Computational Nanoscience and Multiscale Modelling of DNA Molecules

By

Massimo Lai

Supervisor: Prof. Dimitris Drikakis

A thesis submitted to Cranfield University
In Partial Fulfillment of the Requirements
for the Degree of
Doctor of Philosophy

Cranfield University
Department of Aerospace Sciences
Cranfield, UK
October 2010

Examiner's Copy

Supervisor: Prof. Dimitris Drikakis, 2010.

Typeset in LaTeX 2_ε.

Except where acknowledged in the customary manner, the material presented in this thesis is, to the best of my knowledge, original and has not been submitted in whole or part for a degree in any university.

Massimo Lai
Supervisor: Prof. Dimitris Drikakis

Computational Nanoscience and Multiscale Modelling of DNA Molecules

Massimo Lai
Supervisor: Prof. Dimitris Drikakis

October 26, 2010

Abstract

Molecular Dynamics is a very powerful technique for the investigation of matter at nanoscopic level. However, it's application in many fields, such as the investigation of many relevant processes of cell biology, is restricted by issues of computational cost. Therefore, in recent years, a growing interest has been generated by the introduction of Coarse-Grained (CG) models, that allow the investigation of bigger systems for longer timescales.

In this thesis, Molecular Dynamics was used in order to gain a quantitative understanding of mechanical and diffusive processes of DNA molecules in solution, and in order to parametrise a Coarse Grained model of DNA capable of a qualitative description of the mechanical behaviour of the all-atom model at equilibrium.

A software package for the computation of Coarse-Grained interaction force-fields, making use of the recently developed Multiscale Coarse-Grained Method (MSCG) by Izvekov and Voth [1] was implemented.

We tested and validated the method by performing a one-point-per-molecule coarse graining of TIP3P water. The resulting model was able to reproduce the fluid structure (its radial distribution function) in a satisfactory and nearly quantitative way.

Finally, we applied the MSCG method to a more demanding problem, namely the parametrisation of a 3-point-per-residue coarse-grained model of double-stranded DNA.

As a consequence, the agreement of the obtained CG model with the atomistic structure was still not quantitative. In particular, the helical geometry was qualitatively preserved and the Root-Mean-Square Displacement (RMSD) of the coarse-grained model was stable over the trajectory, but higher than its all-atom counterpart.

We suggest several possible routes for future improvements. In particular, the explicit modeling of torsional degrees of freedom of the DNA backbone, and the use of recently introduced methods for the refinement of the MSCG estimation of force-field parameters, and a more accurate treatment of Coulombic interactions.

List of Publications

Part of the work presented in this thesis was published in the following conference and journal papers:

M. Lai, D. Drikakis “Molecular Dynamics investigation of salt-dependent diffusion coefficients of ssDNA oligomers in aqueous solution.” *Proceedings of the 1st International Conference on Mathematical and Computational Biomedical Engineering - CMBE2009*, June 29 - July 1, 2009, Swansea, UK.

M. Lai, D. Drikakis, “Notes on the implementation of brownian motion in mesoscopic fluid-particle models." *Proceedings of the 7th Industrial Simulation Conference - ISC’2009*, June 1-3, 2009, Loughborough, United Kingdom.

N. Asproulis, M. Benke, M. Lai, E. Shapiro, D. Drikakis, D. Brown, M. Dawson, G. Pollard, P. Ioannou, V. Poulopoulos, “Modelling approaches for micro- and nanoscale diffusion phenomena". *International Conference on Process Intensification in Nanotechnology*, Sept. 15-18, 2008, Albany, New York, USA.

M. Lai, M. Kalweit and D. Drikakis, “Temperature and ion concentration effects on the viscosity of Price-Brooks TIP3P water model", *accepted on Molecular Simulation*.

M. Lai and D. Drikakis, “A biomolecule-oriented Python implementation of the Multiscale Coarse-Graining method", *submitted to the Journal of Computational and Theoretical Nanoscience*.

Acknowledgements

First of all, I have to thank my supervisor, Prof. Dimitris Drikakis for his encouragement and support throughout the duration of this thesis, and for always providing a constructive working atmosphere.

My heartfelt thanks to the other present and former members of the Fluid Mechanics and Computational Science group, for their external support and for the many coffee breaks we used to share.

A special thanks goes to Marco Kalweit, for many useful discussions, and to Matyas Benke with whom I've shared so much (maybe too much) office time.

Outside the walls of the School of Engineering, many people have made this time worth remembering, and have helped get me through the occasional (and not so occasional) wee spots of trouble.

They are probably too many to be put in a list that risks to be unfairly incomplete. I want to mention my fArmer housemates (it's not a typo), the old friends who have always been close despite of the distance, and of course our little Silsoe tribe. A big hug to all of you.

Thanks also to the guys and ladies of the Cranfield University Volleyball Team for the few victories and many crushing defeats we have enjoyed together.

Finally, I want to say thanks to my family and to all those who have cared about me, often more deeply than I would have expected. They know who they are.

Contents

List of Figures

List of Tables

Abbreviations

Acronyms

AA	All-Atom
ACF	Autocorrelation function
BAR	Bennet Acceptance Ratio
CF	Correlation function
CG	Coarse-Grained
dsDNA	Double-Stranded DNA
DNA	Deoxyribonucleic Acid
EMD	Equilibrium Molecular Dynamics
LJ	Lennard-Jones (interaction potential)
MBAR	Multistate Bennett Acceptance Ratio
MD	Molecular Dyanmics
MSCG	Multiscale Coarse-Graining
NEMD	Non Equilibrium Molecular Dynamics
PDF	Probability Distribution Function
PMF	Potential of Mean Force
RDF	Radial Distribution Function
ssDNA	Single-Stranded DNA
SMD	Steered Molecular Dynamics
US	Umbrella Sampling
WHAM	Weighted Histograms Analysis Method

Greek Symbols

α	finite-size correction factor for computed viscosity
β	reduced temperature ($=1/k_BT$)
δ	phase angle of harmonic torsion potential
Γ	phace space of a system of particles
ϵ	Lennard-Jones energy parameter
ϵ_0	dielectric constant *in vacuo*
η	kinematic viscosity
θ	bending angle width
θ_0	bending angle rest width
λ	steering velocity for SMD driving potential
Λ	De Broglie's thermal wavelength
ξ	reaction coordinate for PMF calculations
ρ	empirical PDF of a reaction coordinate in US calculations
σ	Lennard-Jones collision radius
χ^2	squared sum of residuals
ϕ	vector of unknowns for the least-squares MSCG problem

Latin Symbols

a	mean free path for a system of particles
$\mathbf{a}_i$	acceleration of particle i
A	Helmholtz free energy
h	perturbing Hamiltonian steered at velocity λ
$\hbar$	Planck's constant
$\mathcal{H}$	Hamiltonian
k_B	Boltzmann constant
K	kinetic energy
$\mathcal{G}$	matrix of the MSCG linear system
m	multiplicity of dihedral torsion angle potential
m_i	mass of particle i
p	system pressure
P_{ij}	pressure tensor components
$\mathbf{p}$	momenta of the AA system
$\mathbf{P}$	momenta of the CG system
$\mathbf{p_i}$	momentum vector of particle i
p_i	momentum component
$\mathbf{q_i}$	coordinate vector for particle i
q_i	cartesian coordinates
$\mathbf{Q}$	coordinates of the CG system
$\mathbf{r}$	configuration of the AA system
$\mathbf{r_i}$	position vector of particle i
r_i	position vector component of particle i
r	harmonic bond length
r_0	harmonic bond rest length
$\mathbf{r_{ij}}$	relative position vector between i and j
$\mathbf{R}$	configuration of the CG system
T	system temperature
$\mathcal{W}$	virial of a system of particles
U	potential energy
$\mathbf{v}$	velocity vector
$\mathbf{v}$	velocity vector
v_i	velocity component
w	weight factor for US calculations
$\mathbf{x}$	vector of positions and momenta of a Hamiltonian system

Part I

Introduction

1 Introduction

Molecular dynamics has evolved over the last five decades as a useful tool for the investigation of matter at nanoscopic level but its applicability is severely conditioned by the available computational power. In particular, the investigation of cell biology by means of molecular simulations has encountered a major obstacle in comparatively large length scale and timescale over which many important biochemical phenomena take place. A typical example is the challenge offered by protein-folding processes and DNA hybridisation, whose description by means of a purely "brute-force" approach would require the simulation of millions of atoms for a time ranging from microseconds to seconds [6]. The timescale accessible to MD simulations keeps growing at a steady pace, from the sub-nanosecond simulations of the 1980s, to the nanosecond barrier in the early 1990s, up to the over 100ns we can afford today. However, despite the enormous growth of the available computational power offered by modern high-performance parallel computing facilities, the timescale of most biological processes remains out of reach by a factor of several order of magnitude, a gap that all-atom simulations will not be able to bridge in the foreseeable future.

Because of the above mentioned limitations, the last decade has seen a growing interest

in coarse-grained (CG) models of biomolecular systems. The idea behind coarse-grained models is that not all the molecular motions reproduced by standard MD are actually relevant, and that a simplified model with lower resolution and fewer particles could be able to simulate the system with sufficient accuracy to retain the necessary structural and dynamic information. Such a system would be much cheaper to simulate, thus increasing the accessible length- and timescale [7].

Another potentially attractive side of CG models is their capability of exploring the phase space of the represented systems, a task that has been shown to be extremely problematic for all-atom simulations of biomolecules [8, 9]. The removal of many degrees of freedom makes the energy landscape smoother, a situation that facilitates the crossing of energy barriers between metastable states [10]. On the other hand, this implies an alteration in the system dynamics, which can be non-trivial to assess [11].

The coarse graining process takes place in two main steps: first, the topological mapping of all-atom models onto simplified CG models, where one single interaction site or "superatom" corresponds to a group of atoms; secondly, the determination of physically consistent force-field for the resulting system of CG "sites" [12, 13].

There exist several methods for the extraction of CG pairwise potentials from AA systems. The first systematic attempts at coarse graining were aimed at fluid systems, such as water, leading to the development of the Boltzmann Inversion technique, which was also fruitfully applied to polymers [13, 14]. However, for the parametrisation of more complicated systems, the method that has probably received the most attention is the force-matching approach.

The first instance of this modeling strategy can be found in the work Ercolessi and Adams [15], and was then reprised, adapted and expanded by Izvekov and Voth [1, 16]. Further contributions were made by Noid,who also gave the method a more detailed theoretical footing [17, 18]. The force-matching was the method chosen for our study.

The exploration of the potential of CG simulations has only started, and the possibility to break the nanometer/nanosecond barrier opens interesting horizons and a set of possible applications for many previously unfeasible systems: for example, Arkhipov and coworkers have managed to produce a model of bacterial flagellum, whilst Freddolino produced a CG model of a viral capside after simulating a whole virus with atomistic detail [19, 20].

The emphasis that the MSCG method has put on the physically consistency of the resulting CG force-fields is remarkable, because it addresses one of the major weaknesses of coarse-graining methods so far, namely the fact that the simulated dynamics and timescale

of the CG system can be different from the atomistic one. This happens mainly because of the geometrical simplification and the smoothing of the energy landscape due to the removal of many degrees of freedom. [11]. However, the number of developments proposed in recent years seems to hint at a very promising future for the rapidly evolving field of multiscale simulations.

1.1 Aims and objectives

The purpose of this thesis is twofold. First, we have used all-atom Molecular Dynamics simulation in order to gather a quantitative understanding of molecular motions and diffusive properties of DNA molecules. Secondly, we have implemented and tested the Multiscale Coarse-Graining method in a flexible and easily extendable software package, with the purpose of developing coarse-grained models of DNA that can capture basic mechanical properties and allow the simulation of bigger systems for longer timescales.

1.2 Thesis Structure

The thesis is organised as follows.
The first chapter gives an overview of Molecular Dynamics, the available methods for the simulation of biomolecules, and current issues and challenges as well as recent developments. It also introduces the basic concepts behind coarse graining and the most widely adopted methods for the parametrisation of CG force fields. A more detailed section is dedicated to the Multiscale Coarse Graining method [1], on which much of the work of this thesis is based.
The second chapter gives a brief summary of the numerical methods used for the postprocessing of our simulations, the related issues and the adopted solutions.
The chapters 4,5 and 6 present the results of all-atom studies of diffusive properties of DNA molecules in solution. In particular, chapter three focusses on the properties of the water model used for our simulations; chapter four uses the gathered information (especially about model water viscosity at different temperatures) in order to simulate and quantify the diffusion of short DNA strands. Chapter five uses the parameters thus quantified in order to test how much of the atomistic behaviour can be captured in a simplified implementation of diffusion in fluid-particle models.
Chapter 7 is dedicated to the mapping of atomistic structures onto reduced coarse-grained

systems, with particular regard to the molecules studied in this thesis.
Chapter 8 and 9, finally, present the results obtained applying the Multiscale Coarse-Graining method to a test-case system of double-stranded DNA in solution. A discussion of the adopted hypotheses and modeling strategies is given. The behaviour of the coarse-system when freely fluctuating at equilibrium is compared with that of the fine-grain all-atom simulations.
Chapter 10 provides a summary of the obtained results and a concise discussion of future developments that can be beneficial for the application of the Multiscale Coarse-Graining method to biomolecular systems.

Part II

Theory

2

Theoretical Background

This chapter gives a summary of the theoretical aspects of atomistic and coarse-grained molecular simulations. At the same time, is gives an overview of the extensive literature search undertaken, with particular regard to the available methods for the calculation of potentials of mean force and coarse-grained molecular interactions. Not all methods described in the following have proven suitable of useful for our purposes, however, they are presented for completeness and for their relevance in many other applications.

2.1 Principles of classical statistical mechanics

This section will concisely outline the general theoretical principles that underlie all methods illustrated in the later chapters of this thesis. Statistical mechanics is a branch of Physics whose goal it to explain the properties of macroscopic (thermodynamic) systems from the motion the nanoscopic particles of which they are made. Historically, it was the results of a long effort to reduce thermodynamics to the laws of mechanics. Classical Molecular Dynamics is essentially a computational tool to investigate the statistical mechanics of molecular systems [21], ideally by simulating the motion of their particles

for a time sufficiently long to ensure stable estimates of observable physical properties. An MD code integrates the system's equations of motion and generates its "trajectory" in time, i.e. the time evolution of all particles' positions and velocities (or momenta). This information is then fed into statistical mechanical equations, in order to calculate structural and thermodynamic properties.

2.1.1 Basic assumptions

Hamiltonian mechanics states that the dynamical properties of a system can be calculated from the knowledge of its total energy, or "hamiltonian", $\mathcal{H}$, expressed as a function of general coordinates q_i and their so-called "conjugate momenta" p_i. Once the Hamiltonian $\mathcal{H}(p_i, q_i)$ is known, the equations of motion for positions and momenta can be expressed as:

$$\frac{\partial p_i}{\partial t} = -\frac{\partial \mathcal{H}}{\partial q_i} \tag{2.1}$$

$$\frac{\partial q_i}{\partial t} = \frac{\partial \mathcal{H}}{\partial p_i} \tag{2.2}$$

In general we will deal with a system of N particles (mass points), which possesses 3N degrees of freedom; for such a system, the index i will go from 1 to 3N. The general coordinates q_i will be the particles' Cartesian coordinates and the conjugate momenta p_i will simply be the "usual" particles' momenta. The total energy will be the sum of potential energy U (which is only a function of the positions) and kinetic energy K (which is only a function of the momenta):

$$\mathcal{H} = K(p_i) + U(q_i). \tag{2.3}$$

In principle, these equations contain all the information needed for the computation of the time evolution of the system. However, thermodynamic systems contain an immense number of atoms, and the direct calculation of all atomistic equations of motions will never be possible.

It is useful at this point to introduce the idea of *ensemble*, an alternative representation of a thermodynamic system first used by Gibbs [22]. Instead of a mechanical system evolving in time, Gibbs suggested to picture it as a collection of copies of itself, where each copy is in a different microstate (i.e. a set of allowed positions and velocities).

The probability with which the system visits its microstates determines the macroscopic properties of the system. It is useful, at this point, to introduce the concept of *phase space*. The mechanical configuration **x** of the system of N particles in a three-dimensional space is univocally determined once the $3N$ coordinates and $3N$ velocity components are known:

$$\mathbf{x} = [p_i, q_i], \quad i = 1 \dots 3N \tag{2.4}$$

Mathematically, we can either describe the system as a collection of N points in a three dimensional space, or as a single point in a 3N-dimensional space. When a many-body system is left free to evolve in time, we can visualising it as a point moving in its phase space, "visiting" different states. Macroscopic bodies, made of many moles of atoms, constantly change their nanoscopic configuration: atoms vibrate (solids) or diffuse (fluids), and it's not improper to say that an object is never, strictly speaking, in the same configuration twice. Nevertheless, observable properties (density, temperature, viscosity) are constant at the macroscale. Boltzmann's great achievement was the proof that the probability with which a system visits a microstate is proportional to the microstates energy, through a term called the Boltzmann factor:

$$P(\mathbf{x}) \propto exp\left(-\frac{E(\mathbf{x})}{T}\right). \tag{2.5}$$

Where E is the energy, T the absolute temperature, and k_B the Boltzmann constant. The absolute value of the probability would be:

$$P(\mathbf{x}) = \frac{exp\,(-E(\mathbf{x})/k_B T)}{\int_\Gamma exp\,(-E(\mathbf{x})/k_B T)\,d\mathbf{x}} \tag{2.6}$$

Where Γ indicates the whole set of possible configurations. The denominator in Eq. 2.6 is called "partition function", usually indicated as Z. The knowledge of the partition function is usually not achievable in practice, but as it will be seen in later sections, there are usually ways around this apparent difficulty [23].

For MD simulations, the three most important ensembles are the microcanonical (NVE), canonical (NVT) and isothermal-isobaric (NpT):

- NVE: constant particle number (N), volume (V), and total energy (E); it's the most "natural" ensemble for MD.
- NVT: constant particle number (N), volume (V), and temperature (T), requires a

thermostatting algorithm that simulated the contact with a heat reservoir at temperature T.

- NpT: constant particle number (N), pressure (p), and temperature (T), requires a thermostatting algorithm and a barostatting algorithm that controls the pressure.

2.1.2 Setup of a typical simulation

When a simulation is started from scratch, the time integrator needs a set of starting velocities for the system's particles. The most natural choice is to assign random velocities sampled from a Maxwell-Boltzmann probability distribution for a given temperature,

$$p(v_{i\alpha}) = \sqrt{\frac{m_i}{2\pi k_B T}} \cdot exp\left(-\frac{1}{2} m_i \frac{v_{i\alpha}^2}{k_B T}\right), \tag{2.7}$$

where i identifies the i-th particle, m_i is the corresponding mass, v_i is the velocity, $\alpha = x, y, z$, k_B is the Boltzmann constant and T the absolute temperature. However, a simpler option is to initialise the system using a uniform distribution, and the velocity distribution will quickly converge to a Maxwell-Boltzmann by effect of molecular collision [23]. The initial configuration of the simulation box may contain artifacts such as strongly overlapping atoms, which generate high repulsive forces and unphysically high atomistic velocities that can cause the time integrator to break down. This quickly leads to computation errors. In order to get rid of the overlapping atoms, all MD codes include a minimization algorithm, that works by minimising the potential energy of the system in the parameter space defined by the particles position. Most commonly used algorithms are steepest descent, conjugated gradient or Levenberg-Marquard [24–26]. The obtained minimised configuration is a safer starting point for a simulation. The minimised box must then be allowed to fluctuate and (hopefully) find a good energy minimum. At the macroscale, this happens spontaneously as the system has plenty of time so sample its configuration space and even fluctuate enough to jump out of unwelcome basins of local energy minimum [27–29]. At the nanosecond timescale, however, the system can easily remain trapped into metastable states and not be able to leave them spontaneously during the simulated time.

2.1.3 Calculation of mechanical and thermodynamic quantities

The simulation of matter at atomistic level gives us access to the "observation" of a length and timescale which is out of reach for experimental methods. At his level of detail, all properties of a system in conditions of thermodynamic equilibrium happen to be continuously fluctuating; this is also true for macroscopic bodies, but in the case of experimental observation, the fluctuations are extremely small in comparison to the average value. Therefore, when considering any property, we have to distinguish between its instant value and its ensemble average over a trajectory, which is the quantity truly related to what is observable in the macroscopic world. It must be further noted that the instant value is properly defined for energetic and mechanical properties, such as potential and kinetic energy, and ill-defined for *thermodynamic* properties, such as temperature and pressure, which are intrinsically statistical in nature, and denote properties of the system in terms of ensemble averages [23, 30]. The energy equipartition theorem allows us to relate the macroscopic absolute temperature with the nanoscopic average kinetic energy. From the generalised equipartition theorem for Hamiltonian systems, which for practical purposes is valid in any ensemble,

$$\langle p_k \cdot \partial \mathcal{H} / \partial p_k \rangle = k_B T \tag{2.8}$$

$$\langle q_k \cdot \partial \mathcal{H} / \partial q_k \rangle = k_B T, \tag{2.9}$$

we can derive, for a system of 3N point particles of mass m_i and momentum p_i, the so called instant "kinetic temperature", defined as:

$$T = \frac{1}{3Nk_B} \sum_{i=1}^{N} |\mathbf{p}_i|^2 / m_i, \tag{2.10}$$

which will fluctuate in time but whose average will be the system temperature. The calculation of pressure is slightly more troublesome. The most widely used method makes use of Clausius virial theorem, which follows from 2.9, and yields, for a system of N particles,

$$\langle \sum_{i=1}^{N} \mathbf{r}_i \cdot \mathbf{f}_i^{TOT} \rangle = -\frac{1}{3} N k_B T = -PV, \tag{2.11}$$

where $\mathbf{f}_i^{TOT}$ is the total force on the i-th particle. If we consider only internal forces, we can define what is called the "internal virial" $\mathcal{W}$,

$$\mathcal{W} = \sum_{i=1}^{N} \mathbf{r}_i \cdot \mathbf{f}_i^{INT} = \sum_{i=1}^{N} \mathbf{r}_i \cdot \nabla_{\mathbf{r}_i} V(\mathbf{r}_i), \tag{2.12}$$

which can be evaluated from the molecular trajectory, and leads to the relation

$$PV = Nk_BT + \langle \mathcal{W} \rangle, \tag{2.13}$$

from which the instantaneous pressure of the system can be calculated. Again, the instantaneous value has no macroscopic significance and can fluctuate heavily (typically even by hundreds of atmospheres during a simulation), and the thermodynamic pressure can only be recovered as an ensemble average. In recent years, some authors have disputed the soundness of this approach, proposing additional corrections; however, such arguments have in turn been disputed [31]. A physically sound MD code must ensure that statistical ensemble is correctly "sampled", or, in other words, that the microstates are generated with the correct probability. This is particularly important when simulating more complicated ensembles, such as the canonical and the isothermal-isobaric. The development of sound methods to simulate the effect of a thermostat and barostat on a MD simulation has been (and still is) a very active field of research in the last 30 years, as it will be discussed in later sections.

2.2 Brief introduction to Classical Molecular Dynamics

"Classical" MD relies on the hypotheses that the correct behaviour of the system, which is described by quantum mechanics, can be well approximated by the far simpler Newtonian mechanics. Of course the accuracy of the approximation gets worse as the quantum effects become relevant. A criterion commonly used to evaluate the applicability of the newtonian approach is given by the *thermal DeBroglie wavelength*, Λ, of the particles involved [21]:

$$\Lambda = \sqrt{\frac{2\pi\hbar^2}{mk_BT}} \tag{2.14}$$

where

$\hbar$ = Planck constant

k_B = Boltzmann constant

T = temperature

m = mass of the particle.

The applicability of the newtonian approximation is considered legitimate when $\Lambda \ll a$, where a is the mean nearest neighbour distance of the system. The approximation is rather poor for very light elements like H or He: their small mass m causes the wavelength to be larger.
The basic elements of a molecular dynamics code are:

A physical model of the system: this is contained in the function and the parameters that define the interaction potential, as well as in the boundary conditions used for the simulation box. The function and the parameters are usually referred to as "*force field*".

A time integration algorithm that, given the positions, velocities and forces, at the current timestep, calculates the new values at the following timestep.

A statistical ensemble where the thermodynamic properties of the system are calculated. An example is the canonical ensemble NVE, where the conserved quantities are number of particles (N), Volume (V) and Energy (E).

They will be discussed separately in the following sections.

2.2.1 Interaction potentials and force calculation

Modelling the physics of a set of particles is essentially modelling their interactions. The approach commonly used is to define a *potential*, a function $U(\mathbf{r_1}, \ldots, \mathbf{r_n})$ that gives the potential energy of the system depending on the relative position of the atoms. The corresponding force on the i-th particle is computed as

$$\mathbf{f_i} = \nabla_{\mathbf{r}_i} U(\mathbf{r}_1, \ldots, \mathbf{r}_n), \tag{2.15}$$

which is the gradient of the potential with respect to the spatial position of the particle.
In biomolecular system and fluid mixtures, a very convenient hypothesis is that the interactions are pairwise additive, which allows the total potential on a particle to be written a

sum of (comparatively simple to compute) pairwise interactions

$$U_i = \sum_{i}^{n} \sum_{j>i}^{n} U_{ij}(|\mathbf{r}_i - \mathbf{r}_j|) \tag{2.16}$$

Where the condition $j > i$ provides that every pair is not counted twice. (This approach fails when three-body forces become relevant, as in the case of metals or semiconductors, were entirely different potentials are required [21, 32]). Historically speaking, the first potential to be implemented was the Lennard-Jones. A Lennard Jones potential mimics the interactions due to Van der Waals forces between atoms. This model was able to reproduce quantitatively the properties of gaseous Argon. Its functional form is:

$$U_{LJ} = 4\epsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right] \tag{2.17}$$

The term $\sim r^{-12}$ models the repulsion at short distances, while the term $\sim r^{-6}$ is the "attractive tail" at longer distances. The parameters ϵ and σ can be fitted on the behaviour of the substance under examination. This potential reproduces a Van der Waals-like interaction (weak attractive forces caused by temporary dipole moments in the electronic cloud of the atoms). It's suitable for noble gasses, but not for structurally complex systems, which require more sophisticate formulations.

When the system contains several different atom types, pairwise interactions should in principle be parametrised for every different pair. This would require tens of different LJ potentials. Instead, the route commonly followed is to determine the interaction of a certain atom type with itself, and then use extrapolations ("mixing rules") to guess the interaction potential for heteroatomic pairs. Several mixing rules have been investigated by White and Al-Matar [33, 34]. The force decreases quickly with distance. Therefore, it is reasonable to assume that only neighbouring particles exert a remarkable effect. It is then possible to introduce a "cut-off distance", over which the force is zero. It can be done keeping a list of every particle's "neighbours", and updating it every few timesteps [35]. This saves a great amount of computational resources, approximately halving the time required for the calculation of pairwise interactions [23].

Anyway, the drawback is that when a particle crosses the cutoff radius, the energy makes a little jump (because of the small effect of the attractive tail) that can jeopardize the energy conservation. The problem can be solved using a potential that goes smoothly to 0 at the

cutoff distance R_c, for example biasing the (2.17), as illustrated in fig. 2.1:

$$U(r) = \begin{cases} U_{LJ} - U_{LJ}(R_c) & \text{if } r < R_c \\ \\ 0 & \text{if } r \geq R_c \end{cases}, \tag{2.18}$$

or using a "switching function" that smoothly takes the potential to 0 at the cutoff, as in the case of the CHARMM potential as implemented in the LAMMPS MD code [25, 36]. Once the potential is chosen, the force exerted on particle i by particle j is calculated as:

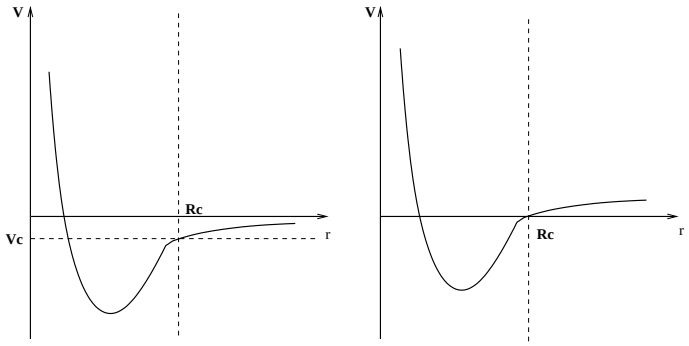


Figure 2.1: Biased Lennard-Jones potential

$$\mathbf{f}_{ij} = -\frac{\partial U_{ij}}{\partial \mathbf{r}_{ij}}, \tag{2.19}$$

Creating empirical potentials from scratch is an enormous task. A force field for biomolecular simulations of proteins, lipids and nucleic acids must contain a collection of parameters that describe the pairwise interactions of all atoms types, as well as the mechanical properties of all the bonds, angles, and dihedral torsional angles that can be found in the complex network of covalent bonds biomolecules are made of. Currently two potentials are the most widely used and validated for nucleic acids: AMBER [37] and CHARMM [36], named after the simulation packages where they were first implemented. The CHARMM potential has been used for all the atomistic simulations described in this thesis.

