LUND UNIVERSITY

Molecular Recognition and Conformational Dynamics in Macromolecules

Bhakat, Soumendranath

2020

Document Version:
Publisher's PDF, also known as Version of record

Link to publication

Citation for published version (APA):
Bhakat, S. (2020). *Molecular Recognition and Conformational Dynamics in Macromolecules*. Biophysical Chemistry (LTH), Lund University.

Total number of authors:
1

Creative Commons License:
CC0

General rights
Unless other specific re-use rights are stated the following general rights apply:
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.
• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal

Read more about Creative commons licenses: https://creativecommons.org/licenses/

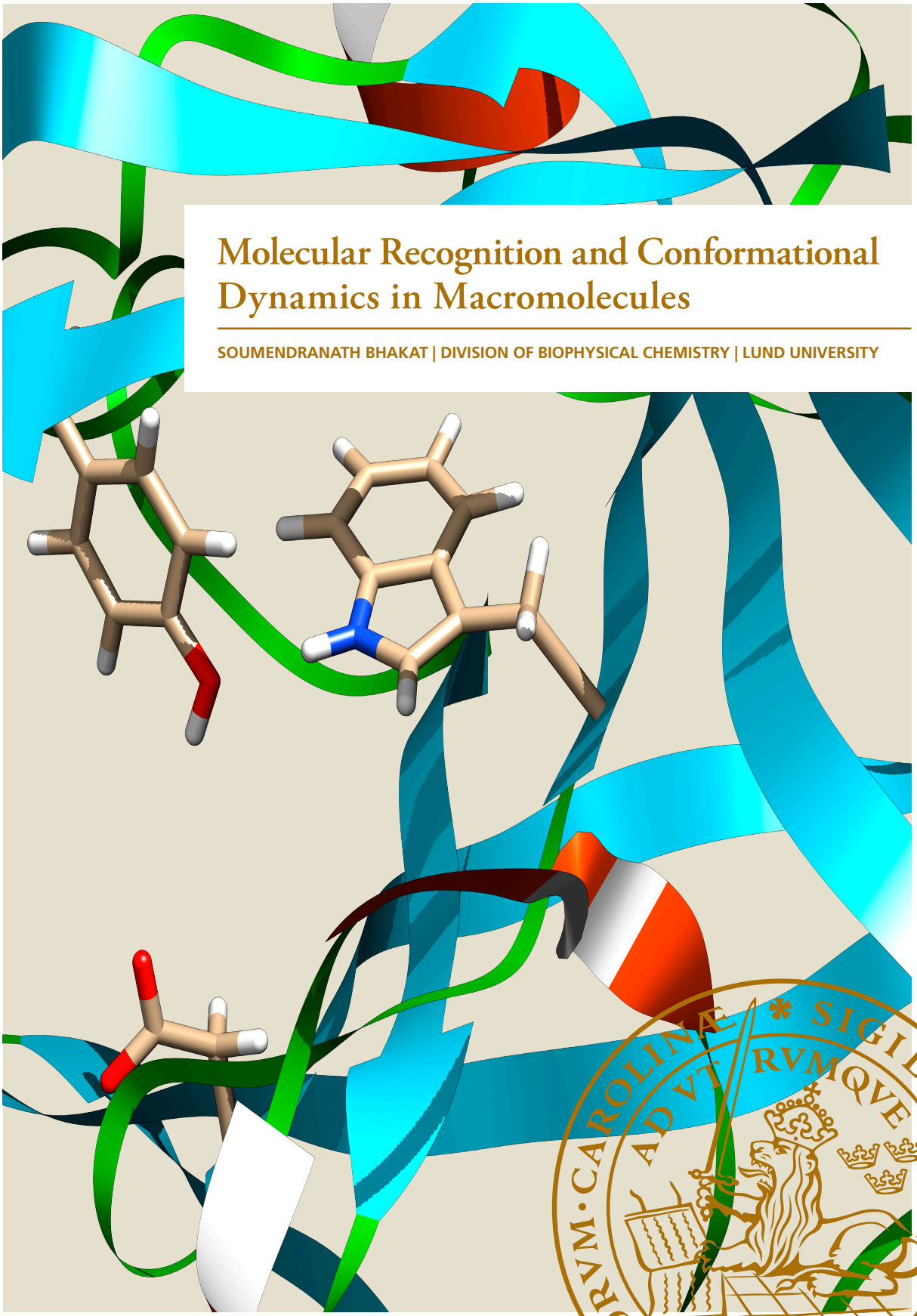
Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

LUND UNIVERSITY

PO Box 117
221 00 Lund
+46 46-222 00 00

Molecular Recognition and Conformational Dynamics in Macromolecules

SOUMENDRANATH BHAKAT | DIVISION OF BIOPHYSICAL CHEMISTRY | LUND UNIVERSITY



Molecular Recognition and Conformational Dynamics in Macromolecules

Molecular Recognition and Conformational Dynamics in Macromolecules

Soumendranath Bhakat

LUND
UNIVERSITY

DOCTORAl DISSERTATION

Faculty opponent:
Dr. Ran Friedman
Linnéuniversitetet, Sweden

By due permission of the Faculty of Engineering of Lund University, Sweden for public criticism in the lecture hall F of Kemicentrum on Thursday, 28th of May 2020 at 09:00.

Organization	Document name
LUND UNIVERSITY Division of Biophysical Chemistry Box 124 SE-221 00 Lund Sweden	DOCTORAL DISSERTATION
	Date of issue 2020-05-06
Author Soumendranath Bhakat	Sponsoring organization The Crafoordska Foundation Swedish Research Council

Title

Molecular Recognition and Conformational Dynamics in Macromolecules

Abstract

Computational methods gained a widespread use in drug discovery. Understanding conformational dynamics of protein and mechanisms of protein-ligand binding are two major areas in drug discovery. Molecular dynamics (MD) simulation have been routinely used to study conformational dynamics of protein and mechanisms of protein-ligand binding. In classical MD simulation, the system often remains stuck in a local free energy minimum for a long time. Hence, conformational changes associated with long timescales (e.g. loop motion, ligand binding/unbinding etc.) are beyond reach of classical MD simulation. Metadynamics is an enhanced sampling method which deposits bias along some chosen reaction coordinate and forces the system to escape local minimum thus, allows better sampling of the conformational space. In this thesis, I have used MD and metadynamics to study protein-ligand binding and conformational dynamics of globular proteins. We found that the presence of trapped water in the binding site of the protein plays a key role ligand binding. Further, we found that the side-chains of binding site residues and flexibility of ligands play a key role in the protein-ligand binding. We also studied how rotation of tyrosine dictates conformational dynamics in a class of protein known as pepsin-like aspartic protease. We found that apo protease remains in a dynamic equilibrium between normal and flipped states due to rotation of tyrosine side-chain. Conformational dynamics also plays a crucial role in hydrogen exchange via solvent penetration. Local fluctuations in protein breaks the hydrogen bond interactions involving backbone amides which allows solvent penetration. We defined this metastable state as *broken* state. In the broken state, the backbone amide forms hydrogen bond interaction with water molecule. Using molecular dynamics and metadynamics we predicted free energy difference between the broken and ground state (backbone amide remains hydrogen bonded with neighboring residue) in a small globular protein.

Key words

Protein, ligand, host-guet, funnel metadynamics, MM/PBSA, well-tempered metadynamics, collective variable, tyrosine, aspartic protease, local fluctuations, hydrogen exchange, time-lagged independent component analysis, principal component analysis, parallel-tempering

Classification system and/or index terms (if any)

Supplementary bibliographical information	Language English
ISSN and key title	ISBN 978-91-7422-745-1 (print) 978-91-7422-746-8 (e-version)
Recipient´s notes	Number of pages 219 / Price
	Security classification

I, the undersigned, being the copyright owner of the abstract of the above-mentioned dissertation, hereby grant to all reference sources permission to publish and disseminate the abstract of the above-mentioned dissertation.

Signature Soumendranath Bhakat Date 2020-04-27

Molecular Recognition and Conformational Dynamics in Macromolecules

Soumendranath Bhakat

LUND
UNIVERSITY

Evaluation Committee

Dr. Lucie Delemotte
Department of Applied Physics
KTH Royal Institute of Technology
Stockholm, Sweden

Dr. Elena Papaleo
Danish Cancer Society Research Center
Copenhagen, Denmark

Prof. Mikael Lund
Department of Theoretical Chemistry
Lund University
Lund, Sweden

Cover: Location of tyrosine, tryptophan and catalytic aspartic acid in pepsin-like aspartic protease. Designed by: Soumendranath Bhakat

Funding: This work is financially supported by the Crafoordska Foundation and the Swedish research council

PUBLIC DOMAIN Pages 1-49 Soumendranath Bhakat, 2020

Division of Biophysical Chemistry
Faculty of Engineering, Lund University, Sweden
ISBN: 978-91-7422-745-1 (print)
ISBN: 978-91-7422-746-8 (pdf)

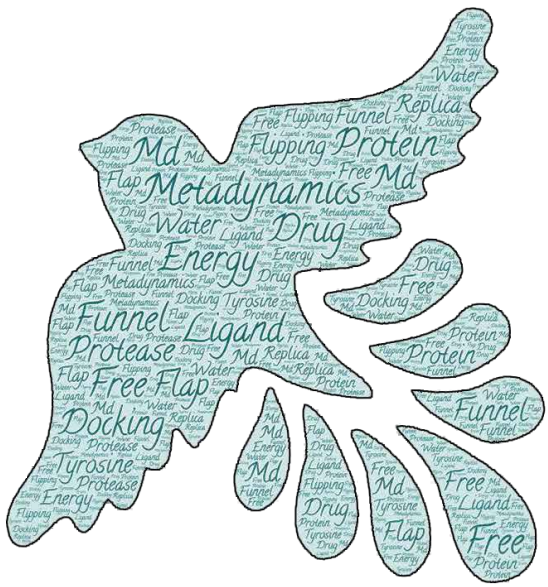
Printed in Sweden by Media-Tryck, Lund University, Lund 2020

NORDIC SWAN ECOLABEL
Printed matter
3041 0903

Media-Tryck is a Nordic Swan Ecolabel certified provider of printed material. Read more about our environmental work at www.mediatryck.lu.se

MADE IN SWEDEN

May everyone be happy and safe



Contents

List of publications

This thesis is based on the following publications [1]:

1 **Resolving the problem of trapped water in binding cavities: prediction of host–guest binding free energies in the SAMPL5 challenge by funnel metadynamics**

Soumendranath Bhakat and Pär Söderhjelm*
J Comput Aided Mol Des, 2017, 31: 119-132

2 **Prediction of binding poses to FXR using multi-targeted docking combined with molecular dynamics and enhanced sampling**

Soumendranath Bhakat, Emil Åberg and Pär Söderhjelm*
J Comput Aided Mol Des, 2018, 32(1): 59-73

3 **Flap dynamics in pepsin-like aspartic proteases: a computational perspective using Plasmepsin-II and BACE-1 as model systems**

Soumendranath Bhakat* and Pär Söderhjelm*
Manuscript, 2020

4 **Computational modelling of local fluctuations causing transient solvent-exposure of protein amides**

Soumendranath Bhakat* and Pär Söderhjelm*
Manuscript, 2020

[1] * denotes corresponding author/s

The following paper is not a part of this thesis:

5 **Flap Dynamics in Aspartic Proteases: A Computational Perspective**

Mukul Mahanti, **Soumendranath Bhakat**, Ulf J. Nilsson and Pär Söderhjelm*
Chem Biol Drug Des, 2016, 88: 159-177

Popular summary

Amino acids are the building blocks of life. They join together via chemical bonding and forms a polymer known as protein. Proteins have an extraordinary property, the ability to perform chemical catalysis. In solution, proteins undergo several conformational changes (known as *conformational dynamics*) necessary for their function. In many cases, proteins are related to certain disease conditions. In order to combat a disease, one can develop a drug which binds to a specific protein and hampers its function. Hence, understanding conformational dynamics and mechanism of drug binding to a protein is necessary for drug discovery.

Drug discovery efforts relies heavily on experimental methods. Several experimental methods have been developed to understand conformational dynamics and mechanism of drug binding (often known as *molecular recognition*). However, these methods are time consuming, costly and often restricted to specialised research facilities. Computational techniques provide an alternative to the experiments. Development of hardware and software makes several computational methods accessible to a common person. Now, one can routinely use computational methods to understand conformational dynamics and mechanism of drug binding, using a fraction of the resources necessary to perform an experiment.

This thesis demonstrates how one can capture conformational dynamics and mechanisms of drug binding using computational methods such as docking, molecular dynamics and metadynamics. *Docking* predicts binding pose and binding affinity between proteins and drug molecules. *Molecular dynamics* samples time-dependent dynamics of a system (such as protein, protein-drug complex etc.) using Newton's second law of motion. *Metadynamics* aims at sampling configurational space of a system along a chosen reaction co-ordinate.

The first paper aims at predicting binding free energies using a set of host-guest molecules. Host-guest systems are frequently used in computational studies as a *toy* model to mimic protein–ligand systems. In this work, we have used a variant of metadynamics, denoted *funnel metadynamics*, and molecular dynamics simulation to predict the binding free energies for these systems. Our prediction matches well with experiment which demonstrates the predictive power of our protocol.

In the second study, we combine docking, molecular dynamics and metadynam-

ics in order to predict the binding poses of 35 different ligands interacting with a particular protein. We managed to predict the correct binding poses for 29 out of the 35 ligands. This shows the capability of computational methods in predicting binding poses of small drug-like molecules.

The last two works deal with understanding conformational dynamics in proteins. In the third paper, we have used molecular dynamics and metadynamics to understand how rotation of one amino acid, tyrosine, dictates the conformational dynamics in plasmepsin-II and BACE-1 (drug targets for malaria and Alzheimer's disease, respectively). Studying conformational dynamics in these proteins is key to understanding drug binding pathways. We predicted that the rotation of the tyrosine side chain dictates the *opening* and *closing* motion of the flap (β-hairpin structure of the protein) that regulates drug binding.

In the fourth paper, we wanted to understand how conformational fluctuations in a protein affects solvent penetration. Here, we mainly focused on *local* fluctuations. The core of a protein is stabilised by hydrogen bond interactions involving backbone amides. Local fluctuations in a protein break these hydrogen bonds and allow solvent penetration, defining a *broken* state. We have used molecular dynamics and metadynamics to sample the *broken* state of a small protein and predicted the free energy difference between the *broken* and *ground* state.

I hope that the predictions made in this thesis will be helpful to guide future experiments.

Acknowledgements

Firstly, I would like to thank all the funding agencies and tax prayers for their support. A special thanks goes to **Lunarc** and High Performance Computing Center North (**HPC2N**) for their generous support with computational resource.

