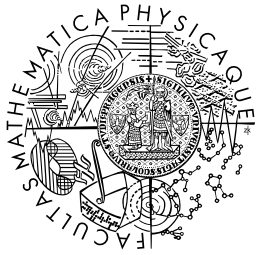


FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

MASTER'S THESIS

Bc. Martin Škorňa

Computational Modeling of Complexes Consisting of Nucleic Acids and Proteins

Division of Biomolecular Physics, Institute of Physics

Supervisor of the master's thesis: RNDr. Ivan Barvík, PhD.

Study programme: Biophysics and Chemical Physics

Study branch: Theoretical Biophysics and Chemical Physics

Prague 2024

UNIVERZITA KARLOVA
Matematicko-fyzikální fakulta

Fyzikální ústav UK

Akademický rok: 2021/2022

ZADÁNÍ DIPLOMOVÉ PRÁCE

Jméno a příjmení: **Martin Škorňa**

Studijní program: **Biofyzika a chemická fyzika**

Studijní obor: **Biofyzika a chemická fyzika se specializací Teoretická biofyzika a chemická fyzika**

Děkan fakulty Vám podle zákona č. 111/1998 Sb. určuje tuto diplomovou práci:

Téma v jazyce práce: **Počítačové modelování komplexů nukleových kyselin a proteinů**

Téma práce v anglickém jazyce: **Computational modeling of complexes consisting of nucleic acids and proteins**

Zásady pro vypracování:

1) Prostudovat určenou literaturu a sepsat rešerši:

- struktura nukleových kyselin a proteinů
- struktura, dynamika, funkce a medicínský význam proteinů interagujících s nukleovými kyselinami (DNA a RNA polymerázy, RNáza H, Argonaute, CRISPR-Cas9, zinkové prsty atd.)
- klasické molekulárně-dynamické simulace biomolekul
- výpočty volné energie
- QMMM výpočty

2) Osvojit si metodiku molekulárně-dynamických simulací - prakticky zvládnout práci se softwarovými balíky VMD, NAMD, Gaussian, ORCA.

3) Provést molekulárně-dynamické simulace modelových systémů sestávajících z vybraného proteinu, nukleové kyseliny a vodní obálky – dohromady cca. 25-100.000 atomů.

4) Nasimulované trajektorie analyzovat.

5) Získané výsledky diskutovat z hlediska potenciálního medicínského využití.

Seznam odborné literatury:

1. Leach A. R.: Molecular Modelling: Principles and Applications. Pearson Education Limited: Harlow, 2001, ISBN 0582382106

2. Frenkel D., Smit B.: Understanding Molecular Simulations: From Algorithms to Applications. Academic Press: San Diego, 2001, ISBN 0122673514

3. Phillips J. C., Braun R., Wang W., Gumbart J., Tajkhorshid E., Villa E., Chipot C., Skeel R. D., Kale L., Schulten K.: Scalable Molecular Dynamics with NAMD. J. Comput. Chem. 26 (2005) 1781-1802

4. Phillips J. C. et al.: Scalable molecular dynamics on CPU and GPU architectures with NAMD. J. Chem. Phys. 153 (2020) 044130

5. Chen H., Maia J. D. C., Radak B. K., Hardy D. J., Cai W., Chipot C. and Tajkhorshid E.: Boosting Free-Energy Perturbation Calculations with GPU-Accelerated NAMD. J. Chem. Inf. Model. 60 (2020) 5301-5307

6. Jensen F.: Introduction to Computational Chemistry. John Wiley & Sons Ltd.: West Sussex, 2007, ISBN: 0470058048

7. Mayne C. G., Saam J., Schulten K., Tajkhorshid E., and Gumbart J. C.: Rapid Parametrization of Small Molecules Using the Force Field Toolkit. J. Comput. Chem. 34 (2013) 2757-2770

8. Chipot Ch., Pohorille A.: Free Energy Calculations: Theory and Applications in Chemistry and Biology. Springer-Verlag: Berlin Heidelberg, 2007, ISBN: 9783540384472

9. Seeliger D., Buelens F. P., Goette M., de Groot B. L. and Grubmuller H.: Towards computational specificity screening of DNA-binding proteins. Nucleic Acids Research 39 (2011) 8281-8290

10. Goette M., Grubmuller H.: Accuracy and Convergence of Free Energy Differences Calculated from Nonequilibrium Switching Processes. J. Comput. Chem. 30 (2009) 447-456

11. Khabiri M. and Freddolino P. L.: Deficiencies in Molecular Dynamics Simulation-Based Prediction of Protein-DNA Binding Free Energy Landscapes. J. Phys. Chem. B 121 (2017) 5151-5161

12. Presnell K. V. and Alper H. S.: Thermodynamic and first-principles biomolecular simulations applied to synthetic biology: promoter and aptamer designs. Mol. Syst. Des. Eng. 3 (2018) 19-37

Vedoucí diplomové práce: **RNDr. Barvík Ivan, Ph.D.**

Navrhovaní oponenti:
Konzultanti:
Datum zadání diplomové práce: 19.4.2022

Termín odevzdání diplomové práce: dle harmonogramu příslušného akademického roku

.....
Vedoucí katedry

.....
za Děkana

V Praze dne 19.4.2022

I declare that I carried out this master's thesis independently, and only with the cited sources, literature and other professional sources. It has not been used to obtain another or the same degree.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Sb., the Copyright Act, as amended, in particular the fact that the Charles University has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 subsection 1 of the Copyright Act.

In ..... date ....

Author's signature

I would like to use this opportunity to express my sincere gratitude to Dr. Barvík for his invaluable assistance and guidance leading me to the successful writing of this thesis. His patience and support during the whole two years of effort were essential. A special thanks goes to my beloved family and to my dear friends for the unceasing support and encouragement.

Title: Computational Modeling of Complexes Consisting of Nucleic Acids and Proteins

Author: Bc. Martin Škorňa

Institute: Division of Biomolecular Physics, Institute of Physics

Supervisor: RNDr. Ivan Barvík, PhD., Division of Biomolecular Physics, Institute of Physics

Abstract: In the introductory chapter, the diploma thesis introduces the reader to the most used tools for the human genome editing (ZFN, TALEN, CRISPR). The following chapters describe in detail the methodology of MD simulations and calculations of hydration and binding free energies. In order to perform our calculations in a massively parallel way, an approach was chosen where a large number of short non-equilibrium MD runs are produced in parallel. From these, work values are obtained, from which the value of the equilibrium hydration or binding free energy is then determined using the Crooks-Gaussian Intersection method. The chosen methodology was first tested in the calculations of hydration free energies of nucleic acid bases and amino acid side chains. In the key results chapter, the complex of the transcription factor Zif268 and a short DNA double helix was studied. The effects of point mutations of individual base pairs in the DNA duplex on the binding free energy values were quantified.

Keywords: molecular dynamics, free energy calculations, nucleic acids, proteins, genetic engineering

Contents

Introduction **4**

1 Aims **5**

2 Biological Motivation – Gene Editing **6**
2.1 Wide Range of Possibilities 6
2.2 Zinc-finger Nucleases (ZFN) 7
2.3 Transcription Activator-like Effector Nucleases (TALEN) 9
2.4 RNA-guided Nucleases (CRISPR) 10

3 Theoretical Background **12**
3.1 Force Fields 12
3.1.1 Bond Term 13
3.1.2 Bond Angle Term 14
3.1.3 Dihedral Term 14
3.1.4 Electrostatic Term 15
3.1.5 Van der Waals Term 15
3.1.6 CHARMM Force Field 16
3.2 Energy Minimization 17
3.2.1 Steepest Descent Method 17
3.2.2 Conjugate Gradient Method 17
3.2.3 Newton-Raphson Method 17
3.3 Molecular Dynamics (MD) Simulations 19
3.3.1 Hamiltonian Mechanics 19
3.3.2 Poisson Bracket Formulation 20
3.3.3 Classical Liouville Operator 20
3.3.4 Trotter Expansion 21
3.3.5 Velocity Verlet Algorithm 22
3.3.6 Thermostats 23
3.3.7 Nosé-Hoover Thermostat 24
3.3.8 Langevin Thermostat (NAMD) 25
3.3.9 Periodic Boundary Conditions 26
3.4 Thermodynamics of Free Energy 28
3.4.1 First Law of Thermodynamics 28
3.4.2 Second Law of Thermodynamics 28
3.4.3 Third Law of Thermodynamics 29
3.4.4 Particle Changes in Closed Systems 29
3.4.5 Helmholtz Free Energy 29
3.4.6 Gibbs Free Energy 30
3.5 Free Energy Calculations 31
3.5.1 Free Energy in (Bio)chemistry and Biology 31
3.5.2 Geometrical vs. Alchemical Methods 32
3.5.3 Thermodynamic Cycles 33
3.5.4 Free Energy and Statistical Partition Function 34
3.5.5 Alchemical Transformations 34

3.5.6 Thermodynamic Integration (TI) . . . 35
3.5.7 Free Energy Perturbation (FEP) . . . 35
3.5.8 Pathway of Intermediate States . . . 36
3.6 Non-equilibrium Approach . . . 38
3.6.1 Free Enegy and Work . . . 38
3.6.2 Jarzynski Identity . . . 39
3.6.3 Crooks Fluctuation Theorem (CFT) . . . 40
3.6.4 Mounting CFT on FEP Framework . . . 43

4 Methodology **46**
4.1 Building of Simulated Systems . . . 46
4.1.1 Protein Data Bank (PDB) . . . 46
4.1.2 VMD – Molefacture . . . 46
4.1.3 VMD – AutoPSF . . . 47
4.1.4 VMD – Solvate . . . 49
4.1.5 VMD – Autoionize . . . 50
4.2 Hybrid Molecular Topologies . . . 51
4.2.1 Single vs. Dual Topology . . . 51
4.2.2 Dual Topology Paradigm . . . 52
4.2.3 Preventing End-point Catastrophes . . . 52
4.3 Non-equilibrium Free Energy Calculations . . . 54
4.4 Hardware – MetaCentrum . . . 56
4.5 Software – NAMD . . . 57
4.6 Data Extraction and Analysis . . . 58
4.6.1 Bash Scripting . . . 58
4.6.2 Custom Python Analysis . . . 58
4.6.3 Scott's Normal Reference Rule . . . 58
4.6.4 Freedman-Diaconis Rule . . . 58
4.6.5 Unified Bin Width . . . 59
4.6.6 Plotting Histogram Distributions . . . 60
4.6.7 Crooks-Gauss Intersection (CGI) . . . 61

5 Hydration Free Energies of Amino Acids **62**
5.1 Simulation Setup . . . 64
5.2 Absolute Hydration Free Energies . . . 65
5.3 Discussion . . . 67

6 Hydration Free Energies of DNA Bases **68**
6.1 Simulation Setup and Method of Analysis . . . 70
6.2 Absolute Hydration Free Energies . . . 71
6.3 Relative Hydration Free Energies . . . 71
6.4 Mutations in Trinucleotides . . . 72
6.5 Discussion . . . 73
6.5.1 Absolute Free Energies . . . 73
6.5.2 Relative Free Energies . . . 73
6.5.3 Mutations in Trinucleotides . . . 73

7 Zif268-DNA Complex **74**
7.1 Simulation Setup . . . 77
7.2 Initial Equilibration . . . 78
7.2.1 Watson-Crick Hydrogen Bonding . . . 78
7.2.2 Amino Acid Side Chains . . . 79
7.3 Relative Binding Free Energies . . . 81
7.4 Mutation Site 2 . . . 83
7.4.1 Mutation C2A . . . 84
7.4.2 Mutation C2G . . . 85
7.4.3 Mutation C2T . . . 85
7.5 Mutation Site 8 . . . 87
7.5.1 Mutation C8A . . . 88
7.5.2 Mutation C8G . . . 88
7.5.3 Mutation C8T . . . 88
7.6 Mutation Site 4 . . . 89
7.6.1 Mutation T4A . . . 90
7.6.2 Mutation T4C . . . 91
7.6.3 Mutation T4G . . . 93
7.6.4 Summary . . . 95
7.7 Mutation Site 3 – G3C . . . 96
7.8 Mutation Site 5 – G5C . . . 97
7.9 Mutation Site 6 – G6C . . . 98
7.10 Mutation Site 7 – G7C . . . 100
7.11 Mutation Site 9 – G9C . . . 102
7.12 Discussion . . . 103
7.12.1 Stability of Mutated Base Pairs . . . 103
7.12.2 Influence of Proximate Amino Acids . . . 104
7.12.3 DNA Sequence Detection by Zif268 . . . 104
7.12.4 Impact of Mutations on Binding Free Energies . . . 106

Conclusion **108**

Bibliography **109**

Introduction

The diploma thesis loosely follows on the long-standing cooperation between the Institute of Physics CU and the IOCB ASCR in the area of targeted interventions in the expression of genetic information [1, 2]. This area has undergone rapid development in recent decades.

First, it was a simple blocking of the expression of genetic information at the DNA or mRNA level by means of so-called antisense or antigene oligonucleotides [3]. They usually have a length of approx. 20 nucleotides and carry various chemical modifications in the sugar phosphate backbone, which make the oligonucleotides resistant to cellular nucleases and which increase their affinity to the target mRNA or DNA. At the same time, chemical modifications should not disrupt the binding of the cellular enzyme RNase H, which in DNA:mRNA complexes degrades the mRNA strand and thus releases the antisense deoxyoligonucleotide, which can then bind to the next copy of the mRNA and the whole cycle can be repeated. At the end of the 1990s, it was discovered that cellular micro-RNAs work similarly to antisense oligonucleotides. Micro-RNAs stimulate the degradation of the target mRNA via the Argonaute enzyme. Andrew Fire and Craig C. Mello won the 2006 Nobel Prize in Physiology or Medicine for the discovery of this RNA interference mechanism. Today there are already many approved oligonucleotide therapeutics [3].

In the course of time, it was also possible to obtain several molecular tools (ZFN, TALEN, CRISPR) that allow for targeted interventions in DNA [4]. In particular, the elucidation of the molecular mechanism of CRISPR in 2012 led to the acquisition of a tool that allows editing of the human genome to be performed essentially routinely in a number of laboratories around the world. Thanks to this, Emmanuelle Charpentier and Jennifer Doudna won the 2020 Nobel Prize in Chemistry. Currently, many clinical trials for therapeutics based not only on CRISPR, but also on TALENs or ZFNs, are underway [4].

In order to make interventions in the expression of genetic information effective, computer modeling tools are widely used [2, 5, 6]. These make it possible to examine and optimize the structure of antisense oligonucleotides or key enzymes used for genome editing so that their action is as specific as possible and no unwanted off-target effects occur. The so-called molecular dynamics (MD) simulations, which enable the binding free energy of complexes to be quantified, are particularly useful for this. However, in the case of large complexes of enzymes and DNA, the given methodology is in the phase of active research, when optimal algorithms are sought and existing versions of the main software packages are modified to give consistent results [6].

1. Aims

- To acquaint readers with the most used tools for editing of the human genome (ZFN, TALEN, CRISPR).
- To describe the methodology of MD simulations and free energy calculations.
- To prepare scripts that will make it possible to perform calculations of hydration and binding free energy in the MetaCentrum supercomputer environment as efficiently as possible using massive parallelization.
- Test the selected algorithms on model systems (nucleic acid components, amino acids).
- Apply the debugged procedures to the ZF-DNA complex. To verify that using the NAMD program and the algorithms implemented in it, it is possible to obtain results comparable with those given by other software packages.
- The obtained binding free energy values should be interpreted in detail at the atomic level based on the evolution of the structures of the ZF-DNA complexes during the alchemical MD transformations.

2. Biological Motivation – Gene Editing

The fundamental step that gave rise to the field of genetic engineering was the discovery of genetic information, first identified in the late 1860s by Swiss chemist Friedrich Miescher [7], and with it the idea of inserting, deleting, or modifying genes within an organism's genome. By introducing specific genetic changes, one could enhance crop resistance to pests and diseases, develop new therapies for genetic disorders, or even engineer microbes to produce valuable pharmaceuticals or biofuels. It holds the promise of personalized treatments, where tailored therapies can target the genetic basis of diseases. Furthermore, genetic engineering enables us to delve deeper into the mysteries of life, unlocking the potential for new discoveries and innovations that could reshape our understanding of biology and the world of nature. However, it also raises important ethical, environmental, and safety concerns that necessitate careful consideration and regulation as the field continues to evolve. It indeed is a double-edged sword.

2.1 Wide Range of Possibilities

In the field of genetic engineering there are various different ways and tools for targeted genetic modifications of not only cultured cells, but also whole living animals and plants. One of the (comparatively) well established classes of molecular tools consists of programmable nucleases, including *zinc-finger nucleases* (ZFNs), *transcription activator-like effector nucleases* (TALENs), and RNA-guided engineered nucleases (RGENs) derived from the publicly well-known bacterial *clustered regularly interspaced short palindromic repeat* (CRISPR)–Cas system. The value of these enzymes in research, medicine and biotechnology arises from their ability to induce site-specific DNA cleavage in the genome, the repair (through endogenous mechanisms) of which allows high-precision genome editing [8].

Each one of these nucleases are characterized by their size, chemical composition, targetable sites and related specificities, mutation signatures, and other important aspects at play. Study of these features pose an essential part of research, allowing us to assign the most suitable machinery for a given range of applications. Experimental measurements can give only so much detail, lacking deeper understanding of the underlying mechanisms, often missing the timescale and resolution on which these events operate. Theoretical and computational approaches can provide the necessary atomistic insight, unlocking the full potential of these molecular machines.

2.2 Zinc-finger Nucleases (ZFN)

Zinc-finger nucleases (ZFNs) have a modular composition consisting of two main domains – DNA-binding zinc-finger protein (ZFP) and the nuclease derived from the *Fok*I restriction enzyme [9]. One such machinery is shown in Fig. 2.1. The wild-type *Fok*I has a structurally separated DNA-binding domain, that can

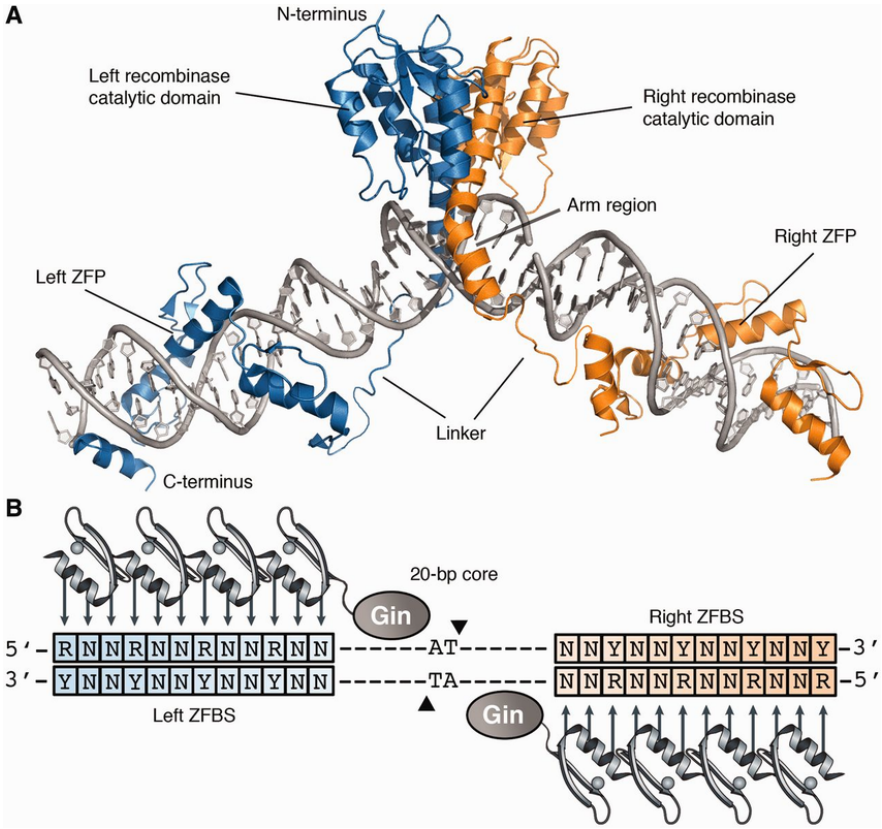


Figure 2.1: Example of a zinc-finger recombinase (ZFR) dimer bound to DNA. (A) Molecular model. Each monomer (blue or orange) consists of an activated serine recombinase catalytic domain linked to a custom-designed, DNA-binding zinc finger protein (ZFP). (B) Sketch of the dimer bound to DNA. Target sites consist of two-inverted ZFBS surrounding a central 20-bp core recognized by the catalytic domain. ZFPs can be designed to recognize distinct 'left' or 'right' half-sites (blue and orange boxes, respectively). Symbols: N represents A, T, C or G; R marks G or A; and Y indicates C or T. Image taken from [10].

be replaced with ZFPs to forge the ZFNs. This artificial structure is made out of two complementary parts, which have to dimerize in order to cleave DNA [11]. For that it is by some referred to as the *zinc-finger recombinase* (ZFR). Each monomer has to bind to adjacent half-sites, separated by spacers of $5 - 7$ bp. This forms a core of up to 20 bp on which the nuclease can operate. Typical ZFP, made out of 3 interconnected finger domains, recognizes 9 bp (3 bp per zinc-finger). Dimerization effectively doubles the length of recognition sites, leading to a considerable increase in specificity of ZFNs as opposed to the wild-type *Fok*I. This requirement ultimately leads to substantial reduction in off-target effects and cytotoxicity of the complex. Such heterodimeric structure is sometimes paired with many ZFPs (by some referred to as a 'train' of ZFPs) specific to different sites near by the taget core, in order to minimize unwanted off-targets even more.

In contrast with other programmable nucleases, ZFNs suffer poor targeting density, limiting its application range. Even though each zinc-finger recognizes 3 bp, there is no open-source collection of 64 zinc-fingers that covers all possible combinations of triplet sites [12]. Another issue stems from the fact that not all ZFNs, especially with three-fingered ZFPs, can cleave chromosomal DNA efficiently [8]. So far it has been used to modify endogenous genes in organisms ranging from viruses, bacteria and cultured cells, to plants, insects, fish, and mammals such as mice or pigs.

2.3 Transcription Activator-like Effector Nucleases (TALEN)

Structure of TALENs is akin to that of ZFNs, see Fig. 2.2. Similarly, they utilize *Fok*I catalytic domain, but for binding to specific DNA sites the so-called *transcription activator-like effectors* (TALEs) are applied. This category of DNA-binding domains is based on proteins from *Xanthomonas* bacterium, pathogenic to plants [13]. TALEs are made out of tandem arrays of $33 - 35$ amino acid repeats (units) [14]. Sequence recognition is provided by amino acids at positions 12 and 13 [15], which detect a single base pair in the major groove. Following the steps of ZFNs, TALENs have to dimerize in order to make any changes to DNA.

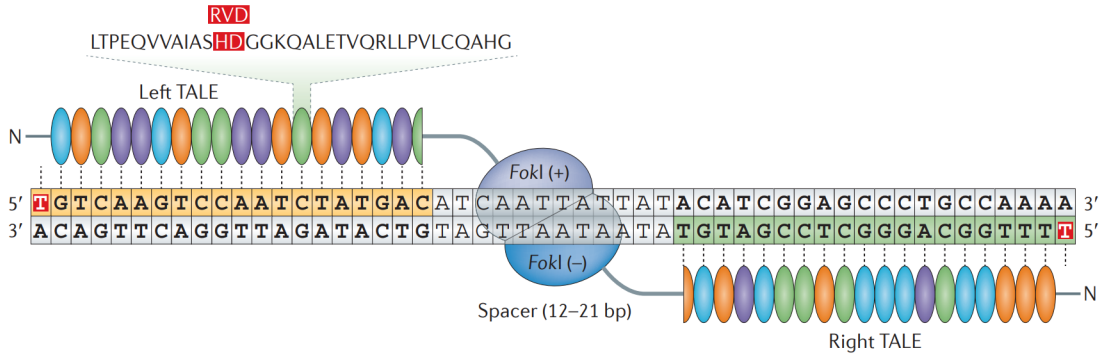


Figure 2.2: Cartoon representation of transcription activator-like effector nuclease (TALEN) bound to DNA sequence. TALENs are composed of transcription activator-like effectors (TALEs) at the amino terminus and the *Fok*I catalytic domain at the carboxyl terminus. Each TALE unit is formed by $33 - 35$ amino acids and recognizes a single base pair by amino acids at positions 12 and 13. This small region is called the repeat variable diresidue (RVD). Target sequences of TALEN dimers are typically $30 - 40$ bp in length, excluding spacers. Picture adapted from [8].

Major advantage of TALENs is that they can be designed to target pretty much any given DNA sequence. On the other hand, their limitation lies in the requirement of thymine (T) at the 5' end of the target sequence, recognized by 2 amino-terminal cryptic repeat folds [14]. Although relatively recently there have been developed TALEs recognizing other bases at the target's 5' end [16], broadening the range of applications. Also conventional TALEs have issues cleaving sequences with methylated cytosines (C) [17]. In order to eliminate this restriction, additional modifications to TALEs have to be done.

Same as ZFNs, TALENs have been successfully used to modify endogenous genes in a wide range of organisms, including viruses and yeast, cultured cells, plants, insects, frogs, fish, and mammals such as mice and pigs.

2.4 RNA-guided Nucleases (CRISPR)

In the recent years, probably the most notoriously known class of tools for genetic engineering belongs to the RNA-guided systems, or RNA-guided engineered nucleases (RGENs). Various RGENs have been successfully used on bacteria, mammalian cells, embryos, plants, nematodes, fruitflies, mice, non-human primates, human pluripotent stem cells, and the list goes on. It is the technology publicly known as CRISPR – the 'hero' of genetic engineering.

It has been discovered [18, 19] that in bacteria and archaea RNA-guided systems for DNA cleavage serve as an adaptive immunity against invading phages and plasmids. The base idea is rather simple. Small ($\sim$ 20 bp) fragments of foreign DNA is captured by the organism and inserted into its very own genome to form a *clustered regularly interspaced short palindromic repeat* (CRISPR). Such repeats are then transcribed and processed to form target-specific CRISPR RNA (crRNA). Invariable target-independent trans-activating crRNA (tracrRNA) is transcribed as well and contributes to the crRNA creation [8]. Both crRNA and tracrRNA then complex with CRISPR-*asociated protein* 9 (Cas9). Mounting RNA on induces conformational changes in Cas9 forming a central channel the target DNA can slip in. This complex is the endonuclease serving as a newly adapted, active immunity against the foreign genetic information.

A connection between crRNA and tracrRNA can be established via a *tetraloop* to form a single-chain guide RNA (sgRNA) [20]; this simplified RGEN is depicted in Fig. 2.3 alongside the cleavage it performs. The whole endonuclease mounts

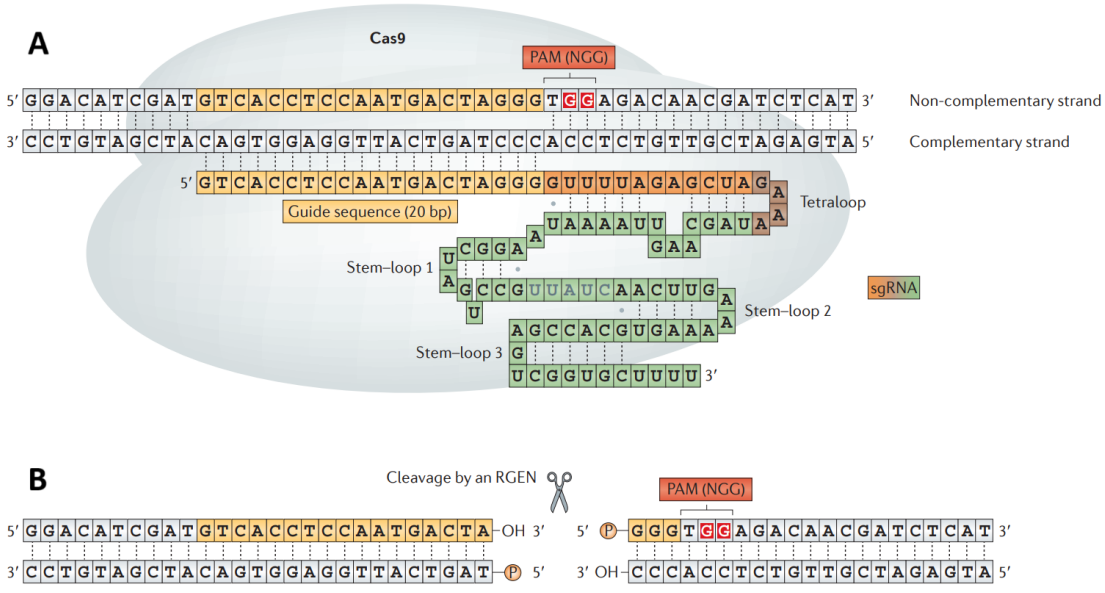


Figure 2.3: Cartoon representation of RNA-guided engineered nuclease (RGEN). (A) RGEN can contain Cas9 protein and a single-chain guide RNA (sgRNA), complementary to a 20-bp target DNA sequence (protospacer). Target sequence is next to the protospacer adjacent motif (PAM) 5'-NGG-3', where N marks any nucleotide. Weak bonding is shown as gray dots. (B) Target DNA sequence is cleaved by the RGEN producing blunt ends. Image adapted from [8].

onto the foreign DNA and slides along, unwinding a little loop, until it finds the target sequence for cleavage. As is shown in the pictures, the RNA-Cas9 complex cleaves a target DNA sequence of 23 bp, composed of 20-bp guide (protospacer) and 5'-NGG-3' sequence called the *protospacer adjacent motif* (PAM).

This little clip is a region which Cas9 protein recognizes itself.

There are of course different variations of RGENs recognizing different PAMs (e.g. 5'-NAG-3'), or involving Cas proteins with modified domains which enhance specific activities of the complex. Their availability in terms of the design and preparation make them a high-stake competitor to systems like ZFNs or TALENs. Once the appropriate (native or modified) Cas protein is chosen for a given application, it remains the same. The only thing that has to be specifically prepared is either the crRNA or sgRNA. Their preparation is done by cloning 20-bp guide DNA sequences in a suitable vector that encodes the given RNA [8].

As is indicated in Fig. 2.3, one of the disadvantages of RGENs derived from Cas proteins is their size which can make it more difficult to deliver them into living cells. This can sometimes be the deciding factor weather to use this or some other, less bulky tool. On the other hand, as opposed to ZFNs and TALENs, RGENs allow for cleaving methylated DNA [21], though with less efficiency inside living cells [22].

3. Theoretical Background

3.1 Force Fields

In the context of molecular dynamics we talk about the combination of a mathematical formula and associated atomic parameters. These force fields are used to describe the energy of a given molecular system as a function of its atomic coordinates and atomic types assigned to each species based on their immediate surroundings. It takes an important role in various computational methods to calculate molecular conformations based on the system's potential energy. Every well-behaved force field contains its own formula describing system's potential energy using variety of different terms based on all approximations the given method was derived with. What we generally end up with is a description where all atomic nuclei in the ground state are moving on the same potential energy hypersurface (PES). This enables us to disentangle the calculation of potential energy from the computation of motion of atomic nuclei. An advantage of this approach lies in elimination of the large number of atoms in exchange for the search of energy hypersurface minimum. This technique requires substantially less computational effort even for systems containing enormous (up to $\sim 10^6$) number of particles. On the contrary, apart from special *ab intio* MD and hybrid QM/MM methods, we are unable to simulate chemical reactions.

Accuracy of the given method is determined by the number and complexity of mathematical terms describing potential energy of the system. We distinguish between 1. and 2. generation force fields. First generation fields are known for their use of less sophisticated potentials, with parameters obtained empirically from experimental measurements. An example could be AMBER force field [23], constructed for nucleic acid and protein modeling. Second generation fields make a use of more complex potentials and their parameters are acquired via *ab initio* calculations, borrowed from quantum chemistry, e.g. CFF [24, 25] or COMPASS [26].

Potential terms are categorized based on whether they describe bonding or non-bonding interactions. Potential energy thus reads

$$E = E_{\mathrm{B}} + E_{\mathrm{NB}}. \tag{3.1}$$

Bonding interactions affecting atoms include atomic bond lengths, angles they form, dihedral angles, or planarity of the surroundings

$$E_{\mathrm{B}} = E_{\mathrm{bon}} + E_{\mathrm{ang}} + E_{\mathrm{dih}} + E_{\mathrm{pla}}. \tag{3.2}$$

Non-bonding component usually relates to atoms separated by at least 3 covalent bonds. It is divided into electrostatic (Coulomb) interactions, Van der Waals interactions, and interactions via hydrogen bonds

$$E_{\mathrm{NB}} = E_{\mathrm{C}} + E_{\mathrm{VdW}} + E_{\mathrm{HB}}. \tag{3.3}$$

Such terms can be of numerous forms. Some force fields may use different guise of similar terms or their modified versions to improve performance for specific types of molecular structures. Nevertheless, the foundation remains largely consistent for practical reasons.

3.1.1 Bond Term

This term models covalent bond between 2 atoms using harmonic oscillator with atomic bond stiffness k_r

$$E_\mathrm{bon}(r) = \frac{k_r}{2}\left(r - r_0\right)^2, \tag{3.4}$$

where r represents an immediate and r_0 the equilibrium radial distance between both atoms. Harmonic model does not allow for proper behavior far from equilibrium point, e.g. high temperature simulations. In such a case it is useful to employ Morse potential [27] instead

$$E_\mathrm{M}(r) = D_\mathrm{e}\left(1 - e^{-a(r-r_0)^2}\right)^2. \tag{3.5}$$

Compared to parabolic approximation, Morse's approach offers an additional parameter, see Fig. 3.1. Though this allows for higher temperature modeling[1], Morse potential fails for larger values of r where the curvature changes too slowly. This can cause a lot of trouble during initial optimization process. In practise, 2. generation force fields use higher-order (usually 3. or 4. order) Taylor expansion of harmonic potential instead.

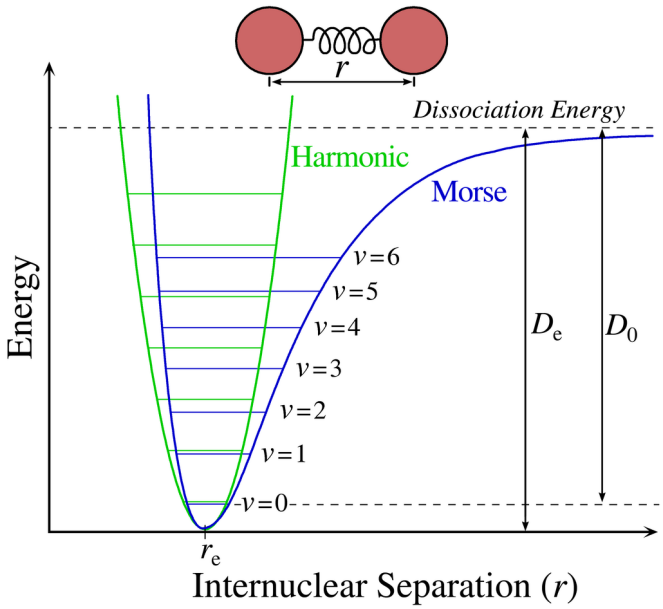


Figure 3.1: Comparison of Morse potential to harmonic oscillator model, as functions of internuclear separation r. Unlike the energy levels of the harmonic oscillator potential, which are evenly spaced by $\hbar\omega$, the Morse potential level spacing decreases as the energy approaches the dissociation energy. The dissociation energy D_e is larger than the true energy required for dissociation D_0 due to the zero point energy of the lowest ($\nu = 0$) vibrational level. Image taken from [28].

[1]For systems with significant anharmonicity (non-negligible deviations from the harmonic approximation) such as in the study of higher-energy vibrational modes, or systems with weak bonds that are more prone to breaking, the Morse potential can provide a more accurate representation of the potential energy surface and is preferred. Nevertheless, modeling such systems and their behavior is a subject of ab initio MD, or even quantum chemistry without the use of force fields in the first place.

3.1.2 Bond Angle Term

Angular term characterizes interaction between 2 nuclei connected to the same atom. For this purpose we usually employ harmonic description

$$E_{\text{ang}}(\theta) = \frac{k_\theta}{2}\left(\theta - \theta_0\right)^2, \tag{3.6}$$

where k_θ is the angular stiffness, θ represents an immediate and θ_0 the equilibrium angle they form.

Another approach is to use Urey-Bradley potential [29], utilizing the distance between the first and the third atom instead of the angle

$$E_{\text{UB}}(r_{13}) = \frac{k}{2}\left(r_{13} - r_{13}^0\right)^2. \tag{3.7}$$

Such a model brings not only manipulation with angle θ_{123}, but also variations of the bond lengths r_{12} and r_{23}. Urey-Bradley term is a cross-term accounting for 1, 3 non-bonded interactions not included in the bond and angle terms. Due to its nature, it is used in combination with both the bond and angle terms, if one wants to incorporate its effects into the calculation.

3.1.3 Dihedral Term

This is a torsion term scanning angle ϕ between atomic planes. Each plane is defined by 3 atoms, while 2 out of these triads form a common axis, see Fig. 3.2. The associated dihedral term renders

$$E_{\text{dih}}(\phi) = \frac{\nu_n}{2}\left[1 + \cos\left(n\phi - \gamma\right)\right], \tag{3.8}$$

where parameter ν_n shows the energy difference between energetically most and least favourable configuration with respect to angle ϕ, n is multiplicity (number of minima in the range from 0 to 2π), and γ is dihedral phase.

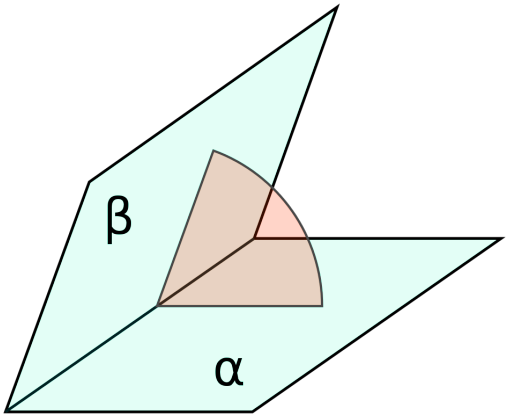


Figure 3.2: Dihedral angle – an angle between two half-planes (α, β) in a third plane (red) which cuts the line of intersection at right angles. The axis common to both planes is formed by 2 atoms, shared between 2 interconnected atomic triads. Image taken from [30].

3.1.4 Electrostatic Term

Notoriously known classical Coulomb [31], [32] term brings the electrostatic interaction as one of the non-bonding component of the potential energy [33]

$$E_{\mathrm{C}} = \frac{1}{4\pi\epsilon_0}\frac{q_i q_j}{r_{ij}}. \tag{3.9}$$

Electric charges q are usually placed at the positions of individual atoms, but can be implemented differently according to specific purposes.

3.1.5 Van der Waals Term

As a Van der Waals (VdW) term we usually choose empirical Lennard-Jones potential [34]

$$E_{\mathrm{LJ}}(r) = 4\varepsilon\left[\left(\frac{\sigma}{r_{ij}}\right)^{12} - \left(\frac{\sigma}{r_{ij}}\right)^{6}\right], \tag{3.10}$$

where ε is the depth of the potential well (or *dispersion energy*), σ the distance at which the particle-particle potential energy is zero (or *size of the particle*), and r_{ij} is the mutual distance between 2 atoms in the system.

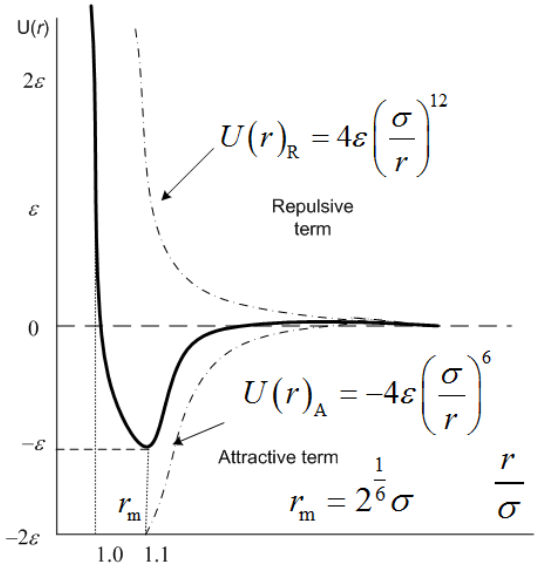


Figure 3.3: Sketch of Lennard-Jones potential side by side with its attractive and repulsive components. Picture taken from [35].

