system further. Since we only focus on the nitrogen transfer between cages, the positions of two triptycenes of each cage were frozen during all MD simulation so that all cages can fix their position with three 1,4-diphenylbenzenes as well as their substituents can be free to move in the crystal.

6.2.2.2. Funnel metadynamics

The free energy surfaces of two possible mechanisms were explored by funnel metadynamics [72, 73] started with the geometries after 50 ns free MD simulation from the last step. The CVs for each mechanism were taken as shown in Fig. 6.12, where the CV x indicated the guest position along with the centre line between the currently hosted cage (green) and the target cage (yellow), the CV y presented the deviation of the current nitrogen position relative to the centre line. The Gaussian bias potential was initially set with height = 1.2 kcal/mol, width σ = 0.2 Å, for both two CVs, and was deposited each picosecond. To ensure that the guests were sampled only in the two cages and the void in them, cylindrical potential restrictions with cross-sectional radii of 6 Å and 7 Å were imposed for the system focusing on the "neighbour cage" mechanism as well as the system focusing on the "void cage" mechanism, respectively. The length of the former restrain is 15 Å and the latter is 36 Å. The initial systems set up and the following all-atom simulations were performed with GROMACS in version 2018.3. [92, 94] The well-tempered metadynamics simulations were performed and analysed via GROMACS 2018.3 interfaced with the Plumed 2.5.1 package [95]. Molecular structures were visualised with VMD 1.9.2 [96].

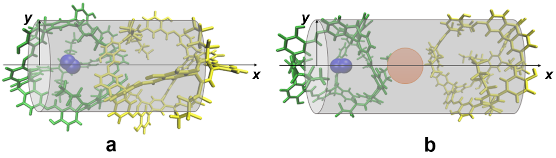


Figure 6.12.: Schematic view of collective variables used to describe the nitrogen position in cages. **a).** For the mechanism when the guest molecule enters into the neighbour cage. **b).** For the mechanism when the guest molecule goes inside the void cage (the orange region). The cylindrical shading represents the restrains from FM. The coordinates value x of nitrogen guest position increases from green cage to yellow cage in both cases.

6.2.3. Results and Discussion

As the crystal is self-assembled from individual cages by π-π interactions, the π-π stacking region exists between two neighbouring cages. When the guest molecules pass through this area for travelling between cages, the "neighbour cage" mechanism emerges. Free energy surfaces of "neighbour cage" mechanism for H-cage crystal, HF-cage crystal and F-cage crystal are shown in Fig. 6.14, 6.15 and 6.16, respectively. In all crystals, it is clear that the guest tends to be located far away from the axis between the two cage centres when passing through the cage junction, indicating that the nitrogen molecule prefers to bypass the π-π stacking region instead of passing directly through this region. This may be due to the nitrogen molecule having a long diameter around 4 Å and a short diameter of 3 Å, which is too large for the window formed by the π-π stacking as shown in Fig. 6.13.

Note that, the activation energy barrier for the H-cage crystals is 5.7 kcal/mol, approximately 1.5 kcal/mol lower than that of the HF-cage crystal and the F-cage crystal. One possible reason is that hydrogen atoms at the end of the butyl side chain of HF-/F-cage crystals are replaced by fluorine atoms, which reduces the tunnel space outside the π-π stacking region, thus increasing the activation energy of the guest transfer.

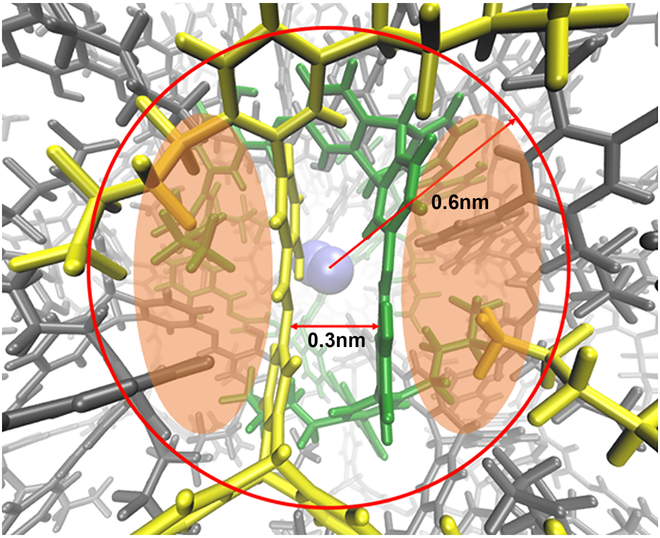


Figure 6.13.: Schematic view of the cross-section for the "neighbour cage" mechanism tunnel. The window produced by the π-π stacking are approximately 3 Å wide. The red circle represents the potential restrain with a radii of 6 Å in FM calculations. The two orange areas indicate the tunnels that guests prefer to cross and are partly blocked by F-/HF-cage side chains.

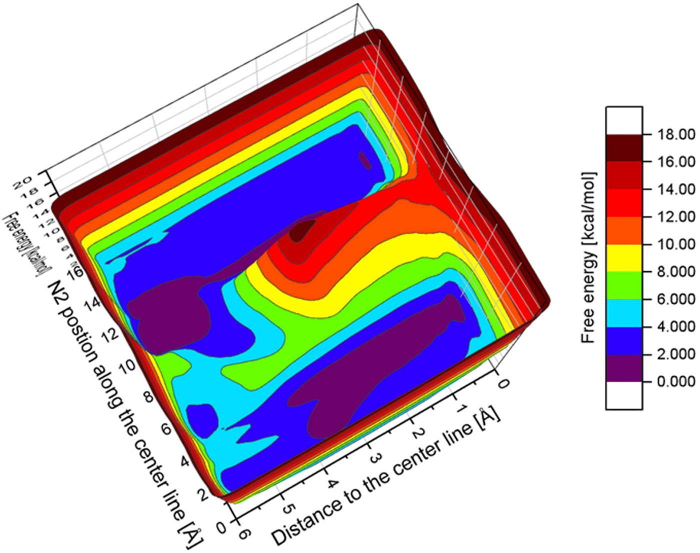


Figure 6.14.: Free energy surface of H-cage crystal with "neighbour cage" mechanism with the guest activation energy of 5.7 kcal/mol.

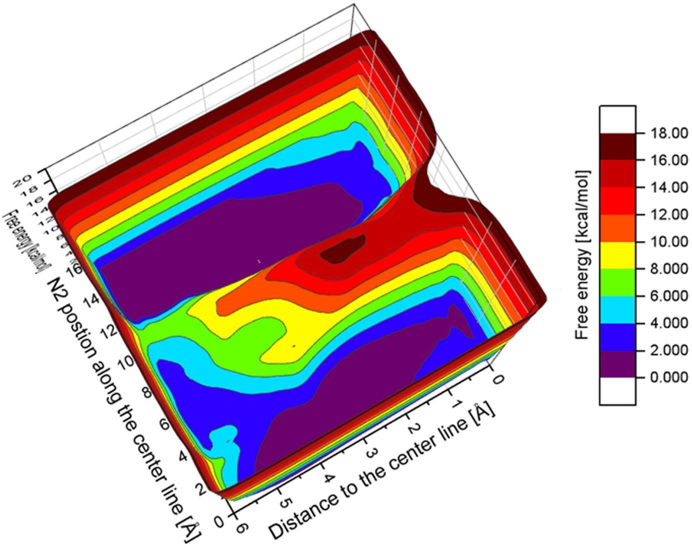


Figure 6.15.: Free energy surface of HF-cage crystal with “neighbour cage” mechanism with the guest activation energy of 7.2 kcal/mol.

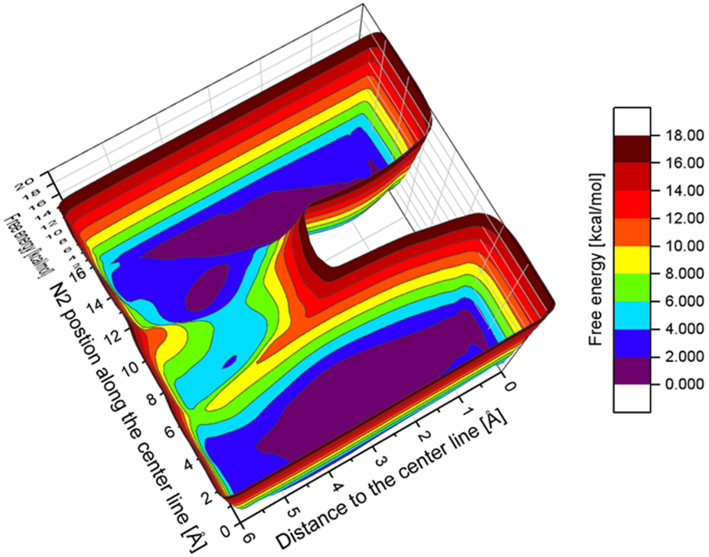


Figure 6.16.: Free energy surface of F-cage crystal with “neighbour cage” mechanism with the guest activation energy of 7.1 kcal/mol.

Fig. 6.17, 6.18 and 6.19 illustrate free energy surfaces of “void cage” mechanism for H-cage crystal, HF-cage crystal and F-cage crystal, respectively. Interestingly, with the “void cage” mechanism, we found different pathways for the guest transfer between diagonal cages in different crystals. Specifically, as shown in Fig. 6.17, for H-cage crystals, the guest molecule tends to move towards the neighbouring cage and then transfers to the diagonal cage, with an activation energy of 4 kcal/mol for both two steps, which means the transfer undergoes more likely with the “neighbour cage” mechanism. Nevertheless, we also found a pathway that proceeds under the “void cage” mechanism, where the guest molecule first crosses a 6 kcal/mol energy barrier into the void cage and then into the diagonal cage with activation energy around 2 kcal/mol. For the HF-cage crystals, however, under the “void cage” mechanism, the guest is unlikely to move towards the neighbouring cage nor the void cage to reach the diagonal cage (Fig. 6.18). In fact, the pathway shows that the nitrogen guest firstly arrives at the edge of the “void cage” and then enters the diagonal cage with an activation energy of 8 kcal/mol. In contrast, for the F-cage crystals with more fluorine atoms on their butyl side chains, the guest pathway indicates that the nitrogen molecule prefers to enter the “void cage” and then transfers to the diagonal cage.

Finally, under the “void cage” mechanism, the larger the butyl side chain, i.e. the more the hydrogen atoms are replaced by the fluorine atoms on side chains, the more likely the guest tends to enter the void cage and thus reach the diagonal cage. One possible reason is that as the side chain size increasing, the space outside the void cage is blocked, thus making it difficult for the guest to pass through.

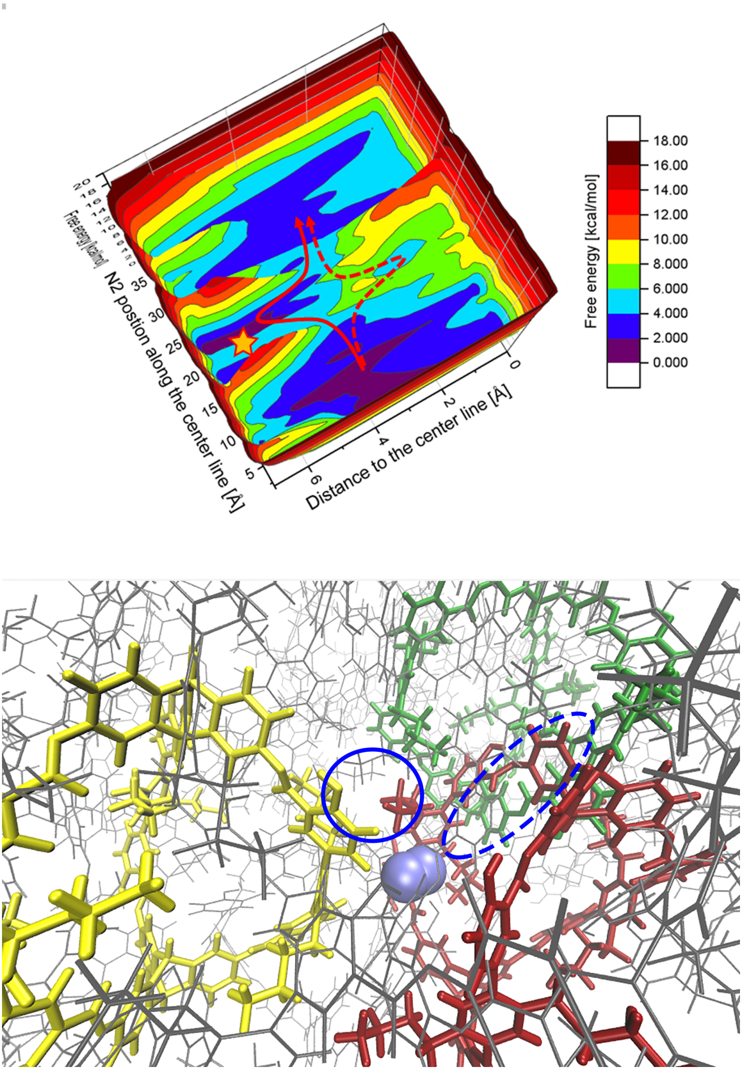


Figure 6.17.: Free energy surface of H-cage crystal with "void cage" mechanism. The red solid arrow represents the pathway close to the "neighbour cage" mechanism from green cage to yellow cage. The red dash arrow indicates the pathway of the guest entering the void cage and reaching the yellow cage from green to yellow. The schematic view shows the guest position with its coordinates labelled as yellow star in the free energy map. The solid blue circle is the void cage and the dash blue circle is the π-π stacking region between green and red cage.

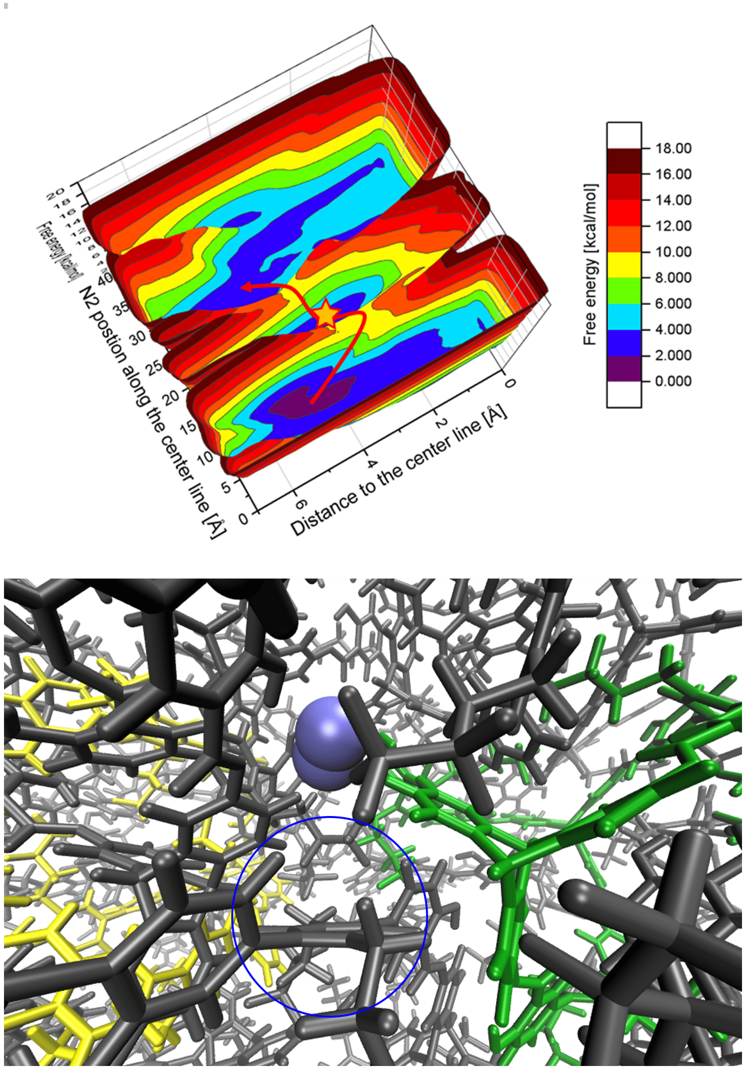


Figure 6.18.: Free energy surface of HF-cage crystal with “void cage” mechanism. The red solid arrow represents the pathway from the green cage to the yellow cage. The schematic view shows the guest position with its coordinates labelled as yellow star in the free energy map. The solid blue circle is the void cage.

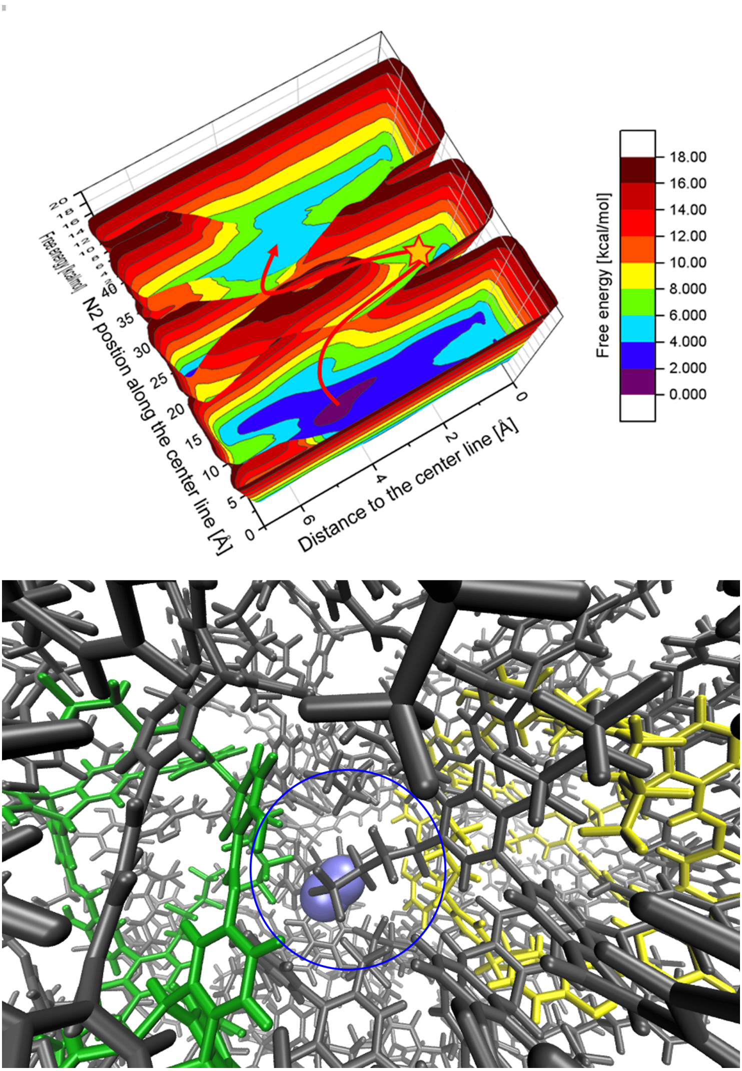


Figure 6.19.: Free energy surface of F-cage crystal with “void cage” mechanism. The red solid arrow represents the pathway from the green cage to the yellow cage. The schematic view shows the guest position with its coordinates labelled as yellow star in the free energy map. The solid blue circle is the void cage.

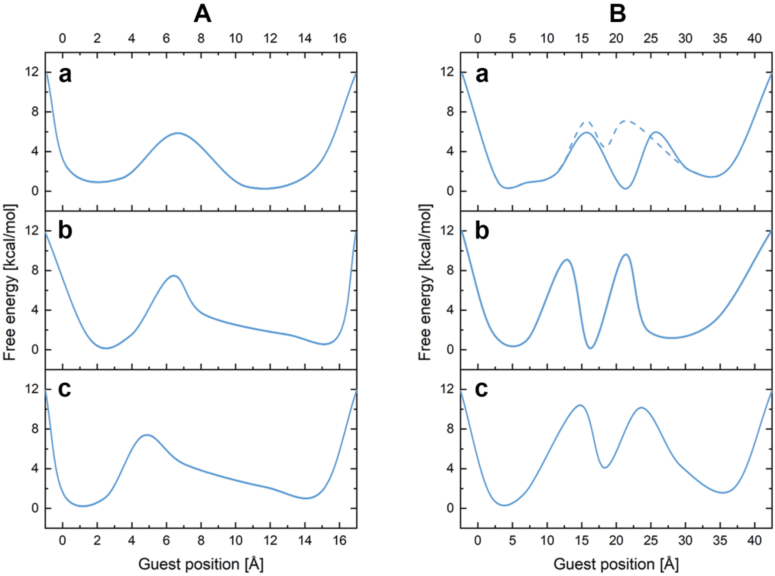


Figure 6.20.: Summary of free energy surfaces for H-cage crystal (**a**), HF-cage crystal (**b**) and F-cage crystal (**c**) with "neighbour cage" mechanism (**A**) and "void cage" mechanism (**B**). The dash line in **B**-**a** represents the pathway of the guest entering the void cage and reaching the yellow cage from green to yellow.

To summarise, as shown in Fig. 6.20, the activation energy required when the guest molecule transfers between cages with the “neighbour cage” mechanism is slightly lower compared to that with the “void cage” mechanism. Specifically, for H-cage crystals, with “neighbour cage” mechanism, an activation energy of 5.7 kcal/mol is required to enter the adjacent cage, whereas with the “void cage” mechanism, the nitrogen guest needs to firstly go over an energy barrier of 6.1 kcal/mol to reach an intermediate state and then over an barrier with the same height to reach the diagonal yellow cage. For HF-cage crystals, an energy barrier of 7.2 kcal/mol needs to be overcome under the “neighbour cage” mechanism, whereas with the “void cage” mechanism, two energy barriers of 9.1 kcal/mol and 9.5 kcal/mol are required to be crossed in steps to reach the diagonal cage. For F-cage crystals, with the former mechanism, the energy barrier is 7.1 kcal/mol, whereas with the latter mechanism, 10.5 kcal/mol is required to enter the void cage from the green cage, and 5.3 kcal/mol is needed from the void cage to the yellow cage.