I would like to thank, **Pär** for making me familiar with the powerful sampling method, metadynamics. Thanks a lot for your patience, guidance and support during this period. **Mikael** and **Kristofer**, thanks for all your help and guidance over the years especially during the thesis writing. **Bertil**, thanks a lot for teaching the fundamentals of biophysical chemistry, story telling and helping me to integrate with the department in a smooth way. I will fondly remember **BPC spring outings** which were always a lot of fun. One of the most enjoyable part of my PhD journey was being involved in teaching. I absolutely cherished it. Thanks to everyone (fellow teaching assistants and students) who were involved in the process.

To the best office mates one can possibly have, **Filip** and **Zhiwei**: thanks for all the laughter, science and friendship. To NMR heroes, **Olof** and **Sven** for great discussions, coffee breaks, friendship and mostly keeping up with me. Shall we have a rematch, team MD (**Filip** and me) vs team NMR (**Olof and Zhiwei**)? **Johan**, thanks for all your advice, tennis practice and thoughts on life. **Ronja**, **Emil**, **Eric**, **Nils**, **Magdalena** and **Johana**, thanks for all the discussions during your time at the department. I really enjoyed it. I would also like to thank **Uli**, **Mandar**, **Santosh** and **Ashar** for great scientific and non-scientific discussions.

Sharing my experience with each members of biochemistry gang will take a lifetime. Hence, I will thank them as a cluster (not in any particular order). Thanks, **Dev, Egle, Samuel, Stefan, Carl Johan, Simon, Mathias, Mattias, Veronica, Björn, Thom, Karin, Rohit, Jennifer, Abhishek, Tamim, Morteza, Signe, Caroline, Mads, Camille, Teun, Helin, Isabella, Eimantas, Yonathan** for all the discussions and fun which plays a significant role in my PhD journey.

To **Magnus** and **Maryam**, thanks for all your help and conversations during coffee breaks.

To **Paula**, thanks for being understanding and all your help during the printing process.

My math teacher from high school, **Ashish Chakraborty**; English teacher, **Shashti**

K. Das; Physics teacher, **Jyoti Sir**, thanks for sharing your knowledge and inspiration. I also have to mention, **Dr. Venkatesan Jayaprakash** for his inspiring lectures during undergraduate days. I hope, we will keep up our collaboration in future.

To all my friends, I love you dearly, you know who you are.

To, **Somnath Bhakat** for inspiring a generation and proving that one can still pursue research even if you have less financial and technical support. You are the **Best Teacher Ever**. **Mallika Bhakat**, nothing would be possible without you. You are the best.

Chapter 1

Introduction

The last decade saw tremendous breakthrough in several frontiers in science [1], ranging from gene editing to gravitational waves, artificial intelligence to quantum computing [2]. However, three supreme mysteries of science, *origin of the universe*, *origin of the life* and *origin of consciousness* are still far from being answered. It is quite extraordinary that two fundamental physical concepts, statistical mechanics and quantum mechanics, developed in early nineteenth century, are intimately involved in explaining some of the nuances associated with these grand scientific mysteries[1]. The origin of life [3, 4], constantly attracted biophysicists for more than a decade.

In the early 20th century, Oparin and Haldane independently proposed an hypothesis which connected chemical evolution with origin of life. Today, the hypothesis is known as *Oparin-Haldane hypothesis* [5, 6]. According to this, early earth atmosphere was reducing in nature and mainly comprised of simple molecules such as hydrogen, methane, ammonia and water vapour. When exposed to a source of energy e.g. lightning, UV-radiation, volcanic eruption etc, these inorganic molecules performed some simple chemical reactions and produced building blocks of life such as amino-acids and nucleotides. These organic molecules accumulated in the sea which acted as a cooking pot and formed a *hot diluted soup* of organic monomers and polymers. Today, it is known as the *primordial soup* [7]. Upon further reactions, these monomers/polymers combined and eventually formed a molecule with an extraordinary property, the ability to perform bio-chemical catalysis.

Unfortunately, this hypothesis remained untested for over two decades. In 1953, graduate student Stanley Miller decided to test the Oparin-Haldane hypothesis in a simulated early-earth environment. Miller simulated the sea by simply putting water in the round-bottom flask, topped up with methane, ammonia, hydrogen and water vapour. He simulated the source of energy with electric sparks. After several days of sparking, Miller analysed the solution and discovered that it managed to synthesise

[1] See The Guardian's list of 20 big questions in science.

small-chain amino acids such as glycine, alanine, and aspartic acid.

Amino acids are fundamental building blocks of proteins. Proteins perform a vast array of functions in an organism, e.g. catalysis, DNA replication, signaling in response to external stimuli, transport of molecules, etc. In solution, proteins undergo conformational changes that are necessary for their function. Conformational changes in a protein are associated with a complex energy landscape, where each basin corresponds to a different conformation. Today, we can leverage upon the concepts of statistical mechanics, quantum mechanics and computer science to understand conformational motions in a protein. In this thesis, I will demonstrate how we can use statistical mechanics and computer simulation to understand protein dynamics.

Chapter 2

Statistical Thermodynamics

> Software is like entropy. It is difficult to grasp, weighs nothing and obey's the second law of thermodynamics; i.e. it always increases
>
> *Norman Ralph Augustine*

Thermodynamics is one of the supreme concepts (the other two are quantum mechanics and Newtonian mechanics) which governs this universe. The principles of thermodynamics were developed in the early eighteenth century[1]. Since then, several articles, books, monographs, and theses have been published which convey the underlying theories of thermodynamics. Going through these vast and somewhat complex theories is out of the scope of this thesis. Here, I introduce some key formulas while keeping the level consistent with an undergraduate chemistry course. The majority of the concepts of this chapter follows an introductory book written by Benjamin Widom [8] and Wereszczynski et al [9].

Statistical thermodynamics is the theoretical framework used to calculate properties of a macroscopic system from the molecular properties of the the vast number of particles that constitutes the system. In statistical thermodynamics, a system is described using an *ensemble* (Figure 2.1). Ensemble is a collection of a vast number of systems in different quantum states. At any instant in time, each of the system in the ensemble will be in different quantum states. When averaging over all the members of the ensemble, the macroscopic variables are obtained [10].

In statistical thermodynamics, the most fundamental relation, as realized by Boltzmann, is the entropy definition

$$S = k_{\mathrm{B}} \ln \Omega \tag{2.1}$$

[1]It all started in 1738 by Bernoulli. The major breakthrough happened in 1859 by James Clerk Maxwell who proposed the Maxwell Distribution.

where Ω is the number of available *microstates* for a system at constant internal energy U and k_{B} is the Boltzmann's constant[2] which has a value of 1.380649×10^{-23} J/K.

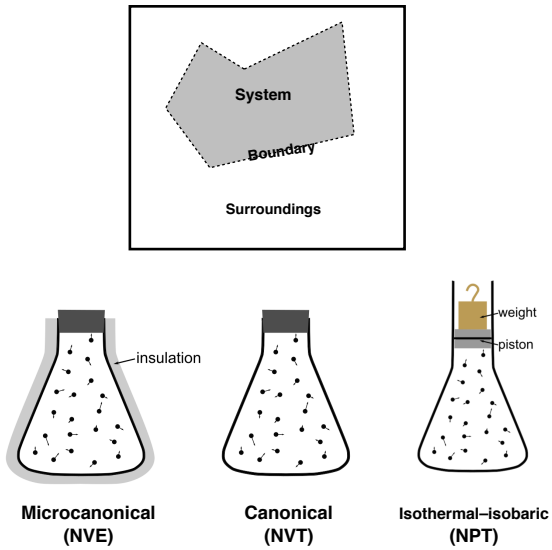


Figure 2.1: An artistic representation of a thermodynamic ensemble and some of the commonly used ensembles in thermodynamics [11]. The difference among these ensembles are due to different degrees of separation from the surroundings.

For a macroscopic system, the internal energy U is the sum of potential plus kinetic energy of all the molecules that make up the system. The change in internal energy, ΔU, is equal to the heat supplied to the system (q) plus the work (w) done on the system:

$$\Delta U = q + w \tag{2.2}$$

According to the first law of thermodynamics, if a system is thermally (no heat is exchanged, q = 0) and mechanically isolated (no work is done, w = 0) from the environment then the internal energy U is constant. Imagine that the system is not isolated but connected to a thermostat which fixes the temperature at T. Then the internal energy is not completely constant since the system is fluctuating among a truly vast number of *microstates* with possibly different energies, E_i.[3]

[2]"Boltzmann himself never introduced it — a peculiar state of affairs, which can be explained by the fact that Boltzmann apparently never gave thought to the possibility of carrying out an exact measurement of the constant": Max Planck's Nobel Lecture

[3]As an example, the number of available microstates for 1 mole of ideal gas at room temperature and normal pressure is typically on the order of $10^{10^{25}}$, that is 1 followed by 10^{25} zeros. To write this value down on paper would require a paper strip with a length of about 3 million light years. As a comparison, Wikipedia states that the number of atoms in universe is only about 10^{80}, which is a number that fits nicely on 24 cm of paper.

The internal energy is the average over the energies of all microstates:

$$U = \sum_i P_i E_i \tag{2.3}$$

where the P_i the probability of finding the system in microstate i. Instead of Eq. 2.1, it is also now more natural to use the equivalent statistical entropy definition

$$S = -k_\mathrm{B} \sum_i P_i \ln P_i \tag{2.4}$$

This relation can be used to calculate entropy more or less directly from the conformational distributions obtained in molecular simulations.

The entropy S and internal energy U of the system can be combined into the master equation that constitutes the definition of *Helmholtz's free energy*

$$A = U - TS \tag{2.5}$$

A fundamental condition in thermodynamics is that A is constant for a system at constant N, V and T (that is, A is constant for a member of a NVT-ensemble). Using this in combination with Equations 2.3, 2.4 and the obvious relation $\sum_i P_i = 1$, allows for the derivation of *Boltzmann's distribution law* for the probability of microstate i with energy E_i:

$$P_i = \frac{e^{-E_i/k_\mathrm{B}T}}{\sum_i e^{-E_i/k_\mathrm{B}T}} \tag{2.6}$$

The normalisation denominator in Eq. 2.6 is known as the (canonical) *partition function* Q. It is a function of the number of molecules (N), volume (V) and temperature (T). Hence $Q(N, V, T)$ can be written as:

$$Q\left(N, V, T\right) = \sum_i e^{-E_i/k_\mathrm{B}T} \tag{2.7}$$

Hence, by combining Eq. 2.3 and Eq. 2.6 the internal energy of the system is given as the (Boltzmann) average over all explored microstates:

$$U = \frac{\sum_i E_i e^{-E_i/k_\mathrm{B}T}}{Q} \tag{2.8}$$

In principle, even though it is very difficult in practice for most systems, it is possible to specify the system's energy levels E_i once the volume V and chemical composition (i.e. number of molecules of each species) and their mutual interactions are known.

Now, from the relations above it is easy to show that the internal energy is given by the derivative of the partition function with respect to temperature:

$$U = k_\mathrm{B}T^2 \left(\frac{\partial \ln Q}{\partial T}\right)_{V,N} \tag{2.9}$$

which in turn is related to the internal energy through the Gibbs–Helmholtz equation of thermodynamics,

$$U = -T^2 \left(\frac{\partial (A/T)}{\partial T} \right)_V \tag{2.10}$$

By comparing Eq. 2.9 and Eq. 2.10, we find that the Helmholtz's free energy can be calculated directly from the partition function as:

$$A = -k_{\mathrm{B}} T \ln Q \tag{2.11}$$

Finally, according to one of the fundamental equations in thermodynamics, the entropy of the system can then be evaluated as

$$S = -\left(\frac{\partial A}{\partial T} \right)_V = k_{\mathrm{B}} \ln Q + \frac{U}{T} \tag{2.12}$$

Of course, this expression for S can also be obtained more directly by inserting the Boltzmann distribution law (Eq. 2.6) into the entropy definition (Eq. 2.4).

In the NVT ensemble, the number of particles, volume and temperature remain constant. In practice, it is desirable to allow fluctuations of the volume so that the pressure can be constant. This ensemble is referred as the isothermal–isobaric or NPT ensemble. The derivation of the partition function for the NPT ensemble is quite similar to that for the NVT ensemble. However, in treating the NPT ensemble we have to take into account the system's volume. The free energy of the NPT ensemble is known as Gibbs free energy (G), which is similar to Helmholtz's free energy (A), except for the addition of a pressure–volume term:

$$G = U + PV - TS \tag{2.13}$$

For a classical system, we describe the accessible energies as a function of momentum ($\boldsymbol{p}$) and position ($\boldsymbol{r}$) vectors ($6N$ coordinates) [9]. Assuming these are continuous variables, we can express the partition function as:

$$Z = \int \int e^{-\beta H(\boldsymbol{r}^N, \boldsymbol{p}^N)} d\boldsymbol{r}^N d\boldsymbol{p}^N \tag{2.14}$$

where $\beta = (k_{\mathrm{B}} T)^{-1}$ and $H(\boldsymbol{r}^N, \boldsymbol{p}^N)$ is the Hamiltonian which is the sum of the kinetic and potential energy (determined by the momentum and position values). Let us consider a macroscopic equilibrium observable O (for example it can be the total energy or pressure of the system). The average value of the observable can be expressed as:

$$\langle O \rangle_{\mathrm{ensemble}} = \frac{1}{Z} \int O(\boldsymbol{r}^N, \boldsymbol{p}^N) e^{-\beta H(\boldsymbol{r}^N, \boldsymbol{p}^N)} d\boldsymbol{r}^N d\boldsymbol{p}^N \tag{2.15}$$

According the *Ergodic* hypothesis, the long time average of an observable is equal to the ensemble average:

$$\langle O \rangle_{\mathrm{ensemble}} = \langle O \rangle_{\mathrm{time}} \tag{2.16}$$

where $\langle O \rangle_{\text{time}}$ is written as:

$$\langle O \rangle_{\text{time}} = \lim_{\tau \to \infty} \frac{1}{\tau} \int_{t=0}^{\tau} O(\boldsymbol{r}^N(t), \boldsymbol{p}^N(t)) dt \approx \frac{1}{M} \sum_{k=1}^{M} O(\boldsymbol{r}_k^N, \boldsymbol{p}_k^N) \tag{2.17}$$

where M is the number of configurations.

One can use different sampling algorithms (e.g. molecular dynamics or Monte Carlo simulation) to sample the configurational space. From Eq. 2.13, one can see that the absolute Gibbs free energy of a system is related to the partition function. Theoretically, it is possible to calculate the free energy of a macroscopic system from its partition function by taking into account each accessible microstate of the system and its corresponding energy. But it will require an extensive sampling of the configurational space, which is impractical even for a small system. We are mainly interested in calculating free energy differences, not the absolute free energies of systems. The *change* in free energy between two states (A and B) can be expressed in terms of the ratio between the corresponding partition functions Z_{A} and Z_{B}:

$$\Delta_{\text{A}\to\text{B}} G = -k_{\text{B}} T \ln \frac{Z_{\text{B}}}{Z_{\text{A}}} \tag{2.18}$$

This is analogous to Eq. 2.11. If one assumes a classical system (Eq. 2.14), then the identical microstates between two macro-states (A and B) cancel out which reduces the problem to sampling the phase space that differs between A and B [9].