2.2.2 Integration of the equations of motion

A time integrator is a numerical scheme that integrates the equations of motions, propagating the trajectory of the system in the time. The most commonly used time integrator in MD is still the Verlet algorithm, which is fairly simple to implement [35]. Its basic form can be easily derived from a forward and backward Taylor expansion of the position vector:

$$\mathbf{r}(t+\Delta t) = \mathbf{r}(t) + \Delta t \cdot \mathbf{v}(t) + \frac{\Delta t^2}{2} \cdot \mathbf{a}(t) + \frac{\Delta t^3}{6} \cdot \mathbf{b}(t) + O(\Delta t)^4 \tag{2.20}$$

$$\mathbf{r}(t-\Delta t) = \mathbf{r}(t) - \Delta t \cdot \mathbf{v}(t) + \frac{\Delta t^2}{2} \cdot \mathbf{a}(t) - \frac{\Delta t^3}{6} \cdot \mathbf{b}(t) + O(\Delta t)^4 \tag{2.21}$$

Summing the two equations, the odd derivatives cancel out; rearranging, we get

$$\mathbf{r}(t+\Delta t) = 2\mathbf{r}(t) - \mathbf{r}(t-\Delta t) + \frac{\Delta t^2}{2} \cdot \mathbf{a}(t) + O(\Delta t)^4 \tag{2.22}$$

The acceleration can be calculated easily as

$$\mathbf{a} = \frac{\mathbf{f}}{m}$$

It is remarkable that this comparatively cheap scheme achieves an error $\sim O(\Delta t)^4$. The drawback is that velocities are not calculated explicitly. They are not needed for the calculation of the trajectory, but they are necessary to calculate the kinetic energy K. The conservation of the total energy $E = K + V$ is an important validity check for a simulation. Extrapolating the velocity from the positions at two different timesteps, for example as a central difference

$$\mathbf{v}(t) = \frac{\mathbf{r}(t+\Delta t) - \mathbf{r}(t-\Delta t)}{2\Delta t} \tag{2.23}$$

would introduce an error $\sim O(\Delta t)^2$. For this reason, the integrator usually implemented (i.e. in LAMMPS) is the so-called Velocity Verlet Algorithm [38]; position and velocities

for every particle are calculated as follows:

$$\mathbf{r}(t+\Delta t) = \mathbf{r}(t) - \Delta t \cdot \mathbf{v}(t) + \frac{\Delta t^2}{2} \cdot \mathbf{a}(t) \quad (2.24)$$

$$\mathbf{v}(t+\frac{\Delta t}{2}) = \mathbf{v}(t) + \frac{\Delta t}{2} \cdot \mathbf{a}(t) \quad (2.25)$$

$$\mathbf{a}(t+\Delta t) = -\frac{1}{m}\nabla \mathbf{v}(\mathbf{r}(t+\Delta t)) \quad (2.26)$$

$$\mathbf{v}(t+\Delta t) = \mathbf{v}(t+\frac{\Delta t}{2}) + \frac{\Delta t}{2} \cdot \mathbf{a}(t+\Delta t) \quad (2.27)$$

Another algorithm that can handle velocities is the *leap-frog* scheme, here omitted for brevity and described for example in [23, 39]. Its accuracy is the same as the Velocity Verlet and should hence produce the same trajectories.

2.2.3 Energy conservation issues: timestep length and constraining of fast degrees of freedom

As mentioned in earlier paragraphs, the generation of a MD trajectory is performed by a "time marching" numerical integration of newtonian equations of motion in time. The physical goodness of the chosen simulation parameters, in terms of energy conservation, is usually assessed by monitoring the behaviour of the system total energy over time, during a (suitably long) simulation in the NVE ensemble. We recall here that an NVE ensemble mimics an adiabatic system where no mass exchange take place. In absence of dissipative effects, the time evolution of the system is completely determined by a conservative force field generated by the chose pairwise interaction potential. Therefore the particles of the system continuously exchange energy (which constantly but the total energy must be a conserved quantity, with the exception of small fluctuations due to numerical and round-off errors [23, 24]. The choice of the timestep is crucial in order to ensure correct energy conservation. Empirically, a good conservation is achieved when the timestep is about $1/10$ of the shortest oscillation period found in the system. For biomolecules in aqueous environment, the fastest motion is the stretching of the covalent bonds involving H atoms, which vibrate very fast because of H small mass. The oscillation period is about 10 fs [40], which would limit the timestep to approx. 1fs. However, for practical purposes, a shorter timestep is a big inconvenience, because it implies that the simulation of the same physical amount of time requires a higher number of steps, and therefore a higher computational cost. Therefore it has become common praxis to apply

holonomic constrains the the bonds involving hydrogens, making them rigid. The first numerical algorithm to successfully implement rigid bonds was SHAKE [41], usually in combination with a Verlet integration scheme, which is also the approach used in all the simulations mentioned in this thesis. After the constraining bonds and angles involving hydrogens, the timestep can be increased up to approx. 2fs, still maintaining good energy conservation. Some authors have also suggested the possibility to further increase the timestep by artificially increasing the mass of H atoms, but this approach has not been widely adopted [40].

2.2.4 Avoiding surface effects: periodic boundary conditions

The well-known high computational cost of MD simulations limits the size of tractable systems to the nanometric scale [23, 24]. For such small systems, the surface effects would be enormous. In order to investigate bulk phenomena and avoid surface effects, it would be necessary to build systems too large to be computationally affordable. This problem is circumvented by using Periodic Boundary Conditions (PBC). The simulation box is surrounded in every direction (or some specific direction) by an infinite array of its mirror images, as depicted in fig. 2.2.4.

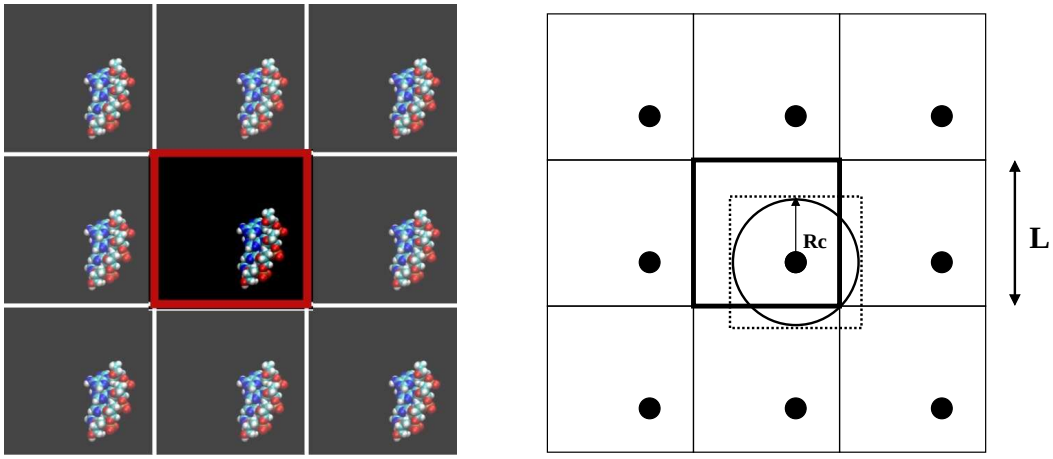


Figure 2.2: Periodic images of a solvated biomolecule (left) and an illustration of the Minimum Image Criterion (right).

A particle that leaves the box is replaced automatically by its image entering from the opposite side. In other words, upon crossing of a boundary, the coordinates of the particle

are modified by adding or subtracting the linear size of the box along the axis perpendicular to the crossed boundary surface. When computing forces on the *i*-th particle, care must be taken to include the interactions coming from particles that belong to the periodic images [42]. This would apparently need to a infinite increase of the involved particles, and hence of the complexity. It is easy to demonstrate that if the smallest dimension of the simulation box is $> 2R_c$, every particle will interact *at most* with one periodic image [24]. This rule is called "Minimum Image Criterion" and ensures that the impact of PBC on system complexity is limited. Usually, the periodicity is enforced along 2 or 3 dimensions, depending on the type on conditions one wants to simulate.

2.2.5 Treatment of long-range electrostatics

The definition of "long range" in MD refer to those interactions that decay with distance as $1/r^{d-1}$, where d is the number of dimensions of the system [24]. The comparatively slow decay rate implies that, unlike the case of Lennard-Jones interactions, the use of truncated potentials is not adequate, because each charged particle is actually interacting with many of its periodic images. The calculation of the electrostatic forces acting on the particles requires the solution of the Poisson equation:

$$\nabla^2 \mathbf{E}(\mathbf{r}) = -\frac{\rho(\mathbf{r})}{\epsilon_0}, \tag{2.28}$$

where $\mathbf{E}(\mathbf{r})$ is the electric field, $\rho(\mathbf{r})$ is the charge density and ϵ_0 is the dielectric constant of the void. For a simulation box with periodic boundary conditions, however, the problem is not simple, because all mirror images generate their own contribution to the total field. The resulting field is that generated by a periodic 3D lattice having the box as elementary cell. However, the theoretical tools for the calculation of the total electrostatic field were already available thanks to the work of Ewald and Madelung [43, 44]. Their methods were adapted to MD simulations under the name of "Ewald summation", and subsequently modified for better numerical efficiency, under the name of Particle-Mesh Ewald (PME) [45], and Particle-Particle Particle-Mesh (PPPM) [46]. The success of the Ewald summation and related methods was crucial in allowing the simulation of highly charged molecules, such as nucleic acids, where the contribution of electrostatic interactions is relevant.

2.2.6 Simulation of different ensembles: thermostatting and barostatting

The most "natural" ensemble for a MD simulation is the NVE, with constant number of particles, constant volume and constant energy, This is equivalent to an adiabatic system where no mass exchange takes place. In this ensemble, however, the systems samples an isoenergetic surface of its phase space, where all states have the same energy and therefore, according to the Boltzmann equation, the same probability [30]. This is however, a rather irrealistic situation in comparison to the conditions under which most macroscopic experiments are performed (typically, at constant temperature and pressure). The correct simulation of ensembles different from the microcanonical is non trivial, and required a considerable amount of research effort to be tackled with success. The first temperature control, and a rather crude one, was a velocity rescaling where every particle's velocity was modified in order to achieve an average which was consistent with Eq. 2.10. Andersen was the first to develop a method of pressure control where the velocity corrections were performed sampling random values from a prescribed distribution [47], a route also followed by Berendsen some years later in order to control temperature [48]. All these methods, however, suffered from a major drawback: while they were very efficient in steering the system to the desired levels of temperature and pressure, they altered the original Newtonian (or Hamiltonian) equations of motion, producing an unknown bias in the statistical ensemble. Therefore, the only statistically sound use of these thermostats and barostats was to drive the system to the desired state, and then switch them off and run a NVE simulation [21]. The first solution to this problem came with an entirely different and deterministic approach, developed by Nose and then perfected by Hoover [49, 50]. The main advantage of their method, suitable for NVT and NpT ensemble simulations, is the correct reproduction of the system partition function, which makes them suitable for simulations that aim at the calculation of thermodynamic properties for which correct phase space sampling in relevant.

The Nose-Hoover thermostat

The original work by Nose [49] was highly praised by Hoover, who was probably the first to understand its potential, and went on the develop and simplify the original thermostatted equations of motion [51]. Nose, in turn, credited Andersen ([47]) with the idea of

controlling the thermodynamic state of the system by using a *deterministic* feedback control, generated by a modified Hamiltonian. Two additional artificial conjugated degrees of freedom (s, p_s) are added to the system. The variable s acts as (in Nose's words) "an external system for the physical system of N particles", whose energy is related to the conjugated momentum p_s, and a particular choice of the potential energy function for s, $\Phi(s) = N_d k_B T \ln s$. The extended Hamiltonian has the form:

$$\mathcal{H}_{Nose} = \left[\frac{p_s^2}{2M}\right] + \sum_{i=1}^{N}\left[\frac{p^2}{2ms^2}\right] + \Phi(\mathbf{q}) + N_d k_B T \ln s \tag{2.29}$$

where the terms $(p_s^2/2M)$ and $(N_d k_B T \ln s)$ can be interpreted as the kinetic and potential energy associated to the additional external system. The parameter M has the dimensions of $energy \cdot (time)^2$, and behaves as a "mass" for the motion of s, Φ is the potential energy, k_B is the Boltzmann constant, T is the absolute temperature and N_d the number of degrees of freedom of the physical system. The resulting equations of motion are:

$$\frac{\partial q}{\partial t} = \frac{p}{ms^2} \tag{2.30}$$

$$\frac{\partial p}{\partial t} = F(\mathbf{q}) \tag{2.31}$$

$$\frac{\partial s}{\partial t} = \frac{p_s}{M} \tag{2.32}$$

$$\frac{\partial p_s}{\partial t} = \sum\left[\frac{p^2}{ms^3} - \frac{k_B T}{s}\right]. \tag{2.33}$$

Nose intepreted s as a "time scaling factor". However, in this case it's physical meaning becomes rather elusive, as pointed out by Hoover ([50]). Nose was able prove that set of equations given above can correctly sample the Canonical ensemble NVT, assuming that the system is ergodic and the times averages keep into account the time rescaling [1][49]. The crucial part of his approach was the choice of the added potential term $\Phi(s) = N_d k_B T \ln s$. Hoover's contribution was the realisation that s is actually decoupled from the dynamics, and it is therefore possible to derive a set of equations of motion equivalent to Nose's, where the "obscure" time scaling factor s does not appear [52]. Hoover's version of the thermostatted equations of motion, which is clearer and had a pivotal role

[1] A system is called "ergodic" when it can sample all the microstates of its phase space. Since this is usually not feasible, what is sought for practical purposes is the "quasi ergodicity", where a system's trajectory will be reasonably close to any given microstate. See for example [30] for a thorough discussion of ergodicity and its implications for statistical mechanics.

in popularising the method, is the following:

$$m\frac{\partial p}{\partial t} = F(\mathbf{q}) - \zeta p \tag{2.34}$$

$$\frac{\partial \zeta}{\partial t} = \left(\frac{K}{K_0} - 1\right)/\tau^2 \tag{2.35}$$

$$K_0 = N_d \cdot k_B T \tag{2.36}$$

$$\zeta = p_s/(N_d \cdot k_B T \tau^2) \tag{2.37}$$

$$M = N_d \cdot k_B T \tau^2 \tag{2.38}$$

where M is the "mass" of the external system, K and K_0 are the current and target kinetic energy, and τ is a relaxation time (a bigger relaxation time decreases the time derivative of the friction coefficient, i.e. the system will be steered to the target temperature more slowly [2]). The time derivative $\partial\zeta/\partial t$ of the friction coefficient ζ, driven by the dicrepancy between the current and target values of the system's kinetic energy, implements a negative feedback that stabilises the temperature. A decade later, Hoover and Dettmann [53] discovered that such equations could also be derived from the Hamiltonian:

$$\mathcal{H}_{Dettmann} = s \cdot \mathcal{H}_{Nose} \tag{2.39}$$

$$= s \cdot \left[\frac{p_s^2}{2M}\right] + s \cdot \sum_{i=1}^{N}\left[\frac{p^2}{2ms^2}\right] + s \cdot \Phi(\mathbf{q}) + s \cdot N_d k_B T \ln s. \tag{2.40}$$

Over the years, the Nose-Hoover equations were generalised for systems of changing size and shape in the isothermal-isobaric NpT ensemble, [51, 54] and also for system that include holonomic constraints, such as rigid bonds [52]. LAMMPS' thermostatting and barostatting algorithms incorporate such developments.

2.2.7 Coarse-Graining: principles and methods

The purpose of this section is to illustrate the general principles and ideas behind the construction of coarse-grained models. The process of coarse-graining has the goal of reducing the level of detail at which the system is simulated, but retaining a correct description of some properties of interest. In molecular simulations, this is usually performed by representing whole groups of atoms with a single point mass, or "site", equipped with

[2] For biomolecular simulations, we empirically found that, for effective thermostatting, a sound value of τ is usually in the order of 100 timesteps.

its own interaction potential. The description of this "global" interaction potential is the main difficulty of the modelling process. This section illustrated some of the ideas and methods which have been used in this area. Eventually, a more detailed treatment is reserved to the Multiscale Coarse-Graining method (MSCG), which has been adopted and implemented in this thesis.

Helmholtz free energy

The concept of free energy has had a central importance in the theoretical fundamentals of coarse-graining (as well as in other fields, not treated here). For a purely mechanical system, the external work L done by the system is given only by the opposite of the corresponding change of internal energy ΔU,

$$L = -\Delta U$$

In the case of thermodynamic systems, the relation must take into account the heat exchange, and we get, according to the first law of thermodynamics

$$L = -\Delta U + Q, \tag{2.41}$$

If we consider now a transformation between two states 1 and 2, for a system that exchanges heat with a source at constant temperature T, we know from the second law of thermodynamics that the corresponding change of entropy can be calculated as

$$\int_1^2 \frac{\delta Q}{T} \leq S(1) - S(2), \tag{2.42}$$

where the equality holds in case of reversible process. Being the temperature constant,

$$Q = \int_A^B \delta Q \leq T[S(2) - S(1)], \tag{2.43}$$

Combining this expression with equation 2.41 and defining

$$\Delta U = U(2) - U(1)$$

we get

$$L \leq U(1) - U(2) + T[S(2) - S(1)] \quad or \quad L \leq -\Delta U + T\Delta S \tag{2.44}$$

This inequality sets the *upper boundary* to the work that can be extracted from the transformation 1-2. We can define a quantity

$$A = U - TS, \tag{2.45}$$

which is called *Helmholtz free energy*, and see that, at constant temperature

$$L \leq A(1) - A(2) = -\Delta A, \tag{2.46}$$

Which means that in the case of an isothermal reversible transformation, the work equals the opposite of the change of the internal energy. If the process is irreversible, *the free energy change is the upper limit for the exchanged work.* This extremely general result of classical thermodynamics has been exploited for the modelling of molecular systems. In particular, the free energy difference between states can be used (although with some limitations) to estimate the *potential energy change* associated with a conformational change of the system, as explained in the next section.

Potential of Mean Force

The Potential of Mean Force (PMF) is a concept first introduced by Kirkwood [55]. The idea is to extract the dependency of the distribution function from a certain coordinate by integrating the probability distribution over all remaining coordinates. From this relation, properties that depend on the configuration integral can be extracted, as a function of the same generalized coordinate.
As seen in the previous chapter, the Helmholtz free energy change is associated to the reversible work done on the system during a transformation between two equilibrium states. A popular approach is to use it in order to estimate the potential energy associated with a conformational change. Early examples of this kind of study can be found for example in the work by Norberg [56–59]. It must noted however, that the Helmholtz energy is a *free* energy and not a potential energy, as it embeds entropic effects and interactions with the environment (for instance with the solvent). The approximation, in principle, is only correct when the entropic contribution is negligible. For instance, when coarse graining a bead-rod polymer model, the free energy typically works well for estimating the effective bond stretching, but less well for torsional or bending degrees of freedom; however, many authors have pursued this route [6, 60], for example fitting a Lennard-Jones-like potential

on the free energy profile, as for example in [19].

Harmonic approximation for bonded interaction

This alternative route to the parametrization of bonded interactions from all-atom simulations has been exploited recently by Arkhipov and coworkers, for a very coarse model of a bacterial flagellum [19]. First, the position of the CG beads is mapped on the all-atom model. Secondly, the assumption is made that CG bonds behave as independent harmonic oscillators. The basic example is the monodimensional oscillator, whose potential energy can be written as

$$U(x) = \frac{1}{2}K_x(x - x_0)^2, \tag{2.47}$$

and the two parameters, K_x and x_0 can be estimated from atomistic simulations as

$$K_x = \frac{RMS\,D(x)}{k_B T},$$

and

$$x_0 = \langle x \rangle,$$

The first relation, the estimation of average length x_0 by means of $\langle x \rangle$, is trivial. The derivation of the second comes from the theory of Brownian motion in a harmonic potential, and can be found for example in [61]. These relations can be extended to the case of bonds and angles. This approach requires however careful consideration when the number of backbone bonds per Kuhn step is low [3], according to some observations on the statistical mechanics of entropic springs, as discussed by Larson, and later by Underhill [62–66]. Moreover, the structural equilibration of very flexible biomolecules like ssDNA can be non-trivial to assess, as pointed out by [8, 9, 28, 67].

2.3 Computation of effective potentials from all-atom simulations

This section presents a concise review of some popular methods for PMF calculation. They can be divided into two main categories, equilibrium and non-equilibrium methods.

[3] The Kuhn step is defined as twice the persistence length of the polymer chain, which is in turn the length scale over which the tangent vector's autocorrelation function decays to zero. It's a very important parameter in bead-rod and bead-spring models of polymers, and it's related to the stiffness of the chain [61]

Most non-equilibrium methods are based on the so-called Umbrella Sampling (US), by Torrie and Valleau ([68], and the Bennet Acceptance Ratio method (BAR). Over three decades, much effort has been put into the refinement of the statistical estimators used to quantify the interaction parameters, which led, for the US, to the introduction of the histogram method by Ferremberg [69], followed by the popular Weighted Histogram Analysis Method (WHAM) proposed by [70]. The approach was then extended to multidimensional energy profiles (along more than one coordinate) by subsequent work by Roux [71] and Souaille [72]. A similar route was followed for the BAR method, expanded twenty years later by Shirts and Chodera [73] into the Multistate BAR (MBAR), again by essentially refining the statistical treatment of the samples collected from equilibrium MD simulations.

For what concerns non-equilibrium methods, the Steered Molecular Dynamics approach (SMD) was made possible by the discovery of the Jarzynski equality [74], for the free energy difference over non-equilibrium processes. There exist two main approaches to SMD, developed by Hummer and Szabo [75] and by Izrailiev, and Park and Schulten [76].

The main downside of both US and SMD methods, is that they calculate a PMF over a specific interaction coordinate (usually a distance between two molecular sites). Therefore, characterising the interactions of complicated molecules would require a very high number of simulations.

More recently, however, a force-matching equilibrium method have been introduced by Izvekov and Voth [1, 11], building upon a previous work by Ercolessi [21] for the construction of MD potentials from ab-initio Molecular Dynamics data. This work was then further refined by Noid [17, 18], who gave the method a more solid theoretical footing. He also demonstrated that, under a certain set of hypotheses (which will be discussed later on), the calculated PMF is optimal with respect to the available simulated data. The force-matching approach is by far the more practical when numerous parameters of the CG potential are to be parametrised simultaneously, and it's the method that has been chosen and implemented for the CG model developed in this work.

2.3.1 Boltzmann inversion

The Boltzmann inversion is a technique that allows the extraction of effective interaction potentials for coarse grained models, from a radial distribution function. The final output is an effective tabulated potential (not an analytical one). The drawback of numerical

potentials is the lack of parameters to which we can associate a physical interpretation [13].

The prerequisite is the availability of a real RDF usually obtained by simulation, and in the case of ssDNA this issue will be addressed in the last section of this report. The method is derived observing that, if we use the distance between two particles as the reaction coordinate, the expression of the Helmholtz free energy in terms of ensemble average (up to an arbitrary additive constant) becomes [77]

$$A(r) = -k_B T ln[g(r)] + const, \tag{2.48}$$

where $g(r)$ is the radial distribution function (RDF). The potential of mean force along r can be expressed as

$$A_1(r) - A_0(r) = -k_B T ln\left[\frac{g_0(r)}{g_1(r)}\right]. \tag{2.49}$$

Initially developed in the case of fluid particle distributions [14], the method has been then applied to polymer coarse graining, for example by [13], with the name of *iterative Boltzmann inversion* for the purpose of polymer coarse graining. The coarse-grained (and hence computationally cheap and rapidly converging) simulation starts with first-guess potential $V_0(r)$ that generates a starting $RDF_0(r)$.

$$V_1(r) = V_0(r) + k_B T ln\frac{RDF_0(r)}{RDF_{target}} \tag{2.50}$$

and then proceeds with further optimization steps

$$V_{n+1}(r) = V_n(r) + k_B T ln\frac{RDF_n(r)}{RDF_{target}} \tag{2.51}$$

This approach assumes uses as correction factor the PMF between the current correlation function and the target one. The convergence criterion, assuming that the desired RDF is computed up to certain cutoff r_c, is defined by means of a penalty function, usually the integral square error [6]:

$$p = \int_0^{r_c} w(r)(RDF_{target} - RDF_n)^2 dr \tag{2.52}$$

The convergence is usually achieved in about ten steps [13]. The biggest issue with Boltzmann inversion is that it requires the knowledge of the radial distribution function, which

is not always available. However, the method has been successfully used by Trovato and Tozzini [10, 78] for a “on bead per residue“ model of DNA, exploiting a collection of DNA structures available in the Protein Data Bank.

2.3.2 Umbrella Sampling

In molecular dynamics simulations of biomolecules, it happens very frequently that the energy landscape presents local minima, that corresponds to what we can call preferential configurations. Such configurations may be separated by energy barriers high enough to "trap" the system thus preventing an effective exploration of the phase space within the accessible simulation timescale. This can be better understood considering the expression:

$$A(r) = -k_B T lnC + const, \tag{2.53}$$

where we can see the logarithmic relation between the free energy and the radial distribution function. Typically, the free energy change is the order of several $k_B T$, and this means that a very thorough sampling of the phase space (which would ensure reliable results) would require a change of several orders of magnitude in the configurational integral. On the affordable simulation timescale it is extremely unlikely (practically, impossible) that this will happen spontaneously by letting the system fluctuate freely. The system would spend most of the simulated time in high-probability microstates, and the resulting estimates of the ensemble averages would be poor.

The most immediate solution would be to use external forces to actively steer the system into a desired configuration (or series of configurations), recover good estimates for each small portion of phase space, and then piecewise reconstruct the whole energy profile. However, the introduction of external forces (e.g. harmonic springs that actively pull the system into the desired conformation) introduce an extra (non-physical) potential energy term in the Boltzmann factor of the system, producing a bias in the probability distribution, and therefore generating a different ensemble, thus invalidating the resulting statistics. One may ask if it’s not somehow possible to manipulate the perturbed ensemble averages in order to recover the properties of the unperturbed system.

The answer is affirmative, and a solution to this problem was devised by Torrie and Valleau [68], in a pioneering work that has had a lasting influence. The key idea behind their method is to use a clever and straightforward analytical manipulation, in order to recover the correct unbiased statistics for a biased simulation. Be $\mathbf{x}$ the configuration of the

system within a phase space Γ, and be $A(\mathbf{x})$ a property that depends on the configuration (positions and velocities). The ensemble average $\langle A \rangle$ is given by definition by:

$$\langle A \rangle = \frac{\int_\Gamma A(\mathbf{x}) \cdot exp(-U(\mathbf{x})/k_B T) d\mathbf{x}}{\int_\Gamma exp(-U(\mathbf{x})/k_B T) d\mathbf{x}}, \tag{2.54}$$

which indicates a weighted average of the "instant" value of A for a given state $\mathbf{x}$, weighted with the probability of the state $\mathbf{x}$. Where as usual T is the absolute temperature and k_B the Boltzmann constant. Now we assume to have a desired perturbing potential $w(\mathbf{x})$ that constrains the system in a close neighbourhood of the region of phase space we are interested in. Eq. 2.54 can be rewritten as

$$\langle A \rangle = \frac{\int_\Gamma (w(\mathbf{x})/w(\mathbf{x})) \cdot A(\mathbf{x}) \cdot exp(-U(\mathbf{x})/k_B T) d\mathbf{x}}{\int_\Gamma (w(\mathbf{x})/w(\mathbf{x})) \cdot exp(-U(\mathbf{x})/k_B T) d\mathbf{x}}, \tag{2.55}$$

and the terms can be rearranged as (omitting the dependency on the configuration, to make the notation lighter)

$$\langle A \rangle = \frac{\langle \frac{A}{w} \rangle_w}{\langle \frac{1}{w} \rangle_w}, \tag{2.56}$$

where the $\langle \cdot \rangle_w$ indicates the ensemble average over the perturbed system, with an extra non-Boltzmann weighting factor $w(\mathbf{x})$: that is to say, the average of the values "as we get them" from our biased simulation. The problem of recovering the unperturbed ensemble average is solved, at least in principle.

Operatively, assuming that we have taken N snapshots from the biased simulation, the ensemble average of any function A of the coordinates, described in equation 2.56, is computed as:

$$\langle A \rangle_0 = \frac{\sum_{i=1}^N A_i / w_i}{\sum_{i=1}^N 1/w_i}, \tag{2.57}$$

where the biasing probability factor w_i can be derived from the constraining potential. Assuming that we have constrained a reaction coordinate ξ (function of the system's coordinates) to a certain value ξ_0 by means of a harmonic potential

$$V(r) = \frac{1}{2} k (\xi - \xi_0)^2,$$

this potential will be added to the system Hamiltonian and will influence the energy and

hence the probability distribution. The new probability density in the constrained ensemble will be

$$p(\mathbf{x}_w) = \frac{1}{Z} exp\left(\frac{-U(\mathbf{x}) + V(\xi)}{k_B T}\right) = \frac{1}{Z} exp\left(\frac{-U(\mathbf{x})}{k_B T}\right) exp\left(\frac{-V(\xi)}{k_B T}\right) \tag{2.58}$$

where the term $exp(-V(\xi)/k_B T) = w_i$ is the sought statistical bias introduced by the constraining potential. The original Umbrella Sampling was devised for trajectories produced by Markov chains, i.e. Monte Carlo simulations, but the same formalism remains valid in the case of snapshots taken from a Molecular Dynamics run, provided that the sampled timesteps are sufficiently far apart to be considered uncorrelated (see section 2.3.6). The aforementioned approach has paved the way to a number of further studies, many of which are essentially a numerical refinement of the original idea. Since the introduction of the Umbrella Sampling approach, one problem became immediately apparent, namely the proper choice of the biasing potential [79]. The US method performs at its best when the biased energy profile is flat, so that all biased microstates have the same probability. In this situation all microstates would have the same probability resulting in a uniform sampling of the phase space. However, this would require a biasing potential which is exactly the opposite of the sought PMF along the investigated coordinate, which is (unfortunately) precisely the unknown of the problem. A possible iterative solution can be found in the work of Mezei [80], who introduced the Adaptive Umbrella Sampling technique, where an initial guess for the biasing potential is iteratively refined.
The second operative problem of the US method is how to optimise the energy profile reconstruction from simulating narrow “windows” of the desired reaction coordinate. A successful a widely used solution is described in the next section.