Note that, the energy barriers given in Fig. 6.20A represent only the transfer of the nitrogen guest within neighbouring cages for the “neighbour cage” mechanism. In fact, when focusing on the transfer of guest molecule between two diagonal cages (from green cage to yellow cage in this study), with the “neighbour cage” mechanism, the guest molecule needs to enter the two neighbour cages before arriving at the diagonal cage, which means that the energy barrier listed in Fig. 6.20A needs to be overcome three times. Alternatively, there may also be a hybrid mechanism in which the guest first enters the neighbour cage and then goes into the void cage, finally reaching the diagonal cage. Unfortunately, to enhance sampling, the potential restrains do not allow one single simulation to describe the complete picture of the guest transfer containing both mechanisms. Therefore, further simulations are required to compare the different mechanisms of guest transfer in the same crystal from a thermodynamic point of view. Nevertheless, we confirm the existence of the two speculations at the beginning of this section, the “neighbour cage” mechanism and “void cage” mechanism. It is also found that as the size of the butyl side chain increases from the H-cage to the F-cage, the tunnel between neighbour cages as well as between diagonal cages are blocked, resulting in an increase in the activation energy of the guest transfer.

6.3. Assessment

In this chapter, we investigated the uptake process of ammonium ions in [4+4] imine cages and the nitrogen transfer in [2+3]imine cages, respectively, using enhanced sampling approaches based on the molecular dynamics simulation. By calculating the free energy surface, we revealed reaction mechanisms that are difficult to justify experimentally in detail. In addition, by applying MetaD-based approaches, we have also verified the reliability and feasibility of these tools when facing with systems with large time scales, which can be employed to study biomacromolecular systems with large system sizes as well as time scales.

7. Unravelling the mechanism of Glucose Binding Protein-based fluorescence probe

In parts reproduced with permission from with permission from the Royal Society of Chemistry.
Pang, Ziwei, Monja Sokolov, Tomáš Kubař, and Marcus Elstner. Unravelling the mechanism of glucose binding in a protein-based fluorescence probe: molecular dynamics simulation with a tailor-made charge model. Physical Chemistry Chemical Physics, 24, pp.2441-2453. DOI: 10.1039/D1CP03733A

7.1. Introduction

The high prevalence of diabetes has launched a challenge to human health. Effective diabetes treatment requires real-time monitoring of a patient's blood glucose concentration. However, such real-time continuous glucose monitoring (CGM) cannot be achieved by traditional monitoring means, such as finger prick methods. Although CGM sensors have been developed considerably, they still suffer from a lack of accuracy and short service time. Therefore, research and development of a new generation of CGM sensors is crucial.

Since the bacterial periplasmic glucose/galactose binding protein (GGBP) of the Gram negative bacteria *E. coli* [100] and *S. typhimorium* [101] can specifically recognise for glucose, it has often been used in recent years as a receptor for fluorescent probes to achieve CGM. These fluorescent probes have environmentally sensitive fluorophores attached to a specific location in the GGBP and the choice of location is decisive for the fluorescent signal of the probe. Usually, this position is required so that the environment of the fluorophore changes in response to the protein conformational change when glucose is bound, resulting in a strong fluorescence change. Therefore, the operating range of glucose concentrations, which is related to the glucose dissociation constant (k_D), and the corresponding changes in probe fluorescence need to be carefully considered for the design of such probes. Since the concentration of galactose in blood is much smaller than that of glucose, the binding of galactose to GGBP has little effect on the binding of glucose to the protein [102].

In 1998, Marvin and his coworkers selected the position of the fluorophores based on an examination of the crystal structure and constructed such sensors for the first time [103]. Subsequently, a large number of experiments have been carried out in terms of the selection of different fluorophores and positions of the fluorophores linked to the protein [102, 104–106]. Later, in 2008, the potential of this sensor for real-time sugar monitoring in microdialysis was demonstrated by Ge et al. [107].

A large number of achievements have been witnessed in the last two decades. In particular, Badan (6-bromoacetyl-2-dimethylaminonaphthalene) has proven as a reliable fluorophore. that can be attached to the H152C single mutant of GGBP, thus allowing for continuous monitoring of glucose [106]. Here, the glucose concentration can be simply estimated by,

$$k_D = \frac{[\text{ protein }][\text{ glucose }]}{[\text{ complex }]} \approx [\text{ glucose }]_{\text{blood}} \,. \tag{7.1}$$

Therefore, if the k_D of a glucose sensor is in or close to the common pathophysiological glycemic range in human bodies (1.7 mM to 30 mM), an optimal signal response will happen. This is because the amount of free and bound glucose is similar when enough fluorescent probes present in the sample, i.e. when the sugar concentration increases there is a sufficient amount of protein to bind new glucose molecules and when the sugar concentration decreases there is a sufficient amount of protein to cause a signal. Although it was shown in this earlier study that the binding of such single mutant had improved the k_D to 0.002 mM compared to the wild-type GGBP with k_D = 0.2 μM, it was still well below the ideal range. To make future clinical measurements possible, probes are need to be optimised to achieve the k_D at the millimolar level.

To this end, Khan et al. subsequently constructed a GGBP triple mutant labelled with Badan as a glucose sensor based on the GGBP single mutant, which exhibited a reasonable k_D = 11 mM in phosphate-buffered saline (PBS) [108]. Later, this high k_D triple mutant was immobilised on a solid surface and demonstrated to measure the fluorescence lifetime after glucose binding in *vitro* [109]. Meanwhile, Tiangco et al. developed a complete fibre-optic biosensor system with the GGBP-Badan single mutant that could measure transdermal glucose, which has lower concentration in blood than glucose [110]. In addition, a double GGBP double mutant with Badan was applied as a detector for changes in liquid glucose concentrations on the airway surface [111]. An overview of the dissociation constants is given in Table. 7.1.

Table 7.1.: Dissociation constant of the wild-type GGBP and its mutants with Badan linked to H152C.

GGBP type	k_D/mM	Reference
Wild-type	0.0002	Vyas et al.[100]
H152C	0.002	Khan et al.[106]
H152C/A213R	0.86	Helassa et al.[111]
H152C/A213R/L238S (in PBS)	11	Khan et al.[108]

Fig. 7.1A shows the wild-type GGBP crystal structure in the closed state with the glucose molecule in the binding pocket. The glucose is hydrogen-bonded (H-bonded) to multiple amino acids as shown in Fig. 7.1E. Four of those amino acids are charged, making the H-bonds to the glucose molecule even stronger. Moreover, the sugar is sandwiched between two aromatic amino acids, phenylalanine and tryptophan. The triple mutant H152C/A213R/L238S with Badan linked to Cys152 is shown in Fig. 7.1B. The mutation H152C eliminates one H-bond to the glucose, while another H-bond is introduced by the mutation A213R. The mutation L238S is also located in the proximity of the binding pocket but has no direct impact on the interaction pattern. Overall, there are fewer H-bonded contacts between the glucose and the protein in the mutant than in wild-type protein, compare Figs. 7.1E,F. This could be the reason for the large change in k_D, and will be investigated in this work.

The chemical structure of Badan (as linked to the side chain of Cys152 in the triple mutant) is shown in Fig. 7.1D. It contains an electron-donating dimethy-

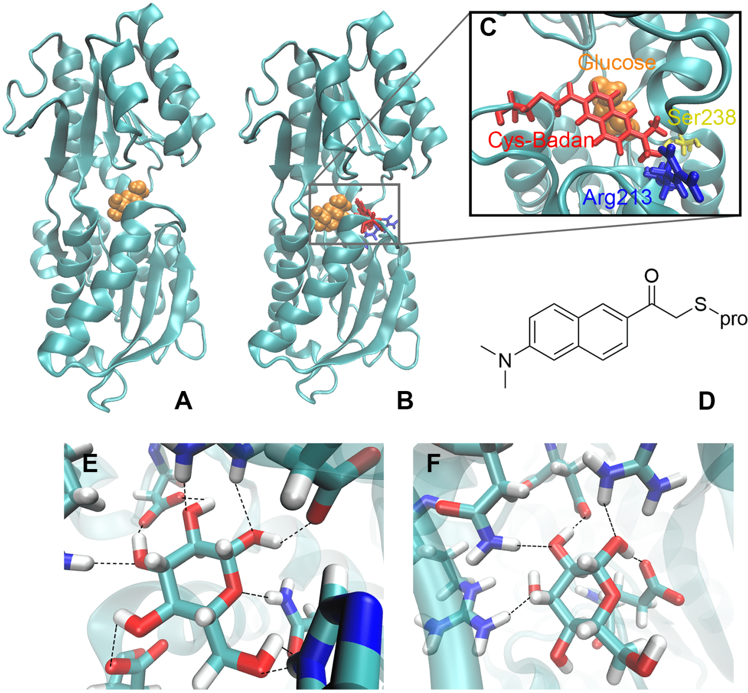


Figure 7.1.: Schematic view of wild-type GGBP and its triple mutant. **A).** Wild-type closed state GGBP. **B).** GGBP triple mutant H152C/A213R/L238S in its closed state. **C).** The binding pocket of the triple mutant. Glucose – orange; Cys-Badan – red; Arg213 – blue; Ser238 – yellow. **D).** Chemical structure of Badan linked to the protein via a cysteine side chain. The H-bonds between glucose and **E).** the wild-type protein or **F).** the GGBP triple mutant are shown by dashed lines. Some of H-bonds are missing due to the mutation.

lamino group and an electron-withdrawing carbonyl group with maximal distance from each other on the two sides of the naphthalene core, making it a push-pull charge-transfer system [112]. The first dye of this kind, Prodan, was synthesized by Weber and Farris [113]. In contrast to Prodan, Badan in its original form contains a thiol-reactive bromine and thus can be linked specifically to a cysteine side chain. Badan is a frequently used environmentally sensitive dye that changes its fluorescence properties depending on the polarity of its surroundings. In the triple mutant, the fluorescence intensity and excited states lifetime change upon glucose binding. So far, it has been assumed that the environment of Badan is more hydrophobic when the protein is in the closed conformation while binding the glucose [108]. However, there is no crystal structure of the triple mutant, therefore, the overall structure, the details of the binding pocket and the actual Badan conformations, depending on the details of the binding site, are unknown. Consequently, the reason for the change of fluorescence upon glucose binding is unclear, too.

With the development of computers and nanoscale modelling, molecular dynamics (MD) simulations on biological macromolecules have become affordable, and it is now possible to study the GGBP mechanism in atomistic detail. Recently, Unione et al. investigated the wild-type GGBP employing steered MD simulations (SMD) [114]. The free energy surface was estimated by the combination of two SMD trajectories, which included the protein opening with glucose and closing without glucose. Due to short simulation times, no continuous free energy map was computed, rather, the information resulting from individual MD simulations of open and closed state was plotted in a single energy scheme. Therefore, the GGBP opening-closing motion and the free energy surface was inferred from two individual hundred-nanosecond free MD trajectories, also including information from NMR measurements. Furthermore, only one collective variable (CV) was used in the SMD, which tends to be a too limited representation of the system's degrees of freedom, since the binding site of the GGBP is buried and the configurational entropy contribution may be large [73]. Panjaitan et al. also reported several short free MD trajectories to study the wild-type GGBP [115]. They found that the protein was unlikely to close after introducing the glucose ligand into the binding pocket, which is in contrast to the experimental findings. The authors argued that this may be due to the insufficient length of the simulations. In fact, it is a challenge for a hundred-nanoseconds free MD simulation to explore a large conformational change when facing a barrier significantly exceeding 3 kcal/mol, especially for large biomacromolecules like the GGBP.

According to the experimental dissociation constants (0.2 μM for the wild-type and 11 mM for the triple mutant), a free energy difference between *apo*-open (without glucose in the binding pocket) and *holo*-closed states (with the glucose in the binding pocket) of 9.2 kcal/mol is expected for the wild-type protein [100]. A value of 2.7 kcal/mol was reported for the triple mutant with Badan [108]. However, our initial simulations did not show a stable bound state over 1 μs, as detailed below, a similar finding as reported in the two previous simulation studies. In fact, a stable bound state for several hundreds of nanoseconds has never been reported so far. In one of our simulations the closed state opens within 100 ns of simulation time, which leads to an immediate glucose unbinding in contradiction to the experimental k_{D}. This may point at a shortcoming in the simulation protocol.

In fact, the H-bonds stabilising the bound state seem to be too weak in the simulations, in order to keep the binding pocket closed on the expected temporal scales. The GLYCAM parameters used here and in the previous computational studies are derived for use with the TIP3P water model [116, 117]. However, the glucose molecule bound in the GGBP pocket is highly polarised by several strong H-bonds as shown in Fig. 7.1B. Therefore, the GLYCAM charge model may not describe glucose in such polarising environments appropriately.

The importance of polarisation effects on force field parameters, in particular on the force field charges was pointed out in several studies of protein-ligand binding [118–121], peptide folding [122, 123], and protein-chromophore complexes [124, 125]. Standard force field parameters, in particular the force field charges, can lead to structural instabilities or even a wrong description of the respective systems. This was previously tackled by explicitly considering the polarisation induced by the specific environment, using one of several approaches. The most accurate but also computationally demanding way is the use of a polarisable force field, where the atomic charges are determined in an iterative procedure. Computationally less demanding is the use of polarised protein-specific charges (PPC), which are computed in order to represent the specific polarised electrostatic state of the protein. The polarised charges are usually determined for one representative structure, but also charge update schemes have been proposed [122, 126]. Typically, PPC are determined using a molecular fractionation with conjugate caps, followed by the calculation of the electron density of the fragments using DFT with a subsequent restrained electrostatic potential fit (RESP) [127]. If PPC for a whole protein are to be determined, an iterative procedure is chosen where the charges of the various fragments are recalculated until convergence is reached, while the charges of the remainder of the protein are represented by point charges. It is also possible to fit only the charge variation (delta-RESP) instead of using the conventional RESP approach for charge fitting [128, 129].

The importance of considering polarisation effects was demonstrated for several examples: Mei et al. reported that the melting temperature of a small Trp-cage protein obtained from the simulation was only in agreement with the experiment when PPC were applied [123]. Tong et al. reported that in contrast to Amber charges, PPC kept the studied light-harvesting complex stable during the simulation and provided also a reliable description of the environment in QM/MM calculations on the chromophores [125]. For protein-ligand binding, Duan et al. showed that the binding energies of complexes of the cycline-dependent kinase with five different ligands agree significantly better with experiment using PPC charges than with the unpolarised Amber charges [120].

In this work, we follow these earlier studies and develop force field parameters that account for the highly polarised protein environment. In a second step, these charges are applied in free MD simulations to investigate bound and unbound structures of wild-type GGBP, as well as the GGBP-Badan triple mutant. These parameters lead to a stable glucose binding pocket with a preserved number of H-bonds compared to the initial structure. Further, the overall structure of the protein remains stable, in contrast to the standard GLYCAM charge model. Metadynamics simulations applying these charge models are then performed to achieve a more detailed insight into the mechanism and energetics of the opening and closing mechanism.

7.2. Computational details

The initial structure of the wild-type GGBP and the GGBP-Badan triple mutant (hereafter referred to as the “triple mutant”) were taken from the closed GGBP crystal structure, PDB ID 2FVY [130]. For the triple mutant, the residues His152, Ala213 and Leu238 were replaced with cysteine, arginine and serine residues, respectively, and the side chain of Cys152 was functionalized with the Badan fluorophore, see Fig. 7.1D. These changes were performed with the xLeap tool from the AmberTools package [131]. The following force fields were employed: Amber14SB for the protein [132]. general Amber force field (GAFF) for the newly parametrized Badan moiety [47, 90], Joung–Cheatham parameters for the ions [133], and GLYCAM 06j for the glucose [117]. Additionally, new atomic charges for the glucose were derived as detailed in the following, and bonded parameters were taken from GAFF.

7.2.1. Polarised force field for the glucose molecule

Initial simulations employing the standard GLYCAM 06j atomic charges for glucose showed no stable binding to the proteins, as detailed in the following section 7.3.1. Similar findings have been repeatedly reported in the literature for a diverse set of systems: there is a large body of evidence that standard force fields with a fixed point charge model fail to describe H-bond strengths with sufficient accuracy for a variety of systems [118–125]. This seems to apply also for the case of GGBP: in pilot simulations, we were unable to find a stable glucose bound state. The H-bonded network broke and set the glucose free, in contrast to the stable bound state found in experiment. This indicates that the GLYCAM 06j charge model may be insufficient to describe the strong H-bonded network of the GGBP binding pocket, as described above. Here, the glucose molecule is located in the highly polar protein binding pocket, restricted to one stable conformation, which seems to be very different from the more weakly bound floppy structure embedded in water solvent described by TIP3P water, for which GLYCAM 06j is parametrised.

For GGBP–glucose binding, we consider the major source of error to be the glucose charges. Therefore, we decided to only update those and leave the protein charges unchanged. This makes the procedure particularly simple as it is non-iterative and glucose charges are determined only for a single conformation, the bound state of the protein. Update schemes developed to follow conformational changes of proteins [122] seem to be less practical for the GGBP–glucose binding case. The focus is on enhanced sampling simulations of the binding–unbinding reaction, during which the glucose is moving back and forth between the protein and water environments which would require a frequent charge update making the calculations computational expensive. Using the charges for the protein-bound state will lead to less accurate charges for the water-solvent case, an error which we will estimate using alchemical free energy calculations as described below.

Two different new sets of atomic charges were created for the glucose polarised in the binding pocket by Monja Sokolov, one for the wild-type protein complex and one for the triple mutant. The procedure started by docking the glucose molecule into the binding pocket of the closed protein, either the wild-type GGBP or the triple mutant, which was subsequently immersed in a TIP3P water box. Standard force field parameters were assigned to the entire system, including the standard GLYCAM 06j atomic charges for the glucose. Then, energy minimisation was performed with steepest descents.

The resulting glucose conformation was taken as input for HF/6-311G* ESP calculations, where atomic charges were subsequently determined with RESP. To account for the polarising environmental effects, we included the force field point charges of the *apo*-protein up to a certain cutoff radius. This was performed twice, for the wild-type protein and the triple mutant, leading to two different sets of binding pocket polarised charges (BPC) for each specific molecular complex. To estimate possible errors due to a hard cut-off, the results from two different approaches were compared: First, atomic charges from atoms within different cutoff distances from 5 Å to 8 Å from the glucose were included. In the other approach, a residue-based cutoff was applied, including all residues for which at least one atom is found within certain distances. No obvious difference was observed in pilot free MD simulations of 50 ns, as shown in Table 7.2, and the charges determined for an atom-based cutoff at 5 Å were used in the following.

Table 7.2.: The average number of H-bonds generated between water molecules and BPC glucose molecules with different cut-off.

Cut-off range	Average number of H-bonds
4Å with full residues	10.80
5Å with full residues	10.77
5Å with broken residues	10.46
6Å with broken residues	10.76
7Å with broken residues	10.44
8Å with broken residues	10.68

In addition, we also computed a third set of charges for the glucose molecule in aqueous solution. Glucose was optimised at the B3LYP/6-31G* level in the presence of implicit water represented by the polarisable continuum model [134]. Then, the electrostatic potential was computed at the HF/6-31G*/PCM level, and a set of atomic charges was obtained with RESP; this charge model will be referred to as water polarised charges (WPC).

We further computed a gas-phase charge model (GPC) using the same methodology, and in addition, a charge model based on DFTB Mulliken charges. For the latter, two QM/MM simulations were performed using GROMACS [92–94, 135]: one with the glucose in water and one with the glucose in the binding pocket. The sugar was treated with the semi-empirical DFTB/3OB method [44, 136] using DFTB+ [137, 138], and the environment was described with a force field. These calculations were intended to investigate charge fluctuations along trajectories, and are detailed in the Appendix.A. However, DFTB Mulliken charges turned out to be largely underpolarised, therefore, they are not further considered in this work.

7.2.2. Free MD simulations

We performed three free, unrestrained MD simulations of the wild-type GGBP with a bound glucose molecule: one with the GLYCAM force field for the glucose, one with the BPC, and one more with the WPC charge model. In addition, free MD simulations under the same conditions were carried out on the GGBP-Badan triple mutant with the BPC as well as the WPC glucose, on the triple mutant without glucose and on the wild-type GGBP without glucose.

In all of these simulations, the protein–glucose complex was embedded in a dodecahedral TIP3P water box keeping a distance of the solute from the edges of the box of at least 20 Å. There is a calcium ion bound to GGBP, and electroneutrality was achieved by replacing six water molecules by sodium ions. No extra salt was added. Periodic boundary conditions were applied and long-range electrostatics was described by the particle–mesh Ewald method [139]. Each simulation was carried out with cut-offs of 1.4 nm for both vdW and real-space PME interactions

The equilibration procedure started with a steepest descents minimisation to reduce all forces below 1000 kJ mol^{-1} nm^{-1}, followed by a conjugate gradient minimisation until all forces dropped below 500 kJ mol^{-1} nm^{-1}. Then, the system was heated to 298 K during an NVT MD simulation of 1 ns with a time step of 2 fs using the Bussi thermostat [66]. Here, the lengths of all bonds were kept constrained to their respective equilibrium values by means of the LINCS algorithm. Subsequently, an NPT simulation of 1 ns with a time step of 2 fs was performed at a temperature of 298 K and a pressure of 1 bar maintained by the Nosé–Hoover thermostat [140, 141] and the Parrinello–Rahman barostat [67, 142], respectively. Position restraints with a force constant of 1000 kJ mol^{-1} nm^{-2} were imposed on all of the protein atoms during the equilibration procedure above. Finally, the system was equilibrated for further 10 ns keeping only bonds involving hydrogen atoms constrained with LINCS. Identical settings were used to carry out the actual production simulations of 1 µs.