Shape of Lennard-Jones potential (3.10) is depicted in Fig. 3.3. Attractive VdW component is mediated by dipole-dipole interaction, divided into *London dispersion forces* [36] (induced and induced dipole), *Debye forces* [36, 37] (induced and permanent dipole), and *Keesom forces* [36, 38] (permanent and permanent dipole). This component is represented by the second term in Eq. 3.10, inversely proportional to the 6. power of the mutual distance between 2 atoms in the system. Repulsive VdW component is based on repulsion of fully occupied orbitals, the so-called *Pauli repulsion* [39]. In Eq. 3.10 it is represented by the first term, inversely proportional to the 12. power of the distance between 2 given atoms. This component operates on very small scales, and its 12. power was chosen manually in order to be easily squared from the already calculated power-6 term.

3.1.6 CHARMM Force Field

Computations done for the purpose of this thesis use exclusively CHARMM36 force field [40]. CHARMM (Chemistry at HARvard Molecular Mechanics) is a widely-used force field in the field of molecular dynamics simulations. Developed and maintained by Martin Karplus and his collaborators, CHARMM is renowned for its accuracy in modeling the behavior of biomolecules, such as nucleic acids, proteins, and lipids.

It is a 2nd-generation force field belonging to the class of semi-empirical force fields. As such, it combines empirical data with quantum chemical calculations to derive the necessary parameters for simulating molecular systems. The use of quantum mechanical principles allows for significant improvement in accuracy of the force field compared to its strictly empirical predecessors.

Key features of CHARMM FF include a comprehensive parameterization for a wide range of biomolecular interactions, including bonded and non-bonded terms, dihedral angles, and electrostatic forces. It is particularly valued for its ability to capture the complex interactions and conformational changes that occur in biological and biochemical systems, though its versatility extends beyond biomolecules, as it has also been adapted for simulations of small organic molecules, liquids, and various materials. This growing width of possible applications secures its place among well-respected force fields of today.

The name CHARMM also stands for the molecular dynamics and analysis software [41, 42] associated with the force fields. Nevertheless, CHARMM FF can easily be used in conjunction with numerous molecular dynamics programs, the most popular of which may be GROMACS [43] or NAMD [44]. Its availability allows for exploration of the dynamics and thermodynamics of molecular systems, making it a valuable tool in areas of biophysics, biochemistry, and drug discovery.

CHARMM36 force field has the following potential energy function [45, 46]

$$\begin{aligned} U = &\sum_{bon} k_r \left(r - r_0\right)^2 + \sum_{ang} k_\theta \left(\theta - \theta_0\right)^2 + \sum_{dih} k_\phi \left[1 + \cos\left(n\phi - \delta\right)\right] \\ &+ \sum_{imp} k_\varphi \left(\varphi - \varphi_0\right)^2 + \sum_{UB} k_u \left(u - u_0\right)^2 \\ &+ \sum_{non} \left(\varepsilon_{ij} \left[\left(\frac{R_{ij}^{\min}}{r_{ij}} \right)^{12} - 2 \left(\frac{R_{ij}^{\min}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{\varepsilon_{\mathrm{r}} r_{ij}} \right), \end{aligned} \tag{3.11}$$

where we identify sum over bond terms (3.4), angular terms (3.6), dihedral terms (3.8), Urey-Bradley potentials (3.7), and non-bonded part formed by sum over LJ potentials (3.10) and Coulomb interactions (3.9). All of these are similar to those found in other force fields such as AMBER [23].

Attentive reader might have noticed that CHARMM36 (3.11) features also an *improper* term accounting for out-of-plane bending, which applies to any set of 4 atoms that are not successively bonded. This term, similarly as other angular parts, is modeled harmonically with k_φ being the force constant, and $\varphi - \varphi_0$ the out-of-plane angle.

3.2 Energy Minimization

In molecular simulations, potential minimization is a technique used to find the most stable arrangement of atoms in a molecular system. This stable arrangement corresponds to the lowest potential energy state. Potential energy states are often visualized by potential energy hypersurfaces (PES) – landscapes with valleys and peaks. Searching through these hypersurfaces can be a challenging task. This process, also known as *conformation analysis*, benefits from a wide range of mathematical algorithms, allowing for systematic scanning of molecular structures. One can imagine them as tiny explorers navigating this landscape of valleys and peaks, until they settle at the lowest valley of all.

Potential minimization always starts from an initial conformation for which an energy calculation is performed. This initial guess can be random or based on some prior knowledge of the system, e.g. crystal structure obtained through x-ray analysis or other experimental method. The molecular structure is then systematically varied according to the given algorithm in terms of bond lengths, angles, and other aspects defining the system's conformation state. Energy calculation is repeated, and the whole procedure is reiterated until the optimal structure, satisfying a given condition (e.g. maximum force acting on system's particles), is successfully reached.

For small enough models it is possible to find the global minimum. To verify its validity we optimize multiple different initial conformations, ultimately leading us to the same final minimal state. Larger systems like proteins are usually optimized into variety of plausible conformations, which can likely be found in nature under specific conditions.

3.2.1 Steepest Descent Method

The Steepest Descent (SD) algorithm [47] iteratively updates atomic positions by following the negative gradient of the PES. This intuitive approach directly steers the system towards lower energy regions. However, its simplicity comes at a cost. SD struggles to navigate narrow valleys on the PES, potentially becoming trapped in local minima.

3.2.2 Conjugate Gradient Method

For enhanced efficiency, the Conjugate Gradient (CG) algorithm [48] incorporates information from previous steps. By employing conjugate directions, it avoids revisiting past minimization paths and exhibits superior convergence compared to SD. This method is particularly advantageous for investigating large molecular systems where computational cost becomes a significant factor.

3.2.3 Newton-Raphson Method

The Newton-Raphson (NR) algorithm [49] stands out for its rapid convergence properties. It leverages the curvature of the PES, incorporating the *Hessian matrix* [50], second derivative of the potential energy

$$(\mathbf{H}_f)_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j} \tag{3.12}$$

for a more informed update of atomic positions. Here $f : \mathbb{R}^n \to \mathbb{R}$ is a general function taking a vector $\mathbf{x} \in \mathbb{R}^n$ as an input and producing a scalar $f(\mathbf{x}) \in \mathbb{R}$. Given that all second-order partial derivatives of f exist, elements (3.12) are to be arranged into the $n \times n$ Hessian matrix $\mathbf{H}_f$. In our case, $f(\mathbf{x})$ is the system's potential energy and $\mathbf{x}$ the atomic coordinates.

Hessian matrices serve as coefficient of the quadratic term of the function's local Taylor expansion

$$f(\mathbf{x} + \Delta\mathbf{x}) \approx f(\mathbf{x}) + \boldsymbol{\nabla} f(\mathbf{x})^{\mathrm{T}} \Delta\mathbf{x} + \frac{1}{2} \Delta\mathbf{x}^{\mathrm{T}} \mathbf{H}_f(\mathbf{x}) \Delta\mathbf{x}, \tag{3.13}$$

where $\boldsymbol{\nabla} \equiv \left(\frac{\partial}{\partial x_1}, \ldots, \frac{\partial}{\partial x_n}\right)$. If $\mathbf{H}_f(\mathbf{x})$ is positive semi-definite, the quadratic approximation is a convex function of $\Delta\mathbf{x}$, and its minimum is then localized by

$$0 = \boldsymbol{\nabla} f(\mathbf{x} + \Delta\mathbf{x}) = \boldsymbol{\nabla} f(\mathbf{x}) + \mathbf{H}_f(\mathbf{x}) \Delta\mathbf{x} + \mathcal{O}\left(\left\|\Delta\mathbf{x}^2\right\|\right), \tag{3.14}$$

with minimum achieved for

$$\Delta\mathbf{x} = -\mathbf{H}_f(\mathbf{x})^{-1} \boldsymbol{\nabla} f(\mathbf{x}). \tag{3.15}$$

Putting all together, NR method performs an iteration scheme[2]

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \Delta\mathbf{x} = \mathbf{x}_k - \mathbf{H}_f(\mathbf{x}_k)^{-1} \boldsymbol{\nabla} f(\mathbf{x}_k), \text{ for } k \geq 0, \tag{3.16}$$

constructing a sequence $\{\mathbf{x}_k\}$ from an initial guess $\mathbf{x}_0 \in \mathbb{R}^n$ that converges towards minimizer $\mathbf{x}_{\min}$ of f.

While this translates to faster minimization, the NR method comes with some caveats. The computational cost associated with calculating the Hessian can be substantial, and the method is far more sensitive to the initial conformation guess compared to its alternatives. Comparison to gradient descent is shown in Fig. 3.4 using a simple contour graph for illustration.

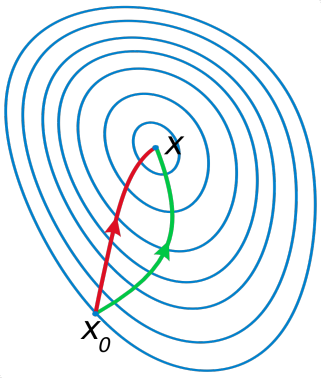


Figure 3.4: A comparison of gradient descent (green) and Newton-Raphson method (red) for minimizing a function (with small step sizes). Function sketched using its contours. NR uses curvature information (i.e. the second derivative) to take a more direct route. Picture taken from [49].

[2]NR is sometimes modified to include a small step size with $0 < \gamma \leq 1$ instead of $\gamma = 1$: $\mathbf{x}_{k+1} = \mathbf{x}_k + \gamma\Delta\mathbf{x} = \mathbf{x}_k - \gamma\mathbf{H}_f(\mathbf{x}_k)^{-1} \boldsymbol{\nabla} f(\mathbf{x}_k)$.

3.3 Molecular Dynamics (MD) Simulations

Everybody's talking about this molecular dynamics, but actually ... what is it? Molecular dynamics (or MD) is a branch of computational physics allowing us to study the behavior of physical, chemical, and biological systems made out of atoms, molecules, and their complexes. It is a body of algorithms (such as numerical integration, thermostats, barostats etc.) using the laws of motion to simulate such systems and provide us with their time evolution as a result. It is a direct product of the 1950's endeavor to solve the N-body problem, which (as we all surely know by now) is analytically unsolvable. For this reason, MD is purely numerical and hence often called the numerical form of statistical mechanics, or Laplace's vision of Newtonian mechanics.

MD comes in many 'shapes and colors', some of them even closely tied to quantum mechanics (ab initio MD, hybrid QM/MM simulations). No matter the form, each of them have to obey the given equations of motion (EOM) and deal with particle interaction potentials. Here we turn our focus to the solely classical version of MD which utilizes Hamiltonian mechanics and its equations of motion to rule over the simulated systems. These are the ones utilizing the so-called *force fields*[3] to deal with interatomic interactions.

3.3.1 Hamiltonian Mechanics

Let us now briefly refresh our brains with what Hamiltonian mechanics [51] is to set ourselves the notation we will be using and some little ground to stand on. In a 1D case a general Hamiltonian could be written as

$$H(p,q) = \frac{p^2}{2m} + U(q), \tag{3.17}$$

where mass m serves as a parameter, p is the momentum, q the position, and $U(q)$ a potential (conservative in the simplest case). Hamiltonian equations of motion (HEOM) now appear as

$$\begin{aligned} \dot{q} &= \frac{\partial H}{\partial p} = \frac{p}{m}, \\ \dot{p} &= -\frac{\partial H}{\partial q} = -\frac{\partial U}{\partial q} \equiv F(q). \end{aligned} \tag{3.18}$$

In the case of N particles in 3D ($3N$ degrees of freedom) the Hamiltonian reads

$$H\left(\mathbf{p},\mathbf{q}\right) = \sum_{i=1}^{N} \frac{\mathbf{p}_i{}^2}{2m_i} + U\left(\{\mathbf{q}_i\}_{i=1}^{N}\right), \tag{3.19}$$

with HEOM taking the form

$$\begin{aligned} \dot{\mathbf{q}}_i &= \frac{\partial H}{\partial \mathbf{p}_i} \equiv \nabla_{\mathbf{p}_i} H, \\ \dot{\mathbf{p}}_i &= -\frac{\partial H}{\partial \mathbf{q}_i} \equiv \mathbf{F}_i. \end{aligned} \tag{3.20}$$

[3]The mathematical models of molecular potentials with associated sets of parameters for every particle of the system studied, dependant on the specific environment each particle is in.

For the sake of simplicity, we will for the most part use the 1D version which (if needed) can be converted into the N-particle 3D case with ease. For our future use we also introduce $x \equiv \{\mathbf{p}_i, \mathbf{q}_i\}_{i=1}^{N}$ as the set of all positions and momenta of an N-particle system.

3.3.2 Poisson Bracket Formulation

The approach could be summarized by several very important words – Poisson bracket formulation of Hamiltonian mechanics, classical Liouville operator, and Trotter expansion leading us to the Trotter version of velocity Verlet computation [52]. Lemme explain ...

Consider general functions $f(p,q)$, $g(p,q)$ on phase space. Poisson bracket [53] of these is

$$\{f,g\} = \frac{\partial f}{\partial q}\frac{\partial g}{\partial p} - \frac{\partial f}{\partial p}\frac{\partial g}{\partial q}. \tag{3.21}$$

If we chose $g(p,q) = H(p,q)$ and f to be either simply q or p, the corresponding Poisson brackets are

$$\{q,H\} = \frac{\partial q}{\partial q}\frac{\partial H}{\partial p} - \frac{\partial q}{\partial p}\frac{\partial H}{\partial q} = \dot{q} \tag{3.22}$$

and

$$\{p,H\} = \frac{\partial p}{\partial q}\frac{\partial H}{\partial p} - \frac{\partial p}{\partial p}\frac{\partial H}{\partial q} = \dot{p} \tag{3.23}$$

when taking account for the independence of phase space coordinates p and q, and inserting HEOM. Based on equations (3.22) and (3.23) we can express HEOM using the Poisson bracket formulation, i.e.

$$\begin{aligned} \dot{q} &= \frac{\partial H}{\partial p} = \{q,H\}, \\ \dot{p} &= -\frac{\partial H}{\partial q} = \{p,H\}. \end{aligned} \tag{3.24}$$

The full set of HEOM can also be written in a compact format[4]

$$\dot{x} = \{x,H\}, \tag{3.25}$$

where x is the set of all p's and q's introduced at the beginning of this section.

3.3.3 Classical Liouville Operator

Now, based on equations (3.25) we can introduce a classical Liouville superoperator as

$$iL \coloneqq \{\blacksquare, H\}, \tag{3.26}$$

[4]As one could remember, equations (3.25) are a powerful tool for uncovering quantities conserved in the system, i.e. the so-called integrals of motion. If a certain quantity y Poisson-commutes with the Hamiltonian, meaning that $\{y,H\} = 0$, the time derivative of that quantity vanishes leaving a constant solution to the given equation. This quantity is therefore one of the requisite integrals. For more detail you can check out Chapter 2 and Appendix B of my bachelor's thesis [54].

where ■ marks a position in which we place the quantity the superoperator acts upon. An example of such an operation would be

$$iLx = \{x, H\} = \dot{x}, \tag{3.27}$$

of which the solution happens to be

$$x(t) = x\left(t_0\right) e^{iL(t-t_0)} \tag{3.28}$$

...sort of. You see, to be really exact, this actually isn't the solution of the differential equation (3.27). What we did (strictly speaking) is that we just hid all the p's and q's inside some exponential time evolution of the Liouville superoperator. The exponential in (3.28) is thus a time-evolution propagator and we will denote it as $U(t)$. It is an analogy to the QM propagator $\exp\left(-\frac{i}{\hbar}\hat{\mathbf{H}}t\right)$ given there by the Schrödinger equation [39].

The Liouville superoperator can be separated into 2 distinct parts according to the phase space coordinates p and q as

$$iL = \dot{q}\frac{\partial}{\partial q} + \dot{p}\frac{\partial}{\partial p} = \frac{p}{m}\frac{\partial}{\partial q} + F\left(q\right)\frac{\partial}{\partial p}. \tag{3.29}$$

We will label both of the constituent parts as iL_q and iL_p which leads us to

$$\exp(iL_q t + iL_p t) \neq \exp(iL_q t)\exp(iL_p t) \neq \exp(iL_p t)\exp(iL_q t), \tag{3.30}$$

where we emphasise that the constituent parts do not commute with each other, and hence the exponential form cannot be separated as a simple product. This is the case since the parts iL_q and iL_p both have a prefactor dependent on the other variable, see separation (3.29). If you don't believe me, just try applying these superoperators on some function of p's and q's one after the other, and vice versa. The results will not match.

3.3.4 Trotter Expansion

In order to create a chain of operations for a computer to work with, we need these exponentials to be separated. So what do we do? To separate the exponential from the left side of (3.30) we can apply the so-called *Trotter expansion* [55, 56]

$$\begin{aligned} \exp(A+B) &= \lim_{P\to\infty}\left[\exp\left(\frac{A}{2P}\right)\exp\left(\frac{B}{P}\right)\exp\left(\frac{A}{2P}\right)\right]^P \\ &= \lim_{P\to\infty}\left[\exp\left(\frac{A}{P}\right)\exp\left(\frac{B}{P}\right)\right]^P. \end{aligned} \tag{3.31}$$

Both of the possible forms of Trotter expansion are valid. They approach the limit differently but ultimately they end up the same. Trotter expansion uses the fact that we can chop up $\exp(A+B)$ into incremental pieces with a factor P, making an infinite ordered product of progressively smaller and smaller constituent exponentials. In limit where $P \to \infty$ such an expansion converges to the desired $\exp(A+B)$.

The result of applying the first version of Trotter expansion (3.31) to the left side of (3.30) is [57, 58]

$$U(\Delta t) = \exp\left(iL_p \frac{\Delta t}{2}\right) \exp(iL_q \Delta t) \exp\left(iL_p \frac{\Delta t}{2}\right), \tag{3.32}$$

where $\Delta t = \frac{t}{P}$ is a discrete time step defining our resolution in time. Equation (3.32) shows the propagation by one singular time step Δt in our simulation. Every such a time-step propagation is then part of the whole iteration process represented by the $[\dots]^P$ in the Trotter expansion. To simplify this for our further (less mathematical) explanation we introduce the following notation

$$U(\Delta t) = U_p\left(\frac{\Delta t}{2}\right) U_q(\Delta t) U_p\left(\frac{\Delta t}{2}\right), \tag{3.33}$$

where U_i are partial propagators with respect to the given phase space coordinate i. It is worth noting that this is a sequential operation in which the order of the partial propagations is important!

3.3.5 Velocity Verlet Algorithm

So how does this actually work in the computer? Without much need of any advanced math, please. We can write one such iteration (by one simulation time step Δt) as a 6-step sequential updating suitable for a computer:

1. Initial conditions: $q(0)$, $p(0)$; $F\left(q(0)\right)$,
2. $U_p\left(\frac{\Delta t}{2}\right)$: $p\left(\frac{\Delta t}{2}\right) = p(0) + F\left(q(0)\right)\frac{\Delta t}{2}$,
3. $U_q\left(\Delta t\right)$: $q\left(\Delta t\right) = q(0) + \frac{1}{m}p\left(\frac{\Delta t}{2}\right)\Delta t$,
4. Updating the forces: $F\left(q(0)\right) \to F\left(q\left(\Delta t\right)\right)$,
5. $U_p\left(\frac{\Delta t}{2}\right)$: $p\left(\Delta t\right) = p\left(\frac{\Delta t}{2}\right) + F\left(q\left(\Delta t\right)\right)\frac{\Delta t}{2}$,
6. Output: $q\left(\Delta t\right)$, $p\left(\Delta t\right)$, $F\left(q\left(\Delta t\right)\right)$.

This can be simplified to a pseudocode using Python notation as:

1. Input q, p, F,
2. $p \mathrel{+}= F\frac{\Delta t}{2}$,
3. $q \mathrel{+}= \frac{p}{m}\Delta t$,
4. Update F,
5. $p \mathrel{+}= F\frac{\Delta t}{2}$,
6. $q\left(\Delta t\right), p\left(\Delta t\right), F\left(q\left(\Delta t\right)\right)$.

Here symbol += means that we simply add the right-hand expression to the old (non-updated) value on the left side. To help us visualize such a process, the single iteration is depicted in Figure 3.5.

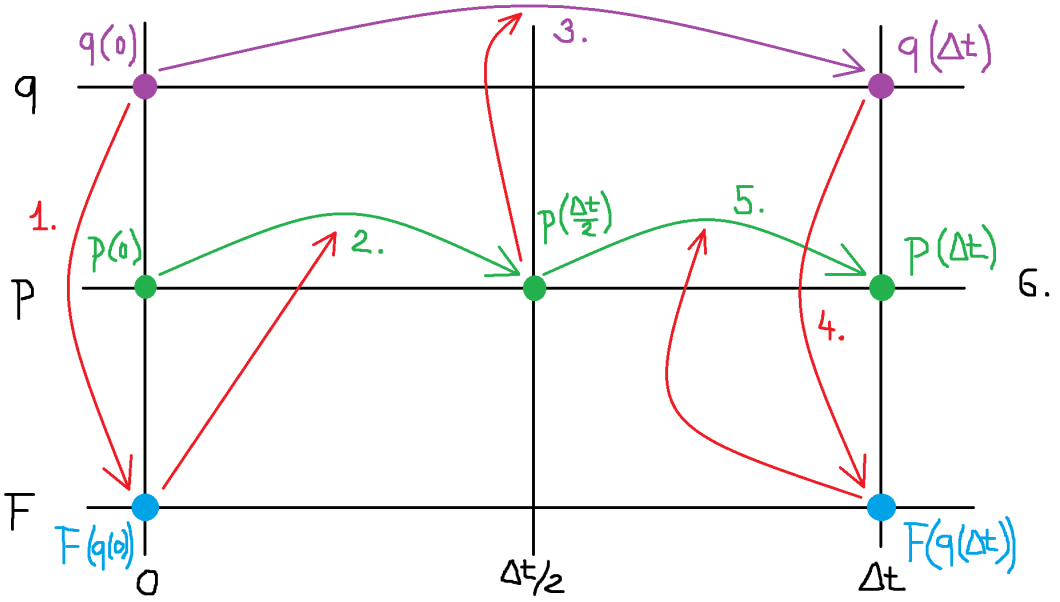


Figure 3.5: Diagram of one iteration (by single time step Δt) in the Trotter version of velocity Verlet computation [52]. Numbers mark the corresponding steps of the sequential updating we presented above. Red arrows represent the transfer of information which serves as the necessary input for each step. This little sequence of operations is then repeated over and over (several million times) in the whole iteration process.

3.3.6 Thermostats

As the name suggests, the goal of a thermostat is to maintain temperature T of the system we are trying to simulate. In molecular dynamics thermostat is a sort of digital heat bath which not only maintains but also creates the temperature of our simulated system in the first place. As we all know, temperature is a collective result of all particles in the system moving around with some kinetic energy given by a certain statistical distribution. Simulations in molecular dynamics introduce this heat bath as a new part of the time dependent side of our problem by taking advantage of the equipartition theorem. This theorem assigns the value $\frac{1}{2}\mathrm{k_B}T$ to every degree of freedom, leaving the average (expectation) value of kinetic energy

$$\langle K \rangle = \frac{1}{2}\mathrm{k_B}T. \tag{3.34}$$

By this relation we are able to connect the temperature T with kinetic energies K in our system – T is directly generated via K's assigned to every each one of the constituent atoms in the system.

Introduction of a good thermostat, suitable for the given problem, is crucial in order for the system to behave as it should – mathematically speaking, so that the system has the right probability distribution describing it. If the thermostat isn't right the overall T might still be correct but the equipartition principle is violated and the resulting probability distribution is therefore wrong.

Thermostats can be classified into 3 standard categories according to the way they operate:

1. Velocity re-scaling,
2. Extended system (explicitly adding degrees of freedom),
3. Stochastic (adding randomness).

The fist one, as its name suggests, simply re-scales the velocities of individual particles in the system after a given period of time. Extended system thermostats relate particles of our NVT system to some extra particles from an artificial heat bath. It is basically a way of exchanging the energy with an outside reservoir represented by extra degrees of freedom explicitly added to our system. The last one, stochastic, adds random perturbations and/or dissipations to the simulated system by adopting some random function, e.g. the white noise.

3.3.7 Nosé-Hoover Thermostat

This is a standard example of an extended ensemble thermostat. Nosé-Hoover thermostat [59, 60] adds 1 or more degrees of freedom (DOF) for the physical system to exchange its E with. As a result, we are left with non-Hamiltonian EOM, but that is no problem since we can still use the same framework for the propagation. For 1 DOF the EOM read

$$\begin{aligned} \dot{q} &= \frac{p}{m}, \\ \dot{p} &= F - \frac{p_\eta}{M} p, \end{aligned} \tag{3.35}$$

where p_η and M are the momentum and mass of the extra DOF. The mass M is not specified which means that we ourselves need to define how heavy the added DOF is by presenting

$$\begin{aligned} \dot{\eta} &= \frac{p_\eta}{M}, \\ \dot{p_\eta} &= \frac{p^2}{m} - \mathrm{k_B}T = F_\eta. \end{aligned} \tag{3.36}$$

Here η is a parameter of the added DOF, and the force F_η represents an instantaneous deviation.

For the case of N particles in 3D ($3N$ DOF) the full set of EOM resembles

$$\begin{aligned} \dot{\mathbf{q}}_i &= \frac{\mathbf{p}_i}{m_i}, \\ \dot{\mathbf{p}}_i &= \mathbf{F}_i - \frac{p_\eta}{M} \mathbf{p}_i, \\ \dot{\eta} &= \frac{p_\eta}{M}, \\ \dot{p_\eta} &= \sum_{j=1}^{N} \frac{\mathbf{p}_j^2}{m_j} - 3N\mathrm{k_B}T. \end{aligned} \tag{3.37}$$

The last equation uses additional factor $3N$ for the thermal element $\mathrm{k_B}T$ to adjust for all DOF of the system. Since a well-defined conserved quantity

$$H^{'}\left(\mathbf{q},\eta,\mathbf{p},p_{\eta}\right)=H\left(\mathbf{q},\mathbf{p}\right)+\frac{p_{\eta}^{2}}{2M}+3N\mathrm{k_B}T\eta \tag{3.38}$$

exists, it is 'just' NVE dynamics of the extended system we are dealing with. In the propagation framework, this all can be thought of as attaching a partial (thermostat) propagator $U_{\mathrm{T}}\left(\frac{\Delta t}{2}\right)$ from both sides to the already existing NVE 1-time-step propagation (3.33) as [57]

$$U\left(\Delta t\right)=U_{\mathrm{T}}\left(\frac{\Delta t}{2}\right)U_{\mathrm{NVE}}\left(\Delta t\right)U_{\mathrm{T}}\left(\frac{\Delta t}{2}\right). \tag{3.39}$$

If we want to add more DOF for the system to interact with, we can simply couple these small 1-DOF thermostats in a chain one after the other as is illustrated in Figure 3.6. This is the reason why Nosé-Hoover thermostat is sometimes referred to as the *chain thermostat*.

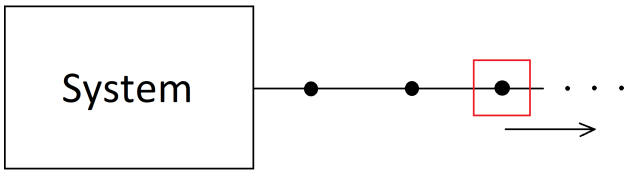


Figure 3.6: Schematic depiction of chaining in Nosé-Hoover thermostat. The box represents our simulated system while the dots in the chain portray individual extra DOF for the system to exchange E with.

Why would we want to chain them? Addition of only 1 DOF as an extended system thermostat is simply not enough for the simulated system to behave correctly. Chaining them like this allows the system with such a 'whole' thermostat to function as it naturally should in a more realistic (non-simulated) scenario. The more extra DOF we chain, the more accurate the results get (to the degree our method's approximations allow) – but at a certain cost of a more resource-demanding computation. In a fairly usual case, 3 additional DOF are enough for the system to start behaving reasonably well.

3.3.8 Langevin Thermostat (NAMD)

Langevin thermostat is a standard example of a stochastic thermostat which combines both dissipation and perturbation so that the T of the particles stays the same throughout the whole simulation. It perturbs and dissipates momenta p by introducing 2 new force components to the system's EOM as [57]

$$\begin{aligned}\dot{q}&=\frac{p}{m},\\ \dot{p}&=F-\gamma p+\sqrt{2\mathrm{k_B}T\gamma m}R(t),\end{aligned} \tag{3.40}$$

where γ is a dissipation factor, and $R(t)$ a Gaussian white (uncorrelated in time) noise term such that $\langle R(t)\rangle=0$ and $\langle R(t_1)R(t_2)\rangle=\delta(t_2-t_1)$. For this reason it is sometimes referred to as the *White Noise Langevin Dynamics* [61].

The truly interesting part is the additional forces this method imposes on our system – the stochastic part

$$\dot{p} = -\gamma p + \sqrt{2\mathrm{k_B}T\gamma m}R(t) \tag{3.41}$$

of which the solution is

$$p(t) = p(0)e^{-\gamma t} + \sqrt{\mathrm{k_B}Tm\,(1 - e^{-2\gamma t})}R(t). \tag{3.42}$$

This solution samples the canonical ensemble. To imagine this in practise, this is how a colloidal particle moves in its environment.

3.3.9 Periodic Boundary Conditions

How do we put our system in proper boundary conditions such that in the end our simulation correctly represents the real-world system we are trying to calculate? For bulk simulations of a system in its natural environment, one of the most used approaches is to utilize periodic boundary conditions [62, 63]. This means we create a box (cell) containing the system of interest (e.g. DNA-protein complex) immersed in its natural environment (water with suitable ions). Periodic boundary conditions then generate infinitely many copies of this box side by side in every possible direction, see Fig. 3.7. By this we do not create any new atoms. We are only producing replicas of the very same atoms and molecules (and their behavior) in the form of repeating boxes filling the whole space.

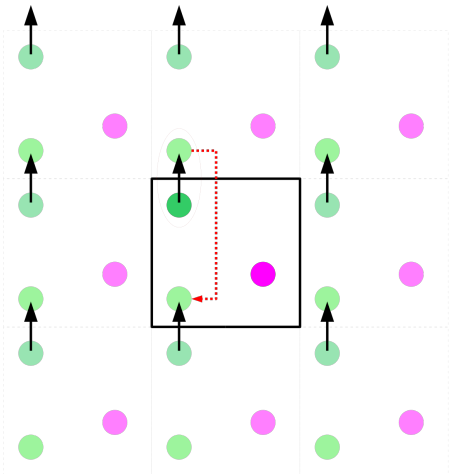


Figure 3.7: Schematic depiction of periodic boundary condition, simplified in 2D. Center square forms the boundary of the initial simulation cell, repeating in both dimensions of the plane. Black arrows indicate evolution of one of the particles (green) crossing the cell's boundary, red arrow highlights one of the copies of the original particle. As the green particle leaves the initial cell it immediately reappears on the opposite side. In that, the initial box is topologically analogous to a 4D torus. Picture taken from [64].

The cell boundaries are no special place. Particles can move through them but symmetrically appear on the other side of the box. However, this is no teleportation. Remember there is the same box (containing replicas of the same atoms and molecules) on every side of the box we are currently looking at.

There are plenty of other methods that can be used for different purposes. For example slab simulations, which make replicas in juts 2 out of 3 dimensions. They form sort of '2D slabs' with both surfaces touching their respective environments – e.g. water slab with air molecules below and above, which can be for instance used to simulate the surface of a droplet of water in the air. We can also go down the dimensions to the so-called chain simulations, which replicate the given system in just a single dimension. This is helpful for calculations of polymer chains that are rigid enough (e.g. carbon nano tubes).

3.4 Thermodynamics of Free Energy

Free energy quantifies the portion of a system's internal energy available to perform useful work. Understanding its variations allows one to predict the spontaneity of processes and the maximum achievable work by a system at hand. In thermodynamics, there are different types of free energy defined. In our thesis we denote a general free energy A, unless deliberately specified. Thermodynamics of free energy is based on the fundamental concepts of internal energy (U) and entropy (S), both of which are subjects of the laws of thermodynamics.

3.4.1 First Law of Thermodynamics

The first law of thermodynamics [65] dictates the conservation of energy through

$$\Delta U = Q - W, \tag{3.43}$$

where ΔU is the change in internal energy, Q is the heat transferred to (from) the system, and W the work done by (on) the system. When a system expands in a *quasistatic* process, the thermodynamic work done by the system on the surroundings is $\delta W = p\mathrm{d}V$, i.e. a product of pressure p and infinitesimal change in volume $\mathrm{d}V$. If the thermodynamic work is done on the system by the surroundings, the sign changes $\delta W = -p\mathrm{d}V$. An infinitesimal change in internal energy of the system is thus

$$\mathrm{d}U = \delta Q - p\mathrm{d}V, \tag{3.44}$$

where δQ marks the so-called *inexact differential* [66] of an infinitesimal amount of heat supplied to the system from its surroundings. Inexact differential is a differential whose integral is path dependent, i.e. $\int_{\gamma_1} \delta u \neq \int_{\gamma_2} \delta u$, for 2 different integrable paths $\gamma_1, \gamma_2 : [0, 1] \to \mathbb{R}$ such that $\gamma_1(0) = \gamma_2(0)$, $\gamma_1(1) = \gamma_2(1)$.

3.4.2 Second Law of Thermodynamics

Entropy, the measure of a system's disorder, plays a crucial role in the second law of thermodynamics [67]

$$\mathrm{d}S \geq \frac{\delta Q}{T}, \tag{3.45}$$

which holds equality for (a) quasistatic irreversible processes without a change in composition, (b) idealized reversible processes in closed systems, and inequality for irreversible processes in closed systems. The notation for infinitesimal amount of heat (δQ) and infinitesimal change in entropy ($\mathrm{d}S$) differ since entropy is a function of state while heat, like work, is not.

Second law of thermodynamics (3.45) states that in an *isolated* system, entropy *always* increases over time. Another, more traditional, interpretation states that heat always flows *spontaneously* from hotter to colder regions of matter, i.e. *downhill* in term of the temperature gradient.

3.4.3 Third Law of Thermodynamics

Third law of thermodynamics [68] states that the entropy of a closed system at thermodynamic equilibrium approaches a constant value when its temperature approaches absolute zero ($T = 0$ K). At absolute zero the system must be in a state with the minimum possible energy. It is equivalent to the claim that it is impossible by any procedure, no matter how idealized, to reduce the temperature of any closed system to 0 K in a finite number of finite operations [69].

3.4.4 Particle Changes in Closed Systems

For a closed system of different types of particles in which chemical reactions may occur, one has to account for the changes in the respective numbers of particles the system has. The fundamental relation for an infinitesimal change in system's internal energy becomes

$$\mathrm{d}U = T\mathrm{d}S - p\mathrm{d}V + \sum_i \mu_i \mathrm{d}N_i, \tag{3.46}$$

where $\mathrm{d}N_i$ is a small change in number of type-i particles, and μ_i the so-called *chemical potential* of the respective particles. Since S, V, and N_i are extensive[5] variables, an Euler relation [70] allows for simple integration yielding

$$U = TS - pV + \sum_i \mu_i N_i. \tag{3.47}$$

3.4.5 Helmholtz Free Energy

Helmholtz free energy (F) [71] is a thermodynamic potential that measures the useful work obtainable from a closed thermodynamic system at a constant temperature (isothermal work). It is defined as

$$F \equiv U - TS, \tag{3.48}$$

where U is the internal energy of the system, T the temperature, and S represents the entropy. It comes from the *Legendre transformation* [72] of internal energy U, in which T replaces S as the independent variable. From the first law of thermodynamics (3.44) and the second law of thermodynamics (3.45) for a reversible process, i.e. $\delta Q = T\mathrm{d}S$, we get

$$\mathrm{d}U = T\mathrm{d}S - p\mathrm{d}V. \tag{3.49}$$

Applying $\mathrm{d}\,(TS) = T\mathrm{d}S + S\mathrm{d}T$ and rearranging yields

$$\mathrm{d}\,(U - TS) = -S\mathrm{d}T - p\mathrm{d}V. \tag{3.50}$$

Definition (3.48) allows for

$$\mathrm{d}F = -S\mathrm{d}T - p\mathrm{d}V, \tag{3.51}$$

[5]Extensive properties depend on the amount of matter in a system – their magnitude is additive for subsystems. Examples include m, N, V, and S.

which is valid even for non-reversible processes, since F is a thermodynamic function of state (i.e. its integral does not depend on the integration path).

In case of chemical reactions, one must allow for changes in the numbers of particles N_i of each type i. Differential Helmholtz free energy then takes a form

$$\mathrm{d}F = -S\mathrm{d}T - p\mathrm{d}V + \sum_i \mu_i \mathrm{d}N_i, \tag{3.52}$$

where μ_i are chemical potentials of the corresponding particles. Such relation is again valid for both reversible and non-reversible changes.

3.4.6 Gibbs Free Energy

In contrast, Gibbs free energy (G) [73] is a thermodynamic potential that can be used to calculate the maximum amount of work, other than pressure-volume work, that may be performed by a thermodynamically closed system at constant temperature and pressure. Gibbs free energy is expressed as

$$G = U + pV - TS = H - TS, \tag{3.53}$$

where U is internal energy of the system, p and V are pressure[6] and volume, T is the system's temperature, S the entropy, and $H \equiv U + pV$ denotes the enthalpy [74] of the system (i.e. the total energy content of the system at constant pressure). It also provides a necessary condition for chemical reactions and similar processes that may occur under these conditions. From (3.48) and (3.53) it is clear that Gibbs free energy relates to Helmholtz free energy through

$$G = F + pV. \tag{3.54}$$

Analogically, an infinitesimal change in Gibbs free energy can be written as

$$\mathrm{d}G = \mathrm{d}U + p\mathrm{d}V - T\mathrm{d}S. \tag{3.55}$$

Taking into account infinitesimal change in internal energy (3.46) alongside G's total derivative

$$\mathrm{d}G = \mathrm{d}U + p\mathrm{d}V + V\mathrm{d}p - T\mathrm{d}S - S\mathrm{d}T, \tag{3.56}$$

one arrives at

$$\mathrm{d}G = V\mathrm{d}p - S\mathrm{d}T + \sum_i \mu_i \mathrm{d}N_i, \tag{3.57}$$

which is the Gibbs free energy total differential with respect to its natural variables p, T, and $\{N_i\}$. Because p and T are intensive[7] variables, $\mathrm{d}G$ cannot be integrated using Euler relations. Instead, one can simply substitute (3.47) into the definition (3.53) to arrive at [73]

$$G = \left(TS - pV + \sum_i \mu_i N_i\right) + pV - TS = \sum_i \mu_i N_i, \tag{3.58}$$

which shows that the chemical potential of a substance i is its (partial) mol(ecul)ar Gibbs free energy. This applies to homogeneous, macroscopic systems only [75].

[6] For mechanical equilibrium, p in the system has to be equal to that of the surroundings.

[7] Their magnitude is independent of the system's size.

3.5 Free Energy Calculations

Unveiling the free energy landscapes of molecular systems is crucial for understanding their thermodynamic properties. Free energy calculations allow us to gain the necessary theoretical insight which could not be acquired through experimental methods. Let us now delve into the computational toolbox for free energy calculations, covering methods such as *Thermodynamic Integration* (TI) and *Free Energy Perturbation* (FEP), as well as some crucial equalities of statistical thermodynamics unlocking valuable non-equilibrium approaches.

3.5.1 Free Energy in (Bio)chemistry and Biology

One of the most valuable forms of free energy in biochemistry and biology is the free energy of solvation for a given molecular system of interest. For carbon-based life as we know it water plays the role of natural solvent. To emphasise the importance of this substance in biological processes we refer to such property as *hydration* free energy instead. Though from know on some of the things will be colored in terms related to water, almost anything we cover can naturally be extended for any given solvent of choice.

The solvation (hydration) free energy of a molecular system provides insights into several important properties:

- *Solubility*: It is an indication of whether the molecule is likely to dissolve in a solvent or not. A lower solvation free energy suggests higher solubility.
- *Stability*: It helps assess the stability of the molecular system in a specific environment. More negative solvation free energy implies greater stability.
- *Chemical Reactivity*: Solvation free energy can influence the rate and thermodynamics of chemical reactions, as it affects the accessibility of reactants and transition states.
- *Hydrophobicity/Hydrophilicity*: Providing information about the molecule's affinity for water. More negative values indicate hydrophilic behavior, while positive values suggest hydrophobic behavior.
- *Protein-Ligand Binding*: In the context of drug discovery, it aids in predicting the binding affinity between a ligand and a target protein in a biological environment.
- *Conformational Changes*: Solvation free energy can influence the preferred conformations of molecules, especially in biological macromolecules like proteins and nucleic acids.