2.3.3 The Weighted Histograms Analysis Method (WHAM)

The WHAM algorithm is an extension of the umbrella sampling approach that optimizes the averages obtained by a series of runs. It was initially proposed by Kumar [70], who reprised the work of Ferrenberg and Swensen [69, 81], and further developed over the years by other authors [72, 82]. In order to sample over a certain excursion of the reaction coordinate, it is useful to split the simulation into partially overlapping "windows", with different biasing potentials that restrict the conformation to specific positions along the reaction path. This allows a "piecewise" sampling of the conformational change, that facilitates the exploration of the corresponding phase space regions of interest [71]. The

WHAM method was conceived as a sensible procedure for the reconstruction of the whole energy profile. If we perform a set of biased simulations with a perturbed potential

$$V_0(\mathbf{r}) + w_i\xi(\mathbf{r}),$$

where w_i is the usual perturbing potential, and $\xi = \xi(\mathbf{r})$ is a scalar variable defined as a function of the configuration of the system (for simple cases, it could be a distance between two specific atoms, or the radius of gyration of a set of atoms, etc.). From these simulations a set of (biased) probability densities $\rho_i^{(b)}(\xi)$ can be obtained constructing an histogram of the values taken by ξ during the simulations, with "bins" (width of the histogram bars) of opportune size. The unbiased distribution is recovered by "removing" the non-Boltzmann part of the probability factor [70]:

$$\rho_i^{(u)}(\xi) = exp\left(\beta[w_i(\xi) - f_i)]\right)\rho_i^{(b)}(\xi), \tag{2.59}$$

where f_i is the free energy contribution coming from the introduction of the perturbing potential, $\rho_i^{(b)}$ and $\rho_i^{(u)}$ are the biased and unbiased probability densities, respectively. The final goal is to recover ρ_0 having at our disposition $\rho_0^{(u)}$. The key idea behind WHAM is to write ρ_0 as a linear combination of the unbiased "window" distributions $\rho_i^{(u)}(\xi)$,

$$\rho_0(\xi) = C\sum_{i=1}^{N} c_i\rho^{(u)}(\xi), \tag{2.60}$$

where C is a normalization constant and c_i a set of normalized weights, so that

$$\sum_{i=1}^{N} c_i = 1.$$

The WHAM equations are then recovered by *minimizing the variance of the total probability distribution* i.e. looking for a stationary point

$$\frac{\partial\sigma^2[\rho_0(\xi)]}{\partial c_i} = 0, \tag{2.61}$$

The minimum of the constrained function can be found for example with the method of Lagrangian multipliers [71]. The resulting equations are:

$$\rho_0(\xi) = C \sum_{i=1}^{N} \frac{n_i \cdot exp(-\beta[w_i(\xi) - f_i])}{\sum_{j=1}^{N} n_i \cdot exp(-\beta[w_i(\xi) - f_j])} \rho_i^{(u)} \tag{2.62}$$

$$= C \sum_{i=1}^{N} \frac{n_i}{\sum_{j=1}^{N} n_i \cdot exp(-\beta[w_i(\xi) - f_j])} \rho^{(b)}(\xi) \tag{2.63}$$

for the probability density, and

$$exp(\beta f_k) = \int exp\left(\frac{-w_i(\xi)}{k_B T}\right) \rho_0 \xi d\xi \tag{2.64}$$

for the free energy. The arbitrary constant C can be dropped if we assume the probability distribution to be normalised. The unknowns in Eq. 2.62 are $\rho_0(\xi)$ and f_i. The set of non-linear equations can be solved iteratively as shown by Roux [71]: starting from an initial guess for the N free energies f_i, equation 2.62 is used to estimate the unbiased distribution, this distribution is then used in Eq. 2.64 to generate the next approximation for the f_i, that are in turn fed back into equation 2.62. The iteration loop is repeated until satisfactory convergence is achieved [72].

Although the presented approach is the one for free energies, it must be noted that no assumption is made about the nature of the reaction coordinate $\rho(\xi(\mathbf{r}))$, so that the same relations hold also in the case of $\rho = (r)$, which is useful for computing averages of arbitrary quantities from the same biased simulations [72]. Currently, several different implementations of WHAM are freely available in the public domain. A general remark for PMF calculations is that most methods assume that the simulations are performed in the NVT ensemble. In such cases where it is mandatory to control the temperature by means of a Nose-Hoover thermostat, as a Berendsen thermostat (or worse, velocity rescaling) wouldn't produce the correct ensemble [49].

2.3.4 The Multistate Bennett Acceptance Ratio method

This method, introduced only very recently, generalizes the Bennett Acceptance Ratio method [83]. It is deemed by its authors to be more efficient than WHAM, removing the necessity of histograms and allowing a straightforward estimation of the uncertainties on

the computed averages.

Our observations (or snapshots of a simulation) behave as a random variable, and we recover expected values and variance by means of opportune averages over the samples we have at hand. However, these averages are sums of random variables, and therefore *behave as random variables themselves*, with their own mean and variance. The mean value is our target, and the variance is the uncertainty (dispersion of measurements around the mean value). The MBAR method constructs estimators with the lowest asymptotic variance (variance in the large sample limit), therefore more accurate, borrowing from recent developments in the field of inferential statistics: the work by Kong [84] has tackled the issue of a systematic analysis of the construction of optimal estimators, while Tan [85] demonstrated that the derived estimators were "optimal" in terms of variance (i.e., they had the lowest variance or all other known estimators).

The starting problem is how to estimate a free energy difference between different molecular configurations, from the data generated by the simulations of multiple equilibrium states. Let's assume we have simulated K different states, and for the i-th state ($i = 1..K$) we have sampled N_i statistically independent snapshots. The configurations $\{\mathbf{x_{in}}\}_{n=1}^{N_i}$ are sampled from an ensemble where the probability distribution ρ_i of a state i is given by

$$\rho_i(\mathbf{x}) = \frac{\hat{\rho}_i(\mathbf{x})}{Z_i}, \tag{2.65}$$

where q_i is the unnormalised probability density, and Z_i is the normalisation factor, which is the partition function for the state i:

$$Z_i = \int_\Gamma d\mathbf{x}\hat{\rho}_i(\mathbf{x}), \tag{2.66}$$

where Γ is the phase space of the system (positions and momenta). It is easy to recognise that the unnormalised would be equal to the Boltzmann factor as defined by Eq. 2.5. The goal is to produce an estimator of the dimensionless free energy $f = A/(k_B T)$. Shirts and Chodera [73], following the approach outlined by Tan [85], have derived the following equations for the dimensionless free energy, called *extended bridge sampling estimators*, for the dimensionless free energy, calculated from a canonical ensemble:

$$\hat{f}_i = -ln \sum_{k=1}^{K} \sum_{n=1}^{N_k} \frac{exp[-V_i(\mathbf{x}_{kn})]}{\sum_{k\prime=1}^{K} N_{K\prime} exp[\hat{f}_{k\prime} - V_{k\prime}(\mathbf{x}_{kn})]} \tag{2.67}$$

where $k = 1 \dots K$ is the number of simulated states ("windows") and $n = 1 \dots N_k$ the number of uncorrelated snapshots (configurations) stored for the $k-th$ state. The system of Eq. 2.67 can be solved iteratively as illustrated by Shirts and Chodera [73]. The free energy is determined up to an additive constant so that only the difference between two energies is physically meaningful.

The uncertainty on the calculated values of the adimensional free energy can be quantified as follows. Be $N = \sum_{k=1}^{K} N_k$ the total number of snapshots; the covariance matrix is given by

$$\hat{\mathbf{\Theta}} = \mathbf{W}^T(\mathbf{I}_N - \mathbf{W}\mathbf{N}\mathbf{W}^T)^+\mathbf{W} \tag{2.68}$$

where: the superscript "+" indicates a generalized inverse, as the matrix might be singular; $\mathbf{I}_N$ is a $N \times N$ identity matrix; $\mathbf{N} = diag(N_1 N_2 ... N_K)$ and $\mathbf{W}$ is a $N \times K$ matrix of weights, whose terms are defined as

$$W_{nk} = \frac{\hat{c}_k^{-1} q_k(\mathbf{x}_n)}{\sum_{k\prime=1}^{K} N_{k\prime}\, \hat{c}_{k\prime}^{-1}\, \hat{\rho}_{k\prime}(\mathbf{x}_n)}, \tag{2.69}$$

where the subscript n runs from 1 to N and the parameters $\hat{c}_{k\prime}$ can be calculated self-consistently from the taken snapshots [73].

The covariance of two arbitrary functions $\phi(\theta_1 ... \theta_K), \psi(\theta_1 ... \theta_K)$ can be calculated as:

$$Cov(\hat{\phi}, \hat{\psi}) = \sum_{i,j=1}^{K} \frac{\partial \phi}{\partial \theta_i} \hat{\Theta}_{ij} \frac{\partial \psi}{\partial \theta_j}, \tag{2.70}$$

and this relation can be used for the computation of the uncertainties: in the large sample limit, the errors can be assumed to be Gaussian-distributed, therefore their dispersion about the mean value is well estimated by their variance. Remembering that for a random variable X, $Cov(X, X) = Var(X)$, we can write the variance of the free energy difference between two states as

$$Var(\Delta f_{ij}) = \langle (\Delta \hat{f}_{ij} - \langle \Delta \hat{f}_{ij} \rangle)^2 \rangle \tag{2.71}$$

$$= Cov\left(-ln\frac{\hat{c}_i}{\hat{c}_j}, -ln\frac{\hat{c}_i}{\hat{c}_j}\right) \tag{2.72}$$

$$= \hat{\Theta}_{ii} - 2\hat{\Theta}_{ij} + \hat{\Theta}_{jj}, \tag{2.73}$$

which provides a convenient way to estimate the accuracy of the computed values and therefore the extent to which the results can be trusted.

2.3.5 Steered Molecular Dynamics (SMD) and Jarzynski equality

The method of PMF calculation through SMD has been reviewed, among others, by Hummer and Szabo [32, 75] and by Park and Schulten [76] to whose approach the following section refers. The evolution of the system is steered along the desired reaction coordinate by an external force, usually large, that allows to overcoming energy barriers easily. Or in other words, the conformational change is enforced by means of opportune harmonic potentials, that quickly bring the system in the desired configuration. At first sight, what described in the previous lines may sound very similar to the use of perturbing potential in equilibrium Umbrella Sampling simulations. Care must be taken to notice that US simulations use harmonic forces to *constrain* the system in a certain configuration, around which the system is allowed to fluctuate by effect of thermal agitation, whilst SMD uses perturbing potentials in order to *steer* the system, rapidly, thorugh a desired configurational change. The result is that US simulation are performed in conditions of thermodynamic equilibrium, whilst SMD simulations are not. The steering force used in SMD makes the simulation an intrinsically non-equilibrium process, but the PMF is in itself an equilibrium property, associated to a state function (the free energy). The theoretical bridge required to overcome this apparent contradiction has become available only in recent times (1997) with the discovery of Jarzynski equality (JE), which has brought a significant progress in the field of non-equilibrium statistical mechanics, providing an "exact" (on average!) relation between the free energy difference and the work done by a non-equilibrium process. The Jarzynski equality has the form

$$\langle e^{-\beta W} \rangle = e^{-\beta \Delta F}, \tag{2.74}$$

where $\beta = k_B T$ is the reduced temperature and the brackets $\langle \cdot \rangle$ denote an average taken over many independent repetitions of the same process. Eq. 2.74 is known in the literature with several alternative names, such as "Jarzynski non-equilibrium work theorem". In comparison to other methods, SMD allows larger conformational changes on the nanosecond timescale, and requires less computation [32]. The analysis method developed by Hummer [75] for the calculation of multidimensional PMF profiles from atomic force microscopy experiment, can be directly applied to extend SMD, as shown by Minh

[86, 87], who also demonstrated the superior efficiency of non-equilibrium methods in comparison to equilibrium ones. The Jarzynski equality, initially derived for canonical ensemble (NVT), has been generalized to isothermal-isobaric ensemble (NpT). It's derivation has been given for both Hamiltonian systems (as in the case of MD simulations) and Markovian processes (Monte Carlo simulations) [76].

Application of SMD and JE to the computation of PMF

Here we follow the description of the SMD method for free energy calculation in the canonical NVT ensemble, as given by Park and Schulten [76]. Given a classical system of N particles, in contact with a thermal bath at temperature T, the phase is defined by 3N positions **q** and 3N momenta **p**. We define our reaction coordinate as a suitable function of the position, and we call it $\xi(\mathbf{q})$. In SMD, a time-dependent driving potential h_λ is applied to the unperturbed Hamiltonian $\mathcal{H}(\mathbf{p},\mathbf{q})$:

$$\tilde{\mathcal{H}}(\mathbf{p},\mathbf{q}) = \mathcal{H}(\mathbf{p},\mathbf{q}) + h_\lambda, \qquad h_\lambda = \frac{k}{2}(\xi(\mathbf{q}) - \lambda)^2, \tag{2.75}$$

and the coupling parameter λ is changed with a certain constant velocity v,

$$\lambda(t) = \lambda(0) + vt, \tag{2.76}$$

over the desired range of ξ, during a time interval from 0 to τ. In other words, we apply a time dependent sequence of harmonic potentials that rapidly pull the scalar variable into a series of values $\lambda(t)$, steering the system into a predetermined conformational change, easily overcoming whatever energy barrier thanks to the magnitude of the perturbation. From the JE we can easily derive that the free energy difference for the state for the initial an final ξ is given by:

$$F_{\lambda(\tau)} - F_{\lambda(0)} = -\frac{1}{\beta} = log\langle exp[-\beta W(\tau)]\rangle, \tag{2.77}$$

where the work over one repetition of the process can be calculated as

$$W(\tau) = \int_0^\tau dt \left[\frac{\partial}{\partial t}\tilde{\mathcal{H}}(\mathbf{q}(t),\mathbf{p}(t))\right], \tag{2.78}$$

by numerical integration of the time derivative of the Hamiltonian over the trajectory steered by the perturbing potential. However, in this way we have calculated the PMF

for the *perturbed* system. The recovery of the original PMF is obtained by means of the so-called "stiff spring approximation": the idea is to choose a very large spring constant for the harmonic driving potential, minimizing the fluctuations of the reaction coordinate among different runs below a certain desired resolution. Several trajectories are then calculated (with different initial configurations but always within a NVT ensemble), the free energy F_λ is calculated as a function of λ. For large k, the PMF along ξ can be calculated, according to [76], as a Taylor expansion around λ,

$$\Phi(\lambda) = F_\lambda + \frac{1}{2k}(\frac{\partial F_\lambda}{\partial \lambda})^2 - \frac{1}{2\beta k}\frac{\partial^2 F_\lambda}{\partial \lambda^2} + O(\frac{1}{k^2}). \tag{2.79}$$

Thermostatting the system

The Jarzynski equality applies to system at constant temperature, i.e. in contact with a heat bath. In MD simulations, the effect of a bath is reproduced consistently by widely applied thermostat algorithms. Jarzynski himself [74] verified the applicability of Nose-Hoover thermostat.

Cumulant expansion

The computational downside of the bare Jarzynski equality is the average $\langle e^{\beta W}\rangle$, where the exponential causes the value to be strongly influenced by those rare repetitions of the process that yield a very small work W. Therefore, the direct calculation of the exponential average tends to give inaccurate results. The problem can be alleviated rewriting equation 2.77, expanding the logarithm of the exponential average as a sum of its cumulants [74]:

$$log\langle e^x\rangle = \langle x\rangle + \frac{1}{2}(\langle x^2\rangle - \langle x\rangle^2) + ..., \tag{2.80}$$

Where only the first two terms are shown. This formula is, in general, affected by a truncation error. However, when the work distribution is Gaussian, the terms beyond the second order vanish (a general property of all Gaussian-distributed quantities), the formula becomes exact, and the only remaining source of uncertainty is the finite number of sampled values. As noted by Park and Schulten, the condition of Gaussian work distribution can be satisfied by applying a stiff driving potential (hence the name of “stiff-spring” approach)[76].

2.3.6 Subsampling and statistical inefficiency

When averaging a quantity over the snapshots of a MD trajectory, we assume that we are averaging over a certain statistical ensemble, i.e. a set of microstates whose probability (usually dependent on the energy) is defined by a known statistical law.
Moreover, we assume the snapshots to be independent on each other. Strictly speaking, this is of course not the case for neighbouring timesteps of a MD run: at every given timestep, the time integration algorithm *deterministically* calculates the next timestep from the information contained in the present one. However, if the timesteps are sufficiently "far apart" in time, we can expect their properties to be decorrelated, and this is what actually happens.
The question is, of course, *how far* the sampled microstates must be: we have to choose a length of the sampling step which is longer than the correlation time, but the correlation time is in general not known *a priori*.
A formal treatment of this issue is given below, closely following the formulation given by [23].
If we want to compute a certain property A, the average of A is given by:

$$\langle A \rangle_{run} = \frac{1}{\tau_{run}} \sum_{\tau=1}^{\tau=\tau_{run}} A(\tau), \tag{2.81}$$

If the sampled values were statistically independent, the variance of the computed mean value over the trajectory would be:

$$\sigma^2(\langle A \rangle_{run}) = \sigma^2(A) \cdot \frac{1}{\tau_{run}}, \tag{2.82}$$

where

$$\sigma^2(A) = \frac{1}{\tau_{run}} \sum_{\tau=1}^{\tau=\tau_{run}} (A(\tau) - \langle A \rangle_{run})^2, \tag{2.83}$$

(The variance gives, by definition, the expected value of the squared error in the mean; We can see that the variance is inversely proportional to the length of the run). In order to estimate the correlation time, we can proceed as follows.
The run is first broken down into n_b series of length τ_b, so that $n_b \tau_b = \tau_{run}$. Now let's

consider the average taken on a certain τ_b chunk. The mean value over this chunk is:

$$\langle A \rangle_b = \frac{1}{\tau_b} \sum_{\tau=1}^{\tau=\tau_b} A(\tau), \tag{2.84}$$

and these averages are used to compute the variance::

$$\sigma^2(\langle A \rangle_b) = \frac{1}{n_b} \sum_{b=1}^{n_b} (\langle A \rangle_b - \langle A \rangle_{run})^2, \tag{2.85}$$

We expect this variance to decrease for longer τ_b, until it goes to 0 in the limit case of $\tau_b = \tau_{run}$.
The aim is to find the proportionality constant that allows the calculation of $\sigma^2(\langle A \rangle_b)$ for the whole trajectory. The "statistical efficiency" is defined as:

$$s = \lim_{\tau_b \to \infty} \frac{\tau_b \sigma^2(\langle A \rangle_b)}{\sigma^2(A)}, \tag{2.86}$$

which is the limit for large τ_b of the ratio between the *observed* variance, and the variance in the hypothesis of statistical independence of the sampled values. This limit can be computed from the available samples, calculating the value of s for increasingly large τ_b, until we reach a plateau. The calculated value tells us that only a snapshot every s adds new information to the average.
The statistical inefficiency sets therefore the upper limit to the “useful” sampling density.

2.4 Coarse-graining by force-matching

All the methods presented so far have proven successful in dealing with the determination of energy profiles for one or few reaction coordinates, as in the field of protein-docking. However, the coarse graining of more complicated systems, such as biomolecules, requires the determination of many interaction parameters. An alternative route, first explored by Ercolessi and coworkers [21], is to perform a so called “force-matching”, i.e. parametrise a coarse-grained potential by minimising the discrepancy between the forces measured by an all-atom simulation, and the forces generated by the CG force field, whose parameters will be the unknowns of the problem.

2.4.1 The Multiscale Coarse-Graining Method

The CG models addressed by the MSCG method are made of classical point masses, called sites, which correspond to one or more atoms in the atomistic representation (e.g. the center of mass of a group of atoms), and interact by means of an opportune CG potential [12]. The goal is the derivation of a CG system which behaves consistently with the underlying all-atom system. What do we mean by *consistency*? The atomistic system is described by classical Hamiltonian equations of motions; the coarse graining procedure involves opportune operators in order to map the coordinates (positions and momenta) of the atoms, onto a reduced set of coordinates of CG sites. The dynamics of the reduced system will then be described by classical Hamiltonian equations for the coordinates (positions and momenta) of the CG sites. It must be noted that the reduced equations of motion are entirely determined by the initial atomistic equations and by the mapping operators [17]. The CG model is called *consistent* if the joint probability distribution function of the reduced positions and momenta is the same as the joint PDF *implied* by the atomistic PDF and the mapping operators. More formally: the state of the atomistic system of n atoms is univocally determined by the vectors of positions and coordinates:

$$\mathbf{r} = \{r_1 \dots r_{3n}\} \quad \textit{and} \quad \mathbf{p} = \{p_1 \dots p_{3n}\}, \tag{2.87}$$

the corresponding Hamiltonian is given by:

$$\mathcal{H}(\mathbf{r}, \mathbf{p}) = \sum_{i=1}^{3n} \frac{1}{2m_i} p_i^2 + u(\mathbf{r}), \tag{2.88}$$

where $u(\mathbf{r})$ is the potential energy. In a completely analogous fashion, the state of the CG system of N sites is univocally determined by the vectors of site positions and coordinates:

$$\mathbf{R} = \{R_1 \dots R_{3n}\} \quad \textit{and} \quad \mathbf{P} = \{P_1 \dots P_{3n}\}, \tag{2.89}$$

the corresponding Hamiltonian is given by:

$$\mathcal{H}(\mathbf{R}, \mathbf{P}) = \sum_{i=1}^{3n} \frac{1}{2m_i} P_i^2 + U(\mathbf{R}), \tag{2.90}$$

where $U(\mathbf{R})$ is the CG potential energy.

The link between the two representations is given by the mapping operators for positions and momenta, namely:

$$M_{RI}(\mathbf{r}) = \sum_{i=1}^{3n} c_{Ii}\mathbf{r}_i \tag{2.91}$$

$$M_{PI}(\mathbf{r}) = M_I \sum_{i=1}^{3n} c_{Ii}\mathbf{p}_i/m_i, \tag{2.92}$$

for $i = 1 \ldots 3N$, which means that every coarse grained coordinate is expressed as a weighted average of the atomistic coordinates, with weights $c_I i$. A condition is imposed on the weighting coefficients of the mapping operators, which makes sure that if the atomic system undergoes a translation, the CG sites will undergo the same translation:

$$\sum_{i=1}^{3n} = 1. \tag{2.93}$$

A simple and intuitive mapping, that satisfies the conditions listed above, is the one that divides the atomistic models into *disjoint* sets of atoms, and places the CG sites on the centres of mass of the sets, provided that no constraints are applied between two atoms that belong to different groups. Without loss of generality, atoms within the same group can be linked by rigid bonds, which is very convenient in most biomolecular simulations, where actually all the fast-moving hydrogen atoms are constrained [17]. The resulting coefficients in the linear mapping operator become, for atoms involved in the I-th group,

$$c_{Ii} = m_i M_I \tag{2.94}$$

$$M_I = \sum_i m_i \tag{2.95}$$

and the total force acting on the corresponding $I - th$ CG site is simply the sum of the forces on the single atoms,

$$\mathbf{f}_I(\mathbf{r}^n) = \sum_i \mathbf{f}_i(\mathbf{r}^n), \tag{2.96}$$

The consistency condition, for positions and momenta, is written as:

$$P_R(\mathbf{R}^\mathbf{N}) = p_R\mathbf{R}^\mathbf{N} \tag{2.97}$$

$$P_P(\mathbf{P}^\mathbf{N}) = p_P\mathbf{P}^\mathbf{N} \tag{2.98}$$

The above relations mean that the CG system will visit its own microstates with the same probability that is implied by the partition function of atomistic system and the chosen mapping. Eventually, the "best fitting" force field is obtained by minimising what we could call the "merit function" of the force field in the space of its parameters. The quantity to be minimised is that proposed by [1], defined as:

$$\chi^2_{MS}[\mathbf{F}^{MS}] = \frac{1}{3n_tN}\sum_{t=1}^{n_t}\sum_{I=1}^{N}|\mathbf{f}_I(\mathbf{r}_t^n) - \mathbf{F}_I^{MS}(\mathbf{M}_{\mathbf{R}}^N(\mathbf{r}_t^n))|^2 \tag{2.99}$$

where, once the mapping is specified, χ^2_{MS} is a function of all the CG force field parameters. The terms that appear in the equation, and their physical meanings, are:

- n_t is the total number of configurations sampled from the atomistic simulation;
- N is the number of sites in the CG model;
- $\mathbf{f}_I$ is the total force, as computed in the atomistic model, on the atoms involved in the I-th CG site;
- $\mathbf{r}_t^n$ is the t-th configuration sampled from the atomistic simulation;
- $\mathbf{F}_I^{MS}$ is the force on the I-th CG site, as a function of the CG sites (this term is a function of the FF parameters to be optimised).
- $\mathbf{M}_{\mathbf{R}}^N(\mathbf{r}_t^n) = \mathbf{R}^N$ is the configuration of the CG model;

all above quantities are known or easy to compute from the atomistic simulation; the crucial modelling step is writing the term $F_I^{MS}(\mathbf{R}^N)$. Schematically, the multiscale Coarse-Graining process can be divided into the following steps:

1. Define a mapping that assigns all atoms of the atomistic model to the appropriate sites of the coarse-grained model;
2. Perform the atomistic simulations, and store n_t snapshots, positions and forces on each atom);
3. Then, for every stored configuration:
   (a) Translate atomistic configurations into CG configurations

(b) Calculate the forces acting on the CG sites, as sum of the atomistic forces acting over all atoms involved in each site.

(c) Calculate the forces acting on the CG sites, as predicted by the CG potential (this will be a function of the unknown CG-FF parameters).

4. Write the resulting residual χ^2 as a sum, over all configurations and all sites, of the squared differences between the CG forces calculated from the atomistic data, and those predicted by the CG-FF.

5. Minimise χ^2, with respect to the force-field parameters.

The problem is recast in terms of least-square minimisation in the parametric space of the CG force field [18]. This method has already been successfully applied to a peptide model [88].

2.4.2 Equations of the MSCG problem

The purpose of the atomistic simulation and the CG-mapping is the construction of the overdetermined linear problem of Eq. 2.99 [18]. The possible strategies for the solution of the systems 2.99 have been extensively discussed by Izvekov and Noid [1, 16, 18]. Since the matrix for the whole system trajectory is usually too big, an alternative is a block-averaging approach, where the n_t sampled configurations are divided into disjointed subsets, which are used to calculate separate residuals, solving the resulting (much smaller) set of equations [18]. The force-field parameters are then recovered by averaging over the estimates obtained from each block. Other numerical developments are being pursued, e.g. combining the MSCG method with techniques of of Bayesian inference, in order to improve the reliability of the parameter estimation [89]. The details on the implementation a possible routes for the numerical solution of the least-squares problem will be given in the next chapters of this thesis.

2.5 Implementation of a generic interaction potential

For the least-squares problem to become an overdetermined linear system, the analytical form of the interaction forces must be linear in the unknown force-field parameters. This is not the case for any of the terms that usually model biomolecular interactions, which are:

- Lennard-Jones potential, that models Van der Waals forces;
- Coulombic potential, that models electrostatic interactions;
- Harmonic bonded potential, that models bond-stretching forces;
- Angular potential, that models angle stretching;
- Dihedral potential, that model dihedral torsional forces.

Since the geometry of the CG system is much more simple than the AA counterpart, the inclusion of improper dihedrals is not necessary. In addition, Forces are calculated as derivatives of the potential energies with respect to the particles' coordinates, and the corresponding expressions can be quite complicated, especially for angles and dihedrals. Let's summarise how the forces on the atoms are computed. Every term of the interaction potential is characterised by an interaction coordinate, which is a function of the positions of two or more atoms:

- Lennard-Jones potential: interatomic distance, r;
- Coulombic potential: interatomic distance, r;
- Harmonic bonded potential: bond length (interatomic distance of the bonded pair), r;
- Angular potential: angle width, θ;
- Dihedral potential: torsional angle ϕ;

The force on the atom is computed as the sum of forces coming from all potential terms. The force associated with a potential term is computed as the negative gradient of the potential function with respect to the position of the atom. For pairwise interactions (LJ, Coulombic and bonds) the force on each atom of a pair ij is a vector acting along the interatomic distance vector $\mathbf{r}_{ij}$, and whose norm is the negative derivative of the potential energy with respect to the distance. For angles and dihedrals, however, the associated linear algebra is less straightforward.
For computational purposes, it is convenient to represent the computed force, which is a function of the interaction coordinate, as a sum of weighted delta functions centred on the discretised values that the scalar variable can take. The weight for each delta function will be one of the parameters to be optimised. The consequence of this approach is that the

functional form becomes very general and formally simple, and most importantly, *linear*, but the number of parameters to fit increases. The actual numerical implementation will be described in the next chapter.

2.5.1 Computation of Lennard-Jones interactions

The pairwise Lennard-Jones (LJ) potential mimics the effect Van der Waals forces. The analytical form is:

$$U_{LJ} = \epsilon \left[\left(\frac{\sigma}{r} \right)^{12} - \left(\frac{\sigma}{r} \right)^{6} \right], \tag{2.100}$$

Where r is the interatomic distance and ϵ and σ represent respectively the depth of the attractive energy well, and what is commonly called "collision radius", i.e. the approximative distance where the potential changes from attractive to repulsive.

2.5.2 Computation of Coulombic forces

Electrostatic interactions are represented by the classical Coulombic potential

$$U_C = \frac{1}{4\pi\epsilon} \frac{q_i q_j}{r} \tag{2.101}$$

where C is the Coulomb constant, q_i and q_j are the atomic charges, r is again the interatomic distance, and the ϵ constant represents the dielectric constant *in vacuo* (not to be confused with the constant in the LJ potential). The resulting forces can be computed either by solving the Poisson equation with a long-range scheme (Ewald summation) or using a cutoff approach, whose usage has recently seen a resurgence [90, 91].

2.5.3 Computation of bond forces

CHARMM represents covalent bonds as harmonic springs, with potential:

$$U_b = K_b (r - r_0)^2. \tag{2.102}$$

where K_b is the spring constant, r_0 is the rest length, and r the interatomic distance. Note that is the CHARMM potential the usual factor $1/2$ (which is seen in most textbook when discussing harmonic potentials) is usually included in the spring constant K_θ and does not

appear explicitly. The same applies to all harmonic potentials used for angles, dihedrals, etc.

2.5.4 Computation of angle forces

In biomolecular simulations, an angle is univocally determined by three bonded atoms. The scalar variable associated with the bending potential is the angle θ. The expression of the potential energy is:

$$U_a = K_a(\theta - \theta_0)^2, \tag{2.103}$$

where K_a is the angular spring constant, θ_0 is the rest width, and θ the angle width. The computation of forces arising from the harmonic bending potential can be computed by the general method suggested by [92] and also presented by [24].

2.5.5 Computation of dihedral forces

A dihedral angle ϕ is defined by four atoms h, i, j, k. The dihedral potential adopted in the CHARMM force field used for the atomistic simulations has the analytical form:

$$U_\phi = K_\phi[1 + cos(m\phi - \delta)], \tag{2.104}$$

Where m is the multiplicity and δ the phase angle. The calculation of the forces on the four atoms implies some very long and rather tedious linear algebra. The interested reader can find an exhaustive derivation of the numerical method in references [23, 24].

3
Numerical Background

The purpose of this chapter is to briefly outline the nature of the main numerical calculations required for the work carried forward for this thesis, and the methods and strategies used to tackle them.

3.1 Numerical implementation of interaction potentials

The choice of the interaction potential is of paramount importance in molecular dynamics, because it defines the physical model of the simulated system. The implementation of interatomic forces is given in the following.