7.2.3. Well-tempered metadynamics simulations

To further investigate the binding of glucose and the opening of the binding pocket, well-tempered metadynamics simulations [**wtmetad**] were performed for both the wild-type GGBP and the triple mutant. These were started from each respective closed structure and involved the atomic charges on the glucose molecule that were polarised for each respective binding situation, as described above. Two collective variables (CV) were employed based on previous work by others [114, 143]: The opening–closing motion is described with the angle θ between centres of mass of the N-terminal domain (residues 3–108 and 258–291), the hinge region (residues 109–111, 255–258 and 292–296) and the C-terminal domain (residues 112–254 and 297–306). The twisting motion is described with the torsion angle ϕ defined by the centres of mass of the N-terminal domain, the N-terminal domain base (residues 109, 258 and 292), the C-terminal domain base (residues 111, 255 and 296) and the C-terminal domain. These CVs are illustrated in Fig. 7.2 and were also used to analyse free MD trajectories in the following discussion. Besides, a set of restraints was imposed on the protein as well as the glucose molecule to maintain the stability of the protein structure and to concentrate the sampling process on the binding/unbinding of glucose and opening/closing of the binding pocket; these are described in detail in Appendix.B via input codes. With a time step of 1 fs and a bias factor of 10, the metadynamics simulations were extended to 2.6 µs for both wild-type and the triple mutant, respectively.

7.2.4. Alchemical free energy calculations and funnel metadynamics simulations

The binding–unbinding simulations were performed with a fixed-charge model. This BPC charge set, however, is optimised for the glucose molecule in the binding pocket, and may yield overestimated solvation free energy of glucose in water. To take into account the change of charge distribution on the glucose molecule upon unbinding from the protein, a series of free energy simulations was carried out for both the wild-type GGBP and the triple mutant, as follows. The resulting free energy shall give an estimate of a possible error in the free energy profiles obtained from the simulations described above.

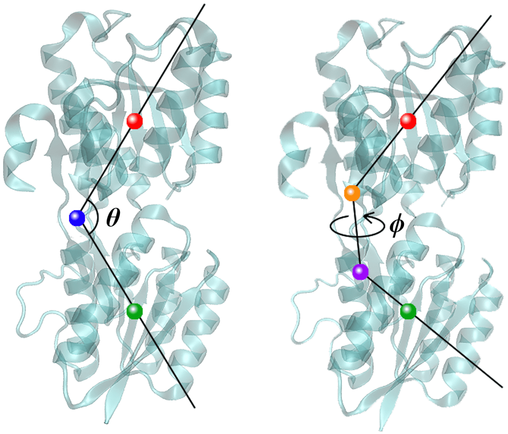


Figure 7.2.: Schematic view of collective variables used to describe GGBP. Left: θ, representing the closing–opening motion; Right: ϕ, representing the protein twisting motion. The centres of mass of respective parts of the protein that define the angle/torsion are shown as solid balls. N-terminal domain–green, C-terminal domain–red, hinge region–blue, N-terminal domain base–purple and C-terminal domain base–orange.

First, alchemical simulations were performed to obtain the glucose binding energy difference when passing from the BPC glucose charges to GLYCAM charges, with the protein binding pocket remaining in the open state. The respective wild-type and triple-mutant BPC charge models were used here. Second, funnel metadynamics simulations [72] were carried out using the GLYCAM glucose charges to obtain the free energy of unbinding. In all of these simulations, the protein structure was constrained at the corresponding open state local minima obtained by well-tempered metadynamics simulations for the respective protein (wild-type or triple mutant GGBP). All sets of alchemical simulations contain 21 λ windows with a spacing of 0.05, each consisting of an MD simulation of 4 ns. All funnel metadynamics simulations were extended to 750 ns to achieve convergence. More details about the settings of funnel metadynamics simulations can be found in Appendix.B.

Free MD simulations and alchemical calculations were carried out with Gromacs 2018.3. Well-tempered metadynamics simulations and funnel metadynamics were performed with PLUMED 2.5 [95, 144] linked to Gromacs 2018.3. Molecular structures were constructed and visualised with VMD 1.9.2 [96].

Quantum chemical calculations were performed with Gaussian09,[91] and RESP was run with Antechamber [90, 131].

7.3. Results and Discussion

7.3.1. Polarisation of the glucose molecule

The wild-type GGBP has a k_D of 0.2 μM [100], corresponding to a binding free energy of ca. 9 kcal/mol, which means that at least hundreds of microseconds would be needed to see a change from the *holo* closed state to the *apo* open state. However, when employing the GLYCAM parameters for glucose, the wild-type GGBP changed from its closed state to an open state within hundred nanoseconds, which is orders of magnitude faster than expected, as shown in Fig. 7.3. This finding agrees with previous studies [114, 115], where no wild-type GGBP closed state simulation for several hundreds of nanoseconds is reported. It further indicates that the ligand in the binding pocket may not be sufficiently stabilised during the MD simulations, most probably due to too weak H-bonds as a result of the applied GLYCAM 06j charge model.

We applied a polarised charge scheme for glucose. Two sets of polarised charges were derived for the wild-type and triple mutant (BPC) binding pockets, one for bulk water (WPC) and one for the gas-phase (GPC). Comparison between these charge models are given in Fig. 7.4 and detail charge parameters are illustrated in the Appendix.A. With our repolarising methods, both hydrogen and oxygen charges were increased compared to the GLYCAM charges, so that the H-bonds based binding between ligand and protein can be strengthened. From another perspective, as shown in Table. 7.3, dipole moments of BPC charges for both wild-type and triple mutant are larger than the GLYCAM, which also proves stronger binding happening with BPC charges. In addition to the larger dipole moment, as shown in Fig. 7.5, the dipole moment direction of WPC glucose is closed to the GLYCAM glucose. However, larger direction deviations have been witnessed for both wild-type and triple mutant BPC glucose, which is caused by the repolarisation in the binding pocket environment. On the other hand, it indicates that the BPC charge compensates for the error caused when using the general GLYCAM force field.

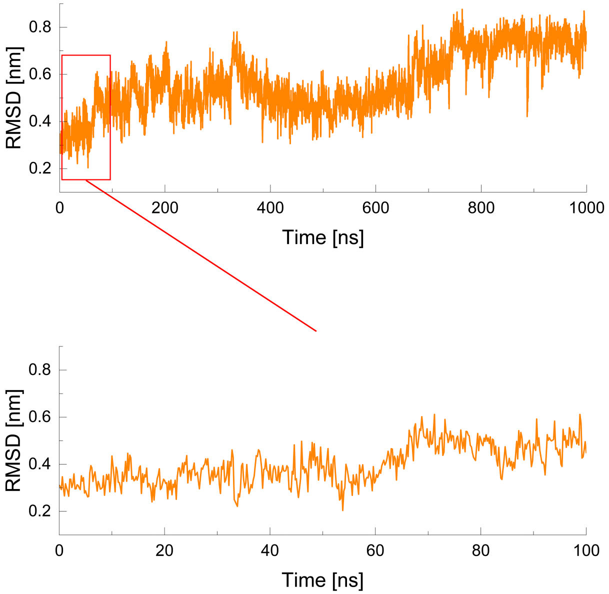


Figure 7.3.: The RMSD with respect to the crystal structure (2FVY) of 1 μs pilot free MD trajectory of the wild-type GGBP.

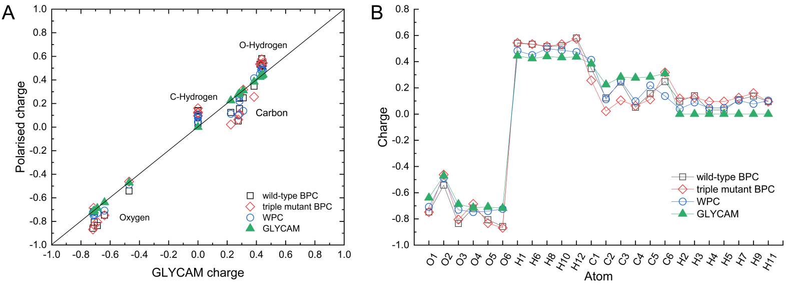


Figure 7.4.: A) Correlation diagram between repolarised glucose charges and GLYCAM glucose charges. B) Charges of each glucose atom. H1,6,8,10,12 are O–Hydrogen atoms, H2–5,7,9,11 are C–Hydrogen atoms.

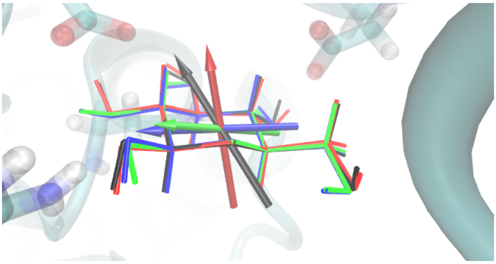


Figure 7.5.: Dipole moment direction of wild-type BPC (black), triple mutant BPC (red), WPC (blue) and GLYCAM (green) charges glucose.To clearly compare the dipole moment directions, triple mutant BPC, WPC and GLYCAM charges glucose is aligned to the wild-type BPC glucose in the wild-type GGBP binding pocket.

We also assigned these polarised charges to glucose and performed a set of 50 ns MD simulations for the glucose in the two binding sites, respectively. As shown in Fig. 7.6, the H-bonded network of the wild type stays intact when using the BPC and WPC charges. We find an average of nine H-bonds between the glucose and the binding pocket, which is the number of H-bonds also found in the crystal structure as indicated in Fig. 7.1E. By contrast, when using the GLYCAM charges, there are much fewer H-bonds, which frequently break so that the binding pocket opens. As discussed below, the

Table 7.3.: Dipole moment of polarised glucose with different basis sets

Basis sets	GPC	WPC	wild-type BPC	triple mutant BPC
6-31G*	3.28	4.15	4.38	4.15
6-311G*	3.39	4.29	4.51	4.26
def2-TZVP	3.09	3.99	4.35	4.08
def2-TZVPP	–	3.94	4.34	4.06
def2-QZVP	–	3.95	4.35	4.06
def2-TZVPD	3.01	3.95	4.37	4.06
def2-TZVPPD	–	3.95	4.38	4.09
def2-QZVPD	–	3.93	4.36	4.06
MD in water	4.59	4.74	5.40	5.20

GLYCAM charge dipole moment = 3.75 with 6-31G* [116]

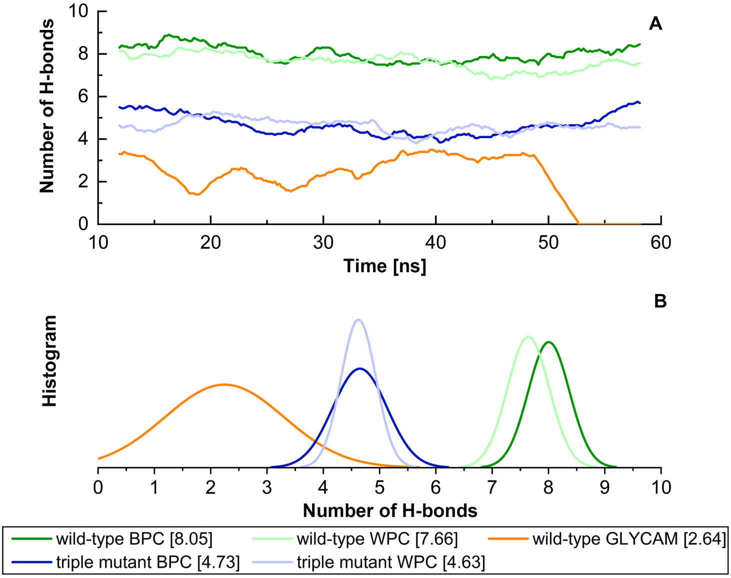


Figure 7.6.: A). The number of H-bonds formed between the binding pocket and the glucose polarised in different environments. **B).** The histogram of the number of H-bonds observed in simulations run with the different glucose charge models. The average number of H-bonds of each corresponding glucose is given in square brackets. Trajectories are taken from corresponding free MD simulations from 10 ns to 60 ns. For the wild-type protein with GLYCAM charge glucose, the ligand escaped from the binding pocket at ca. 50 ns.

free energies of opening/closing of the binding pocket agree very well with the experimental values, supporting the use of such polarised charges.

For the mutant, a smaller number of H-bonds is found as indicated in Fig. 7.1F. The change in the experimental k_D indicates such a behaviour, and again, the agreement with experimental estimates of the unbinding free energies supports the usage of these charges.

The use of fixed polarised charges has some drawbacks, which can only be avoided using a fully polarisable electrostatic model in principle. The dynamical transitions, i.e. the opening and closing of the binding pocket and unbinding and binding of the glucose seem to ask for a change of the charge model during the process, since the polarisation of glucose in solution differs from that in the binding pocket obviously. One way to deal with this could be the usage of an average model, however, we decided to use both BPC and GLYCAM models in the simulations and critically discuss the results. Using the BPC model, glucose could be overstabilised in the solution phase due to the larger dipole moment, while using the GLYCAM model glucose is most probably understabilised in the binding pocket. Since we are interested in the kinetic barrier for unbinding, we use the BPC models for the simulations and discuss corrections to this polarised model below.

7.3.2. Standard MD simulations of the wild-type and the triple mutant

The collective variables, θ and ϕ (see Fig. 7.2), adopted for metadynamics simulations in this study, were also used to analyse free MD trajectories, as described in the methods section. The corresponding values for the crystal closed (PDB-ID: 2FVY) and open (PDB-ID: 2FW0) wild-type GGBP are shown in Tab. 7.4. As discussed above, we performed MD simulations with both charge models, the WPC and BPC charges.

Table 7.4.: Collective variable values for the crystal wild-type GGBP.

Collective variable	closed GGBP	open GGBP
θ	121.6°	143.1°
ϕ	64.9°	89.6°

Fig. 7.7 shows the results of three extended MD simulations of over 1 µs each: In the first two simulations, we used the WPC and BPC glucose charges and started the simulations from a crystal closed state GGBP structure. The protein remained in a stable closed state in both simulations with the glucose molecule inside the binding pocket. Both free MD simulations show a root-mean-squared deviation (RMSD) with respect to the crystal closed wild-type below 0.2 nm, and keep the distance between the glucose and the centre of the binding pocket within 0.3 nm during the simulation of 1 µs. It is interesting to see that both WPC and BPC glucose charges lead to a stable bound state for 1 µs, which is in agreement with the experimental k_D value and is in contrast to the results using the original GLYCAM parameters.

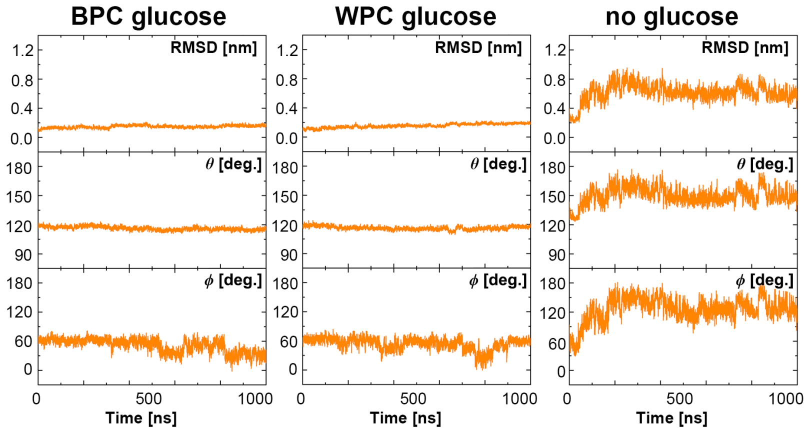


Figure 7.7.: The structure of the wild-type protein in the free MD simulations of the wild-type GGBP. Left: Simulation started at crystal closed state with BPC glucose; Middle: Simulation started at crystal closed state with WPC glucose; Right: Simulation started at crystal open state without glucose. RMSD from the closed GGBP crystal structure (2FVY) is considered. The angle θ describing the opening-closing motion of GGBP and the dihedral ϕ describing the twisting of the domains (details in text) are plotted versus simulation time.

As already mentioned, the glucose molecule is strongly H-bonded in the highly polar binding pocket, therefore, it is expected that the molecule is polarised to a large degree, and parameters derived for a less polar and less strongly H-bonded environment may not optimally describe this situation. The results therefore indicate that the reparametrisation seems to be the right way to go. That alone, however, is not a justification for this approach: more evidence comes from the calculation of the free energies of binding/unbinding, as described below. The fact that the newly derived parameters are able to describe these energies for both, wild-type and mutant with largely different energetics, is highly encouraging.

In the third simulation, which was started from the crystal open state GGBP structure without glucose, the protein immediately closed and then changed its conformation back to the open state after 50 ns. At this point, we like to note that the open structure seen in the X-ray experiment [130] had been crystallised at high salt concentration – different from our simulations that are closer to low-salt, physiological conditions. This may be a reason why the stable open state structure from simulations deviates from the experimentally observed wild-type open structure. The RMSD compared to the wild-type crystal closed structure correspondingly dropped from 0.4 nm to 0.2 nm and then increased up to 0.6 nm, indicating that the wild-type protein is unlikely to remain in a closed state without the glucose.

Besides, as shown in Tab. 7.5, the structures along these three trajectories exhibit a highly similar average θ value of 116° in the closed state and 152° in the open state similar to the θ values reported previously [114, 143]. In the closed state, the ϕ value fluctuates around the crystal structure value of 65°. The regions of lower ϕ values of about 20–30° indicate a twisted closed state, which we will discuss in more detail below. In the open state, the average ϕ value is ca. 40° larger than the crystal open structure value of 90°. This increased flexibility with respect to the crystal structure may be expected due to the removal of the crystal packing constraints in the simulations of GGBP in aqueous solution.

Table 7.5.: Average values of θ and ϕ and their deviations compared to the crystal wild-type GGBP.

Collective variables	θ		ϕ	
	average	**deviation**	**average**	**deviation**
wild-type closed state	116.3°	−5.3°	51.3°	−13.6°
wild-type open state	151.7°	+8.6°	132.7°	+43.1°
triple mutant “cc” state	127.5°	+5.9°	76.8°	+11.9°
triple mutant “cc*” state	124.1°	+2.5°	69.7°	+4.8°
triple mutant “tc” state	116.7°	−4.9°	76.3°	+11.4°
triple mutant “tc*” state	127.1°	+5.5°	85.9°	+21.0°
triple mutant “tc**” state	128.1°	+6.5°	128.6°	+63.7°
triple mutant “op” state	145.1°	+2.0°	126.5°	+36.9°
triple mutant “op*” state	154.0°	+10.9°	137.0°	+47.4°

Just like for the wild type, we performed a set of three MD simulations of 1 µs each for the GGBP triple mutant. The results from the analysis of structures along the trajectories are shown in Fig. 7.8. Recall that since no crystal structure is available for the triple mutant, the crystal structures of the wild-type protein, 2FVY and 2FW0 were considered as initial structures after manual mutation of the three residues. In the following, the label ‘crystal structure’ refers to the mutated 2FVY initial structure. Due to the mutations, the glucose is much less tightly bound, as the experimental k_D value indicates, and we indeed see a more dynamical behaviour already on this microsecond scale.

As seen in Fig. 7.8A for the BPC model, the protein changes from a crystal closed state (cc) to an open state (op, red to blue) after 70 ns. After ca. 200 ns, the protein returns to a metastable twisted closed state (tc, blue to green). Although the protein was open for a significant amount of time, the glucose did not leave the binding pocket entirely. Rather, it interacted with the residues of the C-terminal domain, so that the bound state is recovered after closing. This state shows slightly larger values of the dihedral angle, and is therefore called a twisted closed state labelled ‘tc’. The deviation from the wild-type GGBP can be expected due to less stable H-bonds, as described above. The bound structure is schematically shown in Fig. 7.9D, and comparison with the wild type in Fig. 7.9C shows the dramatically reduced H-bonding of this variant.

In the simulation with WPC glucose (Fig. 7.8B), an unbinding process occurs as well as large conformational changes of the protein: first, an opening and closing process are observed (from red to pink), the glucose leaves the binding pocket at ca. 200 ns when the protein is in the open state during this period. Afterwards, the protein deformed from the crystal closed state (cc*) into a twisted closed state (tc*, from pink to purple), which is stable for almost 500 ns without containing glucose in the binding pocket. Finally, the protein opens again (from purple to yellow). Note that the labels with asterisk denote *apo* states, while the labels without asterisk stand for *holo* states, as found during the different simulations.

To investigate the dynamics without glucose (Fig. 7.8C), a simulation was started from the crystal open structure. The protein immediately changed to the crystal closed state (cc*). After 70 ns, the protein returned to the open state (from pink to yellow) before finally reaching a super-twisted closed state (cc**, from yellow to cyan). Note that – just like in the case of the wild-type – our simulation system has no extra salt, while the X-ray structure had been resolved in experiments performed at high salt conditions [106, 108]. This is a possible reason for the open state structure in simulations deviating from the crystal open state. Furthermore, the mutations in the protein may also have affected the structure and stability of the open state.

Events occurring along single trajectories, however, may not be conclusive to evaluate the different parameter sets used. We therefore use the insight from these simulations merely to determine intermediate structural motifs, which we will use to interpret the free-energy simulations as discussed below. Having now stable structures for sufficiently long temporal scales allows us to characterise these structures in solution and compare to the crystal structure.