Since most of the biological processes occur at approximately constant temperature and pressure, the more relevant form of free energy for us is the Gibbs free energy G and its change ΔG during the thermodynamical transformations describing these events.

3.5.2 Geometrical vs. Alchemical Methods

In order to calculate free energies in a given thermodynamic system one can employ variety of different computational methods. The well established ones can generally be classified into 2 categories according to their strategies for free-energy estimation. One of them uses a *geometrical* transformation while the other is a class of the so-called *alchemical* transformations. Fig. 3.8 portrays their main differences using an example of methane molecule in water environment.

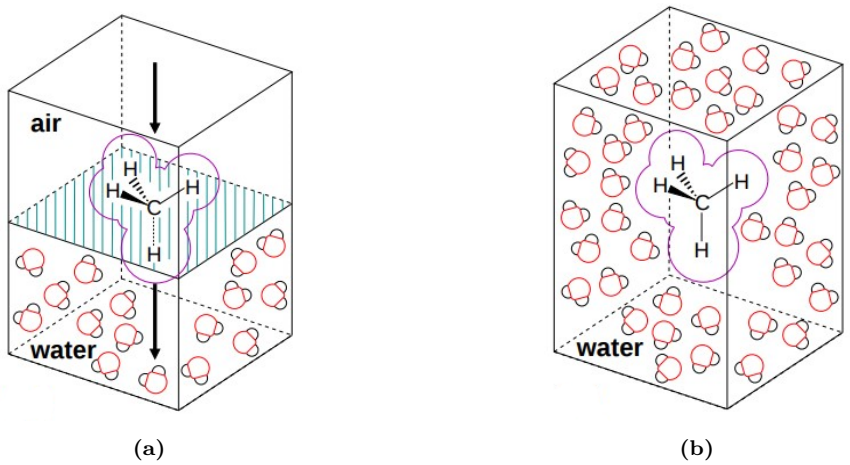


Figure 3.8: Exemplary case of hydration free energy calculation of methane molecule. (a) displays principles of the *geometrical* transformation, (b) shows the *alchemical* counterpart. Pictures adapted from [76].

In simulations based on geometrical transformations, the free energy is estimated by directly changing the system's geometry or environment and calculating the free energy difference from the potential of mean force (PMF) along the reaction coordinate. As can be seen from Fig. 3.8a, the molecule of interest is artificially pushed from reference environment (in this case air) to the relevant medium (here water). Since the investigated molecule is subjected to an influence of artificial forces, one has to treat the resulting data for their considerable effect.

Alchemical transformations involve gradually changing one molecule or state into another. For example, transforming a ligand in a binding site from one chemical species to another. Coming back to our example of hydrated methane, see Fig. 3.8b, there are no artificial forces present. What is being done is that the molecule of interest is gradually decoupled (force-wise) from the environment. At the beginning of the simulation the molecule does fully 'feel' its environment, while at the end it is completely decoupled and behaves like in a vacuum[8].

Due to the obvious advantages over geometrical transformations, alchemical methods gained a bit of traction in the field lately. Keeping that in mind, we chose to explore the nature of proteins and nucleic acids via one of those techniques. The rest of this theoretical section is thus reserved to alchemical methods only.

[8] As we will see later, this process can also go in reverse.

3.5.3 Thermodynamic Cycles

For large and complex systems the scale of proteins and nucleic acids, it is convenient to utilize suitable thermodynamic cycles in order to find and choose the best possible transformation path. Let us take a look at this issue through the scope of an example, illustrated in Fig. 3.9. Here, we are trying to assess the most desirable path towards calculating the binding affinity difference of 2 distinct ligands, a and b, to a given protein. Thermodynamic cycle in Fig. 3.9 yields 2 possible ways of computing the binding affinity difference

$$\Delta\Delta A_{\text{bind}} = \Delta A^{b}_{\text{bind}} - \Delta A^{a}_{\text{bind}} = \Delta A^{\text{B}}_{a\to b} - \Delta A^{\text{UN}}_{a\to b}. \tag{3.59}$$

This means that we could either directly remove each ligand from the binding site by gradual decoupling from the protein environment, or perform a transformation of one molecule into the other both inside the protein's cavity and without the protein's presence (i.e. in the solvent itself).

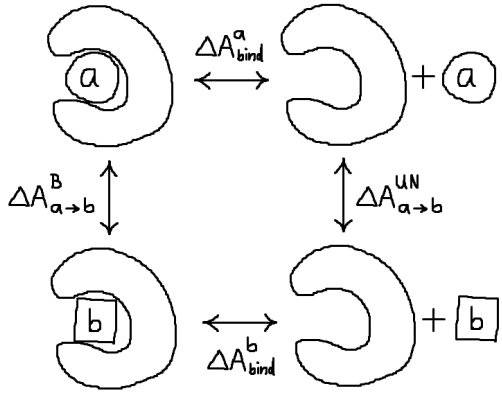


Figure 3.9: Example of applying a thermodynamic cycle to find which molecule, a or b, has a stronger binding affinity to a given protein. One could either remove both molecules from the protein's pocket, or simply transform one molecule to another both inside the protein and without it to find the same value.

Now, which path is more favourable, and why? First equality in Eq. (3.59) would generally lead to a potentially larger disturbance of the system, that is if molecules a and b are significant in size. Imagine that a sizable molecule simply disappears (interaction-wise) from the system. In that case it leaves a considerable vacancy, which is non-negligible and inherently nonphysical, leaving our results negatively affected. On the other hand, second equality in Eq. (3.59) offers a potential alleviation from such undesirable effects, since molecules a and b might share some common motif from which a hybrid topology could be constructed. This way, we would just mutate one molecule into the other inside the protein and then without it to land at the desired free energy differences with less negative impact on the systems overall behavior.

This is juts a single example of many, which could be put to use in practise. One could for instance explore other aspects of a given simulation, such as calculation efficiency of each path. Exploration of alternative thermodynamic pathways, their efficiency and possible effects on the system at hand, can readily be used to study binding properties of protein-DNA complexes, drugs in target sites, or other possibly macromolecular chemical systems.

3.5.4 Free Energy and Statistical Partition Function

Most free energy calculation methods start from a single core equation derived from statistical mechanics, i.e. the relation between free energy and the statistical partititon function

$$A = -\beta^{-1} \ln \mathcal{Z}, \tag{3.60}$$

where $\beta \equiv (\mathrm{k_B}T)^{-1}$, and $\mathcal{Z}$ marks an appropriate partition function, based on the statistical ensemble at play.

For the canonical (NVT) ensemble, corresponding to the Helmholtz free energy F, the partition function takes on a shape [77, 78]

$$Z_{NVT} \propto \int_{\Gamma} \exp(-\beta \mathcal{H}(\mathbf{x}, \mathbf{p}_x))\, \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{p}_x, \tag{3.61}$$

where Γ is the phase space volume over which we sample, and $\mathcal{H}(\mathbf{x}, \mathbf{p}_x)$ is the Hamiltonian of the system featuring canonical positions $\mathbf{x}$ and momenta $\mathbf{p}_x$. Related to the Gibbs free energy G is the isothermal-isobaric (NpT) ensemble, to which a partition function [78, 79]

$$Z_{NpT} \propto \int \mathrm{d}V \exp(-\beta p V) Z_{NVT} \tag{3.62}$$

can be assigned.

Relation (3.60) can be utilized to describe free energy differences

$$\Delta A = A_b - A_a = -\beta^{-1} (\ln \mathcal{Z}_b - \ln \mathcal{Z}_a) = -\beta^{-1} \ln\left(\frac{\mathcal{Z}_b}{\mathcal{Z}_a}\right) \tag{3.63}$$

between 2 thermodynamic states a and b of a given system.

3.5.5 Alchemical Transformations

Let us have a transformation $a \to b$ of a chemical object between 2 thermodynamic states a and b. Considering generic reaction coordinate of the system, one can characterize every point along the coordinate path by a parameter λ, with $\lambda = 0$ and $\lambda = 1$ corresponding to 2 ensembles of microstates for which the reaction coordinate is constrained to different values. Transformation $a \to b$ is referred to as the *alchemical transformation* with perturbation parameter λ. *Forward* transformation is thus externally driven process $\lambda : 0 \to 1$, while its time reversal path $\lambda : 1 \to 0$ is referred to as *backward* transformation. As such, states a and b are represented by the distributions of microstates having $\lambda = 0$ and $\lambda = 1$, respectively.

Using alchemical parameter λ, one could construct hybrid Hamiltonian of the system as a linear combination

$$\mathcal{H}(\mathbf{x}, \mathbf{p}_x; \lambda) = \mathcal{H}_0(\mathbf{x}, \mathbf{p}_x) + \lambda \mathcal{H}_b(\mathbf{x}, \mathbf{p}_x) + (1 - \lambda)\mathcal{H}_a(\mathbf{x}, \mathbf{p}_x), \tag{3.64}$$

where $\mathbf{x}$ and $\mathbf{p}_x$ are the canonical positions and momenta of the particles. $\mathcal{H}_0$ is the Hamiltonian describing the unperturbed part of the system, i.e. atoms that do not undergo any transformation. Interactions of atomic groups of the initial and the final state with the rest of the system are represented using corresponding subscripts a, b.

3.5.6 Thermodynamic Integration (TI)

One of the well-known alchemical methods, and probably the most common one, is the so-called *Thermodynamic Integration* (TI). For simplicity, let us work within the canonical (NVT) framework. The starting point is Eq. (3.60) which we take a partial derivative of with respect to the parameter λ

$$\frac{\partial A}{\partial \lambda} = -\beta^{-1} \frac{\partial}{\partial \lambda} \ln \int e^{-\beta \mathcal{H}(\mathbf{x},\mathbf{p}_x;\lambda)} \, \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{p}_x = -\beta^{-1} \frac{\frac{\partial}{\partial \lambda} \int e^{-\beta \mathcal{H}(\mathbf{x},\mathbf{p}_x;\lambda)} \, \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{p}_x}{Z}. \quad (3.65)$$

This can be written as

$$\frac{\partial A}{\partial \lambda} = -\beta^{-1} \frac{-\beta \int \frac{\partial \mathcal{H}(\mathbf{x},\mathbf{p}_x;\lambda)}{\partial \lambda} e^{-\beta \mathcal{H}(\mathbf{x},\mathbf{p}_x;\lambda)} \, \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{p}_x}{Z} = \left\langle \frac{\partial \mathcal{H}(\mathbf{x},\mathbf{p}_x;\lambda)}{\partial \lambda} \right\rangle_\lambda . \quad (3.66)$$

Finally, one can do integration over the whole range of λ to arrive at the final TI equation [80, 81]

$$\Delta A = \int_0^1 \left\langle \frac{\partial \mathcal{H}(\mathbf{x},\mathbf{p}_x;\lambda)}{\partial \lambda} \right\rangle_\lambda \mathrm{d}\lambda. \quad (3.67)$$

It is a comparison of free energy levels between 2 given states a and b, whose potential energies have generally different dependencies on the spatial coordinates. Since the free energy of a system is not simply a function of the system's phase space coordinates, but rather a function of Boltzmann-weighted integral over the phase space (i.e. partition function) [82], ΔA cannot be calculated directly from potential energies of just 2 coordinate sets for states a and b. The free energy difference is thus computed using a defined thermodynamic path[9] connecting the states, and integrating ensemble-averaged changes in enthalpy along this path.

The above derivation points to a rather simple way of estimating free energies. The inside of the integral (3.67), the derivative, can be calculated from information of just a single state. However, since we can only perform simulations at a finite number of λ states, numeric integration schemes are required. In practise, integral (3.67) is performed as [83, 84]

$$\Delta A \approx \sum_{i=1}^{N} w_i \left\langle \frac{\partial \mathcal{H}(\mathbf{x},\mathbf{p}_x;\lambda)}{\partial \lambda} \right\rangle_i , \quad (3.68)$$

where the weights $\{w_i\}$ will depend on which numeric integration style is chosen (e.g. trapezoid rule).

3.5.7 Free Energy Perturbation (FEP)

An alternative approach is the so-called *Free Energy Perturbation* (FEP), or alternatively *Exponential Averaging* (EXP). It is one of the earliest free energy methods available in this field. Let us again, for simplicity, work within the canonical (NVT) ensemble. Starting from free energy difference (3.63), adding and subtracting $\beta \mathcal{H}_a(\mathbf{x},\mathbf{p}_x)$ inside the exponential of the upper partition function $\mathcal{Z}_b$, and rearranging the exponentials, we get

$$\Delta A_{a\to b} = -\beta^{-1} \ln \left[\frac{\int e^{-\beta(\mathcal{H}_b(\mathbf{x},\mathbf{p}_x)-\mathcal{H}_a(\mathbf{x},\mathbf{p}_x))} e^{-\beta \mathcal{H}_a(\mathbf{x},\mathbf{p}_x)} \, \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{p}_x}{\mathcal{Z}_a} \right]. \quad (3.69)$$

[9] Such paths can either be alchemical or real chemical processes.

Following the statistical definition of an average value, the free energy difference is governed by the *Zwanzig equation* [85, 62]

$$\Delta A_{a\to b} = -\frac{1}{\beta} \ln \left\langle \exp\{-\beta \left[\mathcal{H}_b(\mathbf{x}, \mathbf{p}_x) - \mathcal{H}_a(\mathbf{x}, \mathbf{p}_x)\right]\}\right\rangle_a , \tag{3.70}$$

where $\beta \equiv (\mathrm{k_B}T)^{-1}$, and the angular brackets denote an ensemble average over configurations representing the initial (reference) state a, i.e. averaging over the simulation run for state a.

Zwanzig relation (3.70) reveals a two state method, estimating free energies straightforwardly from 2 endpoints a and b we are trying to calculate the free energy difference for. Although this is an exact solution and probably the simplest free energy method to understand, it is also one of the poorest methods in terms of efficiency. Convergence of Eq. (3.70) relies on low-energy configurations of the target state b forming a subset of configurations corresponding to the reference state a. If the ensembles are too disparate, Eq. (3.70) will not converge. This issue is depicted in Fig. 3.10. Difficulties reflected in Fig. 3.10a can be alleviated by constructing a thermodynamic path which takes the system through a set of intermediate states, improving the phase space overlap.

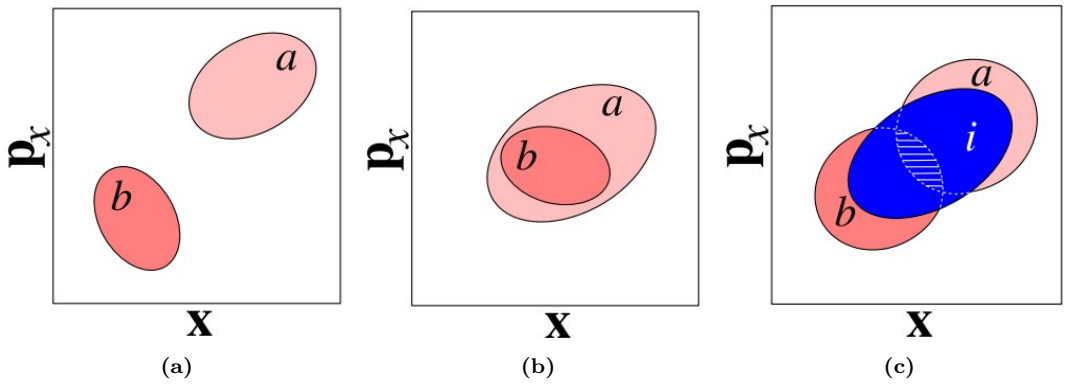


Figure 3.10: Convergence of FEP calculations. (a) ensembles are too separated, Eq. (3.70) will not converge. (b) ideal scenario, where configurations of state b form a subset of the ensemble belonging to state a – the simulation is expected to converge. (c) helping convergence using an overlap of non-physical intermediate states i connecting ensembles a and b. Pictures adapted from [76].

3.5.8 Pathway of Intermediate States

The phase space overlap of two states of interest a and b can be almost non-existent. As a result, free energy calculations for the two states alone will either suffer significant errors or end up not converging at all, recall Fig.3.10a. Such an issue is even more pronounced for more complicated molecules and transformations between them. That being said, Zwanzig relation (3.70) converges relatively good only if the unsampled target state b is a subset of the reference state a, see Fig. 3.10b (e.g. inserting a rigid molecule into a dense fluid).

Since free energy is a state function, one can construct a thermodynamic path taking our system through series of intermediate states connecting both endpoints

of interest. Putting this into a mathematical perspective, we can improve the convergence of our calculations by creating high phase space overlap intermediates and calculate the free energy difference as

$$\Delta A_{a\to b} = \sum_{i=1}^{N} \Delta A_{i\to i+1}. \tag{3.71}$$

This is illustrated in Fig. 3.10c. Note that our intermediate states do not have to be real, experimentally observable states. In fact, due to the actions of parameter λ, recall Hamiltonian (3.64), the vast majority of cases feature non-physical intermediate states. They are a mere computational tool to aid a proper convergence.

Let us immerse this into the FEP framework completely. System subjected to FEP transformation, whether forward or backward, goes from a to b (or vice versa) through series of non-physical intermediate states along a well-defined pathway connecting a and b. This path is characterized by the alchemical parameter λ, introduced in Hamiltonian (3.64). The free energy is thus a continuous function of parameter λ on the pathway connecting a to b, and the free energy difference reads

$$\Delta A_{a\to b} = -\frac{1}{\beta} \sum_{i=1}^{N} \ln \left\langle \exp\{-\beta \left[\mathcal{H}(\mathbf{x}, \mathbf{p}_x; \lambda_{i+1}) - \mathcal{H}(\mathbf{x}, \mathbf{p}_x; \lambda_i)\right]\}\right\rangle_i , \tag{3.72}$$

where N represents the number of intermediate stages (or λ windows) between the initial and the final state.

Practically speaking, in each one of the independent[10] windows, equally separated by increment λ, the system undergoes FEP calculation procedure to evaluate the free energy difference up until that particular point of the transformation. This is done in an open simulation setup (similar to random walker), averaging the values out for each λ window. Collected non-physical intermediate states i then form an ensemble mutually overlapping generally disparate ensembles belonging to states a and b – see Fig. 3.10c. This allows for successful convergence to the free energy difference we seek.

Now, simple application of Eq. (3.72) searches for the desired free energy value directly. This means that through 2 possible paths, forward or backward with respect to parameter λ (i.e. $a \rightleftarrows b$), the calculation is set up to land exactly at the equilibrium free energy value, within the error of the method used. Such an approach is not always an easy piece of cake, especially for larger structures like nucleic acids and proteins. Computational efficiency of Eq. (3.72) drops dramatically with increasing size of the simulated system. Running calculations in a simple setup like this can cost a lot of valuable resources and time.

[10]The process can be trivially parallelized by executing each window on a separate CPU, as there is no need for communication between the simulation of one window and the next.

3.6 Non-equilibrium Approach

There are many different routes towards the estimation of ΔA, which are in their core computationally more effective in comparison to brute-force application of TI (3.68) or FEP (3.72) formulae. An example could be the *Bennett Acceptance Ratio* (BAR) [86], which benefits from both forward and backward transformations. Unlike BAR, which draws upon equilibrium simulations, we intend to tap into the power of non-equilibrium thermodynamics. Since the method we are about to present is not well established, compared to techniques the likes of BAR, we intend to call it the *Non-equilibrium Overlap Sampling* (NOS).

3.6.1 Free Enegy and Work

Consider a finite classical system in contact with a heat reservoir. When some external parameters[11] of the system are made to change with time, a work W is performed on the system. The external parameters are switched along some path γ in phase space from a to b at a *finite* rate t_s. The work will on average exceed the equilibrium free energy difference $\Delta A \equiv A_b - A_a$ between the initial and final states a and b [87], i.e.

$$\overline{W} = \int \mathrm{d}W \rho(W; t_\mathrm{s}) W \geq \Delta A, \tag{3.73}$$

where the overline denotes an average over an ensemble of measurements of W, each made after the other allowing the system and the heat bath to equilibrate at temperature T. Individual values of W depend on the microscopic initial conditions of the system and the reservoir. Difference $\overline{W} - \Delta A \equiv W_\mathrm{D}$ is known as the *dissipated work* associated with the increase of entropy during an irreversible process. Equality of (3.73) holds only in the case of *quasistatic process*, i.e. taking the system from a to b *infinitely slowly* such that all intermediate states are in thermodynamic equilibrium. For quasistatic processes W_D vanishes.

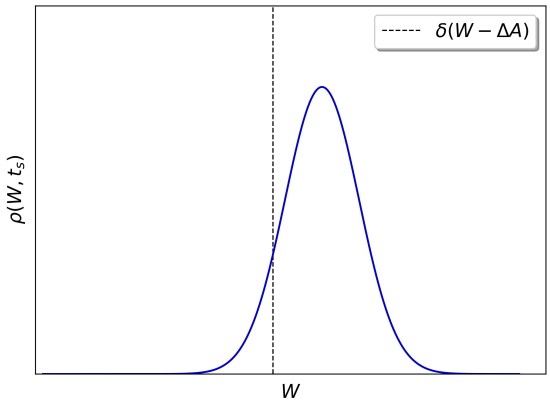


Figure 3.11: Distribution $\rho(W; t_\mathrm{s})$ of work values acquired through independent switching measurements at a set switching time t_s. Dashed line represents a δ function at $\overline{W} = \Delta A$, corresponding to $t_\mathrm{s} \to \infty$. Note that ensemble average $\overline{W}$ exceeds ΔA due to energy dissipated in a finite-time, irreversible process.

[11] For example volume V within which the system is confined, strength of an external field, or some particle-particle interactions modulated throughout an MD simulation.

A simple illustration of Eq. (3.73) can be made, see Fig. 3.11. Let us have many independent processes switching between 2 states a and b during a given switching time t_s. Performing measurements of work W for each transformation results in work distribution $\rho(W;t_\mathrm{s})$, parametrically modulated by t_s. For a finite value of t_s, ensemble average $\overline{W}$ will *always* exceed equilibrium free energy difference ΔA due to energy dissipated during such irreversible transformations. Increasing t_s shifts $\overline{W}$ closer to ΔA. Only in a limit case $t_\mathrm{s}\to\infty$, $\overline{W}=\Delta A$.

In practise, one would perform a number of simulations of slow switching $a\to b$, and the work W obtained from each simulation would then be treated as an estimate of the free energy difference ΔA. Such an estimate contains not only statistical errors (W differs from one simulation to another), but also systematic errors (as per the above inequality, any finite-rate simulation has a bias) [87]. Averaging over many simulations eliminates statistical errors, however the systematic error remains. $\overline{W}$ thus represents an upper bound of ΔA.

3.6.2 Jarzynski Identity

It was not until the very end of the last century that new expressions, relating non-equilibrium work and free energy, were discovered. The first of which relates an ensemble average over an exponential Boltzmann distribution of work with free energy, the so-called *Jarzynski equation* [87, 88]

$$\overline{e^{-\beta W}} = e^{-\beta\Delta A}. \tag{3.74}$$

Here again $\beta\equiv(\mathrm{k_B}T)^{-1}$, and $\Delta A\equiv A_\mathrm{b}-A_\mathrm{a}$ is the equilibrium free energy difference between 2 thermodynamic states a and b of a given system. This equality remains valid for all paths γ connecting a to b, independent of the rate at which the external parameters (mentioned above) are switched along the path. Jarzynski derived this relation under the usual assumption of weak coupling between the system and its heat bath, but otherwise it follows directly from Hamilton's equations [87]. At its core, the equality highlights that fluctuations in the work adhere to specific constraints separately from the average work, recall Eq. (3.73).

Left-hand side of Eq. (3.74) averages over all possible realizations of an external process that takes the system from the equilibrium state a to a generally non-equilibrium state under the same external conditions as that of the equilibrium state b. In other words, it is an average over different fluctuations that could occur during the process, each of which will result in a slightly different value of the work W done on the system at hand. The work is done on the system through inequality $\overline{W}\geq\Delta A$, which immediately falls out of (3.74) when considering mathematical identity $\overline{\exp(x)}\geq\exp(\overline{x})$ [89]. Unlike this inequality, Jarzynski holds no matter how fast the process happens.

Jarzynski (3.74) implies that using $e^{-\beta W}$ as an estimate for $e^{-\beta\Delta A}$ (instead of W for ΔA) is an *unbiased* estimate – there are *only* statistical errors [87]. One can therefore use an exponential average $W^\mathrm{x}\equiv-\beta^{-1}\ln\overline{\exp(-\beta W)}$, rather than the standard average $\overline{W}$, as an estimate for ΔA. The overline now marks an average over a finite number of simulations N_s. It can be shown [88] that the systematic error in W^x is smaller than the one of $\overline{W}$, and vanishes for $N_\mathrm{s}\to\infty$. Using W^x, rather than $\overline{W}$, therefore provides a tighter upper bound of ΔA. Reversing the direction ($a\leftarrow b$) establishes the lower bound as well.

Relation (3.74) provides a statistical mechanics framework to compute equilibrium free energy differences between 2 states from measurements of *irreversible* work along an ensemble of trajectories joining these states, i.e. as

$$\Delta A = -\beta^{-1} \ln \int \mathrm{d}W \rho(W, t_\mathrm{s}) e^{-\beta W}, \tag{3.75}$$

valid for any switching time t_s and any path γ joining both states, cf. Eq. (3.73).

In principle, Eq. (3.75) is exact for any path between equilibrium states. However, in practise it is highly computationally demanding to evaluate exponential averages. Additionally, since we take an exponential of work, rare, negative work events may have devastating effects on the resulting free energy value. For that reason, Jarzynski averaging needs a vast number of simulations in order to yield statistically significant results.

3.6.3 Crooks Fluctuation Theorem (CFT)

An alternative way to bypass the systematic error of inequality (3.73) is to combine the information from both forward and backward transformations. We again consider a finite classical system coupled to large, equilibrium, thermodynamic baths (e.g. T, p). The system is driven by a (possibly time-dependent) process out of equilibrium. The utilization of both transformation directions is allowed through the so-called *Crooks Fluctuation Theorem* (CFT) [90], which in its most general form reads

$$\frac{P_\mathrm{f}(+\omega)}{P_\mathrm{b}(-\omega)} = e^{+\omega}, \tag{3.76}$$

where ω is an *entropy production* of the driven system measured over some time interval, $P_\mathrm{f}(\omega)$ is the probability distribution of the entropy production, and $P_\mathrm{b}(\omega)$ corresponds to the system driven in a time-reversed manner.

Relation (3.76) was derived as a somewhat generalized version of an entropy production fluctuation theorem (EPFT) [90, 91], for stochastic, *microscopically reversible* dynamics. Unlike most relations of non-equilibrium statistical dynamics, valid *only* in the near-equilibrium (linear) regime, EPFTs present a group of exceptions valid for systems perturbed *arbitrarily far* from equilibrium. In that sense, CFT follows their applicability.

Let us define a particular work process by the phase-space distribution $\rho(x_{-\tau})$ at time $-\tau$. Here, for the purpose of simple notation, we consider $x \equiv \{\mathbf{p}_i, \mathbf{q}_i\}_{i=1}^N$ to be the set of all positions and momenta of an N-particle system. Each and every realization of this process is thus given by the path $x(t) \equiv \gamma$ the system follows through phase space. Entropy production ω must be a functional of this path [90]. Associated with this process there must be a change in entropy ΔS due to interactions of our system with the baths. The form of ΔS depends on conditions the system with the heat bath is in, i.e. whether it can be described by canonical, NpT, or some other ensemble. One should also consider the change in entropy associated with the change in the microscopic state of the system, the so-called *information entropy*. Entropy of a microscopic state of a system is given as $s(x) = -\ln \rho(x)$ [92], and is a measure of information required to describe the state occurring with probability $\rho(x)$. Thus the (non-equilibrium) ensemble of microstates describing the system has entropy $S = -\sum_x \rho(x) \ln \rho(x)$.

Assuming a single realization of such process taking the system from $\rho(x_{-\tau})$ at time $-\tau$ to a final state described by $\rho(x_{+\tau})$ at some future time $+\tau$, the general shape of ω would be [90]

$$\omega = \ln \rho(x_{-\tau}) - \ln \rho(x_{+\tau}) + \Delta S, \tag{3.77}$$

which describes the change in the amount of information necessary to characterize the microstate of the system plus the change in entropy due to system-bath interactions. In the following we will explore different forms of ΔS based on distinct possible scenarios, and its consequences on the shape of CFT formula.

Entropy change due to system-bath interactions, ΔS, depends on how many and what kinds of baths our system is coupled to. As we mentioned earlier, all baths are considered to be large, equilibrium, thermodynamic systems. The most commonly considered one would be the heat bath. In such a case $\Delta S = -\beta Q$ is the change in the entropy of the bath, where Q is the amount of energy that flows out of the heat reservoir and into the system. The system would then be described by the *canonical* ensemble, and the free energy difference associated with the transformation would be of the Helmholtz free energy, i.e. ΔF. If one were to consider an isothermal-isobaric system, the system would have to be coupled to a volume bath as well, and $\Delta S = -\beta Q - \beta p \Delta V$. Here p is the pressure and ΔV the change in volume of our system. The free energy difference would then be of the Gibbs free energy instead, i.e. ΔG. It is possible to extend the application of CFT formula (3.76) for systems with any standard set of baths, so long as the microscopic reversibility holds true and the definition of the entropy production ω is consistent. We can thus happily continue using the general ΔA.

Tying it back to our previous sections, transformation $a \to b$ of the system, driven from equilibrium state a to a generally non-equilibrium state under the same external conditions as that of the equilibrium state b, is called the *forward* transformation[12]. Such a transition is carried out by a controlled change in alchemical parameter $\lambda(t)$, recall Section 3.5.5. The system is in the equilibrium state a in $t \in (-\infty, -\tau]$. From $t = -\tau$ the system is actively perturbed up to $t = +\tau$. In $t \in [+\tau, +\infty)$ the system is allowed to relax and reach equilibrium. The perturbation is thus executed in a finite amount of time. For $\lambda : 0 \to 1$, ensembles of states a and b correspond to $\lambda = 0$ and $\lambda = 1$, respectively. It can be shown [90] that the entropy production of such forward transformation, $a \to b$, reads

$$\omega_{\mathrm{f}} = -\beta \Delta A + \beta W_{a \to b}, \tag{3.78}$$

where $\Delta A = A_b - A_a$ is the free energy difference between the states, and $W_{a \to b}$ is the work done on the system during the forward transformation.

For the case of a single transformation between 2 microstates a and b, plugging entropy production (3.78) into CFT relation (3.76), leads to

$$\frac{P(a \to b)}{P(a \leftarrow b)} = \exp[\beta (W_{a \to b} - \Delta A)], \tag{3.79}$$

where $P(a \to b)$ is the joint probability of taking equilibrium microstate a from the ensemble corresponding to $\lambda = 0$ and performing the *forward* transformation

[12]Remember that the process, driving our system out of equilibrium, can possibly be time-dependent. For that reason alone it is important to generally distinguish between the forward and the backward route.

to the microstate b corresponding to $\lambda = 1$. Similarly $P\left(a \leftarrow b\right)$ is its counterpart reversing the process via the *backward* route. Difference $W_{a\to b} - \Delta A$ in Eq. (3.79) can be interpreted as the work W_{D} dissipated during the forward transformation. If the transformation is performed *infinitely slowly*, $P\left(a \to b\right) \equiv P\left(a \leftarrow b\right)$, and thus $W_{a\to b} = \Delta A$. In other words, for equilibrium transformations the dissipated work W_{D} vanishes.

For systems, satisfying that ω is odd under time reversal, one can entertain a relation $W_{a\to b} = -W_{a\leftarrow b}$, and group together all the trajectories yielding the same work W in the forward and the backward transformation to arrive at

$$\frac{P_{\mathrm{f}}\left(W\right)}{P_{\mathrm{b}}\left(-W\right)} = e^{\beta(W-\Delta A)}, \tag{3.80}$$

which is the work-formulated CFT we intend to use in our calculations. As such, we are rewriting Eq. (3.79) in terms of probability distribution $P_{\mathrm{f}}\left(W\right)$ of work W exerted by a random forward trajectory $a \to b$, and $P_{\mathrm{b}}\left(-W\right)$ taking a random backward route.

Formulation (3.80) implies that the 2 work distributions cross at $W = \Delta A$, meaning that the non-equilibrium forward and backward runs yield the equilibrium free energy difference ΔA right at this very intersection. Illustration is provided in Fig. 3.12. This phenomenon has been experimentally verified [93] using optical tweezers for the process of unfolding and refolding of a small RNA hairpin. Practical implications related to the finding of these equilibrium values in real non-equilibrium simulation data will be covered in the following chapter, featuring our methodology.

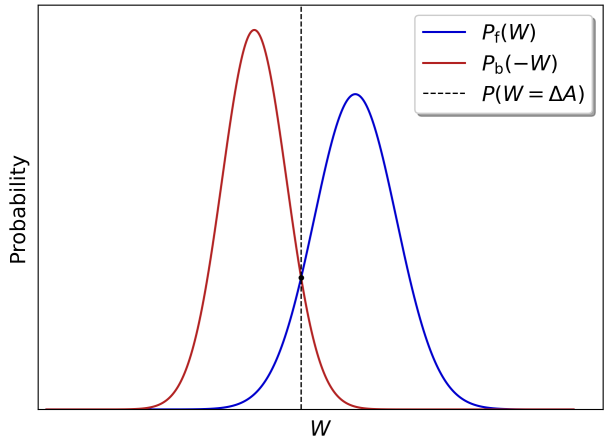


Figure 3.12: Demonstration of CFT, Eq. (3.80). $P_{\mathrm{f}}\left(W\right)$ is the probability distribution of work W exerted by a random forward trajectory $a \to b$. $P_{\mathrm{b}}\left(-W\right)$ is its time-reversed counterpart taking a random backward route, $a \leftarrow b$.

Before we move onto the next section, note that whenever Eq. (3.76) holds, the relation

$$\overline{e^{-\omega}} = \int_{-\infty}^{+\infty} P_{\mathrm{f}}(+\omega)e^{-\omega}\mathrm{d}\omega = \int_{-\infty}^{+\infty} P_{\mathrm{b}}(-\omega)\mathrm{d}\omega = 1 \tag{3.81}$$

remains true [90]. If one considers systems that start in equilibrium, of which $\omega = -\beta\Delta A + \beta W$, Jarzynski equality (3.74) readily falls out. An important move here is to acknowledge that ΔA is a state function and can thus be taken outside the average. CFT therefore implies Jarzynski relation.

3.6.4 Mounting CFT on FEP Framework

Our approach (NOS) mounts the principles of CFT equality (3.80) onto the FEP framework given by the modified Zwanzig formula (3.72). We launch a series of many short-lasting non-equilibrium FEP simulations (runs), the results of which have no reporting value on their own whatsoever. However, when grouped together and plotted as probability distributions the form of histograms, the equilibrium free energy value starts surfacing up. This allows us to achieve a great level of parallelization since all the non-equilibrium FEP runs can run in parallel to each other, given sufficient amount of computational resources.

In practise, alchemical parameter $\lambda(t) : 0 \rightarrow 1$ has to be varied in a discrete manner throughout each FEP simulation. Partition into λ windows is provided by the user, and is in the vast majority of cases done equidistantly by some integer N. An example of a forward transformation $a \rightarrow b$ is illustrated in Fig. 3.13; backward shift would just reverse the course of parameter λ. In each window the value of parameter λ is fixed, and the system is evolving under the Hamiltonian (3.64) in an open MD simulation for # number of (FEP) steps. Individual free energy differences (samples) between the current $(i + 1)$ and the preceding (i) intermediate state are computed according to Zwanzig equation (3.70). The protocol samples different possible microstates of the system as the simulation progresses (sampling through the underlying MD simulation) and averages the free energy differences acquired from Zwanzig relation 'on the fly' to arrive at a single $\Delta A_{i \rightarrow i+1}$ for each window. Grouping all the sampled states from all the windows together effectively creates the ensemble of non-physical, intermediate states forging the overlap between generally disparate ensembles of endpoint states a and b, recall convergence issue of FEP calculations in Fig. 3.10.

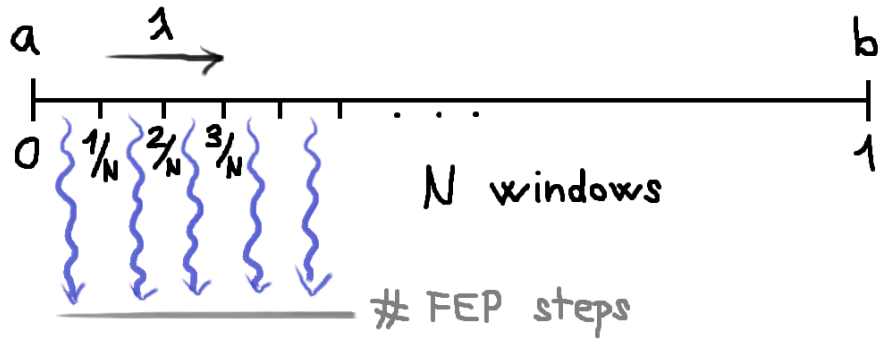


Figure 3.13: Sketch representing a single forward FEP run, transforming a given system between states a and b. Alchemical parameter λ is varied between 0 and 1, creating N intermediate windows. In each window the value of λ is fixed, and the system is evolving in an open MD simulation for # number of FEP steps, in parallel to all the other λ windows. This allows us to sample the ensemble of intermediate states, facilitating a successful convergence of the transformation.

The final (non-equilibrium) free energy difference $\Delta A_{a \rightarrow b}$ of a single realization of $a \rightarrow b$ is acquired through accumulation, Eq. (3.71) , of these partial differences $\Delta A_{i \rightarrow i+1}$. As such, we utilize the FEP formula (3.72), but for a non-equilibrium transformation instead (i.e. running FEP for 'insufficient' amount of time).

We run many such short-lasting, non-equilibrium FEP runs in both directions (we are talking about $10^2 - 10^3$ simulations per direction). An illustration of NOS simulation scheme is shown in Fig. 3.14. First, we perform energy minimization and equilibration of the given system, to reach the desired equilibrium state a. The reference state a is then fed into each and every forward run as an initial configuration. To ensure unique starting points for each run, we perform an additional short equilibration prior to each FEP simulation. After forward transformations are completed, final states b are fed into backaward runs, and the whole process is reversed. Collected data is processed, and the resulting values of non-equilibrium work, done one the system throughout each transformation, then form probability distributions $\rho(W; t_\mathrm{s})$, where $\overline{W_\mathrm{f}}$ and $\overline{W_\mathrm{b}}$ represent the upper and lower bounds on equilibrium free energy difference $\Delta A = A_b - A_a$ respectively, as per Eq. (3.73). Following CFT (3.80), the equilibrium free energy difference ΔA is then extracted from the intersection of the forward and the backward distributions. Technical details of the data processing and analysis will be covered in the following chapter, featuring our methodology.

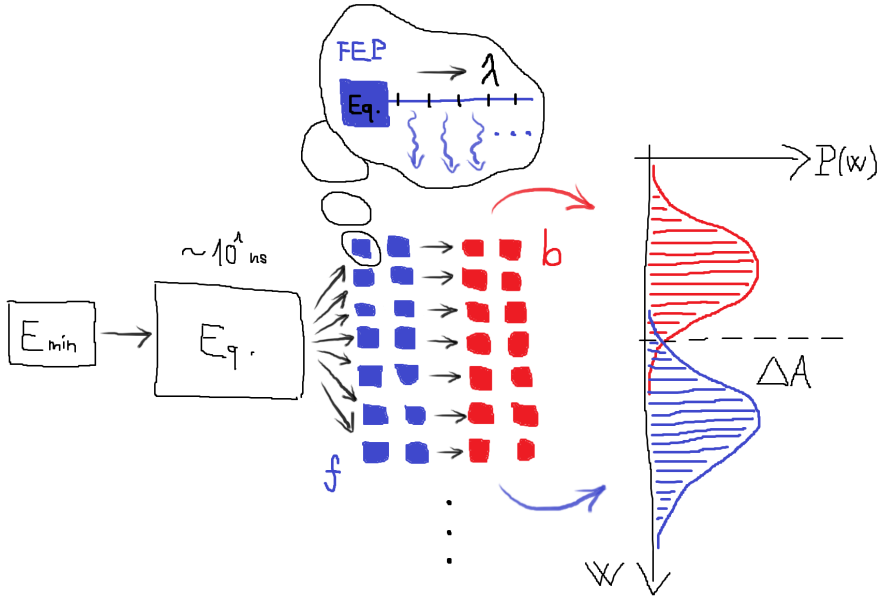


Figure 3.14: Sketch of NOS simulation scheme. First, minimization and equilibration of the system are performed. The equilibrium state a is then fed into each and every forward run. Prior to FEP simulation, a short equilibration is carried out to ensure different starting points for each run. Final states b are then fed into backaward runs, and the whole process is reversed. Data is collected, processed, and plotted as work probability distributions, overlapping around $\Delta A = A_b - A_a$.