3.1.1 Pairwise non-bonded interactions: Lennard-Jones and Coulombic interactions

For pairwise Lennard-Jones potentials, the interaction between atom i and j produces a contribution to the potential energy :

$$v_{LJ}(r_{ij}) = 4\epsilon\left[\left(\frac{\sigma}{r_{ij}}\right)^{12} - \left(\frac{\sigma}{r_{ij}}\right)^{6}\right] \tag{3.1}$$

where v is the potential energy, $r_{ij} = |\mathbf{r}_i - \mathbf{r}_j|$ is the interatomic distance, σ is the collision radius and ϵ is the depth of the energy well.
The corresponding pairwise force $\mathbf{F}_i$ on atom the i can be calculated as:

$$\mathbf{f}_i = -\nabla_{\mathbf{r}_i} v_{LJ}(r_{ij}) \tag{3.2}$$

where $\nabla_{\mathbf{r}_i}$ must be interpreted as a partial gradient of the potential with respect to the coordinates of the atom i. The resulting force components are (dropping the potential's dependency on r_{ij}, for brevity):

$$f_{ix} = -\frac{\partial v_{LJ}}{\partial x_i} = -\frac{dv_{LJ}}{dr_{ij}} \cdot \frac{\partial r_{ij}}{\partial x_i} \tag{3.3}$$

$$f_{iy} = -\frac{\partial v_{LJ}}{\partial y_i} = -\frac{dv_{LJ}}{dr_{ij}} \cdot \frac{\partial r_{ij}}{\partial y_i} \tag{3.4}$$

$$f_{iz} = -\frac{\partial v_{LJ}}{\partial z_i} = -\frac{dv_{LJ}}{dr_{ij}} \cdot \frac{\partial r_{ij}}{\partial z_i} \tag{3.5}$$

The derivative with respect to r_{ij} is immediately evaluated as:

$$\frac{dv_{LJ}}{dr_{ij}} = -48\epsilon\left[\frac{\sigma^{12}}{r_{ij}^{13}} - \frac{1}{2}\frac{\sigma^{6}}{r_{ij}^{7}}\right], \tag{3.6}$$

whilst the terms containing the partial derivatives are:

$$\frac{\partial r_{ij}}{\partial x_i} = \frac{(x_i - x_j)}{r_{ij}} \tag{3.7}$$

$$\frac{\partial r_{ij}}{\partial y_i} = \frac{(y_i - y_j)}{r_{ij}} \tag{3.8}$$

$$\frac{\partial r_{ij}}{\partial z_i} = \frac{(z_i - z_j)}{r_{ij}}, \tag{3.9}$$

which result in the following expressions:

$$f_{ix} = 48\frac{\epsilon}{\sigma^2}\left[\left(\frac{\sigma}{r_{ij}}\right)^{14} - \frac{1}{2}\left(\frac{\sigma}{r_{ij}}\right)^{8}\right](x_i - x_j) \tag{3.10}$$

$$f_{ix} = 48\frac{\epsilon}{\sigma^2}\left[\left(\frac{\sigma}{r_{ij}}\right)^{14} - \frac{1}{2}\left(\frac{\sigma}{r_{ij}}\right)^{8}\right](y_i - y_j) \tag{3.11}$$

$$f_{ix} = 48\frac{\epsilon}{\sigma^2}\left[\left(\frac{\sigma}{r_{ij}}\right)^{14} - \frac{1}{2}\left(\frac{\sigma}{r_{ij}}\right)^{8}\right](z_i - z_j). \tag{3.12}$$

for the three force components. The derivation of the Coulombic forces is analogous, with the potential function given by

$$v_C(r_{ij}) = C\frac{q_i q_j}{r_{ij}}, \tag{3.13}$$

where $C = 1/4\pi\epsilon$, with ϵ being the dielectric constant *in vacuo*, and q_i, q_j the charges of the atoms i anf j. Again the force can be computed from the general principle expressed in Eq. 3.2, as:

$$f_{ix} = -\frac{\partial v_C}{\partial x_i} = \frac{dv_C}{dr_{ij}} \cdot \frac{\partial r_{ij}}{\partial x_i} \tag{3.14}$$

$$f_{iy} = -\frac{\partial v_C}{\partial y_i} = \frac{dv_C}{dr_{ij}} \cdot \frac{\partial r_{ij}}{\partial y_i} \tag{3.15}$$

$$f_{iz} = -\frac{\partial v_C}{\partial z_i} = \frac{dv_C}{dr_{ij}} \cdot \frac{\partial r_{ij}}{\partial z_i}, \tag{3.16}$$

The partial derivatives are exactly the same as in the case of the Lennard-Jones potential, and the derivative with respect to the interatomic distance is trivial. The resulting forces are:

$$f_{ix} = C\frac{q_i q_j}{r_{ij}^3}(x_i - x_j) \tag{3.17}$$

$$f_{iy} = C\frac{q_i q_j}{r_{ij}^3}(y_i - y_j) \tag{3.18}$$

$$f_{iz} = C\frac{q_i q_j}{r_{ij}^3}(z_i - z_j), \tag{3.19}$$

The implementation is slightly more complicated when the implementation includes a switching function that brings the potential smoothly to 0 at the cutoff distance. This eliminates the discontinuity produced by the truncation of the potential function at the

cutoff distance.

3.1.2 Bond stretching

The computation of the internal forces arising from bond stretching closely resembles the case of the pairiwise potential. For biomolecular simulations, the energy of the bonds is usually modelled by a harmonic potential:

$$v_b(r_{ij}) = K(r_{ij} - r_0)^2, \tag{3.20}$$

which after a derivation that closely resembles the one in the previous section, yelds the following force contributions:

$$f_{ix} = -2(r_{ij} - r_0)\frac{(x_i - x_j)}{r_{ij}} \tag{3.21}$$

$$f_{iy} = -2(r_{ij} - r_0)\frac{(y_i - y_j)}{r_{ij}} \tag{3.22}$$

$$f_{iz} = -2(r_{ij} - r_0)\frac{(z_i - z_j)}{r_{ij}}, \tag{3.23}$$

for the atom i, and force components of opposite sign for the atom j.

3.1.3 Angle bending

In a molecular simulation of a systems that contains covalent bonds, and "angle" is individuated by three mutually bonded atoms i, j, k (here we will assume that the bonds are i-j and j-k). The potential energy associated to angle bending is also usually a harmonic relation,

$$v_\theta(\theta_{ijk}) = K(\theta_{ijk} - \theta_0)^2. \tag{3.24}$$

Defining the two vectors $\mathbf{b}_1 = \mathbf{r}_i - \mathbf{r}_j$ and $\mathbf{b}_2 = \mathbf{r}_k - \mathbf{r}_j$, the angle θ can be calculated from the relations

$$\theta = cos^{-1}(cos\theta) \tag{3.25}$$

$$cos\theta = \frac{\mathbf{b_1} \cdot \mathbf{b_2}}{|\mathbf{b_1}||\mathbf{b_2}|}. \tag{3.26}$$

and the corresponding forces can be calculated again by means of Eq. 3.2. The analytical route suggested by [23] is to use the chain rule of derivation to rewrite Eq. 3.2 as:

$$\mathbf{f}_i = -\frac{dv(\theta)}{d(cos\theta)} \cdot \nabla_{\mathbf{r}_i} cos\theta, \tag{3.27}$$

In comparison to the pairwise and the bonded potentials, however, the computation of the atomic forces requires some extra linear algebra.

First, we notice that for a sistem of three atoms the forces arising from the angle bending potential are internal, therefore the total force on the three atoms must be zero along the three directions. It is enough to compute the components on atoms i and k, and the forces on atom 2 can be calculated by difference. This saves us the effort of computing all derivatives explicitly, and the resulting equations for the force components are:

$$\mathbf{f}_{ix} = -\frac{dv(\theta)}{d(cos\theta)} \cdot \frac{\partial}{\partial x_i} cos\theta \tag{3.28}$$

$$\mathbf{f}_{iy} = -\frac{dv(\theta)}{d(cos\theta)} \cdot \frac{\partial}{\partial y_i} cos\theta \tag{3.29}$$

$$\mathbf{f}_{iz} = -\frac{dv(\theta)}{d(cos\theta)} \cdot \frac{\partial}{\partial z_i} cos\theta \tag{3.30}$$

$$\mathbf{f}_{kx} = -\frac{dv(\theta)}{d(cos\theta)} \cdot \frac{\partial}{\partial x_k} cos\theta \tag{3.31}$$

$$\mathbf{f}_{ky} = -\frac{dv(\theta)}{d(cos\theta)} \cdot \frac{\partial}{\partial y_k} cos\theta \tag{3.32}$$

$$\mathbf{f}_{kz} = -\frac{dv(\theta)}{d(cos\theta)} \cdot \frac{\partial}{\partial z_k} cos\theta \tag{3.33}$$

$$\mathbf{f}_{jx} = -\mathbf{f}_{ix} - \mathbf{f}_{kx} \tag{3.34}$$

$$\mathbf{f}_{jy} = -\mathbf{f}_{iy} - \mathbf{f}_{ky} \tag{3.35}$$

$$\mathbf{f}_{jz} = -\mathbf{f}_{iz} - \mathbf{f}_{kz} \tag{3.36}$$

The first derivative with respect to the cosine, for the harmonic potential given by Eq. 3.24, is easily evaluated as:

$$\frac{dv(\theta)}{d(cos\theta)} = \frac{2K(\theta - \theta_0)}{sin\theta}. \tag{3.37}$$

It is possible to derive simple relations for the partial derivatives of the numerator and denominator of the cosine in Eq. 3.25, with respect to the coordinates of atoms i and k.

For the numerator $(\mathbf{b}_1 \cdot \mathbf{b}_2)$ we get:

$$\frac{\partial}{\partial x_i}(\mathbf{b}_1 \cdot \mathbf{b}_2) = (x_k - x_j) \tag{3.38}$$
$$\frac{\partial}{\partial y_i}(\mathbf{b}_1 \cdot \mathbf{b}_2) = (y_k - y_j) \tag{3.39}$$
$$\frac{\partial}{\partial z_i}(\mathbf{b}_1 \cdot \mathbf{b}_2) = (z_k - z_j) \tag{3.40}$$
$$\tag{3.41}$$
$$\frac{\partial}{\partial x_i}(\mathbf{b}_1 \cdot \mathbf{b}_2) = (x_j - x_k) \tag{3.42}$$
$$\frac{\partial}{\partial x_i}(\mathbf{b}_1 \cdot \mathbf{b}_2) = (y_j - y_k) \tag{3.43}$$
$$\frac{\partial}{\partial x_i}(\mathbf{b}_1 \cdot \mathbf{b}_2) = (z_j - z_k), \tag{3.44}$$

and for the denominator $|\mathbf{b}_1| \cdot |\mathbf{b}_2|$ we get:

$$\frac{\partial}{\partial x_i}(|\mathbf{b}_1| \cdot |\mathbf{b}_2|) = (x_i - x_j)\frac{|\mathbf{b}_2|}{|\mathbf{b}_1|} \tag{3.45}$$
$$\frac{\partial}{\partial y_i}(|\mathbf{b}_1| \cdot |\mathbf{b}_2|) = (y_i - y_j)\frac{|\mathbf{b}_2|}{|\mathbf{b}_1|} \tag{3.46}$$
$$\frac{\partial}{\partial z_i}(|\mathbf{b}_1| \cdot |\mathbf{b}_2|) = (z_i - z_j)\frac{|\mathbf{b}_2|}{|\mathbf{b}_1|} \tag{3.47}$$
$$\tag{3.48}$$
$$\frac{\partial}{\partial x_k}(|\mathbf{b}_1| \cdot |\mathbf{b}_2|) = (x_k - x_j)\frac{|\mathbf{b}_1|}{|\mathbf{b}_2|} \tag{3.49}$$
$$\frac{\partial}{\partial y_k}(|\mathbf{b}_1| \cdot |\mathbf{b}_2|) = (x_y - y_j)\frac{|\mathbf{b}_1|}{|\mathbf{b}_2|} \tag{3.50}$$
$$\frac{\partial}{\partial z_k}(|\mathbf{b}_1| \cdot |\mathbf{b}_2|) = (z_k - z_j)\frac{|\mathbf{b}_1|}{|\mathbf{b}_2|} \tag{3.51}$$

Now if we define:

$$\mathbf{b}_1 \cdot \mathbf{b}_2 = A \tag{3.52}$$
$$|\mathbf{b}_1| \cdot |\mathbf{b}_2| = B \tag{3.53}$$
$$\frac{|\mathbf{b}_2|}{|\mathbf{b}_1|} = C \tag{3.54}$$
$$\frac{|\mathbf{b}_1|}{|\mathbf{b}_2|} = D, \tag{3.55}$$

the partial derivatives of the cosine can be calculated as:

$$\frac{\partial}{\partial x_i} cos\theta = \frac{(x_k - x_j)B - (x_i - x_j)A}{B^2} C \tag{3.56}$$

$$\frac{\partial}{\partial y_i} cos\theta = \frac{(y_k - y_j)B - (y_i - y_j)A}{B^2} C \tag{3.57}$$

$$\frac{\partial}{\partial z_i} cos\theta = \frac{(z_k - z_j)B - (z_i - z_j)A}{B^2} C \tag{3.58}$$

$$\tag{3.59}$$

$$\frac{\partial}{\partial x_k} cos\theta = \frac{(x_i - x_j)B - (x_k - x_j)A}{B^2} D \tag{3.60}$$

$$\frac{\partial}{\partial y_k} cos\theta = \frac{(y_i - y_j)B - (y_k - y_j)A}{B^2} D \tag{3.61}$$

$$\frac{\partial}{\partial z_k} cos\theta = \frac{(z_i - z_j)B - (z_k - z_j)A}{B^2} D. \tag{3.62}$$

Furthermore, if the width of the angle is expected to deviate considerably from the equilibrium value θ_0, a harmonic potential is added between atoms i and k to keep into account the increase in angular spring stiffness caused by the collision between the electronic shells of finite-sized atoms. This corrective term usually goes under the name of Urey-Bradley potential, and is implemented as a additional harmonic bond as described in the previous section. However, it is only computed when the bond is "compressed", i.e. when the bond length is lesser than its rest length.

3.1.4 Dihedral torsion potential

The derivation of the forces arising from dihedrals is extremely cumbersome and omitted for the sake of brevity. A very accurate derivation is detailed in Ref.[23].

3.2 Force-matching equations

The calculation of the pairwise interactions is performed using the MSCG method, minimising the merit coefficient χ^2_{MS} as written in Eq. 2.99.
If we rewrite Eq. 2.99 as follows:

$$\chi^2_{MS} = (\mathbf{f} - \mathcal{G}\phi)^T(\mathbf{f} - \mathcal{G}\phi), \tag{3.63}$$

where $\mathcal{G}$ is a matrix of $3n_tN \times N_D$ elements, and $\mathbf{f}$ is a vector of $3n_tN$ known terms, and ϕ is the vector of N_D unknown force-field parameters, it becomes apparent that the set of parameters that minimise χ^2_{MS} can be calculated as a least-squares solution of the overdetermined linear system:

$$\mathbf{f} - \mathcal{G}\phi = \mathbf{0}, \quad or \quad \mathcal{G}\phi = \mathbf{f}, \tag{3.64}$$

with $3n_tN$ equations and N_D unknowns [18].
The system in Eq. 3.64 has $3n_tN$ equations and N_D unknowns. We require such system to have more equations than unknowns, therefore the condition on the minimum number of configurations that we must postprocess is:

$$n_t > \frac{N_D}{3N}. \tag{3.65}$$

This condition must be satisfied for each block in case a Block-Averaging technique [18] is used.

3.2.1 Pre-calculation of bonded interactions using a harmonic approximation

In our CG force field we have decided to neglect the torsional degrees of freedom of the polymer backbone, as their energy contribution is usually more limited than for bonds and bending angles [13]. The chosen analytical form for CG bonds and angles was that of a harmonic potential,

$$U_b = K_b(r - r_0)^2 \tag{3.66}$$

$$U_a = K_a(\theta - \theta_0)^2 \tag{3.67}$$

Where U_b, U_a, are the potential energies, K_b, K_a the spring constants, and b_0, θ_0 the rest length and rest angle width, respectively. The force-field parameters were obtained from the equilibrium fluctuations of the atomistic system, following the method suggested by Zhou and coworkers [88]. At equilibrium, the probability distribution of the energy associated with one bonded degree of freedom must obey Boltzmann statistics, and therefore the probability density of the associated scalar variable x must be

$$P(x) = C \cdot exp(-U(x)/k_B T), \tag{3.68}$$

where x is the appropriate scalar variable that depends on the type of potential (i.e. bond length or angle width),T is the absolute temperature, k_B is the Boltzmann constant, $U(x)$ is the potential energy and C a normalisation constant to ensure the the integral of over all possible values of x is equal to 1, i.e. ensure sure that P(x) is actually a probability distribution. The empirical probability histograms for each CG bonded degree of freedom were generated from the all-atom simulations, using the appropriate topological mapping. The parameters were then determined by fitting Eq. 3.69 onto the empirical distribution. This approach has the downside of requiring a starting assumption for the analytical shape of the potential energy. However, harmonic potentials seem to work well with most systems [18, 88]. If we substitute a harmonic potential in Eq. 3.69 we can see that the resulting probability density of x has the form:

$$P(x) = C \cdot exp\left(-\frac{K_x(x - x_0)^2}{k_B T}\right) = C \cdot exp\left(-\frac{1}{2}\frac{(x - x_0)^2}{k_B T/2K_x}\right), \tag{3.69}$$

which is a Gaussian distribution with mean x_0 and variance $k_B T/2K_x$. Therefore, after checking that the shape of the distribution is actually Gaussian, the parameters could also be estimated as:

$$x_0 = \langle x \rangle \tag{3.70}$$

$$K_x = \frac{k_B T}{2\langle (x - x_0)^2 \rangle} \tag{3.71}$$

where $\langle \cdot \rangle$ represents the averaging over the available samples. The possibility to determine bonded interactions in the comparatively simple way described above, can help simplify the complexity of the more demanding least-squares problem associated with the determination of the non-bonded pairwise potential between each possible pair of CG sites.

Namely, the forces contribution arising from bonds, angles and dihedrals can be computed separately and *subtracted* from the vector of known terms (i.e. known forces on the sites, according to the atomistic model and the chosen mapping). The advantage of precalculating the bonded interactions, instead of leaving them as unknowns and computing them along the pairwise interactions from the general MSCG least-squares problem, is that the new system has about only half of the original unknowns. In the CG study presented in later sections, about 1000 unknowns out of over 2000 were removed in this fashion.

3.2.2 Solution of the least-squares problem

The original formulation of the least-squares problem associated with the MSCG method, (Eq. 3.64) [18]. can be cumbersome and can lead to very big matrices. In particular, several factors can cause a rapid increase in size of the matrix G, beyond the size that can be directly addressed with the available memory and storage space.
The first is the necessity to determine explicitly *all* pairwise interactions between different types of CG sites, can be problematic in systems with many different types of sites: with three site types, the possible combinations are only 6, whilst with 5 site types (as in the case of the solvated DNA double strand) the combinations are already 15. With a cutoff distance of 10Å and a grid step of 0.1Å, every type of pairwise interaction contributes with 100 further unknowns to the complexity of the problem. Therefore a simple system of 2000 sites and 5 site types, even using a block averaging approximation with a conservative block size of 10 configurations, would generate a matrix 60000×1500, which uses about 1GB of memory (storing elements in double precision). We are therefore already close to the upper limit of complexity that can be dealt with using an average desktop computer.
A possible way to circumvent this problem, suggested by Noid and coworkers, is to rewrite the system in the following form, equivalent to Eq. 3.64:

$$\chi^2_{MS} = \phi^T \mathbf{G}\phi - 2\mathbf{b}^T\phi + \mathbf{f}^T\mathbf{f}, \tag{3.72}$$

Where $\mathbf{b} = \mathcal{G}^T\mathbf{f}$, and $\mathbf{G} = \mathcal{G}^T\mathcal{G}$. The suggested route to solve this problem is look for the stationary point of $\chi^2_{MS}(\phi)$, zeroing all partial derivatives $\partial\chi^2/\partial\phi_i$, for $i = 1...N_D$, obtaining

the system of N_D equations and N_D unknowns

$$\mathbf{G}\phi = \mathbf{b}, \tag{3.73}$$

that can be attacked directly using standard methods for linear systems (as the size of the coefficient matrix will be typically in the order of about 1000×1000). This formulation has the disadvantage that the condition number of the **G** matrix is the square of the original condition number of **G**, and therefore the corresponding linear system requires preconditioning in order to yield accurate results; on the other hand, the size of the matrix is greatly reduced ($N_D \times N_D$) and consequently the memory requirements are much lower [18].

The strategies for the solution of the linear system 3.64 will be discussed further in the last chapters of this thesis, dedicated to the coarse graining of DNA molecules.

Part III

Simulation of the diffusive properties of DNA oligomers

4

Viscosity of Price-Brooks TIP3P model water

In this chapter we present the results of a parametric study of the effect of different temperatures and salt concentrations on the dynamic viscosity of a specific type of water model, the Price-Brooks TIP3P water [93].
This kind of water model is a refinement of the original TIP3P water model [94], and is optimised for MD simulations performed with Periodic Boundary Conditions and long-range solvers for electrostatic interactions, such as Ewald summation, and particle-particle particle-mesh (PPPM) [46]. The necessity of this study arises because a reliable knowledge of the solvent viscosity is crucial in the estimation of molecular diffusion coefficients from MD simulation. Moreover, the data available from the literature did not take into consideration the effects of temperature and salt concentration on the viscosity.
Like most empirical models, the Price and Brooks TIP3P is parametrised for pure water at 298K, and its behaviour under different conditions is not known a priori. We have investigated the effects of different temperatures and salt concentrations on the viscosity of the Price-Brooks variety of TIP3P (which we will call PB-TIP3P), in order to quantify the

deviation from the behaviour in standard conditions. This is important for any simulation performed at non-standard temperature and ion concentrations.

4.1 Introduction

The existence of a large number of water models implicitly underlines the difficulty to capture all water properties with only one parametrisation [95]. The usage of a model is therefore limited to the set of properties that it can reproduce within an acceptable accuracy level. Simple models such as the 3-point TIP3P and its subsequent modifications have encountered success, because they are computationally less demanding than more sophisticated ones, such as the TIP5P [95]. Further interest in the TIP3P arose from the fact that the widely used CHARMM force field for biomolecular simulations was parameterised using TIP3P, and it was unclear whether the solvation properties would be preserved when using a different water model [96]. Only recently, a study performed by Nutt and Smith using the CHARMM potential has shown that the behaviour of biomolecules in solution is fairly similar when other water models are adopted: this applies to several versions of TIP3P, TIP4P, and to a lesser extent to TIP5P [96]. Despite of this, TIP3P still remains the most widely adopted water model. One of the reasons is that not all MD packages implement the numerical optimisations that allow to compensate most of the additional computational cost of TIP4P and TIP5P [97]. Well-known limitations of the TIP3P model include its poor structure, high self-diffusion coefficient [95], and low viscosity [98]. A modified set of parameters (PB-TIP3P) was proposed a few years ago by Price and Brooks that improves the structural properties of the model and its suitability in simulations with periodic boundary conditions and long-range electrostatics [93].

4.2 Methods and simulated cases

The PB-TIP3P has the same rigid 3-point geometry of the original TIP3P model, and the same stiffness for the O-H bonds and the H-O-H angle (see Tab. 4.2), but it modifies the fixed electrical charges on the atoms, as well as the parameters of the pairwise Lennard-Jones interactions for oxygen and hydrogen atoms. The interaction parameters are summarised in Tab. 4.1. For biomolecular simulations the molecule is usually kept rigid applying holonomic constraints by means of algorithm such as SHAKE [41]. Constraining all bonds and angles involving the fast-vibrating H atoms is common practice,

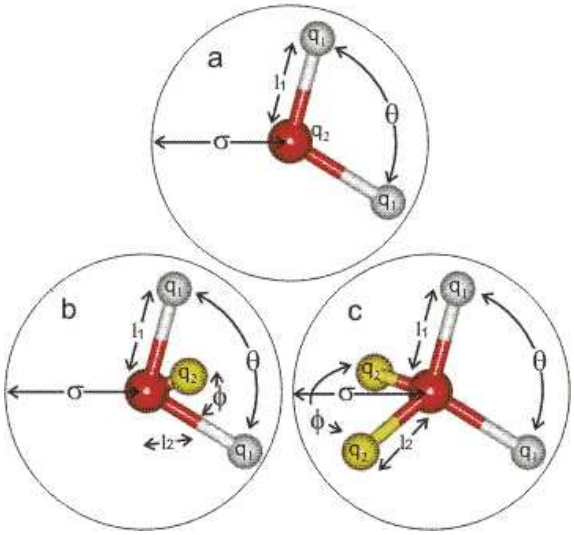
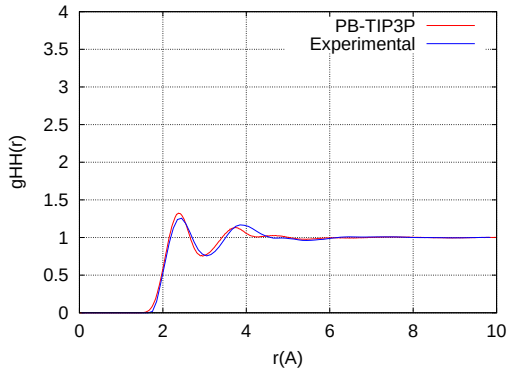


FIGURE 4.1: Geometry of TIP3P (a), TIP4P (b) and TIP5P (c) water models. TIP4P and TIP5P add extra massless charges in order to better reproduce the charge distribution around the molecule.

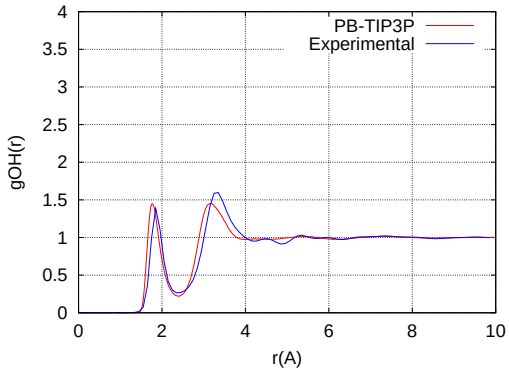
in biomolecular simulations, in order to allow a the use of a longer timestep and reduce the computational cost.

The PB-TIP3P model has a good capability to reproduce the radial distribution function of oxygen atoms which is found in real water, as shown in Fig. 4.2.

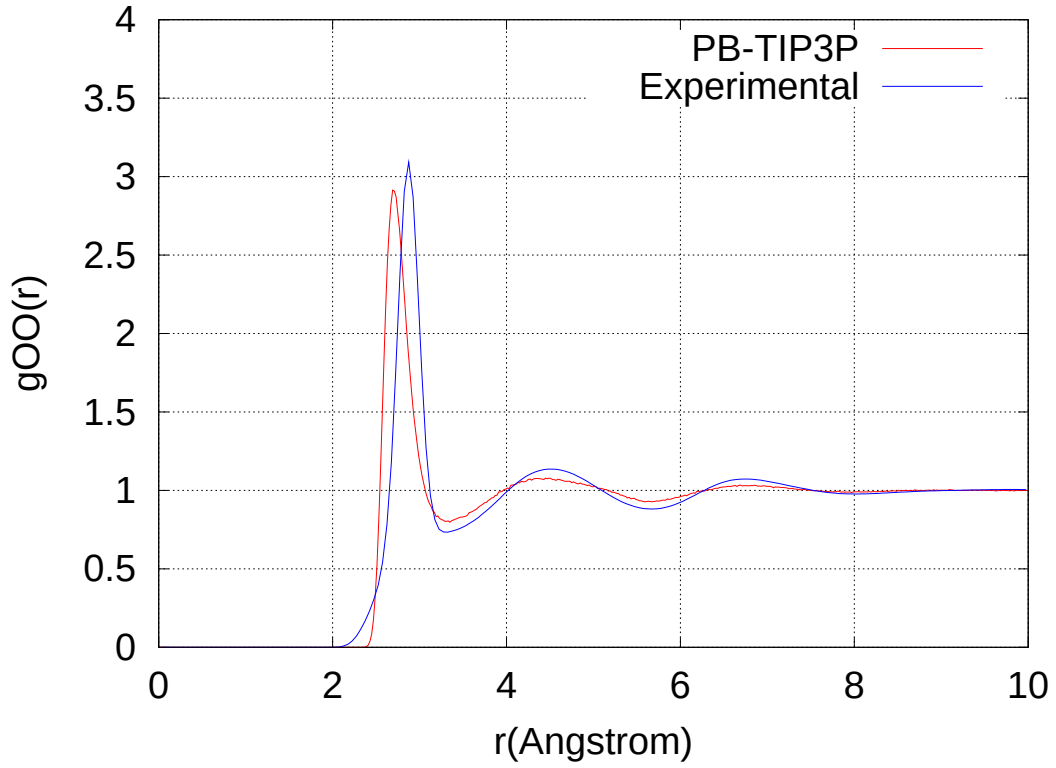
Molecular dynamics (MD) [23] was used to simulate cubic domains (60x60x60Å) filled with PB-TIP3P water in a NPT ensemble that was controlled by a Nose-Hoover thermostat and barostat [51, 99]. Periodic boundary conditions were applied in all directions and a particle-particle particle-mesh (PPPM) solver was used for calculating the long-range electrostatics forces. Bonds and angles involving hydrogen atoms were constrained with the SHAKE algorithm. The simulations were performed for all combinations of the temperatures 298K, 323K and 348K; and the ions molar concentrations (total moles of Na^+ and Cl^- ions per liter water) of 0M , 0.1M, and 1.0M. The target pressure was 1atm for all simulations. An overview of the composition of the simulated systems is presented in Tab. 4.3. After assigning the initial positions and velocities, the potential energy of each system was minimised before equilibrating the systems to the target temperature and pressure for 1ns with a timestep of 0.5fs. The equilibrated systems were then used for the production runs of 4ns with a timestep of 2fs. Every combination of temperature and



(a) RDF for H-H pairs



(b) RDF for O-H pairs



(c) RDF for O-O pairs

FIGURE 4.2: Radial distribution functions of PB-TIP3Pwater, compared with experimental values [2]

.

Table 4.1: Atomic charges and Lennard-Jones parameters for TIP3P and PB-TIP3P water models

Parameter	TIP3P	PB-TIP3P
q_O	-0.8340	-0.8300
q_H	0.4170	0.4150
ϵ_{OO}	0.1521	0.1020
σ_{OO}	3.1507	3.1880
ϵ_{OH}	0.0460	0.0000
σ_{OH}	0.4000	0.0000
ϵ_{HH}	0.0836	0.0000
σ_{HH}	1.7753	0.0000

Table 4.2: Geometry and mechanical parameters for TIP3P models.