For both simulations with glucose, average θ values at crystal closed (red and pink region, ca. 125°) and open (blue and yellow region, ca. 150°) state agree well with the crystal wild-type GGBP ($\theta = 122°$ and 143°), as shown in Tab. 7.5. The θ value at "tc" state (green region, ca. 117°) is, like the wild-type closed state, ca. 5° smaller than the crystal closed state, which indicates that the twisted state is slightly more closed.

Larger deviations from the wild-type crystal structure are found for ϕ, see Tab. 7.5. The crystal closed states have average ϕ values between 51°–77°. The open states have ϕ values beyond 90° extending to 180° (Fig. 7.8C, yellow region); the twisted closed states have an average ϕ value of 76° and 86°; the super-twisted closed state has an average ϕ value of 129°. In summary, compared to the closed crystal structure, the wild-type as well as the triple mutant in solution show metastable states which are similarly closed but twisted, while both twist directions are possible. Compared to the wild-type open crystal structure, the structures of the wild-type and triple mutant proteins in solution are more open and significantly more twisted.

Fig. 7.9 illustrates the average number of H-bonds between the glucose and the side chains of the proteins. Using the BPC parameters, nearly nine H-bonded interactions are found for the wild-type protein in the crystal closed state, which is close to the experimental estimate [100, 145]. Only half of the H-bonds remain between the BPC glucose and the GGBP triple mutant in the twisted closed state. In the "tc" state, Arg213 forms an H-bond to the glucose molecule as shown in Fig. 7.9. Additional H-bonds are formed with Arg158, Asp236, Asn211 and Glu93. The latter is not interacting with glucose in the wild-type. Compared to the wild-type, the triple mutant's twisted closed state misses H-bonds to Asp14, Asn91, Asp154 and the mutated residue 152. As a result, the glucose hydroxyl groups are not interacting with protein residues, and they form H-bonds with solvent water molecules instead. In the wild-type, the glucose is additionally stabilised by two aromatic residues Phe16 and Trp183. In the triple mutant, Phe16 moves away from the glucose molecule and the stabilisation due to the aromatic sandwich structure is missing, as seen in Fig. 7.9E&F.

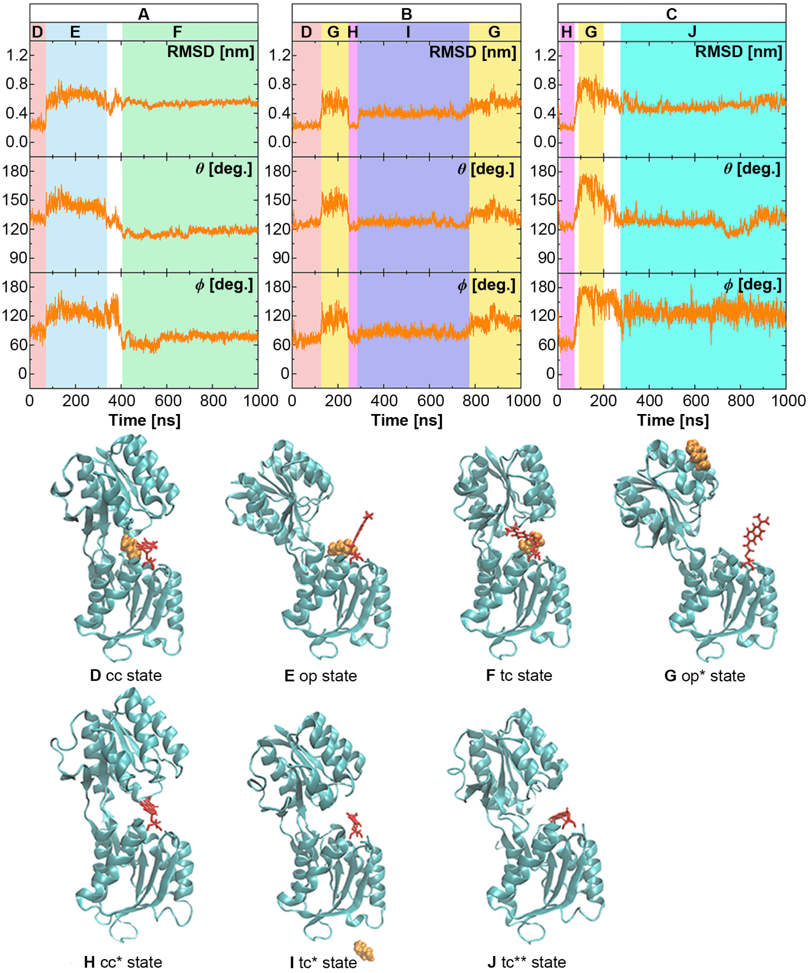


Figure 7.8.: The structure of the GGBP triple mutant protein in the free MD simulations of the GGBP triple mutant. The RMSD was calculated compared to the backbone of closed wild-type GGBP crystal structure (2FVY). The glucose is coloured with orange and the Cys-Badan residue with red. A) Simulation with the BPC glucose started with a *holo* crystal closed state. B) Simulation with the WPC glucose started with a *holo* crystal closed state. C) Simulation without the glucose started with a *apo* crystal open state. The red region represents *holo* crystal closed (cc) state; The pink region is *apo* crystal open (cc*) state; The blue region represents *holo* open (op) state; The yellow region represents *apo* open (op*) state; The green region represents *holo* twisted closed (tc) state; The purple region represents *apo* twisted closed (tc*) state; The cyan region represents *apo* super-twisted closed (tc**) state. Schematic representation structures D, E and F were taken from trajectory A, G and I from trajectory B, H and J from trajectory C, respectively.

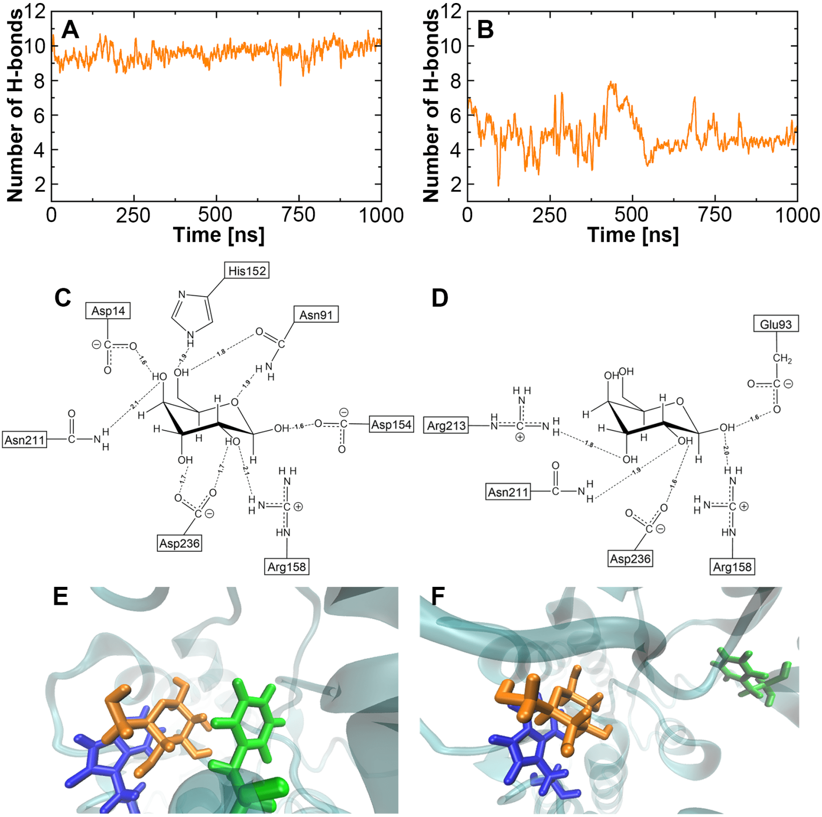


Figure 7.9.: The average number of H-bonds between glucose and the GGBP side chains along the free MD trajectories of wild-type GGBP with BPC glucose (A) and triple mutant with BPC glucose (B). C) schematic representation of the H-bond pattern in wild-type at 500 ns, the protein is at a crystal closed state and there are 9 strong H-bonds; D) schematic representation of the H-bond pattern in the triple mutant at 750 ns, the protein is at "tc" state and there are 5 strong H-bonds; E) The carbohydrate–π interaction between glucose and the wild-type binding pocket; F) The carbohydrate–π interaction between glucose and the triple mutant binding pocket. The glucose–Phe16 interaction is missing due to the mutation. Glucose – orange; Phe16 – green; Trp183 – blue.

7.3.3. Free energy surfaces

To explore a full free energy surface (FES) of the closing–opening motion of the protein, metadynamics simulations of wild-type and triple mutant were performed beyond 2.5 μs until convergence. We chose the BPC charge model because a proper description of the bound state has to be assured for the unbinding barrier to be overcome. This means, however, that the free energy of solvation in the water bulk phase may be described less accurately.

The FES of wild-type GGBP is shown in Fig. 7.10A, with well defined closed and open states denoted by *a* and *b*. The CV values at the global minimum *a*, $\theta \approx 120°$ and $\phi \approx 30°$, agree well with the results from the free MD simulations discussed above. The CV values of state *b*, $\theta \approx 160°$ and $\phi \approx 170°$, are much larger than those reported for the crystal structure. As discussed in Ref. [114], a possible reason for this deviation is that the ligand-free crystal structure is stabilised in a more closed state due to the crystallisation reagents. There is a free energy difference of $\Delta G = 9.4$ kcal/mol between the closed state and open state, and the reaction barrier is 11 kcal/mol.

We also find a "semi-closed" state *c* (Fig. 7.10A), which is twisted compared to the minimum *a*. Such a state has never been reported before and could be part of an alternative pathway for the closing–opening motion in the wild-type GGBP: the closed protein firstly twists from state *a* towards state *c*, which can be seen as an intermediate, and then opens to state *b*.

The FES of the triple mutant is shown in Fig. 7.10B with closed state *a* and open state *b*. The free energy difference of 0.8 kcal/mol is slightly less than the experimental value of ca. 2.7 kcal/mol [108], and the energy barrier of 9 kcal/mol is only slightly lower than that of the wild-type. The states D–J discussed for the free MD simulations are distributed along the reaction coordinate connecting the states *a* and *b*. Compared to the wild-type crystal closed structure, the state *a* is a slightly more closed and twisted conformation with $\theta = 115°$ and $\phi = 50°$.

Comparing wild-type and mutant, a slight difference in conformation and opening mechanism is visible in Fig. 7.10: The closed state *a* spans a much wider range of ϕ, i.e., this structure can exist in more twisted conformations, while the global minimum of the mutant is much more localised in the CV space. Further, the opening motion seems to follow slightly different pathways: While a twisting motion along ϕ is followed by an opening of the

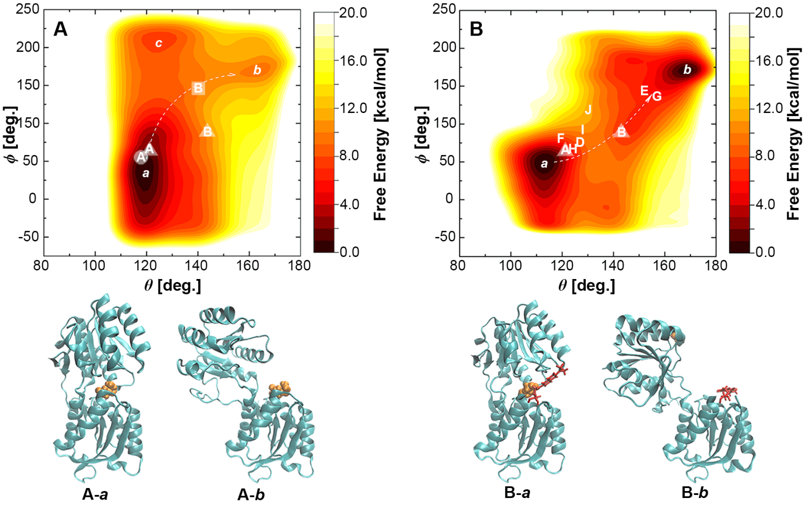


Figure 7.10.: A) The free energy landscape of the wild-type GGBP with collective variables θ and ϕ. The label with circle represents the *holo* state, square represents the *apo* state. State A in circle and B in the square is picked up from free MD trajectories of the wild-type GGBP. B) The free energy landscape of the GGBP-Badan triple mutant with collective variables θ and ϕ. States D to J are taken from three free MD trajectories of the GGBP triple mutant, and represent the crystal 'cc', 'cc*', 'op', 'op*', 'tc', 'tc*' and 'tc**' state as shown in Fig. 7.8, respectively. States A and B in the triangles indicate wild-type crystal closed and open structures, and the white arrows indicate pathways between the closed and the open states. The corresponding schematic representations of local minima *a* and *b* for wild-type and triple mutant are shown as A–a, A–b, B–a and B–b.

pocket with increasing θ in the wild-type, the motion in the mutant follows the opposite order.

There are two potentially small imperfections to note. First, the computed energetics clearly depend on the force field parameters. The glucose molecule is located in an unusually strongly polarising environment, which is an extreme situation to deal with. In such a case, the general purpose parameter set does not depict the true distribution of charge, leading to wrong energetics and potentially to qualitative errors in simulations. Here, we have reparametrised the force field charges, which fixed the qualitative failure of the previous simulations. However, with the BPC charges, a glucose molecule is overpolarised in water, which may result in an overstabilisation in water. Second, the position of the glucose molecule is not accessible from the applied CVs.

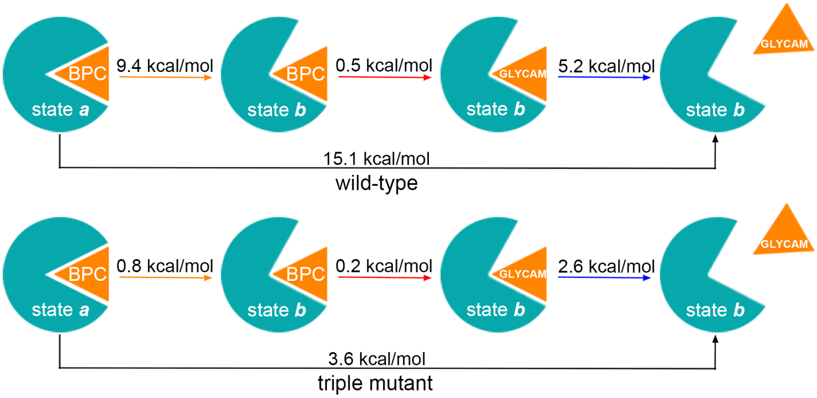


Figure 7.11.: Two series of free energy simulations for protein structural changes from *holo* closed state to *apo* open state. Orange arrows – Free energies of *holo* protein opening from metadynamics. Red arrows – $\Delta\Delta G_{\text{binding}}$ from alchemical simulations. Blue arrows – GLYCAM glucose unbinding energies from funnel metadynamics.

As seen in Fig. 7.12, the glucose remained at the binding site in both open and close state. The advantage of this is that the closed state is always the glucose bound state in the simulations, and there is no mixture of closed *holo* and *apo* states. Therefore, the barrier and free energy of binding in the closed state are described correctly. This is probably due to the fact that the simulations were started with the closed *holo* states and the reparametrised force field charges were taken for the glucose, and the glucose remained in the binding site during the process of protein opening and closing. For the open state, however, a small error may arise because the unbinding is not described fully.

To account for the over-polarisation of glucose in water and the failure of glucose to unbind in the open state, we compute the glucose binding free energy difference from BPC to GLYCAM charge sets, and the glucose unbinding energy with the GLYCAM charge set, as shown in Tab. 7.6. In this way, the free energies of glucose unbinding accompanied by opening of the protein pocket were corrected, see Fig. 7.11 for the results.

We find a reaction free energy of 15.1 kcal/mol for the wild-type and 3.6 kcal/-mol for the triple mutant. Notably, in the *holo* open state *b*, when the glucose leaves the binding pocket, the glucose polarisation by the protein will grad-

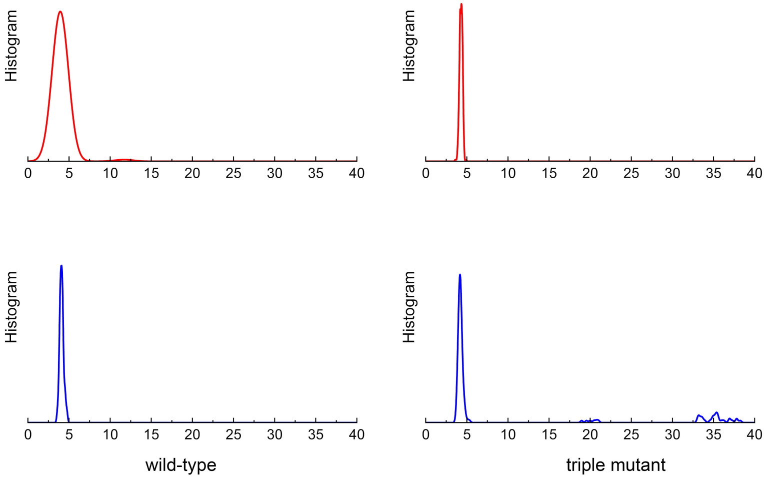


Figure 7.12.: Histogram of distance (Å) between glucose and binding pocket when sampling at state *a* (red) and state *b* (blue).

Table 7.6.: Glucose binding free energy difference and unbinding energy (kcal/mol).

	$\Delta\Delta G_{\text{binding}}$	$\Delta G_{\text{unbinding}}$
wild-type	0.5	5.2
triple mutant	0.2	2.6

ually decrease, and hence the glucose charges will reduce to the GLYCAM charges. This means, with GLYCAM charges in our free energy simulations series, the glucose unbinding energy is slightly underestimated when the glucose starts to leave. Therefore, the reaction free energy between *holo* closed state and *apo* open state may be slightly higher than obtained from free energy simulations series.

Note that the experimentally measured k_{D} may not only describe the process from *holo* close state to *apo* open state, rather it may also correspond to the processes from *holo* close state to *apo* close state or from *holo* open state to *apo* open state, indicated by the large error bars [106, 108, 130]. For both of the latter situations, the protein conformational changes are missing, leading to a smaller reaction free energy. Therefore, in these cases, the experimental ΔG will be smaller than our binding free energy results. Nevertheless, the agreement with experiment is remarkable, and the computed free energy differences are in line with the qualitative mechanistic picture emerging from the free simulations discussed above: the H-bonded network is destabilised in the mutant, leading to a much weaker binding of glucose, which qualitatively explains the difference in the experimentally reported k_D values.

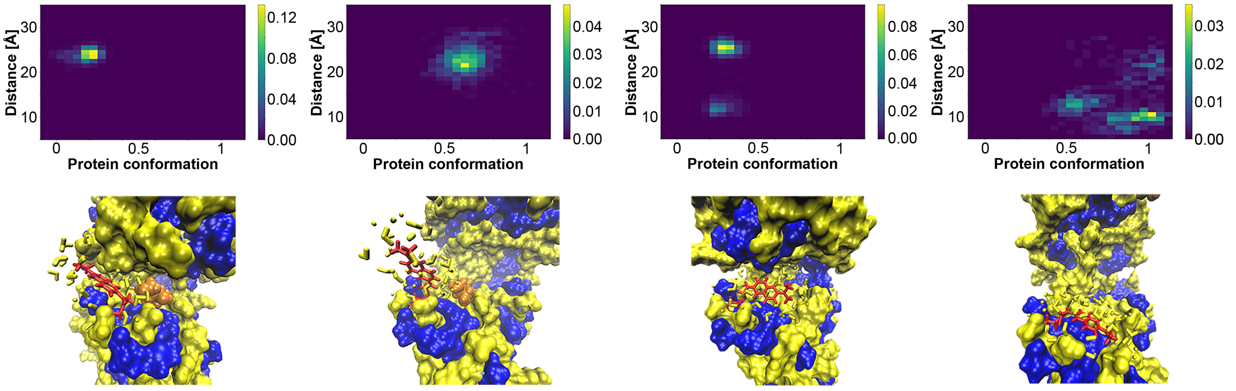
To resolve the problem altogether, an additional CV describing the glucose position is necessary. However, for that, either a 3D metadynamics simulation needs to be performed (which is computationally highly expensive) or the protein motion needs to be described by only one CV, which carries the risk of missing important regions in the conformational space or of losing the easy interpretability of the result by an abstract CV. Further, a change in glucose polarisation along the reaction coordinate would have to be considered in this case as well, which is a difficult task that would require an explicitly polarisable force field. Still, we believe that our estimates are sufficiently accurate to allow for an insight into the mechanisms, and in particular, to understand the differences between the wild-type and the mutant.

7.3.4. Conformations of Badan

Besides understanding the change in binding, another key aim of this study is the investigation of Badan conformations and their possible relation to its fluorescence properties. The conformational changes in the protein impact the properties of the excited states of Badan, and it was suggested that the dye resides inside a hydrophobic environment in the protein if and only if a glucose molecule is bound [108, 146]. Analysis of the MD trajectories allows to investigate the dynamics of the chromophore coupled to the protein conformational changes in more detail.

To this end, 2D histograms were obtained with one variable describing the protein conformation and the other variable describing the orientation of Badan by Monja Sokolov. The the relative frequencies of appearance of conformations of protein and Badan, together with representative conformations of Badan are presented in Fig. 7.13. These data are derived from the MD simulations of the GGBP triple mutant, see also Fig. 7.8. The binding of glucose correlates with the conformation of Badan clearly: Badan is outside the binding pocket when a glucose molecule is bound, being exposed to a more hydrophilic environment (Fig. 7.13A&B) For the protein *apo* state, Badan is mostly located inside the binding pocket when the binding pocket is open (Fig. 7.13D), in which case it is exposed to a probably more hydrophobic environment (Fig. 7.13C&D). For the *apo* closed state (Fig. 7.13C), the dye is found in- and outside of the binding pocket.