In the next chapters, I will introduce two sampling algorithms, molecular dynamics and metadynamics. I will also touch upon different methods to calculate free energy differences in a macro-molecular systems.

Chapter 3

Molecular Dynamics

> everything that living things do can be understood in terms of the jigglings and wigglings of atoms
>
> *Richard P. Feynman*

Molecular dynamics simulation (MD) is a sampling method that generates time-series of configurations corresponding to the thermal fluctuations of an equilibrium system. A sufficiently long MD simulation, i.e. one that samples enough of the possible conformations of a system, can be used to extract experimentally relevant information, such as kinetics, lifetime distributions, etc. However, one might ask whether it is possible for classical MD to go enough of the conformations of a biological macromolecule within reasonable time (Figure 3.1) [12, 13]? We will address this question later, but for now let us focus on MD simulations.

The first MD simulation of a simple protein folding was published in 1977 [15]. Since then, MD simulations have been routinely used to investigate structure, conformational dynamics and thermodynamics associated with biological macromolecules [16, 17, 18, 19]. This section mainly deals with some key principles behind MD simulation.

Classical MD simulations involve numerical integration of Newton's equations of motion. According to Newton's second law, the force ($\boldsymbol{F}$) acting on a particle of mass m is:

$$\boldsymbol{F} = m\boldsymbol{a} = m\frac{d\boldsymbol{v}}{dt} = m\frac{d^2\boldsymbol{r}}{dt^2} \tag{3.1}$$

where $\boldsymbol{a}$ is the acceleration, $\boldsymbol{v}$ is the velocity, and $\boldsymbol{r}$ represents the coordinates of the particle. The force can also be expressed as the gradient of the potential energy $\mathcal{V}$:

$$\boldsymbol{F} = -\boldsymbol{\nabla}\mathcal{V} \tag{3.2}$$

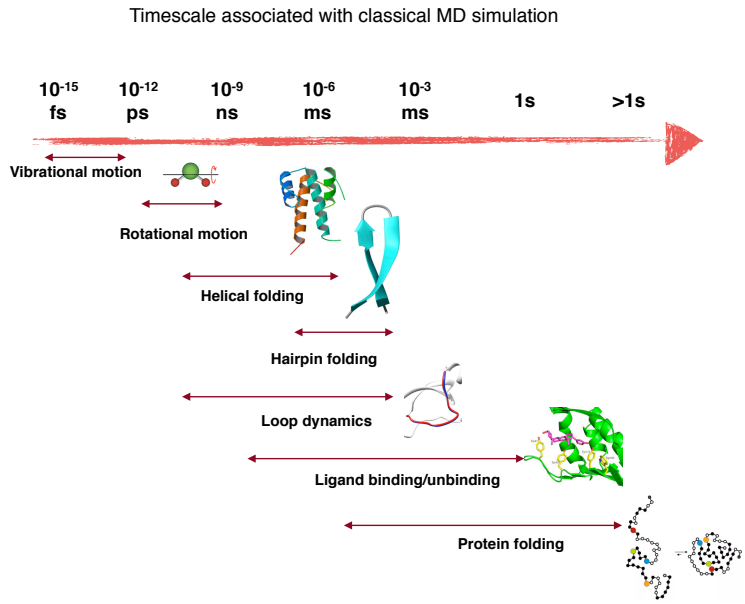


Figure 3.1: Timescale of specific biological processes that one can capture with classical MD simulation [14].

Using Eq. 3.1 and Eq. 3.2, acceleration $\boldsymbol{a}$ can be expressed in terms of the gradient of the potential energy:

$$\boldsymbol{a} = -\frac{1}{m}\boldsymbol{\nabla}\mathcal{V} \tag{3.3}$$

Hence, in order to generate a molecular dynamics trajectory one needs to know the initial position of atoms, the initial distribution of velocities and the potential energy surface. The equation of motion is deterministic, which means that positions and velocities at $t = 0$ determine the positions and velocities at some other time, t. In biomolecular simulation, the initial positions (co-ordinates) are typically obtained from experiments such as X-ray, NMR, or Cryo-EM.

3.1 Integration Algorithms

Integration algorithms assume that the position $\boldsymbol{r}$ and velocity $\boldsymbol{v}$ of an atom can be approximated by Taylor series [20]:

$$\boldsymbol{r}(t + \Delta t) = \boldsymbol{r}(t) + \boldsymbol{v}(t)\Delta t + \frac{1}{2}\boldsymbol{a}(t)\Delta t^2 + \ldots \tag{3.4a}$$

$$\boldsymbol{v}(t + \Delta t) = \boldsymbol{v}(t) + \boldsymbol{a}(t)\Delta t + \frac{1}{2}\boldsymbol{b}(t)\Delta t^2 + \ldots \tag{3.4b}$$

The choice of time-step (Δt) is critical to perform MD simulations. Often, the fastest motion is the vibrations of bonds involving hydrogen atoms. One can choose a time-step of typically 0.5 fs which will allow such vibrations. Alternatively, if the bond lengths associated with hydrogen atoms are kept fixed using some constraint algorithm (e.g. SHAKE [21] or LINCS [22], then one can use a slightly larger time-step (typically 2 fs). Over the years, several algorithms have been developed for integrating the equations of motions e.g. *verlet*, *leap-frog*, *velocity-verlet* etc [20]. As an example, I will briefly describe the leap-frog algorithm, which is a commonly used algorithm.

In case of the leap-frog algorithm, the velocities at time $t + 1/2\Delta t$ are calculated using velocities at time $t - 1/2\Delta t$ and the acceleration $\boldsymbol{a}$ at time t [23]. The positions $\boldsymbol{r}$ at time $t + \Delta t$ are calculated using positions at time t together with previously calculated velocities. The velocities leap over the positions and the positions leap over the velocities just like a frog, hence the name leap-frog (Figure 3.2). An advantage of this algorithm is that the velocities are explicitly calculated, whereas a disadvantage is that the velocities are not calculated at the same time as the positions, which compromises its precision. Thus, the leap-frog algorithm is

$$\boldsymbol{v}\left(t + \frac{1}{2}\Delta t\right) = \boldsymbol{v}\left(t - \frac{1}{2}\Delta t\right) + \boldsymbol{a}(t)\Delta t \tag{3.5a}$$

$$\boldsymbol{r}(t + \Delta t) = \boldsymbol{r}(t) + \boldsymbol{v}\left(t + \frac{1}{2}\Delta t\right)\Delta t \tag{3.5b}$$

and the velocities at time t can be approximated as:

$$\boldsymbol{v}(t) = \frac{1}{2}\left[\boldsymbol{v}\left(t - \frac{1}{2}\Delta t\right) + \boldsymbol{v}\left(t + \frac{1}{2}\Delta t\right)\right] \tag{3.5c}$$

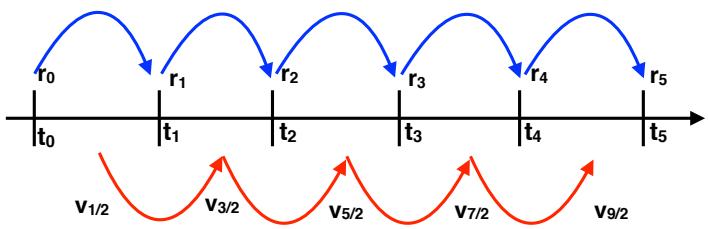


Figure 3.2: Schematic diagram of leap-frog algorithm.

3.2 Force Fields: potential energy functions

Biomolecules consist of many atoms, which makes it very difficult to study them using full quantum-mechanical calculations. Empirical potential energy functions

provide an attractive alternative that is computationally cheap compared to quantum mechanics. Potential energy functions are often referred to as force fields [24, 25, 26]. The functional form of these force fields defines the potential energy of the system. The current generation of force fields provides a reasonably good compromise between accuracy and computational efficiency.

A typical potential energy function can be divided into two terms, representing the *bonded* and *nonbonded* interactions:

$$\mathcal{V}(\boldsymbol{R}) = \mathcal{V}_{\text{bonded}}(\boldsymbol{R}) + \mathcal{V}_{\text{nonbonded}}(\boldsymbol{R}) \tag{3.6}$$

The bonded interactions comprise three terms:

$$\begin{aligned}\mathcal{V}_{\text{bonded}}(\boldsymbol{R}) = \sum_{\text{bonds}} k_b(l-l_0)^2 + \sum_{\text{angles}} k_a(\theta-\theta_0)^2 \\ + \sum_{\text{torsions}} k_\phi[1+\cos(n\phi-\gamma)]\end{aligned} \tag{3.7}$$

The three bonded terms of the potential energy represent bond stretching, angle bending and rotation around torsion angles, with l being the distance between two covalently bound atoms, θ the angle between three atoms (Figure 3.3), ϕ the torsional angle, and k_b, l_0, k_a, θ_0, k_ϕ, n, and γ fixed parameters (summation indices have been omitted for simplicity).

The nonbonded term is the sum of Lennard–Jones and Coulomb interactions between all pairs of atoms:

$$\begin{aligned}\mathcal{V}_{\text{nonbonded}}(\boldsymbol{R}) = \sum_i \sum_{j\neq i} 4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] \\ + \sum_i \sum_{j\neq i} \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}}\end{aligned} \tag{3.8}$$

The Lennard–Jones potential energy describes the exchange repulsion and dispersion attraction between all pairs of atoms i and j, with r_{ij} being the distance between two atoms and σ_{ij} and ϵ_{ij} being fixed parameters. The Coulomb interaction describes attraction or repulsion between two atoms with partial atomic charges q_i and q_j separated by distance r_{ij}.

Over the years, several force fields were developed for organic molecules, proteins, nucleic acids, lipids etc. In our study, the protein was treated using Amber FF14SB [27] and CHARMM36 [28] force fields. The organic molecules were described using GAFF [29] and OPLS [30] force fields.

The most time-consuming part in a MD simulation is to calculate the non-bonded energy terms. As can be seen in Eq. 3.8, an explicit calculation of the non-bonded energy term between every pair of atoms increases the complexity as the square of

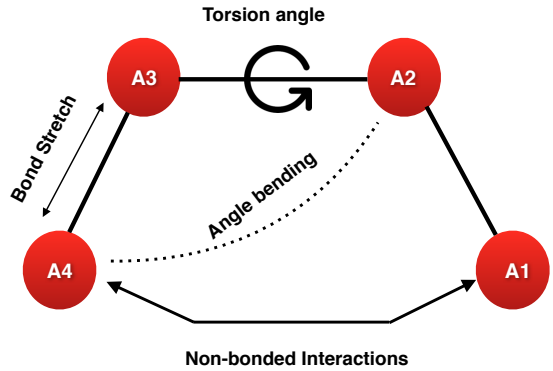


Figure 3.3: Schematic diagram of different types of interaction energies accounted in a force field.

the number of atoms (N^2). A popular strategy is to set a cutoff distance beyond which interactions are ignored. Accounting for long-range interactions just by increasing the cutoff is highly computationally demanding [24]. In recent years, several models have been developed which permit the inclusion of long-range interactions in biomolecular simulation [31]. Ewald summation is considered to be one of the better approximations to treat long-range electrostatic interactions for a periodic system. A variant of Ewald summation, known as particle-mesh Ewald has been used frequently [32, 31, 33].

3.3 Periodic boundary conditions (PBC)

Enabling periodic boundary conditions (PBC) makes it possible to run simulations on a relatively small number of particles, in such way that every particle still experiences forces as if it were in bulk solution. A central box is constructed by immersing the solute in water molecules. The box is then replicated in all directions. During simulation, if a particle drifts out of the central box it ends up in the replica box. Forces on the particle are calculated from particles within same box as well as replica boxes. The minimum image convention is used to avoid double counting[1]. The simplest box is a cube. For globular proteins, a truncated octahedron box is often preferred over the cubic box. The shape of the truncated octahedron reduces the number of water molecules that need to be simulated compared to the cubic box, which speeds up the calculation.

[1]Only the shortest distances between a pair of atoms are counted, irrespective of their position in the same box or replica box.

3.4 Water models

Water plays an important role in screening of electrostatic interactions. The *implicit* way to treat the water is to include an effective dielectric screening constant. This is a very crude approximation. In *explicit* treatment of water, the electrostatic interactions are expressed in terms of Coulomb's law and the dispersion and repulsive forces are expressed in terms of Lennard-Jones potential [34]. Figure 3.4 shows a representation of some typical water models used in MD simulations [35].

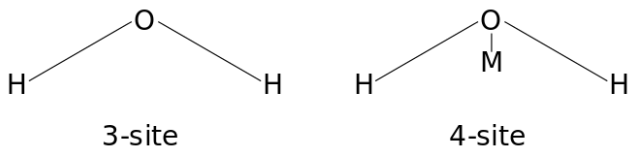


Figure 3.4: Shape of different water models used in MD simulation [36]. In 4-site water model the dummy atom, M has a negative charge to improve the electrostatic distribution.

MD simulation has found its application in several fields of science, such as biochemistry, materials science, atmospheric chemistry, solution chemistry, toxicology, etc. In this thesis, we will strictly limit our discussion to biomolecular simulations, which can be used to study conformational dynamics, protein folding, ligand binding/unbinding, effect of mutations, allosteric regulations, etc [16].

3.5 Limitations of MD simulation

In recent years we have seen a significant improvement in computational power. MD simulations have leveraged upon the development of powerful hardware to understand complex motions associated with macromolecules. However, MD simulation still suffers from several shortcomings. A couple are:

1. The timescale problem: The integration time-step of MD is usually in the order of femtoseconds (fs). However, many interesting slow conformational changes (e.g. ligand binding/unbinding, loop dynamics, protein folding/unfolding) happen in the timescale of micro/milliseconds or longer (Figure 3.1). It is not routinely feasible to sample conformational changes in the millisecond regime for any system with more than a thousand atoms [13].

   The energy landscape of a biomolecule is characterised by different metastable states that are separated by high kinetic barriers. Due to an integration time-step of a few femtoseconds for a classical MD simulation, crossing these kinetic

barriers within reasonable computational resources becomes a daunting challenge [37].

2. The accuracy of the force fields: Empirical force fields are approximations and one needs some kind of experience to know what to trust in a MD simulation [38, 39]. Moreover, classical MD simulations cannot capture formation and breaking of covalent bonds. Hence, it is impossible to study reaction mechanisms using classical MD simulation.

In the next chapter I will discuss some methods that were developed to address the first limitation, the *timescale problem*.