	Symbol	Value
Bond stiffness	K	450
Bond length	r_0	0.9572
Angle stiffness	K_θ	55
Angle width	θ_0	104.52

salt conditions was simulated in three independent runs with randomised initial conditions, in order to increase the number of samples to average over. Pressure, temperature and volume were stored at every timestep. The fluid viscosity, η, was calculated with the Green-Kubo relations [100, 101] from the time integral of the autocorrelation function of the off-diagonal components of the pressure tensor,

$$\eta = \frac{\langle V \rangle}{k_B \langle T \rangle} \int_0^\infty \langle P_{ij}(0) P_{ij}(t) \rangle dt \,, \tag{4.1}$$

where $\langle V \rangle$ and $\langle T \rangle$ are the run averages of temperature and pressure, k_B is the Boltzmann constant and P_{ij} one of the off-diagonal components of the pressure tensor. The autocorrelation function (ACF) was computed using a window-averaging procedure analogous

TABLE 4.3: Number of molecules used in the simulations at different salt concentrations

Ion conc.	H_2O	Na^+	Cl^-
0M	7224	0	0
0.1M	7210	7	7
1.0M	7088	68	68

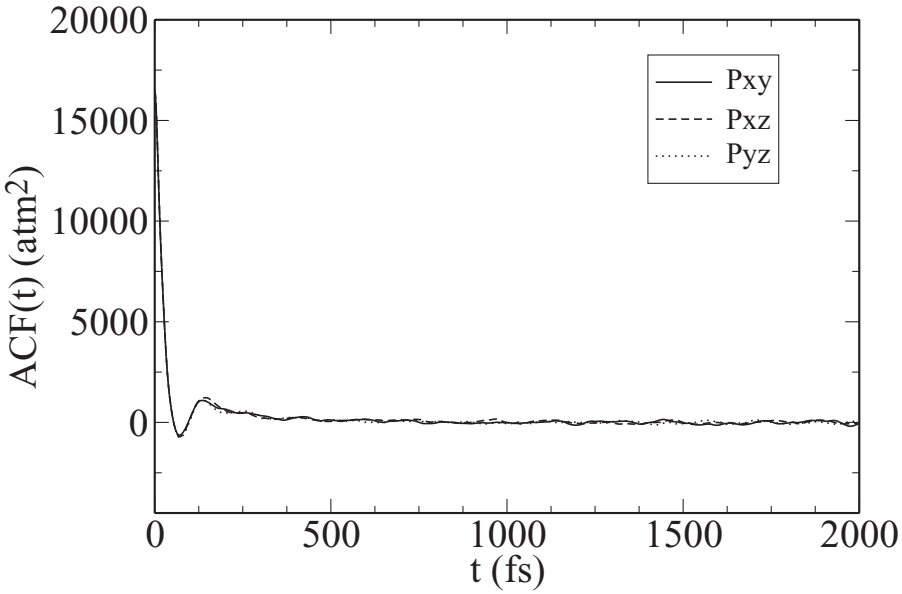
to that described by Nevins and Spera [102] with a time window of 2ps and integrated numerically using the trapezoid rule with a step of 2fs. Through the noisy tail of the ACF, the integral function reaches a plateau after about 1.5ps, but is then subject to fluctuations. Therefore, the limit value was estimated by averaging the value over the last 0.5ps of the time window.

An example of the pressure ACF and its integral function is given in Figs. 4.3(a) and 4.3(b). For every simulation, the procedure was performed for the three independent pressure components P_{xy}, P_{xz} and P_{yz}. Since every condition was simulated three times, each reported value is the average of nine independent estimates. All 27 independent runs were performed using LAMMPS [25], and the results were post-processed using custom C and Octave scripts.

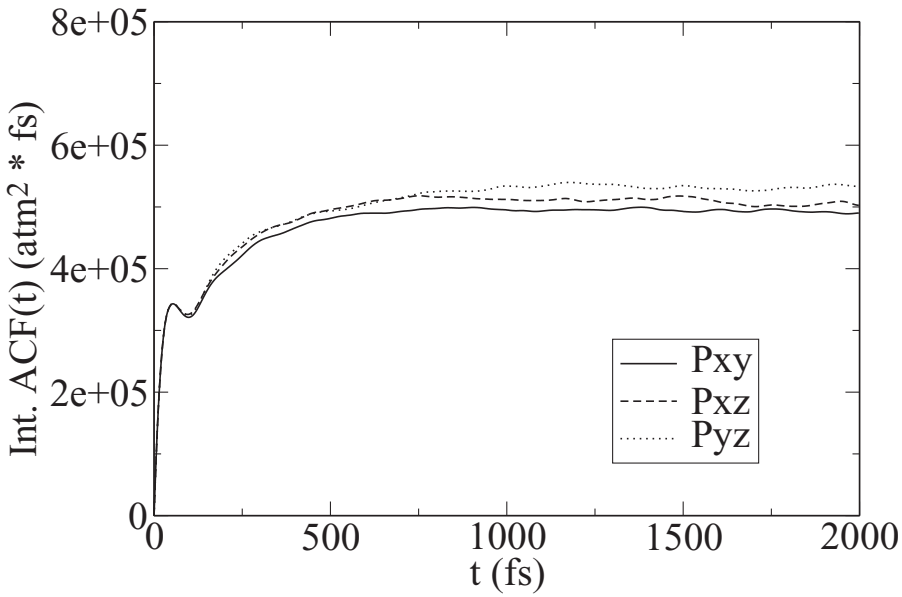
4.3 Results

The results obtained for water viscosity are summarised in Tab. 4.4 and compared with available experimental data for both pure water and saline mixtures. In particular, accurate measurements at 298.15K have been produced for pure water and saline solution by Zhang [5]. For several different conditions of temperature and salt concentration, further data could be found in the papers by Kestin and coworkers [103]. However, for some of the simulated conditions, we were not able to locate suitable experimental values and we have relied on extrapolation. The available measurements were reported to be accurate within 1% of their absolute values [3, 5, 103].

As we can see from Tab. 4.4, the viscosity of the PB-TIP3P model is consistently lower than that of real water and higher than the TIP3P model [98, 104]. The absolute values are about half of the experimental ones. This discrepancy is partly due to the



(a) Plot of window-averaged autocorrelation function over a time-window of 2ps



(b) Time integral of autocorrelation function over a time-window of 2ps

FIGURE 4.3: (a) Computed autocorrelation function; (b) corresponding time integral.

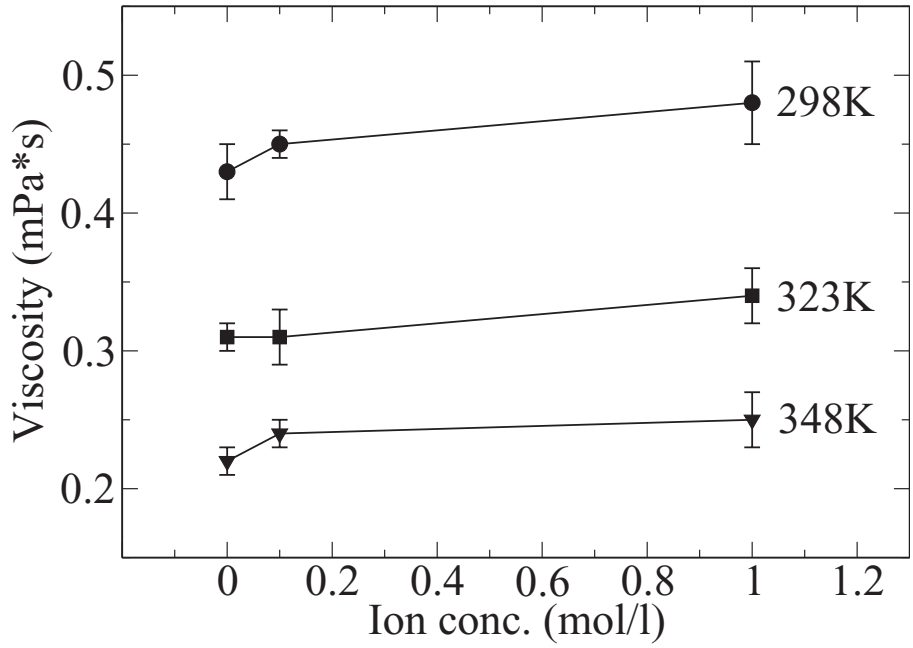


FIGURE 4.4: Summary of resulting viscosity measurements at 298K, 323K and 348K.

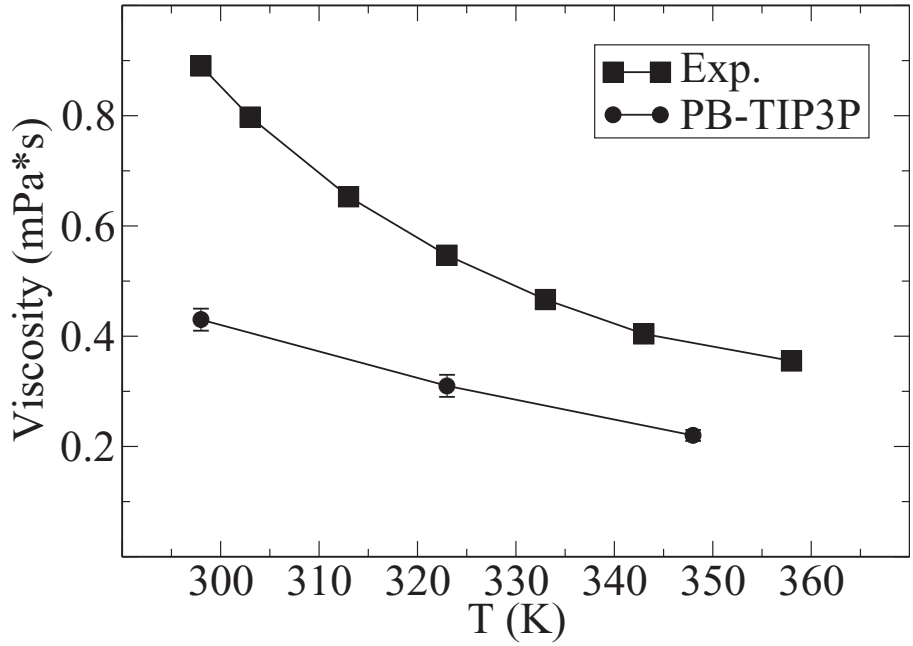


FIGURE 4.5: Viscosity versus temperature, for real water and PB-TIP3P, in no salt conditions.Experimental datapoints were taken from [3].

TABLE 4.4: Results for PB-TIP3P water viscosities and comparison with experimental values

$T(K)$	$M(ions)$	$\eta(mPa \cdot s)$	$\eta_{exp}(mPa \cdot s)$	η/η_{exp}
298	0	0.43±0.02	0.890 [a]	0.48
298	0.1	0.45±0.03	0.893 [a†]	0.50
298	1.0	0.48±0.03	0.914 [a†]	0.53
323	0	0.31±0.01	0.547 [b]	0.57
323	0.1	0.31±0.02	0.550 [c†]	0.56
323	1.0	0.34±0.02	0.576 [c†]	0.59
348	0	0.22±0.01	0.379 [c†]	0.58
348	0.1	0.24±0.01	0.381 [c†]	0.63
348	1.0	0.25±0.02	0.403 [c†]	0.62

[a] Experimental value, from [5]
[b] Experimental value, from [3]
[c] Experimental value, from [103]
† Extrapolated values

simplifications of the model. In particular, the charge distribution of the electron cloud cannt be accurately represented by three point charges whose positions are fixed relative to the atoms. Models like TIP4P and TIP5P, while still rigid and therefore incapable of reproducing polarisation effects, add extra massless charges which mimic the charge distribution more accurately, and are (almost) always more successful in reproducing a wider range of water properties, although greater complexity is not always associated with better results [97].

In addition, the empirical and simplified nature of the model implies a poor predictive capability for those properties that, like viscosity, were not included in the original parametrisation [94]. Nevertheless, by comparison with the experiments [5], the PB-TIP3P model is able to reproduce at least qualitatively the trend of the the viscosity increase due to ions in solution, as shown in Fig. 4.4. For the simulated conditions, the effect of temperature is much more visible than that of the salt concentration.

The viscosity decreases considerably at higher temperatures: at 348K its absolute value is nearly half of the value measured at 298K. For pure water, experimental data are available for a range of temperatures and a comparison is plotted in Fig. 4.5 showing that the viscosity of the PB-TIP3P model follows the same trend as the experimentally obtained

values.

The discrepancy between the results for real water and TIP3P water, shown in Fig. 4.5 is due to two main reasons: first, the model is geometrically simplified, rigid and non-polarisable, which limits its capability to reproduce the correct charge distribution around the molecule; secondly, the *empirical* natural of this class of models means that their parameters are "fitted" in order to reproduce a given set of structural or thermodynamical properties. This implies however that no warranty is given that properties not introduced in the original parametrisation will be reproduced correctly (and this was indeed the case with the viscosity of TIP3P).

It is worth mentioning that our results for TIP3P viscosity in salf-free conditions are in excellent agreement with the values given by both Yeh and Hummer ($0.31 \pm 0.01 Mpa \cdot s$) and Gonzalez and Abascal ($0.321 MPa \cdot s$, although no standard deviation was given in their paper) [98, 105]. Amongst the available models, the best overall performance seems to belong to the TIP4P/2005 model by Abascal and Vega, but it's use has been somewhat limited, at least in part for the reasons outlined in the introduction of this chapter [106].

Summarising, he presented results can be useful in simulations that aim at the extraction of quantitative information about the diffusive properties of small molecules in aqueous saline solution: if the deviation from the behaviour of real water is known, the too low viscosity can be kept into account and the unrealistically high diffusion coefficients can be rescaled accordingly, thus recovering good estimates of the real values.

5

Diffusion coefficients of ssDNA oligomers

5.1 Introduction

The aim of the work presented in this chapter was to evaluate the effect of salt concentration on the diffusive properties of DNA oligomers of different sequence. Salt concentration plays a double role, modifying the carrier fluid viscosity and the mechanical properties of the ssDNA chain [5, 107]. This interplay determines the resulting diffusive behaviour. The effect on the conformation has been investigated calculating the equivalent hydrodynamic radius of the strands. The way the salt concentration affects the solvent viscosity wa kept in to account thanks to the results obtained in the previous chapter. The averaged properties of short oligomers can be used as constitutive elements of a finely discretised model of ssDNA for implementation in particle-fluid models for the simulation of transport phenomena in DNA biosensors.

5.2 Method and simulated cases

All simulations have been carried using the LAMMPS parallel MD simulator using the CHARMM27 force field [25, 36]. All bonds and angles involving H atoms were constrained with SHAKE. The cutoff for non-bonded interactions was set to 10 Angstrom and long-range electrostatics were computed by PPPM method using a grid space of 1 Angstrom. The 3' and 5' ends of the nucleic acid chain were capped with hydroxyl groups. The tetramers $(pA)_4$ and $(pT)_4$, and the octamers $(p8)_4$ and $(p8)_4$ were solvated in cubic boxes of explicit TIP3P water, modified according to the parametrization suggested by Price and Brooks in order to reproduce the solvent structure more accurately [93].
The systems have been neutralized with 3 (for tetramers) and 7 (octamers) Na^+ions and NaCl was added up to two different concentrations, 0.1M and 1.0M respectively.

5.2.1 Simulation boxes

The boxes for $(pA)_4$ in 0.1 and 1.0M solutions included 6755 waters, $8Na^+$ and 5 Cl^- and 6631 waters, $65Na^+$ and 62 Cl^- respectively. The boxes for $(pT)_4$ contained 6761 waters, $8Na^+$ and 5 Cl^-, and 6646 waters, $65Na^+$ and 62 Cl^-. The boxes for $(pA)_8$ contained 6431 waters, $10Na^+$ and 3 Cl^-, and 6646 waters, $66Na^+$ and 59 Cl^-. The boxes for $(pT)_4$ contained 6761 waters, $8Na^+$ and 3 Cl^-, and 6646 waters, $66Na^+$ and 59 Cl^-. The composition of the simulation boxes is summarised in Tab. 5.1.

All runs were performed at 298K and 1 atm in the NpT ensemble using Nose-Hoover thermostat and barostat [54]. The systems have been minimized and then equilibrated for 1ns with a timestep of 0.5fs, and then run for 4 more nanoseconds to allow ion equilibration around the molecule [108]. Each of the 8 cases has been run in 3 independent simulations with randomized initial configurations, for a total simulated time of 600ns. The position of the ssDNA centre of mass was stored every 0.2 ps.

Table 5.1: Composition of the simulation boxes. All runs were performed at 298K and 1atm.

Seq.	C(NaCl)	*Atoms*	H_2O	Na^+	CL^-
4A	0.1M	20409	6755	8	5
4A	1.0M	20181	6631	65	62
4T	0.1M	20423	6761	8	5
4T	1.0M	20195	6646	65	62
8A	0.1M	20305	6678	10	3
8A	1.0M	20081	6566	66	59
8T	0.1M	20317	6682	10	3
8T	1.0M	20093	6570	66	59

5.2.2 Calculation of the diffusion coefficient

The diffusion coefficients were calculated from the random fluctuations of the molecule centre of mass using the well-known Einstein relation for diffusion in three dimensions,

$$D = \lim_{t\to\infty} \frac{\partial}{\partial t} \frac{\langle MS\, D(t)\rangle}{6} \tag{5.1}$$

The slope of a linear fitting of the Mean Square Displacement (MSD) over a time window Δt was evaluated after computing the MSD by window-averaging over the production run, as illustrated for example by [102]. The chosen window length was $\Delta t = 100$ps, while the spacing between the time origins of the windows was 10ps. The MSD converges very neatly to the expected linear behaviour, and the uncertainty on the slope is several orders of magnitude smaller than the absolute value, thus providing a solid estimat, as shown in Fig. 5.1.

5.2.3 Correction for finite-size effects

A suitable treatment for keeping account of the effect of PBC and the limited size of the simulation box was developed by Yeh and Hummer, based on an empirical correction of the original theoretical analysis for point-like particles by Dünweg and Kremer [98, 109]. The simulated diffusion coefficient D_s computed by means of Eq. (5.1) is first corrected

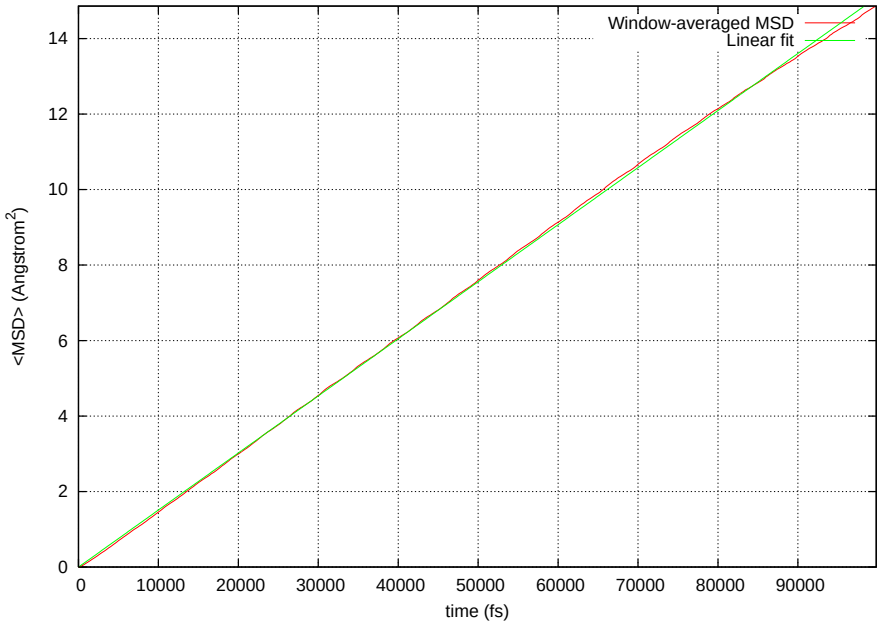


FIGURE 5.1: Summary of corrected hydrodynamic radiuses at different salt concentrations.

for finite-size effects as:

$$D = D_s + \frac{k_B T \xi \alpha}{6\pi\eta L}, \tag{5.2}$$

Where $k_B T$ is the reduced temperature, η is the dynamic viscosity of the used water model, $\xi = 2.873$ is a constant that synthetises the correlation effect summed over all periodic images of a cubic lattice, $\alpha = 0.76$ is an empirical correction for finite-size molecules, and L is the linear size of the box [98, 109].

5.2.4 Correction for low solvent viscosity

Assuming that viscosity alone doesn't have a remarkable effect on the hydrodynamic radius of the molecules, the Stokes-Einstein equation provides the scaling relationship for diffusion in two solvents of different viscosity (in our case, real and model water), as

$$\eta_w D_w = \eta D, \quad or \quad D_w = \frac{\eta}{\eta_w} D, \tag{5.3}$$

where D_w is the estimated diffusion coefficient in real water. The dynamic viscosity of TIP3P water is known to be very low, and it has been reported by some authors to be 0.31±

$0.01 mPa \cdot s$ [98]. However, as shown in the previous chapter, the Price-Brooks TIP3P water model has a comparatively higher viscosity. The experimental values for water viscosity in no-salt, 0.1M, and 1.0M NaCl solutions are respectively $0.89 \pm 0.02 mPa \cdot s$, $0.90 \pm 0.02 mPa \cdot s$ and $0.97 \pm 0.02 mPa \cdot s$. [5]. The comparison with model water in similar conditions, and the corresponding correction factors as summarized in Tab. (5.2). For the solutions at 0.1M and 1.0M NaCl the solvent viscosity correction factors η/η_w are respectively 0.50 and 0.49. The corrected diffusion coefficients are summarised in Tab. (5.3).

5.2.5 Correction for constant drift

Every simulation was initialised with randomised atomic velocities; however, numerical errors can cause the center of mass of the simulation box to possess a small residual velocity. The result is a system slowly drifting in space, and the resulting motion of the molecule in solution is the sum of a random brownian motion and a translation at constant velocity. It is known that, in the case of brownian motion with drift, the MSD becomes quadratic instead of linear [110]; moreover, no simple analytical relation exists between a quadratic MSD and the corresponding diffusion coefficients. Therefore it was necessary to correct the drift artifact by subtracting the total velocity of the system's centre of mass. This way we were able to recover the expected linear shape of the window-averaged MSD.

5.3 Results

Salt concentration is known to have an effect on both fluid viscosity [5] and ssDNA persistence length [107], which influences the chain persistence length and therefore its conformation in solution and, potentially, its hydrodynamic radius. We have used molecular dynamics simulations to investigate the effect of salt concentrations on the diffusive properties of ssDNA oligomers. Finite-size effects have been taken into consideration by means of opportune correction coefficients. The discrepancy between the viscosity of real and model water has also been considered.

The nucleotide sequences of the simulated ssDNA strands were chosen as two limit cases, polyadenine and polythymine, which show respectively the maximum and minimum base stacking tendency and therefore the maximum and minimum expected persistence length.

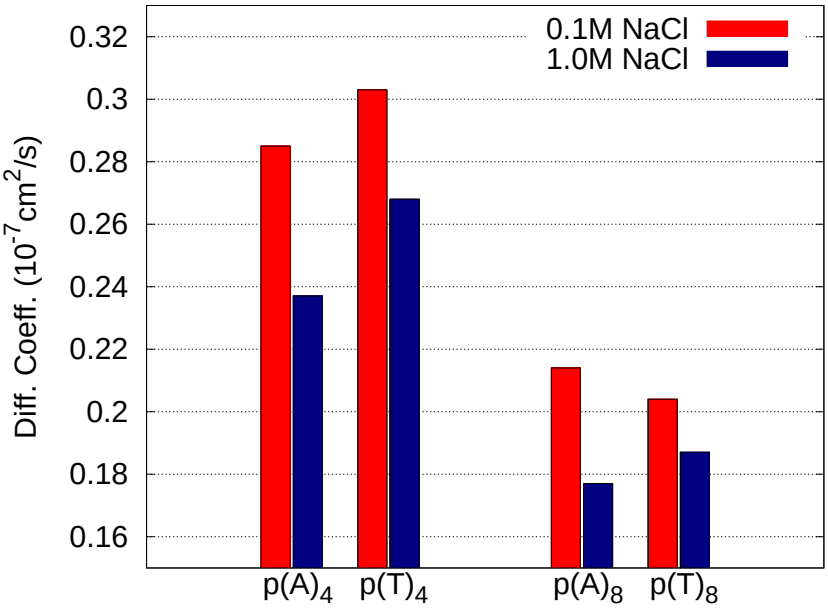


Figure 5.2: Summary of computed diffusion coefficients at different salt concentrations.

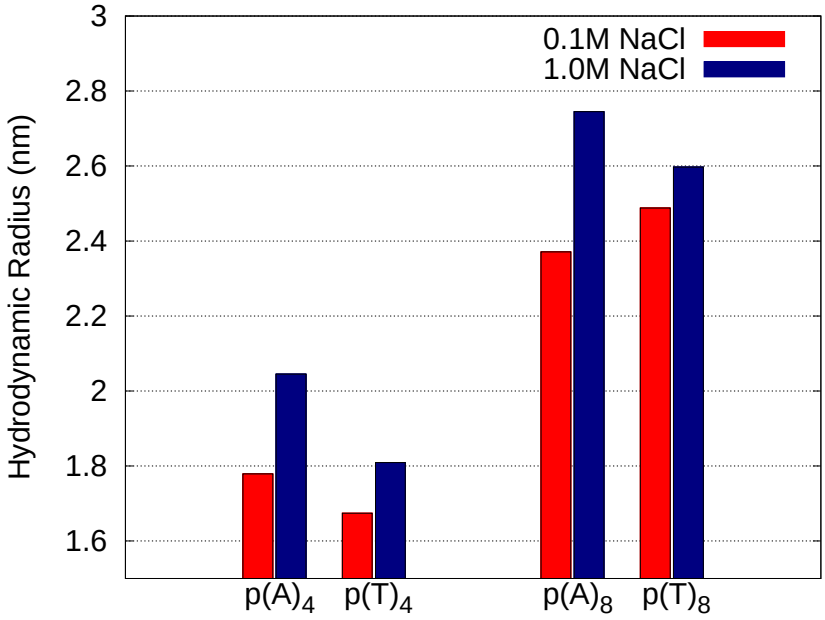


Figure 5.3: Summary of corrected hydrodynamic radiuses at different salt concentrations.

Table 5.2: Viscosity of real water, (extrapolated from [5]) and Price-Brooks TIP3P water at 298K and 1 atm at different NaCl molar concentrations (values in $mPa \cdot s$).

C(NaCl)	$\eta_{PB-TIP3P}$	η_{H_2O}(exp.)	$\eta_{PB-TIP3P}/\eta_{H_2O}$
0	0.43 ± 0.02	0.89	0.48
0.1M	0.45 ± 0.01	0.90	0.49
1.0M	0.48 ± 0.01	0.97	0.50

Table 5.3: Computed (D_s) and corrected (D_w) diffusion coefficients (in $10^{-5} cm^2/s$) and corrected hydrodynamic radius R_h (in nm)

Seq.	C(NaCl)	D_s	D_w	R_h
4A	0.1M	0.394 ± 0.025	0.285 ± 0.012	1.779 ± 0.008
4A	1.0M	0.319 ± 0.020	0.237 ± 0.010	2.045 ± 0.088
4T	0.1M	0.431 ± 0.015	0.303 ± 0.008	1.674 ± 0.043
4T	1.0M	0.382 ± 0.025	0.268 ± 0.012	1.809 ± 0.082
8A	0.1M	0.253 ± 0.003	0.214 ± 0.002	2.371 ± 0.019
8A	1.0M	0.197 ± 0.020	0.177 ± 0.009	2.745 ± 0.148
8T	0.1M	0.233 ± 0.011	0.204 ± 0.005	2.488 ± 0.066
8T	1.0M	0.217 ± 0.008	0.187 ± 0.004	2.597 ± 0.055

The lack of experimental data for short oligonucleotide doesn't allow a direct validation of the simulation results. However, experimental measurements have been performed by Nkodo and coworkers for slightly longer strands [4], at higher salt conditions, as shown in Fig 5.4. Our results in high-salt conditions are very close to the values extrapolated from the experiment, whose least-square fitting gives:

$$D = 5.525 \cdot N^{-0.585} \tag{5.4}$$

where D is the diffusion coefficient and N the number of nucleotides. Although the solvent conditions are not exactly the same, our results can be compared quantitatively, with decent accuracy, with values extrapolated from Eq. 5.4, because over 1.0M the viscosity of the solvent can be assumed not to be massively affected by higher salt concentration [5].

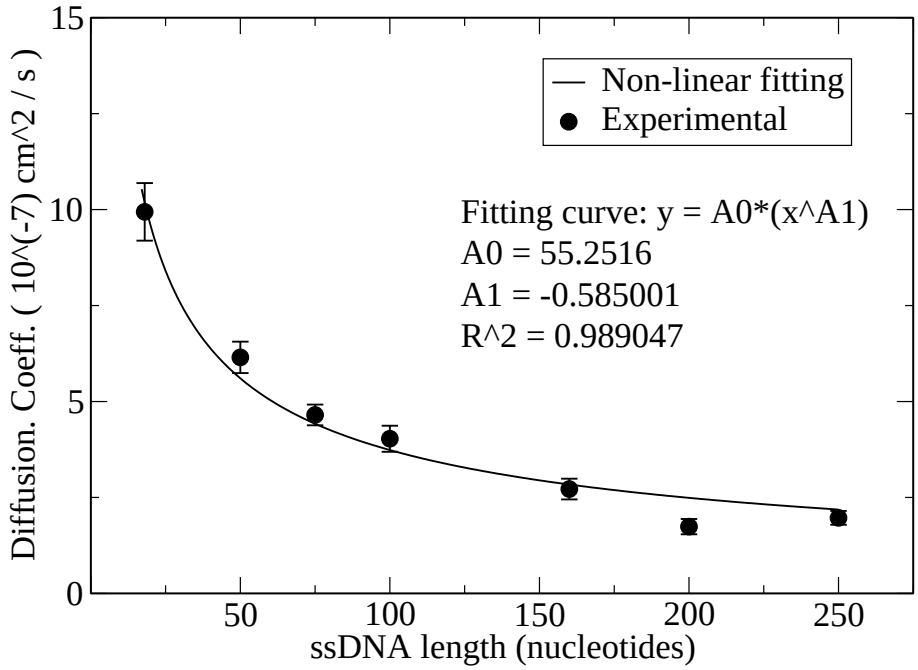


FIGURE 5.4: Least square fitting of the experimental results by Nkodo and coworkers, for ssDNA in high-salt conditions [4].