Therefore, the environment of Badan changes upon glucose binding clearly, and is more polarisable, which most probably is responsible for the increase of fluorescence observed experimentally. It was also observed that in the stable conformation of Badan folded inside the binding pocket in absence of glucose, the aromatic core of the dye and Trp183 are in close proximity. This points to another factor for the intensity increase upon glucose binding – the presence of Trp183 in the binding pocket, which is a known quencher of the fluorescence of Badan [147, 148].



A ***holo*-closed state** **B** ***holo*-open state** **C** ***apo*-closed state** **D** ***apo*-open state**

Figure 7.13.: 2D histogram of protein conformation and Badan orientation (created with Python 3.9) with correspond schematic presentations of Badan in hydrophobic/hydrophilic environment. In the normalised histograms, the x-axis represents the line connecting minima *a* and *b* from the metadynamics simulations. It is scaled such that 0 corresponds to *a* and 1 corresponds to *b*. The y-axis shows the distance between the top of Badan and Pro239. A distance smaller than 15 Å indicates that Badan is inside the binding pocket and a larger distance means that Badan is outside. Protein hydrophobic region–blue; Protein hydrophilic region–yellow; Cys-Badan–red; Glucose–orange. Water molecules within 4 Å away from the Cys-Badan molecule are described as yellow licorice shapes. Structure A to D were take from Fig. 7.8A-E,A-F,B-G and B-I respectively. The 2D histograms were done by Monja Sokolov

7.4. Conclusion and Outlook

In this work, we aimed at a detailed explanation of the mechanisms of glucose binding in GGBP using classical MD simulations and enhanced sampling techniques. A particular goal is to understand the changes upon mutation, which removes four H-bonds, leading to a drastic increase of k_D.

Up to now, no crystal structure of the triple mutant is available, therefore, our simulations uncover the molecular details of the mutant including Badan conformations for the first time. In particular, the simulations show how the conformation of Badan is coupled to the opening and closing of the binding pocket, and to the presence of glucose. So far, it had been assumed that the environment of Badan is more hydrophobic when the protein is in the closed conformation while binding the glucose [108]. Our simulations indicate that the opposite is true in fact: in the unbound state, Badan interacts with a Trp side chain, leading to the quenching of fluorescence.

Force field charges turned out to be critical parameters. The standard charge set, developed for bulk solvent, was unable to describe the binding situation in this highly polar and charged environment. Correction of charges by means of reparametrisation lead to a stable binding pocket as well as free energies of pocket opening in a very good agreement with experimental estimates. The series of free energy simulations designed in this work provided additional insight into the processes of interest taking place in such complex protein systems. Hence, such simulations appear capable of supporting the efforts of rational design of new glucose sensors.

The knowledge of the conformational dynamics of Badan allows for further work to investigate the changes of fluorescence of Badan in detail. That will involve excited-state QM/MM simulations using the semi-empirical TD-LC-DFTB method to describe the dynamics of Badan in its excited states.

In this work, for the glucose polarisation, the GPC as well as QM/MM glucose charge models were calculated by Monja Sokolov, Ziwei Pang computed the WPC and the GPC model, and conducted the evaluation of the repolarised charge sets. We declare that the simulation and analysis work of the classical MD simulations were carried out by both of us. The localisation of the Badan fluorophore was mainly done by Monja Sokolov, and Ziwei Pang only undertook the definition of the hydrophilic/hydrophobic environment in which the fluorophore was located.

8. Machine Learning on the 4-Aminophthalimide Fluorophore

8.1. Introduction

In principle, quantum chemical methods can quantitatively describe almost all the properties of various molecular systems. However, for molecular systems such as biomolecules, which contain a large number of atoms, the computational effort required to solve Schrödinger's equation becomes so large that even with large storage capacity and high computing speed tools, in practice, the solution process is still difficult and in many cases almost impossible. Although, with the development of the computational chemistry discipline, semi-empirical methods such as DFTB have emerged that can be used for QM calculations of biomolecular systems. However, as summarised in chapter 3, fast computational speed is often accompanied by a loss of accuracy.

Thanks to recent rapid developments in machine learning, we now have new means to model results of QM calculations on large systems. Current applications of machine learning methods to quantum chemical calculations include potential energy fitting, optimisation of the QM calculation accuracy, etc. The algorithms employed contain linear regression, statistical data analysis, artificial neural networks, etc. In this chapter, we applied the machine learning approach to learn the predictions of the LC-DFTB QM method to explore and optimise the excited state potential energy surfaces of 4-Aminophthalimide (4-AP) fluorescent molecules in different environments.

8.1.1. The 4-Aminophthalimide Fluorophore

The 4-AP system is a well known fluorescent probe molecule that is widely used in nanomaterials, microtissues and biological systems due to its fluorescence lifetime, quantum yield and sensitivity of the emission wavelength to the environment [149–151]. The fluorescence spectra of these molecules were reported to be significantly red-shifted in polar solvents compared to non-polar solvents, especially in aqueous solutions, suggesting a strong interaction between the molecules and the water [151].

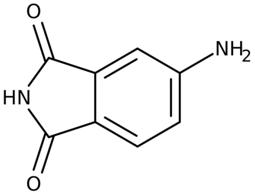


Figure 8.1.: Chemical structure of 4-Aminophthalimide.

Specifically, in non-polar solvents, 4-AP is a strongly fluorescent molecule. The fluorescence quantum yield of 4-AP changes slightly with increasing solvent polarity, remaining at 0.73 to 0.63. In polar solvents, the fluorescence quantum yield of 4-AP decreases rapidly due to hydrogen bonding between the excited state molecules and the solvent molecules, with a quantum yield of only 0.01 in water. In addition, the excited state of 4-AP has a long lifetime of about 14-15 ns in non-polar solvents, which decreases rapidly in polar solvents, to about 1 ns in water [152, 153]. As highly efficient fluorescent dyes, 4-AP and its derivatives have great potential to replace Badan in the design of novel GGBP-based blood glucose sensors due to their high sensitivity to the environment.

8.1.2. Motivation

In earlier studies, Monja Sokolov identified the problem of LC-DFTB inaccurately ordering excited states. Specifically, as shown in Table. 8.1, when calculating the first excited state of the 4-AP, the largest oscillator strength was achieved in the S_1 state using the DFT method with B3LYP functional or using the LC-DFT with CAM-B3LYP functional, whereas when applying LC-DFTB, the largest oscillator strength appeared in the S_3 state. Therefore, when calculating the S_1 state of 4-AP with LC-DFTB, we need to focus on the S_3 state rather than the S_1 state in the results.

Table 8.1.: Excited state energies (eV) of 4-AP with oscillator strengths (OS) optimised with LC-DFTB (pure electronical parameters) in vacuo computed with different methods.*

Time (ps)	LC-DFTB**		B3LYP		CAM-B3LYP	
	Energy	OS	Energy	OS	Energy	OS
S_1	3.78	0.0000	3.74	0.0660	4.16	0.0760
S_2	4.10	0.0002	3.77	0.0074	4.20	0.0018
S_3	4.14	0.0680	4.38	0.0000	4.72	0.0210

*done by Monja Sokolov. **with el.params. [154]

Besides, pilot simulations with the LC-DFTB method also showed another problem for the 4-AP in vacuo. With pure electronical parameters (el. parameters), for 4-AP in vacuo, it was able to follow the excited state during the QM/MM calculation. However, when the excitation energies were subsequently computed with the el. parameters on the same QM/MM trajectory, it can not follow the same state. Table. 8.2 illustrates changes in oscillator strengths corresponding to S_1, S_2 and S_3 states over simulation time from 7 fs to 11 fs. It is obvious that the LC-DFTB began to follow the S_2 state after 9 fs, i.e. the energy of the S_2 state after 9 fs is the energy of the S_3 state in the LC-DFTB results (the energy of the S_1 state in B3LYP/CAM-B3LYP results). Therefore, the energy of the excited state after this time needs to be revised. To this end, we corrected the S_3 energy with the S_2 energy when the S_2 state has the largest oscillator strength, and the corrected state (hereafter called as “target” state) is presented as an orange curve in Fig. 8.2. Note that this problem does not only occur when 4-AP is in vacuum, but also when 4-AP is in DMSO as well as in water, as detailed in the Appendix.A.

Such a crossing of excited states does not arise for commonly used QM approaches like DFT/B3LYP. However, since LC-DFTB is two to three orders of magnitude faster than DFT [155], and showed to be able to reproduce trends correctly [75], it is advantageous to optimise the LC-DFTB method.

To efficiently address the state-unfollowing issue and further speed up the calculations, we used artificial neural networks to learn the data sets from

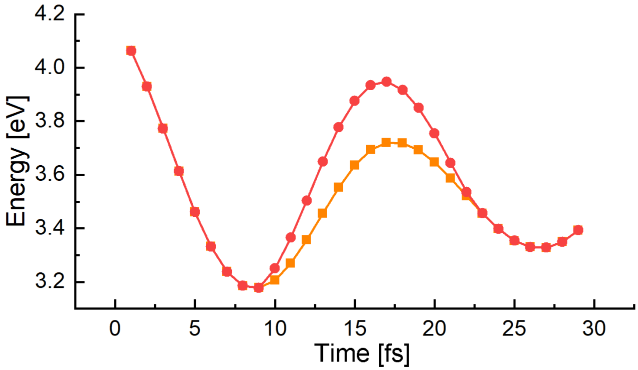


Figure 8.2.: The presence of energy level crossing of 4-AP in vacuo. The red curve indicates the third excited state (S_3) energy surface from the QM/MM calculation with LC-DFTB approach. The orange curve indicates the S_3 state after correction.

Table 8.2.: Excited state energies (eV) of 4-AP with oscillator strengths (OS) from LC-DFTB in vacuo.

Time (ps)	S_1		S_2		S_3	
	Energy	OS	Energy	OS	Energy	OS
7	2.932	0.0001	3.160	0.0000	3.238	0.0631
8	2.919	0.0002	3.139	0.0002	3.185	0.0629
9	2.955	0.0002	3.168	0.0145	3.178	0.0495
10	3.034	0.0005	3.205	0.0641	3.25	0.0014
11	3.144	0.0009	3.269	0.0667	3.365	0.0006

the pilot calculations and trained machines that can predict the excited state energies of 4-AP by feeding excited state structures.

8.2. Workflow

In pilot QM/MM simulations, 4-AP was calculated in vacuo, dimethyl sulfoxide (DMSO), and water. Due to the different state-ordering in LC-DFTB as mentioned, the focus on experimental S_1 state thus became a focus on

S_3 in vacuo, S_2 and S_3 in DMSO, and S_1 in water. All four simulation were performed for 1 ns producing 10,000 frames each with Gromacs 2021 [94] for the MM part and DFTB+ 20.1 [138] for the QM part. The repulsive OB2 (base) parameters for the LC-DFTB were taken from Ref. [154].

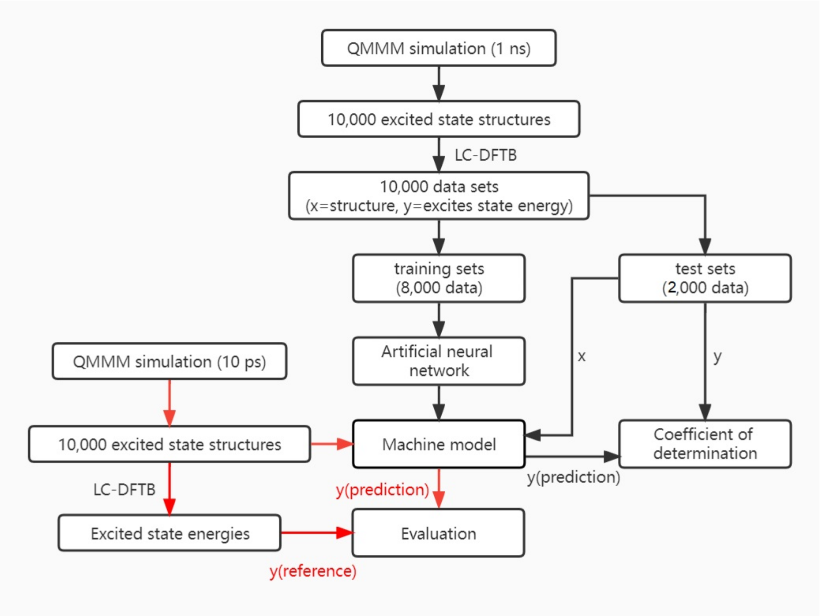


Figure 8.3.: Schematic view of the workflow. The black arrows represent the process of training machines. The red arrows represent the process of applying trained machines to predict excited state energies that can follow correct excited states.

Afterwards, for each case, excited state calculations were carried out for each of these 10,000 structures by LC-DFTB with el. parameters. Subsequently, we applied 8,000 LC-DFTB output data as training sets to train the machine (10% of which was used for validation) and 2,000 output data as test sets. The machine was coded by Li et al. [156], and corresponding scripts were created by Mila Krämer. The ANN initially contains in total four hidden layers, each with 30 neurons. The training stage had 1,000 epochs, and the batch size was set to 64.

To evaluate the machine model, four sets of QM/MM trajectories and obtained 10,000 structures for each model with a time interval of one femtosecond were additionally performed. Their corresponding excited states were calculated by LC-DFTB with el. parameters and the excitation energies with largest oscillator strength were selected as “target” excited state energies. Meanwhile, the same structures were sent to the trained machine model, and the excited state energies were produced as “predictions”.

8.3. Results and Discussion

In machine learning, the *coefficient of determination* (R^2) is commonly used to reflect the accuracy of a model by comparing the error of the evaluated model with the error of the *null* model as,

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y}_i)^2}. \tag{8.1}$$

Here, the error of the null model (denominator) indicates the mean of the reference data distribution. The better the model fit, the closer the R^2 value is to 1, the more accurate the model being evaluated. Conversely, the worse the model, the closer the R^2 value is to 0, or even negative. Table. 8.3 illustrates the coefficient of determination for the machine test sets. It is clear that both machines for the 4-AP in vacuo at S_3 state and DMSO at S_3 states were well-trained. Besides, compared to the former two machines, the one trained in DMSO at S_2 state has lower correlations. It is worth noting that the test sets of the machine that trained 4-AP in water at S_1 state showed a very low correlation, indicating a bad machine model in this case.

Table 8.3.: Coefficient of determination for machine test sets

Environment	State of interest	R^2_{test}
vacuo	S_3	0.9228
DMSO	S_2	0.8557
DMSO	S_3	0.9486
water	S_1	0.6657

To further evaluate the well-trained machine and investigate the possible reasons for producing bad machines, all four models were employed to predict corresponding excited state energies as follows.

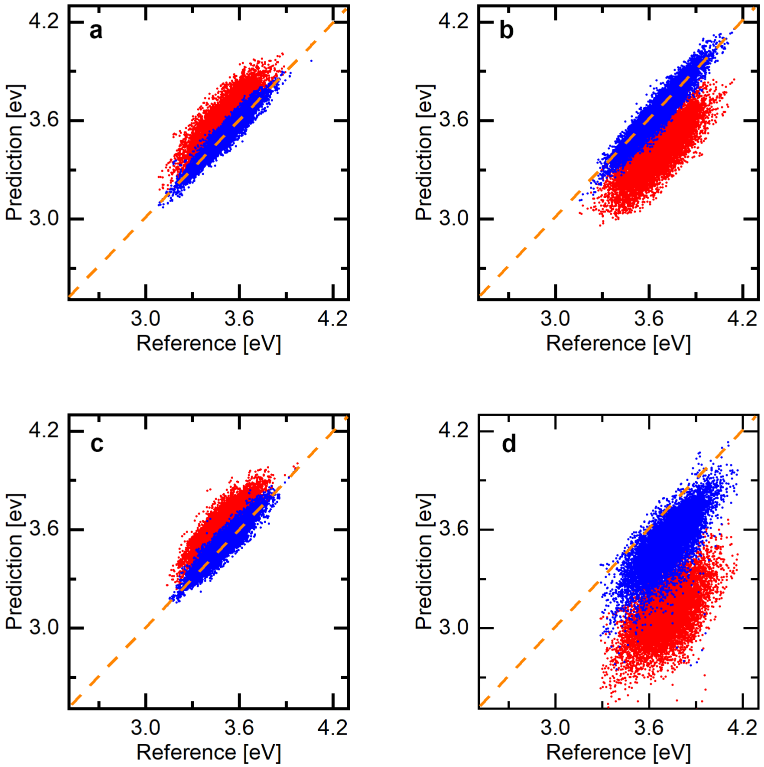


Figure 8.4.: Correlation diagrams between predictions from DFTB+ (red) as well as the trained machine (blue), and the reference "target" excited state energies. **a).** in vacuo, focusing on S_3 state. **b).** in DMSO, focusing on S_2 state. **c).** in DMSO, focusing on S_3 state. **d).** in water, focusing on S_1 state.

Correlations between excited state energy predictions from LC-DFTB and from the trained machine model as well as excited state energy references are shown in Fig. 8.4. Compared with the LC-DFTB results, all the excited states predicted by the machine model fit the reference excited states better. In particular, the machine models trained in vacuo and DMSO predict excited state energies very accurately, although the machine trained in DMSO for the S_2 state performed poorly in its test sets. Interestingly, as shown in

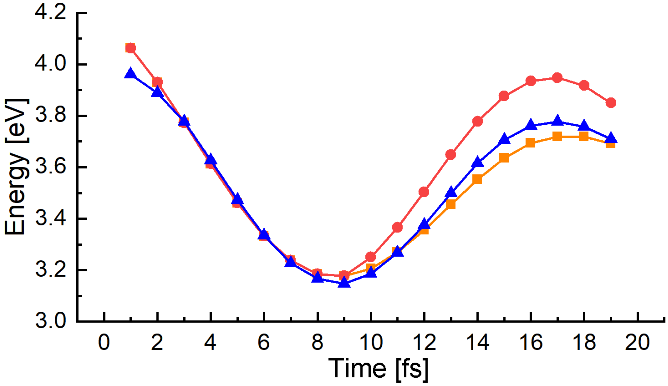


Figure 8.5.: Correction for the energy level crossing of 4-AP in vacuo (S_3). The red curve is from LC-DFTB calculations using DFTB+. The blue curve is prediction from the trained machine. The orange curve represents the "target" energy surface of the excited state and is from the DFTB+ output for states with the largest oscillation strength.

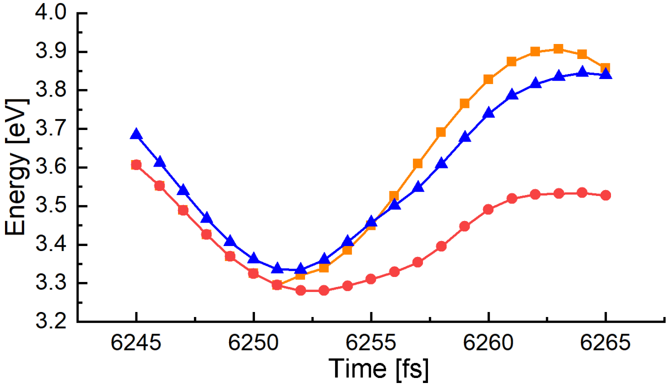


Figure 8.6.: Correction for the energy level crossing of 4-AP in DMSO (S_2). The red curve is from LC-DFTB calculations using DFTB+. The blue curve is prediction from the trained machine. The orange curve represents the "target" energy surface of the excited state and is from the DFTB+ output for states with the largest oscillation strength.

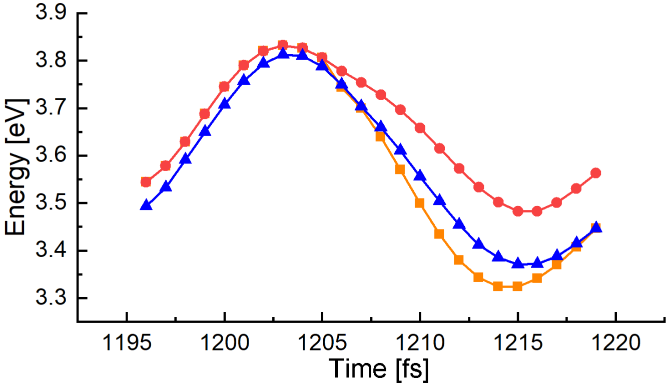


Figure 8.7.: Correction for the energy level crossing of 4-AP in DMSO (S_3). The red curve is from LC-DFTB calculations using DFTB+. The blue curve is prediction from the trained machine. The orange curve represents the “target” energy surface of the excited state and is from the DFTB+ output for states with the largest oscillation strength.

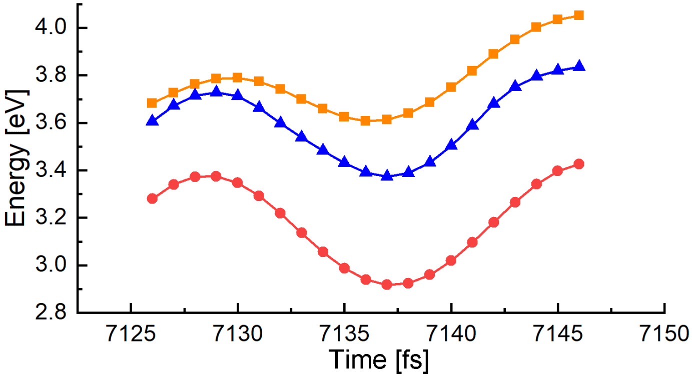


Figure 8.8.: Correction for the energy level crossing of 4-AP in water (S_1). The red curve is from LC-DFTB calculations using DFTB+. The blue curve is prediction from the trained machine. The orange curve represents the “target” energy surface of the excited state and is from the DFTB+ output for states with the largest oscillation strength.