Chapter 4

Metadynamics: overcoming barriers

Now everybody's sampling

Missy Elliot, American Musician

The energy landscape of a biomolecule is *rugged*, meaning that it is characterised by numerous metastable basins which are separated by high kinetic barriers. Crossing a kinetic barrier to sample a metastable state is therefore a *rare event*[1], and can be inaccessible in classical MD simulations due to the timescale problem. In last two decades, several methods have been developed to accelerate the sampling of rare events. These methods are known as *enhanced sampling* methods [40, 41, 42, 43, 44]. Enhanced sampling methods can be divided into two categories: collective-variable based (such as metadynamics, umbrella sampling, steered MD) and collective-variable free methods (such as parallel-tempering MD, accelerated MD). Collective variables (CVs) or reaction co-ordinates are functions of atomic co-ordinates that differ between two or more metastable states within the configurational space [45].

Here, I will mainly discuss *metadynamics*, which is a CV-based enhanced-sampling method. Over the years, several reviews have been written on metadynamics which sums up the key concepts and applications [46, 47, 48, 43, 49].

4.1 Metadynamics

Metadynamics involves the idea of *filling the free energy minima* [50, 45] with an external bias potential. Addition of bias pushes the system away from local free energy minima that have been explored by the simulation and thus accelerates sampling of configurational space. The bias is applied along pre-defined CVs. CVs are generally low-dimensional representations of atomic coordinates. A good CV should be able to

[1] events that occur with low frequency

distinguish key metastable states along *slow* degrees of freedom [45]. The selection of a suitable CV is not trivial and is still an active area of research that I will touch upon in a later section.

In metadynamics the bias is deposited as a sum of Gaussian shaped hills. The metadynamics bias potential at time t along a set of d chosen CVs, collectively denoted by $\boldsymbol{s}$ ($\boldsymbol{s}$ is a function of atomic coordinates $\boldsymbol{R}$) can be written as:

$$V(\boldsymbol{s},t) = \sum_{k\tau<t} W(k\tau) \exp\left(-\sum_{i=1}^{d} \frac{(s_i - s_i(\boldsymbol{R}(k\tau)))^2}{2\sigma_i^2}\right) \tag{4.1}$$

where $W(k\tau)$ is the Gaussian height, τ is the Gaussian deposition stride and σ_i is the Gaussian width of the i^{th} CV. The Gaussian width is usually chosen by monitoring the fluctuation of the CV in a MD simulation.

Assume that a free energy surface (FES) is described by two local minima A and B. A metadynamics simulation starts with the system being in free energy minimum B. As time goes by, the bias is deposited in basin B which increases the underlying potential. After some time t, the system jumps out of basin B and falls into basin A. Now, the bias starts accumulating in basin A. When basin A is also filled up by bias potential, the free energy surface flattens out and the system can fluctuate freely along the flattened FES (Figure 4.1).

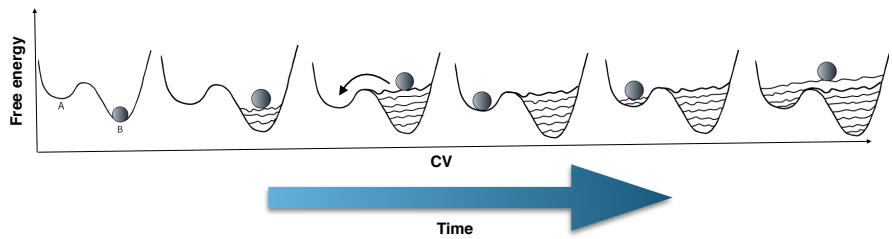


Figure 4.1: Time evolution of a typical metadynamics simulation. Basin A and B are separated by kinetic barrier. Over the time, external bias potential builds up in basin B and the system escapes to basin A. Once the bias fills up basin A, the system shows diffusive behaviour along CV space.

If the external bias potential $V(\boldsymbol{s})$ converges to a particular value, then one can estimate the underlying unbiased free energy surface from metadynamics using the following expression:

$$V(\boldsymbol{s}, t \to \infty) = G(\boldsymbol{s}) + C \tag{4.2}$$

Here, C is an irrelevant constant and $G(\boldsymbol{s})$ is the free energy surface along the CV $\boldsymbol{s}(\boldsymbol{R})$:

$$G(\boldsymbol{s}) = -\frac{1}{\beta} \ln\left(\int \delta(\boldsymbol{s} - \boldsymbol{s}(\boldsymbol{R})) e^{-\beta \mathcal{V}(\boldsymbol{R})}\, d\boldsymbol{R}\right) \tag{4.3}$$

where $\mathcal{V}(\boldsymbol{R})$ is the potential energy. In theory, at the end of a metadynamics simulation, the underlying free energy surface should be constant (Eq. 4.2). However, as

the repulsive bias potential is continuously deposited during the simulation, it really never converges but oscillates around a particular value.

In order to solve this problem, a smoothly converging variant of metadynamics known as *Well-tempered* metadynamics (WT-Metad) [51] was developed, in which the Gaussian height W decreases with increasing bias potential $V(\boldsymbol{s}, t)$:

$$W(t) = W_0 e^{-\frac{1}{\gamma - 1}\beta V(\boldsymbol{s},t)} \tag{4.4}$$

where W_0 is the initial Gaussian height and γ is the *bias factor* which can be expressed as

$$\gamma = \frac{T + \Delta T}{T} \tag{4.5}$$

T is the temperature and ΔT is an adjustable input parameter with the dimension of temperature. The choice of ΔT regulates the exploration of the free energy surface. When $\Delta T \to 0$, the simulation corresponds to a MD simulation, whereas when $\Delta T \to \infty$ it corresponds to a standard metadynamics (non well-tempered) simulation. For $\gamma > 1$ and $t \to \infty$, we have that $W(t) \to 0$ and the bias potential $V(\boldsymbol{s}, t)$ converges to:

$$V(\boldsymbol{s}, t) = -\left(1 - \frac{1}{\gamma}\right) G(\boldsymbol{s}) + c(t) \tag{4.6}$$

where $c(t)$ can be expressed as:

$$c(t) = \frac{1}{\beta} \ln \frac{\int e^{-\beta G(\boldsymbol{s})}\, d\boldsymbol{s}}{\int e^{-\beta(G(\boldsymbol{s}) + V(\boldsymbol{s},t))}\, d\boldsymbol{s}} = \frac{1}{\beta} \ln \frac{\int e^{-\frac{\gamma}{\gamma - 1}\beta V(\boldsymbol{s},t)}\, d\boldsymbol{s}}{\int e^{-\frac{1}{\gamma - 1}\beta V(\boldsymbol{s},t)}\, d\boldsymbol{s}} \tag{4.7}$$

Addition of bias in metadynamics alters the unbiased probability distribution $P(\boldsymbol{R})$. One can express the time dependent biased probability distribution, P_V as:

$$P_V(\boldsymbol{R}, t) = \frac{e^{-\beta(\mathcal{V}(\boldsymbol{R}) + V(\boldsymbol{s}(\boldsymbol{R}),t))}}{\int e^{-\beta(\mathcal{V}(\boldsymbol{R}) + V(\boldsymbol{s}(\boldsymbol{R}),t))}\, d\boldsymbol{R}} \tag{4.8}$$

One can extract the unbiased probability distribution from a biased distribution by re-weighting it according to the Boltzmann distribution law:

$$P(\boldsymbol{R}) = P_V(\boldsymbol{R}, t) e^{\beta(V(\boldsymbol{s}(\boldsymbol{R}),t) - c(t))} \tag{4.9}$$

Different algorithms have been developed to recover unbiased probability distribution from a biased simulation [52, 53, 54]. Together, they are known as *re-weighting* algorithms. The possibility to extract the unbiased probability distribution of any reaction-coordinate using re-weighting makes WT-MetaD a powerful sampling approach.

4.2 Convergence of metadynamics

System starts in a local minimum where the bias starts depositing. As the simulation progress, the bias starts to grow and the Gaussian height decreases as in Eq. 4.4. After sometime, the system escapes the local minimum and stars sampling new regions in the conformational space. When this happens the Gaussian height is readjusted to its initial value and starts decreasing again. In the long run, Gaussian height gets smaller and smaller and the system shows a diffusive behaviour in the CV space. The free energy of WT-Metad along a good CV (discussed in a later section) should converge as in Eq 4.6. Voth and co-workers demonstrated that WT-Metad converges asymptotically [55]. However, the time required for convergence cannot be predicted.

At any point in time in a metadynamics simulation, one can calculate the free energy difference between two local minima along a chosen CV as a function of simulation time (Figure 4.2). In the long run, Gaussian heights gets smaller and smaller and free energy fluctuates asymptotically. A converged free energy profile can be obtained by averaging over a time-interval where the system shows diffusive behaviour along CV space [56].

4.3 Choice of collective variable

Metadynamics is dependent on the choice of CV. Theoretically, one can use any atomic co-ordinates as CV in a metadynamics simulation. In practice, a badly chosen CV can cause irreversible changes in a system by pushing the system towards an unphysically high free energy region. A good CV should able to discriminate between different metastable states, i.e. each metastable state should correspond to different values of the CV. If this condition is violated, the system remains stuck in a local free energy minimum during the metadynamics simulation [45]. If one chooses a CV that ignores orthogonal degrees of freedom (separated by high free energy barriers), then metadynamics experiences *hysteresis*, meaning that it gets stuck in some intermediate free energy basin along orthogonal variables. For example, assume that we want to capture a protein–ligand binding process. A natural choice of CV would be the distance between the ligand and the binding site of the protein. Imagine that the entry of the ligand is occasionally blocked by the presence of a long-lived water molecule in the binding site. In this case, an ideal second CV should capture the dynamics of the water molecule in the binding site. Failure to incorporate such a second CV will create hysteresis. Thus, selection of good CVs is far from trivial. One needs some amount of prior information in order to develop an optimal set of CVs. This can be achieved by monitoring some interesting fluctuations in a MD simulation or using information from experiments.

The problem of selecting optimum CVs is associated with the complex high dimensional configurational space. One way to solve this problem is to project the

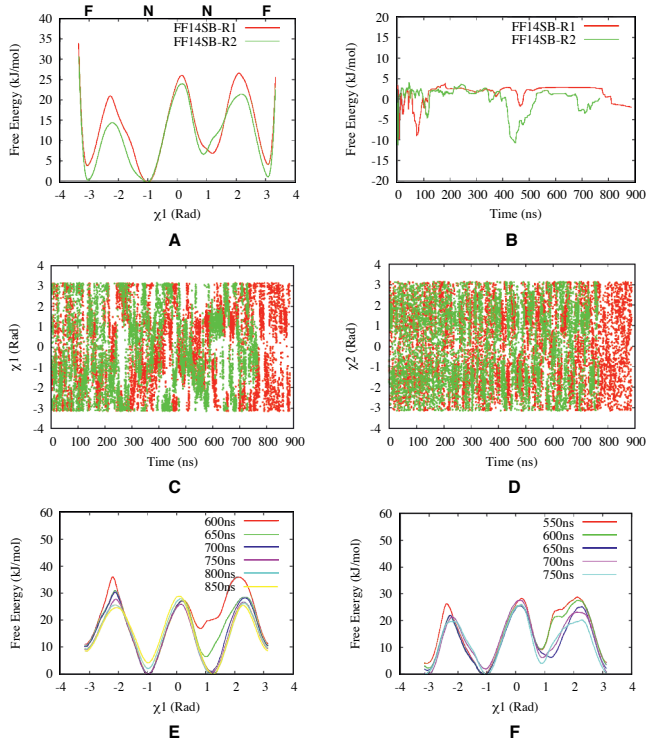


Figure 4.2: Reweighted free energy surface along χ_1 angle (A) in two independent (different initial starting velocities) WT-Metad simulations using χ_1 and χ_2 angles as CVs. Distribution of χ_1 angle centred around $\pm\frac{\pi}{3}$ radian is denoted as normal, whereas distribution centred around $\pm\pi$ radian is denoted as flipped). The first step to check convergence is to calculate free energy difference between normal (N) and flipped (F) states along χ_1 as a function of simulation time (B). One can see that the free energy is fluctuating around an average value for the two independent metadynamics runs. Sampling of χ_1 and χ_2 angles during metadynamics simulations shows diffusive behaviour along CV space (C and D). Free energy profile of χ_1 as a function of simulation time from the last ∼200 ns of metadynamics simulation (E and F). In the last part, the free energy profiles look similar, apart from a constant offset. Using all these observations, we can say that these two independent metadynamics simulations reached convergence.

high dimensional space onto a low-dimensional sub-space (i.e. defined by few eigenvectors), using dimensionality reduction and machine learning algorithms [57, 58]. Dimensionality reduction methods such as principal component analysis (PCA) and time-lagged independent component analysis (tICA) have been used to generate CVs for metadynamics.

PCA does a maximal variance projection of the high-dimensional data onto a low-dimensional subspace. The orthogonal axes (principal components) of the low-dimensional subspace represents directions of maximum variance. In contrast to PCA, tICA captures high autocorrelation linear combinations of the high-dimensional data. The directions of maximal autocorrelation are referred to as time-lagged independent components (tICs). Let's say that in a MD simulation the loop region of a small protein remains highly flexible and the helix region undergoes a rare transition due

to rotation of side chains. In this case, the motion with high variance (loop motion) will be captured by the first few principal components. On the other hand, the rotation of side-chains, which is the motion with high autocorrelation, will be captured by the first few tICs. Because of their ability to capture conformational motions in biomoecules (Figure 4.3), PCs and tICs are used as CVs in metadynamics [59, 60, 61].

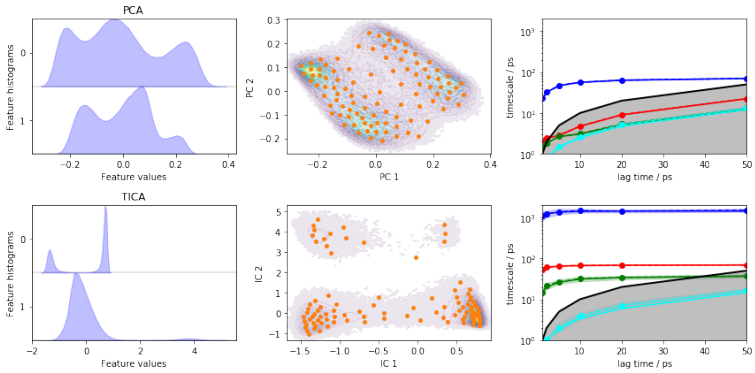


Figure 4.3: PCA and tICA performed on the pairwise distance between heavy atoms in a MD simulation of alanine dipeptide. Both PCA and tICA yields projections with some defined basins. However, PCA resolves only one slow process (see the timescale plot) whereas, tICA captured three slow process in MD simulation (source: PyEMMA [62] tutorial)

4.4 Reconnaissance metadynamics

Incorporation of many low-dimensional CVs in metadynamics remains a challenge. Reconnaissance metadynamics (Recon-Metad) leverages upon dimensionality reduction (PCA) and clustering (Gaussian mixture model) algorithms in order to be effective with a larger number of CVs [63]. In Recon-Metad, the bias potential is deposited along a mixture of basins which are a low-dimensional representation of the underlying high-dimensional FES. The basins are identified dynamically at regular intervals using a combination of PCA and the Gaussian mixture clustering algorithm [64, 65]. The biasing leads to escape from the already sampled basins and exploration of new areas of conformational space. Recon-Metad has been used mainly for prediction of binding poses [66, 67] and sampling the conformational space of small proteins [63].