As we explored earlier, switching time t_s affects the upper and lower bounds of equilibrium difference ΔA, generated by distributions $\rho(W; t_\mathrm{s})$. Switching time t_s amounts to the total simulation time per run, and is thus represented by the number of FEP steps performed, i.e.

$$t_\mathrm{s} \propto \sum_{i=1}^{N} \#_i = N \cdot \#, \tag{3.82}$$

where N is the number of λ windows, in which $\#_i$ number of FEP steps were performed. Typically, the number of FEP steps is the same for every λ window

(i.e. $\#_i = \#$), hence the sum in Eq. (3.82) simplifies to a mere multiplication. Raising switching time t_s forces $\overline{W}$ closer to ΔA, and for $t_\mathrm{s} \to \infty$ we get $\overline{W} = \Delta A$; recall quasistatic limit of relation (3.73). One can thus modulate the degree of systematic error (made through estimating ΔA with $\overline{W}$) simply by changing $\#$. In other words, by raising $\#$ of each run the overlap of both distributions will increase, improving the precision of NOS.

It is important to note that for large $\#$ FEP simulations are approaching equilibrium simulations, eventually rendering NOS counterproductive. In order for NOS to be truly effective, a balanced tuning of $\#$ alongside N and the number of non-equilibrium runs has to be made.

4. Methodology

Every well-behaved diploma thesis ought to have its methods adequately described and explained. This chapter does exactly that. We cover our sources of computational power alongside general simulation setup and analysis technique to cultivate a solid ground to stand on later throughout the rest of the thesis. Common simulation setup and methods of analysis are shared among all of our (bio)chemical systems of interest. Every following chapter thus contains only a light section further specifying features unique to the systems at hand.

4.1 Building of Simulated Systems

Setting up our calculations is a tedious and time consuming task, requiring many consecutive steps. Same as any experimentalist has to prepare his samples before any measurement can be performed, we need to create an initial structure of the chemical system we are trying to simulate. This can go one of two ways, depending on the complexity of our system – using experimentally determined structures, or by manual labor.

4.1.1 Protein Data Bank (PDB)

For macromolecules (proteins, nucleic acids etc.) manual modeling is practically impossible. In that case we have to turn our sight to experimentally determined[1] structures provided in RCSB Protein Data Bank (PDB) [94]. There is an enormous amount of crystallized molecular structures stored here, each coded by its unique alphanumeric PDB ID, ready do be downloaded for free by any researcher who might deem them to be useful. Every molecular asset is accompanied by its own database page filled with important information like chemical composition, symmetries, biological function, bound ligands, related structures etc.

Since some of these experimental methods cannot detect majority of hydrogen atoms, the following compulsory procedure on the downloaded structure is its neutralization by addition of necessary hydrogen atoms. This can be achieved via hydrogen addition routine implemented in UCSF Chimera [95], which uses well-known (bio)chemical rules for carbon-based systems to correctly attach every missing particle. If necessary, we can further modify this structure manually according to our needs.

4.1.2 VMD – Molefacture

If we are dealing with a simple-enough molecular structure, the straight forward path is to model the system manually atom-to-atom, group-by-group inside VMD's *Molefacture module* [96]. Inerface of VMD's Molefacture module is shown in Fig. 4.1. Such an approach is pretty easy, unless the molecule is substantially bigger, potentially taking many possible intricate conformations.

[1] Experimental methods to determine macromolecular assembly include X-ray crystallography, NMR spectroscopy, and (cryo)electron microscopy.

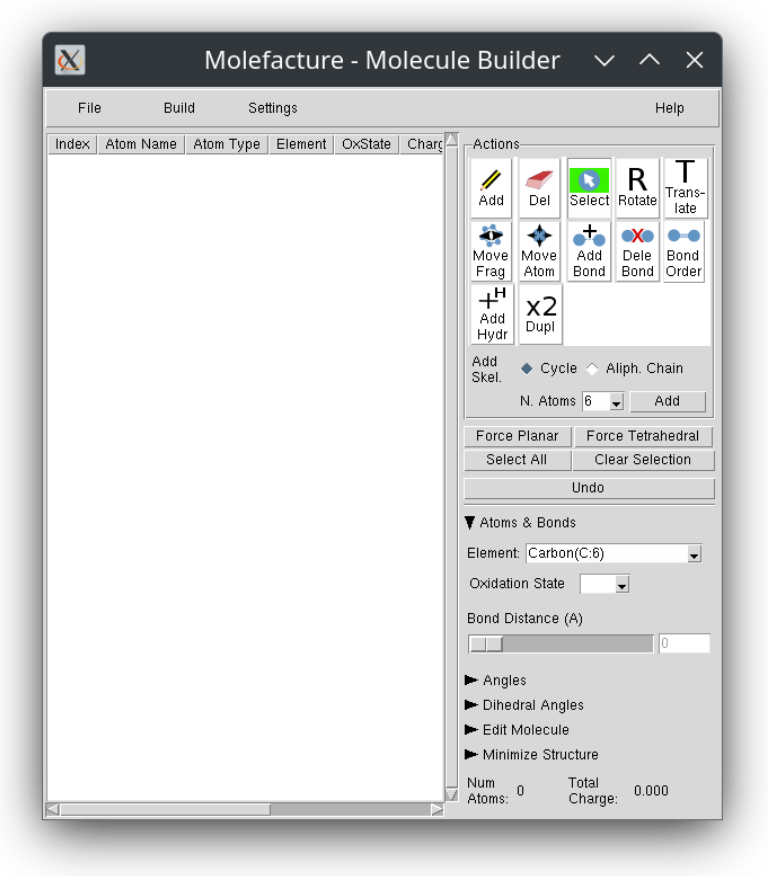


Figure 4.1: Interface of VMD's Molfacture module [96]. Any molecule can be created here atom-by-atom, group-by-group using varieties of convenient tools. One can also load in already existing structures for modifications.

4.1.3 VMD – AutoPSF

Every such a molecular system needs to be aptly parametrized by the force field we intend to use. For that we need the system's coordinates and topology. CHARMM FF [45, 46] offers 2 possibilities to parametrize the system. Smaller molecules (up to $\sim$ 100 atoms) can be automatically topologized promptly via *CGenFF* web application [97]. Coordinates alongside topology then have to be driven through VMD's (or equivalent) parametrization process using CGenFF's parametric file. For this we utilize the so-called *Automatic PSF Builder* implemented in VMD. Interface of the AutoPSF module can be seen in Fig. 4.2.

Larger structures (proteins, nucleic acids, lipids) have to be treated differently. Due to the 'LEGO-like' nature of these macromolecular structures it is possible to create general topologies based on their individual building blocks and standard chemical connections between them. Extensive general topology files intended specifically for such macromolecular systems are pre-made by the authors of CHARMM FF [98, 99], and are available for download together with force field's parametric files. System's coordinates and general topologies are again to be ran through AutoPSF module to form the parametrized structure.

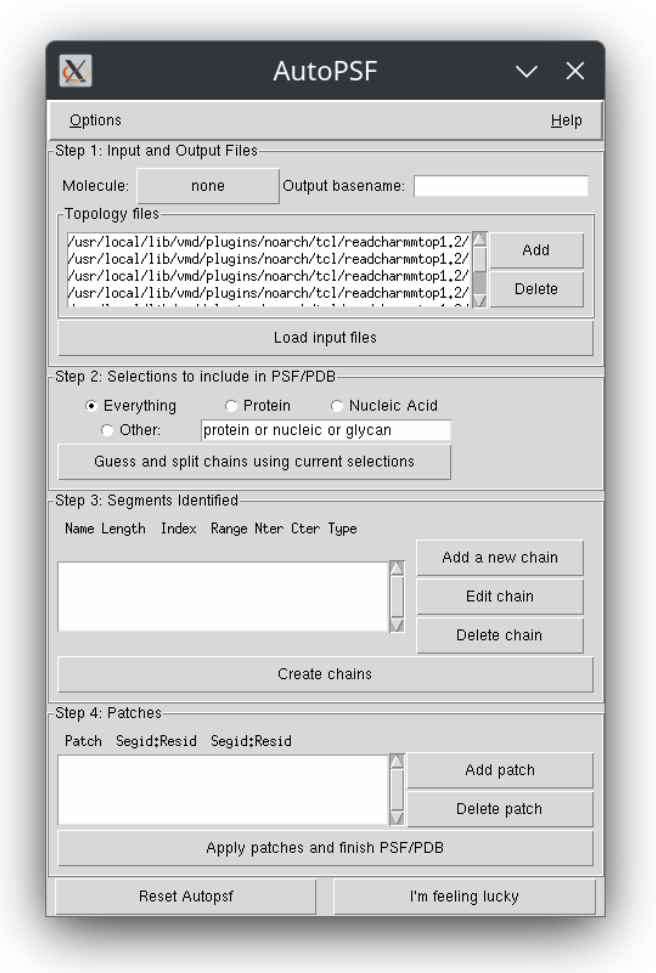


Figure 4.2: Interface of VMD's AutoPSF module [96]. This tool allows for automatic parametrization of molecular systems based on provided topology and force field files. The outcome is a formatted coordinate file (pdb) and protein structure file (psf), both required as the input to NAMD [44] program.

Among our various interests, mutations of specific residues are a focal point, and in the context of FEP we thus also deal with non-physical hybrid structures. These unnatural hybrids need special care. To successfully parametrize them we have to manufacture *hybrid topologies*, cautiously in accordance with the inner workings of the simulations and the underlying force field. We crafted topologies for all possible DNA base mutants and inscribed them into the general nucleic acid topologies of CHARMM FF [45], hybrid topolgies are provided in Appendix **??**. Problematics of non-physical hybrid structures is to be covered in the following Section 4.2. Modified topologies can readily substitute the original CHARMM files in the AutoPSF module. The same had to be done separately for every amino acid mutation performed, though we did not explore every possible amino acid mutation – that would be another very long story.

Fully parametrized system is sufficiently described by its coordinate (pdb) and structure (psf) files. Translated version of topology (top) is included in the structure file, for it was used to generate the structure in the first place. As an input, NAMD simulations therefore require only coordinates and structure, alongside force field's parameter files.

4.1.4 VMD – Solvate

Since we are investigating biochemical structures naturally occurring in physiological environment, the next step is to immerse them in water, possibly (if necessary) alongside native ions that might be of importance to their function. For that we wrap our system in adequately sized box[2] of TIP3 water molecules and employ appropriate periodic boundary conditions. This can be done inside the *Solvate* module of VMD, see interface in Fig. 4.3. One has to provide the coordinates (pdb) and the parametrized structure (psf) from the previous step.

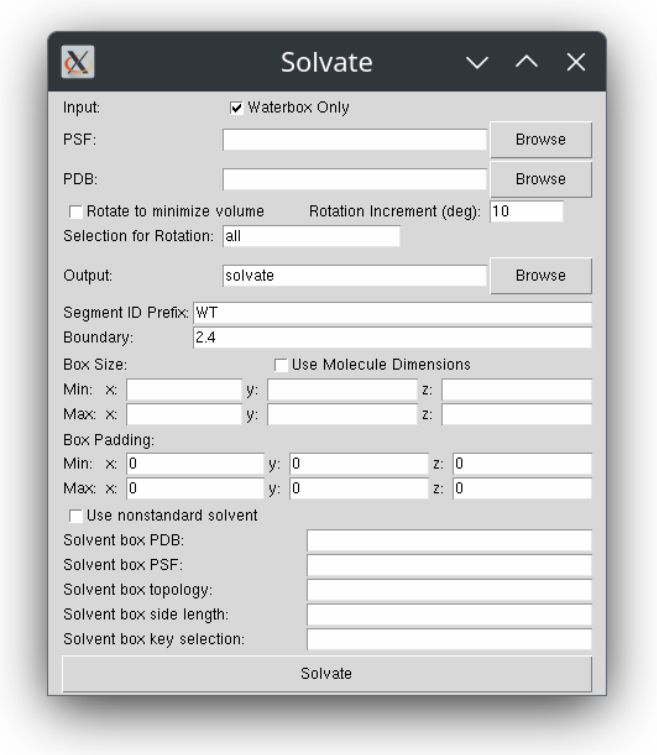


Figure 4.3: Interface of VMD's Solvate module [96]. This tool allows for automatic solvation of the system. The default solvent is water, but other solvents can be chosen given that the appropriate files are provided.

Parametrization as described above has to be done explicitly to *vacuo* structures prior to their immersion in water. *Aqua* systems inherently possess these parameters, and the extra TIP3 water is automatically included. The choice of TIP3 water model is based on its implementation in CHARMM. Although the default solvent in this module is water, the user can specify a different one given that the appropriate files are provided.

[2]By *adequately* the author tactfully evades explanation of common practise involving careful selection of suitable water cell. The box has to be big enough for our structure to not interact with its periodic copies in near by cells, but we don't need to simulate unnecessary amount of water. A healthy balance is needed. It takes some experience and intuition to build, but eventually it becomes almost automatic. Usually ~ 10 Å distance from the edge of the box to the molecule(s) of interest is safely enough, though the size and shape of the system play an important and non-negligible role. Also, one might need to take into account the cutoff and related parameters embedded in the simulation itself.

4.1.5 VMD – Autoionize

One can also add ions in an automatic manner using VMD's *Autoionize* module, interface is given in Fig. 4.4. This tool allows the user to add varieties of ions, naturally present in aqueous systems. It can be done for mimicking the real solvent with a given salt concentration or just to neutralize the system at hand. The only thing we do is to make sure our systems are charge neutral.

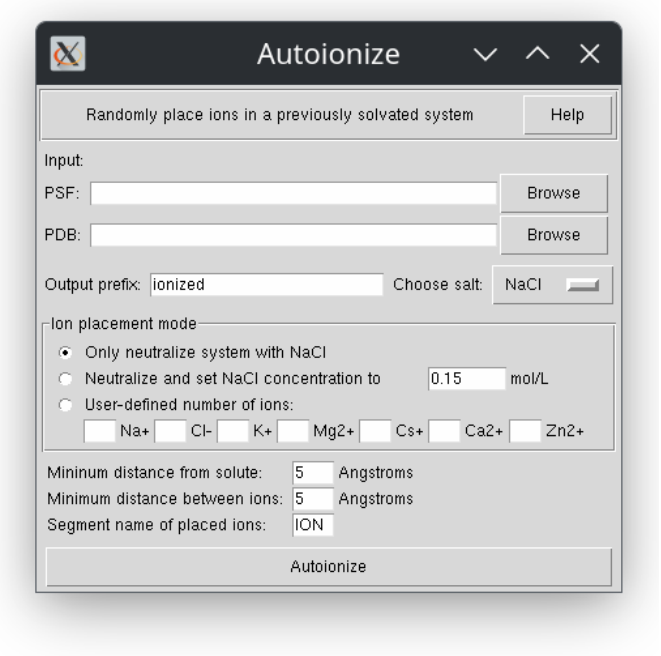


Figure 4.4: Interface of VMD's Autoionize module [96]. It allows for automatic addition of ions in a random seed manner, replacing the already existing solvent molecules in a given system.

Autoionize replaces individual solvent molecules of an already solvated system with specified ionic species in a randomized manner. For that it again needs the coordinate file (pdb) and parametrized structure (psf) of the solvated system.

4.2 Hybrid Molecular Topologies

In a typical alchemical transformation setup, involving the alteration of one chemical species into another, one must somehow construct the alchemical path between the chemical states of interest. This is done using hybrid molecular topologies. Atoms of any hybrid molecular topology can be classified into three groups:

- atoms of the unperturbed part of the system (e.g. environment),
- atoms describing the reference state a,
- atoms belonging to the target state b.

Atoms of state a should *never* interact with those of state b throughout the course of the simulation. Such interactions are thus turned completely *off*.

4.2.1 Single vs. Dual Topology

There are two widely used hybrid topologies – single and dual topology. An example comparing both approaches is depicted in Fig. 4.5. Single topology has only one largest common motif shared between the end states, and then *dummy* atoms to make up for any unique sites. During the transformation, the dummy

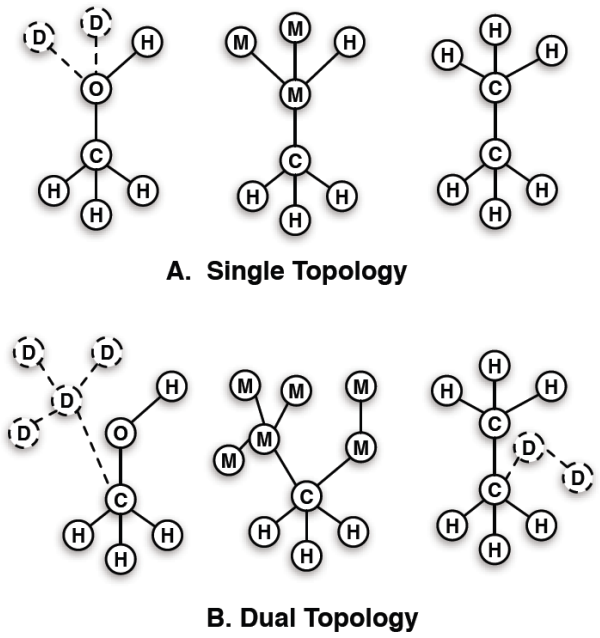


Figure 4.5: Single topology (A, top) and dual topology (B, bottom) techniques to construct an alchemical path between ethane and ethanol. D represents non-interacting dummy atoms, while M correspond to nonphysical intermediates. Image taken from [100].

atoms are transformed into fully interacting atoms, and the shared site atom is transformed directly to a new atom. Dual topology tackles the problem differently. Shared sites between states do not share atoms. No atoms change type, only their interactions are gradually turned on/off from the rest of the system. However, more atoms need to be altered in order to mutate from the initial to the final state of the system. On the other hand, dual topologies have a strong

advantage in that their dummies can simultaneously explore considerably larger space while decoupled.

It is important to note that even though the dummies may have nonbonded interactions turned off at some point of the simulation, they still are bound to the rest of the molecule. This in fact renders the end states a and b slightly nonphysical. Such undesirable interactions can be treated for by simulating in both molecular medium (e.g. natural solvent) and in vacuum. In the rigid rotor approximation with bond lengths fixed, the impact of these dummies on the free energy is eliminated [101]. Nevertheless, even without the bond constraints, the difference is usually negligible (~ 0.01 kcal/mol).

4.2.2 Dual Topology Paradigm

In our simulations we will use exclusively dual topologies. Take a look at yet another example featured in Fig. 4.6, a point mutation of an alanine side chain into that of serine. Both side chains coexist with one another, both connected to the central carbon C_α, yet without actually 'seeing' each other.

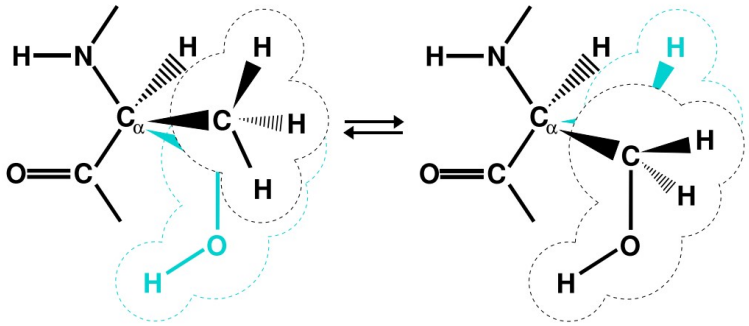


Figure 4.6: Example of a simple dual-topology amino acid mutation Ala → Ser. Currently non-interacting side chains are depicted in cyan. Image taken from [76].

As we discussed earlier within Section 3.5.5, such systems are described by the hybrid Hamiltonian (3.64). The energy and forces are thus defined as a function of alchemical parameter λ. In the case of our simple Ala → Ser example, interactions of the methyl side chain of alanine with the rest of the system (except for the other side chain) are fully effective at the beginning of the simulation ($\lambda = 0$), while the serine side chain feels nothing. Conversely, at the other end of the simulation ($\lambda = 1$) their roles are reversed. For intermediate values of λ, both the side chains participate in nonbonded interactions with the rest of the system, scaled appropriately by the current value of λ.

4.2.3 Preventing End-point Catastrophes

Endpoints of alchemical transformations carried out in the framework of the dual–topology paradigm have been shown to be prone to numerical instabilities from MD simulations [76]. These instabilities usually occur for λ approaching 0 or 1, when incoming atoms instantly appear where other particles are already present, resulting in practically infinite potential as the interatomic distance tends towards zero. This is referred to as the *end-point catastrophes*.

One can alleviate these effects and prevent such scenarios to happen by introduction of a so-called *soft-core potential* [102, 103]

$$U_{\mathrm{NB}}\left(r_{ij}\right) = \lambda_{\mathrm{LJ}}\varepsilon_{ij}\left[\left(\frac{R_{ij}^{\min\,2}}{r_{ij}^2+\delta\left(1-\lambda_{\mathrm{LJ}}\right)}\right)^6 - \left(\frac{R_{ij}^{\min\,2}}{r_{ij}^2+\delta\left(1-\lambda_{\mathrm{LJ}}\right)}\right)^3\right] + \lambda_{\mathrm{elec}}\frac{q_i q_j}{\varepsilon_1 r_{ij}}, \tag{4.1}$$

performing a gradual scaling of the short-range nonbonded interactions of incoming atoms with their environment. As the simulation progresses, the gradual decoupling is done through parameters λ_{LJ} and λ_{elec}, modulating VdW and electrostatic interactions respectively. Two examples of possible linear coupling of nonbonded interactions though parameters λ_{LJ} and λ_{elec}, and their relation to user-defined alchemical parameter λ, are shown in Fig. 4.7. Any bonded potentials (e.g. valence angle, bond stretch, torsions) are not altered at all.

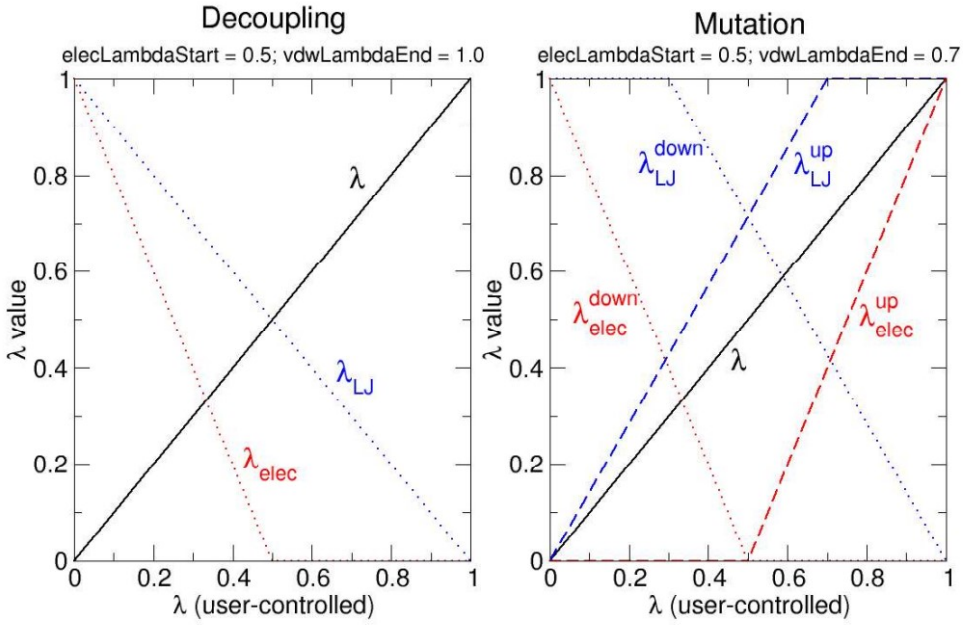


Figure 4.7: Two examples of typical coupling parameters λ_{LJ} and λ_{elec} during a FEP simulation (as implemented in NAMD [99]), and their relationship to user-defined alchemical parameter λ. Image taken from [76].

4.3 Non-equilibrium Free Energy Calculations

Our simulation process generally consists of 2 subsequent segments:

1. Minimization & Equilibration,
2. Free Energy Perturbation (FEP).

The first one takes the system we created in its initial conditions, minimizes it in terms of energy via the conjugate gradient method [48], recall Section 3.2, to find a conformation close to optimum, and performs a standard thermalization molecular dynamics to sample multiple conformations bearing similar energies. Throughout the whole simulation process the temperature is maintained at 300 K using Langevine thermostat [61].

Normally, the following would be to perform FEP simulation by austere application of the FEP formula (3.72), and if one felt productive enough, he would have performed the simulation in both the forward and the backward direction with respect to perturbation parameter λ to check if the results came out in agreement. As was discussed earlier in Section 3.5.8, brute-force approach such as this searches directly for an equilibrium result via the 2 possible routes, both ideally landing exactly at the desired free energy value (within the error of the method used). Running calculations in this simple, yet inefficient setup would cost us a lot of valuable resources and time, none of which can we spare. Large structures like proteins and nucleic acids are tremendously more expensive. In such setting it would take us days or even weeks to simulate a single mutation inside a DNA-protein complex, and with no guaranteed success of convergence to the desired value with the required precision.

To enhance convergence and accuracy of our calculations we employ a non-equilibrium approach governed by the CFT [90], recall Section 3.6. This allows us to achieve a great level of parallelization by launching a series of many short-lasting, non-equilibrium FEP simulations simultaneously. Each non-equilibrium simulation consists of a forward and a backward run. Forward run utilizes the last configuration of the preceding equilibration process as one of its initial conditions. Backward run directly follows taking the final conformation of the forward transformation, running the simulation literally (and independently) backwards.

As the name of this fairly unconventional technique suggests, the results of such runs have no reporting value on their own whatsoever. This is where the CFT relation, Eq. (3.80), steps in. The free energy value we seek lies right at the intersection of the forward and the backward distributions of work done on the system throughout the non-equilibrium transformations. In order to sufficiently sample both the requisite distributions, it is appropriate to carry out from hundreds up to thousands non-equilibrium runs. In most of our cases, 100 forward, each followed by their backward counterpart, is enough to achieve desirable precision, though individual systems may require specialized treatment. Separate minor methodology sections, tailored to each system at hand, will unveil the specific details.

Efficiency of this method highly relies on available computational resources, since it involves potentially hundreds or even thousands of simulations running in parallel. With cluster system and well-made allocation order the likes of MetaCentrum [104], this method is not only with its precision but also in its efficiency

comparable to techniques like BAR [86]. As such, what would normally take us days or even weeks is now a matter of hours, for mutations of small molecules like amino acids even minutes.

4.4 Hardware – MetaCentrum

First and foremost we would like to express our sincere gratitude to MetaCentrum [104] for allowing us to use their valuable resources in order to conduct all of our calculations, playing a crucial and necessary part of this thesis. Without an access to such computational power and titanic data storage we would not be able to carry out simulations of these scales.

MetaCentrum [104] is a state-of-the-art cluster system located in the heart of Czechia, serving as a critical tool for scientific research and innovation in the region. The map of Czechia as seen through eyes of MetaCentrum is shown in Fig. 4.8. The infrastructure comprises a vast network of high-performance computing (HPC) clusters, data storage facilities, and specialized computing resources. Equipped with thousands of CPU cores and GPU accelerators, it boasts a substantial computational capacity, enabling researchers to tackle complex and computationally intensive problems efficiently with clever parallelization schemes or sheer brute-force power. The cluster system also integrates with cloud computing resources, offering users a flexible and scalable environment to meet their computational needs.

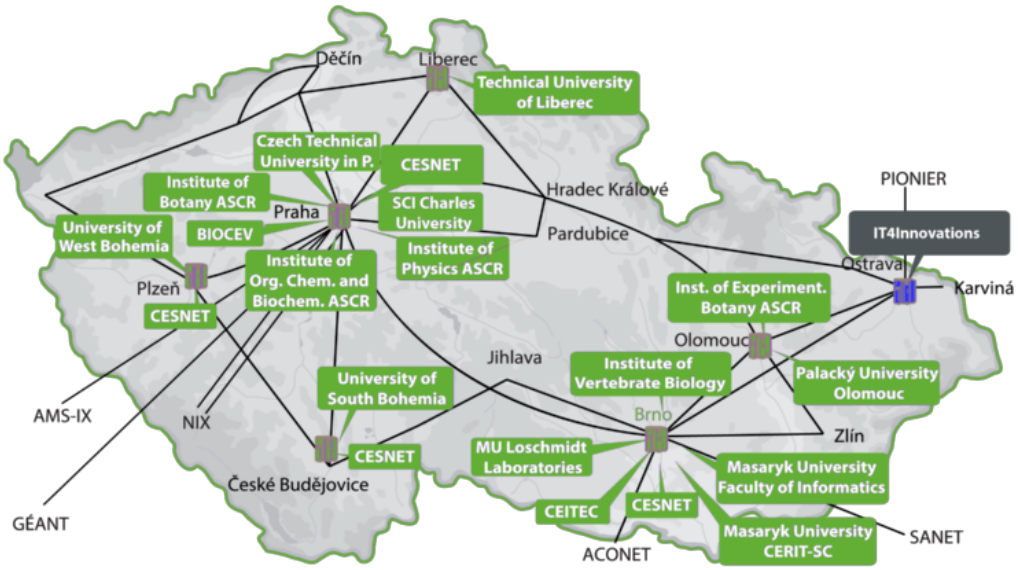


Figure 4.8: Home sweet home as seen through eyes of MetaCentrum [104]. The land of Czechia is interwoven by many internet lines connecting various HPC clusters and storage facilities, accessible to research groups and individuals from all over the country.

One of the defining features of MetaCentrum is its accessibility and relatively userfriendly approach. It caters to a diverse user base, ranging from individual researchers and small research teams to large scientific institutions and universities. Researchers from across the Czech Republic and Central Europe can gain remote access to any MetaCentrum cluster through SSH client of their choice. A well-defined allocation process ensures that computational resources are distributed fairly and transparently between all its users.

4.5 Software – NAMD

NAMD [44], short for *Nanoscale Molecular Dynamics*, is a powerful and widely used program for conducting molecular dynamics simulations of large biomolecular systems. Developed and maintained by the Theoretical and Computational Biophysics Group at the University of Illinois at Urbana-Champaign, NAMD has become an indispensable tool for researchers in the field of computational biology, chemistry, and related disciplines.

NAMD's primary strength lies in its ability to simulate the dynamic behavior of complex biological macromolecules at the atomic and molecular level. It employs highly efficient parallel algorithms that leverage the capabilities of modern supercomputers and high-performance computing clusters. This allows scientists to investigate a wide range of biological processes, including protein folding, molecular binding, and membrane dynamics, with remarkable accuracy and computational efficiency.

One of the standout features of NAMD is its scalability. It can effectively harness the computational power of multi-core processors and GPU accelerators, making it suitable for simulating systems ranging from thousands to millions of atoms. This scalability enables to explore increasingly larger and more complex biological structures, providing valuable insights into the function and behavior of biomolecules. Furthermore, NAMD is supported by an active user community and a wealth of documentation. It also integrates seamlessly with visualization and analysis tools, facilitating the interpretation of simulation results and the generation of meaningful insights into biological processes.

4.6 Data Extraction and Analysis

4.6.1 Bash Scripting

In order to properly monitor the behavior of our simulated systems we store a lot of outcoming data, including *dcd* trajectories, extended system information, and FEP data. The latter is in a single *fepout* file separately for the forward and the backward run, written down for every non-equilibrium simulation performed. Fepout files contain many different columns of data. Since the scope of our interest aims mainly on final free-energy differences, it would be a madness to go through every single simulation output manually. For us to dig through such a pile of information effectively we have written a few bash scripts that do the necessary extraction automatically. The data we acquired via our extraction scripts form 2 (one for each FEP direction) one-column files of final free-energy differences. These files are the subjects of our main analysis.

4.6.2 Custom Python Analysis

For we desire to have a total control over the analysis process alongside its possible modulation and flexibility, we have written a short program in Python that does an automatic FEP analysis tailored to exactly fit our needs. Our program takes the extracted files (forward and backward), formats them in order to be easily processable by the packages of our choice[3], and rearranges the loaded data into suitable Python arrays. Such constructs are then subjected to our histogram distribution analysis.

4.6.3 Scott's Normal Reference Rule

There are many formulas providing a wide range of rules to estimate a suitable number of bins a histogram should have, based on the type of data one might apply it on. An example can be the *Scott's normal reference rule* [105]

$$b_{\mathrm{W}} = 3.49\frac{\sigma}{\sqrt[3]{n}}, \tag{4.2}$$

where σ is the sample standard deviation (STD), and n is the number of observations in the sample. Scott's estimation of bin width is optimal for random samples of data governed by an *ideal* normal distribution.

4.6.4 Freedman-Diaconis Rule

In our analysis we implement the so-called *Freedman-Diaconis rule* [106] which replaces 3.5σ of Scott's rule (4.2) with 2IQR, i.e.

$$b_{\mathrm{W}} = 2\frac{\mathrm{IQR}(x)}{\sqrt[3]{n}}, \tag{4.3}$$

where the $\mathrm{IQR}(x)$ is an *interquartile range* of the data, and n is the number of observations in the sample x. In our case $x = W$.

[3]Our code operates within the boundaries of *Scipy*, *Numpy*, and *Matplotlib*.

Interquartile range can be understood as the midspread or middle 50 %, defined by the difference between the 75th and 25th percentiles of the data, see Fig. 4.9. Practically speaking, the data set is divided into quartiles, denoted as Q_1 (*lower quartile*), Q_2 (*median*), and Q_3 (*upper quartile*). Since lower quartile corresponds to the 25th percentile and upper quartile to the 75th percentile, $\mathrm{IQR} = Q_3 - Q_1$ [107]. IQR's can also be used to build box plots – simple graphical representations of probability distributions, see Fig. 4.9. Using IQR instead of STD makes the histogram less sensitive to outliers in data.

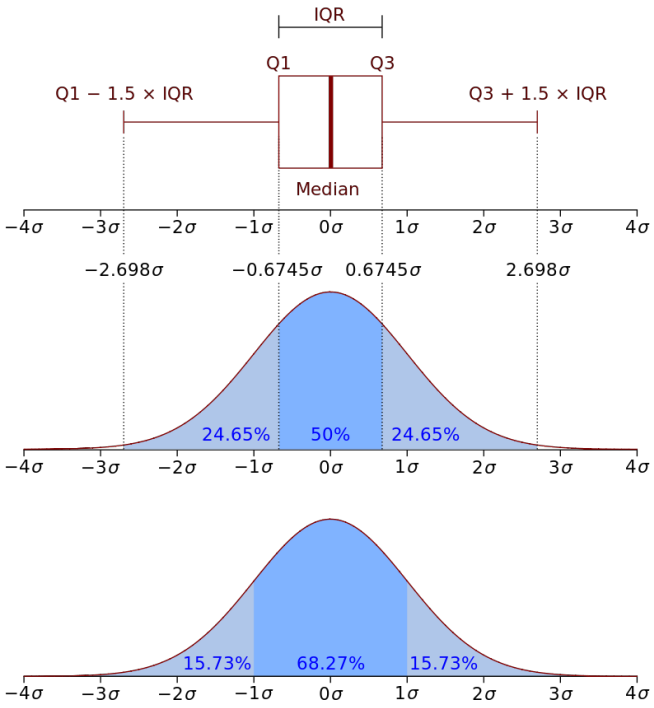


Figure 4.9: Box plot with an interquartile range (IQR) and a probability density function of a normal $\mathrm{N}\left(0, \sigma^2\right)$ population. Graph is divided into quartiles Q_1 (25th percentile), Q_2 (median), and Q_3 (75th percentile). Image taken from [107].

Our choice of this rule instead of the Scott's normal reference rule is based on the fact that it is more versatile and robust for a wide range of distributions, making it a good choice whenever we are uncertain about the exact nature of the data behavior. That includes possible presence of outliers or slight deviations from normality. However, if the sample is indeed normally distributed, Freedman-Diaconis rule should still provide a good histogram distribution without any pronounced issues.

4.6.5 Unified Bin Width

Bin widths provided by Eq. (4.3) and length of the data interval allow us to determine the corresponding number of bins for each histogram. Though the bin widths of forward and backward runs are generally not the same, they are not that far apart from each other. Our code takes an average of these 2 to create a new, common bin width. This allows for unification of forward and backward data into single data set W, which comes quite handy for certain parts of our

implementation and also gives our data unified appearance to avoid any possible confusion. Nevertheless, for the most part we still treat both samples separately.

A curious reader might have an objection – averaging 2 bin widths into a single common one to use for both merged data sets inevitably leads to literal shifting of histograms away from their form estimated by Freedman-Diaconis rule. That is correct. However, unless the 2 widths dramatically differ from each other, shifting (though still present) is negligible and thus contributes to the analysis error very faintly. The only possible cause of any significant difference in bin widths of forward and backward runs would be caused either by insufficient conformation sampling, or by an extreme case of mutation where states a and b occupy regions of conformation space that are radically different in size. An example of such an extremum could be mutation of hydrogen atom into a much larger and highly flexible side chain.

Another fact that supports not only our choice of histogram rule but also the bin width averaging is that there is no ideal number of bins a histogram should have. As was hinted earlier, there exists a whole mathematical branch studying the problematics of histogram distributions and the related statistics of discrete values. Our averaging thus constitutes only a slight modification to one of the possible choices, the Freedman-Diaconis rule, and should faithfully work unless we enter the realm of extreme cases.

4.6.6 Plotting Histogram Distributions

The final histogram features bars representing the probability of work values W calculated during the non-equilibrium FEP transformations, an exemplary graph is shown in Fig. 4.10. Both distributions are separately normalized such that the sum of all their bins yield 1. Similarly as in the provided example, all of our distributions exhibit normal behavior, though we occasionally report outliers in data and slight deviations from normality. Since these minor deviations are a natural result of statistical errors present in real data, we assume our histograms can indeed be faithfully represented by Gaussian distributions.

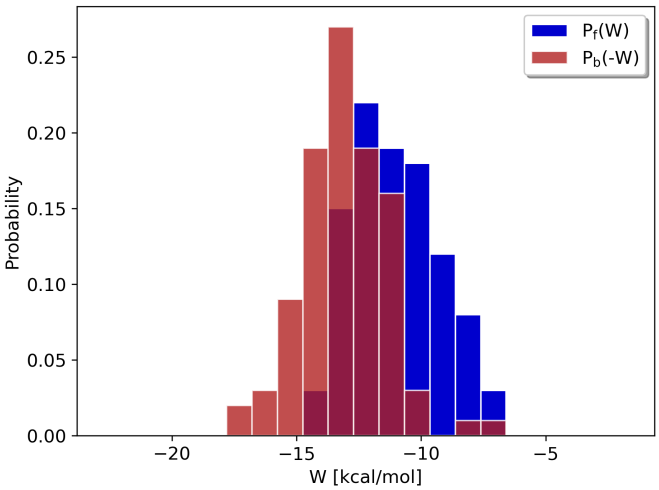


Figure 4.10: An example of histograms produced by our program up until now, analyzing 100 forward and 100 backward non-equilibrium FEP runs.

4.6.7 Crooks-Gauss Intersection (CGI)

According to the Crooks fluctuation theorem [90], recall Section 3.6, the free energy value we seek lies right at the intersection of the forward and the backward distribution. But how exactly do we determine the exact point of intersection between two histograms? Since our distributions follow normal behavior, the next step is to fit our histograms with Gaussian curves

$$g(x) = A \exp\left[-\frac{(x - x_0)^2}{2\sigma^2}\right], \tag{4.4}$$

where A is the height of the curve's peak, x_0 the central expected value, and σ the standard deviation (FWHM $= 2\sigma\sqrt{2\ln 2}$). These curves were fitted onto our data using the *method of least squares*.