Fig. 8.4d, despite the machine trained in water for the S_1 state has a bad accuracy, it is still possible to substantially correct energy deviations by LC-DFTB-calculations due to the “target” state unfollowing issue.

Fig. 8.5–8.8 illustrate applying different machine models to correct the excited state unfollowing problems generated by LC-DFTB in different environments. It was found that the machine correctly predicts the target state for all the structures although the LC-DFTB states cross. Note that, for the 4-AP in water, the 10 ps evaluation trajectory showed that the largest oscillator strengths for all frames were found for the S_2 and S_3 states, while we focused on the S_1 state in LC-DFTB calculations. That is, the calculation followed the wrong state during the whole simulation. Nevertheless, by tracking the excited state energy surface predicted by the machine model, as shown in Fig. 8.8, it can be inferred that the machine model can effectively fix the excited state unfollowing problem.

It is worthy to point out that for poor machine models trained in water, one possible reason is that underfitting has happened. As elaborated in section 8.1.1, the 4-AP can interact strongly with water, which leads to the presence of some substable 4-AP-water complexes (4-AP-$(H_2O)_{1,2}$) in aqueous solutions [157]. In these complexes, alternatives of H-bonds length play a key role in excited state energy changes. It has been reported that, the H-bond of 4-AP-$(H_2O)_2$ is shorter than that of 4-AP-H_2O, resulting in a significant red shift of the excited state [158]. However, in different 4-AP complexes, the 4-AP has only slight structural changes. For example, when H-bonds are formed as $C=O\cdots H-O$ between 4-AP and water, from 4-AP-H_2O to 4-AP-$(H_2O)_2$, the $C=O$ length changes only from 1.250 Å to 1.256 Å, whereas the $O\cdots H$ length changes from 1.828 Å to 1.779 Å, resulting in a red-shift of wavelength of 29 nm at excited state. Therefore, in this case, it is hard for machine to learn the excited state energies with only difference structures due to the solvent. To train a machine model comparable to those trained in vacuo and DMSO, either a super large training sets is required or solvent environment is included in the training sets.

Note that in our trained machine models, the excited state energies of the 4-AP were predicted by only inputting their structures. However, in practice, training machines that can predict excited state energies in different solvent environments requires additional inputs of environment parameters, such as the point charges used in Chapter 7. However, including the environment into ANN algorithms is still a challenge. For technical reasons, the input representing the environment should be as compact as possible. One possibility could be even to learn the difference between the excitation energies in vacuum and within the environment by a neural network only on the basis of the geometry of the dye and then add this contribution during a simulation. Another possibility could be to learn the excitation energies within the environment or the environmental shift by means of the electrostatic potential induced by the environment on the atoms of the dye as additional input to the machine. The work presented here represents the basis for such further development. It also demonstrated the feasibility of using the ANN algorithm to predict accurate excited states based on the molecular structure of 4-AP, which is still of considerable interest.

8.4. Assessment

In this work, we applied the ANN algorithm to successfully train a set of machines to correctly predict the target excited state of 4-AP based on LC-DFTB data. In contrast to LC-DFTB, the machine only knows the target state on which it was trained while in LC-DFTB calculations the state ordering changes during a simulation which makes it difficult to follow the target state. The models trained in vacuo as well as in DMSO solvents can accurately correct the excited states when such problem happens. A larger training data set is necessary for the machine training in water.

In addition, further studies could incorporate gradients as well as solvent point charges to achieve the production of continuous ML/MM trajectory with correct excited state description in different environments. This improves not only for the computational time needed by LC-DFTB, but also avoids the problem of crossing excited states.

9. Conclusion and Outlook

In this work, we investigated biomolecular systems using several multiscale modelling approaches. The study explored and revealed thermodynamic as well as kinetic related problems that can not be explained in laboratory-based experiments. Thus, this work has implications not only for the development of the discipline of computational chemistry, but also for the continuation of relevant laboratory experiments.

In Chapter 6, we explored the reaction mechanism of host-guest chemistry using classical MD simulations. In the study for [4+4] imine cages, we performed well-tempered MetaD to calculate the free energy surface of the uptake of ammonium guest ions with different sizes in 3-H, 3-Me and 3-Et imine cages, respectively. By comparing the activation energy of the guest uptake, we concluded that as the window size of the host cage became larger, the smaller the guest ion, the easier it was for the guest to leave the cage. This result was consistent with the laboratory experiments (the larger the size of the guest, the smaller the size of the cage and the longer it took for the guest to dissociate completely). Furthermore, we identified two possible mechanisms behind the ammonium ion uptake process. The first is a "solvent competition" mechanism, which often acts when the guest is located within the cage, and there is enough space in the cavity for the solvent molecules to enter, i.e. the guest volume is smaller, and the cage cavity is larger. In this case, more polar solvent molecules will crowd out the ammonium ions and thus gradually occupy the entire cage, eventually causing the guest to be squeezed out. The second mechanism occurs when the guest occupies a larger proportion of the cage cavity. In this case, the ammonium ions located inside the cage are pulled out due to the solvation effect of the surrounding solvent molecules outside the cage. In addition, we identified a cage deformation mechanism that often happens when the window size of the cage is small and the guest is large. In this case, when the guest crosses the window, the cage is locally deformed, thus inhibiting the process of the guest uptake.

In addition, the transfer of liquid nitrogen in [2+3] imine cages has been investigated using the FM method. The free energy surface shows that nitrogen molecules can transfer between hosts in two pathways. In the pathway that follows the “neighbour cage” mechanism, we demonstrated that the nitrogen molecules did not pass directly through the π-π stacking region but passed through the tunnels on the side. Besides, we found that the guest molecule did not easily pass through such pathway when the hydrogen atoms on the butyl substituent of the cage were replaced by fluorine atoms for F-cages and HF-cages. One possible reason is as such atomic substitution occurs, the cage side chain becomes larger and thus blocks the passage on either side of the π-π stacking region. In addition, we identified a pathway whereby the guest entered the cage gap in the crystal before entering other cages. Such process is called as the “void cage” mechanism. It is important to note that there is a “void cage” consisting of six side chains of neighbouring cages in the gap. In fact, we found that only the nitrogen molecule in the F-cage crystal entered this cage strictly before entering the other cages. However, in the HF- and H-cages, the guest tended to enter the gap region outside the “void cage”. We concluded that the reason for this phenomenon was similar to that of the “neighbour cage” mechanism, the blocking effect of substituted fluorine atoms.

In Chapter 7, we revealed the mechanism of a glucose binding protein-based fluorescence probe using QM calculations as well as MD simulations. Due to the high polarity of the protein binding pocket, we repolarised the glucose charges and solved the problem of the glucose unstable presence at the binding sites. In the following classical MD simulations, we obtained molecular structures similar to the crystal structure in the *holo*-closed state, and also, for the first time, we modelled the GGBP-Badan triple mutant and successfully predicted its multiple metastable structures. Subsequently, we performed well-tempered MetaD, alchemical calculation, and FM approaches for wild-type GGBP and GGBP triple mutant to calculate the glucose unbinding process in steps and obtained slightly higher free energies than those obtained in the experiments. We point out that the main reason for the discrepancy is the failure of the experiment to focus purely on the change of the protein from its *holo* closed state to its *apo* open state. Also, the solvents used in the experiments differ from the simulations, and the different solvation effects affect the glucose unbinding energy as well as the protein conformation. Finally, we tentatively explored the position of the Badan fluorophore for different states of the GGBP triple mutant. We found that when glucose is in the bound state, the fluorophore tends to be located outside the binding pocket, whereas when glucose is absent, the fluorophore prefers to be enveloped inside the binding pocket. This finding contradicts previous hypotheses, and we conclude that the fluorescence quenching seen in the experiments may be due to tryptophane in the binding pocket. To conclude, in this work, we have successfully explained the mechanism of this fluorescent probe. However, further research is still required to optimise this type of fluorescent probe (e.g. the position of the fluorophore, or the use of alternative fluorophores, etc.), which may be effectively implemented in the future through machine learning. Nevertheless, our work is a milestone in theoretically setting a precedent for studying such fluorescent probes.

In Chapter 8, we investigated the conformational dynamics of the 4-AP excited state using a machine learning with ANN algorithm. We found that, with the same training set volume, machine models trained in vacuo and DMSO solvent have good accuracy, while those trained in aqueous solutions have poor accuracy. We then performed evaluations on the four machine models, and the results showed that all machines could solve the “target” excited states unfollowing issue caused by LC-DFTB. Similar to their performance in corresponding test sets, the machines trained in vacuo and DMSO can give convincing results, while the machines trained in aqueous solution can not accurately predict the excited states. We deduce that one possible reason is that the strong hydrogen bonding interaction of 4-AP with water molecules in an aqueous solution makes 4-AP presents as 4-AP-water complexes, leading to excited state energies more related to the H-bond length than the 4-AP structure. Thus, a super large training sets or a training sets includes solvent environment is necessary to improve the accuracy of the machine. Nevertheless, as a preliminary attempt, the present work successfully verified the feasibility of using ANN algorithm to follow and calculate excited states accurately. Furthermore, energy gradients as well as solvent environment could be introduced in future machines to obtain continuous trajectories of molecular excited states with solvent effects.

In summary, the above three aspects of this work form a blueprint for studying biomolecular systems from molecular mechanics to quantum mechanics. At the molecular mechanics level, we attempted enhanced sampling methods based on MD simulations to reveal receptor-ligand interactions at larger time scales and in larger systems; at the quantum mechanics level, we applied a combination of machine learning and semi-empirical LC-DFTB to ensure computational accuracy while being 2-3 orders of magnitude faster than classical DFT calculations. Together, the two have theoretical implications from a broad perspective, for instance, the design of fluorescent probes, targeting drugs and so on.

Bibliography

1. Fischer, E. Influence of the Configuration on the Action of the Enzymes. *Reports of the German Chemical Society* **27,** 2985–2993 (1894).
2. Koshland Jr, D. Application of a theory of enzyme specificity to protein synthesis. *Proceedings of the National Academy of Sciences of the United States of America* **44,** 98 (1958).
3. Boehr, D. D., Nussinov, R. & Wright, P. E. The role of dynamic conformational ensembles in biomolecular recognition. *Nature chemical biology* **5,** 789–796 (2009).
4. Csermely, P., Palotai, R. & Nussinov, R. Induced fit, conformational selection and independent dynamic segments: an extended view of binding events. *Nature Precedings,* 1–1 (2010).
5. Acuña, A. & Amat-Guerri, F. in *Fluorescence of supermolecules, polymers, and nanosystems* 3–20 (Springer, 2007).
6. Briggs, E. A., Besley, N. A. & Robinson, D. QM/MM excited state molecular dynamics and fluorescence spectroscopy of BODIPY. *The Journal of Physical Chemistry A* **117,** 2644–2650 (2013).
7. Lázaro, E., San Andrés, M. & Vera, S. Determination of five polycyclic aromatic hydrocarbons in aqueous micellar media by fluorescence at room temperature. *Analytica Chimica Acta* **413,** 159–166 (2000).
8. Pulgarın, J. M. & Bermejo, L. G. Determination of napropamide in technical formulations, soil and vegetable samples by sensitised fluorescence: validation of the method. *Analytica chimica acta* **491,** 37–45 (2003).
9. Lakowicz, J. R. *Principles of fluorescence spectroscopy* (Springer, 2006).
10. De Silva, A. P. *et al.* Signaling recognition events with fluorescent sensors and switches. *Chemical reviews* **97,** 1515–1566 (1997).
11. Liu, Z., He, W. & Guo, Z. Metal coordination in photoluminescent sensing. *Chemical Society Reviews* **42,** 1568–1600 (2013).

12. Gee, K. R., Zhou, Z.-L., Qian, W.-J. & Kennedy, R. Detection and imaging of zinc secretion from pancreatic β-cells using a new fluorescent zinc indicator. *Journal of the American Chemical Society* **124,** 776–778 (2002).

13. Boens, N., Leen, V. & Dehaen, W. Fluorescent indicators based on BODIPY. *Chemical Society Reviews* **41,** 1130–1172 (2012).

14. Steinmeyer, J., Rönicke, F., Schepers, U. & Wagenknecht, H.-A. Synthesis of Wavelength-Shifting Fluorescent DNA and RNA with Two Photostable Cyanine–Styryl Dyes as the Base Surrogate Pair. *ChemistryOpen* **6,** 514–518 (2017).

15. Zhou, J., Liu, Z. & Li, F. Upconversion nanophosphors for small-animal imaging. *Chemical Society Reviews* **41,** 1323–1349 (2012).

16. Fan, L.-J. & Jones, W. E. A highly selective and sensitive inorganic/organic hybrid polymer fluorescence "turn-on" chemosensory system for iron cations. *Journal of the American Chemical Society* **128,** 6784–6785 (2006).

17. Chalfie, M., Tu, Y., Euskirchen, G., Ward, W. W. & Prasher, D. C. Green fluorescent protein as a marker for gene expression. *Science* **263,** 802–805 (1994).

18. Grabowski, Z. R., Rotkiewicz, K. & Rettig, W. Structural changes accompanying intramolecular electron transfer: focus on twisted intramolecular charge-transfer states and structures. *Chemical reviews* **103,** 3899–4032 (2003).

19. Wang, H., Zhang, B. W. & Cao, Y. Intramolecular charge transfer and exciplex formation in anthracene bichromophoric compounds. *Journal of Photochemistry and Photobiology A: Chemistry* **92,** 29–34 (1995).

20. Wu, J., Liu, W., Ge, J., Zhang, H. & Wang, P. New sensing mechanisms for design of fluorescent chemosensors emerging in recent years. *Chemical Society Reviews* **40,** 3483–3495 (2011).

21. Olsen, S. Locally-excited (LE) versus charge-transfer (CT) excited state competition in a series of para-substituted neutral green fluorescent protein (GFP) chromophore models. *The Journal of Physical Chemistry B* **119,** 2566–2575 (2015).

22. Hartree, D. R. *The wave mechanics of an atom with a non-coulomb central field. Part II. Some results and discussion* in *Mathematical Proceedings of the Cambridge Philosophical Society* **24** (1928), 111–132.

23. Slater, J. C. The theory of complex spectra. *Physical Review* **34,** 1293 (1929).

24. Leach, A. R. & Leach, A. R. *Molecular modelling: principles and applications* (Pearson education, 2001).

25. Hohenberg, P. & Kohn, W. Inhomogeneous electron gas. *Physical review* **136,** B864 (1964).

26. Kohn, W. & Sham, L. J. Self-consistent equations including exchange and correlation effects. *Physical review* **140,** A1133 (1965).

27. Wang, H. & Zhang, Y. Density-functional theory for the spin-1 bosons in a one-dimensional harmonic trap. *Physical Review A* **88,** 023626 (2013).

28. Hu, Y., Murthy, G., Rao, S. & Jain, J. Kohn-Sham density functional theory of Abelian anyons. *Physical Review B* **103,** 035124 (2021).

29. Perdew, J. P. & Schmidt, K. *Jacob's ladder of density functional approximations for the exchange-correlation energy* in *AIP Conference Proceedings* **577** (2001), 1–20.

30. Becke, A. D. Density-functional exchange-energy approximation with correct asymptotic behavior. *Physical review A* **38,** 3098 (1988).

31. Perdew, J. P. *et al.* Atoms, molecules, solids, and surfaces: Applications of the generalized gradient approximation for exchange and correlation. *Physical review B* **46,** 6671 (1992).

32. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Physical review letters* **77,** 3865 (1996).

33. Lee, C., Yang, W. & Parr, R. G. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Physical review B* **37,** 785 (1988).

34. Burke, K., Perdew, J. P. & Ernzerhof, M. Why semilocal functionals work: Accuracy of the on-top pair density and importance of system averaging. *The Journal of chemical physics* **109,** 3760–3771 (1998).

35. Becke, A. D. A new mixing of Hartree–Fock and local density-functional theories. *The Journal of chemical physics* **98,** 1372–1377 (1993).

36. Stephens, P. J., Devlin, F. J., Chabalowski, C. F. & Frisch, M. J. Ab initio calculation of vibrational absorption and circular dichroism spectra using density functional force fields. *The Journal of physical chemistry* **98,** 11623–11627 (1994).

37. Vosko, S. H., Wilk, L. & Nusair, M. Accurate spin-dependent electron liquid correlation energies for local spin density calculations: a critical analysis. *Canadian Journal of physics* **58,** 1200–1211 (1980).

38. Adamo, C. & Barone, V. Toward reliable density functional methods without adjustable parameters: The PBE0 model. *The Journal of chemical physics* **110,** 6158–6170 (1999).

39. Heyd, J., Scuseria, G. E. & Ernzerhof, M. Hybrid functionals based on a screened Coulomb potential. *The Journal of chemical physics* **118,** 8207–8215 (2003).

40. Pople, J. A. & Segal, G. A. Approximate self-consistent molecular orbital theory. III. CNDO results for AB2 and AB3 systems. *The Journal of Chemical Physics* **44,** 3289–3296 (1966).

41. Dewar, M. J. & Thiel, W. Ground states of molecules. 38. The MNDO method. Approximations and parameters. *Journal of the American Chemical Society* **99,** 4899–4907 (1977).

42. Stewart, J. J. Optimization of parameters for semiempirical methods I. Method. *Journal of computational chemistry* **10,** 209–220 (1989).

43. Elstner, M. The SCC-DFTB method and its application to biological systems. *Theoretical Chemistry Accounts* **116,** 316–325 (2006).

44. Gaus, M., Cui, Q. & Elstner, M. DFTB3: extension of the self-consistent-charge density-functional tight-binding method (SCC-DFTB). *Journal of chemical theory and computation* **7,** 931–948 (2011).

45. Elstner, M. *et al.* Self-consistent-charge density-functional tight-binding method for simulations of complex materials properties. *Physical Review B* **58,** 7260 (1998).

46. Brooks, B. R. *et al.* CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *Journal of computational chemistry* **4,** 187–217 (1983).

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

48. Allinger, N. L. Conformational analysis. 130. MM2. A hydrocarbon force field utilizing V1 and V2 torsional terms. *Journal of the American Chemical Society* **99,** 8127–8134 (1977).

49. Allinger, N. L., Yuh, Y. H. & Lii, J. H. Molecular mechanics. The MM3 force field for hydrocarbons. 1. *Journal of the American Chemical Society* **111,** 8551–8566 (1989).

50. Allinger, N. L., Chen, K. & Lii, J.-H. An improved force field (MM4) for saturated hydrocarbons. *Journal of computational chemistry* **17,** 642–668 (1996).

51. Lifson, S. & Warshel, A. Consistent force field for calculations of conformations, vibrational spectra, and enthalpies of cycloalkane and n-alkane molecules. *The Journal of Chemical Physics* **49,** 5116–5129 (1968).

52. Weiner, S. J., Kollman, P. A., Nguyen, D. T. & Case, D. A. An all atom force field for simulations of proteins and nucleic acids. *Journal of computational chemistry* **7,** 230–252 (1986).

53. Weiner, S. J. *et al.* A new force field for molecular mechanical simulation of nucleic acids and proteins. *Journal of the American Chemical Society* **106,** 765–784 (1984).

54. MacKerell Jr, A. D. *et al.* All-atom empirical potential for molecular modeling and dynamics studies of proteins. *The journal of physical chemistry B* **102,** 3586–3616 (1998).

55. Hagler, A., Huler, E. & Lifson, S. Energy functions for peptides and proteins. I. Derivation of a consistent force field including the hydrogen bond from amide crystals. *Journal of the American Chemical Society* **96,** 5319–5327 (1974).

56. Lifson, S., Hagler, A. & Dauber, P. Consistent force field studies of intermolecular forces in hydrogen-bonded crystals. 1. Carboxylic acids, amides, and the C: O. cntdot.. cntdot.. cntdot. H-hydrogen bonds. *Journal of the American Chemical Society* **101,** 5111–5121 (1979).

57. Stocker, U. & van Gunsteren, W. F. Molecular dynamics simulation of hen egg white lysozyme: a test of the GROMOS96 force field against nuclear magnetic resonance data. *Proteins: Structure, Function, and Bioinformatics* **40,** 145–153 (2000).

58. Oostenbrink, C., Villa, A., Mark, A. E. & Van Gunsteren, W. F. A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force-field parameter sets 53A5 and 53A6. *Journal of computational chemistry* **25,** 1656–1676 (2004).

59. Jorgensen, W. L. & Tirado-Rives, J. The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *Journal of the American Chemical Society* **110,** 1657–1666 (1988).

60. Jorgensen, W. L., Maxwell, D. S. & Tirado-Rives, J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *Journal of the American Chemical Society* **118,** 11225–11236 (1996).

61. Verlet, L. Computer" experiments" on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules. *Physical review* **159,** 98 (1967).

62. Van Gunsteren, W. F. & Berendsen, H. J. A leap-frog algorithm for stochastic dynamics. *Molecular Simulation* **1,** 173–185 (1988).

63. Di Pierro, M., Elber, R. & Leimkuhler, B. A stochastic algorithm for the isobaric–isothermal ensemble with Ewald summations for all long range forces. *Journal of chemical theory and computation* **11,** 5624–5637 (2015).

64. Berendsen, H. J., Postma, J. v., Van Gunsteren, W. F., DiNola, A. & Haak, J. R. Molecular dynamics with coupling to an external bath. *The Journal of chemical physics* **81,** 3684–3690 (1984).

65. Evans, D. J. & Holian, B. L. The nose–hoover thermostat. *The Journal of chemical physics* **83,** 4069–4074 (1985).

66. Bussi, G., Donadio, D. & Parrinello, M. Canonical sampling through velocity rescaling. *The Journal of chemical physics* **126,** 014101 (2007).