The surprisingly good agreement shown in Fig. 5.5 suggests that the method described in this chapter can actually be used for the quantitative calculation of diffusion coefficients in saline solution. The same approach, in principle, can be fruitfully applied to other small molecules (i.e. drugs).

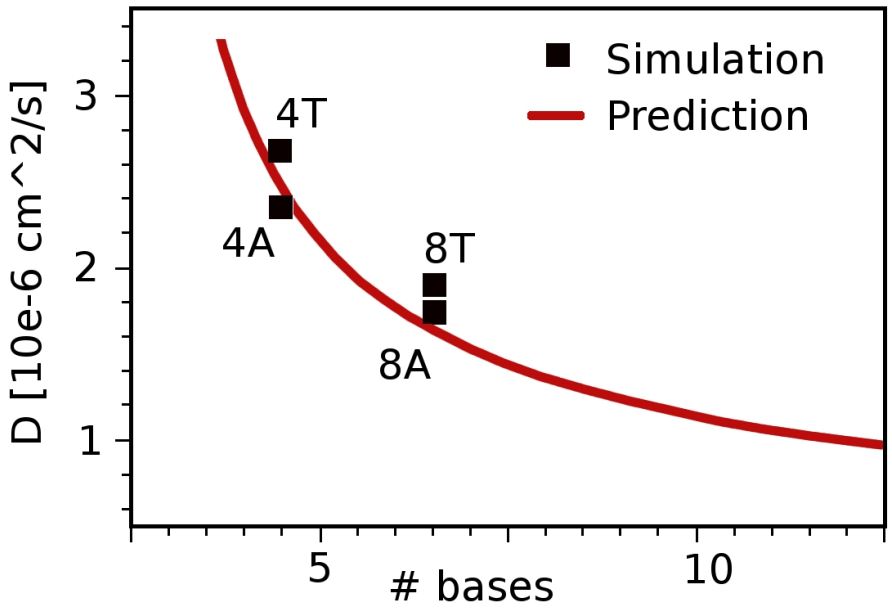
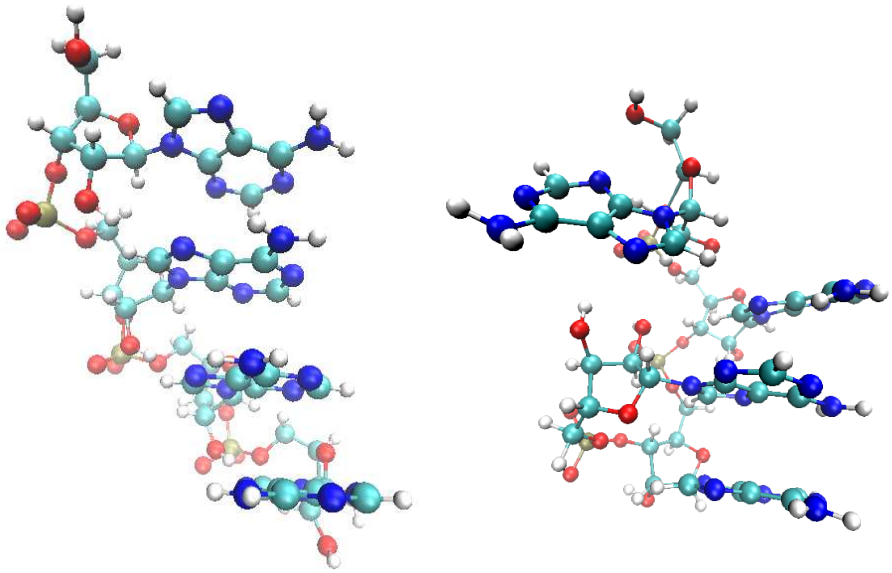


Figure 5.5: Comparison between the diffusion coefficient predicted by a least-square fit of experimental results, and the results of our simulations in high-salt conditions

.



(a) (b)

Figure 5.6: An example of configurational fluctuation of ssDNA in solution: initial orderly configuration of a polyadenine tetramer, (left), and configuration after 10ns equilibration, with unstacking of one of the terminal residues (right).

6

Implementation of brownian perturbations for isolated particles

6.1 Introduction

Modelling the brownian motion is and important part of a mesoscopic simulation of particles and polymer transport, because random agitation is the physical mechanism that triggers diffusive processes.

The explanation and description of Brownian motion was studied and formalized by the likes of Einstein, Langevin, Smoluchowski [61]. Langevin's approach is to include explicitly the random force in the hamiltonian equations of motion of the colloid particle (assumed to be spherical) thus obtaining a stochastic ordinary differential equation which can be solved formally using the boundary condition $\mathbf{v}(0) = \mathbf{v}_0$,

$$\frac{d\mathbf{v}}{dt} = -\beta\mathbf{v} + \mathbf{A}(t) \Rightarrow \mathbf{v} - \mathbf{v}_0 e^{-\beta t} = e^{-\beta t} \int_0^t e^{\beta\tau}\mathbf{A}(\tau)d\tau \tag{6.1}$$

where $\beta = 6\pi R\eta/M$ is the friction coefficient, m and R are the mass and the equivalent hydrodynamic radius of the particle, η is the newtonian viscosity of the solvent, and $A(t)$ is the stochastic acceleration, that needs to be characterised for Eq. 6.1 to be of any practical use. Given a certain simulation timestep Δt, what we need to know is the *cumulative* effect of all the random accelerations caused by collisions occurred in the interval $[t, t + \Delta t]$, on the resulting position and velocity. When the "random flight" condition $\Delta t \gg \beta^{-1}$ does *not* apply, the displacement over the timestep is no longer purely random, but it's affected by the velocity at the previous timestep, which must be calculated as well. This can be often the case in mesoscopic fluid-particle simulations [111]. If we call $\mathbf{r}$ and $\mathbf{r}_0$ the particle's position vector and initial position, under the hypotheses that

- the timescale of the random force variation is much faster than any other physical quantity in play;
- energy equipartition will hold for both solvent molecules and particles, so the particle velocity distribution will also converge to a Maxwellian curve

the joint probabilty distribution function (PDF) $f(\mathbf{d}, \mathbf{u})$ for the vectors

$$\mathbf{d} = \mathbf{r} - \mathbf{r}_0 - \beta^{-1}\mathbf{v}_0(1 - e^{-\beta\Delta t}) \tag{6.2}$$

$$\mathbf{u} = \mathbf{v} + \mathbf{v}_0 e^{-\beta\Delta t}, \tag{6.3}$$

(from which displacement and velocity can be readily obtained) over a timestep Δt, is given by [112]:

$$f(\mathbf{d}, \mathbf{u}) = \frac{1}{(2\pi)^3(fg - h^2)^{3/2}} \exp\left[-\frac{g|\mathbf{d}|^2 - 2h\mathbf{d}\cdot\mathbf{u} + f|\mathbf{u}|^2}{2(fg - h^2)}\right], \tag{6.4}$$

where the coefficients f, g, h are given by

$$f = \frac{1}{\beta^2}\frac{k_B T}{M}(2\beta\Delta t - 3 + 4e^{-\beta\Delta t} - e^{-2\beta\Delta t}) \tag{6.5}$$

$$g = \frac{k_B T}{M}(1 - e^{-2\beta\Delta t}) \tag{6.6}$$

$$h = \frac{1}{\beta}\frac{k_B T}{M}(1 - e^{-\beta\Delta t})^2 \tag{6.7}$$

where T is the absolute temperature and k_B the Boltzmann constant. Modelling brownian fluctuations involves sampling six random numbers $d_x, d_y, d_z, u_x, u_y, u_z$, from the joint PDF

given by Eq. 6.4. In the following, a numerical implementation is described.

6.2 Method

For numerical implementation, Eq. 6.4 must first be recast in the canonical form of a Gaussian multivariate.
When the vector of mean values is **0** (as in brownian fluctuations) its general form is:

$$f(\mathbf{w}) = \frac{1}{(2\pi)^{n/2}|\mathbf{C}|^{1/2}} exp\left[-\frac{1}{2}\mathbf{w}\mathbf{C}^{-1}\mathbf{w}^T\right] \tag{6.8}$$

where **w** is an n-dimensional row vector of random variables with zero mean, and **C** is the $n \times n$ covariance matrix, simmetric and positive definite.The off-diagonal terms of **C** are responsible for the mutual correlations between the stochastic variables. If we write the vector **w** as:

$$\mathbf{w} = \begin{bmatrix}\mathbf{d} & \mathbf{u}\end{bmatrix}^T = \begin{bmatrix}d_x & d_y & d_z & u_x & u_y & u_z\end{bmatrix}^T \tag{6.9}$$

the exponent in Eq. 6.4 can be rewritten as:

$$\mathbf{w}^T\mathbf{C}^{-1}\mathbf{w},$$

where the inverse matrix $\mathbf{C}^{-1}$ and the covariance matrix **C** are found to be

$$\mathbf{C}^{-1} = \frac{1}{fg-h^2}\begin{bmatrix} g & 0 & 0 & -h & 0 & 0 \\ 0 & g & 0 & 0 & -h & 0 \\ 0 & 0 & g & 0 & 0 & -h \\ -h & 0 & 0 & f & 0 & 0 \\ 0 & -h & 0 & 0 & f & 0 \\ 0 & 0 & -h & 0 & 0 & f \end{bmatrix} \tag{6.10}$$

and

$$\mathbf{C} = \begin{bmatrix} f & 0 & 0 & h & 0 & 0 \\ 0 & f & 0 & 0 & h & 0 \\ 0 & 0 & f & 0 & 0 & h \\ h & 0 & 0 & g & 0 & 0 \\ 0 & h & 0 & 0 & g & 0 \\ 0 & 0 & h & 0 & 0 & g \end{bmatrix} \tag{6.11}$$

The covariances (i.e. off-diagonal terms) are non-zero only between components of position and velocity along the same cartesian direction. Components along different directions are non correlated, and actually independent, their joint PDF being Gaussian [113]. Consequently, the distribution in Eq. 6.4 can be factorized into three identical distributions:

$$f(d_i, u_i) = \frac{1}{(2\pi)^2(fg - h^2)} exp\left[-\frac{gd_i^2 - 2hd_iu_i + fu_i^2}{2(fg - h^2)}\right], \quad i = x, y, z. \tag{6.12}$$

Now if we split $\mathbf{w}$ into three vectors

$$\mathbf{w}_i = \begin{bmatrix} d_i & u_i \end{bmatrix}, \quad i = x, y, z, \tag{6.13}$$

the Eq. 6.12 can be rewritten as canonical Gaussian bivariate (consisting of a displacement and a velocity component each)

$$\frac{1}{(2\pi)^2|\mathbf{B}|} exp\left[-\frac{1}{2}\mathbf{w}_i^T\mathbf{B}^{-1}\mathbf{w}_i\right], \quad i = x, y, z, \tag{6.14}$$

where the covariance matrix and its inverse are:

$$\mathbf{B} = \begin{bmatrix} f & h \\ h & g \end{bmatrix}, \quad \mathbf{B}^{-1} = \frac{1}{fg - h^2}\begin{bmatrix} g & -h \\ -h & f \end{bmatrix}, \tag{6.15}$$

and their coefficients are already known. A common method for the implementation of Gaussian multivariates requires the Cholesky factorization of the covariance matrix [23]. Namely, if $\mathbf{x} = [x_1, x_2]^T$ is a vector of independent normal variables, and L a lower triangular matrix such that $\mathbf{L}\mathbf{L}^T = \mathbf{B}$, with $\mathbf{B}$ positive definite, then the vector $\mathbf{y} = \mathbf{L}\mathbf{x}$ follows a Gaussian multivariate with 0 mean vector and covariance matrix $\mathbf{B}$ [113]. The matrix $\mathbf{L}$ is the so-called Cholesky factor of $\mathbf{B}$.
Our bidimensional case is quick to derive, and we get:

$$\mathbf{L} = \begin{bmatrix} \sqrt{f} & 0 \\ -\frac{h}{\sqrt{f}} & \sqrt{g - \frac{h^2}{f}} \end{bmatrix}, \tag{6.16}$$

that yelds the recipe for each of the component of displacement and velocity, for $i = x, y, z$:

$$d_i = x_{i1}\sqrt{f} \tag{6.17}$$

$$u_i = x_{i1}\left(-\frac{h}{\sqrt{f}}\right) + x_{i2}\sqrt{g - \frac{h^2}{f}}, \tag{6.18}$$

where x_{i1} and x_{i2} are independent normal variables that can be generated for example by a Box-Muller or Marsaglia polar algorithm [114], provided that a good-quality random generator is available [115].
The definition of the two random vectors $\mathbf{d}, \mathbf{u}$, and the knowledge of their PDF, provide the numerical scheme for the generation of a brownian trajectory. Be Δt the timestep length, and $\mathbf{r}^t$ and $\mathbf{v}^t$ position and velocity at the timestep t, rearranging the terms in the Eq. 6.2 we get:

$$\mathbf{r}^{t+\Delta t} = \mathbf{r}^t + \beta^{-1}(1 - e^{-\beta\Delta t})\mathbf{v}^t + \mathbf{d}^t \tag{6.19}$$

$$\mathbf{v}^{t+\Delta t} = (e^{-\beta\Delta t})\mathbf{v}^t + \mathbf{u}^t, \tag{6.20}$$

where $\mathbf{d}^t$ and $\mathbf{u}^t$ are correlated random vectors whose components are generated as described above.

6.3 Calculation of relevant parameters from all-atom simulation

The model described in the previous section was implemented in a custom C code using the WELL random number generator [115], and parametrised according to the data collected from an all-atom 40ns molecular dynamics MD simulation of a ssDNA polyadenine tetramer in saline water at 1.0M ion concentration, with timestep 1fs in order to allow a comparison between the predicted diffusion coefficients. The resulting model parameters, in SI units, are summarised in Tab. 6.1.

The resulting elements of the covariance matrix

$$\mathbf{B} = \begin{bmatrix} f & h \\ h & g \end{bmatrix} \tag{6.21}$$

are respectively:

TABLE 6.1: Summary of brownian motion parameters for a particle simulated at 298K and 1atm.

Parameter	Symbol	value	Units
Hydrodinamic radius	R_h	$0.818 \cdot 10^{-9}$	m
Particle mass	M	$2.108 \cdot 10^{-24}$	Kg
Temperature	T	298.0	K
Friction coeff.	β	$7.095 \cdot 10^{-12}$	s^{-1}
Solvent viscosity	η	$0.97 \cdot 10^{-3}$	Pa·s.

- $f = 5.601 \cdot 10^{3} \mathrm{pm}^2$
- $g = 1.473 \cdot 10^{3} \mathrm{pm}^2 \mathrm{ps}^{-2}$
- $h = 7.075 \cdot 10^{1} \mathrm{pm}^2 \mathrm{ps}^{-1}$

The diffusion coefficient predicted by the all-atom simulation was $D = (0.275 \pm 0.003) \cdot 10^{-9} m^2 s^{-1}$.
The decorrelation time of the displacement vector autocorrelation was $\tau = 1.2ps$.
The decorrelation time and the diffusion coefficient from the MD simulation have been used to validate the model.
First, the autocorrelation function (ACF) for the displacement vector was computed. The resulting decay curve has a time constant of $0.136ps$, very close to the value of $0.122ps$ given by the all-atom simulation, as shown in Fig. 6.1.
Secondly, the diffusion coefficient of the brownian particle was calculated from the slope of the window-averaged the mean square displacement (MSD) of a trajectory of 10^6 steps, applying Einstein's relation:

$$D = \lim_{t \to \infty} \frac{\langle |\mathbf{r(t)} - \mathbf{r_0}|^2 \rangle}{6t}, \tag{6.22}$$

and validated against the results from the all-atom simulation, again with good agreement (Fig. 6.2).

6.4 Notes on the simulation of diffusing particles

The model here described is could be defined as purely kinematic, because apparently no numerical integration is performed on the newtonian equations of motion of the particle.

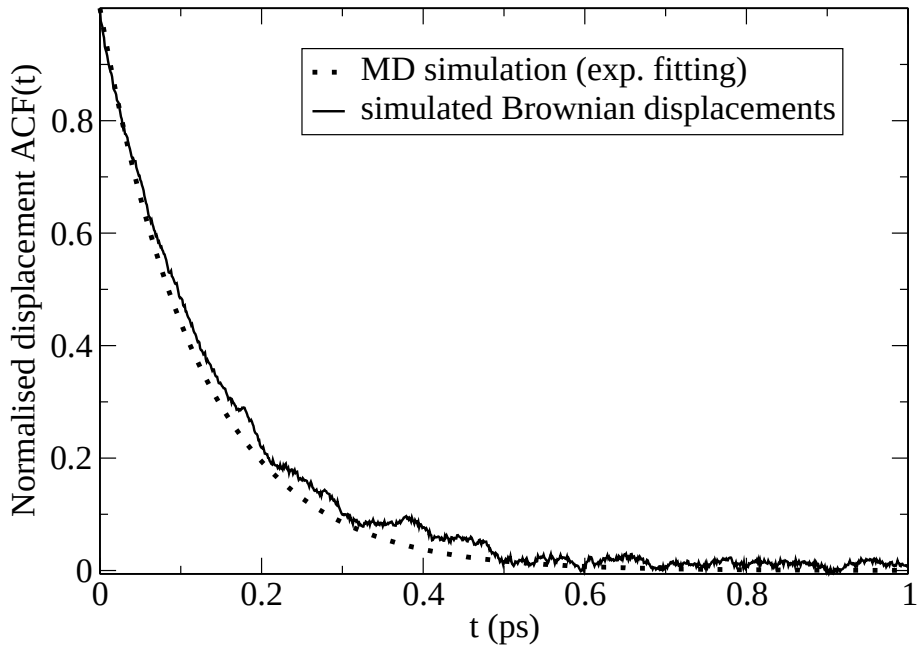


FIGURE 6.1: Computed autocorrelation (solid line) and exponential fitting of the ACF from MD simulation (dotted) for the brownian displacements calculated with a timestep of $0.001ps = 1fs$.

In fact, the brownian perturbation applied at every timestep Δt mimics the total effect of a large number fluctuations (collisions) over a span of time equal to the timestep itself. The Chandrasekhar model is an analytical integration of the stochastic Langevin equations of motion, which yelds the resulting displacements and velocity perturbations for a given timestep. The solution of a stochastic differential equation is not a function, but a probabilty density. Therefore, the resulting brownian displacements and velocity perturbations are opportune random variables, whose probabilty density is a function of several parameters (fluid viscosity particle diameter, and timestep). In this respect, the model can be seen as a direct implementation of an analytical solution, rather than a numerical integration. Now if we take a closer look at the shape of Eq. 6.19

$$\begin{aligned} \mathbf{r}^{t+\Delta t} &= \mathbf{r}^t + \beta^{-1}(1 - e^{-\beta\Delta t})\mathbf{v}^t \quad + \quad \mathbf{d}^t \\ \mathbf{v}^{t+\Delta t} &= \underbrace{(e^{-\beta\Delta t})\mathbf{v}^t}_{\text{“DETERMINISTIC”}} \quad + \underbrace{\mathbf{u}^t}_{RANDOM} \end{aligned}$$

we can see that for timesteps comaprable with $1/\beta$, the perturbations are not purely random, but is the sum of two contributions: a random one, and one that depends on the

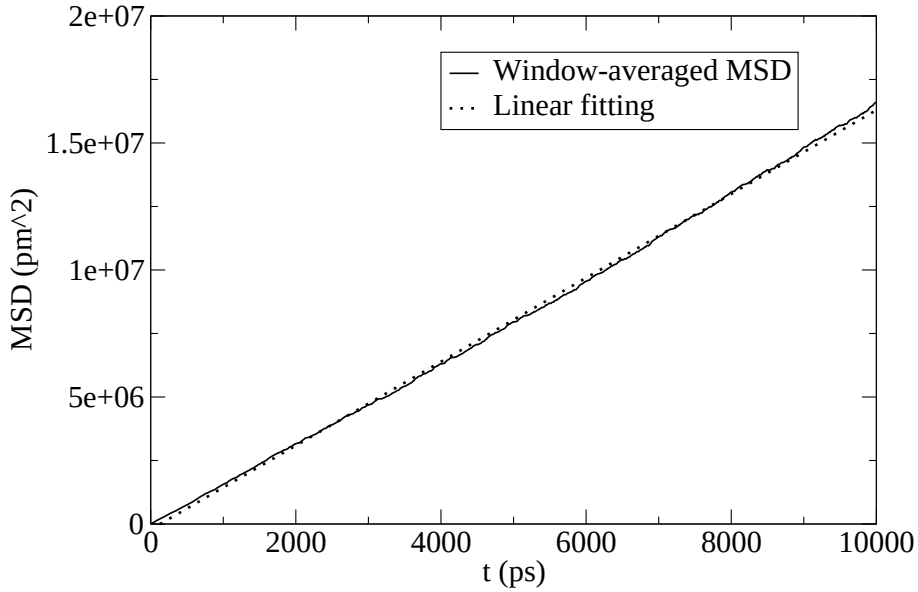


FIGURE 6.2: Window-averaged MSD (solid line) and linear fitting (dotted) for a trajectory of 10^5 timesteps of 10ps each. The slope of the line is proportional to the diffusion coefficeint. The theoretical diffusion coefficient, $0.275 \cdot 10^{-9} m^2/s$ is correctly reproduced.

velocity at the previous timestep. That implies that the particle partially remembers its previous velocity, but this effect decays rapidly (exponentially) as the timestep becomes longer (and therefore we integrate in time over a larger timestep, where more collisions happen with solvent molecules). When the timestep length increases, the exponential terms become very small and we reach the random flight regime. As a result, trajectories simulated with very short timesteps look much more like regular lines, whereas for longer timesteps the trajectory becomes extremely noisy, as shown in Fig. 6.4.

It can be noted that the necessity to include the perturbations on the velocity only arises when the timestepping in comparable with the characteristic time. When the timestep gets "long enough", the exponential in Eq. 6.19 decay to zero, i.e. during the timestep the particle undergoes enough collision to lose memory of the initial velocity, the random flight condition holds and the velocity at each timestep becomes purely random. In order to quantify the timestep length that ensures random-flight conditions, we propose to use the comparison between two estimates of the diffusion coefficient D. On on hand we have the Stokes-Einstein formula:

$$D = \frac{k_B T}{6\pi\eta R_h}, \tag{6.23}$$

and on the other hand, we can use the Einstein relation, which is only correct under the

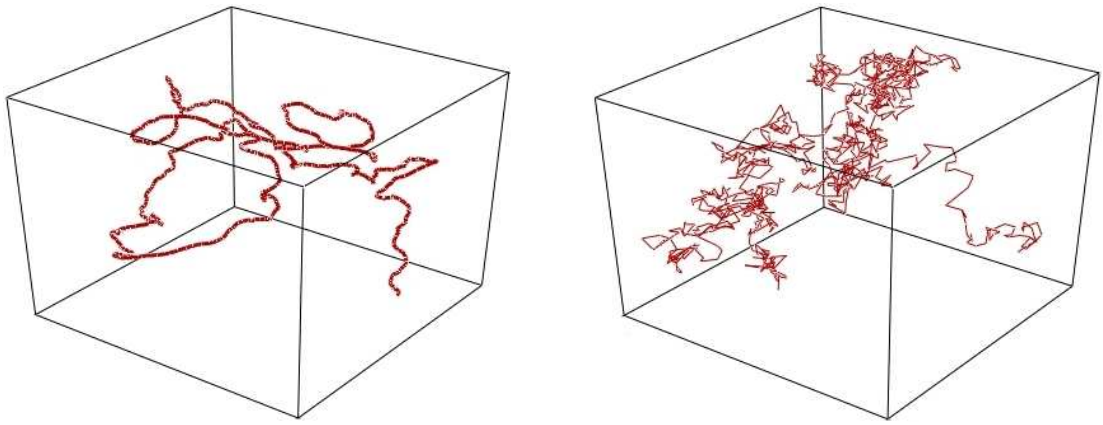


FIGURE 6.3: Effect of the timestep on the level of correlation for brownian perturbation: a trajectory generated with short timestep of 1fs, (left) and another generated with a longer timestep of 2500fs (right). In both cases the characteristic time of the motion of the brownian particle was $1/\beta = 0.12ps$.

assumption that the random-flight condition holds:

$$D_{ER} = \lim_{t\to\infty} \frac{\langle |\mathbf{r}(t) - \mathbf{r}_0|^2 \rangle}{6t}. \tag{6.24}$$

The idea is that the two estimates will agree in random-flight conditions, i.e. $D_{ER}/D \simeq 1$, and disagree otherwise. The results are shown in Fig. 6.4. The empirical rule is that the random flight hypotheses is satisfied for $\Delta t \geq 20\tau$. The results obtained for timesteps much shorter than the characteristic time, curiously, show a good agreement with the correct value. However, this is more an insight in the mathematical behaviour of the chosen model, rather than a physically meaningful result: timesteps so short would contradict the hypotheses under which the equations used here were derived.

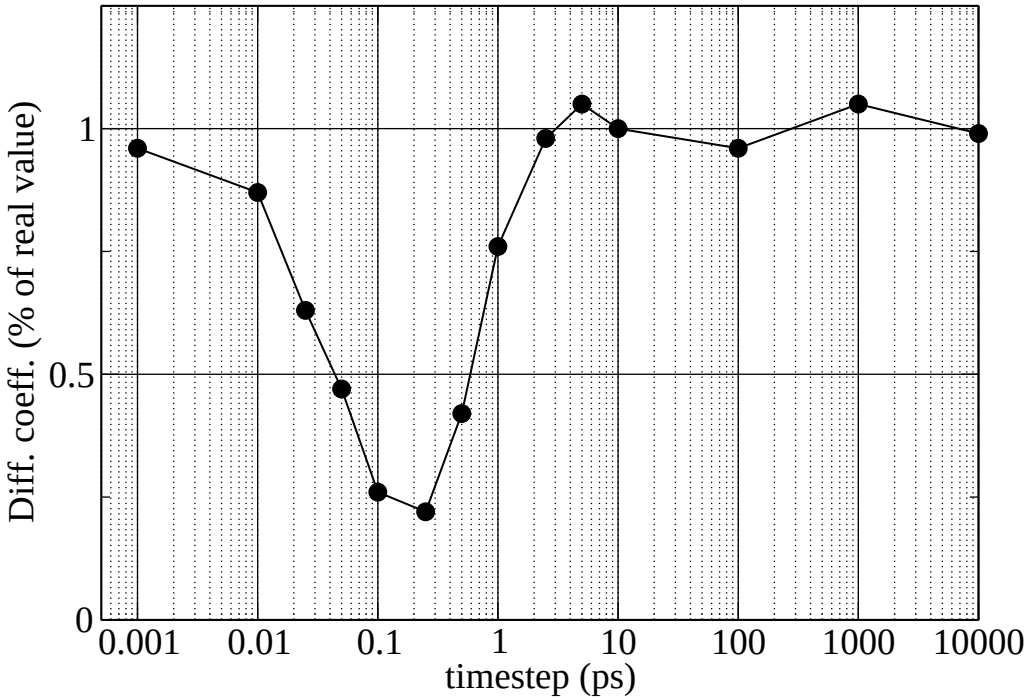


FIGURE 6.4: Plot of the ratio D_{ER}/D between the diffusion coefficients predicted by the Einstein relation and the Einstein-Stokes formula, for different timesteps.

Part IV

Multiscale modelling of nucleic acids

7

Mapping atoms onto coarse-grained sites

The coarse graining procedure can be divided into two main steps. First, the atomistic model is mapped topologically onto a reduced set of CG "sites" (or superatoms). Secondly, the interaction potentials are determined for the reduced model in terms of bonded and non-bonded interactions. In the following we describe the approach we have used to model water and DNA.

7.1 Definition of the reduced model

A very straightforward method for the choice of the CG sites position is to locate them onto the centre of mass of a molecule or chemical group. In this way, the "mapping operator", i.e. the function that links the atomistic coordinates the the coarse-grained coordinates, is linear (more precisely, it's a weighted average of the atomistic positions).
Nucleic acids are biopolymers whose monomers, called nucleotides, are made of a phosphate group, a ribose or deoxyribose, and a nitrogenous base. DNA is a polymer of adenine, guanine, cytosine and thymine; RNA replaces thymine with uracil. Here we will focus on DNA only.

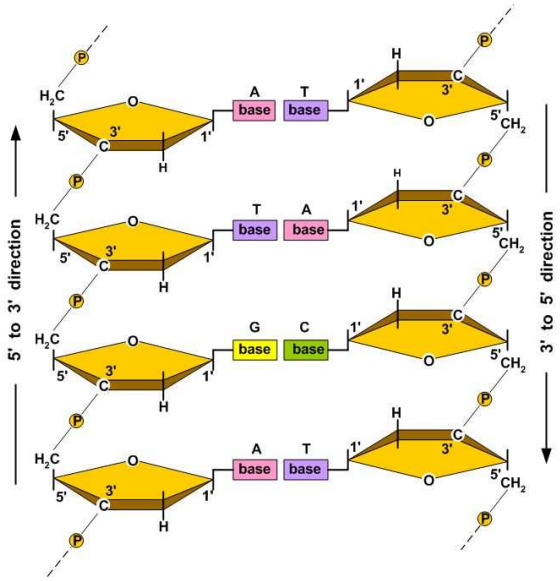


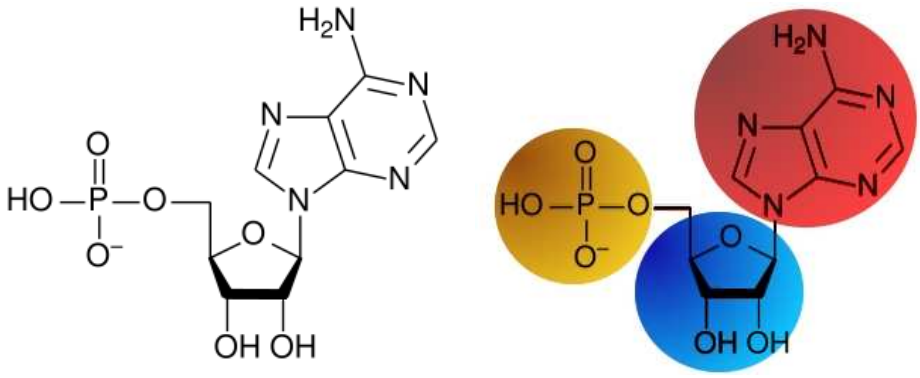
FIGURE 7.1: Schematic representation of DNA double strand structure, and base pairing. URL: http://rh.healthsciences.purdue.edu/vc/theory/dna/index.html

Phosphate and sugar build the polymer backbone, whilst the bases are lateral groups. Each base can interact in a highly specific way with its complementary, forming hydrogen bonds that act as a "glue" between the single strands to form a double helix. Complementary base pairs are Adenine - Thymine, (A-T) and Guanine - Cytosine (G-C). A schematic representation of a DNA double strand is given in Fig. 7.1

The sugar ring is made of four carbon atoms and one oxygen atom. The carbons are usually labelled as C1'-C2'-C3'-C4'. A fifth carbon bound to C4, and called C5'. (The apostrophes are not a standard chemical notation, but are used in the CHARMM notation to distinguish sugar atoms from the others). All sugar groups along a single strand are always aligned along a specific direction, identified by the position of C3' and C5' carbon atoms; the direction can be either 3'-5' or 5'-3'. The two strands of a double helix are always antiparallel, i.e. one is always a 3'-5' and the other a 5'-3' type [116].