The choice of general Gaussian models (4.4) without the standard normalization factor $\sigma^{-1}(2\pi)^{-1/2}$ lies in the fact that they inherit the normalization from the histograms they were fitted on. Using the standard prefactor here would mean applying 2 norms at the same time, resulting in curves that are too narrow in width to correctly fit the underlying histograms.

After the fitting procedure the intersections can be readily computed as

$$\Delta G_{\mathrm{CGI}} = \frac{\frac{W_\mathrm{f}}{\sigma_\mathrm{f}^2} - \frac{-W_\mathrm{b}}{\sigma_\mathrm{b}^2} \pm \sqrt{\frac{1}{\sigma_\mathrm{f}\sigma_\mathrm{b}}\left(W_\mathrm{f} + W_\mathrm{b}\right)^2 + 2\left(\frac{1}{\sigma_\mathrm{f}^2} - \frac{1}{\sigma_\mathrm{b}^2}\right)\ln\frac{\sigma_\mathrm{b}}{\sigma_\mathrm{f}}}}{\frac{1}{\sigma_\mathrm{f}^2} - \frac{1}{\sigma_\mathrm{b}^2}} \tag{4.5}$$

where $W_{\mathrm{f,b}}$ are the work values of the forward and the backward runs, and $\sigma_{\mathrm{f,b}}$ are the standard deviations of the respective distributions. This method of finding the intersection between normal forward and backward distributions of non-equilibrium work is known as the *Crooks-Gauss intersection* [108] (CGI).

Formula (4.5) yields 2 distinct intersection points, unless the distributions are exactly the same. Which one do we choose? For that we implemented a simple criterion, taking the one closer to the midpoint of the interval W. Final graph, featuring Gaussian fits and the relevant intersection, is shown in Fig. 4.11.

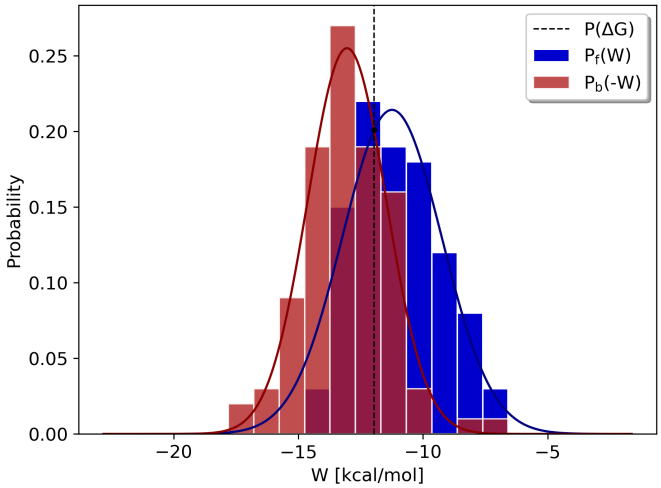


Figure 4.11: Histograms completed with Gaussian fits and pinpointed relevant intersection, providing us with the free energy difference ΔG.

5. Hydration Free Energies of Amino Acids

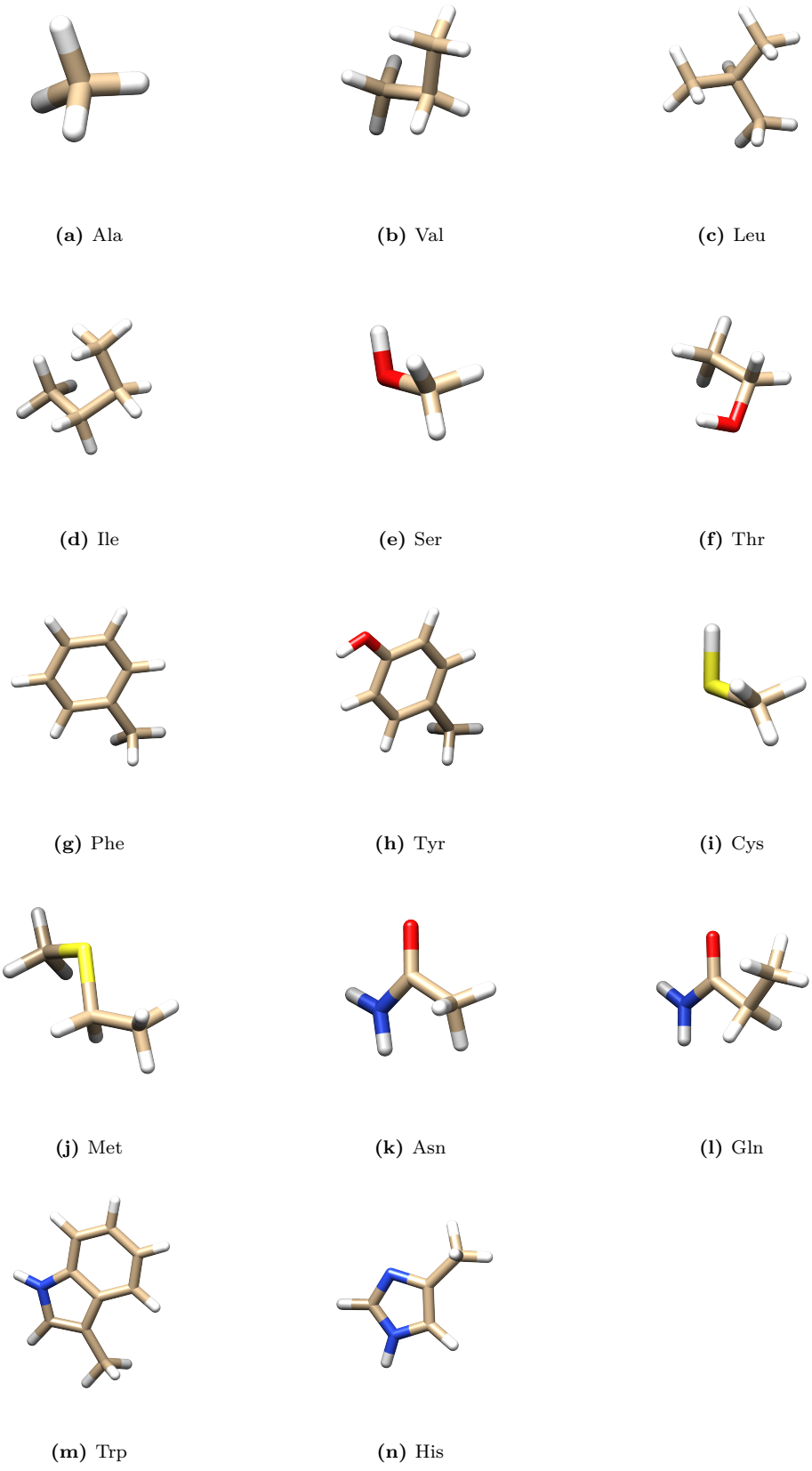
In this chapter we explore the nature of the very building blocks of proteins, the so-called *amino acids*. For an amino acid (AA), as for any other naturally condensed-state molecule, one of the most valuable chemical properties is its solvation (hydration) free energy. Study of amino acids plays a pivotal role in the field of genetic engineering, as their behavior dictates the proteins' features and abilities. Rationally designed modifications to protein structures can promote the reliability of their actions, possibly altering genes inside any living organism according to our needs.

Since in proteins the only part of amino acids sticking outwards from the peptide chain to the surroundings are the side chains, the reasonable way to study the behavior of these small molecules is to introduce their side chain analogues, see Tab. 5.1. This is what our reference studies [109, 110] do, including [111] providing us with invaluable experimental results. Models of these molecular mimetics are shown in Fig. 5.1.

Table 5.1: Amino acid side-chain molecular mimetics, and their hydration free energies according to experiments [111]. Values listed in kcal/mol.

Res.	Side chain analogue	ΔG_{exp}
Ala	Methane	1.94
Val	Propane	1.99
Leu	Isobutane	2.28
Ile	Butane	2.15
Ser	Methanol	−5.06
Thr	Ethanol	−4.88
Phe	Toluene	−0.76
Tyr	p-Cresol	−6.11
Cys	Methanethiol	−1.24
Met	Ethylmethylsulfide	−1.48
Asn	Acetamide	−9.68
Gln	Propionamide	−9.38
Trp	Indole	−5.88
His	4-Methylimidazole	−10.27

Simplification through side-chain mimetics also serves to eliminate the need to simulate the amino-acid head, which would normally get charged at physiological pH, posing possible complications. We are not really interested in simulating whole charged amino acids, but rather in the behavior of the side chains themselves. They will be the ones directly interacting with DNA bases inside protein-DNA complexes.



(a) Ala **(b)** Val **(c)** Leu

(d) Ile **(e)** Ser **(f)** Thr

(g) Phe **(h)** Tyr **(i)** Cys

(j) Met **(k)** Asn **(l)** Gln

(m) Trp **(n)** His

Figure 5.1: Models of all 14 amino acid side-chain molecular mimetics studied in this chapter.

5.1 Simulation Setup

We have built each one of the 14 amino acid side chain analogues (listed in Tab. 5.1 and displayed in Fig. 5.1) manually inside the environment of VMD [96] using the Molefacture module. Each system was immersed in an aqueous box of TIP3 water approx. 15^3 Å^3 in volume.

MD simulations were carried out in NAMD [44] software package using the CHARMM36 [45, 46] force field. Throughout the whole simulation, temperature was maintained at 300 K by Langevin thermostat [61] while Langevin piston targeted pressure at 1.0 bar. Systems were simulated using RESPA algorithm [58] with 1.0 fs time step, and PME inside periodic boundary conditions. Inside the NAMD's FEP module the parameter *alchDecouple* was turned *on*.

In contrast, the reference study [109] uses thermodynamic integration (TI) in a standard equilibrium setting using Folding@Home distributed computing infrastructure [112] with MBER(*ff*94), CHARMM22, and OPLS-AA force fields. Our results therefore serve as an addition to their collection of results, with different method used and an updated version of the CHARMM FF.

First, we performed a minimization and equilibration procedure (500 ps), after which followed non-equilibrium FEP calculations (recall Section 4.3) consisting of swarms of 100 independent MD runs, each composed of forward and backward stages. Each independent MD run included an additional short 10 ps thermalization to ensure slightly different starting conformations of forward MD runs. Backward MD runs used the final conformations of forward MD runs as a starting point. Each of the 20 λ windows consisted of $5 \cdot 10^3$ FEP steps. The final work values from non-equilibrium forward and backward MD runs were analyzed via our Python program (mentioned above in Section 4.6) that determines CGI of histograms pinpointing out the equilibrium binding free energy.

5.2 Absolute Hydration Free Energies

Histogram depicted in Fig. 5.2 reveals the absolute hydration free energy values

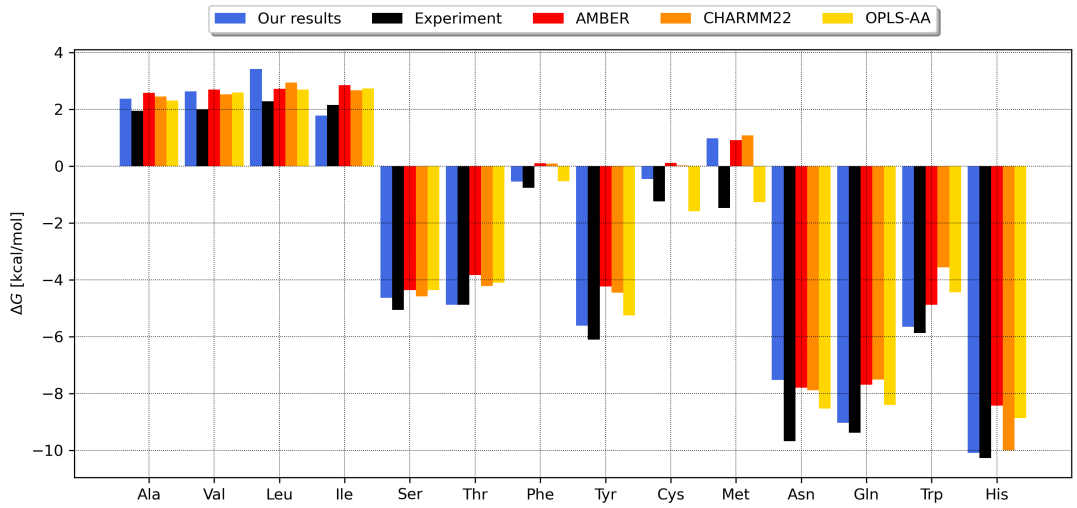


Figure 5.2: Absolute hydration free energies of 14 amino acid side chain analogues in comparison to reference simulation data [109] and experimental values provided in [111]. All values shown in kcal/mol. Our data are a result of non-equilibrium CGI FEP in CHARMM36 force field, reference done by TI in AMBER(*ff*94), CHARMM22, and OPLS-AA.

we obtained for each of the 14 amino acid side chain analogues studied. We contrast our free energy values with reference MD simulations [109] as well as with experimental data [111]. Reference calculations simulated all the given molecular systems using 3 force fields (AMBER, CHARMM22, and OPLS-AA) different from the one we used in our approach (CHARMM36).

Our MD simulations were able to achieve similar performance as that of reference calculations. We provide the reader with graph of differences from the experimental values [111] to inspect the necessary details, see Fig. 5.3. Apart from

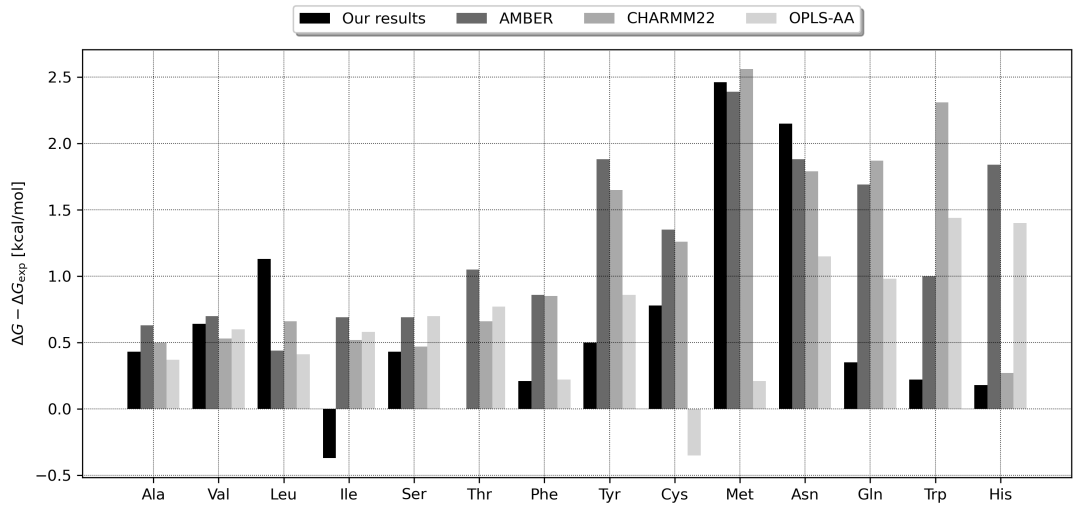


Figure 5.3: Differences from experiment [111] for absolute hydration free energies of 14 amino acid side chain analogues, presented in Fig. 5.2 We contrast our free energy values with [109]. All values shown in kcal/mol.

one singular case (Leu), our calculations are in better or at least similar accordance with experimental results compared to our reference counterpart. This is a fair success since some of our results (Phe, Tyr, Gln, Trp, and His) are far closer than any of their reference competitors. Though such achievement gives no surprise taking account for our use of refined version of CHARMM force field. Nevertheless, FEP method with our non-equilibrium approach is clearly sufficient enough to yield satisfactory results.

5.3 Discussion

Overall our absolute hydration free energies (shown in Fig. 5.2) report differences from the experiment [111] within a standard chemical accuracy of ± 1 kcal/mol, recall Fig. 5.3 for details. The performance of our non-equilibrium scheme is comparable to that of the reference MD simulations [109], and thus yielding satisfactory outcomes. The only results at issue could be those of Met and Asn, with little over double the desirable difference from experiment. Though such differences were also recorded by the reference calculations [109].

The slight deviation from experimental data (observed for Met and Asn) might stem from one particular choice we made at the beginning of our simulations, based on the official documentation of NAMD [76]. As the manual suggested, in each of our simulations performed the so-called *alchDecouple* parameter was turned *on*. The reason behind their discrepancies may lie in the fact that these are a bit larger and more flexible side chains, potentially taking slightly different conformations in vacuum compared to their hydrated liquid form.

We can conclude that it is not an incorrect choice to simplify the calculation process of absolute hydration free energies with *alchDecouple on* for the case of small rigid molecules like the amino acid side chain analogues. For more flexible moieties the likes of Met and Asn such an option could be debatable. A further investigation is needed.

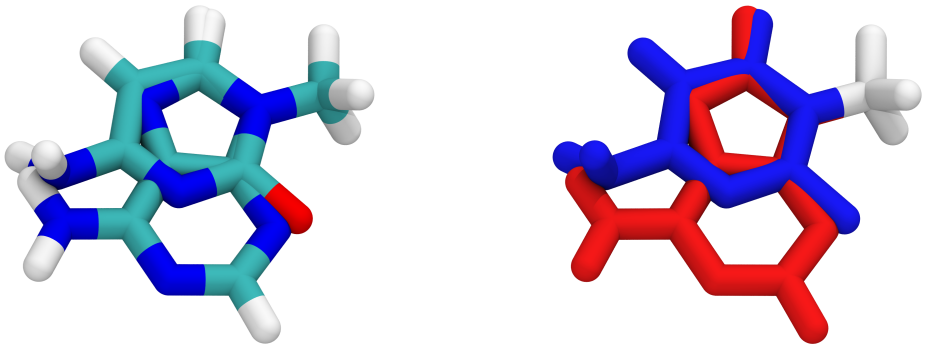
6. Hydration Free Energies of DNA Bases

Before we dive into the study of protein-DNA complexes, let us examine the very building blocks of nucleic acids. We will use very the same approach but this time to explore the nature of DNA bases via their N-methylated forms, listed in Tab. 6.1. Same as for amino acids, we will study their hydration energies.

Table 6.1: N-methylated DNA bases and our labels. Their hydration free energies according to experiments are taken from Ref. [113]. Values are listed in kcal/mol.

Base	Methylated structure	Tag	ΔG_{exp}
A	9-Methyladenine	9MA	-13.60
G	9-Methylguanine	9MG	
T	1-Methylthymine	1MT	$-(9.1 - 12.7)$
C	1-Methylcytosine	1MC	

In the filed of genetic engineering mutations in DNA sequence are one of the main focal points allowing for assessment of features related to DNA sequence recognition by various proteins, e.g. their ability to successfully bind to the right region of a given DNA double helix. Before we perform any mutation inside a protein-DNA complex we need to perfect our methodology on simple systems. We will start by exploring all possible base mutations using N-methylated DNA base hybrids, one of which is shown in Fig. 6.1. All hybrids used are made out of



(a) 9MA-1MC **(b)** 9MA-1MC: beta

Figure 6.1: Hybrid molecular structure for 9MA $\rightarrow$ 1MC mutation (dual topology). Model (a) is colored by atomic species, *beta* coloring in (b) shows disappearing (initial) 9MA in red and appearing (final, target) 1MC in blue; the untouched methyl moiety is bleached.

methylated bases presented in Tab. 6.1, sharing a conjoint terminal methyl group. Mutations of DNA bases will provide us with valuable relative free energies, which will be compared to results of a reference study [114] done on systems similar[1]

[1]Some of the systems featured in [114] are mutated using a single-topology frame instead.

to our own. Later on during our future studies of protein-DNA complexes these simulations can serve as a reference frame for evaluating the behavior of mutants inside a DNA double helix. At the end of this chapter we will also try our methodology on an ATT trinucleotide to see how well it performs in a slightly more complicated environment.

6.1 Simulation Setup and Method of Analysis

We have built each one of the methylated DNA bases listed in Tab. 6.1 and their respective hybrid structures (10 unique systems in total[2]) manually inside the environment of VMD [96] using the Molefacture module. Each system was immersed in an aqueous box of TIP3 water approximately 15^3 Å^3 in volume. Depending on specifics of the given simulation process, some of the systems were simulated also in gas phase without any water molecules present.

MD simulations were carried out in NAMD [44] software package using the CHARMM36 [45, 46] force field. Throughout the whole simulation, temperature was maintained at 300 K by Langevin thermostat [61] while Langevin piston targeted pressure at 1.0 bar. Systems were simulated using RESPA algorithm [58] with 1.0 fs time step, and PME inside periodic boundary conditions. Inside the NAMD's FEP module the parameter *alchDecouple* was turned *on*.

In contrast our reference MD simulations [114] run equilibrium thermodynamic integration (TI) transformations with AMBER 4.1. Our results therefore serve as an addition to their collection of results, with different method and software package used together with different force field.

First, we performed a minimization and equilibration procedure (500 ps), after which followed non-equilibrium FEP calculations (recall Section 4.3) consisting of swarms of 100 independent MD runs, each composed of forward and backward stages. Each independent MD run included an additional short 10 ps thermalization to ensure slightly different starting conformations of forward MD runs. Backward MD runs used the final conformations of forward MD runs as a starting point. Each of the 20 λ windows consisted of $5 \cdot 10^3$ FEP steps. The final work values from non-equilibrium forward and backward MD runs were analyzed via our Python program (mentioned above in Section 4.6) that determines CGI of histograms pinpointing out the equilibrium binding free energy.

[2] 4 methylated DNA bases, and all 6 possible hybrids – 10 systems in total. As will be shown later, we also explored some mutations in trinucleotides.

6.2 Absolute Hydration Free Energies

Let us first probe absolute hydration free energies of DNA bases via MD simulations of their methylated forms. Due to the lack of experimental data we mostly limit ourselves to reference simulations. The resulting hydration free energies are listed in Tab. 6.2. The reference study [114] simulated the N-methylated bases

Table 6.2: Calculated absolute free energies of hydration for all 4 N-methylated DNA bases. ΔG_{ref} are results of reference MD simulations [114], experimental data [113] were available only for some of the bases studied. All values listed in kcal/mol.

Base	ΔG	ΔG_{ref}	ΔG_{exp}
9MA	-12.35	-12.00	-13.60
9MG	-21.71	-22.44	
1MT	-10.28	-12.44	$-(9.1-12.7)$
1MC	-14.85	-18.40	

via TI using AMBER 4.1. Wherever it was available we used reference experimental values from [113]. As is shown in Tab. 6.2, apart from 1MC none other of our calculated values differ from experiment or reference simulation data by any significant extent, taking account for standard chemical accuracy of ± 1 kcal/mol.

6.3 Relative Hydration Free Energies

Relative free energy calculations were done on every possible mutation of N-methylated DNA bases. The resulting free energy differences $\Delta\Delta G_{\mathrm{alch}}$ are listed in Tab. 6.3. We compare them with values $\Delta\Delta G_{\mathrm{hyd}}$ obtained through our previous

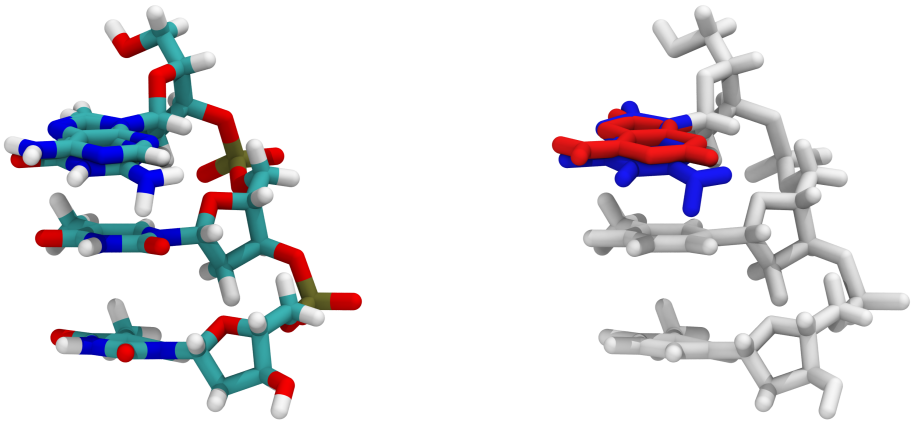
Table 6.3: Calculated differences in hydration free energy for mutations of N-methylated DNA bases. Subscript *alch* marks results of our relative calculations, *hyd* label values we obtained through absolute hydration free energies. References [114] with superscript *s* correspond to MD simulations performed with perturbations in single topology systems. Values listed in kcal/mol.

Mutation	$\Delta\Delta G_{\mathrm{alch}}$	$\Delta\Delta G_{\mathrm{hyd}}$	$\Delta\Delta G_{\mathrm{ref}}$	$\Delta\Delta G_{\mathrm{ref}}^{\mathrm{s}}$	$\Delta\Delta G_{\mathrm{ref}}^{\mathrm{hyd}}$
1MC $\rightarrow$ 1MT	4.58	4.57		5.51	5.96
1MC $\rightarrow$ 9MG	-4.98	-6.86			-4.04
9MG $\rightarrow$ 1MT	12.60	11.43	10.07		10.00
9MA $\rightarrow$ 1MC	-4.49	-2.50	-6.43		-6.40
9MA $\rightarrow$ 1MT	1.25	2.07			-0.44
9MA $\rightarrow$ 9MG	-10.41	-9.36		-11.00	-10.44

absolute calculations, cf. Tab. 6.2. Due to the lack of reference experimental data we put our results against reference MD simulations [114] only.

6.4 Mutations in Trinucleotides

Let us now step a bit further and investigate mutations of DNA bases in a slightly larger system. For that we chose trinucleotide ATT. An example of such mutation can be seen in Fig. 6.2. As a demonstration we performed 2 different base



(a) Trinucleotide hybrid **(b)** Trinucleotide hybrid: beta

Figure 6.2: Hybrid molecular structure for Ade → Gua mutation inside ATT trinucleotide. Model (a) is standardly colored by atomic species. Beta coloring in (b) features disappearing (red) and appearing (blue) residues while the untouched remains are bleached.

mutations (A→G and T→C) in 2 different positions inside the ATT sequence. tThe results can be seen in Tab. 6.4. Due to the lack of references we decided to compare our results to values obtained from our previous N-methylated DNA bases. This will allow us to at least capture a hint of the influence the rest of the trinucleotide has on the final free energy values we observe.

Table 6.4: Calculated differences in hydration free energy of ATT trinucleotide after a single base mutation occurred. Subscript *alch* marks values obtained through relative calculations of trinucleotides, *base* label our results of corresponding methylated DNA base mutations. Reference values acquired from methylated base calculations of [114]. Values listed in kcal/mol.

Mutation	$\Delta\Delta G_{\text{alch}}$	$\Delta\Delta G_{\text{base}}$	$\Delta\Delta G_{\text{ref}}$
ATT → **GTT**	−9.63	−9.36	−11.00
AT**T** → **A**C**T**	−6.41	−4.57	−5.51

6.5 Discussion

6.5.1 Absolute Free Energies

Overall the results of our absolute calculations (listed in Tab. 6.2) are in a good agreement with reference MD simulations [114] and available experimental data [113]. Apart from base 1MC, all of our hydration free energies of N-methylated DNA bases lie within the standard chemical accuracy of ± 1 kcal/mol with respect to the reference values. This means that the performance of our non-equilibrium approach is comparable to that of the reference MD simulations. Our protocol produces satisfactory outcomes.

The base at issue (1MC) resulted in approx. 3.5 kcal/mol higher hydration free energy as compared to the reference calculation. This might be due to the outdated force field of Ref. [114] and (or) due to different computational method used. Due to the lack of experimental data, we are unable to further assess the roots causing 1MC to bear such inconsistency.

6.5.2 Relative Free Energies

Our results of DNA base mutations show strong conformity with our previous absolute calculations, cf. $\Delta\Delta G_{\text{alch}}$ and $\Delta\Delta G_{\text{hyd}}$ in Tab. 6.3. Apart from mutations involving 1MC, the agreement lies within the standard chemical accuracy of ± 1 kcal/mol. The same applies to comparison with reference MD simulations [114]. The fact that only the 2 mutations involving 1MC are having trouble connecting to any reference within the chemical accuracy lets us assume that base 1MC is on fault. Though the difference is not more than ± 2 kcal/mol. On the other hand, mutation 1MC $\rightarrow$ 1MT lies perfectly within the chemical accuracy with respect to all reference values. With our current data it is hard to tell whether this single agreement is a matter of errors cancelling each other out. Due to the lack of any experimental data on this issue, we are unable to further assess the roots causing mutations of 1MC to bear these inconsistencies. Further investigation of 1MC base is needed. Nonetheless, the overall agreement suggests that our non-equilibrium approach is able to yield satisfactory results.

6.5.3 Mutations in Trinucleotides

The effects of additional molecular structure in the solvation envelope are captured in Tab. 6.4. The final change in free energy we observe during a mutation shifts based on the immediate surroundings of the mutated residue. Differences among various yet similar systems vary from approximately 0.3 to 2.1 kcal/mol. The sugar-phosphate backbone along with non-mutated bases are now part of the solvation envelope of the mutated residues. Presence of additional molecular structures changes the system, possibly leading to markedly different behaviors of the mutated bases. The differences in relative free energies we captured are traces of this influence.

7. Zif268-DNA Complex

In this chapter we explore the nature of DNA-binding proteins through investigation of zinc-finger transcription factor Zif268 bound to a short, CG-rich DNA double helix. We have covered the importance of zinc finger proteins (ZFPs) in Section 2.2 regarding zinc-finger nucleases (ZFNs), promising for genetic modifications and enhancements in various medical and industrial fields. The structure of Zif268-DNA complex (PDB: 1AAY) is shown in Fig. 7.1. It is a three-fingered

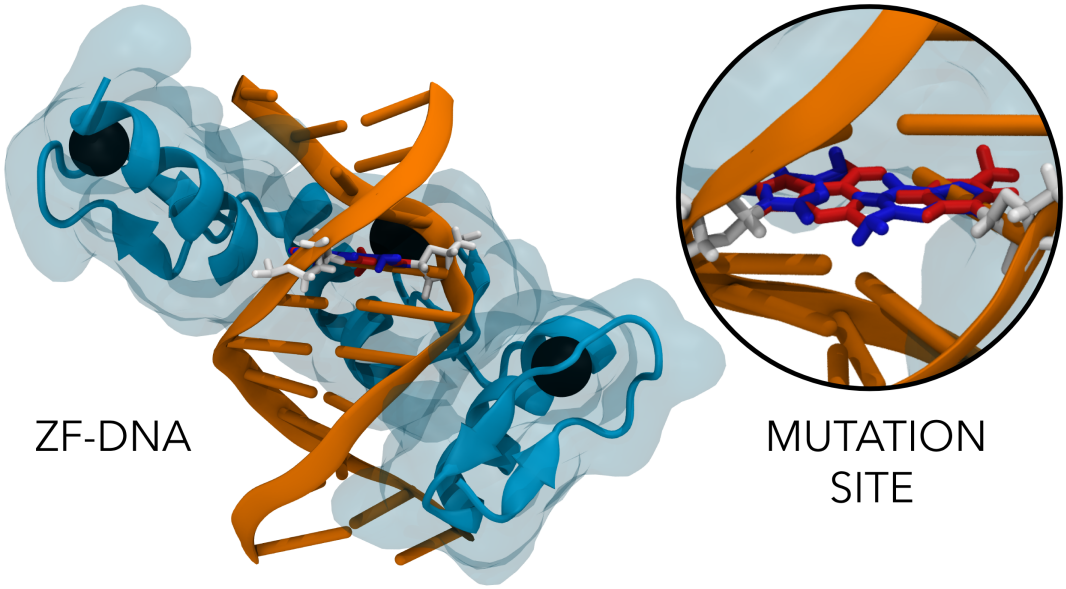


Figure 7.1: ZF-DNA complex (PDB: 1AAY, modified in mutation site) with detail zoomed in on one of the possible mutations T=A → G≡C. Transcription factor Zif268 (darker cyan) is displayed in 2 mutually overlayed representations – ribbons demonstrate the underlying protein structure while the transparent surface showcases its outer shape. Inside the protein there are 3 Zn^{+2} ions (black spheres). The orange DNA double helix harbors sequence GCGTGGGCG, recognized by the transcription factor. Mutated residues are shown in beta coloring – red disappearing (initial) and blue appearing (final) base pair, while the untouched sugar-phosphate moieties are bleached.

variant from the class of ZFPs. Each zinc finger domain comprises of one β-turn connected to an α-helix, between which a single Zn^{+2} ion is non-covalently bound. Zif268 is known to recognize DNA sequence GCGTGGGCG. Detection of nucleobases is done via specific amino acid side chains pointing inwards into the major groove of DNA, see Fig. 7.2 for schematic depiction of all binding sites.

As is introduced in Fig. 7.2, we will distinguish between the main (1S) and the complementary (2S) DNA strands in the following text. Individual positions of base pairs in the DNA double helix will be referred to according to the red numbers in Fig. 7.2. Given these rules, position 4 (or equivalently *mutation site* 4) features nucleobases 1ST and 2SA. Direct interaction with proximate amino acid side chains is depicted by arrows pointing to the bases. Position 4 thus harbors 2SA base directly interacting with Asp76. Close by Arg74 influences the site only indirectly through interactions with its neighbour Asp76. 1ST base does not have any amino acid in its immediate vicinity. The rest of the sites $2-9$ follow this example.

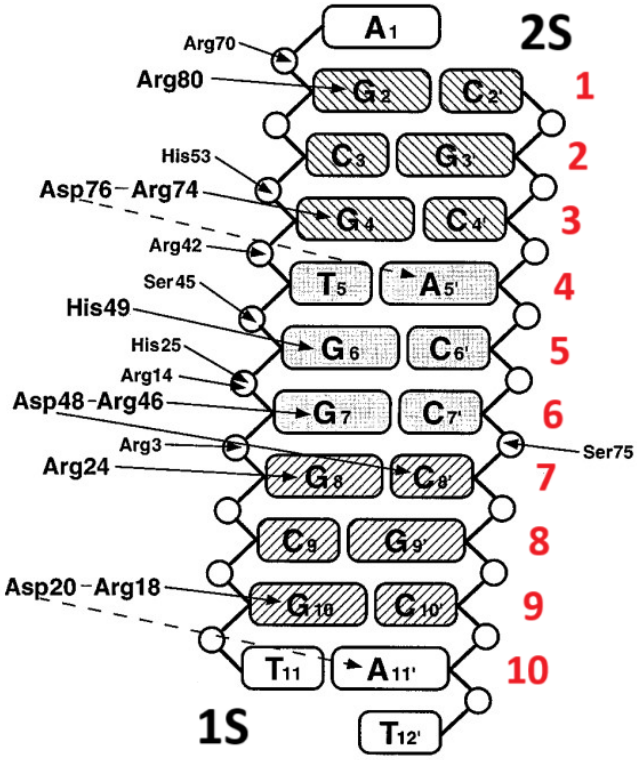


Figure 7.2: Base pairs of short DNA double helix (sequence GCGTGGGCG) recognized by transcription factor Zif268, and proximate amino acid side chains forming the protein's binding sites. We introduce numbering of base pair positions in red, and abbreviations (1S and 2S) for each strand of the DNA double helix. Individual bases in each site will be marked by these prefixes, e.g. position 4 hosts bases 1ST and 2SA. Image adapted from [??].

Even though the Zif268 protein ought to recognize the given DNA sequence, binding to slightly different sequences is not completely ruled out. The most likely candidates for unwanted off-target binding are DNA sequences carrying single base pair mutations, some of which may still disrupt Zif268-DNA binding substantially. Impact of single base pair mutations on the stability of the Zif268-DNA complex will be quantified using MD simulations in this chapter.

As we have discussed earlier within Section 3.5.3, thermodynamic cycles are an efficient way to gain relative binding free energies. Fig. 7.3 presents such a thermodynamic cycle for our Zif268-DNA complex. In principle, it yields 2 alternative pathways for acquiring differences in binding free energies for original and mutated DNA double helix. Speaking the language of thermodynamics

$$\Delta\Delta G_{\text{bind}} = \Delta G_1 - \Delta G_2 = \Delta G_3 - \Delta G_4 \tag{7.1}$$

is the change in binding free energy of the complex due to a single base pair mutation inside the DNA double helix. One of the possibilities features alchemical transformation of decoupling the whole protein from its environment, i.e. computing separately ΔG_3 and ΔG_4. Such a dramatic change would require a lot of computer time. Therefore, the only reasonable pathway is by performing the mutation in DNA bound to the protein (ΔG_1) and separately in water without the protein's presence (ΔG_2).

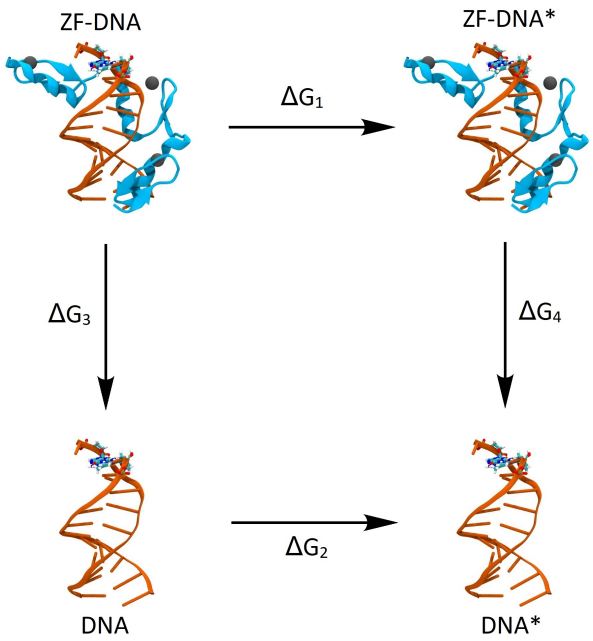


Figure 7.3: Thermodynamic cycle for a single DNA base pair mutation inside the Zif268-DNA complex. Values ΔG_i represent changes in free energy of the system after a given alchemical transformation has occurred. Symbol * marks structures with the final (target) DNA base pair, i.e. product of the mutation.

Relative binding free energy calculations for protein-DNA complexes emerged relatively recently [115, 116]. It is a subject of an ongoing research. Not so long ago, different software packages yielded different results [117]. Specifically, the AMBER software package vastly overestimated the magnitude of binding free energy changes and often with incorrect signs. A very recent study has provided consistent results for different software packages [6]. By chance one of the leading MD software packages, NAMD, escaped the scope of this study. This chapter of the present thesis should fill this gap.

7.1 Simulation Setup

Following the literature [115], we use non-equilibrium MD calculations to gain free-energy differences for the Zif268-DNA complexes with different single base pair mutations. The reference study [115] applied two fundamentally different approaches: non-equilibrium Crooks-Gaussian intersection (CGI) [108] coupled with Thermodynamic Integration (TI) [82] and equilibrium Hamiltonian Replica Exchange [118, 119] with Multistate Bennet's Acceptance Ratio (RE/MBAR) method [86]. Further, they used a linear soft core potentials. All of their MD simulations were carried out employing the Gromacs software package [43] with the AMBER99SB force field [23].

In contrast, we used the CHARMM36/CgenFF force fields [45], free energy perturbation (FEP) implemented in the NAMD software package [44] together with our version of the CGI protocol (described in Chapter 4). Unlike the reference study [115], we do not utilize any constraints that would artificially stabilize Watson-Crick hydrogen bonding in mutated base pairs.

The temperature was maintained at 300 K by the Langevin thermostat [61] while the Langevin piston barostat targeted pressure at 1.0 bar. Simulated systems were propagated in time using the RESPA algorithm [58] with a 1.0 fs time step. PBC and PME were applied.