67. Parrinello, M. & Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied physics* **52,** 7182–7190 (1981).

68. Torrie, G. M. & Valleau, J. P. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics* **23,** 187–199 (1977).

69. Laio, A. & Parrinello, M. Escaping free-energy minima. *Proceedings of the National Academy of Sciences* **99,** 12562–12566 (2002).

70. Sugita, Y. & Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chemical physics letters* **314,** 141–151 (1999).

71. Barducci, A., Bussi, G. & Parrinello, M. Well-tempered metadynamics: a smoothly converging and tunable free-energy method. *Physical review letters* **100,** 020603 (2008).

72. Limongelli, V., Bonomi, M. & Parrinello, M. Funnel metadynamics as accurate binding free-energy method. *Proceedings of the National Academy of Sciences* **110,** 6358–6363 (2013).

73. Raniolo, S. & Limongelli, V. Ligand binding free-energy calculations with funnel metadynamics. *Nature Protocols* **15,** 2837–2866 (2020).

74. Mey, A. S. *et al.* Best practices for alchemical free energy calculations [Article v1. 0]. *Living journal of computational molecular science* **2** (2020).

75. Sokolov, M. *et al.* Analytical Time-Dependent Long-Range Corrected Density Functional Tight Binding (TD-LC-DFTB) Gradients in DFTB+: Implementation and Benchmark for Excited-State Geometries and Transition Energies. *Journal of Chemical Theory and Computation* **17,** 2266–2282 (2021).

76. Schütt, K. T., Gastegger, M., Tkatchenko, A., Müller, K.-R. & Maurer, R. J. Unifying machine learning and quantum chemistry with a deep neural network for molecular wavefunctions. *Nature communications* **10,** 1–10 (2019).

77. Gómez-Flores, C. L. *et al.* Accurate Free Energies for Complex Condensed-Phase Reactions Using an Artificial Neural Network Corrected DFT-B/MM Methodology. *Journal of Chemical Theory and Computation* (2022).

78. Dral, P. O. Quantum chemistry in the age of machine learning. *The journal of physical chemistry letters* **11,** 2336–2347 (2020).

79. Prechelt, L. in *Neural Networks: Tricks of the trade* 55–69 (Springer, 1998).

80. Shorten, C. & Khoshgoftaar, T. M. A survey on image data augmentation for deep learning. *Journal of big data* **6,** 1–48 (2019).

81. Wan, L., Zeiler, M., Zhang, S., Le Cun, Y. & Fergus, R. *Regularization of neural networks using dropconnect* in *International conference on machine learning* (2013), 1058–1066.

82. Géron, A. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems* (" O'Reilly Media, Inc.", 2019).

83. Pedersen, C. J. Cyclic polyethers and their complexes with metal salts. *Journal of the American Chemical Society* **89,** 7017–7036 (1967).

84. Pedersen, C. J. The discovery of crown ethers (Noble Lecture). *Angewandte Chemie International Edition in English* **27,** 1021–1027 (1988).

85. Dietrich, B., Lehn, J. & Sauvage, J. Diaza-polyoxa-macrocycles et macrobicycles. *Tetrahedron Letters* **10,** 2885–2888 (1969).

86. Späth, A. & König, B. Molecular recognition of organic ammonium ions in solution using synthetic receptors. *Beilstein journal of organic chemistry* **6,** 32 (2010).

87. Lauer, J. C., Zhang, W.-S., Rominger, F., Schröder, R. R. & Mastalerz, M. Shape-Persistent [4+ 4] Imine Cages with a Truncated Tetrahedral Geometry. *Chemistry–A European Journal* **24,** 1816–1820 (2018).

88. Lauer, J. C. *et al.* Host-Guest Chemistry of Truncated Tetrahedral Imine Cages with Ammonium Ions. *ChemistryOpen* **9,** 183–190 (2020).

89. Case, D. A. *et al.* AMBER 2015 (2015).

90. Wang, J., Wang, W., Kollman, P. A. & Case, D. A. Automatic atom type and bond type perception in molecular mechanical calculations. *Journal of molecular graphics and modelling* **25,** 247–260 (2006).

91. Frisch, M. J. *et al. Gaussian 09, Revision A.02* 2009.

92. Van Der Spoel, D. *et al.* GROMACS: fast, flexible, and free. *Journal of computational chemistry* **26,** 1701–1718 (2005).

93. Hess, B., Kutzner, C., Van Der Spoel, D. & Lindahl, E. GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. *Journal of chemical theory and computation* **4,** 435–447 (2008).

94. Abraham, M. J. *et al.* GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1,** 19–25 (2015).

95. Tribello, G. A., Bonomi, M., Branduardi, D., Camilloni, C. & Bussi, G. PLUMED 2: New feathers for an old bird. *Computer Physics Communications* **185,** 604–613 (2014).

96. Humphrey, W., Dalke, A. & Schulten, K. VMD: visual molecular dynamics. *Journal of molecular graphics* **14,** 33–38 (1996).

97. Houk, K., Nakamura, K., Sheu, C. & Keating, A. E. Gating as a control element in constrictive binding and guest release by hemicarcerands. *Science* **273,** 627–629 (1996).

98. Ro, S., Rowan, S. J., Pease, A. R., Cram, D. J. & Stoddart, J. F. Dynamic hemicarcerands and hemicarceplexes. *Organic letters* **2,** 2411–2414 (2000).

99. Davis, A. V. & Raymond, K. N. The big squeeze: Guest exchange in an M4L6 supramolecular host. *Journal of the American Chemical Society* **127,** 7912–7919 (2005).

100. Vyas, N. K., Vyas, M. N. & Quiocho, F. A. Sugar and signal-transducer binding sites of the Escherichia coli galactose chemoreceptor protein. *Science* **242,** 1290–1295 (1988).

101. Mowbray, S. L., Smith, R. & Cole, L. Structure of the periplasmic glucose/galactose receptor of Salmonella typhimurium. *Receptor* **1,** 41–53 (1990).

102. Salins, L. L., Ware, R. A., Ensor, C. M. & Daunert, S. A novel reagentless sensing system for measuring glucose based on the galactose/glucose-binding protein. *Analytical Biochemistry* **294,** 19–26 (2001).

103. Marvin, J. S. & Hellinga, H. W. Engineering biosensors by introducing fluorescent allosteric signal transducers: construction of a novel glucose sensor. *Journal of the American chemical society* **120,** 7–11 (1998).

104. De Lorimier, R. M. *et al.* Construction of a fluorescent biosensor family. *Protein Science* **11,** 2655–2675 (2002).

105. Scognamiglio, V. *et al.* D-galactose/D-glucose-binding Protein from Escherichia coli as Probe for a Non-consuming Glucose Implantable Fluorescence Biosensor. *Sensors* **7,** 2484–2491 (2007).

106. Khan, F., Gnudi, L. & Pickup, J. C. Fluorescence-based sensing of glucose using engineered glucose/galactose-binding protein: a comparison of fluorescence resonance energy transfer and environmentally sensitive dye labelling strategies. *Biochemical and biophysical research communications* **365,** 102–106 (2008).

107. Ge, X., Rao, G. & Tolosa, L. On the possibility of real-time monitoring of glucose in cell culture by microdialysis using a fluorescent glucose binding protein sensor. *Biotechnology progress* **24,** 691–697 (2008).

108. Khan, F., Saxl, T. E. & Pickup, J. C. Fluorescence intensity-and lifetime-based glucose sensing using an engineered high-Kd mutant of glucose/galactose-binding protein. *Analytical biochemistry* **399,** 39–43 (2010).

109. Saxl, T., Khan, F., Ferla, M., Birch, D. & Pickup, J. A fluorescence lifetime-based fibre-optic glucose sensor using glucose/galactose-binding protein. *Analyst* **136,** 968–972 (2011).

110. Tiangco, C. *et al.* Fiber optic biosensor for transdermal glucose based on the glucose binding protein. *Sensors and Actuators B: Chemical* **242,** 569–576 (2017).

111. Helassa, N. *et al.* A novel fluorescent sensor protein for detecting changes in airway surface liquid glucose concentration. *Biochemical Journal* **464,** 213–220 (2014).

112. Adhikary, R., Barnes, C. A. & Petrich, J. W. Solvation dynamics of the fluorescent probe PRODAN in heterogeneous environments: contributions from the locally excited and charge-transferred states. *The Journal of Physical Chemistry B* **113,** 11999–12004 (2009).

113. Weber, G. & Farris, F. J. Synthesis and spectral properties of a hydrophobic fluorescent probe: 6-propionyl-2-(dimethylamino) naphthalene. *Biochemistry* **18,** 3075–3078 (1979).

114. Unione, L. *et al.* Unraveling the conformational landscape of ligand binding to glucose/galactose-binding protein by paramagnetic NMR and MD simulations. *ACS chemical biology* **11,** 2149–2157 (2016).

115. Panjaitan, B. M., Kubiak-Ossowska, K., Birch, D. & Chen, Y. *Study of Glucose Binding Protein Encapsulated Gold Nanoclusters by Molecular Dynamic Simulation* in *Materials Science Forum* **948** (2019), 133–139.

116. Basma, M., Sundara, S., Çalgan, D., Vernali, T. & Woods, R. J. Solvated ensemble averaging in the calculation of partial atomic charges. *Journal of computational chemistry* **22,** 1125–1137 (2001).

117. Kirschner, K. N. *et al.* GLYCAM06: a generalizable biomolecular force field. Carbohydrates. *Journal of computational chemistry* **29,** 622–655 (2008).

118. Chang, L., Ishikawa, T., Kuwata, K. & Takada, S. Protein-specific force field derived from the fragment molecular orbital method can improve protein–ligand binding interactions. *Journal of Computational Chemistry* **34,** 1251–1257 (2013).

119. Liu, J., He, X. & Zhang, J. Z. Improving the scoring of protein–ligand binding affinity by including the effects of structural water and electronic polarization. *Journal of chemical information and modeling* **53,** 1306–1314 (2013).

120. Duan, L., Feng, G., Wang, X., Wang, L. & Zhang, Q. Effect of electrostatic polarization and bridging water on CDK2–ligand binding affinities calculated using a highly efficient interaction entropy method. *Physical Chemistry Chemical Physics* **19,** 10140–10152 (2017).

121. Zhong, S. *et al.* Binding Mechanism of Thrombin–Ligand Systems Investigated by a Polarized Protein-Specific Charge Force Field and Interaction Entropy Method. *The Journal of Physical Chemistry B* **123,** 8704–8716 (2019).

122. Wei, C., Tung, D., Yip, Y. M., Mei, Y. & Zhang, D. Communication: The electrostatic polarization is essential to differentiate the helical propensity in polyalanine mutants. *The Journal of chemical physics* **134,** 171101 (2011).

123. Mei, Y. *et al.* Folding and thermodynamic studies of Trp-cage based on polarized force field. *Theoretical Chemistry Accounts* **131,** 1–7 (2012).

124. Jia, X., Mei, Y., Zhang, J. Z. & Mo, Y. Hybrid QM/MM study of FMO complex with polarized protein-specific charge. *Scientific reports* **5,** 1–10 (2015).

125. Tong, Z., Huai, Z., Mei, Y. & Mo, Y. Influence of the Protein Environment on the Electronic Excitation of Chromophores in the Phycoerythrin 545 Light–Harvesting Complex: A Combined MD-QM/MM Method with Polarized Protein–Specific Charge Scheme. *The Journal of Physical Chemistry B* **123,** 2040–2049 (2019).

126. Xu, Z., Lazim, R., Mei, Y. & Zhang, D. Stability of the β-structure in prion protein: A molecular dynamics study based on polarized force field. *Chemical Physics Letters* **539,** 239–244 (2012).

127. Bayly, C. I., Cieplak, P., Cornell, W. & Kollman, P. A. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *The Journal of Physical Chemistry* **97,** 10269–10280 (1993).

128. Zeng, J., Duan, L., Zhang, J. Z. & Mei, Y. A numerically stable restrained electrostatic potential charge fitting method. *Journal of computational chemistry* **34,** 847–853 (2013).

129. Duan, R., Lazim, R. & Zhang, D. Understanding the basis of I50V-induced affinity decrease in HIV-1 protease via molecular dynamics simulations using polarized force field. *Journal of computational chemistry* **36,** 1885–1892 (2015).

130. Borrok, M. J., Kiessling, L. L. & Forest, K. T. Conformational changes of glucose/galactose-binding protein illuminated by open, unliganded, and ultra-high-resolution ligand-bound structures. *Protein science* **16,** 1032–1041 (2007).

131. Case, D. A. *et al. AmberTools 2016* University of California, San Francisco. 2016.

132. Lindorff-Larsen, K. *et al.* Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins: Structure, Function, and Bioinformatics* **78,** 1950–1958 (2010).

133. Joung, I. S. & Cheatham III, T. E. Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations. *The journal of physical chemistry B* **112,** 9020–9041 (2008).

134. Cossi, M., Rega, N., Scalmani, G. & Barone, V. Energies, structures, and electronic properties of molecules in solution with the C-PCM solvation model. *Journal of computational chemistry* **24,** 669–681 (2003).

135. Pronk, S. *et al.* GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* **29,** 845–854 (2013).

136. Gaus, M., Goez, A. & Elstner, M. Parametrization and benchmark of DFTB3 for organic molecules. *Journal of Chemical Theory and Computation* **9,** 338–354 (2013).

137. Aradi, B., Hourahine, B. & Frauenheim, T. DFTB+, a sparse matrix-based implementation of the DFTB method. *The Journal of Physical Chemistry A* **111,** 5678–5684 (2007).

138. Hourahine, B. *et al.* DFTB+, a software package for efficient approximate density functional theory based atomistic simulations. *The Journal of chemical physics* **152,** 124101 (2020).

139. Darden, T., York, D. & Pedersen, L. Particle mesh Ewald: An N log (N) method for Ewald sums in large systems. *The Journal of chemical physics* **98,** 10089–10092 (1993).

140. Nosé, S. A molecular dynamics method for simulations in the canonical ensemble. *Molecular physics* **52,** 255–268 (1984).

141. Hoover, W. G. Canonical dynamics: Equilibrium phase-space distributions. *Physical review A* **31,** 1695 (1985).

142. Nosé, S. & Klein, M. Constant pressure molecular dynamics for molecular systems. *Molecular Physics* **50,** 1055–1076 (1983).

143. Ortega, G., Castaño, D., Diercks, T. & Millet, O. Carbohydrate affinity for the glucose–galactose binding protein is regulated by allosteric domain motions. *Journal of the American Chemical Society* **134,** 19869–19876 (2012).

144. Bonomi, M. Promoting transparency and reproducibility in enhanced molecular simulations. *Nature methods* **16,** 670–673 (2019).

145. Flocco, M. & Mowbray, S. L. The 1.9 A x-ray structure of a closed unliganded form of the periplasmic glucose/galactose receptor from Salmonella typhimurium. *Journal of Biological Chemistry* **269,** 8931–8936 (1994).

146. Saxl, T. *et al.* Fluorescence lifetime spectroscopy and imaging of nano-engineered glucose sensor microcapsules based on glucose/galactose-binding protein. *Biosensors and Bioelectronics* **24,** 3229–3234 (2009).

147. Pospıšil, P. *et al.* Fluorescence quenching of (dimethylamino) naphthalene dyes badan and prodan by tryptophan in cytochromes P450 and micelles. *The Journal of Physical Chemistry B* **118,** 10085–10091 (2014).

148. Fonin, A. V. *et al.* Photophysical Properties of BADAN Revealed in the Study of GGBP Structural Transitions. *International Journal of Molecular Sciences* **22,** 11113 (2021).

149. Kim, T. G., Wolford, M. F. & Topp, M. R. Ultrashort-lived excited states of aminophthalimides in fluid solution. *Photochemical & Photobiological Sciences* **2,** 576–584 (2003).

150. Paul, A. & Samanta, A. Solute rotation and solvation dynamics in an alcohol-functionalized room temperature ionic liquid. *The Journal of Physical Chemistry B* **111,** 4724–4731 (2007).

151. Saroja, G. & Samanta, A. Polarity of the micelle-water interface as seen by 4-aminophthalimide, a solvent sensitive fluorescence probe. *Chemical physics letters* **246,** 506–512 (1995).

152. Saroja, G., Soujanya, T., Ramachandram, B. & Samanta, A. 4-Aminophthalimide derivatives as environment-sensitive probes. *Journal of Fluorescence* **8,** 405–410 (1998).

153. Soujanya, T., Fessenden, R. & Samanta, A. Role of nonfluorescent twisted intramolecular charge transfer state on the photophysical behavior of aminophthalimide dyes. *The Journal of physical chemistry* **100,** 3507–3512 (1996).

154. Schieschke, N. *et al.* Geometry dependence of excitonic couplings and the consequences for configuration-space sampling. *Journal of Computational Chemistry* **42,** 1402–1418 (2021).

155. Kranz, J. J. *et al.* Time-dependent extension of the long-range corrected density functional based tight-binding method. *Journal of Chemical Theory and Computation* **13,** 1737–1747 (2017).

156. Li, J. *et al.* Automatic discovery of photoisomerization mechanisms with nanosecond machine learning photodynamics simulations. *Chemical science* **12,** 5302–5314 (2021).

157. Chen, Y. & Topp, M. R. Infrared-optical double-resonance measurements of hydrogen-bonding interactions in clusters involving aminophthalimides. *Chemical Physics* **283,** 249–268 (2002).

158. Wang, R. *et al.* Time-dependent density functional theory study on the electronic excited-state hydrogen-bonding dynamics of 4-aminophthalimide (4AP) in aqueous solution: 4AP and 4AP–(H2O) 1, 2 clusters. *Journal of computational chemistry* **31,** 2157–2163 (2010).

A. Appendix

A.1. Evaluation of the BPC glucose with different cut-off

The atoms within different cut-off range were selected on the "closed" wild-type GGBP crystal structure in the VMD, and were replaced by their respective force field point charges. After a set of QM calculation at the HF/6-311G* level on the glucose molecule, these polarised glucose molecules have undergone a set of 50 ns MD simulations in an explicit aqueous solution. The average number of H-bonds generated between water molecules and these PPC glucose molecules are shown in Table. A.1. The variance of population standard deviation of these results is 0.02, hence the cut-off distance at such range has few effect on the glucose charge polarisation.

Table A.1.: The average number of H-bonds generated between water molecules and BPC glucose molecules with different cut-off.

Cut-off range	Average number of H-bonds
4Å with full residues	10.80
5Å with full residues	10.77
5Å with broken residues	10.46
6Å with broken residues	10.76
7Å with broken residues	10.44
8Å with broken residues	10.68

A.2. Repolarised glucose charges with evaluation in free MD with water

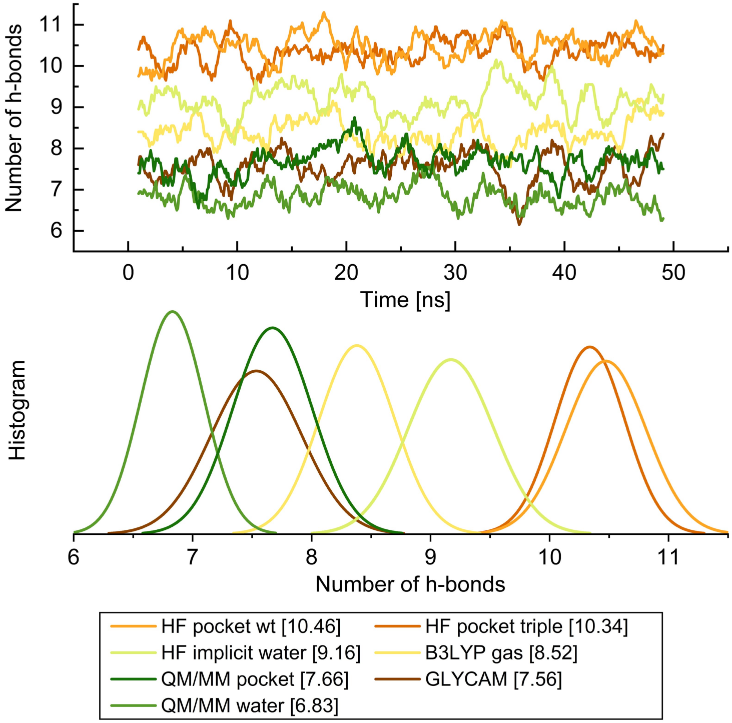


Figure A.1.: **A).** The number of hydrogen bonds generated between water molecules and the glucose polarised in different environments with all approaches. **B).** The normal distribution histogram of hydrogen bonds obtained by the different glucose charges. The number in the square brackets indicates the average number of hydrogen bonds of each corresponding glucose.