4.5 Replica-exchange with metadynamics

Replica-exchange MD (REMD) is a CV free method for enhanced sampling, where the system is accelerated by modifying the original Hamiltonian of the system. One of the most popular variants of REMD is parallel tempering [68]. In parallel tempering (PT), several replicas are simulated with the same potential energy function but at

different temperatures. During the simulation, exchange of configurations between two neighbouring replicas (Figure 4.4) are attempted using the following acceptance probability:

$$p(i \to j) = \min\left\{1, e^{\Delta^{\mathrm{PT}}_{i,j}}\right\} \tag{4.10}$$

$\Delta^{\mathrm{PT}}_{i,j}$ is written as:

$$\Delta^{\mathrm{PT}}_{i,j} = \left(\frac{1}{k_{\mathrm{B}}T_i} - \frac{1}{k_{\mathrm{B}}T_j}\right)\left(\mathcal{V}(\boldsymbol{R}_i) - \mathcal{V}(\boldsymbol{R}_j)\right) \tag{4.11}$$

where $\boldsymbol{R}_i$ and $\boldsymbol{R}_j$ are the coordinates of two replicas at temperatures T_i and T_j respectively and $\mathcal{V}(\boldsymbol{R}_i)$ and $\mathcal{V}(\boldsymbol{R}_j)$ are the potential energies of the two replicas i and j. The efficiency of exchange depends on how much overlap there is between the potential energy distributions of the replicas.

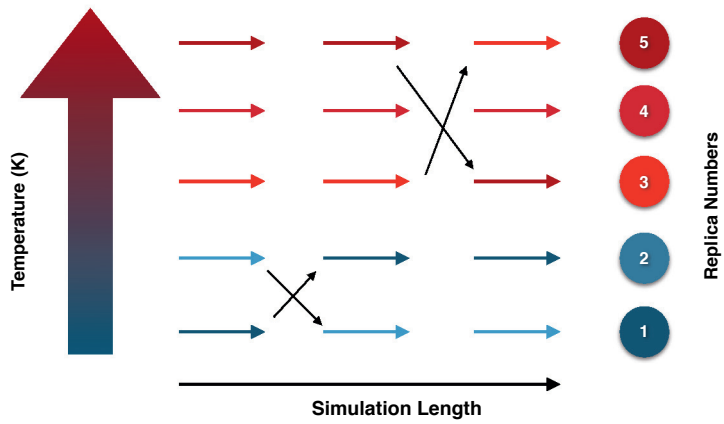


Figure 4.4: Schematics diagram of parallel tempering. The black arrows describe the exchange process between replicas.

One can easily combine a CV based method such as metadynamics with PT. The resulting PT-MetaD algorithm has the following modified acceptance probability which takes into account the metadynamics bias:

$$\begin{aligned}\Delta^{\mathrm{PTMetaD}}_{i,j} = \Delta^{\mathrm{PT}}_{i,j} &+ \frac{1}{k_{\mathrm{B}}T_i}\left[V^i_G(s(R_i),t) - V^i_G(s(R_j),t)\right] \\ &+ \frac{1}{k_{\mathrm{B}}T_j}\left[V^j_G(s(R_j),t) - V^j_G(s(R_i),t)\right]\end{aligned}$$

where V^i_G and V^j_G are metadynamics bias potentials acting on the i^{th} and j^{th} replicas, respectively. The effect of neglecting slow degrees of freedom in metadynamics (due to a limited number of CVs) can be compensated by PT, which increases the

probability to cross moderate to high free energy barriers along all degrees of freedom [69]. In my study, I used a variant of PT-Metad that enhances fluctuation within a well-tempered ensemble (WTE). In the WTE, bias is applied to the system's potential energy, which increases the fluctuations while keeping the average energy close to that of the canonical ensemble [70]. It increases the overlap between the potential energy distributions of neighboring replicas, so that fewer replicas are needed.

PT-Metad scales poorly with system size. Hence, running PT-Metad for a big system is highly computationally demanding.

Chapter 5

Molecular recognition

> our cells engage in protein production, and many of those proteins are enzymes responsible for the chemistry of life
>
> *Randy Schekman*

Molecular recognition is a process by which two or more molecules bind to each other through non-covalent interactions. In this thesis, I mainly focus on protein–ligand binding. Binding of a protein P and ligand L forms a protein–ligand PL complex. The binding process can be expressed as:

$$\mathrm{P} + \mathrm{L} \underset{k_{\mathrm{off}}}{\overset{k_{\mathrm{on}}}{\rightleftharpoons}} \mathrm{PL} \tag{5.1}$$

where k_{on} and k_{off} are the binding and unbinding rate constants. If [PL], [L] and [P] denote the equilibrium concentrations of the protein–ligand complex, the protein and the free ligand, respectively, the binding constant K_{b} is defined as:

$$K_{\mathrm{b}} = \frac{k_{\mathrm{on}}}{k_{\mathrm{off}}} = \frac{[\mathrm{PL}]}{[\mathrm{P}][\mathrm{L}]} = \frac{1}{K_{\mathrm{d}}} \tag{5.2}$$

where K_{d} is the dissociation constant. The Gibbs free energy of binding, ΔG_b can be written as a function of binding constant, K_b:[1]

$$\Delta_{\mathrm{b}} G = -RT \ln K_{\mathrm{b}} \tag{5.3}$$

[1]Standard concentration is implicitly assumed in all these equations. Hence, I somewhat sloppy omit the $^{\ominus}$-symbol that should be present at all standard state differences.

where T is the temperature and R is the gas constant. A more negative free energy corresponds to more favourable binding. $\Delta_{\mathrm{b}}G$ is further divided into two components, the enthalpy $\Delta_{\mathrm{b}}H$ and entropy $\Delta_{\mathrm{b}}S$ of binding:

$$\Delta_{\mathrm{b}}G = \Delta_{\mathrm{b}}H - T\Delta_{\mathrm{b}}S \tag{5.4}$$

The enthalpic part mainly depends on the strength of interactions between the protein and the ligand. These contributions include hydrogen bonds, electrostatic interactions, ionic interactions, van-der Waals interactions etc [71, 72]. However, there might also be significant contributions from solvation processes. The binding entropy $\Delta_{\mathrm{b}}S$ can be decomposed into three terms:

$$\Delta_{\mathrm{b}}S = \Delta S_{\mathrm{solv}} + \Delta S_{\mathrm{conf}} + \Delta S_{\mathrm{r/t}} \tag{5.5}$$

where ΔS_{solv} is the change in solvent entropy upon ligand binding, mainly due to release of tightly-bound/buried water molecules. ΔS_{conf} is the change in conformational degrees of freedom of protein and ligand upon binding. $\Delta S_{\mathrm{r/t}}$ is the change in translational and rotational degrees of freedom for both protein and ligand upon binding.

Over the years, several computational methods have been developed to predict protein–ligand binding-free energy [73, 74]. These methods can be roughly divided into three categories: (i) pathway methods, which involve rigorous free-energy paths and thus would in principle give the exact result if the force field was perfect and the sampling sufficient, (ii) endpoint methods, which are also based on extensive sampling, but with an approximate statistical-mechanical treatment that only considers the end-states, and (iii) molecular docking methods that are based on empirical free-energy expressions that are faster to evaluate.

5.1 Alchemical transformation

The most commonly used pathway methods are based on so-called *alchemical* transformations and typically involve the calculation of a relative binding free energy, $\Delta\Delta_{\mathrm{b}}G$ between two similar ligands (A and B). This process can be visualised as a thermodynamic cycle as in Figure 5.1. The relative binding free energy can be written as:

$$\Delta\Delta_{\mathrm{b}}G = \Delta_{\mathrm{b}}G^{\mathrm{B}} - \Delta_{\mathrm{b}}G^{\mathrm{A}} = \Delta G^{\mathrm{A}\to\mathrm{B}}_{\mathrm{bound}} - \Delta G^{\mathrm{A}\to\mathrm{B}}_{\mathrm{solv}} \tag{5.6}$$

The idea is to calculate the free energy along the vertical lines, i.e. to alchemically transform one ligand into the other in the bound ($\Delta G_{\mathrm{bound}}$) and solvated state ($\Delta G_{\mathrm{solv}}$), respectively. The protein without ligand does not need to be simulated, which facilitates convergence. Each transformation works by dividing the path between the end states into a series of intermediate, unphysical states, in which one ligand is changed

into the other by turning off the interactions of one ligand with the surroundings, while turning on the interactions of the other ligand with the surroundings [75]. After extensive sampling of all intermediate states, the free energy can be calculated by free energy perturbation [76], thermodynamic integration [77], or Bennett acceptance ratio (BAR) [78].

Relative binding free energy calculations are more efficient when two ligands are similar to each other. However, a slight modification in ligand structure can make a big change in its binding mode. Knowledge of the binding mode is necessary for a reliable estimation of the free energy.

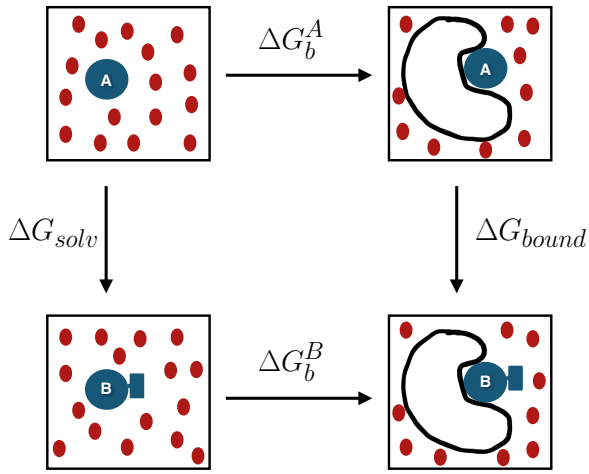


Figure 5.1: Pictorial representation of a thermodynamic cycle for relative binding free energy, $\Delta\Delta G_b$ between two ligand A and B. Explicit waters are indicated as red circles.

5.2 Funnel metadynamics

Direct calculation of the binding free energy along the horizontal lines in Figure 5.1 is more computationally expensive due to the large difference between the end states. The binding of a ligand to a protein is a slow process where changes in solvation plays a key role. In practice, sampling along horizontal lines needs to be accelerated by biasing along carefully chosen reaction co-ordinates that promote frequent binding/unbinding.

Funnel metadynamics is an enhanced sampling method that aims at accurate estimation of binding free energy by sampling along the ligand binding path (thus, it is also a pathway method). In regular well-tempered metadynamics, one can choose a CV such as the distance between binding site of the protein and ligand that facilitates ligand binding/unbinding. As soon as the ligand leaves the binding site, it starts

sampling all possible conformations in the solvated state, which takes a long time to converge. Funnel metadynamics facilitates frequent binding/unbinding by using a funnel like restraint potential (Figure 5.2) that reduces the sampling of the unbound state [79]. The effect of the restraint potential can be rigorously taken into account and the free energy difference between the bound and unbound state, $\Delta_{\mathrm{b}}G$ can be written in terms of one-dimensional PMF $w(z)$:

$$e^{-\beta\Delta_{\mathrm{b}}G} = C^{\ominus} S_u e^{-\beta\Delta G_{\mathrm{site}}} \int_{\mathrm{site}} e^{-\beta[w(z)-w_{\mathrm{ref}}]} dz \tag{5.7}$$

where $C^{\ominus} = 1/1.660\,\text{Å}^{-3}$ is the standard concentration, S_u is the cross-section of the funnel cylinder, ΔG_{site} is the change in the free energy for restraining the bound ligand. w_{ref} corresponds to the reference value of the PMF in the unbound state, which in practice is calculated by taking the average of $w(z)$ over some chosen interval along z. The radius of the cylindrical section of the funnel should be such that it doesn't affect the natural fluctuation of the ligand in the binding site. In that case, $\Delta G_{\mathrm{site}} = 0$. A large radius increases the sampling of the unbound state, whereas a small radius affects the equilibrium dynamics of the binding state. In practice, test calculations with a few different choices are performed and the time-evolution of ligand binding/unbinding (along the z axis) is monitored to select an optimal radius. Funnel metadynamics assumes that one has previous knowledge of the binding site. However, in principle it does not require *a priori* information about the binding mode of the ligand.

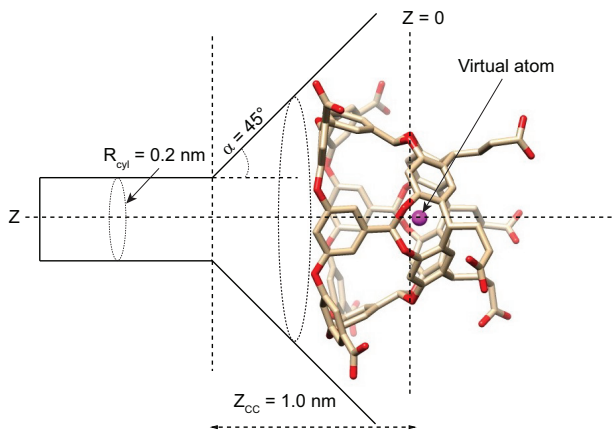


Figure 5.2: Pictorial representation of a model funnel potential used to calculate binding free energy in a host-guest system [80]. α defines the amplitude of the cone, R_{cyl} defines the radius of the cylindrical section, Z is the axis defined to study binding/unbinding and Z_{CC} is the distance where the potential switches from cone to cylindrical shape.

A typical CV in funnel metadynamics is the distance between the heavy atoms of the binding site and the ligand. However, the binding process may involve other slow

degrees of freedom such as rotation of the ligand, conformational dynamics of the protein, desolvation of buried water from the binding site, etc. Ignoring orthogonal slow degrees of freedom introduces *hysteresis*, which prevents frequent sampling of unbinding/binding and makes the simulation take an infinitely long time to converge [80].

5.3 MM/PBSA

End-point methods sample the protein–ligand complex as well as the protein and ligand in the unbound states, and calculate the free energy difference in an approximate way by taking the difference between absolute free energies corresponding to these states. One of the most popular end-point methods is molecular mechanics Poisson–Boltzmann surface area (MM/PBSA) [81]. In MM/PBSA method, the binding free energy of a protein–ligand complex is written as:

$$\Delta_{\mathrm{b}}G = \Delta E_{\mathrm{MM}} + \Delta G_{\mathrm{sol}} - T\Delta S_{\mathrm{conf}} \tag{5.8}$$

ΔE_{MM}, ΔG_{sol} and $T\Delta S_{\mathrm{conf}}$ corresponds to changes in gas-phase molecular mechanics energy, solvation free energy and conformational entropy upon ligand binding, respectively.