The most spontaneous starting point for the coarse-graining of a polymer is, of course, the modelling of its monomers. In the case of DNA, usually the most widely adopted CG schemes have 1,2, or 3 points per nucleotide [117].
For our CG model, we have chosen a 3 point mode, where phosphate, deoxyribose and nitrogenous base are modelled by one site each, as illustrated in Fig. 7.1.

A further consideration is needed when coarse-graining an atomistic model such as



(a) (b)

Figure 7.2: Example of three-point-per-residue model of a nucleotide (here, Adenosine): the atomistic structures of phospate, deoxyribose and nitrogenous base (left) are mapped onto one bead each (right).

those produced for the CHARMM atomistic force field. It is customary, in such cases, to "patch" (modify) the terminal monomers of the chain, namely "capping" the (3') and (5') ends (that would normally be bonded to the neighbouring residues) with hydroxyl groups (O-H) after removing the phosphate at the 5' end of the chain. In the present work, for the parametrisation of the CG interaction sites, no distinction was made between the terminal and non-terminal residues along the DNA chains. Water molecules were modelled with 1 site per molecule, located in the centre of mass. The all-atom DNA chains used in this work were created using the MakeNA server, based on the NAB language for molecular manipulation [118].

7.1.1 Calculation of atomic groups' centres of mass

In the computation of the group centre of mass, every atom coordinate is weighted by the atom mass. Luckily, like most biomolecules, nucleic acids are also made of few elements (in this case carbon, nitrogen, oxygen, hydrogen, and phosphorus), which simplifies the computation. All four nucleotides share two common groups, phosphate and deoxyribose, and are characterised by the third, the nitrogenous base. The atomic masses and the resulting masses of the CG sites obtained using a 3-point-per-residue mapping are listed in Tab 7.1 and Tab. 7.2. Sodium ions were modelled explicitly, and clearly the only option was a 1-point-per-atom map.

Element	Symbol	Mass (g/mol)
Carbon	C	12.0107
Hydrogen	H	1.0079
Nitrogen	N	14.0067
Oxygen	O	15.9994
Phosphorus	P	30.9737
Sodium	Na	22.9897
Chlorine	Cl	35.4532

Table 7.1: List of elements that form nucleic acids, water and the most common ions used in MD simulations. Masses are expressed in g/mol.

CG site	Symbol	Mass (g/mol)
Adenine	A	134.122
Thymine	T	125.108
Guanine	G	150.121
Cytosine	C	112.101
Deoxyribose	SUG	116.117
Phosphate	PHO	62.973
Water	WT3	18.05
Sodium	SOD	22.990

Table 7.2: List of coarse-grained sites for the CG model, and their masses in g/mol.

7.1.2 Topology of the CG model

The term “topology”, when applied to particle models such as those used in molecular dynamics, refers to the way particles are interconnected by covalent bonds. The mechanical properties of the bonds are modelled as opportune “springs” (e.g. harmonic potentials) between pairs of bonded atoms. Likewise, angular and torsional rigidity are described by angular springs, involving triplets of atoms (necessary to univocally identify a bending angle), and torsional springs, involving quadruplets of atoms (to identify two planes and the corresponding torsion angle).

If we see the molecule as a “graph”, it is easy to notice that whilst the identification of bonds and angles is quite straightforward, the identification of dihedrals becomes more and more demanding as the topological complexity increases, because of the high number of possible combinations of quadruplets of consecutively bonded atoms. This task is

usually carried forward with the help of dedicated software libraries, such as the Psfgen package available in VMD [119].
An example: a single adenine nucleotide on a ssDNA chain is represented in an all-atom simulation as a set of 30 atoms, whose bonded interactions are modelled by 33 bonds, 57 angles and 92 dihedrals. In the case of a 3-point coarse grained models,reducing the number of particles leads to a great topological simplification. The topology recognition is more conveniently performed in three steps: first, a bond list is generated from the known residue list; secondly, the bond list is used to generate an angle list (every angle can be seen as formed by two distinct bonds sharing one atom); thirdly, the angle list is used to generate the list of dihedrals (every dihedral can be seen as formed by two distinct angles sharing two atoms).

7.2 Implementation

In practical application, the coarse-graining procedure takes as input the atomistic data, in terms of sets of trajectory snapshots, each containing the coordinates of the atomistic system, the forces acting on each atom, and some other thermodynamic parameters, such as temperature. The output is a reduced system of CG sites, and their mutual interaction potentials. The MD software used in this work (LAMMPS, by Sandia National Laboratories, USA) is open-source, constantly updated and rather mature, having been developed for over a decade. It is designed to be easy to modify, but it's already extremely flexible "as it is". For the purposes of this work, almost all needed simulation capabilities were already available. However, LAMMPS only comes as a highly efficient parallel simulation engine, with no building, pre- or post-processing capability whatsoever. A considerable effort was hence devoted to the development of the required pre- and post-processing tools.

7.2.1 Estimation of datafile size

Our simulation scenario comprised long atomistic trajectories, in the order of tens of nanoseconds, performed with timesteps of 1-2 femtoseconds, with snapshots taken with a sampling stride in the order of 1 picosecond. The resulting datafiles contained in the order of approx. 10^5 configurations, for a total file size up to few tens of Gigabytes, which can be cumbersome to handle but does not cause any serious feasibility problem.

7.2.2 Format conversion

LAMMPS' output routines are very flexible, and allow the dumping of prescribed subsets of system particles and their properties. This is extremely useful to define a taylor-made output that only stores what is strictly necessary. This flexibility also implies that a dump file can only be parsed in a context-dependent way, and post-processing routines must take into account the chosen format for each simulation. The output can be in binary or text format, the former being much faster, but also prone to compatibility problems because of the possible architecture differences between the CPUs of HPC supercomputers, used for the simulation, and desktop PCs,used for the post-processing. Because of such problems we have used text output only for the dumping of simulation snapshots. Appropriate tools have been implemented in Python and C for the conversion from LAMMPS data files to the commonly used PDB and PSF file formats for molecular coordinates and topology, which can be read by VMD.

7.2.3 Physical units

For biomolecular simulations, LAMMPS operates with a peculiar set of physical units. We have used such units for all the quantities computed by our code, and presented in the remaining chapters of this thesis. The units are summarised in Tab. 7.2.3.

Quantity	Units
Distance	Angstroms
Time	femtoseconds
Mass	grams/mole
Energy	Kcal/mole
Velocity	Angstroms/femtoseconds
Force	Kcal/mole-Angstroms
Temperature	Kelvin
Pressure	atmospheres
Charge	proton charge e

TABLE 7.3: Summary of physical units used by LAMMPS and also adopted for the simulations performed in this thesis.

8

Coarse-Graining of Water

In this chapter we describe the modelling of the coarse-grained solvent, generated from the initial all-atom simulation of the Price-Brooks TIP3P model that was described and studied in the previous parts of this thesis.

8.1 Coarse-Grained mapping of water onto a one-point model

The triatomic, rigid TIP3P water molecule was mapped onto a single CG site situated in its centre of mass. The electrostatic charges were not included explicitly in the CG model, and their effect was therefore implicitly incorporated in the overall pairwise potential to be determined. Other authors have pursued a different route and located the geometric center of the molecule, in order to account for the unbalanced mass distribution within a water molecule's atoms (an oxygen atom is about 16 times heavier than a hydrogen). As it will be shown in the next sections, however, our approach is nevertheless effective in reproducing the fluid structure.

8.2 Method and simulated cases

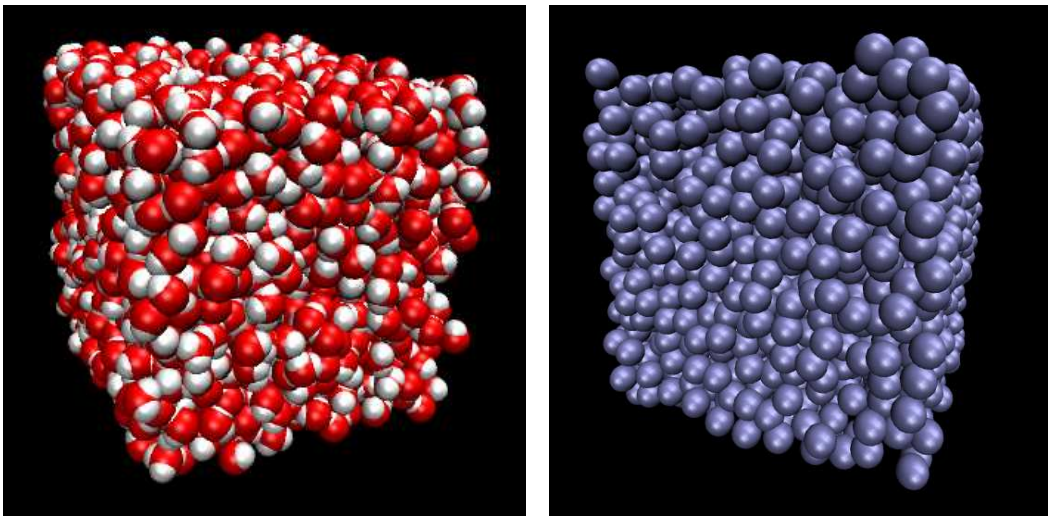
The all-atom simulations were performed using the LAMMPS MD code by Sandia National Laboratories (USA) in combination with the CHARMM force field. Pressure and temperature were controlled using a Nose-Hoover thermostat and barostat. Long-range electrostatic interactions were calculated with a PPPM algorithm, whilst the cutoff radius for Lennard-Jones interactions was set at 10.0Å.

8.2.1 All-atom simulation

The all-atom simulations were performed using the LAMMPS MD code by Sandia National Laboratories (USA) in combination with the CHARMM force field. Pressure and temperature were controlled using a Nose-Hoover thermostat and barostat. Long-range electrostatic interactions were calculated with a PPPM algorithm, whilst the cutoff radius for Lennard-Jones interactions was set at 10.0Å. A cubic water box of 40Å, containing 5943 atoms (1981 molecules), was equilibrated for 1ns in the NpT ensemble at 298K and 1 atm, with a timestep of 0.5 fs,and then simulated for 4ns in the NVT ensemble at 298K, with a timestep of 2 fs. Configurations and atomistic forces were saved every 2ps, for a total of 2000 configurations.

8.2.2 Coarse-graining procedure

Mapping molecules onto 1-point models implies that the resulting CG system contains no bonded interactions, and its physical behaviour depends uniquely on the non-bonded pairwise interactions. We have used our implementation of the MSCG method in order to parametrise the resulting interaction potential for CG water molecules. The chosen cutoff radius was 10.0Å, and the interactions were discretised on a grid with step 0.01Å, for a total of 1000 unknowns. Notice that in this case, having only one possible type of pair interaction (water-water) greatly reduces the number of unknowns. The MSCG problem was solved using a block-averaging approach as described by [18]. The 2000 configurations were partitioned into 200 blocks of 10 configurations each, and the size of the matrix associated with each block's linear system was 178290×1000. The overdetermined system associated with each block was solved in the least-squares sense using the solver provided by the Linalg package in the Python numerical module Numpy. The resulting solution vectors were then averaged over all blocks.



(a) (b)

FIGURE 8.1: A snapshot of the all-atom (left) water box used for the atomistic simulations, and the resulting 1-point coarse grained water box (right).

8.2.3 Validation of the coarse-grained potential

The output of the MSCG method is a tabulated distance-dependent force that models the CG sites interactions by trying to reproduce the overall force arising from the underlying all-atom interactions. The computed tabulated force was then fed back into a LAMMPS simulation of a CG water box of $40 \times 40 \times 40$Å, which was then simulated at 298K in the NVT ensemble with a timestep of 2fs in order to allow an easier comparison with the atomistic simulation. LAMMPS has the capability to take as input a force table and then perform cubic spline interpolation in order to calculate forces at distances comprised between two grid points. The original tabulated potential and the resulting interpolation are shown in Fig. 8.2.3.

The shape of the force-distance relation shows one apparently curious feature. The repulsive force becomes very strong approaching an intersite distance of 2Å, but then drops considerably for even shorter values of r. This is clearly unphysical, as we would expect the finite size of the molecules to cause even greater repulsion once the electronic shells of the atoms are close enough to collide. The reason can be readily seen in Fig. 8.2.3, where the empirical density function of r has been computed over all molecules and sampled configurations. Values of r which are smaller that 2Å, i.e. the position of the "collision peak" in the force table, almost never occur. Therefore the corresponding regions of

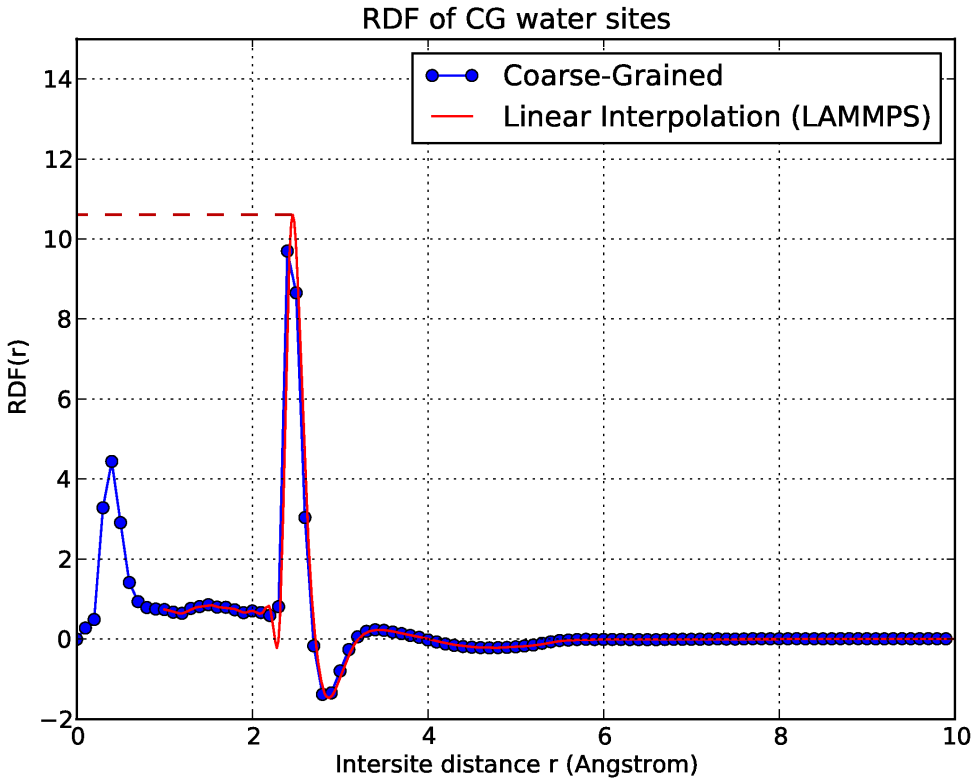


FIGURE 8.2: Tabulated force as computed using the MSCG method, and as interpolated by LAMMPS; the constant extrapolation to the core region is indicated with a dashed line.

force-field parametric space are very poorly guessed. On one hand, this is an unavoidable limitation of the force-matching approach: the centres of mass of the coarse-grained chemical groups will never spontaneously sample those regions of phase space where their positions are strongly overlapped, because of the huge repulsive forces and the steric hindrance caused by the finite size of the molecules in the underlying all-atom model. On the other hand, however, the problem can be easily solved by "manually" patching the computed force table with highly repulsive short-distance forces. This is for example the course of action suggested by Noid and coworkers [18]. The parameter chosen for the validation of the model is the radial distribution function (RDF) of water's oxygen atoms. The radial distribution function has been chosen because it's been a long-standing tradition to use it as the "structural fingerprint" for fluid systems [2, 120, 121]. As discussed previously, in Chapter 4,the RDF of particular atoms is a distinctive property of a fluid.

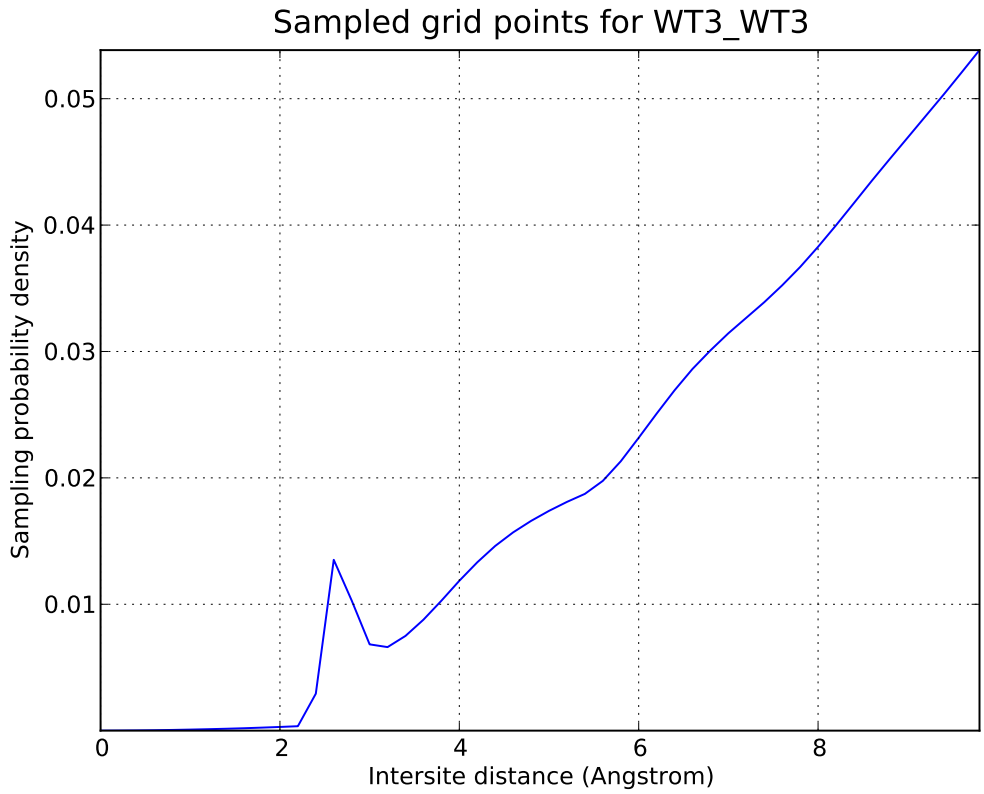


Figure 8.3: probability density function of a certain intersite distance to occur between two coarse-grained water molecules.

Moreover, experimental data are available for comparison, although it is also well-known that TIP3P models fail to perfectly reproduce water structure, characterised by a RDF with three characteristic peaks of decreasing height [97]. Being simpler than the atomistic representation, a CG model cannot hope to improve over the limitations of the underlying all-atom model. However, as shown in Fig 8.2.3, the resulting RDF approximates very well the structure of Price-Brooks TIP3P, which is in turn a reasonable approximation of real water.

8.3 Results and discussion

The overall agreement between the structure of the all-atom and CG water models is good, with a good positioning of the first peak and a loss of the second and third peak that can

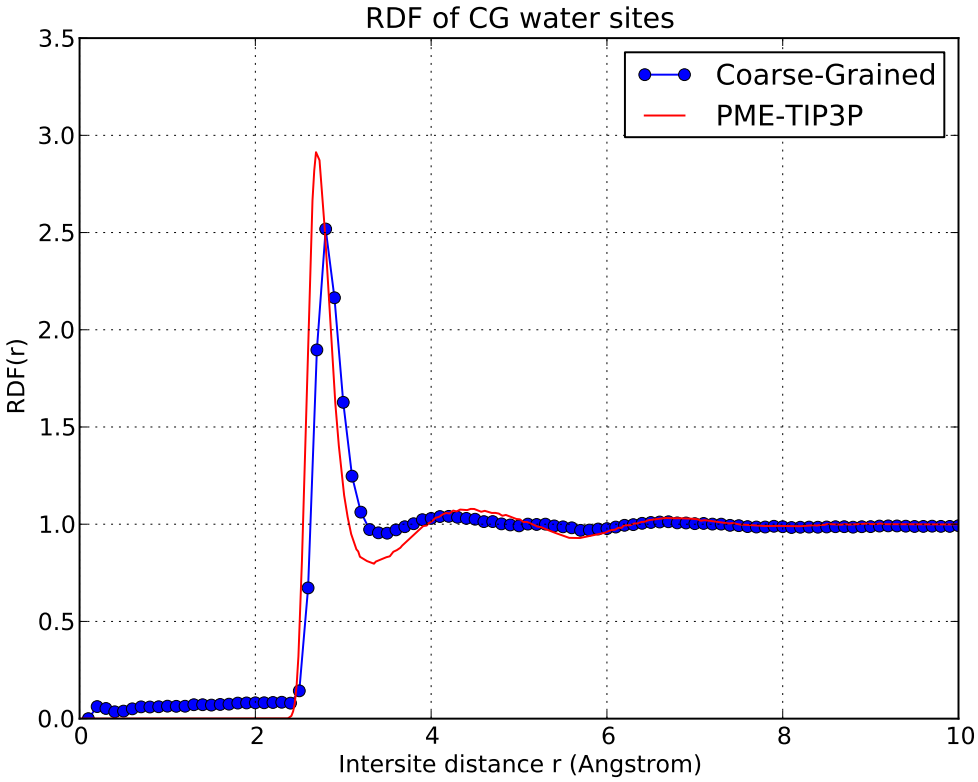


Figure 8.4: radial distribution function of all-atom and coarse-grained water models

be considering acceptable, considering that even some all-atom models cannot capture these features [95]. The resulting water model has been used for the simulation of coarse-grained models of DNA. The characterisation of the interactions of water sites with other CG sites of the DNA model will be discussed in the next chapter.

9

Coarse-Graining of DNA

In this chapter we describe the coarse-graining strategy adopted for DNA molecules. In the case of nucleic acids, the structural complexity is clearly much greater than for simple water molecules. Several have been exploited in order to reduce the computational cost required for the parametrisation of the coarse-grained force field. In particular, the expensive MSCG approach has only been used for pairwise interactions; Bonded interactions (bond stretching, angle bending and dihedral torsional forces) have been parametrised separately according to the method previously described in Chapter 3. With this (reasonable) simplifying hypothesis, the number of unknowns in the least-squares problem (Eq. 3.64 was reduced to half of the original value.

9.1 Method and simulated cases

All-atom simulations for the present chapter have been performed using the LAMMPS molecular simulator [25] and the CHARMM force-field [36]. For all simulations, the cutoff for the pairwise interactions was set at 10Å, and the long-range electrostatic forces were solved using a Ewald summation, and bonds containing hydrogens were constrained

with SHAKE [41]. The resulting trajectories were postprocessed using custom C and Octave scripts, and the VMD visualisation software [119]. The Force-Matching technique was applied using our in-house implementation of the MSCG method, written in Python, making extensive use of the Numpy and SciPy numerical libraries [17, 18]. Numerically demanding subroutines, such as the loops for the calculation of pairwise interactions, were implemented in C and linked to the Python code using the Weave subpackage contained in SciPy.

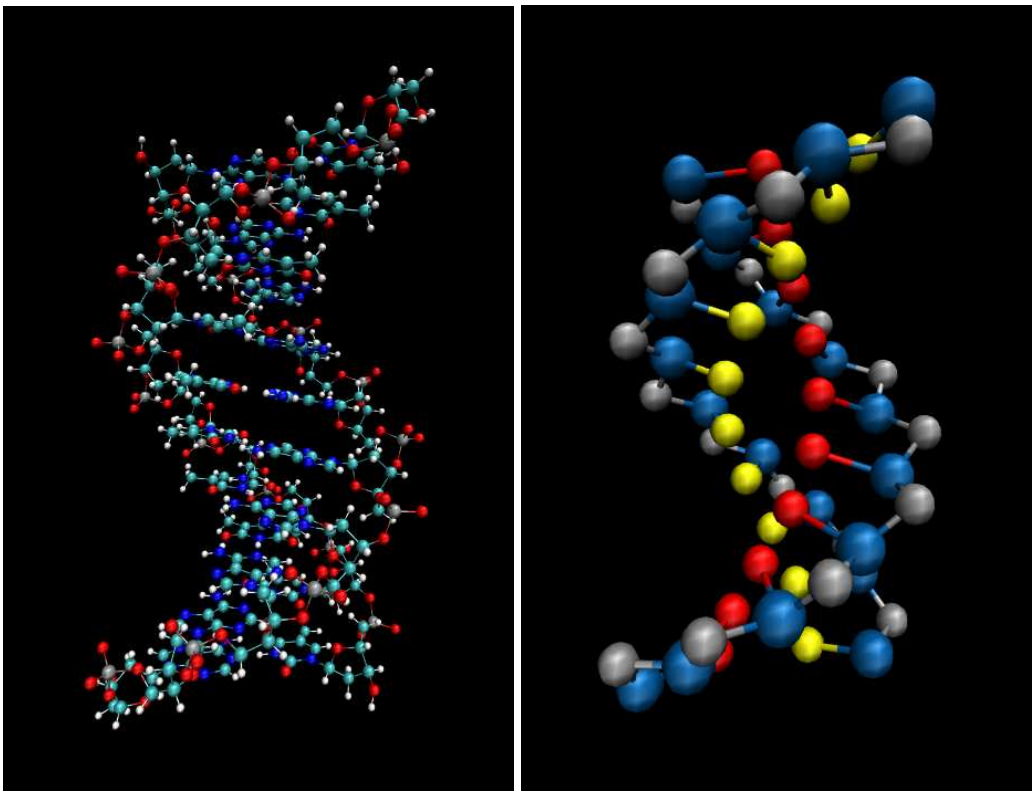
All-Atom simulation of dsDNA

For this study, a decamer of double-stranded B-DNA made of Adenine-Thymine base pairs was minimised and equilibrated in a rectangular box of 40×40×60Å, containing 3487 PME-TIP3P water molecules [93] and 18 Na^+ ions. The inclusion of sodium ions is absolutely necessary in order to neutralise the high negative charge of the DNA molecule. The downside is the introduction of one additional site type in the system, which causes the number of possible heteroatomic pairs to raise from 15 to 21. The number of unknowns in the linear system described by Eq.3.64 grows accordingly. The system was then simulated for 10ns at a constant temperature and pressure of 298K using a Nose-Hoover thermostat, with a timestep of 1fs. Atomistic forces and positions were stored every 2ps, for a total of 5000 configurations.

9.1.1 Coarse-Grained mapping of DNA onto a 3-point-per-residue model

For the coarse grained solvent, we decided to use a 1-point model with a single interaction site for the whole molecule, located in its center of mass, as discussed in the previous chapter. It must be noted that the results of the previous chapter are not sufficient for the parametrisation of the solvent: all pairwise interactions between water sites and coarse-grained DNA sites had also to be determined. However, the pre-parametrisation of water-water interactions allowed a great simplification in the numerical solution of the least-squares problem, because the contributions of all those water molecules that were only interacting with other water molecules (which were the vast majority) could be excluded. The possibility to use the geometric center instead was discussed for example by Noid [18]. The pairwise interaction potential was determined by force matching. Coulombic interactions were not included explicitly in the CG potential. For the DNA,

the choice was a 3-point per residue model, where the interaction sites model respectively the phosphate group, the ribose rings and the nitrogenous bases. Other authors have pursued different routes, for example 1-point-per-residue models. However in such cases the drastic geometrical simplification requires ad-hoc adjustments in order to mimic the base-pairing interaction, which is highly directional in nature, and cannot be captured by radial interactions in a 1-point model (and actually, it proves a challenge even for 3-point models) [78].



(a) All-atom DNA model (b) 3-point-per-residue CG model

FIGURE 9.1: Comparison between the All-Atom and Coarse-Grained DNA models

9.1.2 Parametrisation of bonded interactions

All bonded interactions were parametrised in advance from their equilibrium fluctuations, assuming a harmonic functional form for both bond-stretching and angle-bending potentials, as explained in Chapter 3. As a first approximation, the contribution of backbone dihedral angles was neglected, since the involved forces are usually less relevant than for

bonds and angles [13]. The resulting probability distribution and least-squares fitting for each bond and angle are shown in Fig. 9.2 and 9.3. The overall good agreement demonstrates the validity of the harmonic approximation, as already pointed out by [88]. The numerical values for the parameters are summarised in Tab 9.1 and Tab 9.2 . One observation is that in some cases the probability distribution deviates from the expected Gaussian behaviour and shows two peaks instead of one. This can be explained with the fact that the underlying atomistic structure is fluctuating around *two* possible metastable configurations, which correspond to two (slightly) different positions for the centres of mass of the involved atomic groups.