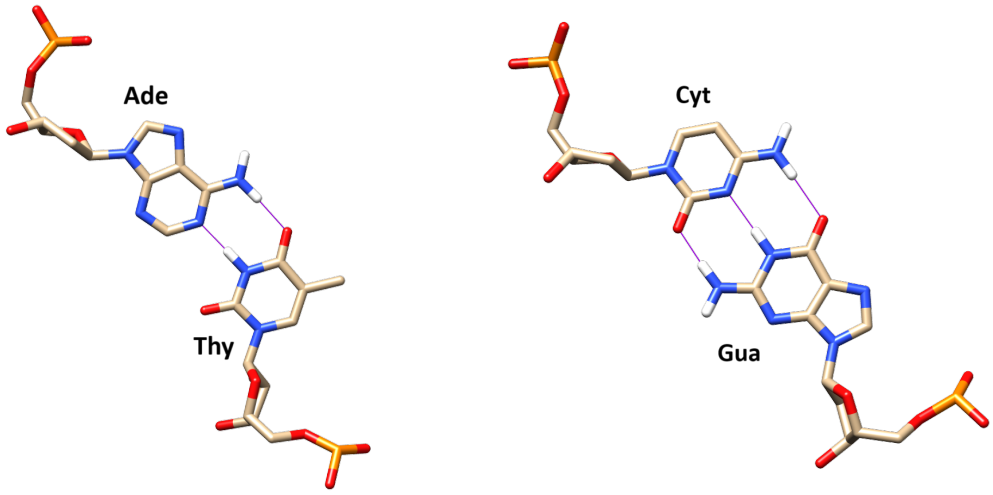
First, we performed a minimization and equilibration procedure (500 ps), after which followed non-equilibrium FEP calculations (recall Section 4.3) consisting of swarms of 100 independent MD runs, each composed of forward and backward stages. Each independent MD run included an additional short 10 ps thermalization to ensure slightly different starting conformations of forward MD runs. Backward MD runs used the final conformations of forward MD runs as a starting point. Each of the 20 λ windows consisted of $5 \cdot 10^4$ FEP steps. The final work values from non-equilibrium forward and backward MD runs were analyzed via our Python program (mentioned above in Section 4.6) that determines CGI of histograms pinpointing out the equilibrium binding free energy.

7.2 Initial Equilibration

First, we examine the initial equilibration MD trajectories. More specifically, we are interested in whether the Watson-Crick hydrogen bonding was preserved in mutated base pairs. Since equilibration MD runs were produced with $\lambda = 0$, mutated (final, target) base pairs were detached from their environment. Of course, they were still covalently bound to the sugar-phosphate backbone of DNA and felt each other.

7.2.1 Watson-Crick Hydrogen Bonding

In each mutation site, the length of the central Watson-Crick hydrogen bond was measured for each MD run. Gathered data were plotted against the recorded trajectory frames. The average values and standard deviations (STD) were computed as well. Let us describe general patterns observed in all equilibration MD runs. Example snapshots of DNA base pairs taken from our MD trajectories are shown in Fig. 7.4. Their structures differ only slightly due to thermal fluctuations



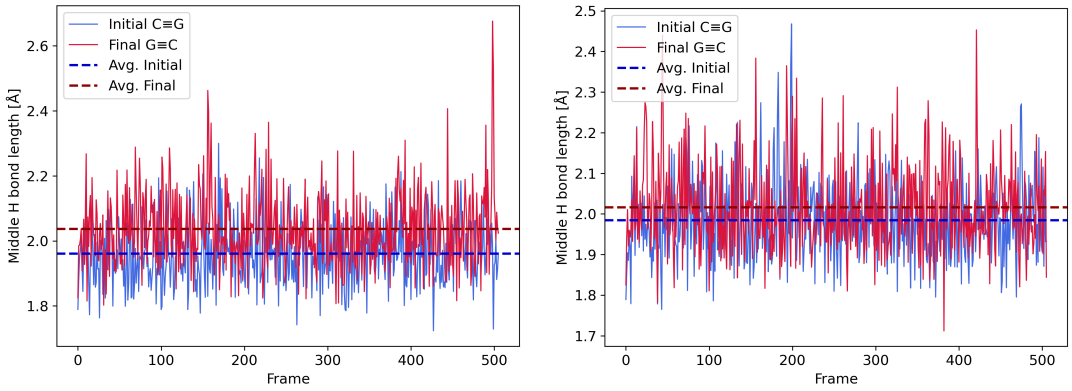
(a) T=A **(b)** G≡C

Figure 7.4: Hydrogen bonding (purple lines) in selected DNA base pairs. The rest of the DNA double helix as well as the whole protein structure are hidden for clarity. We monitored the central Watson-Crick hydrogen bond to assess the base pair stability within initial equilibration MD runs with $\lambda = 0$ when the mutated base pairs were decoupled from their environment.

in the simulated systems. The average length of the central Watson-Crick hydrogen bond is always approx. 2.0 Åwith STD ranging from 0.1 to 0.3 Å. This applies to both *in aqua* and *in protein* cases, no matter whether it is the initial (coupled) or the final (decoupled) base pair. An example of typical sampled distances is given in Fig. 7.5. Thermal fluctuations yield no more than a 0.6 Ådifference from the average value.

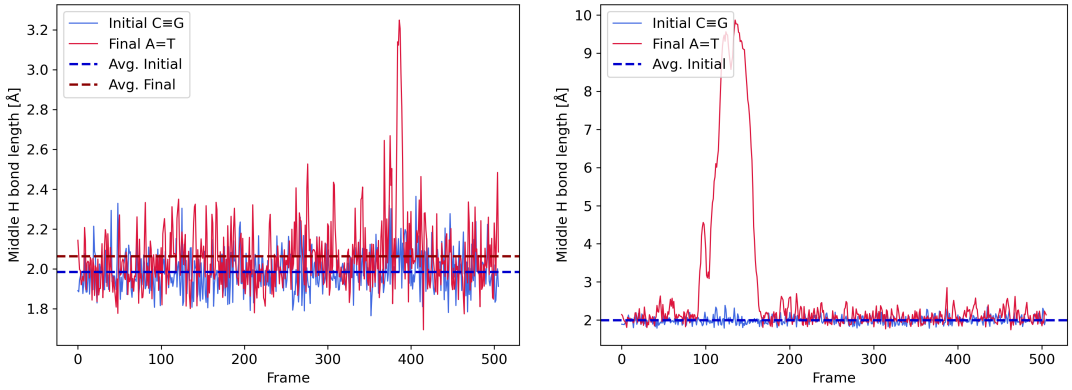
Overall, the hydrogen bonding of mutated bases was stable. All plots closely resembled the example in Fig. 7.5. However, there is one particular exception, where the mutated base pair was for a brief moment broken. It occurred for the decoupled (final) A=T base pair within the *in protein* C8A equilibration MD run, see Fig. 7.6. Nevertheless, after about 80 frames the base pair was re-established

again and remained stable for the rest of the MD run. This was the only transient instability in mutated base pairs we encountered in equilibration MD runs.



(a) In aqua **(b)** In protein

Figure 7.5: Time evolution of lengths of typical central hydrogen bonds in mutated DNA base pairs within equilibration MD runs: (a) *in aqua* i.e. DNA double helix in water; (b) *in protein* i.e. solvated complex of DNA double helix bound to the ZF protein. In all cases, average values fall close to 2.0 Å with STD ranging from 0.1 to 0.3 Å.



(a) In aqua **(b)** In protein

Figure 7.6: Time evolution of central hydrogen bond length in mutated DNA base pair during initial equilibration C8A MD run: (a) *in aqua* i.e. just DNA double helix in water; (b) *in protein* i.e. solvated complex of DNA double helix bound to the ZF protein. The latter chart reveals a brief transient disconnection of the final (i.e. mutated and from its environment decoupled) A=T base pair.

Summarized, we did not observe any differences in base pairing among *in aqua* and *in protein* simulated systems, nor between original (to its environment coupled) and mutated (from its environment decoupled) base pairs. Therefore, we concluded that it would not be necessary to apply any artificial constraints that would additionally stabilize the mutated base pairs (as they did in the reference study [115]).

7.2.2 Amino Acid Side Chains

We also studied interactions between spatially close amino acids – mostly attraction to each other. Generally, the strongest interactions were observed between

side chains of amino acids with opposite charges that form the so-called salt bridges. Fig. 7.7 shows the Arg46-Asp48 salt bridge from the G7C equilibration MD run. Similar salt bridges were found in many other binding sites (3, 4, 6, 9).

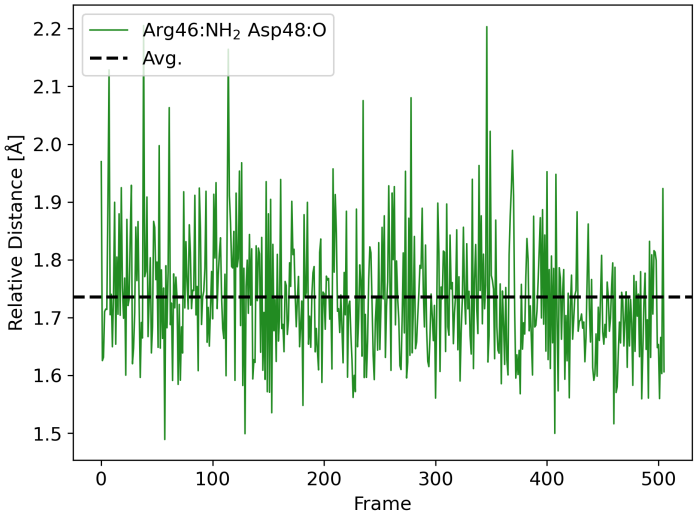


Figure 7.7: Attraction between neighboring Arg46 and Asp48 side chains during the G7C equilibration MD run. The average mutual distance of (1.7 ± 0.1) Å is even a little bit shorter than the typical average hydrogen bond length in DNA base pairs.

Moreover, we monitored interactions between amino acids and mutated base pairs. Fig. 7.8 captures the switch between two oxygen atoms of the Asp76 carboxyl group as regards their binding to the $-NH_2$ group of Adenine base.

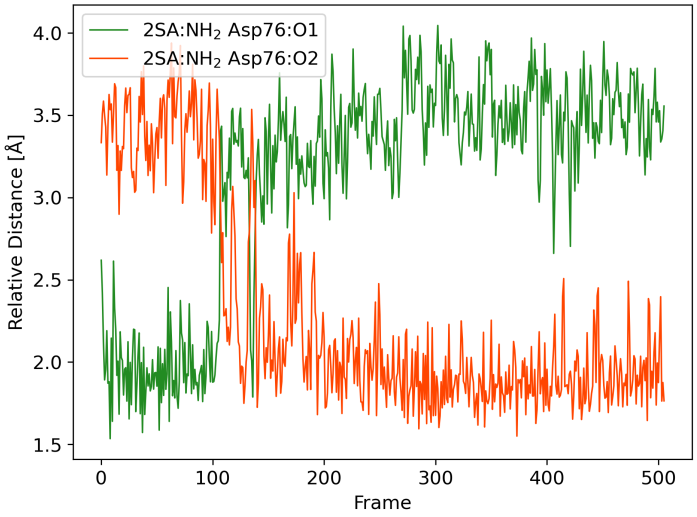


Figure 7.8: Switch between two oxygen atoms of the Asp76 side chain as regards binding to the $-NH_2$ group of an Adenine base, captured in mutation site 4 during the T4C equilibration MD run.

7.3 Relative Binding Free Energies

We performed 14 different mutations in positions $2-9$ of the DNA double helix recognized by the Zif268 transcription factor, recall Fig. 7.2 for the schematic depiction of binding sites. In positions 2, 4, and 8 we performed all possible mutations starting from the original original base pairs. Positions 2 and 8 do not directly interact with any AA side chains. Mutation site 4 represents a typical binding site of Zif268, and serves us as an exemplary case to test most of the effects of base pair mutations. In the rest of the binding sites (3, 5, 6, 7, and 9) we performed transformations G≡C → C≡G, switching the positions of original bases from one DNA strand to another in very similar environments of proximate amino acids. The resulting differences in binding affinity of the complex are plotted in Fig. 7.9. Mutations G6C and G7C tend to decrease the

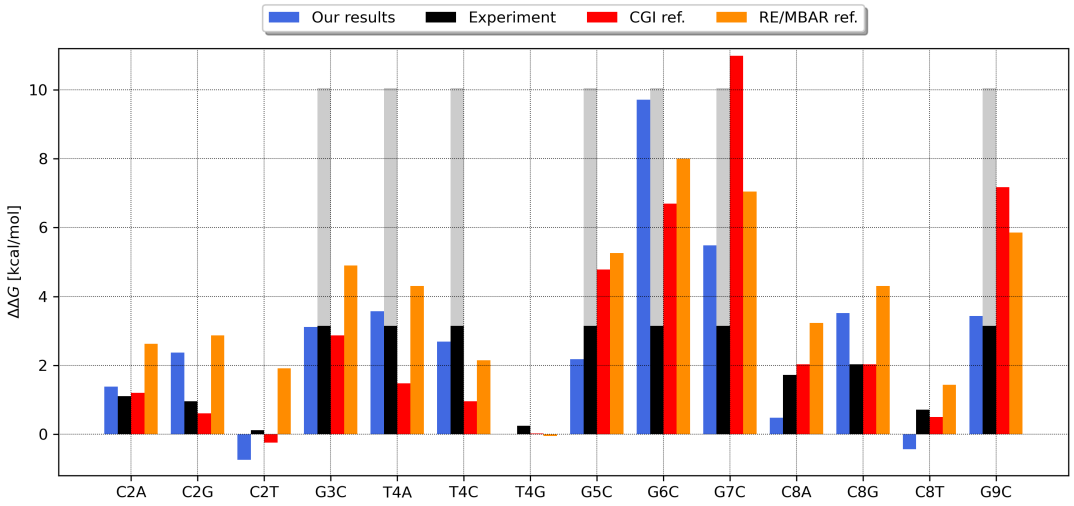


Figure 7.9: Binding affinity differences for Zif268-DNA complex, featuring 14 different bp mutations. Results of our MD simulations are shown in blue, against reference simulations [115] (CGI in red, RE/MBAR in orange). Experimental results [120] are colored in black, with lighter shades indicating binding affinity decrease of more than 3.15 kcal/mol where exact values could not be extracted.

binding affinity by the largest amount, approx. 9.8 and 5.6 kcal/mol respectively. Certain mutations performed in positions 2 (C2A, C2T) and 8 (C8A, C8T) have only a small effect, yielding values around 1 kcal/mol and less. A special case of mutation, T4G, has de facto zero effect on the binding affinity. The rest of mutations leads to free energy differences of moderate values from little above 2 to around 3.5 kcal/mol.

Overall, our results are in a good agreement with the reference MD simulations [115] as well as with experimental data provided in [120]. Though certain mutations lead to discrepancies with some of the reference values; namely C2G, C2T, C8G, and C8T. In order to get an idea of what is actually happening due to all of these mutations we have to look into each one of the mutation sites in detail.

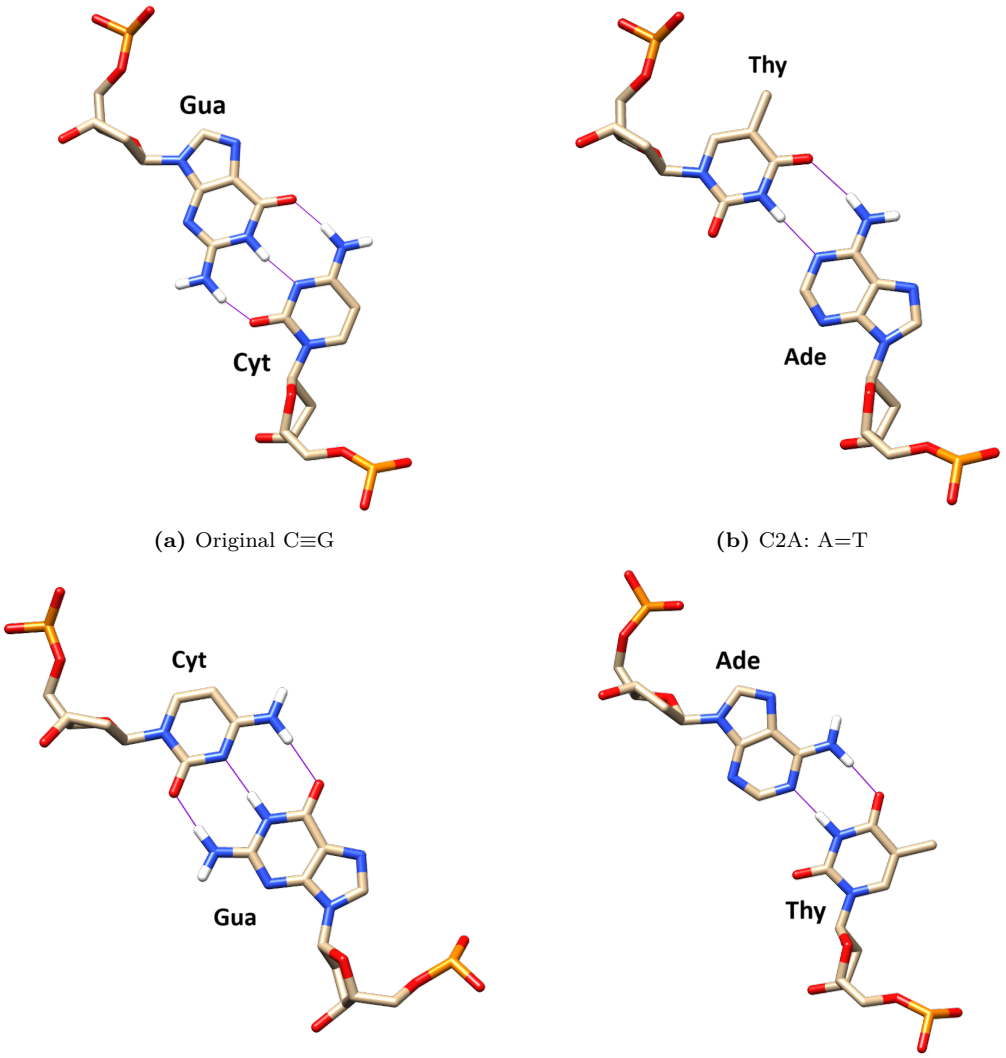
We investigate stability of base pairs in the ZF-DNA system in regards to the alchemical transformations performed. In sites where amino acid side chains are present the effort is put towards finding traces of their possible influence on

the free energy difference $\Delta\Delta G$ we observe during base pair mutations. In the following we will show and discuss examples of repeating patterns we observe throughout all of the mutations performed as well as some of the individual cases pointing us to the explanation of the protein's sequence detection mechanism.

As was mentioned times and times again, FEP transformations vary interactions between chosen (mutated) residues and their surroundings via alchemical parameter $\lambda : 0 \leftrightarrows 1$. If one does not wish to study the alchemical transformation itself, the only relevant conformations holding true physical meaning are the endpoints of those transformations. This leads us to an analysis of mere pseudo-trajectories created by snapshots with base pairs that are fully coupled to their environment, i.e. in alchemical windows where λ is either exactly 0 or 1. Other windows offer structures where the mutated residues 'feel' their environment only partially, which is inherently non-physical. Such conformations would provide us with no relevant information to draw any meaningful conclusions from. Our analysis is thus closer to that of Monte Carlo data instead of standard molecular dynamics trajectories.

7.4 Mutation Site 2

Here we performed all possible mutations from the initial C≡G pair none of which resulted in a major disruption of binding affinity of the complex; recall Fig. 7.9, mutations C2A, C2G, and C2T. Snapshots of all the base pairs from our simulations can be seen in Fig. 7.10. Since there are no amino acid side chains in the vicinity of this site, the most important information we can extract from the sampled conformations is their overall structure and base pair stability.



(a) Original C≡G

(b) C2A: A=T

(c) C2G: G≡C

(d) C2T: T=A

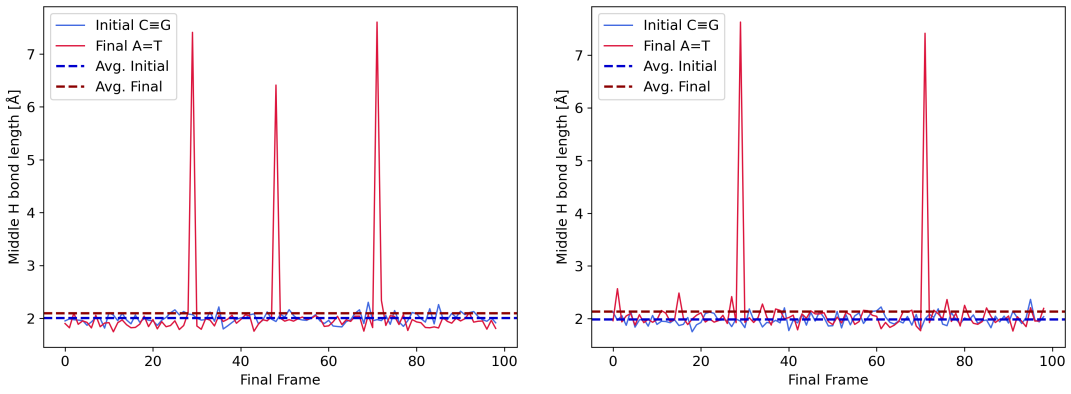
Figure 7.10: Detail of mutation site 2 – recall Fig. 7.2. For clarity, the remaining parts of the DNA double helix and the protein are hidden. We show the original C≡G base pair in (a), the rest feature snapshots of the site after a base pair mutation has occurred, in our notation (b) C2A, (c) C2G, and (d) C2T. In the vicinity of this site there are no amino acid side chains for the nucleobases to directly interact with. Purple lines mark hydrogen bonding.

Structure can be investigated visually from the snapshots of base pairs that are fully coupled to their environment, i.e. in alchemical windows where λ is either exactly 0 or 1. Other windows offer structures that are inherently non-physical, hence providing us with no relevant information to draw any conclusions from. As is demonstrated in Fig. 7.10, DNA bases are overall planar though they are

not completely rigid. Base pairs twist and bend a bit in thermal fluctuations and due to stacking interactions with surrounding bases. Nothing abnormal was observed for any of the studied base pairs. Similar limitations as those mentioned above apply to the study of hydrogen bonding between bases – the only relevant structures are the endpoints of our FEP transformations.

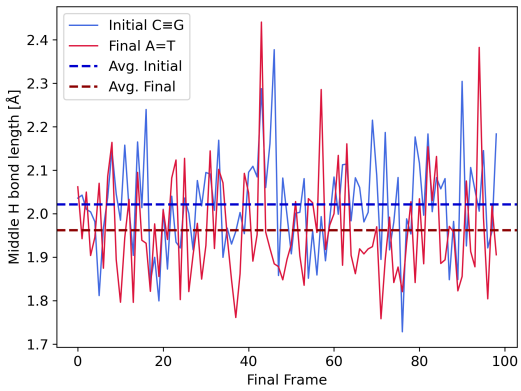
7.4.1 Mutation C2A

As a measure of base pair stability we chose the central hydrogen bond length. Fig. 7.11 shows lengths measured for mutation C2A. In water without the protein

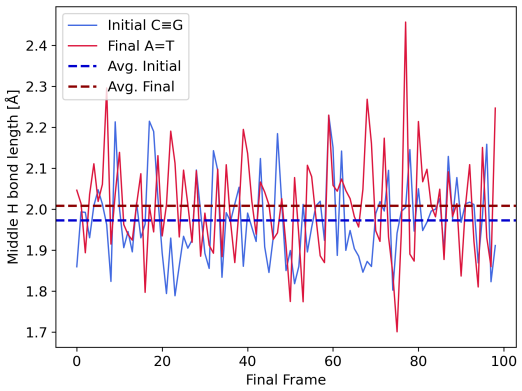


(a) Forward runs in aqua

(b) Backward runs in aqua



(c) Forward runs in protein



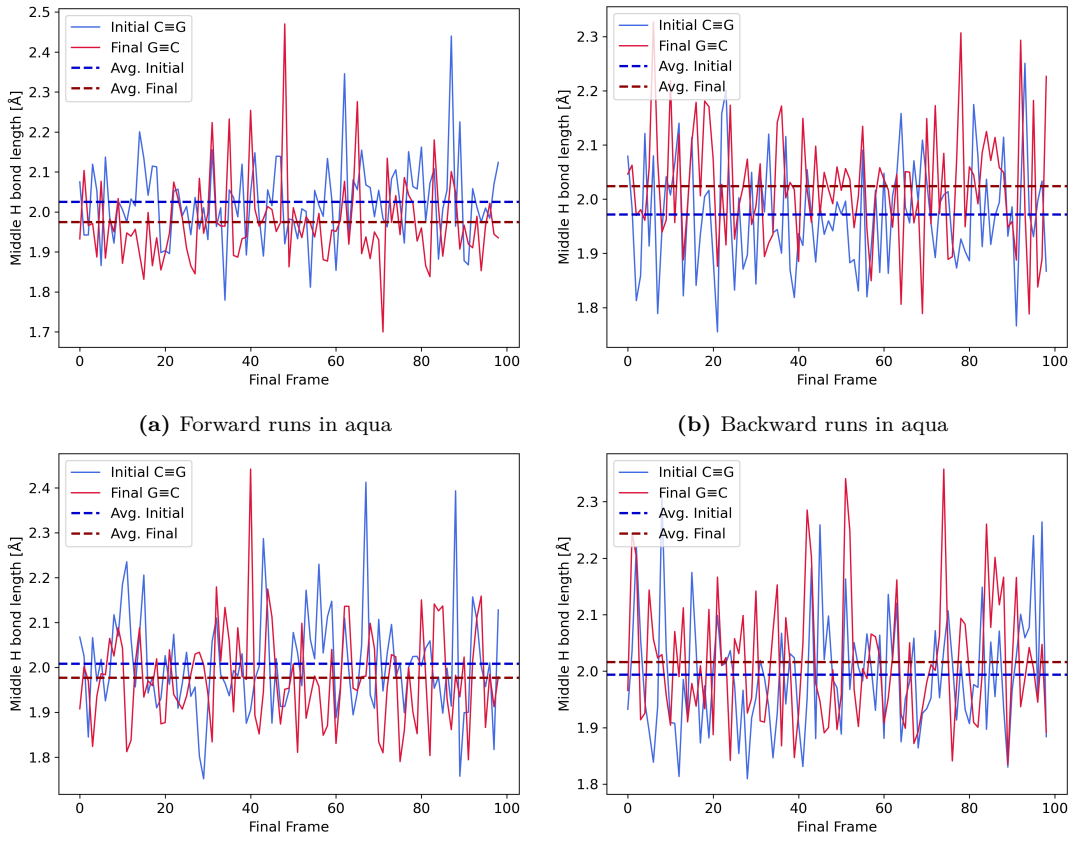
(d) Backward runs in protein

Figure 7.11: Analysis of mutation C2A – central hydrogen bond length between DNA bases against the number of the corresponding FEP run (final frame). Blue color is used for the initial base pair, i.e. before the mutation ($\lambda = 0$), and red color indicates the final base pair, i.e. after the mutation ($\lambda = 1$). Horizontal lines mark the sample average values of the central hydrogen bond length.

we encountered 3 disconnections for the final A=T pair (Fig. 7.11a), backward transformation was able to reconnect one of them back but the rest remained broken all through (Fig. 7.11b). The separation of disconnected bases lies within a range of 6.5 and 7.5 Å. Sample average value of central hydrogen bond lengths of both pairs (excluding the disconnections) is approx. (2.0 ± 0.2) Å no matter the coupling to environment. When the protein is present no broken base pairs are recorded and the sample average value of central hydrogen bond lengths is approx. (2.0 ± 0.1) Å in all cases, see Fig. 7.11c and 7.11d.

7.4.2 Mutation C2G

When performing mutation C2G no base pair disconnections were recorded, see Fig. 7.12 for the results. In the case where the double helix is floating in water without the ZF protein forward runs lead to average central hydrogen bond lengths to be (2.0 ± 0.1) Å, backward runs feature (2.0 ± 0.2) Å. Both alchemical transformation directions in the MD simulation with the protein present give average central hydrogen bong lengths of (2.0 ± 0.2) Å. In all cases coupling of base pairs to their environment makes no significant difference in hydrogen bonding.



(a) Forward runs in aqua

(b) Backward runs in aqua

(c) Forward runs in protein

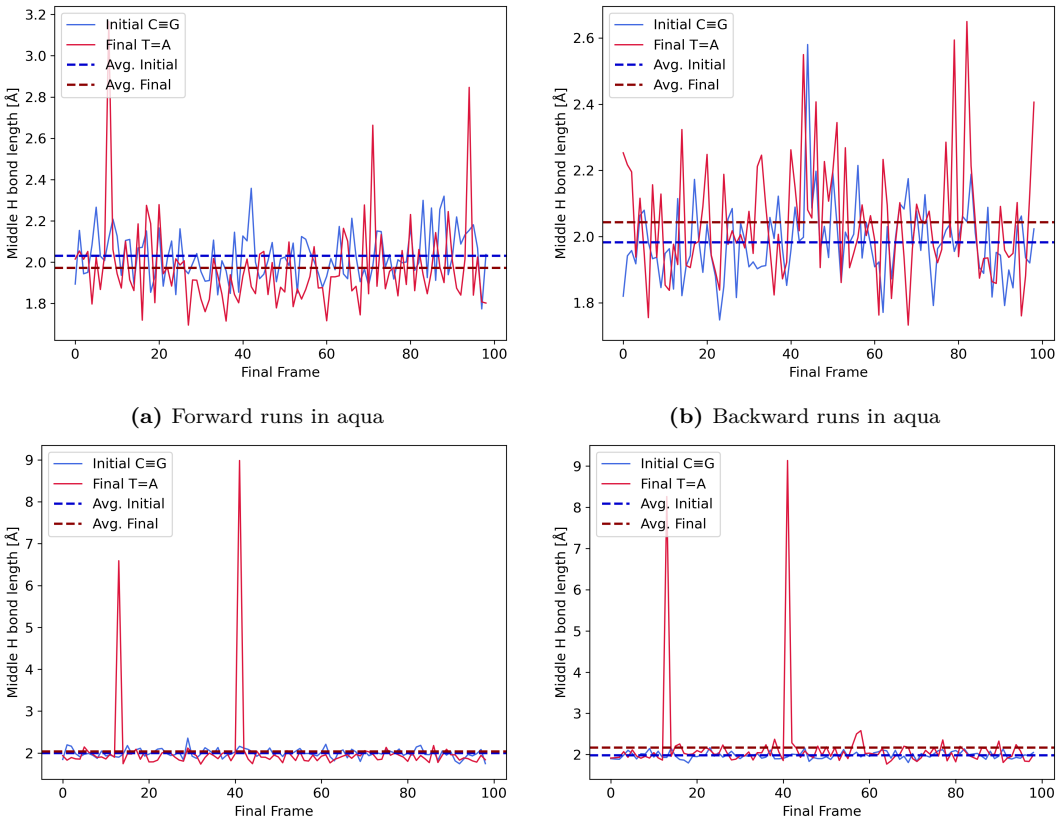
(d) Backward runs in protein

Figure 7.12: Analysis of mutation C2G – central hydrogen bond length between DNA bases against the number of the corresponding FEP run (final frame). Blue color is used for the initial base pair, i.e. before the mutation ($\lambda = 0$), and red color indicates the final base pair, i.e. after the mutation ($\lambda = 1$). Horizontal lines mark the sample average values of the central hydrogen bond length

7.4.3 Mutation C2T

The last possible mutation in this site, C2T, gives base pair hydrogen bonding shown in Fig. 7.13. Here breaking of hydrogen bonds happened only for final T=A pair in the case of the whole protein complex. Both disconnections formed during forward transformations and persisted even through the following backward runs, cf. Fig. 7.13c and 7.13d. Magnitudes of these mutual distances range from around 7.0 and 9.0 Å. Sample average values of the central hydrogen bonds

are (2.0 ± 0.2) Å no matter the protein's presence or coupling of bases to their environment.



(a) Forward runs in aqua

(b) Backward runs in aqua

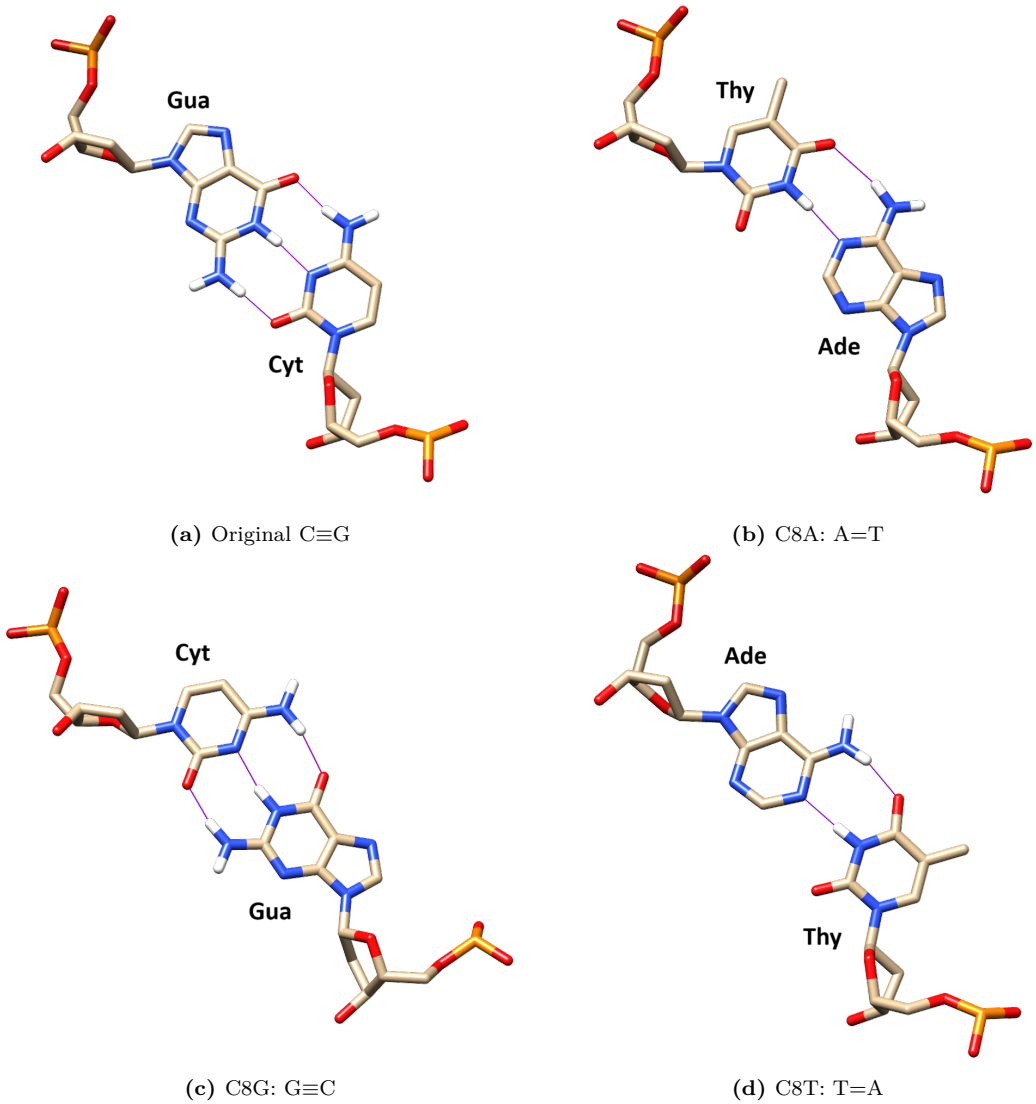
(c) Forward runs in protein

(d) Backward runs in protein

Figure 7.13: Analysis of mutation C2T – central hydrogen bond length between DNA bases against the number of the corresponding FEP run (final frame). Blue color is used for the initial base pair, i.e. before the mutation ($\lambda = 0$), and red color indicates the final base pair, i.e. after the mutation ($\lambda = 1$). Horizontal lines mark the sample average value of the central hydrogen bond length.

7.5 Mutation Site 8

The second mutation site without any amino acid side chains in its vicinity is mutation site 8, originally with C≡G base pair; cf. Fig. 7.2. Performing all possible mutations also resulted in only minor disruptions towards the binding affinity of the complex, recall mutations C8A, C8G, and C8T in Fig. 7.9. This makes it a complete analogy to mutation site 2. Snapshots of all base pair structures are given in Fig. 7.14. Due to the almost identical nature of these mutations to those discussed earlier we will cover this section briefly in an analogy to its twin above.



(a) Original C≡G

(b) C8A: A=T

(c) C8G: G≡C

(d) C8T: T=A

Figure 7.14: Detail of mutation site 8 – recall Fig. 7.2. For clarity, the remaining parts of the DNA double helix and the protein are hidden. We show the original C≡G base pair in (a), the rest feature snapshots of the site after a base pair mutation has occurred, in our notation (b) C8A, (c) C8G, and (d) C8T. In the vicinity of this site there are no amino acid side chains for the nucleobases to directly interact with. Purple lines mark hydrogen bonding.

7.5.1 Mutation C8A

Mutation C8A causes disconnections only during transformations with DNA bound to the ZF protein. Forward runs broke 3/100 final A=T pairs, the subsequent backward transformations reconnected one of them back. This makes it again around 2 to 3 % disconnection rate. The disconnected bases were separated by a distance of 4.5 to 6.0 Å. C8A base pair mutations of DNA duplex in water (i.e. without bound protein) led to only a single final A=T pair being on a verge of being disconnected with around 3.5 Å mutual distance. In all cases, whether the protein was present or not, the average central hydrogen bond length is (2.0 ± 0.2) Å (broken base pairs were excluded).

7.5.2 Mutation C8G

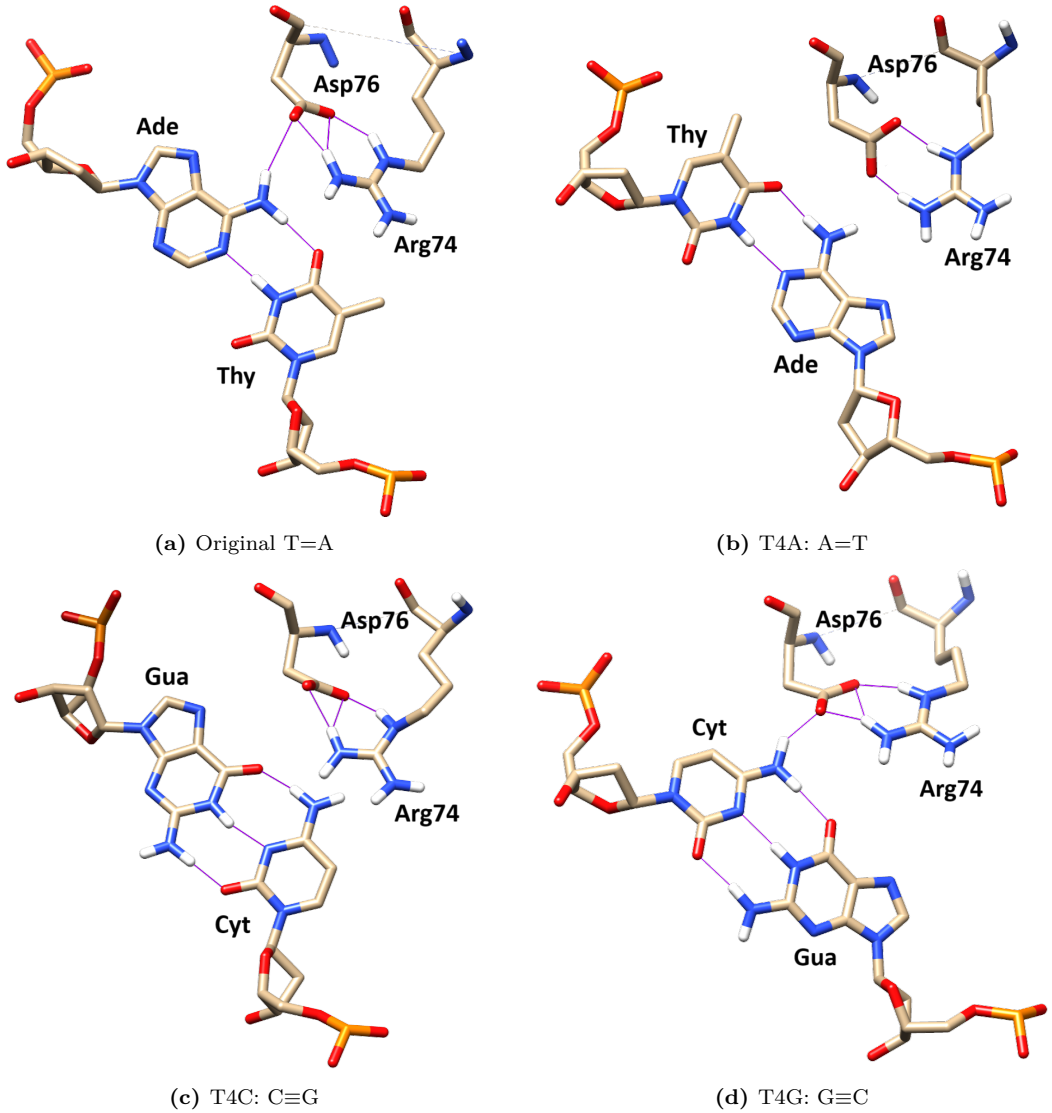
Performing mutations C8G leave no base pairs divided. Both alchemical transformation directions in water without the protein give an average central hydrogen bond bond length between bases (2.0 ± 0.2) Å. Forward MD runs with ZF protein present in the simulated system yield the central hydrogen bond of (2.0 ± 0.1) Å, backward runs give (2.0 ± 0.2) Å.