The individual components of Eq. 5.8 can be further expanded as follows:

$$\Delta E_{\mathrm{MM}} = \Delta E_{\mathrm{int}} + \Delta E_{\mathrm{elec}} + \Delta E_{\mathrm{vdW}} \tag{5.9}$$

where ΔE_{int}, ΔE_{elec} and ΔE_{vdW} are the changes in internal (bond angles and torsion angles), electrostatic and van-der Waals energy, respectively. Furthermore,

$$\Delta G_{\mathrm{sol}} = \Delta G_{\mathrm{PB}} + \Delta G_{\mathrm{SA}} \tag{5.10}$$

where ΔG_{PB} and ΔG_{SA} denotes the polar and non-polar contributions respectively. The polar contribution is approximated by the Poisson–Boltzmann (PB) method, whereas the non-polar contribution is estimated from the solvent accessible surface area (SASA):

$$\Delta G_{\mathrm{SA}} = \gamma \cdot \mathrm{SASA} + b \tag{5.11}$$

where γ and b are empirical parameters.

The change in conformational entropy (Eq. 5.8) is estimated using Normal Mode Analysis (NMA). In practice, MM/PBSA analysis between similar complexes often ignores the entropic term.

Each individual energy term in the previous equations is evaluated as an average over snapshots along the MD trajectory. In principle, MM/PBSA requires independent MD simulations for the protein, ligand, and protein–ligand complex, which is computationally demanding. In practice, one usually makes the approximation that

no conformational changes happen upon binding, so that snapshots of all three species can be obtained from a single MD simulation of the protein–ligand complex. The main advantage of this approach is that the simulations can be much shorter, because the calculation of averages converges much faster due to error cancellation. On the other hand, it ignores changes in the conformation of ligand and protein upon binding. It also ignores entropy of water molecules in the binding site before and after ligand binding [81, 82]. Ignoring these contributions leads to larger error and poor reliability.

MM/PBSA also suffers from convergence problem. MM/PBSA analysis of a single long MD simulation underestimates the statistical error in the result. One needs to perform many independent simulations (with different starting velocities) using the same starting structure in order to generate reliable precision [83]. The performance of MM/PBSA is highly dependent on the studied system. MM/PBSA has been used in conjunction with docking to evaluate docking poses, refine docking scores and determine structural stability.

5.4 Molecular docking

Docking is a computational method which uses a combination of scoring functions and search algorithms to predict binding modes and affinities of ligands [84, 85, 86, 87, 88]. Conformational degrees of freedom associated with ligand and protein sidechains (in the binding site) make the pose prediction a conformational search problem. Molecular docking uses *search algorithms* to perform the conformational search. The majority of the search algorithms deal with ligand flexibility, whereas a few algorithms have been developed to consider flexibility of binding site residues within a framework known as *flexible docking*. During the conformational search, the algorithm generates several conformations (poses) of the ligand. The poses are ranked based on binding free energy. Lower the binding free energy better the pose.

In docking, binding free energy calculations are carried out by simplified energy functions known as *scoring functions*. There are three main types of scoring functions: force-field based, empirical and knowledge-based [89, 90]. A typical force-field based scoring function computes the enthalpic contribution to the binding as a sum of van der Waals and electrostatic interactions. The solvation effect is approximated by implicit solvent models [91]. Entropy plays a key role in protein–ligand binding. However, computing the entropic contribution is time consuming. Incorporation of entropy in the scoring function is still a challenge in molecular docking. Efforts have been made in order to incorporate the conformational entropy of ligand using a clustering approach [92].

Approximations introduced in scoring functions and conformational search algorithms affect the accuracy of docking outcomes. However, it provides a computationally cheap way to screen large-scale ligand libraries against a particular protein

(known as *virtual screening*). Docking has been routinely used to predict binding poses of ligands when the experimental structure of the complex is unknown. Often, protein–ligand complexes generated by docking are subjected to MD simulations and end-point free-energy calculations (such as MM/PBSA), in order to check the stability of the complex and improve the free energy estimation [93]. The combination of docking, MD simulation and enhanced sampling in order to predict the binding pose and free energy of binding is an attractive area of research [94].

Chapter 6

Summary of the papers

6.1 Paper I

In this paper, we have used funnel metadynamics and MM/PBSA to predict binding free energies between a set of six guest molecules and two octa-acid hosts (OAH and OAMe) (Figure 6.1).

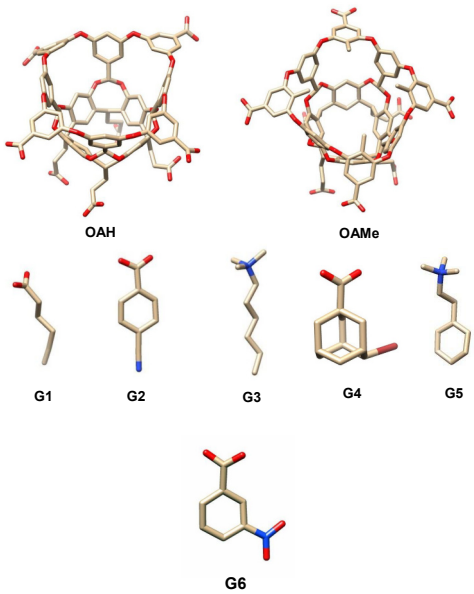


Figure 6.1: Structural representation of two octa-acid hosts, OAH and OAMe along with six guest molecules, G1-G6 used in our study [80].

Funnel metadynamics was performed using the distance between host-guest and orientation of guest molecule as CVs. The binding of the guest molecule was hin-

dered by the presence of water molecule trapped inside the host's cavity. This created *hysteresis*, which prevented frequent sampling of the binding event. We introduced a restraint potential that prevented the water molecules from getting trapped in the binding site of the host. The effect of the restraint potential was rigorously calculated by free-energy perturbation. For OAH and OAMe systems, our predicted relative binding free energies from funnel metadynamics agreed well with experimental results (Figure 6.3). However, we observed poor convergence of the funnel metadynamics in the case of OAMe-G4 (Figure 6.2). The bulky G4 guest had difficulty finding its way back into the binding pocket of OAMe during funnel metadynamics. The convergence can be improved by choosing an auxiliary CV that takes into account rotational degrees of freedom of this bulky ligand.

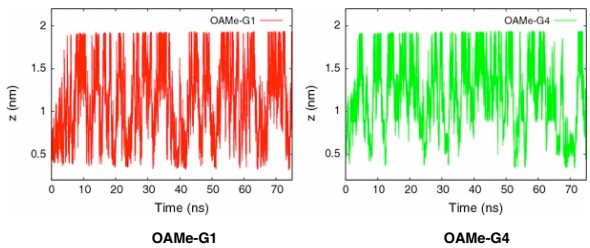


Figure 6.2: Fluctuation of binding/unbinding event (projection of ligand on z axis) during well-tempered funnel metadynamics with OAMe-G1 and OAMe-G4 complex using GAFF force field. The bound state was defined as: $z = 0.5 - 0.6nm.$, whereas for unbound state $z > 1.3nm$. OAMe-G4 shows fewer transitions between unbound and bound states compared to OAMe-G1. All other complexes with OAMe resembles the fluctuation of OAMe-G1.

We also performed MM/PBSA analysis on the host-guest systems. No significant correlation was observed between MM/PBSA and the experiment. We concluded that the approximations used in the MM/PBSA were not accurate enough to calculate differences in binding free energy among various guest molecules.

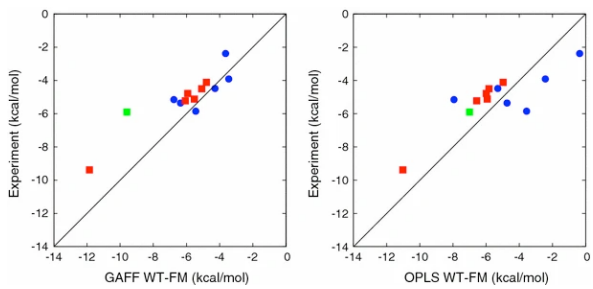


Figure 6.3: Correlation in binding free energy between experiment and funnel metadynamics using GAFF and OPLS force-fields. Blue, red and green represent OAMe, OAH and OAH-Gu2 complex respectively [80].

6.2 Paper II

In this paper, we predicted the binding pose of 35 ligands with *farnesoid X receptor* (FXR), using a combination of molecular docking, molecular dynamics and reconnaissance metadynamics. We docked all ligands against all available crystal structures of FXR (apo and another 18 crystal structures retrieved from PDB) using AutoDock Vina. For each ligand, the top predicted docking pose (the prediction with the best score among all included crystal structures) was used to perform MD simulations. The end-point of the MD simulation was then used as a starting point for the reconnaissance metadynamics (Recon-Metad) simulation. An RMSD less than 2 Å towards the crystal structure (released by D3R team) was used to define the correct pose.

Our approach using multiple crystal structures for docking allowed us to find the correct binding poses for 21 out of 35 ligands. Our submission was one of the most successful submissions to the D3R Grand challenge 2 (Figure 6.4). For 8 of these ligands, the correct pose was not the top pose. Inclusion of experimental protein structures (crystal structures released after the submission deadline) allowed us to predict the correct binding poses for 29 ligands. The docked conformations (both correctly docked and mis-docked) remained stable during the following MD simulations. MD simulations did not provide any significant improvement in coordinate RMSD versus the experimental structure. Using Recon-Metad simulations we explored new binding poses for the ligands. However, if started from a mis-docked pose, Recon-MetaD simulations failed to predict the correct pose in all cases. In Recon-MetaD, the bias was applied on the dihedrals of the ligand, which did not promote the rotation of the ligand with respect to the protein. Protein side chains present in the binding site also sterically hindered the ligand exploration. We hypothesised that the incorporation of such CVs will improve docking poses using Recon-MetaD simulations.

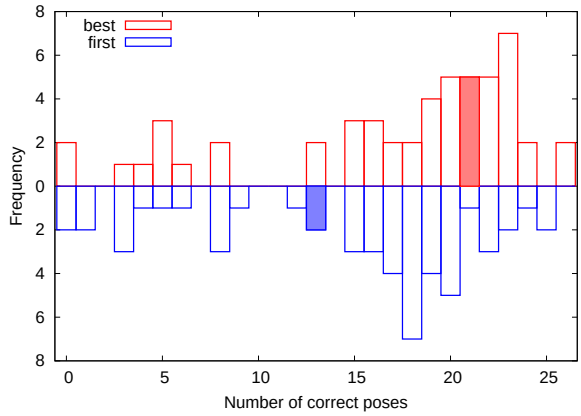


Figure 6.4: Performance of the submissions in D3R Grand Challenge 2. The upper panel of the histogram shows the distribution of the number of correctly predicted ligand poses (out of 35) over all the submissions in D3R challenge. Whereas, the lower panel of the histogram shows the results when only the first (top-predicted) pose was considered. The filled part shows the performance of our docking result [95].

6.3 Paper III

In paper III, we used a combination of molecular dynamics and metadynamics to understand flap dynamics of two pepsin-like aspatic proteases, Plm-II and BACE-1. Previous computational studies suggested that a tyrosine residue present in the flap region governs different flap conformations.

A χ_1 angle distribution of the tyrosine centred around $+\frac{\pi}{3}$ radian or $-\frac{\pi}{3}$ radian is denoted as *normal*, whereas a distribution centred around $\pm\pi$ radian is denoted as *flipped* (Figure 6.5).

We performed independent (different starting velocities) MD simulations on apo Plm-II and BACE-1. The starting conformations of BACE-1 differ in terms of the tyrosine orientation. MD simulations of apo Plm-II showed a tendency for the tyrosine to remain stuck in the normal state (Figure 6.6). On the other hand, simulations starting with BACE-1 remained trapped either in the normal or flipped state (Figure 6.6). Metadynamics using torsional angles (χ_1 and χ_2) as CVs sampled the transition between the normal and flipped states. The free energy surface reweighted along the torsional angles showed that the flap remains in a dynamic equilibrium between the normal and flipped states (Figure 6.6). Hydrogen bond interactions between the tyrosine and neighbouring residues, Trp (tryptophan) and Asp (aspartate), were predicted to be the dominant interactions that stabilise these states (Figure 6.5). Both MD and metadynamics simulations sampled spontaneous flap opening in Plm-II and BACE-1.

Mutation of the tyrosine to alanine resulted in a complete flap collapse in Plm-II and BACE-1. This is in accordance with previous experimental studies which showed that mutation to alanine resulted in loss of activity in pepsin-like aspartic proteases.

Most of the pepsin-like aspartic proteases possess conserved tyrosine residue in

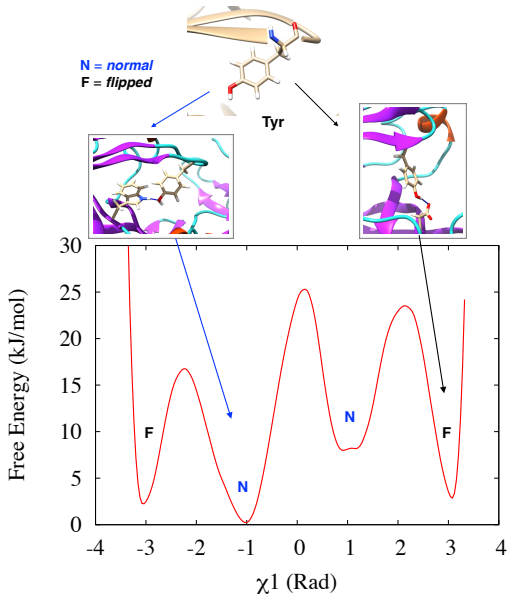


Figure 6.5: Reweighted free energy surface projected along χ_1 angle of tyrosine in case of metadynamics simulation starting with apo Plm-II. Basins corresponds to normal and flipped states were also highlighted. H-bond interactions with Trp and Asp are the key interactions stabilising these two states.

their flap region. Using observations from our study combined with previous experimental calculations, we predicted that the flap dynamics in pepsin-like aspartic proteases is governed by the rotation of the tyrosine side chain.

6.4 Paper IV

In paper IV, we attempted to understand the role of local fluctuation in hydrogen exchange (HX) of backbone amides with solvent. The core of a protein is held together by H-bond interactions between backbone amide (NH) and neighbouring residues. Local fluctuations in a protein break the H-bond interactions and allows solvent penetration. This results in HX between NH and hydrogen atoms of solvent water molecules [96]. The free-energy difference between exchange competent (open) conformation and the dominant (closed) conformation can be calculated from an MD simulation by counting the number of conformations belonging to the open (O) and closed (C) state, respectively [97]. We hypothesised that the O state, more precisely an exchange compatible conformation (ECC) is a rare fluctuation within the metastable broken (B) state (Figure 6.7). In the B state, backbone H-bond interaction involving amide is broken which leads to influx of at least one water molecule close to the amide hydrogen.