Table A.2.: Polarised charges computed for the glucose molecule with the various methods.

atom	pocket wt* (6-311G*)	pocket wt* (6-311G**)	pocket triple* (6-311G*)	implicit water	gas phase	QM/MM* pocket	QM/MM* water	Glycam
H1	0.540	0.521	0.544	0.483	0.449	0.437	0.413	0.445
O1	−0.747	−0.719	−0.748	−0.708	−0.649	−0.619	−0.625	−0.639
C1	0.347	0.338	0.257	0.414	0.303	0.464	0.451	0.384
H2	0.098	0.098	0.119	0.040	0.053	0.007	0.029	0.
O2	−0.542	−0.544	−0.462	−0.487	−0.400	−0.499	−0.486	−0.471
C2	0.124	0.128	0.022	0.113	0.057	0.144	0.155	0.225
H3	0.140	0.139	0.126	0.089	0.108	0.056	0.056	0.
C3	0.245	0.232	0.105	0.255	0.172	0.137	0.131	0.282
H4	0.031	0.032	0.096	0.047	0.072	0.045	0.058	0.
H5	0.031	0.032	0.096	0.047	0.072	0.059	0.055	0.
O3	−0.835	−0.809	−0.805	−0.730	−0.661	−0.640	−0.626	−0.688
H6	0.533	0.515	0.534	0.450	0.416	0.419	0.406	0.424
C4	0.052	0.052	0.063	0.098	0.077	0.166	0.157	0.276
H7	0.110	0.110	0.125	0.103	0.114	0.045	0.064	0.
O4	−0.717	−0.692	−0.685	−0.748	−0.687	−0.609	−0.626	−0.714
H8	0.516	0.497	0.518	0.500	0.478	0.421	0.407	0.440
C5	0.156	0.135	0.112	0.219	0.116	0.176	0.160	0.284
H9	0.138	0.143	0.160	0.078	0.107	0.053	0.062	0.
O5	−0.806	−0.778	−0.832	−0.741	−0.677	−0.690	−0.624	−0.709
H10	0.522	0.503	0.533	0.487	0.464	0.449	0.406	0.432
C6	0.248	0.246	0.318	0.138	0.110	0.129	0.120	0.310
H11	0.095	0.093	0.095	0.103	0.122	0.080	0.073	0.
O6	−0.858	−0.838	−0.868	−0.724	−0.669	−0.684	−0.623	−0.718
H12	0.581	0.568	0.575	0.474	0.451	0.455	0.407	0.437

*done by Monja Sokolov

A.3. Alchemical calculations for glucose binding free energy

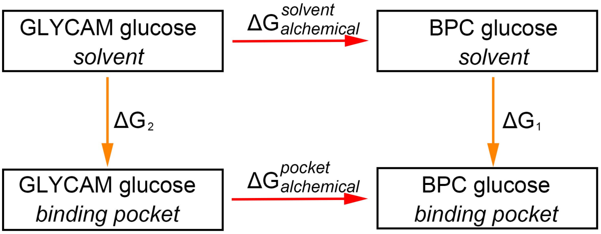


Figure A.2.: Thermodynamics cycle for glucose binding free energy. ΔG_1 and ΔG_2 is the GGBP binding free energy for GLYCAM glucose and BPC glucose. $\Delta G^{pocket}_{alchemical}$ is the energy difference for bound glucose when passing from the GLYCAM charges to BPC charges in alchemical calculations. $\Delta G^{solvent}_{alchemical}$ is the energy difference for unbound glucose when passing from the GLYCAM charges to BPC charges in alchemical calculations.

According to the thermodynamics cycle,

$$\Delta G_1 + \Delta G^{\text{pocket}}_{\text{alchemical}} = \Delta G_2 + \Delta G^{\text{pocket}}_{\text{alchemical}} \tag{A.1}$$

Hence, the glucose binding free energy difference between BPC glucose and GLYCAM glucose is,

$$\Delta\Delta G_{\text{binding}} = \Delta G_1 - \Delta G_2 = \Delta G^{\text{pocket}}_{\text{alchemical}} - \Delta G^{\text{solvent}}_{\text{alchemical}} \tag{A.2}$$

Table A.3.: Binding free energy difference (kcal/mol) from BPC glucose to GLYCAM glucose.

	$\Delta G^{\text{pocket}}_{\text{alchemical}}$	$\Delta G^{\text{solvent}}_{\text{alchemical}}$	$\Delta\Delta G_{\text{binding}}$
wild-type	2.0	1.5	0.5
triple mutant	1.2	1.0	0.2

A.4. Excited state wrong following by LC-DFTB with el.parameters

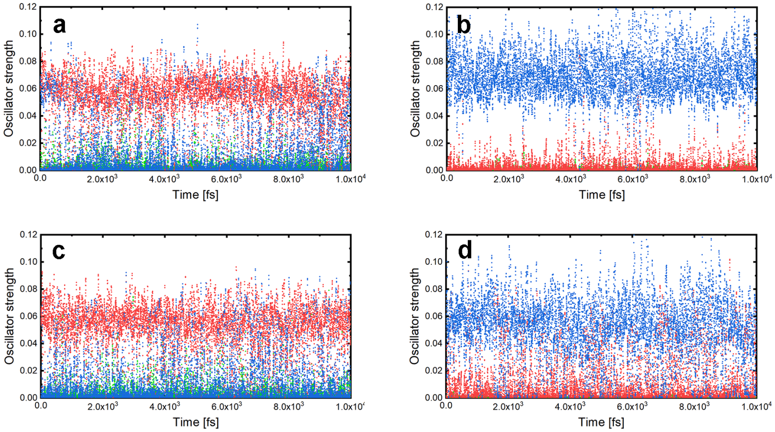


Figure A.3.: Oscillator strengths computed with LC-DFTB (el.parameters) along the trajectory of 4-AP in different environments. a). In vacuo, computed along the S3 trajectory, with correct state following rate of 17.26%. b). In DMSO, computed along the S2 trajectory with correct state following rate of 0.76%. c). In DMSO, computed along the S3 trajectory with correct state following rate of 13.62%. d). In water, computed along the S1 trajectory. The correct state was not followed. S_1–green, S_2–red, S_3–blue.

B. Appendix: Codes

B.1. Plumed inputs for the Host-Guest Chemistry

Plumed input file for 3-Et cage with NEt_4^+ (Chapter 6.1.2):

```
UNITS LENGTH=A TIME=0.001 # Angstrom and femtoseconds

# Definition for the center of mass of the host and the guest
cage_center: COM ATOMS=4,5,13,17,25,29,40,41,49,53,61,65,76,77,85,89,97,101,112,
113,121,125,133,137,148,149,157,161,169,173,184,185,193,197,205,209,220,221,229,
233,241,245,256,257,265,269,277,281
ligand_center: COM ATOMS=289-317

d1: DISTANCE ATOMS=cage_center,ligand_center
d2: DISTANCE ATOMS=cage_center,3998

# keep the guest not sample the outside of the host
UPPER_WALLS ARG=d1 AT=+7.0 KAPPA=500.0 EXP=2 LABEL=uwall

# keep the counter ions out side the host
LOWER_WALLS ARG=d2 AT=+5.0 KAPPA=500.0 EXP=2 LABEL=lwall

METAD ...
 LABEL=metad
 ARG=d1
 PACE=1000
 HEIGHT=0.5
 SIGMA=0.2
 GRID_MIN=0.0
 GRID_MAX=30.0
 GRID_BIN=1000
 FILE=HILLS
 BIASFACTOR=100
 TEMP=298.0
... METAD

PRINT STRIDE=100 FILE=COLVAR ARG=d1,d2,metad.*,uwall.bias,lwall.bias
```

Plumed input file for 3-Et cage with NMe_4^+ (Chapter 6.1.2):

```
UNITS LENGTH=A TIME=0.001 # Angstrom and femtoseconds

# Definition for the center of mass of the host and the guest
cage_center: COM ATOMS=4,5,13,17,25,29,40,41,49,53,61,65,76,77,85,89,97,101,112,
113,121,125,133,137,148,149,157,161,169,173,184,185,193,197,205,209,220,221,229,
233,241,245,256,257,265,269,277,281
ligand_center: COM ATOMS=289-305

d1: DISTANCE ATOMS=cage_center,ligand_center
d2: DISTANCE ATOMS=cage_center,4051

# keep the guest not sample the outside of the host
UPPER_WALLS ARG=d1 AT=+7.0 KAPPA=500.0 EXP=2 LABEL=uwall

# keep the counter ions out side the host
LOWER_WALLS ARG=d2 AT=+5.0 KAPPA=500.0 EXP=2 LABEL=lwall

METAD ...
 LABEL=metad
 ARG=d1
 PACE=1000
 HEIGHT=0.5
 SIGMA=0.2
 GRID_MIN=0.0
 GRID_MAX=30.0
 GRID_BIN=1000
 FILE=HILLS
 BIASFACTOR=50
 TEMP=298.0
... METAD

PRINT STRIDE=100 FILE=COLVAR ARG=d1,d2,metad.*,uwall.bias,lwall.bias
```

Plumed input file for nitrogen transfer in F-cage crystal with “neighbour cage” mechanism (Chapter 6.2.2):

```
UNITS LENGTH=A TIME=0.001 ENERGY=kcal/mol # Angstrom and femtoseconds

WHOLEMOLECULES ENTITY0=3-5458
lig: CENTER ATOMS=1,2

#funnel restrains
fps: FUNNEL_PS LIGAND=lig REFERENCE=start.pdb ANCHOR=85
     POINTS=12.62,17.35,50.85,14.22,19.52,19.81
FUNNEL ARG=fps.lp,fps.ld ZCC=40.0 ALPHA=0.01 RCYL=8.0 MINS=0.0 MAXS=40.0
     KAPPA=50000 NBINS=500 NBINZ=500 FILE=BIAS LABEL=funnel

```

```
LOWER_WALLS ARG=fps.lp AT=2.0 KAPPA=50000 EXP=2 OFFSET=0 LABEL=lwall
UPPER_WALLS ARG=fps.lp AT=38.0 KAPPA=50000 EXP=2 OFFSET=0 LABEL=uwall
UPPER_WALLS ARG=fps.ld AT=7.0 KAPPA=50000 EXP=2 OFFSET=0 LABEL=uwall1

METAD ...
 LABEL=metad
 ARG=fps.lp,fps.ld
 PACE=1000
 HEIGHT=1.2
 SIGMA=0.2,0.2
 GRID_MIN=0,0
 GRID_MAX=40,8
 GRID_BIN=1000,200
 FILE=HILLS
 BIASFACTOR=20
 TEMP=77.0
 CALC_RCT
... METAD

PRINT STRIDE=100 FILE=COLVAR ARG=fps.lp,fps.ld,metad.*
```

Plumed input file for nitrogen transfer in F-cage crystal with “void cage” mechanism (Chapter 6.2.2):

```
UNITS LENGTH=A TIME=0.001 ENERGY=kcal/mol # Angstrom and femtoseconds

WHOLEMOLECULES ENTITY0=3-5458
lig: CENTER ATOMS=1,2

#funnel restrains
fps: FUNNEL_PS LIGAND=lig REFERENCE=start.pdb ANCHOR=96
     POINTS=12.98,17.84,43.84,9.82,29.65,35.77
FUNNEL ARG=fps.lp,fps.ld ZCC=40.0 ALPHA=0.01 RCYL=6.5 MINS=-2.0 MAXS=17.0
     KAPPA=50000 NBINS=500 NBINZ=500 FILE=BIAS LABEL=funnel

LOWER_WALLS ARG=fps.lp AT=0.0 KAPPA=50000 EXP=2 OFFSET=0 LABEL=lwall
UPPER_WALLS ARG=fps.lp AT=15.0 KAPPA=50000 EXP=2 OFFSET=0 LABEL=uwall
UPPER_WALLS ARG=fps.ld AT=6.0 KAPPA=50000 EXP=2 OFFSET=0 LABEL=uwall1

METAD ...
 LABEL=metad
 ARG=fps.lp,fps.ld
 PACE=1000
 HEIGHT=1.2
 SIGMA=0.2,0.2
 GRID_MIN=-2,0
 GRID_MAX=17,8
 GRID_BIN=475,200
 FILE=HILLS
 BIASFACTOR=20
 TEMP=77.0
 CALC_RCT
... METAD

PRINT STRIDE=100 FILE=COLVAR ARG=fps.lp,fps.ld,metad.*
```

B.2. Plumed inputs for the GGBP

Plumed input file for wild-type GGBP well-tempered metadynamics (Chapter 7.2.3):

```
UNITS LENGTH=A TIME=0.001 # Angstrom and femtoseconds

# Definition for the center of mass of the N-domain (n_center), C-domain (c_center), junction between N-domain & hinge region (n_base), junction between C-domain & hinge region (c_base), and the hinge region (h_center).
n_center: COM ATOMS=15-1673,3897-4393
c_center: COM ATOMS=1707-3851,4485-4632
n_base: COM ATOMS=1674-1680,3897-3906,4394-4417
c_base: COM ATOMS=1695-1706,3852-3870,4469-4484
h_center: COM ATOMS=1674-1706,3852-3906,4394-4484

# Definition for the center of mass of the binding pocket without Badan side-chain (p_center) and the glucose (g_center)
p_center: COM ATOMS=204-215,1377-1390,2318-2334,2349-2360,2396-2419,3236-3249,3592-3603,3871-3884
g_center: COM ATOMS=4633-4656

# Definition for the center of mass of three residues at junction between N-domain & hinge region (n_a, n_b, n_c), three residues at junction between C-domain & hinge region (c_a, c_b, c_c),
n_a: COM ATOMS=1674-1680
n_b: COM ATOMS=3897-3906
n_c: COM ATOMS=4394-4417
c_a: COM ATOMS=1695-1706
c_b: COM ATOMS=3852-3870
c_c: COM ATOMS=4469-4484

# Collective variables definition for the theta (cv1) and phi (cv2)
cv1: ANGLE ATOMS=n_center,h_center,c_center
cv2: TORSION ATOMS=n_center,n_base,c_base,c_center

# limit glucose not too far away from GGBP
```

```
restrain1: DISTANCE ATOMS=p_center,g_center
UPPER_WALLS ARG=restrain1 AT=+30.0 KAPPA=150.0 EXP=2 LABEL=uwall1

# limit protein not over-twists
restrain2: CUSTOM ARG=cv2 FUNC=sin(x) PERIODIC=NO
LOWER_WALLS ARG=restrain2 AT=-0.34 KAPPA=500.0 EXP=4 LABEL=lwall2

# protect hinge region to avoid protein deconstruction
restrain3: DISTANCE ATOMS=n_a,n_c
restrain4: DISTANCE ATOMS=n_b,n_c
restrain5: DISTANCE ATOMS=c_a,c_c
restrain6: DISTANCE ATOMS=c_b,c_c
UPPER_WALLS ARG=restrain3 AT=+15.0 KAPPA=500.0 EXP=4 LABEL=uwall3
UPPER_WALLS ARG=restrain4 AT=+20.0 KAPPA=500.0 EXP=4 LABEL=uwall4
UPPER_WALLS ARG=restrain5 AT=+20.0 KAPPA=500.0 EXP=4 LABEL=uwall5
UPPER_WALLS ARG=restrain6 AT=+15.0 KAPPA=500.0 EXP=4 LABEL=uwall6

# limit protein not over-opened
restrain7: DISTANCE ATOMS=c_center,n_center
UPPER_WALLS ARG=restrain7 AT=+36.0 KAPPA=500.0 EXP=4 LABEL=uwall7
UPPER_WALLS ARG=cv1 AT=+2.80 KAPPA=500.0 EXP=4 LABEL=uwall8
LOWER_WALLS ARG=cv1 AT=+1.74 KAPPA=500.0 EXP=4 LABEL=lwall8

METAD ...
 LABEL=metad
 ARG=cv1,cv2
 PACE=1000
 HEIGHT=1.2
 SIGMA=0.2,0.2
 GRID_MIN=-pi,-pi
 GRID_MAX=pi,pi
 GRID_BIN=150,150
 FILE=HILLS
 BIASFACTOR=10
 TEMP=300.0
... METAD
PRINT STRIDE=100 FILE=COLVAR ARG=cv1,cv2,metad.*
```

Plumed input file for GGBP triple mutant well-tempered metadynamics (Chapter 7.2.3):

```
UNITS LENGTH=A TIME=0.001 # Angstrom and femtoseconds

# Definition for the center of mass of the N-domain (n_center), C-domain (c_center), junction between N-domain & hinge region (n_base), junction between C-domain & hinge region (c_base), and the hinge region (h_center).
n_center: COM ATOMS=15-1673,3926-4422
c_center: COM ATOMS=1707-3880,4514-4661
```

```
n_base: COM ATOMS=1674-1680,3926-3935,4423-4446
c_base: COM ATOMS=1695-1706,3881-3899,4498-4513
h_center: COM ATOMS=1674-1706,3881-3935,4423-4513

# Definition for the center of mass of the binding pocket without Badan side-chain (p_center) and the glucose (g_center)
p_center: COM ATOMS=204-215,1377-1390,2318-2326,2372-2383,2419-2442,3259-3272,3629-3640,3900-3913
g_center: COM ATOMS=4662-4685

# Definition for the center of mass of three residues at junction between N-domain & hinge region (n_a, n_b, n_c), three residues at junction between C-domain & hinge region (c_a, c_b, c_c),
n_a: COM ATOMS=1674-1680
n_b: COM ATOMS=3926-3935
n_c: COM ATOMS=4423-4446
c_a: COM ATOMS=1695-1706
c_b: COM ATOMS=3881-3899
c_c: COM ATOMS=4498-4513

# Collective variables definition for the theta (cv1) and phi (cv2)
cv1: ANGLE ATOMS=n_center,h_center,c_center
cv2: TORSION ATOMS=n_center,n_base,c_base,c_center

# limit glucose not too far away from GGBP
restrain1: DISTANCE ATOMS=p_center,g_center
UPPER_WALLS ARG=restrain1 AT=+30.0 KAPPA=150.0 EXP=2 LABEL=uwall1

# limit protein not over-twists
restrain2: CUSTOM ARG=cv2 FUNC=sin(x) PERIODIC=NO
LOWER_WALLS ARG=restrain2 AT=-0.34 KAPPA=500.0 EXP=4 LABEL=lwall2

# protect hinge region to avoid protein deconstruction
restrain3: DISTANCE ATOMS=n_a,n_c
restrain4: DISTANCE ATOMS=n_b,n_c
restrain5: DISTANCE ATOMS=c_a,c_c
restrain6: DISTANCE ATOMS=c_b,c_c
UPPER_WALLS ARG=restrain3 AT=+15.0 KAPPA=500.0 EXP=4 LABEL=uwall3
UPPER_WALLS ARG=restrain4 AT=+20.0 KAPPA=500.0 EXP=4 LABEL=uwall4
UPPER_WALLS ARG=restrain5 AT=+20.0 KAPPA=500.0 EXP=4 LABEL=uwall5
UPPER_WALLS ARG=restrain6 AT=+15.0 KAPPA=500.0 EXP=4 LABEL=uwall6

# limit protein not over-opened
restrain7: DISTANCE ATOMS=c_center,n_center
UPPER_WALLS ARG=restrain7 AT=+36.0 KAPPA=500.0 EXP=4 LABEL=uwall7
UPPER_WALLS ARG=cv1 AT=+2.80 KAPPA=500.0 EXP=4 LABEL=uwall8
LOWER_WALLS ARG=cv1 AT=+1.74 KAPPA=500.0 EXP=4 LABEL=lwall8

METAD ...
```

```
 LABEL=metad
 ARG=cv1,cv2
 PACE=1000
 HEIGHT=1.2
 SIGMA=0.2,0.2
 GRID_MIN=-pi,-pi
 GRID_MAX=pi,pi
 GRID_BIN=150,150
 FILE=HILLS
 BIASFACTOR=10
 TEMP=300.0
... METAD
PRINT STRIDE=100 FILE=COLVAR ARG=cv1,cv2,metad.*
```

Plumed input file for wild-type GGBP funnel metadynamics (Chapter 7.2.4):

```
UNITS LENGTH=A TIME=0.001 ENERGY=kcal/mol # Angstrom and femtoseconds
WHOLEMOLECULES ENTITY0=1-4632

lig: COM ATOMS=4633-4656

#funnel restrains
fps: FUNNEL_PS LIGAND=lig REFERENCE=start.pdb ANCHOR=656 POINTS=85,80,32,-20,40,10
FUNNEL ARG=fps.lp,fps.ld ZCC=25.00 ALPHA=0.2 RCYL=4 MINS=0 MAXS=30 KAPPA=35100
     NBINS=500 NBINZ=500 FILE=BIAS LABEL=funnel
LOWER_WALLS ARG=fps.lp AT=2.0 KAPPA=35100 EXP=2 OFFSET=0 LABEL=lwall
UPPER_WALLS ARG=fps.lp AT=25.0 KAPPA=35100 EXP=2 OFFSET=0 LABEL=uwall

METAD ...
 LABEL=metad
 ARG=fps.lp
 PACE=1000
 HEIGHT=1.2
 SIGMA=0.2
 GRID_MIN=0
 GRID_MAX=30
 GRID_BIN=750
 FILE=HILLS
 BIASFACTOR=10
 TEMP=300.0
... METAD
PRINT STRIDE=100 FILE=COLVAR ARG=fps.lp,metad.*
```

Plumed input file for GGBP funnel triple mutant metadynamics (Chapter 7.2.4):

```
UNITS LENGTH=A TIME=0.001 ENERGY=kcal/mol # Angstrom and femtoseconds
WHOLEMOLECULES ENTITY0=1-4681,4662-4685
```

```

lig: COM ATOMS=4662-4685

#funnel restrains
fps: FUNNEL_PS LIGAND=lig REFERENCE=start.pdb ANCHOR=2379 POINTS=90,75,45,90,60,20
FUNNEL ARG=fps.lp,fps.ld ZCC=20 ALPHA=0.2 RCYL=4 MINS=0 MAXS=30 KAPPA=35100 NBINS=500 NBINZ=500 FILE=BIAS LABEL=funnel
LOWER_WALLS ARG=fps.lp AT=2.0 KAPPA=35100 EXP=2 OFFSET=0 LABEL=lwall
UPPER_WALLS ARG=fps.lp AT=25.0 KAPPA=35100 EXP=2 OFFSET=0 LABEL=uwall

METAD ...
 LABEL=metad
 ARG=fps.lp
 PACE=1000
 HEIGHT=1.2
 SIGMA=0.2
 GRID_MIN=0
 GRID_MAX=30
 GRID_BIN=750
 FILE=HILLS
 BIASFACTOR=10
 TEMP=300.0
... METAD
PRINT STRIDE=100 FILE=COLVAR ARG=fps.lp,metad.*
```

List of Figures

List of Tables