7.5.3 Mutation C8T

Mutations C8T disconnected only a single final T=A base pair during a forward transformation without the protein's presence, backward run was able to reconnect it afterwards. Another base pair split appeared during one of the forward transformations in the context of the protein complex and remained disconnected all through. In water the sample average central hydrogen bond linking any base pair is always (2.0 ± 0.1) Å. With the transcription factor the bond reads an average of (2.0 ± 0.2) Å for the final T=A pairs while initial C≡G pairs give (2.0 ± 0.1) Å (broken pairs excluded).

7.6 Mutation Site 4

This position features the 2 most common AA residues (Asp, Arg) appearing in binding sites of transcription factor Zif268, rendering this to be the ideal case for studying general traits we observe in any of our mutation sites. Here we performed all possible base pair mutations (T4A, T4C, and T4G), see Fig. 7.15. Asp76 directly interacts with the 2S base while Arg74 resides one level above the mutation site, interacting with these bases only indirectly via pulling on its closest neighbour Asp76. By looking at distances between spatially close chemical groups of DNA bases and amino acids, and the charges at play, we can get an insight into why certain mutations are disruptive and others rather benign as regards the binding affinity of the complex.



(a) Original T=A **(b)** T4A: A=T

(c) T4C: C≡G **(d)** T4G: G≡C

Figure 7.15: Detail of mutation site 4 – recall Fig. 7.2. For clarity, the remaining parts of the DNA double helix and the protein are hidden. We show the original T=A base pair in (a), the rest feature snapshots of the site after a base pair mutation has occurred, in our notation (b) T4A, (c) T4C, and (d) T4G. Nearby amino acids Asp76 and Arg74 affect these base pairs. Note that Arg74 resides one level above the present site, i.e. on the level of the 3rd position in the DNA double helix. It influences the base pairs only indirectly by pulling on its closest neighbour Asp76. Purple lines mark hydrogen bonding.

7.6.1 Mutation T4A

Switching the original base pair through T=A → A=T results in a non-negligible disruption towards the binding affinity of the complex, recall T4A in Fig. 7.9. Comparing the site before and after the mutation, i.e. 7.15a with 7.15b, reveals that Asp76 tends to prefer binding to Arg74 fully after the base pair switch as opposed to the original sharing of hydrogen bonding with $-NH_2$ group of 2SA.

Using our simulation setup we sampled 100 different final conformations for both forward and backward transformation (T=A ⇆ A=T) and measured the distances between proximate chemical groups of fully coupled 2S bases and the side chain of Asp76; results can be seen in Fig. 7.16. The initial 2SA base interacts

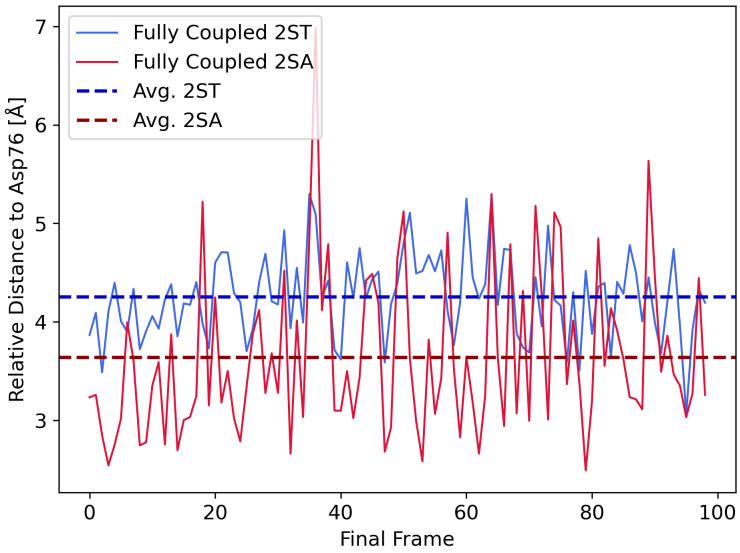


Figure 7.16: Mutation T4A – interactions between 2S bases and nearby Asp76. Reference point is represented by the carbon of $-COO^-$ side chain. For 2SA the distance is measured to the nitrogen of its $-NH_2$, for 2ST to its =O. Keep in mind that these are relative distances measured in the endpoint conformations ($\lambda = 1$) with bases fully coupled to the environment.

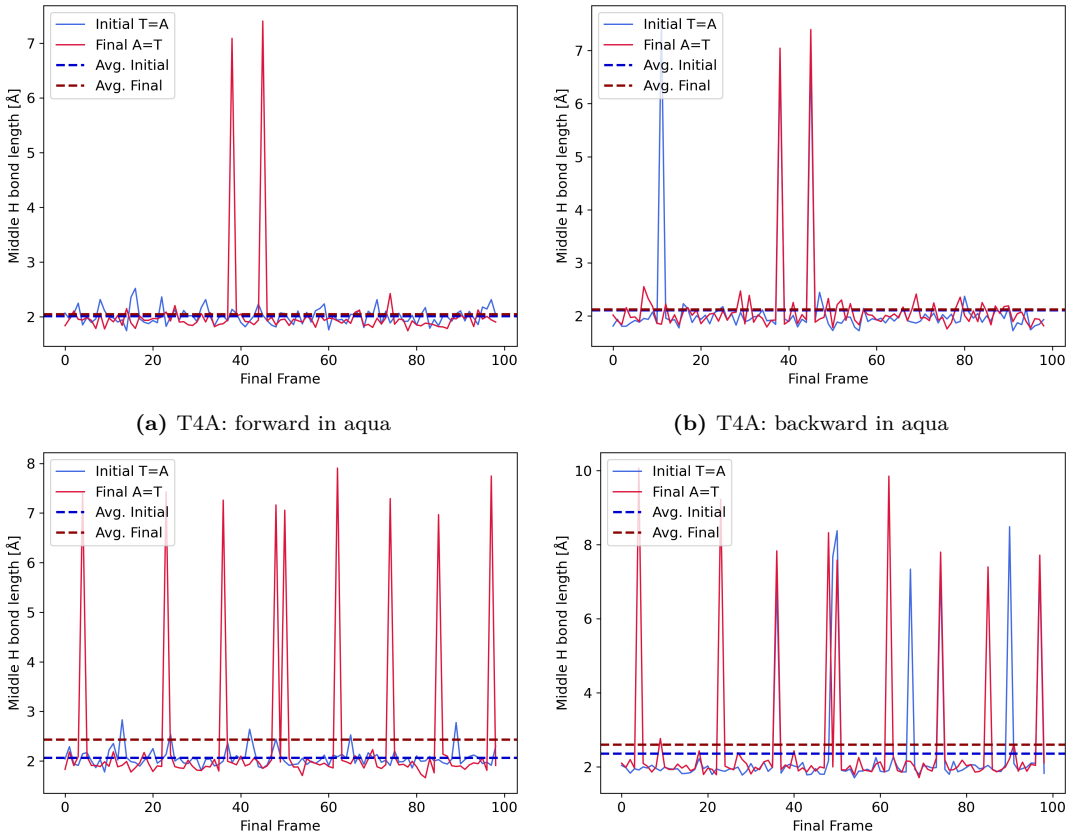
with its environment fully during the final $\lambda = 1$ window of the backward run. The sample average distance between the nitrogen of its $-NH_2$ group and the carbon of the $-COO^-$ side chain is approx. (3.6 ± 0.8) Å. Final conformations of the forward runs have the target 2ST base fully coupled, with the sample average distance (4.3 ± 0.4) Å between 2ST =O group and the $-COO^-$ carbon of Asp76.

Looking at the distances in Fig. 7.16 one can spot a hint of repulsion between target 2ST base and Asp76 in a form of a 0.7 Å gap among the averages marked by dashed lines. This is a proof of preference towards the original base pair as opposed to the mutation performed.

Visual inspection of the sampled interaction sites, recall snapshots in Fig. 7.15, confirms that 2SA indeed tends to form hydrogen bond with the carboxylic side chain via the proximate $-NH_2$ group. The average hydrogen bond length is approx. 3.6 Å, though individual sampled values vary greatly. This is due to switching (recall Fig. 7.8) between different donor and acceptor atoms of these groups, which makes it rather complicated to capture with reasonable STD. Nearby Arg74 competes for the hydrogen bonding with Asp76. When the mutation to 2ST base is performed Asp76 side chain commits to Arg74 all the more, since there is no

proximate positive partial charge offered by the 2ST base.

Another point of interest is the pairing of DNA bases. For probing the stability of base pairs we again make a use of the central hydrogen bond lengths between the given bases. Each of the lengths measured in the final conformations are plotted against the number of the corresponding run, see Fig. 7.17.



(a) T4A: forward in aqua

(b) T4A: backward in aqua

(c) T4A: forward in protein

(d) T4A: backward in protein

Figure 7.17: Mutation T4A – central hydrogen bonds of DNA bases. Forward runs (a) show 2 disconnected final A=T pairs, each separated by 7.0 to 7.5 Å. The following backward runs (b) disconnected an additional initial T=A pair. Forward runs (c) feature 9 broken final A=T pairs, each separated by approx. 7.5 Å. Backward runs (d) drove them apart even a bit more (approx. 8 to 10 Å), and added 4 more disconnections from the pool of initial T=A pairs.

Base pairs during this particular mutation are overall stable, though there are some simulations which resulted in disconnected pairs. Central hydrogen bond breaking is more pronounced for final A=T pairs, and especially while bound to the protein where the disconnection rate goes up to 10 %. The sample average value of the central hydrogen bond length between DNA bases is (2.0 ± 0.2) Å in water, with the protein present it reads (2.0 ± 0.1) Å – excluding defective pairs.

7.6.2 Mutation T4C

Mutating through T=A → C≡G, cf. Fig. 7.15 for structures, has disruptive effects of the same magnitude as mutation T4A, recall Fig. 7.9. Similarly as in the previous case, the target 2SG base has no positive partial charge in the vicinity of Asp76 side chain to offer for hydrogen bonding as opposed to the initial 2SA

base. The result is Asp76 preferring non-covalent bonding with nearby Arg74, leaving the base pair without a direct link to the protein.

We again sampled 100 different final conformations for both transformation directions (T=A ⇆ C≡G) and measured the distances between proximate chemical groups of fully coupled 2S bases and the side chain of Asp76, see Fig. 7.18. The initial 2SA base interacts with its environment fully during the final $\lambda = 1$

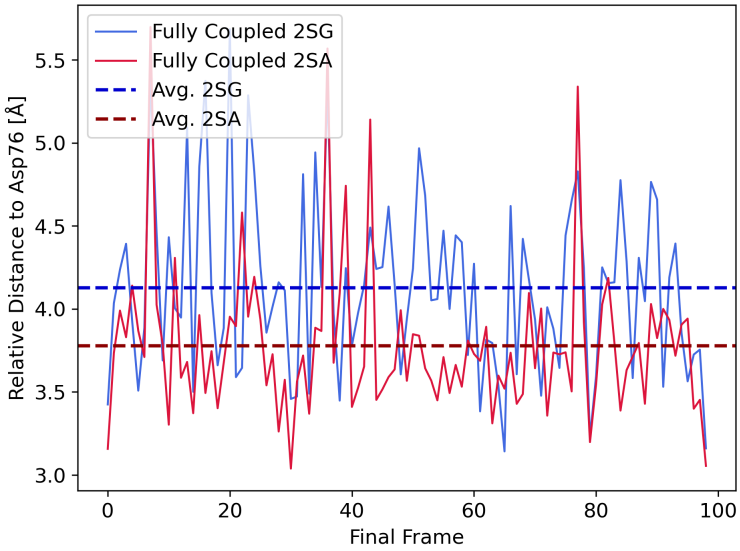


Figure 7.18: Mutation T4C – interactions between 2S bases and nearby Asp76. Reference point is represented by the carbon of $-COO^-$ side chain. For 2SA the distance is measured to the nitrogen of its $-NH_2$, for 2SG to its =O. Keep in mind that these are relative distances measured in the endpoint conformations ($\lambda = 1$) with bases fully coupled to the environment.

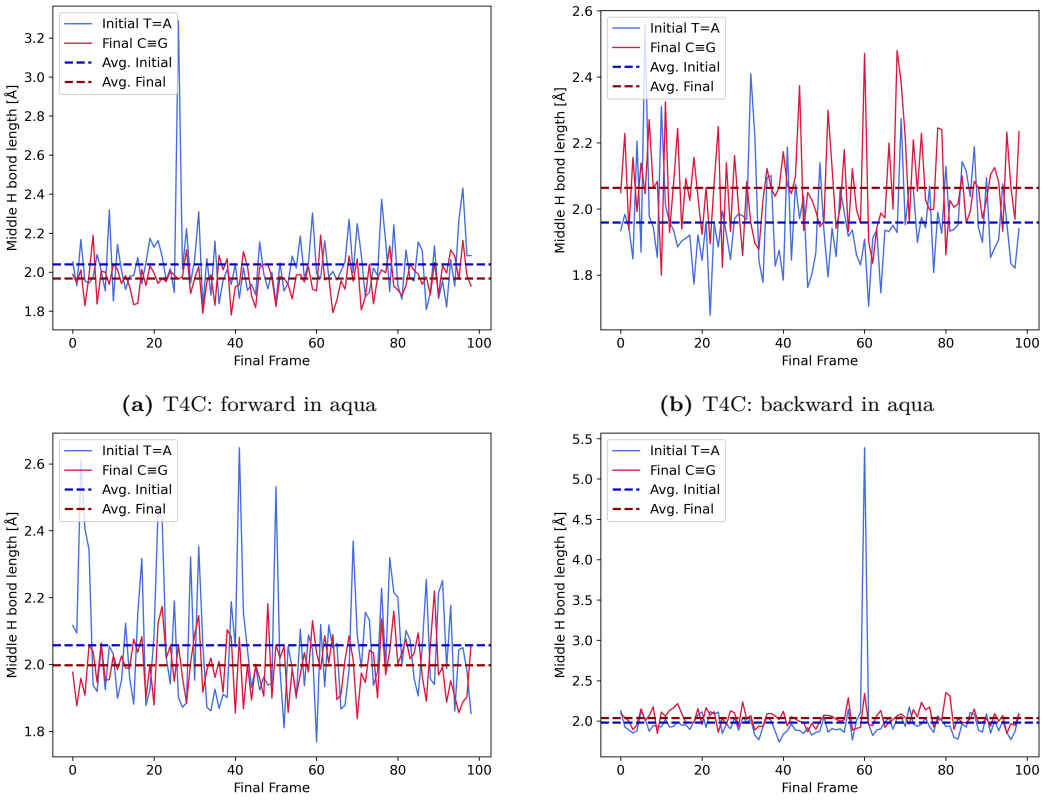
window of the backward run. The sample average distance between the nitrogen of its $-NH_2$ group and the carbon of the $-COO^-$ side chain is (3.7 ± 0.4) Å. Final conformations of the forward runs have the target 2SG base fully coupled, with the sample average distance (4.1 ± 0.5) Å between 2SG =O group and the $-COO^-$ carbon of Asp76.

Looking at the distances in Fig. 7.18 one can spot a hint of repulsion between target 2SG base and Asp76 in a form of a 0.4 Å gap among the averages marked by dashed lines. This is a proof of preference towards the original base pair as opposed to the mutation performed.

Visual inspection of the sampled interaction sites, recall snapshots in Fig. 7.15, confirms that 2SA indeed tends to form hydrogen bond with the carboxylic side chain via the proximate $-NH_2$ group. The average hydrogen bond length is approx. 3.3 Å, though individual sampled values vary greatly. This is due to switching between different donor and acceptor atoms of these groups (recall Fig. 7.8), which makes it rather complicated to capture with reasonable STD. Nearby Arg74 competes for the hydrogen bonding with Asp76. Similarly as in mutation T4A, when the mutation to 2SG base is performed Asp76 side chain commits to Arg74 all the more, since there is no proximate positive partial charge offered by the 2SG base.

Examination of base pair stability via central hydrogen bonding, Fig. 7.19, reveals that almost none of the pairs ended up disconnected. The only case of a

broken base pair was recorded for an initial T=A pair during a backward transformation with the double helix bound to the transcription factor, see the single spike of approx. 5.5 Å in Fig. 7.19d. Another outlier, close to being disconnected, was a case of another initial T=A pair separated by approx. 3.5 Å during a forward transformation without the protein's presence, see Fig. 7.19a. Forward structures in protein feature T=A pairs with (2.0±0.2) Å central hydrogen bonds, identical results are captured for both base pairs during backward runs without the protein's presence. Remaining cases give hydrogen bonding of (2.0 ± 0.1) Å.



(a) T4C: forward in aqua

(b) T4C: backward in aqua

(c) T4C: forward in protein

(d) T4C: backward in protein

Figure 7.19: Mutation T4C – central hydrogen bonds between DNA bases. Forward runs (a) show single outlier initial T=A pair, separated by approx. 3.5 Å. The following backward runs (b) features no outliers at all. Forward runs (c) feature well connected pairs. Backward runs (d) remain well connected apart from a single broken initial T=A pair separated by ~ 5.5 Å.

7.6.3 Mutation T4G

Last possible mutation in this site is T=A $\rightarrow$ G≡C, cf. structures in Fig. 7.15. This is a special case of mutation which is completely non-disruptive towards the binding affinity of Zif268-DNA, recall T4G in Fig. 7.9. The fact that this particular mutation has negligible effects on the binding affinity of the complex can be seen in our computed relative distances, plotted in Fig. 7.20.

We again sampled 100 different final conformations for both transformation directions (T=A $\leftrightarrows$ G≡C) and measured the distances between proximate chemical groups of fully coupled 2S bases and the side chain of Asp76. The reference point is again the carbon of the $-\mathrm{COO}^-$ side chain. Initial 2SA base interacts with its environment fully during the final $\lambda = 1$ window of the backward run.

The sample average distance to the nitrogen of its $-NH_2$ group is (3.8 ± 0.4) Å. Final conformations of the forward runs have the target 2SC base fully coupled, and the sample average distance to its nitrogen of $-NH_2$ group is (3.5 ± 0.2) Å

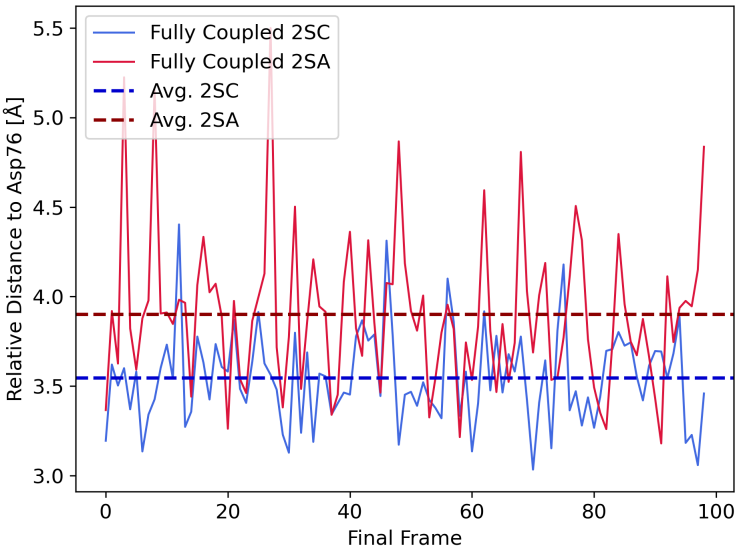
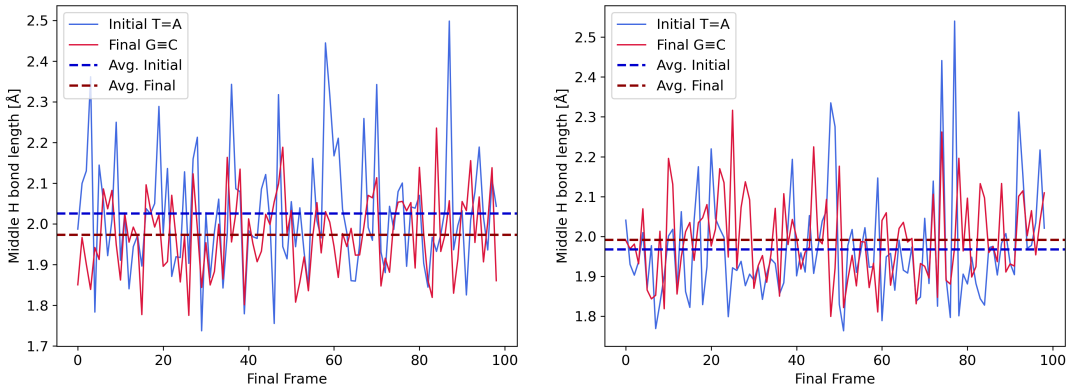


Figure 7.20: Mutation T4G – interactions between 2S bases and nearby Asp76. Reference point is represented by the carbon of $-COO^-$ side chain. For both 2SA and 2SC the distance is measured to the nitrogen of their $-NH_2$ exposed to the side chain. Keep in mind that these are relative distances measured in the endpoint conformations ($\lambda = 1$) with bases fully coupled to the environment.

Comparison of the averages in Fig. 7.18 reveals an *attraction* gap of 0.3 Å. This shows that unlike all other mutations performed in this site, T4G creates target 2SC base which is 'accepted' by the Zif268 transcription factor. The average hydrogen bond length between the bases and Asp76 is approx. 3.3 and 2.9 Å for 2SA and 2SC respectively. Again, individual sampled values vary greatly, which is due to switching (recall Fig. 7.8) between different donor and acceptor atoms of these groups. This makes it rather complicated to capture the average hydrogen bond lengths with reasonable STDs.

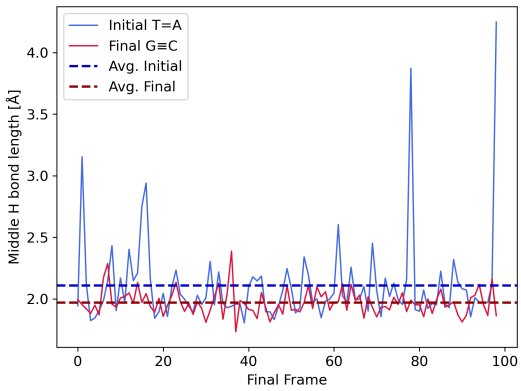
Visual inspection of the site before and after the mutation has occurred, recall snapshots in Fig. 7.15, reveals that both 2S bases expose their $-NH_2$ group to nearby Asp76, leaving it available for hydrogen bonding with the carboxyl side chain. Final 2SC base simply fits in well. Any other mutation in this site (T4A, T4C) produced final 2S base which repelled the carboxyl side chain, making it instead preferable for Asp76 to bind with Arg74 only.

Pairing of bases features defects only during backward transformations of the double helix bound to the ZF protein, see Fig. 7.21. In that case, we encounter a disconnection rate of 9 % for initial T=A pairs, separated by up to 8.5 Å distance. Preceding forward transformations capture 2 of those defects forming, see spikes around 80th and 100th conformation. These 2 initial base pairs did not fully separate (distance around 4.0 Å) during the forward run, but were driven further away by the following backward transformation. Without the protein's presence the sample average central hydrogen bond lengths between nucleobases are in all cases approx. (2.0 ± 0.2) Å. Systems with the protein present feature hydrogen bonds of (2.0 ± 0.1) Å, defective pairs were excluded out.

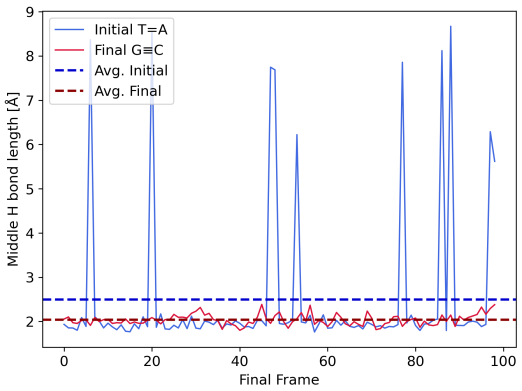


(a) T4G: forward in aqua

(b) T4G: backward in aqua



(c) T4G: forward in protein



(d) T4G: backward in protein

Figure 7.21: Mutation T4G – central hydrogen bonds between DNA bases. All pairs of both forward (a) and backward (b) transformations in vacuo remain well connected. Forward transformations in protein (c) feature 2 initial T=A pairs close to being disconnected with around 4.0 Å separation, backward runs in (d) drove a total of 9 initial T=A pairs up to 8.5 Å apart.

7.6.4 Summary

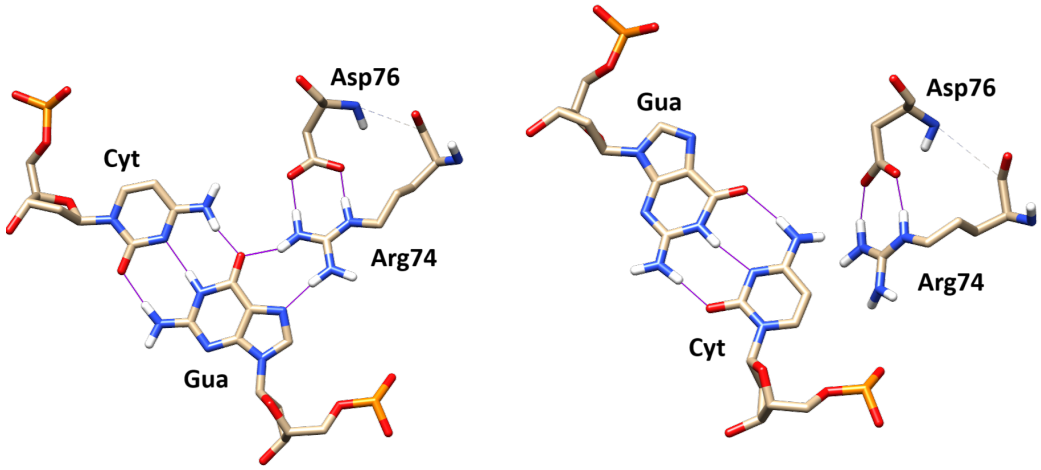
To summarize the analysis of mutation site 4, in Tab. 7.1 we provide the reader with all the sampled average values of distances between the $-COO^-$ carbon of proximate Asp76 and heavy atoms of DNA bases exposed to it. This shall facilitate the comparison between performed base pair mutations

Table 7.1: Mutation site 4 – sample average distances between $-COO^-$ carbon atom of Asp76 and heavy atoms of donors of 2S DNA bases exposed to it. Distances are measured in final sampled conformations of both FEP transformation directions (forward and backward). Dimmed (gray) values correspond to the structures with the bases decoupled from their environment. Values are given in Å with the appropriate sample STD.

Mut.	Base	Dist.	Forward	Backward
T4A	2SA	N–C	4.6 ± 0.7	3.6 ± 0.8
	2ST	O–C	4.3 ± 0.4	4.1 ± 0.7
T4C	2SA	N–C	4.1 ± 0.5	3.7 ± 0.4
	2SG	O–C	4.1 ± 0.5	4.0 ± 0.5
T4G	2SA	N–C	4.2 ± 0.3	3.8 ± 0.4
	2SC	N–C	3.5 ± 0.2	3.4 ± 0.5

7.7 Mutation Site 3 – G3C

This position hosts side chains of Arg74 and Asp76 in a similar manner as was observed in mutation site 4, see Fig. 7.22. The only difference here is that the roles



(a) Original G≡C **(b)** G3C: C≡G

Figure 7.22: Detail of mutation site 3 – recall Fig. 7.2. For clarity, the remaining parts of the DNA double helix and the protein are hidden. (a) shows the original G≡C base pair, (b) features the final C≡G pair. Note that Asp76 resides one level below the present site, i.e. on the level of the 4th position in the DNA double helix. It influences the base pairs only indirectly by pulling on its closest neighbour Arg74. Purple lines mark hydrogen bonding.

of amino acids have switched, cf. Fig. 7.15. Arg74 is now the one in direct contact with the 1S base while Asp76 resides one level below, influencing the mutation site only indirectly. We performed mutation G3C (G≡C → C≡G) which resulted in a moderate disruption towards the binding affinity of the complex, see Fig. 7.9.

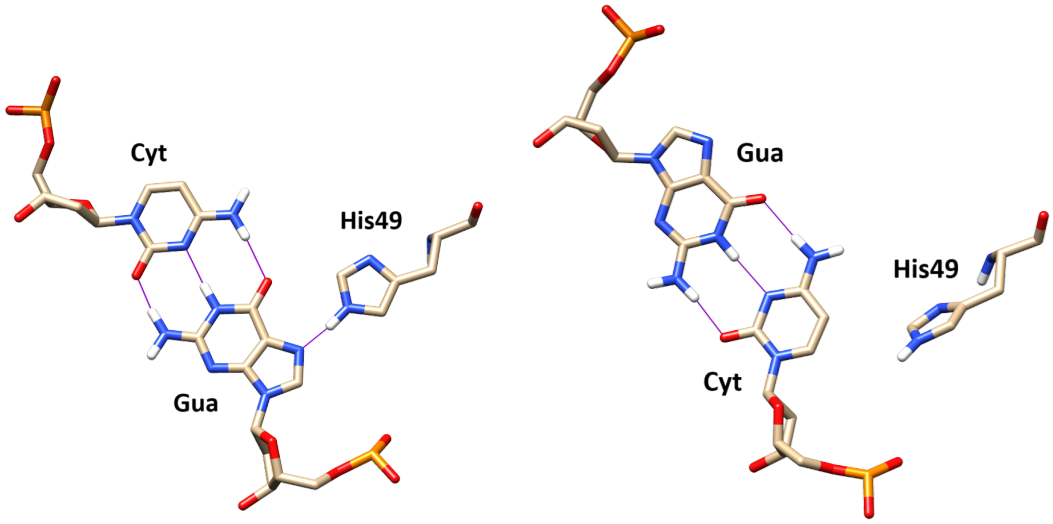
The original 1SG base interacts with proximate Arg74 via formation of 2 hydrogen bonds. 1SG =O group links with the closest $-NH_2$ hydrogen of guanidinium side chain via an average hydrogen bond of 3.3 Å, second hydrogen bonding (connecting 1SG nitrogen with side chain's hydrogen) is done on average at a relative distance of 3.1 Å. Asp76 competes for the hydrogen bonding and pulls Arg74 a little below the level of the mutation site. The distance between their donor and acceptor atoms is on average 2.5 Å, in value comparable to our typical hydrogen bonding between DNA base pairs.

In contrary, target 1SC base offers no acceptor atoms exposed to Arg74 to offer for hydrogen bonding. The result is the guanidinium group tilting away in favor of non-covalent bonding with nearby Asp76 only, see snapshot in Fig. 7.22b. This leaves the target base pair without a direct link to the protein, lowering binding affinity of the protein complex. After the mutation the distances between amino acids are on average approx. 2.4 Å, which presents de facto no change at all keeping in mind our typical sample STD for hydrogen bonds.

We report no base pair disconnections in all of our final conformations sampled, no matter the protein's presence. In water without the transcription factor average central hydrogen bond lengths are always approx. (2.0 ± 0.2) Å, when protein is introduced to the system central hydrogen bonding reads (2.0 ± 0.1) Å.

7.8 Mutation Site 5 – G5C

Mutation site 5 has a single amino acid (His49) side chain close enough to affect the binding of ZF protein to the given DNA sequence, see Fig. 7.23. This is the



(a) Original G≡C **(b)** G5C: C≡G

Figure 7.23: Detail of mutation site 5 – recall Fig. 7.2. For clarity, the remaining parts of the DNA double helix and the protein are hidden. Picture (a) shows the original G≡C pair with His49 aligned in the plane of the base pair, (b) features the final C≡G pair forcing the side chain to tilt out of the plane. The tilt was observed to be always between 60° and 80° with respect to the base pair plane. Purple lines mark hydrogen bonding.

only case of a binding site in Zif268 transcription factor having a different amino acid other than Asp or Arg, cf. Fig. 7.2.

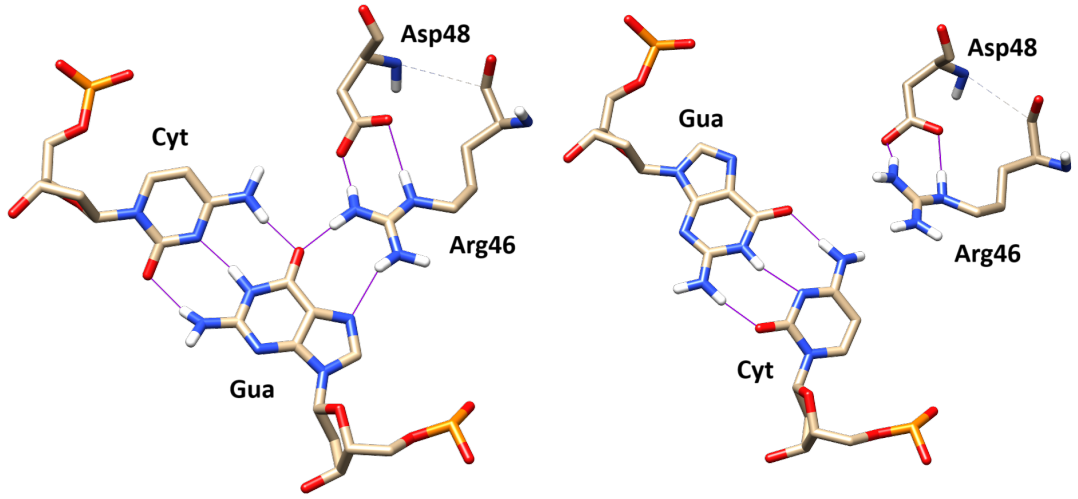
Visually we observed that His49 tends to be more aligned in the plane of the original G≡C pair, hydrogen bonding to the exposed 1SG nitrogen acceptor via its proximate hydrogen atom (see Fig. 7.23a). The sample average value of this non-covalent link is approx. 2.3 Å. Conformations with the original G≡C pair fully coupled to the environment display a preference of His49 to stay oriented towards the 1SG base with out-of-plane angle no more than 30°.

After mutation G5C (G≡C → C≡G) His49 ends up tilted away from the 1SC base in all 100 sampled conformations with the target C≡G pair fully coupled to its environment, see snapshot in Fig. 7.23b. Unlike for original 1SG base, there is no way for the side chain to hydrogen bond with mutant 1SC base. As a result the DNA duplex looses a direct link to the ZF protein in this binding site, and His49 is left to wiggle around freely in thermal fluctuations. These effects explain the moderate decrease in binding affinity of the Zif268-DNA complex due to mutation G5C, observed in Fig. 7.9.

Both of the base pairs are stable in all of the sampled conformations whether the ZF protein is present or not, no matter their coupling to the environment. In water without the protein average central hydrogen bond lengths are always approx. (2.0 ± 0.2) Å. Inside the whole protein complex the central hydrogen bonding reads again (2.0 ± 0.1) Å.

7.9 Mutation Site 6 – G6C

Mutation site 6 represents a similar environment to that of mutation site 3. Here the directly interacting amino acid is Arg46 while Asp48 competes for its hydrogen bonding from one level below the site, see Fig. 7.24. Keep in mind that this site is



(a) Original G≡C **(b)** G6C: C≡G

Figure 7.24: Detail of mutation site 6 – recall Fig. 7.2. For clarity, the remaining parts of the DNA double helix and the protein are hidden. (a) shows the original G≡C base pair, (b) features the final C≡G pair. Note that Asp48 resides one level below the present site, i.e. on the level of the 7th position in the DNA double helix. It influences the base pairs only indirectly by pulling on its closest neighbour Arg46. Purple lines mark hydrogen bonding.

inside the most amino-acid-rich region of all the binding sites of Zif268; cf. Fig. 7.2, base pair position 7 and its surroundings. This will become more apparent once we reach the following mutation site.

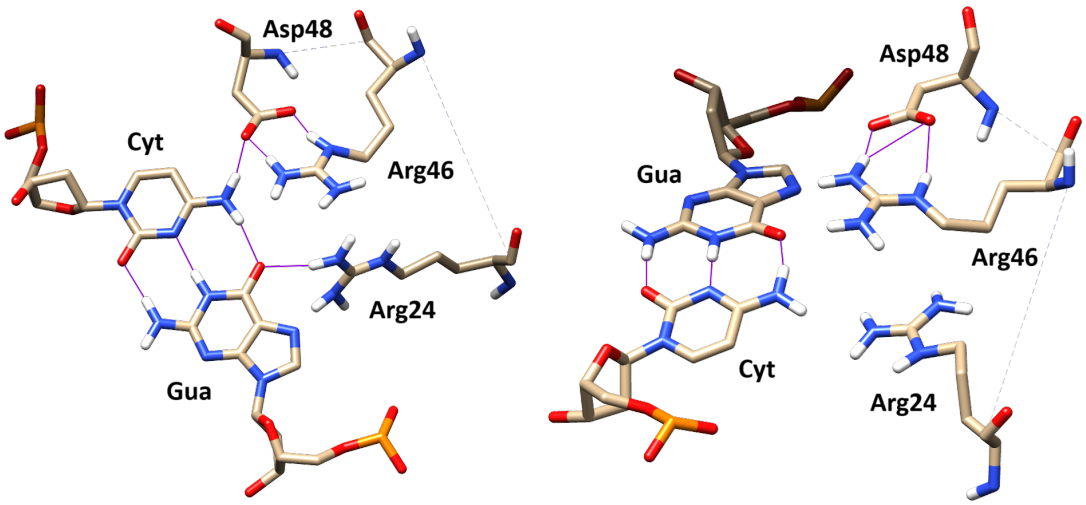
Same as in mutation site 3, Arg46 links to initial 1SG base via 2 hydrogen bonds, see structure in Fig. 7.24a. Hydrogen bond to 1SG oxygen is on average 3.1 Å long while the second one towards the nitrogen acceptor reads an average of 3.4 Å. Asp48 again pulls the guanidinium group with 2 non-covalent bonds of an average length of 2.1 Å, this time a little below the level of the base pair. These salt bridges are in value comparable to our typical Watson-Crick hydrogen bonding between paired DNA bases.

When mutation G6C is performed (switching G≡C → C≡G) Arg46 is tilted away above the plane of the final base pair while its bonding to Asp48 is preserved with the sample average relative distances between the donor and acceptor atoms unchanged, see snapshot in Fig. 7.24b. This created the largest decrease in binding affinity of the Zif268-DNA complex among all of the performed mutations, recall Fig. 7.9. Such a major disruption may be due to amino-acid-rich nature of the protein region this mutation site is part of. The intricate network of amino acids surrounding it is most likely behind its heightened sensitivity to changes, since this is the only difference we observe as compared to otherwise very similar mutation site 3, cf. Fig. 7.22. Note that the very same mutation (G3C) performed there led to only a moderate disruption, in value roughly 1/3 that of the present mutation G6C.

Concerning pairing of nucleobases, none of the sampled cases ended up disconnected. Same as in previous sites, the average central hydrogen bond lengths in water without the protein's presence are (2.0 ± 0.2) Å. When the transcription factor is bound to the double helix, central hydrogen bonds read (2.0 ± 0.2) Å for the initial G≡C pair and (2.1 ± 0.3) Å for the final C≡G pair fully coupled to the environment, featuring a wider spread of measured distances as compared to all previous mutations analyzed.

7.10 Mutation Site 7 – G7C

Mutation site 7 is the most complicated environment recorded in Zif268 transcription factor. It hosts Arg24 directly interacting with 1S bases, Asp48 directly interacting with 2S bases, and Arg46 influencing both of these amino acids from one level above the site, see Fig. 7.25. Arg46 can compete for hydrogen bonding



(a) Original G≡C **(b)** G7C: C≡G

Figure 7.25: Detail of mutation site 7 – recall Fig. 7.2. For clarity, the remaining parts of the DNA double helix and the protein are hidden. (a) shows the original G≡C base pair, (b) features the final C≡G pair. Note that Arg46 resides one level above the present site, i.e. on the level of the 6th position in the DNA double helix. It influences the base pairs only indirectly by pulling on its closest neighbour Asp48. Purple lines mark hydrogen bonding.

with Asp48 while repelling nearby Arg24. Note that Arg46 and Asp48 are also shared between this site and the site above (i.e. mutation site 6), see Fig. 7.24. Changes in one mutation site could thus influence the other more easily.