In the millisecond MD trajectory of BPTI, the β hairpin region remained stable.

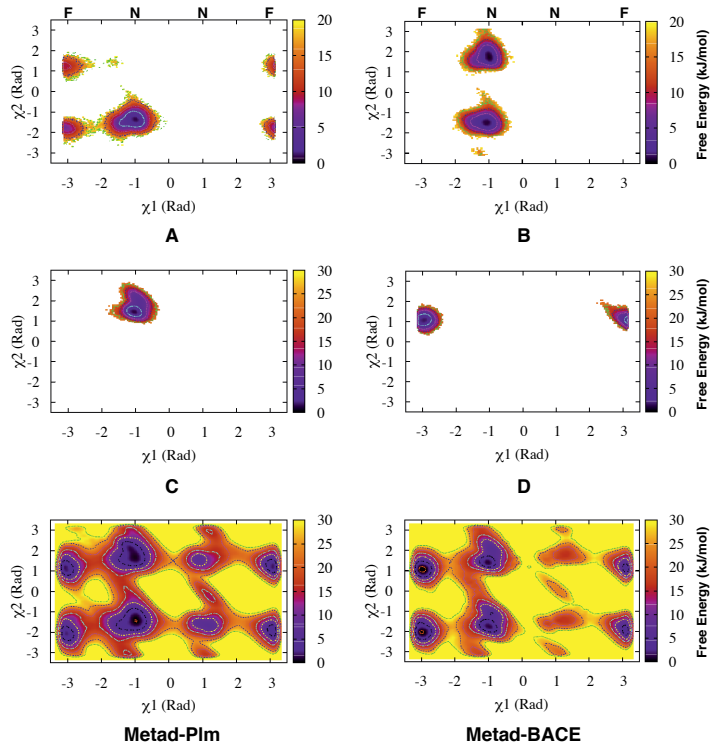


Figure 6.6: Sampling of normal (N) and flipped (F) conformations in MD simulations using Plm-II (A-B) and BACE-1 (C-D). Reweighted free energy surface projected on χ_1 and χ_2 shows that the flap remains in a dynamic equilibrium between normal and flipped states in metadynamics simulations. Here, we presented a few representative free energy surfaces from MD and metadynamics simulations.

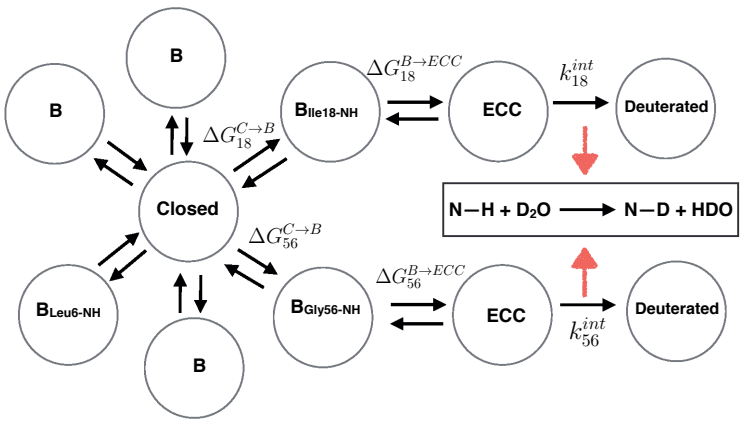


Figure 6.7: Pictorial representation of different conformational states of BPTI. In closed state backbone amide remains H-bonded with neighbouring residue. Breaking of H-bond interaction leads to formation of metastable B state. In B state, amide hydrogen forms H-bond interactions with solvent water molecules.

Hence, the amides located in this region didn't sample O conformation [97]. However, one amide ($Ile18$) located at the end part of the hairpin region, accessed the open state due to local fluctuations involving neighbouring loop (residues $11-19$ and $34-40$). This is a typical example which connects local fluctuation in a protein with solvent penetration, resulting in HX. In this study, we mainly focused on Ile18. However, we also provided a general overview of the dynamics of other residues in context of HX.

In this study, we analysed millisecond long simulation of BPTI provided by D. E shaw group [98]. Further, we performed enhanced sampling calculations (well-tempered metadynamics and PT metadynamics) and several short MD simulations starting with broken conformations of Ile18. Our hypothesis that local fluctuations in protein defines the metastable state which is responsible for HX can be seen from Figure 6.8. Looking at the free energy surface we can say that the time-independent component (tIC1) able to capture the transition between broken and closed state. Breaking of H-bond interaction leads to water penetration which can be seen from the distribution of FW in Figure 6.8. We also performed similar calculations for other amides e.g. Gly36 and Met52. Observations from our study validates the hypothesis that the local fluctuations in a protein defines the metastable states which is responsible for HX in most residues of BPTI.

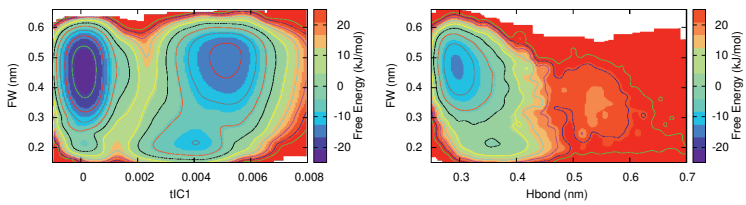


Figure 6.8: 2D FES for Ile18 calculated from PT metadynamics simulations. Left: tIC1 plotted against first water; right: H-bond distance plotted against first water. tIC1 was used as a measure to capture local fluctuation of the loop.

Bibliography

[1] M. Greshko, "These are the top 20 scientific discoveries of the decade," *National Geographic*, 2019.

[2] F. Arute, K. Arya, R. Babbush, D. Bacon, J. C. Bardin, R. Barends, R. Biswas, S. Boixo, F. G. S. L. Brandao, D. A. Buell, B. Burkett, Y. Chen, Z. Chen, B. Chiaro, R. Collins, W. Courtney, A. Dunsworth, E. Farhi, B. Foxen, A. Fowler, C. Gidney, M. Giustina, R. Graff, K. Guerin, S. Habegger, M. P. Harrigan, M. J. Hartmann, A. Ho, M. Hoffmann, T. Huang, T. S. Humble, S. V. Isakov, E. Jeffrey, Z. Jiang, D. Kafri, K. Kechedzhi, J. Kelly, P. V. Klimov, S. Knysh, A. Korotkov, F. Kostritsa, D. Landhuis, M. Lindmark, E. Lucero, D. Lyakh, S. Mandrà, J. R. McClean, M. McEwen, A. Megrant, X. Mi, K. Michielsen, M. Mohseni, J. Mutus, O. Naaman, M. Neeley, C. Neill, M. Y. Niu, E. Ostby, A. Petukhov, J. C. Platt, C. Quintana, E. G. Rieffel, P. Roushan, N. C. Rubin, D. Sank, K. J. Satzinger, V. Smelyanskiy, K. J. Sung, M. D. Trevithick, A. Vainsencher, B. Villalonga, T. White, Z. J. Yao, P. Yeh, A. Zalcman, H. Neven, and J. M. Martinis, "Quantum supremacy using a programmable superconducting processor," *Nature*, vol. 574, no. 7779, pp. 505–510, 2019.

[3] L. E. Orgel, "The origin of life–a review of facts and speculations," *Trends in Biochemical Sciences*, vol. 23, pp. 491–495, 2020/02/12 1998.

[4] I. Fry, "The origins of research into the origins of life," *Endeavour*, vol. 30, no. 1, pp. 24 – 28, 2006.

[5] A. Lazcano, "Alexandr i. oparin and the origin of life: A historical reassessment of the heterotrophic theory," *Journal of Molecular Evolution*, vol. 83, no. 5, pp. 214–222, 2016.

[6] S. Tirard, "J. b. s. haldane and the origin of life," *Journal of Genetics*, vol. 96, no. 5, pp. 735–739, 2017.

[7] J. Pereto, "Out of fuzzy chemistry: from prebiotic chemistry to metabolic networks," *Chem. Soc. Rev.*, vol. 41, pp. 5394–5403, 2012.

[8] B. Widom, *Statistical Mechanics: A Concise Introduction for Chemists*. Cambridge University Press, 2002.

[9] J. Wereszczynski and J. A. McCammon, "Statistical mechanics and molecular dynamics in evaluating thermodynamic properties of biomolecular recognition," *Quarterly Reviews of Biophysics*, vol. 45, no. 1, p. 1–25, 2012.

[10] X. Han, "Chapter 11 - mechanism of nanomachining semiconductor and ceramic blades for surgical applications," in *Engineering of Nanobiomaterials* (A. M. Grumezescu, ed.), pp. 329 – 358, William Andrew Publishing, 2016.

[11] Wikipedia contributors, "Statistical ensemble (mathematical physics) — Wikipedia, the free encyclopedia," 2019. [Online; accessed 18-March-2020].

[12] T. J. Lane, D. Shukla, K. A. Beauchamp, and V. S. Pande, "To milliseconds and beyond: challenges in the simulation of protein folding," *Current Opinion in Structural Biology*, vol. 23, no. 1, pp. 58 – 65, 2013. Folding and binding / Protein-nucleic acid interactions.

[13] F. Noé, "Beating the millisecond barrier in molecular dynamics simulations," *Biophysical Journal*, vol. 108, pp. 228–229, 2020/02/14 2015.

[14] S. Kalyaanamoorthy and Y.-P. P. Chen, "Modelling and enhanced molecular dynamics to steer structure-based drug discovery," *Progress in Biophysics and Molecular Biology*, vol. 114, no. 3, pp. 123–136, 2014.

[15] J. A. McCammon, B. R. Gelin, and M. Karplus, "Dynamics of folded proteins," *Nature*, vol. 267, no. 5612, pp. 585—590, 1977.

[16] R. O. Dror, R. M. Dirks, J. P. Grossman, H. Xu, and D. E. Shaw, "Biomolecular simulation: A computational microscope for molecular biology," *Annual Review of Biophysics*, vol. 41, no. 1, pp. 429–452, 2012.

[17] S. A. Hollingsworth and R. O. Dror, "Molecular dynamics simulation for all," *Neuron*, vol. 99, pp. 1129—1143, 2020/02/14 2018.

[18] M. Karplus and J. A. McCammon, "Molecular dynamics simulations of biomolecules," *Nature Structural Biology*, vol. 9, no. 9, pp. 646–652, 2002.

[19] S. A. Adcock and J. A. McCammon, "Molecular dynamics: Survey of methods for simulating the activity of proteins," *Chemical Reviews*, vol. 106, pp. 1589—1615, 05 2006.

[20] C. W. Gear, *Numerical Initial Value Problems in Ordinary Differential Equations*. USA: Prentice Hall PTR, 1971.

[21] J.-P. Ryckaert, G. Ciccotti, and H. J. Berendsen, "Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes," *Journal of Computational Physics*, vol. 23, no. 3, pp. 327 – 341, 1977.

[22] B. Hess, H. Bekker, H. J. C. Berendsen, and J. G. E. M. Fraaije, "Lincs: A linear constraint solver for molecular simulations," *Journal of Computational Chemistry*, vol. 18, no. 12, pp. 1463–1472, 1997.

[23] R. W. Hockney, "The potential calculation and some applications," *Methods Comput. Phys.*, vol. 9, p. 136, 1970.

[24] L. Monticelli and D. P. Tieleman, *Force Fields for Classical Molecular Dynamics*, pp. 197—213. Totowa, NJ: Humana Press, 2013.

[25] A. D. Mackerell Jr., "Empirical force fields for biological macromolecules: Overview and issues," *Journal of Computational Chemistry*, vol. 25, no. 13, pp. 1584–1604, 2004.

[26] M. A. Gonzalez, "Force fields and molecular dynamics simulations," *Collection SFN*, vol. 12, pp. 169–200, 2011.

[27] Maier James A., Martinez Carmenza, Kasavajhala Koushik, Wickstrom Lauren, Hauser Kevin E., and Simmerling Carlos, "ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB," *Journal of Chemical Theory and Computation*, vol. 11, no. 8, p. 3696–3713, 2015. doi: 10.1021/acs.jctc.5b00255.

[28] J. Huang and A. D. MacKerell Jr, "CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data," *Journal of Computational Chemistry*, vol. 34, no. 25, p. 2135–2145, 2013.

[29] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case, "Development and testing of a general amber force field," *Journal of Computational Chemistry*, vol. 25, no. 9, p. 1157–1174, 2004.

[30] Jorgensen William L. and Tirado-Rives Julian, "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*, vol. 110, no. 6, p. 1657–1666, 1988. doi: 10.1021/ja00214a001.

[31] B. A. Luty, M. E. Davis, I. G. Tironi, and W. F. V. Gunsteren, "A comparison of particle-particle, particle-mesh and ewald methods for calculating electrostatic interactions in periodic molecular systems," *Molecular Simulation*, vol. 14, no. 1, pp. 11–20, 1994.

[32] R. Salomon-Ferrer, A. W. Götz, D. Poole, S. Le Grand, and R. C. Walker, "Routine microsecond molecular dynamics simulations with amber on gpus. 2. explicit solvent particle mesh ewald," *Journal of Chemical Theory and Computation*, vol. 9, pp. 3878—3888, 09 2013.

[33] T. E. I. Cheatham, J. L. Miller, T. Fox, T. A. Darden, and P. A. Kollman, "Molecular dynamics simulations on solvated biomolecular systems: The particle mesh ewald method leads to stable trajectories of dna, rna, and proteins," *Journal of the American Chemical Society*, vol. 117, pp. 4193—4194, 04 1995.

[34] P. Mark and L. Nilsson, "Structure and dynamics of the tip3p, spc, and spc/e water models at 298 k," *The Journal of Physical Chemistry A*, vol. 105, pp. 9954–9960, 11 2001.

[35] D. van der Spoel, P. J. van Maaren, and H. J. C. Berendsen, "A systematic study of water models for molecular simulation: Derivation of water models optimized for use with a reaction field," *The Journal of Chemical Physics*, vol. 108, no. 24, pp. 10220–10230, 1998.

[36] Wikipedia contributors, "Water model — Wikipedia, the free encyclopedia," 2020. [Online; accessed 30-March-2020].

[37] J. D. Durrant and J. A. McCammon, "Molecular dynamics simulations and drug discovery," *BMC Biology*, vol. 9, no. 1, p. 71, 2011.

[38] P. E. M. Lopes, O. Guvench, and A. D. MacKerell, *Current Status of Protein Force Fields for Molecular Dynamics Simulations*, pp. 47–71. New York, NY: Springer New York, 2015.

[39] K. Lindorff-Larsen, P. Maragakis, S. Piana, M. P. Eastwood, R. O. Dror, and D. E. Shaw, "Systematic validation of protein force fields against experimental data," *PLOS ONE*, vol. 7, pp. 1–6, 02 2012.

[40] P. Tiwary and A. van de Walle, *A Review of Enhanced Sampling Approaches for Accelerated Molecular Dynamics*, pp. 195–221. Cham: Springer International Publishing, 2016.