Initial G≡C pair is held by 2 hydrogen bonds formed with nearby amino acids. 1SG base shares hydrogen bonds with nearby Arg24. In most of the cases, only a single hydrogen bond is formed – either with the oxygen or nitrogen acceptor atom of 1SG, see structure in Fig. 7.25a. The length of such bonds is on average approx. 3.3 Å. The reason for the second bond not to be formed lies behind the presence of Arg46, residing one level above the mutation site. Guanidinium groups are known to possess special stacking properties despite their positive charge [121], the so-called *Arginine Magic*. The like-charge ion pairing forces Arg24 closer to Arg46, the result of which is breaking of one of the hydrogen bonds with initial 1SG base. This is demonstrated in Fig. 7.25a where Arg24 is tilted towards Arg46. Similar patterns of attraction are observed all through our sampled structures. Initial 2SC base pulls in Asp48 via hydrogen bonding between its $-NH_2$ hydrogen and the closest oxygen of the carboxyl side chain. The average distance between them is approx. 2.9 Å. Similarly as before, Arg46 competes for the hydrogen bonding with Asp48, the average relative distance between the donor and acceptor atoms being approx. 2.6 Å.

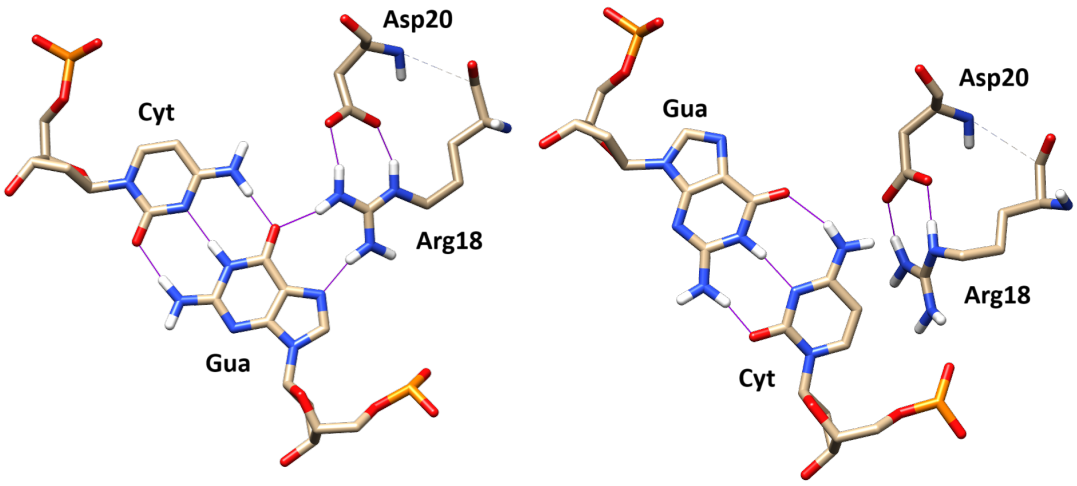
After the mutation G7C, again switching G≡C → C≡G, all hydrogen bonds between nucleobases and proximate amino acids are broken, see Fig. 7.25b. Asp48

fully commits to Arg46 with hydrogen bonds of approx. 2.5 Å on average. Arg24 is floating little below the mutation site, repelled by the $-NH_2$ hydrogen atoms of the final 1SC base. The resulting relative free energy difference (recall Fig. 7.9) reports highly disruptive nature of mutation G7C, which can be assigned to the complicated network of amino acids surrounding the mutation site. The amino-acid-rich character of this region is most likely behind the increased sensitivity to mutations. Furthermore, changes made in this site can readily propagate to mutation site 6 (Fig. 7.24) due to their sharing of proximate amino acid residues, rendering them both susceptible to base pair alterations.

Pairing of nucleobases is overall stable with no disconnections detected in any of our sampled conformations. The average central hydrogen bond lengths in water without the protein are reported to be (2.0 ± 0.2) Å. Inside the protein pairing of bases depends on the specific case. For the initial G≡C pair the central hydrogen bond is always (2.0±0.2) Å. After the mutation fully coupled final C≡G pair is linked via an average (2.2 ± 0.2) Å long hydrogen bond, fully decoupled reads (2.0 ± 0.2) Å.

7.11 Mutation Site 9 – G9C

Mutation site 9 is the last one explored in our study. Its composition is exactly the same as that of mutation site 3, cf. Fig. 7.22 and Fig. 7.26. Arg18 directly



(a) Original G≡C **(b)** G9C: C≡G

Figure 7.26: Detail of mutation site 9 – recall Fig. 7.2. For clarity, the remaining parts of the DNA double helix and the protein are hidden. (a) shows the original G≡C base pair, (b) features the final C≡G pair. Note that Asp20 resides one level below the present site, i.e. on the level of the 10th position in the DNA double helix. It influences the base pairs only indirectly by pulling on its closest neighbour Arg18. Purple lines mark hydrogen bonding.

interacts with 1S bases while Asp20 lies one level below this site. In this sense Asp20 competes for hydrogen bonding with Arg18, indirectly influencing the mutation site at hand.

Same as in mutation site 3 we observe 2 hydrogen bonds between the original 1SG base and the guanidinium side chain, see structure in Fig. 7.26a. Each of these bonds are on average approx. 3.2 Å. Asp20 pulls Arg18 little below the plane of the present site. The links between their donor and acceptor atoms are on average around 2.7 Å, which is in value comparable to the typical Watson-Crick hydrogen bonding we observe between paired DNA bases.

After the mutation G9C (i.e. G≡C → C≡G) the presence of 1SC base causes Arg18 to tilt away from the base pair plane, attached to Asp20 via hydrogen bonds of approx. 2.6 Å, see Fig. 7.26b. 1SC base simply repels the positively charged guanidinium side chain via its $-NH_2$ hydrogens ($+\delta$). This results in a moderate decrease in binding free energy of the Zif268-DNA complex, recall Fig. 7.9. The disruptivity of the present mutation G9C is comparable to that of mutation G3C. Both of them represent the same mutation in exactly the same amino-acid environment, the only difference being the position in the DNA duplex (see Fig. 7.2). This further confirms the consistency of our MD simulations.

Base pair stability remains the same as in previous cases, no disconnections reported in any of the sampled conformations. In water without the protein we encounter average central hydrogen bonds of (2.0 ± 0.2) Å, systems with the ZF protein present give an average of (2.0 ± 0.1) Å.

7.12 Discussion

7.12.1 Stability of Mutated Base Pairs

The sample average value of the central hydrogen bond length is in almost all cases approx. 2.0 Å with STD ranging from 0.1 to 0.3 Å, no matter the base pair and its coupling to the environment. This applies to both initial equilibrations and FEP transformations. The only 2 instances of slightly different average hydrogen bonding are mutations G6C (Fig. 7.24) and G7C (Fig. 7.25) with (2.1 ± 0.3) Å and (2.2 ± 0.2) Å respectively. Both of these mutations happen inside the most side-chain rich region of Zif268; cf. Fig. 7.2, positions 6 and 7. According to experimental data provided in Ref. [122, 123] the distance between the donor and acceptor atoms in the hydrogen bond between paired DNA bases typically falls within a range of 2.8 to 3.4 Å. In our MD simulations we measure distances between the hydrogen and acceptor atoms. Reference experiments do not detect hydroges. The mutual distances they report are thus measured between the heavy atom of the donor and the heavy acceptor, rendering the values inherently longer by length of the omitted covalent bonds.

Naturally, hydrogen bond lengths in DNA can vary due to factors such as base pair stacking, base pair mismatches, and local structural variations. Additionally, different studies and experimental conditions may report slightly different values within the general range mentioned above. Because our system works with a DNA double helix of just 11 pairs, and not with the whole stable DNA, it is reasonable to assume that this may also have an effect on base pair stacking – ergo our hydrogen bond lengths could be affected.

Overall, base pairs of our systems are stable. There are a few cases in which some of the simulations resulted in disconnected pairs. Breaking of base pairs happens no matter the protein's presence and is an issue of A=T pairs only. The fact that we never see this happen to any C≡G pair can be assigned to their stronger connection through 3 hydrogen bonds compared to the A=T pairs which are formed via 2 bonds only. The latter thus require less energy to break. Typically the disconnection rate ranges between 2 and 3 % (e.g mutation C2A in Fig. 7.11) and the separation lies within 5 to 10 Å, measured between the central donor and acceptor atoms. The typical 2 to 3 % is regarded by us to be acceptable. Additionally we report 2 specific instances (T4A, T4G) in which the rate is as high as 10 % – cf. Fig. 7.17 and 7.21. Mutation site 4 seems to be no special place among binding sites of Zif268, see Fig. 7.2. It features the 2 most common amino acid side chains, Arg and Asp. On top of that, mutation T4A poses only a moderate disruption towards the binding affinity of the complex and T4G does not influence the complex at all, recall Fig. 7.9. These facts are strong indicators that the observed disconnections have nothing to do with the disruptivity of the mutations. They are a mere statistical matter of our simulations getting stuck in conformations where Ade and Thy are too separated to link. Though 10 % is far from an ideal rate, cutting these disconnections out of the sampled ensembles leaves still a reasonable statistics to work with.

7.12.2 Influence of Proximate Amino Acids

Generally, the presence of amino acids does not appear to influence hydrogen bonding between paired DNA bases, since the sample average value of the central hydrogen bond length between bases does not change with their coupling to the environment. Visible from spreads in distances recorded, cf. Fig. 7.18 and 7.19, interactions between amino acids and nucleobases tend to fluctuate more compared to nucleobase pairing. When a DNA base interacts with some amino acid attractively, the hydrogen bonds tend to switch the participating donor and acceptor atoms. An example of a typical time evolution of these bonds, sampled through one of the equilibration MD runs is given in Fig. 7.8. This does not happen in Watson-Crick pairing of DNA bases. The switching makes it rather complicated to capture mutual distances with reasonable STD using donor and acceptor atoms directly. This made our decision to measure the distances between amino acid side chains and proximate chemical groups of DNA bases through heavy atoms (O, N, C), e.g. Fig. 7.18. Indeed we were readily able to capture the lengths with acceptable STDs, see values listed in Tab. 7.1 in the case of mutations inside position 4.

Individual mutation sites can host multiple amino acids. Their influence over the binding site depends on how far away from the DNA bases they are. Side chains close enough to form hydrogen bonds with the bases (up to ~ 4 Å) interact directly with the binding site. More distant ones tend to interact with the proximate one, influencing the binding site only indirectly. The more side chains are present, the more sensitive to changes the binding site tends to be. The most amino-acid-rich binding region of Zif268 contains sites 6 and 7, cf. Fig. 7.2. Mutations inside these 2 binding sites resulted in the largest free energy differences (G6C with 9.8 kcal/mol and G7C with 5.6 kcal/mol, see Fig. 7.9).

Binding sites are also interlinked. A lot of the positions share amino acid residues, e.g. structures in Fig. 7.24 and 7.25 sharing Arg46 and Asp48. When a mutation happens and the proximate amino acid gets repelled by the target DNA base, its preference for hydrogen bonding is shifted towards the neighbouring side chain instead. If this neighbour resides on a level above/below the site, such mutation can possibly propagate its effects onto the surrounding sites. Change in one of the sites can facilitate changes in others also through base pair stacking. Base pairs can wobble up and down, left to right under the influence of nearby amino acids and in thermal fluctuations. These changes can propagate through the whole structure by pushing and pulling, effects of which will fade away with increasing distance from the epicenter. The effects we observe and study are thus mere fragments of a complex interplay of amino acid regions and DNA bases, possibly playing out on multiple levels at the same time. Such issues are a matter of debate and a more sophisticated analysis is needed.

7.12.3 DNA Sequence Detection by Zif268

We were able to capture potential hints to the sequence detection mechanism of Zif268 transcription factor. Individual binding sites of the ZF protein host positively (Arg, His) or negatively (Asp) charged side chains and their combinations, see Fig. 7.2. Patterns repeating throughout all of the performed mutations are exemplified and thoroughly investigated in mutation site 4, Section 7.6. This

site has been chosen specifically for its environment featuring Arg and Asp, typical for most of the protein's binding sites, and also for the fact that one of the mutations performed here (T4G) caused de facto zero change in the binding free energy of the complex, cf. Fig. 7.9.

Interactions between mutated bases and nearby amino acids are carried out based on charges at play. DNA bases tend to interact with charged AA side chains via partially charged groups, $-NH_2$ ($+\delta$ at H) or $=O$ ($-\delta$), in their immediate vicinity. As expected, if the signs of both species differ the interaction is attractive, and vice versa. Examples of such interactions are explored in Fig. 7.16, 7.18 and 7.20. It is important to note that interactions with initial pairs have always been reported by us to be attractive. Since the DNA double helix we use as initial structure before the mutations is actually the one exactly recognized by the Zif268, it leads us to suspicion that these attractive treatments by AA side chains are precisely the way by which the transcription factor recognizes the correct DNA sequence. Repulsive effects should thus hurt the consequent binding affinity of the complex. Based on our observations we can safely say that if such $\pm\delta$ group of a DNA base is on the opposite side to where the charged AA side chain is, there is no such repulsion or attraction. The distance is simply enough for the base not to get influenced by the presence of this charge.

Arg tends to form 2 hydrogen bonds with Gua, Asp prefers Cyt and Ade to which it binds via a single hydrogen bond. Examples of these repeating patterns can be seen in Fig. 7.15. In all of the performed mutations we did not report any attractive treatments towards Thy. This might simply be due to lack of cases that would allow for such interaction to happen. Even though Zif268 is a transcription factor recognizing sequence GCG**T**GGGCG, Thy in 4th position is preferred indirectly via detection (Asp76) of Ade at the complementary strand. Since there is in principle no reason for Thy not to be attached to these amino acid side chains, we do not rule out the existence of a binding site with Thy as a preference for a given position in DNA sequence. Based on our observations, a criterion for such detection to happen would be one of the $=O$ groups of Thy being exposed to a guanidinium side chain (or similar species featuring $-NH_2$) in its immediate vicinity. In this way Thy would be added to the collection of preferences for amino acids such as Arg.

These preferences may also depend on the overall shape of the given protein. Zif268 wraps around the DNA double helix such that these AA side chains are exposed to the nucleobases in a particular way. We oriented all of the structure snapshots of sites in an aligned way such that they can be easily comparable. From the perspective of our alignment, 1S bases are always at the bottom of the snapshots which are taken from the top of the helix down, i.e. with accordance to our site ordering (Fig. 7.2). We then observe all of the amino acid side chains reside to the right of any base pair. If a different protein were to interact with this DNA segment, exposing the same side chains from the opposite side, the preferences of amino acids towards the bases would have to change accordingly. This can be seen by comparison of exposed base chemical groups ($-NH_2$ vs. $=O$) to the proximate AA side chains, cf. Fig. 7.15. Sequence detection thus appears to be a complex mechanism involving higher order protein structures.

7.12.4 Impact of Mutations on Binding Free Energies

When a base pair mutation happens the free energy content of the protein-DNA complex may change. Whether a mutation is disruptive towards the binding affinity of the complex is determined by the immediate surroundings of the mutated base pair. Since the sequence detection mechanism involves attraction between amino acid side chains and specific chemical groups of DNA bases directly exposed to them, the disruptivity of mutations is given by their mutual repulsion (or simply by the lack of attraction).

Typical binding sites of Zif268 involve a combination of a single Arg and a single Asp side chain (cf. Fig. 7.2, positions 3, 4, 6, and 9) and a vast majority of the mutations performed in these sites resulted in a moderate binding affinity disruption ranging between 2.8 and 3.5 kcal/mol, recall Fig. 7.9. These values are also confirmed by reference simulations [115] and experiments [120]. Looking at the local structures of each of the mutation sites, such disruptions correspond to at least one final (target) nucleobase repelling the proximate AA side chain which was previously linked with the original base through hydrogen bonding. If at least one target nucleobase forms a hydrogen bond with the amino acid in site, the mutation process should be a non-disruptive one towards the binding affinity of the complex. An example of such a case is mutation T4G, cf. Fig. 7.9. The net zero change in binding free energy is confirmed on all fronts of our references [115, 120]. The final G≡C pair fits in the binding site well, see structures in Fig. 7.15. Both 2S bases (before and after mutation T4G) expose their $-NH_2$ group to the nearby Asp side chain, forming a hydrogen bond. All other mutations in this site (T4A, T4C) repulse Asp such that it prefers binding to its neighbouring Arg side chain only. Since there is virtually no free energy difference, mutations the likes of T4G pose a danger of high off-target binding of the Zif268 protein.

Mutations highly disruptive towards the binding affinity take place inside the most amino-acid-rich regions. The largest decrease in binding affinity was recorded during mutation G6C with the value of 9.8 kcal/mol, see Fig. 7.9. Results given by reference simulations [115] are slightly lower, though still representing a major disruption. Experiments [120] give an undeclared decrease in a range going as high as 10 kcal/mol. Although position 6 hosts the typical Arg and Asp, see Fig. 7.24, the site is localized in the most side-chain-rich region of Zif268, cf. Fig. 7.2. This is most probably the reason behind its heightened sensitivity to mutations. The Arg46 and Asp48 residues are also shared between this site and the following position 7, see Fig. 7.25. Site 7 is the most complicated binding environment of Zif268; featuring Arg24, Arg46, and Asp48. Mutation G7C scores the second highest disruption, 5.6 kcal/mol, of all mutations we performed. Similar values are reported by Ref. [115, 120].

Whenever there is a site which does not directly interact with any amino acids, its mutations have only a small effect on the binding affinity of the complex. This is demonstrated in Fig. 7.9, cf. structures in Fig. 7.10 and 7.14. Corresponding mutations in both of the sites give very similar $\Delta\Delta G$, with 4 out of 6 mutations (C2A, C2T, C8A, C8T) yielding values around 1 kcal/mol or less. These binding affinity differences are in value comparable to the standard chemical accuracy. Mutations C2T and C8T both give $\Delta\Delta G < 0$, which is in contradiction with our references [115, 120]. Negative values mean that after these mutations the complex gets more stable. Even though this is the case, magnitudes of their dif-

ferences still lie within the chemical accuracy of ±1 kcal/mol. Since there are no amino acids to directly interact with, no strong preference of base pairs should be present. It can be that the differences in values we observe may fluctuate from case to case based on events happening in sites close by. In order to resolve this question, larger statistics of mutations in these sites is needed. Nevertheless, Zif268 should still with high-enough efficiency be able to recognize DNA sequence altered through any of these mutations. The remaining 2 (C2G and C8G) give approx. 2.2 and 3.5 kcal/mol respectively. These values significantly differ from most of the references, including simulations [115] and experiments [120]. Since we did not observe any major structural deficiencies, such disparity could possibly be a result of insufficiently sampled ensemble of conformations. A further investigation is needed.

Conclusion

In the introductory chapter, the diploma thesis introduces the reader to the most used tools for editing of the human genome (ZFN, TALEN, CRISPR). The following chapters describe in detail the methodology of MD simulations and calculations of hydration and binding free energy.

Scripts were prepared that made it possible to perform a large number of hydration and binding free energy calculations in the environment of the supercomputing MetaCentrum. In order to be able to perform the calculations in a massively parallel way, an approach was chosen where a large number of short non-equilibrium FEP MD runs are produced in parallel. From these, a large number of work values are obtained, from which the value of the equilibrium hydration or binding free energy is then determined using the CGI method.

The chosen methodology was first tested in the calculations of hydration free energies of nucleic acid components and amino acid side chains, as a large amount of historical reference data was available.

In the key results chapter, the complex of the transcription factor Zif268 and a short DNA double helix was investigated, in which mutations of individual base pairs were carried out. At the same time, calculations of changes in binding free energy were performed. A comparison with the literature showed that using the NAMD software package, the algorithms implemented in it and the chosen methodology for calculating binding free energies using non-equilibrium MD simulations, results comparable to those that can be obtained using other established software packages (Gromacs, AMBER) can be obtained. In the future, it will be interesting to debug our methodology also for the alternative value of the alchDecouple parameter — *off*. In that case, the mutual interaction of the mutated bases will be also scaled during alchemical transformations. This will allow to remove potential artifacts that can occur with the alchDecouple *on* option, where mutated bases that are decoupled (as regards non-covalent interactions) from the rest of the simulated system, however, still interact with it somehow due to their covalent bonding to the sugar phosphate backbone of the DNA and due to retained mutual non-covalent interaction between mutated bases. It means that these mutated/decoupled bases still affect the geometry of the Zif268-DNA complex, which could affect resulting values of binding free energies.

Last but not least, the detailed interpretation of the obtained binding free energy values at the atomic level provides the basis for the rational design of ZFNs (design of point mutations of amino acids) so that they can be applied to any DNA sequence and without the risk of off-target effects.

Bibliography

[1] J. Hanus et al. $-CH_2-$ lengthening of the internucleotide linkage in the ApA dimer can improve its conformationa compatibility with its natural polynukleotide counterpart. *Nucleic Acid Res.*, 29:5182–5194, 2001.

[2] H. Sipova et al. 5'-O-Methylphosphonate nucleic acids — new modified DNAs that increase the Escherichia coli RNase H cleavage rate of hybrid duplexes. *Nucleic Acid Res.*, 42:5378–5389, 2014.

[3] M. Egli and M. Manoharan. Chemistry, structure and function of approved oligonucleotide therapeutics. *Nucleic Acid Res.*, 51:2529–2573, 2023.

[4] C. Ashmore-Harris and G. O. Fruhwirth. The clinical potential of gene editing as a tool to engineer cell-based therapeutics. *Clin. Trans. Med.*, 9:15, 2020.

[5] D. M. Ichikawa et al. A universal deep-learning model for zinc finger design enables transcription factor reprogamming. *Nat. Biotechnol.*, 41:1117–1129, 2023.

[6] V. Gapsys and et al. Comment on "Deficiencies in Molecular Dynamics Simulation-Based Prediction of Protein-DNA Binding Free Energy Landscapes". *J. Phys. Chem. B*, 124:1115–1123, 2020.

[7] R. Dahm. Discovering DNA: Friedrich Miescher and the early years of nucleic acid research. *Human Genetics*, 122(6):565–581, 2008.

[8] K. Hyongbum and K. Jin-Soo. A guide to genome engineering with programmable nucleases. *Nature Reviews Genetics*, 15:321–334, 2014.

[9] Y. J. Kim et al. Hybrid restriction enzymes: zinc finger fusions to Fok I cleavage domain. *Proc. Natl. Acad. Sci.*, 93(3):1156–1160, 1996.

[10] T. Gaj et al. A comprehensive approach to zinc-finger recombinase customization enables genomic targeting in human cells. *Nucleic Acids Research*, 41, 2013.

[11] J. Bitinaite et al. FokI dimerization is required for DNA cleavage. *Proc. Natl. Acad. Sci.*, 95(18):10570–10575, 1998.

[12] K. H. Bae et al. Human zinc fingers as building blocks in the construction of artificial transcription factors. *Nature Biotech*, 21:275–280, 2003.

[13] J. C. Miller et al. A TALE nuclease architecture for efficient genome editing. *Nature Biotech*, 29:143–148, 2011.

[14] A. J. Bogdanove and B. L. Stoddard. The crystal structure of TAL effector PthXo1 bound to its DNA target. *Science*, 335:716–719, 2012.

[15] D. Deng et al. Structural basis for sequence-specific recognition of DNA by TAL effectors. *Science*, 335:720–723, 2012.

[16] Q. Ding et al. A TALEN genome-editing system for generating human stem cell-based disease models. *Cell Stem Cell*, 12(2):238–251, 2013.

[17] Y. Kim et al. A library of TAL effector nucleases spanning the human genome. *Nature Biotech*, 31:251–258, 2013.

[18] K. S. Makarova et al. A putative RNA-interferencebased immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. *Biol. Direct*, 1(7), 2006.

[19] R. Barrangou et al. CRISPR provides acquired resistance against viruses in prokaryotes. *Science*, 315:1709–1712, 2007.

[20] M. Jinek et al. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science*, 337:816–821, 2012.

[21] P. D. Hsu et al. DNA targeting specificity of RNA-guided Cas9 nucleases. *Nature Biotech.*, 31:827–832, 2013.

[22] H. Koike-Yusa et al. Genome-wide recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library. *Nature Biotech.*, 32:267–273, 2013.

[23] P. A. Kollman et al. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.*, 117(19):5179–5197, 1995.

[24] A. Warshel et al. Consistent force field for calculation of vibrational spectra and conformations of some amides and lactam rings. *Journal of Molecular Spectroscopy*, 33(1):84–99, 1970.

[25] H. Sun et al. An ab Initio CFF93 All-Atom Force Field for Polycarbonates. *J. Am. Chem. Soc.*, 116(7):2978–2987, 1994.

[26] H. Sun. COMPASS: An ab Initio Force-Field Optimized for Condensed-Phase Applications – Overview with Details on Alkane and Benzene Compounds. *J. Phys. Chem. B*, 102(38):7338–7364, 1998.

[27] P. M. Morse. Diatomic molecules according to the wave mechanics. II. Vibrational levels. *Phys. Rev.*, 34(1):57–64, 1929.

[28] M. Somoza. Morse potential (2006). *Wikipedia, The Free Encyclopedia*, accessed Dec. 2023.

[29] H. C. Urey and C. A. Bradley Jr. (1931). The Vibrations of Pentatonic Tetrahedral Molecules. *Phys. Rev.*, 38(11), 1969.

[30] L. Antonelli. Dihedral angle (2018). *Wikipedia, The Free Encyclopedia*, accessed Dec. 2023.

[31] Charles-Augustin de Coulomb. Premier mémoire sur l'électricité et le magnétisme [First dissertation on electricity and magnetism]. *Histoire de l'Académie Royale des Sciences [History of the Royal Academy of Sciences]*, page 569–577, 1785.

[32] Charles-Augustin de Coulomb. Second mémoire sur l'électricité et le magnétisme [Second dissertation on electricity and magnetism]. *Histoire de l'Académie Royale des Sciences [History of the Royal Academy of Sciences]*, page 578–611, 1785.

[33] D. Halliday; R. Resnick; J. Walker. *Fundamentals of Physics*. John Wiley and Sons, 2013.

[34] J. E. Lennard-Jones. Cohesion. *Proc. Phys. Soc.*, 43(5):461–482, 1931.

[35] Scientific Figure on ResearchGate. Parameters (σ, ε) of Lennard-Jones Potential for Fe, Ni, Pb and Cr based on Melting Point Values Using the Molecular Dynamics Method of the LAMMPS Program, 2020. `https://www.researchgate.net/figure/Potential-Lennard-Jones-2_fig1_341976308` [Accessed: Dec. 2023].

[36] F. L. Leite. Theoretical Models for Surface Forces and Adhesion and Their Measurement Using Atomic Force Microscopy. *Int. J. Mol. Sci.*, 13(10):12773–12856, 2012.

[37] J. K. Roberts and W. J. Orr. Induced dipoles and the heat of adsorption of argon on ionic crystals. *Trans. Faraday Soc.*, 34:1346–1349, 1938.

[38] W. G. Keesom. The second virial coefficient for rigid spherical molecules whose mutual attraction is equivalent to that of a quadruplet placed at its center. *Proc. R. Neth. Acad. Arts Sci.*, 18:636–646, 1915.

[39] D. J. Griffiths and D. F. Schroeter. *Introduction to Quantum Mechanics*. 3rd Edition. Cambridge University Press, Cambridge, 2018.

[40] A. D. MacKerell Jr. et al. Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *J. Comput. Chem.*, 25(11):1400–1415, 2004.

[41] M. Karplus et al. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.*, 4(2):187–217, 1983.

[42] A. D. MacKerell Jr. et al. CHARMM: The biomolecular simulation program. *J. Comput. Chem.*, 30(10):1545–1614, 2009.

[43] S. Pronk et al. GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics*, 29(7):845–854, 2013.

[44] J. C. Phillips et al. Scalable molecular dynamics with NAMD. *J. Comp. Chem.*, 26:1781–1802, 2005.

[45] A. D. MacKerell Jr. et al. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B*, 102(18):3586–3616, 1998.

[46] A. D. MacKerell Jr. et al. CHARMM additive and polarizable force fields for biophysics and computer-aided drug design. *Biochimica et Biophysica Acta (BBA) – General Subjects*, 1850(5):861–871, 2015.

[47] The Free Encyclopedia Wikipedia. Gradient descent. `https://en.wikipedia.org/wiki/Gradient_descent` [Accessed: Mar. 2024].

[48] The Free Encyclopedia Wikipedia. Conjugate gradient method. `https://en.wikipedia.org/wiki/Conjugate_gradient_method` [Accessed: Mar. 2024].

[49] The Free Encyclopedia Wikipedia. Newton's method in optimization. `https://en.wikipedia.org/wiki/Newton's_method_in_optimization` [Accessed: Mar. 2024].

[50] The Free Encyclopedia Wikipedia. Hessian matrix. `https://en.wikipedia.org/wiki/Hessian_matrix` [Accessed: Mar. 2024].

[51] Sir W. R. Hamilton. On a General Method of Expressing the Paths of Light, and of the Planets, by the Coefficients of a Characteristic Function. *Dublin University Review and Quarterly Magazine*, 1:795–826, 1833.

[52] L. Verlet. Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. *Phys. Rev.*, 159(1):98–103, 1967.

[53] The Free Encyclopedia Wikipedia. Poisson bracket. `https://en.wikipedia.org/wiki/Poisson_bracket` [Accessed: Mar. 2024].

[54] M. Škorňa. *Spontaneous Symmetry Breaking and the Non-relativistic Goldstone Theorem*. Bachelor Thesis. Masaryk University, Brno, 2021.

[55] H. Trotter. On the Product of Semigroups of Operators. *Proc. Amer. Math. Soc.*, 10(4):545–551, 1959.

[56] B. C. Hall. *Lie Groups, Lie Algebras, and Representations: An Elementary Introduction*. 2nd Edition. Springer Cham, 2015.

[57] M. E. Tuckerman. *Statistical Mechanics: Theory and Molecular Simulation*. 1st Edition. Oxford University Press, Oxford, 2010.

[58] M. Tuckerman et al. Reversible multiple time scale molecular dynamics. *J. Chem. Phys.*, 97(2):1990–2001, 1992.

[59] S. Nosé. A unified formulation of the constant temperature molecular dynamics methods. *J. Chem. Phys.*, 81(1):511–519, 1984.

[60] W. G. Hoover. Canonical dynamics: Equilibrium phase-space distributions. *Phys. Rev. A*, 31(3):1695–1697, 1985.

[61] P. Langevin. Sur la théorie du mouvement brownien [On the Theory of Brownian Motion]. *C. R. Acad. Sci. Paris*, 146:530–533, 1908.

[62] M. P. Allen and D. K. Tildesley. *Computer Simulation of Liquids*. 1st Edition. Oxford University Press, New York, 1989.

[63] B. Frenkel and B. Smit. *Understanding molecular simulation: from algorithms to applications*. 2nd Edition. Academic Press, San Diego, 2002.

[64] The Free Encyclopedia Wikipedia. Periodic boundary conditions. `https://en.wikipedia.org/wiki/Periodic_boundary_conditions` [Accessed: Mar. 2024].

[65] The Free Encyclopedia Wikipedia. First law of thermodynamics. `https://en.wikipedia.org/wiki/First_law_of_thermodynamics` [Accessed: Mar. 2024].

[66] The Free Encyclopedia Wikipedia. Inexact differential. `https://en.wikipedia.org/wiki/Inexact_differential` [Accessed: Mar. 2024].

[67] The Free Encyclopedia Wikipedia. Second law of thermodynamics. `https://en.wikipedia.org/wiki/Second_law_of_thermodynamics` [Accessed: Mar. 2024].

[68] The Free Encyclopedia Wikipedia. Third law of thermodynamics. `https://en.wikipedia.org/wiki/Third_law_of_thermodynamics` [Accessed: Mar. 2024].

[69] E. A. Guggenheim. *Thermodynamics: An Advanced Treatment for Chemists and Physicists*. 8st Edition. North Holland, Amsterdam, 1986.

[70] The Free Encyclopedia Wikipedia. Thermodynamic potential. `https://en.wikipedia.org/wiki/Thermodynamic_potential` [Accessed: Mar. 2024].

[71] The Free Encyclopedia Wikipedia. Helmholtz free energy. `https://en.wikipedia.org/wiki/Helmholtz_free_energy` [Accessed: Mar. 2024].

[72] The Free Encyclopedia Wikipedia. Legendre tranformation. `https://en.wikipedia.org/wiki/Legendre_transformation` [Accessed: Mar. 2024].

[73] The Free Encyclopedia Wikipedia. Gibbs free energy. `https://en.wikipedia.org/wiki/Gibbs_free_energy` [Accessed: Mar. 2024].

[74] The Free Encyclopedia Wikipedia. Enthalpy. `https://en.wikipedia.org/wiki/Enthalpy` [Accessed: Mar. 2024].

[75] M. K. Brachman. Chemical Potential, and Gibbs Free Energy. *J. Chem. Phys.*, 22(6):1152, 1954.

[76] R. Bernardi et al. *NAMD User's Guide*. Version 2.14. University of Illinois, Urbana, IL, 2020.

[77] The Free Encyclopedia Wikipedia. Partition function (statistical mechanics). https://en.wikipedia.org/wiki/Partition_function_(statistical_mechanics) [Accessed: Mar. 2024].

[78] K. Mehran. *Statistical Physics of Particles*. 1st Edition. Cambridge University Press, New York, 2007.

[79] The Free Encyclopedia Wikipedia. Isothermal–isobaric ensemble. https://en.wikipedia.org/wiki/Isothermal-isobaric_ensemble [Accessed: Mar. 2024].

[80] D. Frenkel and B. Smit. *Understanding Molecular Simulation: From Algorithms to Applications*. 2nd Edition. Academic Press, 2001.

[81] Alchemistry Wiki. Thermodynamic integration. http://alchemistry.org/wiki/Thermodynamic_Integration [Accessed: Mar. 2024].

[82] The Free Encyclopedia Wikipedia. Thermodynamic integration. https://en.wikipedia.org/wiki/Thermodynamic_integration [Accessed: Mar. 2024].

[83] M. Jorge et al. Effect of the Integration Method on the Accuracy and Computational Efficiency of Free Energy Calculations Using Thermodynamic Integration. *J. Chem. Theo. Comp.*, 6:1018–1027, 2010.

[84] C. Shyu and F. M. Ytreberg. Reducing the Bias and Uncertainty of Free Energy Estimates by Using Regression to Fit Thermodynamic Integration Data. *J. Comp. Chem.*, 30:2297–2304, 2009.

[85] R. W. Zwanzig. High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. *J. Chem. Phys.*, 22(8):1420–1426, 1954.

[86] C. H. Bennett. Efficient estimation of free energy differences from Monte Carlo data. *J. Comput. Phys.*, 22:245–268, 1976.

[87] C. Jarzynski. Nonequilibrium equality for free energy differences. *Phys. Rev. Lett.*, 78(14):2690, 1997.

[88] C. Jarzynski. Equilibrium free-energy differences from nonequilibrium measurements: A master-equation approach. *Phys. Rev. E*, 56(5):5018, 1997.

[89] D. Chandler. *Introduction to Modern Statistical Mechanics*. 1st Edition. Oxford University Press, New York, 1987.

[90] G. E. Crooks. Entropy production fluctuation theorem and the nonequilibrium work relation for free energy differences. *Physical Review E*, 60:2721, 1999.

[91] G. E. Crooks. Path-ensemble averages in systems driven far from equilibrium. *Phys. Rev. E*, 61:2361–2366, 2000.

[92] C. E. Shannon. A Mathematical Theory of Communication. *Bell Syst. Tech. J.*, 27(3):379–423, 1948.

[93] D. Collin et al. Verification of the Crooks fluctuation theorem and recovery of RNA folding free energies. *Nature*, 437(7056):231–234, 2005.

[94] RCSB. Protein Data Bank. `https://www.rcsb.org` [Accessed: Mar. 2024].

[95] E. F. Pettersen et al. UCSF Chimera – a visualization system for exploratory research and analysis. *Journal of Molecular Graphics*, 14(1):33–38, 1996.

[96] A. Dalke W. Humphrey and K. Schulten. VMD: Visual Molecular Dynamics. *Journal of Molecular Graphics*, 14(1):33–38, 1996.

[97] SilcsBio. CGenFF. `https://cgenff.silcsbio.com` [Accessed: Mar. 2024].

[98] CHARMM – official site. `https://www.charmm.org` [Accessed: Mar. 2024].

[99] MacKerell Lab. CHARMM Force Field Files. `https://mackerell.umaryland.edu/charmm_ff.shtml` [Accessed: Mar. 2024].

[100] R. Baron. *Computational Drug Discovery and Design*. 1st Edition. Humana Press, New York, 2012.

[101] S. Boresch and M. Karplus. The Role of Bonded Terms in Free Energy Simulations. 2. Calculation of Their Influence on Free Energy Differences of Solvation. *J. Phys. Chem. A*, 103(1):119–136, 1999.

[102] T. C. Beutler et al. Avoiding singularities and numerical instabilities in free energy calculations based on molecular simulations. *Chem. Phys. Lett.*, 222(6):529–539, 1994.

[103] M. Zacharias et al. Separation-shifted scaling, a new scaling method for Lennard-Jones interactions in thermodynamic integration. *J. Chem. Phys.*, 100(12):9025–9031, 1994.

[104] Cesnet. MetaCentrum. `https://metavo.metacentrum.cz` [Accessed: Mar. 2024].

[105] W. D. Scott. On optimal and data-based histograms. *Probability Theory and Related Fields*, 66(3):605–610, 1979.

[106] D. Freedman and P. Diaconis. On the histogram as a density estimator: L_2 theory. *Probability Theory and Related Fields*, 57(4):453–476, 1981.

[107] The Free Encyclopedia Wikipedia. Interquartile range. `https://en.wikipedia.org/wiki/Interquartile_range` [Accessed: Mar. 2024].

[108] M. Goette and H. Grubmuller. Accuracy and Convergence of Free Energy Differences Calculated from Nonequilibrium Switching Processes. *J. Comput. Chem.*, 30:447–456, 2009.

[109] S. R. Michael et al. Extremely precise free energy calculations of amino acid side chain analogs: Comparison of common molecular mechanics force fields for proteins. *J. Chem. Phys.*, 119:5740, 2003.

[110] V. A. Ngo et al. Comparative Analysis of Protein Hydration from MD simulations with Additive and Polarizable Force Fields. *Adv. Theory Simul.*, 2:1800106, 2019.

[111] R. Wolfenden et al. Affinities of amino acid side chains for solvent water. *Biochemistry*, 20(4):849–855, 1981.

[112] M. Shirts and V. S. Pande. Screen Savers of the World Unite! *Science*, 290:1903–1904, 2000.

[113] B. R. Brooks at al. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.*, 4:187, 1983.

[114] J. L. Miller and P. A. Kollman. Solvation Free Energies of the Nucleic Acid Bases. *J. Chem. Phys.*, 100:8587–8594, 1996.

[115] D. Seeliger et al. Towards computional specificity screening of DNA-binding proteins. *Nucleic Acids Research*, 39(19):8281–8290, 2011.

[116] V. Gapsys and B. L. de Groot. Alchemical Free Energy Calculations for Nucleotide Mutations in Protein-DNA Complexes. *J. Chem. Theory Comput.*, 13:6275–6289, 2017.

[117] M. Khabiri and P. L. Freddolino. Deficiencies in Molecular Dynamics Simulation-Based Prediction of Protein-DNA Binding Free Energy Landscapes. *J. Phys. Chem. B*, 121:5151–5161, 2017.

[118] Y. Sugita et al. Multidimensional replica-exchange method for free-energy calculations. *J. Chem. Phys.*, 113:6042–6051, 2000.

[119] H. Fukunishi et al. On the Hamiltonian replica exchange method for efficient sampling of biomolecular systems: application to protein structure prediction. *J. Chem. Phys.*, 116:9058–9067, 2002.

[120] T. Hamilton et al. Comparison of the DNA binding characteristics of the related zinc finger proteins WT1 and EGR1. *Biochemistry*, 37:2051–2058, 1998.

[121] M. Vazdar et al. Arginine "Magic": Guanidinium Like-Charge Ion Pairing from Aqueous Salts to Cell Penetrating Peptides. *Acc. Chem. Res.*, 51(6):1455–1464, 2018.

[122] W. Saenger. *Principles of Nucleic Acid Structure*. 1st Edition. Springer New York, New York, 2013.

[123] G. N. Parkinson et al. Crystal structure of parallel quadruplexes from human telomeric DNA. *Nature*, 417(6891):876–880, 2002.