[41] V. Spiwok, Z. Sucur, and P. Hosek, "Enhanced sampling techniques in biomolecular simulations," *Biotechnology Advances*, vol. 33, no. 6, Part 2, pp. 1130–1140, 2015. BioTech 2014 and 6th Czech-Swiss Biotechnology Symposium.

[42] C. Camilloni and F. Pietrucci, "Advanced simulation techniques for the thermodynamic and kinetic characterization of biological systems," *Advances in Physics: X*, vol. 3, no. 1, pp. 885–916, 2018.

[43] Y. I. Yang, Q. Shao, J. Zhang, L. Yang, and Y. Q. Gao, "Enhanced sampling in molecular dynamics," *The Journal of Chemical Physics*, vol. 151, no. 7, p. 070902, 2019.

[44] C. Abrams and G. Bussi, "Enhanced sampling in molecular dynamics using metadynamics, replica-exchange, and temperature-acceleration," *Entropy*, vol. 16, no. 1, pp. 163–199, 2014.

[45] G. Bussi and A. Laio, "Using metadynamics to explore complex free-energy landscapes," *Nature Reviews Physics*, vol. 2, no. 4, pp. 200–212, 2020.

[46] A. Laio and F. L. Gervasio, "Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science," *Reports on Progress in Physics*, vol. 71, p. 126601, nov 2008.

[47] A. Barducci, M. Bonomi, and M. Parrinello, "Metadynamics," *WIREs Computational Molecular Science*, vol. 1, no. 5, pp. 826–843, 2011.

[48] O. Valsson, P. Tiwary, and M. Parrinello, "Enhancing important fluctuations: Rare events and metadynamics from a conceptual viewpoint," *Annual Review of Physical Chemistry*, vol. 67, no. 1, pp. 159–184, 2016. PMID: 26980304.

[49] L. Sutto, S. Marsili, and F. L. Gervasio, "New advances in metadynamics," *WIREs Computational Molecular Science*, vol. 2, no. 5, pp. 771–779, 2012.

[50] H. Grubmüller, "Predicting slow structural transitions in macromolecular systems: Conformational flooding," *Phys. Rev. E*, vol. 52, pp. 2893–2906, Sep 1995.

[51] A. Barducci, G. Bussi, and M. Parrinello, "Well-tempered metadynamics: A smoothly converging and tunable free-energy method," *Phys. Rev. Lett.*, vol. 100, p. 020603, Jan 2008.

[52] P. Tiwary and M. Parrinello, "A time-independent free energy estimator for metadynamics," *The Journal of Physical Chemistry B*, vol. 119, pp. 736–742, 01 2015.

[53] M. Bonomi, A. Barducci, and M. Parrinello, "Reconstructing the equilibrium boltzmann distribution from well-tempered metadynamics," *Journal of Computational Chemistry*, vol. 30, no. 11, pp. 1615–1621, 2009.

[54] J. Smiatek and A. Heuer, "Calculation of free energy landscapes: A histogram reweighted metadynamics approach," *Journal of Computational Chemistry*, vol. 32, no. 10, pp. 2084–2096, 2011.

[55] J. F. Dama, M. Parrinello, and G. A. Voth, "Well-tempered metadynamics converges asymptotically," *Phys. Rev. Lett.*, vol. 112, p. 240602, Jun 2014.

[56] G. Bussi and G. A. Tribello, *Analyzing and Biasing Simulations with PLUMED*, pp. 529–578. New York, NY: Springer New York, 2019.

[57] S. Mittal and D. Shukla, "Recruiting machine learning methods for molecular simulations of proteins," *Molecular Simulation*, vol. 44, no. 11, pp. 891–904, 2018.

[58] Y. Wang, J. M. L. Ribeiro, and P. Tiwary, "Machine learning approaches for analyzing and enhancing molecular dynamics simulations," *Current Opinion in Structural Biology*, vol. 61, pp. 139 – 145, 2020.

[59] F. Sicard and P. Senet, "Reconstructing the free-energy landscape of met-enkephalin using dihedral principal component analysis and well-tempered metadynamics," *The Journal of Chemical Physics*, vol. 138, no. 23, p. 235101, 2013.

[60] M. M. Sultan and V. S. Pande, "tica-metadynamics: Accelerating metadynamics by using kinetically selected collective variables," *Journal of Chemical Theory and Computation*, vol. 13, pp. 2440–2447, 06 2017.

[61] J. McCarty and M. Parrinello, "A variational conformational dynamics approach to the selection of collective variables in metadynamics," *The Journal of Chemical Physics*, vol. 147, no. 20, p. 204109, 2017.

[62] M. K. Scherer, B. Trendelkamp-Schroer, F. Paul, G. Pérez-Hernández, M. Hoffmann, N. Plattner, C. Wehmeyer, J.-H. Prinz, and F. Noé, "PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models," *Journal of Chemical Theory and Computation*, vol. 11, pp. 5525–5542, Oct. 2015.

[63] G. A. Tribello, M. Ceriotti, and M. Parrinello, "A self-learning algorithm for biased molecular dynamics," *Proceedings of the National Academy of Sciences*, vol. 107, no. 41, pp. 17509–17514, 2010.

[64] M. E. Tipping and C. M. Bishop, "Mixtures of probabilistic principal component analyzers," *Neural Computation*, vol. 11, no. 2, pp. 443–482, 1999.

[65] P. Meinicke and H. Ritter, "Resolution-based complexity control for gaussian mixture models," *Neural Computation*, vol. 13, no. 2, pp. 453–475, 2001.

[66] P. Söderhjelm, G. A. Tribello, and M. Parrinello, "Locating binding poses in protein-ligand systems using reconnaissance metadynamics," *Proceedings of the National Academy of Sciences*, vol. 109, no. 14, pp. 5170–5175, 2012.

[67] S. Bhakat, E. Åberg, and P. Söderhjelm, "Prediction of binding poses to fxr using multi-targeted docking combined with molecular dynamics and enhanced sampling," *Journal of Computer-Aided Molecular Design*, vol. 32, no. 1, pp. 59–73, 2018.

[68] D. J. Earl and M. W. Deem, "Parallel tempering: Theory, applications, and new perspectives," *Phys. Chem. Chem. Phys.*, vol. 7, pp. 3910–3916, 2005.

[69] G. Bussi, F. L. Gervasio, A. Laio, and M. Parrinello, "Free-energy landscape for β hairpin folding from combined parallel tempering and metadynamics," *Journal of the American Chemical Society*, vol. 128, pp. 13435–13441, 10 2006.

[70] M. Bonomi and M. Parrinello, "Enhanced sampling in the well-tempered ensemble," *Phys. Rev. Lett.*, vol. 104, p. 190601, May 2010.

[71] X. Du, Y. Li, Y.-L. Xia, S.-M. Ai, J. Liang, P. Sang, X.-L. Ji, and S.-Q. Liu, "Insights into protein–ligand interactions: Mechanisms, models, and methods," *International Journal of Molecular Sciences*, vol. 17, no. 2, pp. 1–34, 2016.

[72] R. Perozzo, G. Folkers, and L. Scapozza, "Thermodynamics of protein–ligand interactions: History, presence, and future aspects," *Journal of Receptors and Signal Transduction*, vol. 24, no. 1-2, pp. 1–52, 2004. PMID: 15344878.

[73] M. R. Shirts, D. L. Mobley, and S. P. Brown, *Free-energy calculations in structure-based drug design*, p. 61–86. Cambridge University Press, 2010.

[74] N. Hansen and W. F. van Gunsteren, "Practical aspects of free-energy calculations: A review," *Journal of Chemical Theory and Computation*, vol. 10, pp. 2632–2647, 07 2014.

[75] D. L. Mobley and P. V. Klimovich, "Perspective: Alchemical free energy calculations for drug discovery," *The Journal of Chemical Physics*, vol. 137, no. 23, p. 230901, 2012.

[76] R. W. Zwanzig, "High-temperature equation of state by a perturbation method. ii. polar gases," *The Journal of Chemical Physics*, vol. 23, no. 10, pp. 1915–1922, 1955.

[77] J. G. Kirkwood, "Statistical mechanics of fluid mixtures," *The Journal of Chemical Physics*, vol. 3, no. 5, pp. 300–313, 1935.

[78] C. H. Bennett, "Efficient estimation of free energy differences from monte carlo data," *Journal of Computational Physics*, vol. 22, no. 2, pp. 245 – 268, 1976.

[79] V. Limongelli, M. Bonomi, and M. Parrinello, "Funnel metadynamics as accurate binding free-energy method," *Proceedings of the National Academy of Sciences*, vol. 110, no. 16, pp. 6358–6363, 2013.

[80] S. Bhakat and P. Söderhjelm, "Resolving the problem of trapped water in binding cavities: prediction of host–guest binding free energies in the sampl5 challenge by funnel metadynamics," *Journal of Computer-Aided Molecular Design*, vol. 31, pp. 119–132, Jan 2017.

[81] S. Genheden and U. Ryde, "The mm/pbsa and mm/gbsa methods to estimate ligand-binding affinities," *Expert Opinion on Drug Discovery*, vol. 10, no. 5, pp. 449–461, 2015. PMID: 25835573.

[82] E. Wang, H. Sun, J. Wang, Z. Wang, H. Liu, J. Z. H. Zhang, and T. Hou, "End-point binding free energy calculation with mm/pbsa and mm/gbsa: Strategies and applications in drug design," *Chemical Reviews*, vol. 119, pp. 9478–9508, 08 2019.

[83] S. Genheden and U. Ryde, "How to obtain statistically converged mm/gbsa results," *Journal of Computational Chemistry*, vol. 31, no. 4, pp. 837–846, 2010.

[84] J. Fan, A. Fu, and L. Zhang, "Progress in molecular docking," *Quantitative Biology*, vol. 7, pp. 83–89, Jun 2019.

[85] N. S. Pagadala, K. Syed, and J. Tuszynski, "Software for molecular docking: a review," *Biophysical Reviews*, vol. 9, pp. 91–102, Apr 2017.

[86] N. Brooijmans and I. D. Kuntz, "Molecular recognition and docking algorithms," *Annual Review of Biophysics and Biomolecular Structure*, vol. 32, no. 1, pp. 335–373, 2003. PMID: 12574069.

[87] E. Yuriev, J. Holien, and P. A. Ramsland, "Improvements, trends, and new ideas in molecular docking: 2012–2013 in review," *Journal of Molecular Recognition*, vol. 28, no. 10, pp. 581–604, 2015.

[88] S. F. Sousa, P. A. Fernandes, and M. J. Ramos, "Protein–ligand docking: Current status and future challenges," *Proteins: Structure, Function, and Bioinformatics*, vol. 65, no. 1, pp. 15–26, 2006.

[89] J. Li, A. Fu, and L. Zhang, "An overview of scoring functions used for protein–ligand interactions in molecular docking," *Interdisciplinary Sciences: Computational Life Sciences*, vol. 11, no. 2, pp. 320–328, 2019.

[90] S.-Y. Huang, S. Z. Grinter, and X. Zou, "Scoring functions and their evaluation methods for protein−ligand docking: recent advances and future directions," *Phys. Chem. Chem. Phys.*, vol. 12, pp. 12899–12908, 2010.

[91] A. Kumar and K. Y. J. Zhang, "Investigation on the effect of key water molecules on docking performance in csardock exercise," *Journal of Chemical Information and Modeling*, vol. 53, pp. 1880–1892, 08 2013.

[92] A. M. Ruvinsky, "Role of binding entropy in the refinement of protein-ligand docking predictions: analysis based on the use of 11 scoring functions.," *J Comput Chem*, vol. 28, pp. 1364–1372, Jun 2007.

[93] F. Chen, H. Liu, H. Sun, P. Pan, Y. Li, D. Li, and T. Hou, "Assessing the performance of the mm/pbsa and mm/gbsa methods. 6. capability to predict protein−protein binding free energies and re-rank binding poses generated by protein-protein docking," *Phys. Chem. Chem. Phys.*, vol. 18, pp. 22129–22139, 2016.

[94] A. J. Clark, P. Tiwary, K. Borrelli, S. Feng, E. B. Miller, R. Abel, R. A. Friesner, and B. J. Berne, "Prediction of protein–ligand binding poses via a combination of induced fit docking and metadynamics simulations," *Journal of Chemical Theory and Computation*, vol. 12, pp. 2990–2998, 06 2016.

[95] S. Bhakat, E. Åberg, and P. Söderhjelm, "Prediction of binding poses to fxr using multi-targeted docking combined with molecular dynamics and enhanced sampling," *Journal of Computer-Aided Molecular Design*, vol. 32, no. 1, pp. 59–73, 2018.

[96] S. W. Englander, T. R. Sosnick, J. J. Englander, and L. Mayne, "Mechanisms and uses of hydrogen exchange," *Current Opinion in Structural Biology*, vol. 6, no. 1, pp. 18 – 23, 1996.

[97] F. Persson and B. Halle, "How amide hydrogens exchange in native proteins," *Proceedings of the National Academy of Sciences*, 2015.

[98] D. E. Shaw, P. Maragakis, K. Lindorff-Larsen, S. Piana, R. O. Dror, M. P. Eastwood, J. A. Bank, J. M. Jumper, J. K. Salmon, Y. Shan, and W. Wriggers, "Atomic-level characterization of the structural dynamics of proteins," *Science*, vol. 330, no. 6002, pp. 341–346, 2010.

This is a doctoral thesis. Not a novel, textbook, monograph or a comprehensive review article. The introductory part of the thesis was not invented by me. It was written by consulting several books, articles, tutorials etc. which I have cited. If I skipped some deserving citation/s, I am sorry for that.

I am against the notion of *intellectual copyright*. It is similar to the philosophy of free software or free music where "free" refers to freedom, not price. As a supporter of free science, I give full permission to copy, distribute and modify this work without any restrictions.

The main part of this work deals with how one can use computational methods to study protein. One may ask, what is the social relevance of this study? When I was writing this thesis, DE Shaw research published a long molecular dynamics simulation of the SARS-CoV-2 virus. The trajectories were made *open* access so that it will help other scientists to develop effective drug against Covid-19. I think most people will agree that developing drugs for diseases such as Covid-19, Ebola, malaria etc. has large scale social impact. I believe in the near future, computational methods/models will gradually substitute experiments in the field of drug discovery. Unfortunately, this work will not able to generate any new drugs, but parts of the work might influence efforts to understand mechanism of drug action.

So, if you received this book and planned to read it, I hope some of the methods used in our study will be useful for your work. If you are only looking for your name in the acknowledgement then, I suggest you to use it to raise the IO device or laptop. Either way, stay happy and be safe!

Printed by Media-Tryck, Lund 2020 NORDIC SWAN ECOLABEL 3041 0903

ISBN 978-91-7422-745-1

Biophysical Chemistry
Department of Chemistry
Faculty of Engineering (LTH)
Lund University