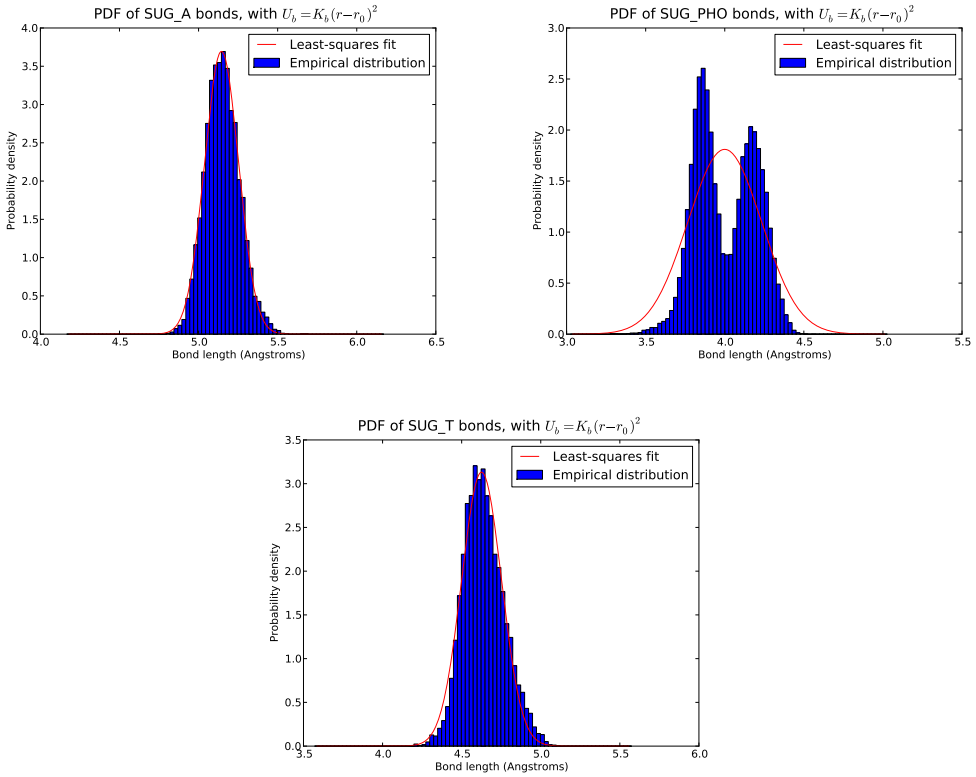


FIGURE 9.2: Empirical probability distribution for the bonds of a coarse-grained 3-point-per-residue model of DNA.

9.1.3 Solution of the least-squares problem

The cutoff radius for the pairwise interactions was set to 10.0Å, with a grid step of 0.2Å. The 6 different types of sites generate 21 possible pairwise interactions and the resulting

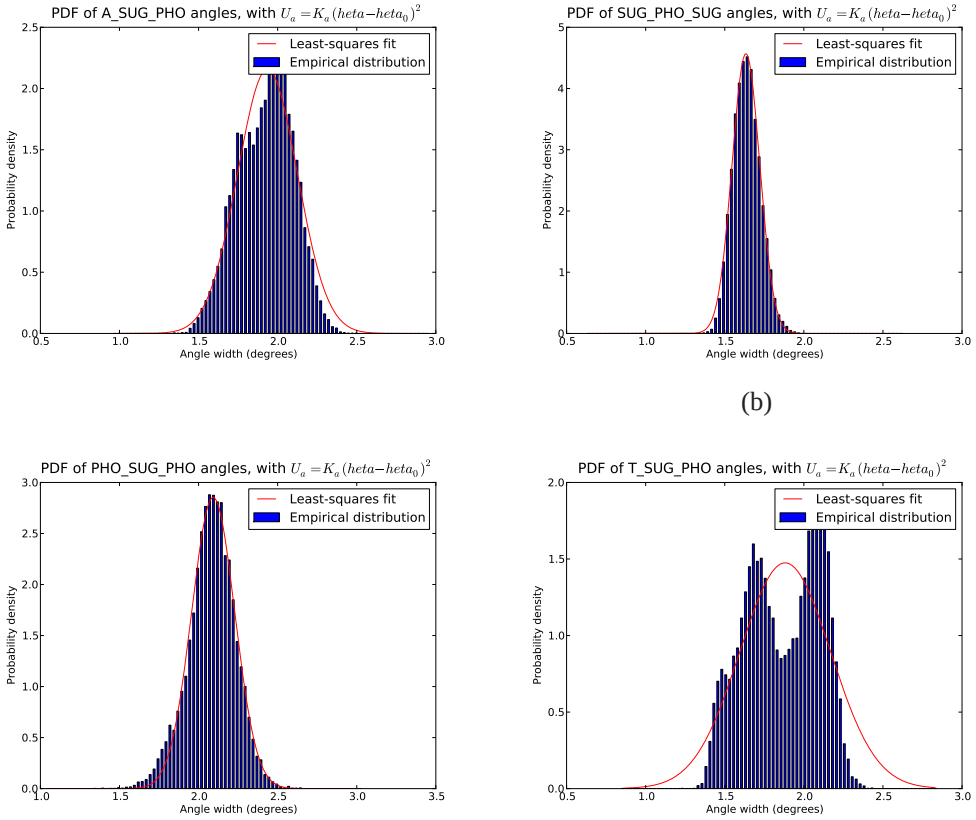


FIGURE 9.3: Empirical probability distribution for the angles of a coarse-grained 3-point-per-residue model of DNA.

system of MSCG equations had 1050 unknowns. Each site generates 3 independent equations, one for each Cartesian direction. However, it is useful to note that nearly 99% of the system's sites are CG waters, and most of them are only interacting with other CG waters. The situation is depicted more clearly in Fig. 9.4. Since most water molecules only interact with other water molecules, the corresponding matrix rows would contain non-zero coefficients only for that (small) portion of the ϕ vector that contains the parameters of the water-water intersite forces, which are the parameters we *already* known from the separate parametrisation of the solvent performed beforehand in the previous chapter. The adopted approach removes "redundant" equations, but greatly reduces the number of equations that each saved configuration can contribute. The minimum number of configurations per block must be increased accordingly, in order to ensure overdetermination of the linear system of Eq. 3.64.

ID	Bond type	K_b	r_0
1	SUG-A	25.539264	5.144872
2	SUG-PHO	5.426890	3.999550
3	SUG-T	18.627567	4.625684

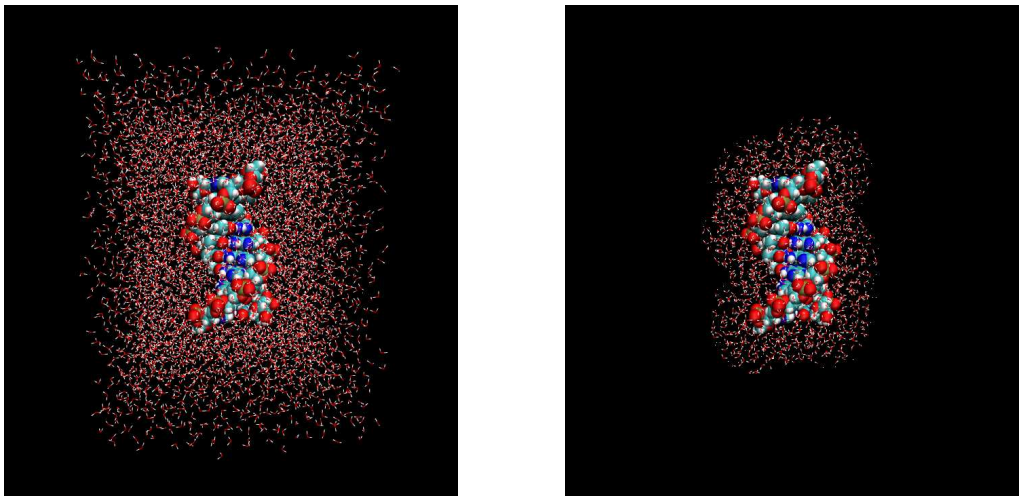
TABLE 9.1: Bond parameters for the CG DNA model; K_b is in $Kcal/(mol \cdot Angstrom^2)$, r_0 in Angstroms.

ID	Angle type	K_a	θ_0
1	A-SUG-PHO	8.296963	1.935816
2	SUG-PHO-SUG	38.551760	1.633703
3	PHO-SUG-PHO	15.942489	2.092407
4	T-SUG-PHO	3.603392	1.881817

TABLE 9.2: Angle parameters for the CG DNA model; K_a is in $Kcal/(mol \cdot rad^2)$, θ_0 in radiants.

Once we excluded water sites, the resulting system had only 76 sites (58 for the DNA and 18 for sodium ions), therefore each snapshot generated 228 equations. The 5000 configuration were partitioned into 125 blocks of 40 configurations each. Each block of 40 timesteps contained 9120 equations in 1050 unknowns, enough to make the linear system overdetermined. The 5000 configurations were partitioned into 125 blocks of 40 configurations each. After the exclusion of the matrix rows corresponding to water molecules, each block contained 10120 equations in 1050 unknowns. It must be noted that we only excluded the *equations* (matrix rows) generated by water molecules, but their contributions to the forces on all remaining CG sites were computed as usual. Therefore this apparent simplification doesn't cause any loss of physical consistency.

The forces associated with CG bonded interactions (already parametrised as explained in the previous section) were precomputed for each timestep and subtracted from the vector of known terms. Each of the 125 blocks was solved independently and the results were averaged over all blocks. The resulting potential of mean force are summarised in



(a) DNA decamer surrounded by water

(b) Water molecules actually interacting with the DNA

Figure 9.4: Portion of water box which is within 1 cutoff radius from the DNA decamer

Figg. 9.5, 9.6, 9.7, 9.8, 9.9. One final remark is that all discretised pairwise forces were represented as arrays of opportune delta functions. The resulting profile of the force table is a stepwise continue function. Other authors produced tabulated potentials using spline interpolations, but we discarded this option for practical reasons related to the MD code we used for all simulations: LAMMPS only supports tables of discretised values and automatically produces its own spline interpolation.

9.2 Simulation of the CG system

The CG system, mapped and parametrised as described in the previous sections, was simulated for 1ns at 298K in the NVT ensemble, with a timestep of 1fs. The temperature was controlled using a Nose-Hoover thermostat. The obtained configurations were used to calculate the molecule's Root-Mean-Square-Displacement (RMSD) over time. Usually the RMSD is commonly used as a measure of the level of equilibration reached by a molecule [23]. A stabilised value for the RMSD is related to the molecule's capability to fluctuate around an equilibrium configuration. We have decided to use the RMSD value over time in order to compare the equilibrium behaviours of the CG and AA systems over a time window of 1ns, and the results are shown in Fig 9.11. The helical geometry is at least qualitatively preserved as shown in Fig. 9.12.

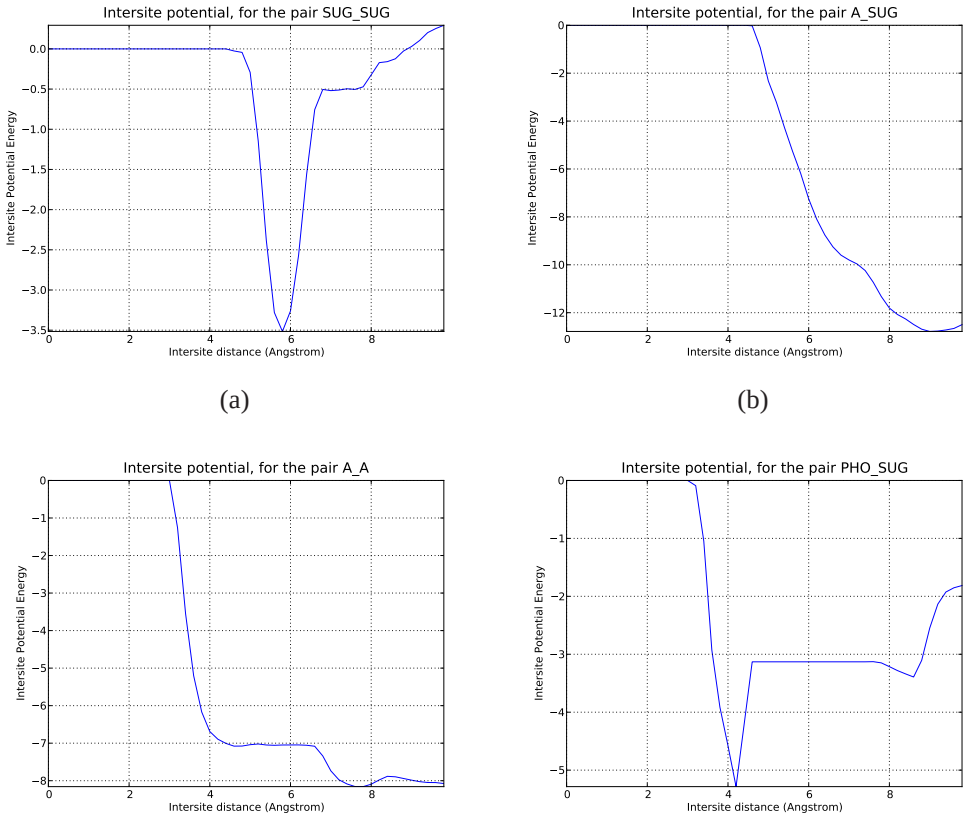


Figure 9.5: Potentials of mean force for pairwise interactions of the coarse-grained system

9.3 Discussion

The results for the CG system show that the equilibrium configuration is still not completely stable and tends to unzip after about 1ns. At the nanosecond timescale, the CG structure preserves its helical shape but it tends to be farther away from the initial crystal structure in comparison to the all-atom model (the RMSD stabilised on a higher value). This can be explained by the higher flexibility of the coarse-grained chain due to the exclusion of torsional dihedral potentials, which were excluded on the ground of the low energy associated with polymer torsional degrees of freedom [13, 122]. They should be included in future improvements of this model.

However, our approach succeeded in his main task, i.e. proving capable of handling a structurally complex system and generate a plausible force field for bonded and pairwise

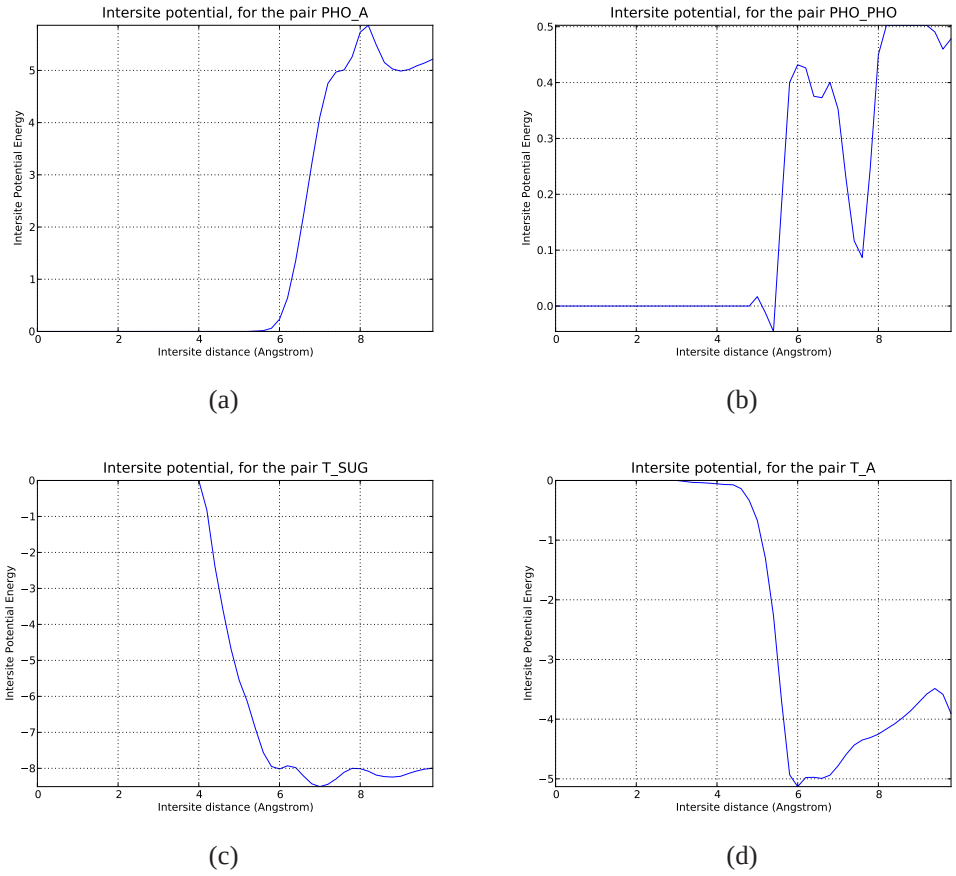


Figure 9.6: Potentials of mean force for pairwise interactions of the coarse-grained system

interactions. The comparative stability of the double helix is remarkable especially because the MSCG was capable of generating the expected attractive potential for Adenine-Thymine interactions at short range, without relying on any ad-hoc hypotheses of sort, unlike other models [78]. In particular, for systems in the order of 10000-20000 particles, which are big enough for the simulation of many small biomolecules (DNA strand and small proteins), the postprocessing of tens of thousands of configurations can be performed in a matter of hours even on a normal desktop PC.

Moreover, DNA can be considered as a particularly hard molecule to coarse-grain, and the method can be much easier to apply to other types of biomolecules. Nitrogenous bases (in our model, Adenine and Thymine) are planar chemical groups, and therefore the loss of geometrical detail, when replacing them with spherical beads, is considerable. Furthermore, the two strands of the double helix are hold together by highly directional

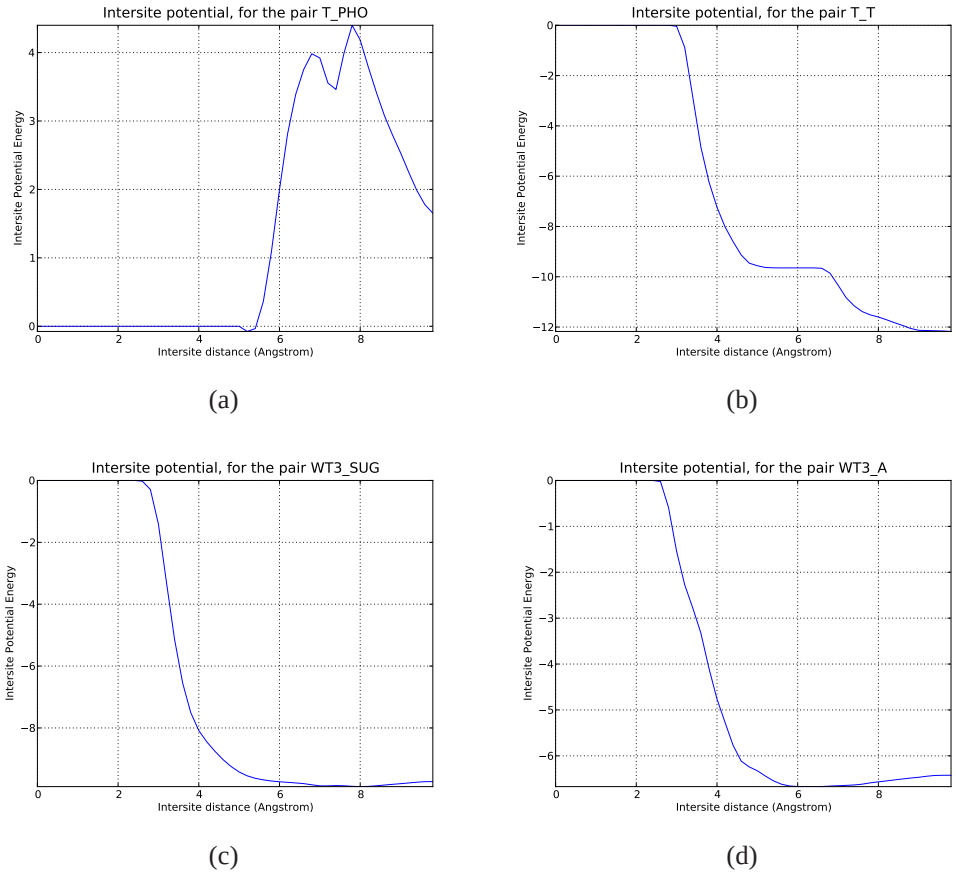


Figure 9.7: Potentials of mean force for pairwise interactions of the coarse-grained system

hydrogen bonds between complementary bases, which are replaced in the CG model by much simpler radial interactions. Therefore, base-pairing interactions are not easy to reproduce accurately in a CG system.

Taking into account the above mentioned difficulties, the CG DNA model performs promisingly, especially because it's possible to see room for improvement in several respects. The explicit modelling of the torsional degrees of freedom of the DNA backbone will help keep the alignment of the complementary bases and stabilise the base pairing and therefore the double helix.

Further refinement of the interactions can be achieved by processing more atomistic configurations. The upper limit is set by storage space and processing speed: with our setup, for the system investigated in this chapter, 1000 configurations take approx. 1GB of storage space and can be processed in about 1h.

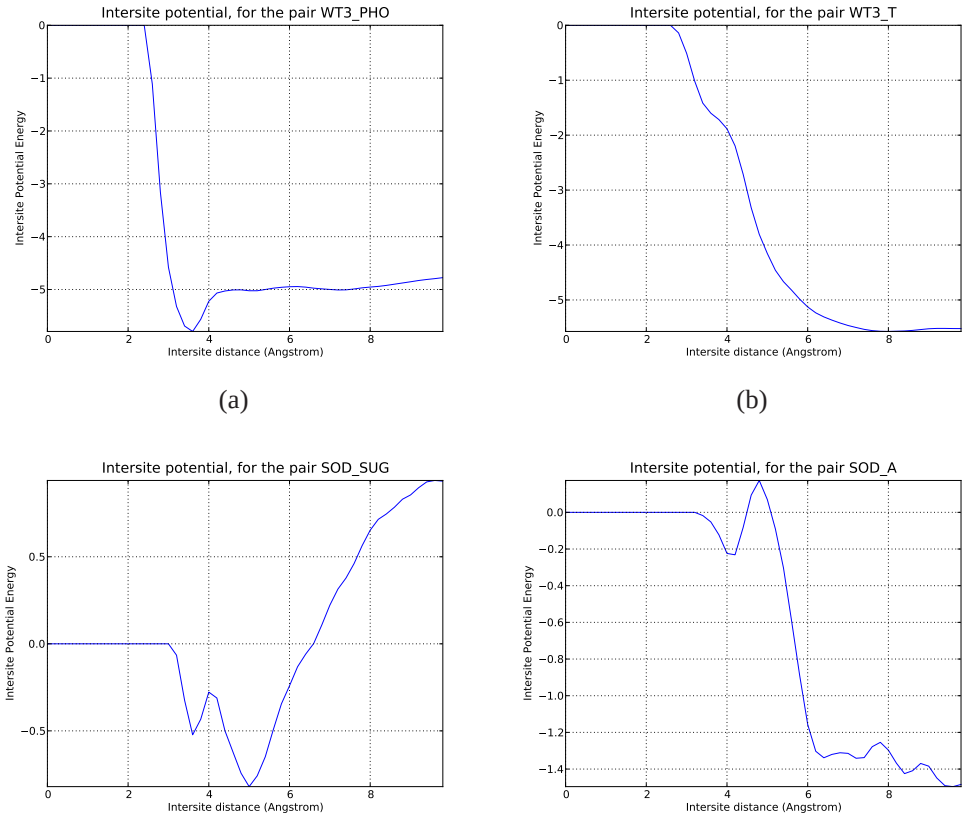


Figure 9.8: Potentials of mean force for pairwise interactions of the coarse-grained system

In addition, other authors have recently proposed a method based of Bayesian inference for the refinement of the force-field produced by the MSCG method, which can improve the physical consistency of the CG model [91].

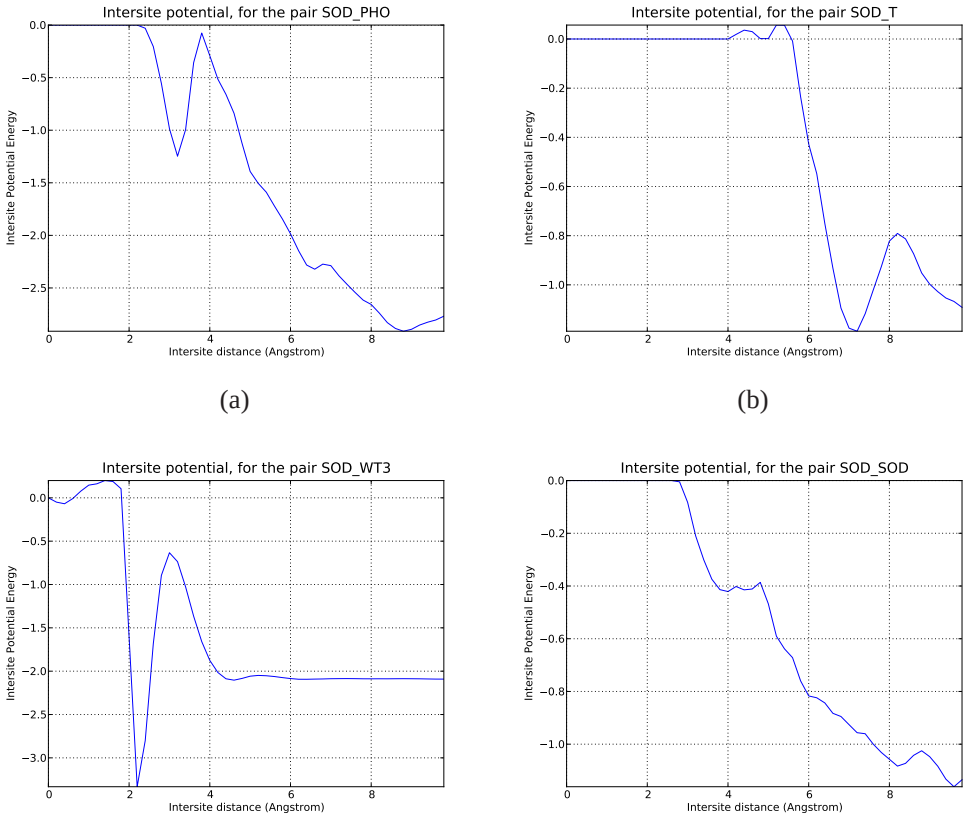
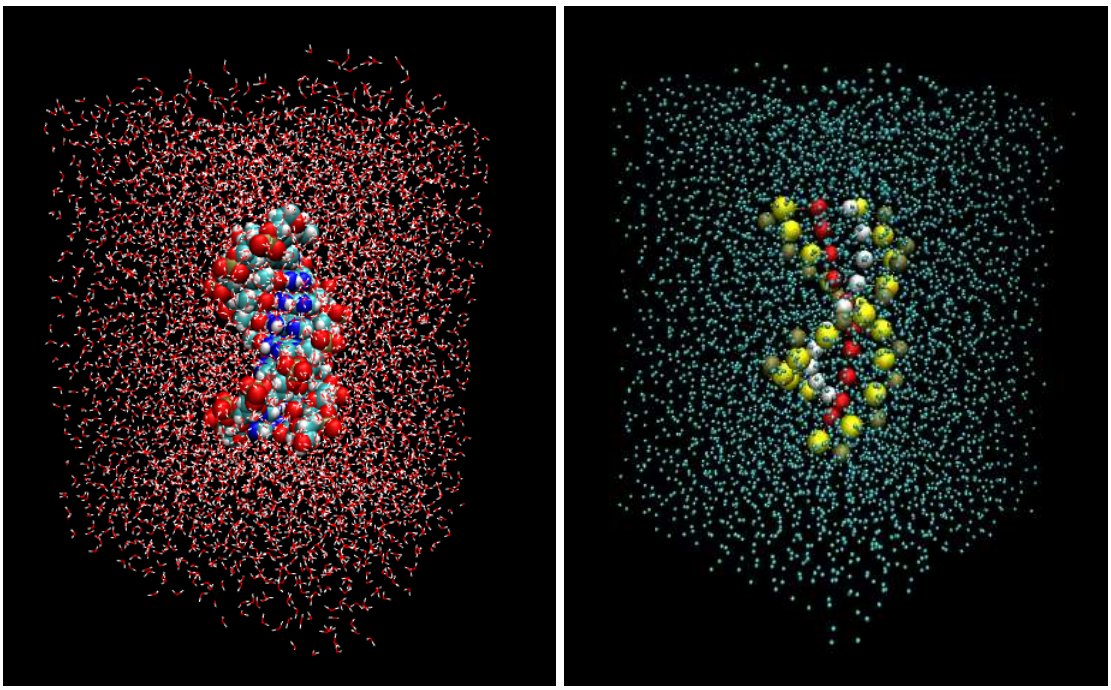


Figure 9.9: Potentials of mean force for pairwise interactions of the coarse-grained system



(a) (b)

Figure 9.10: Comparison between the all-atom and the CG simulation boxes. The number of particles was reduced from over 15000 to about 5000.

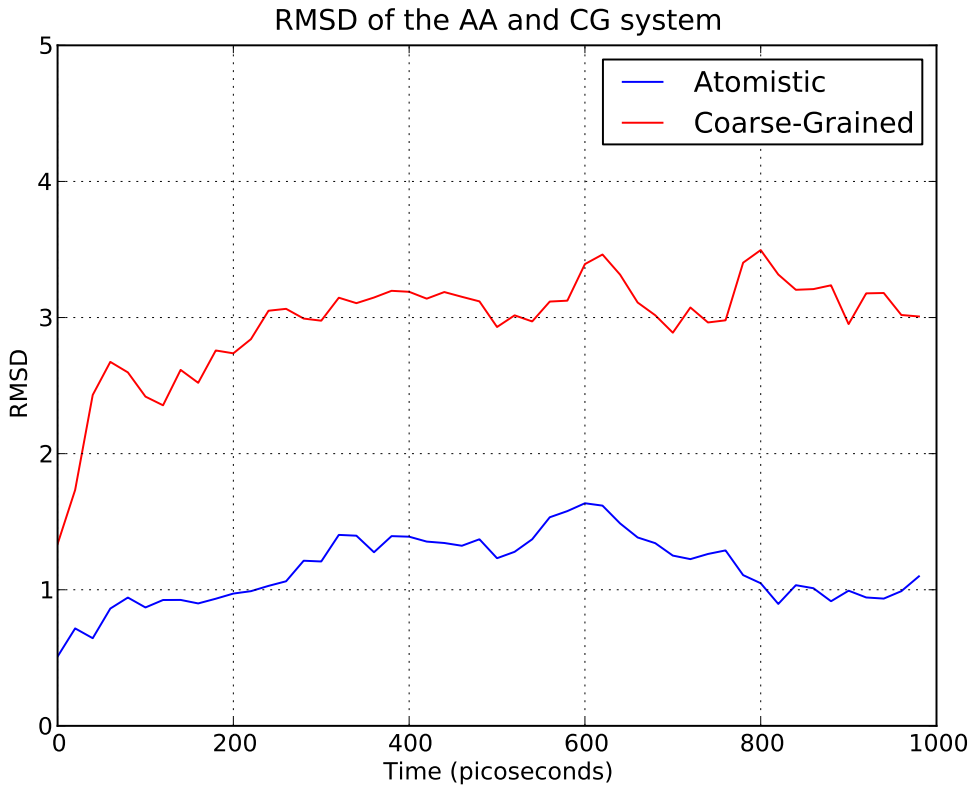
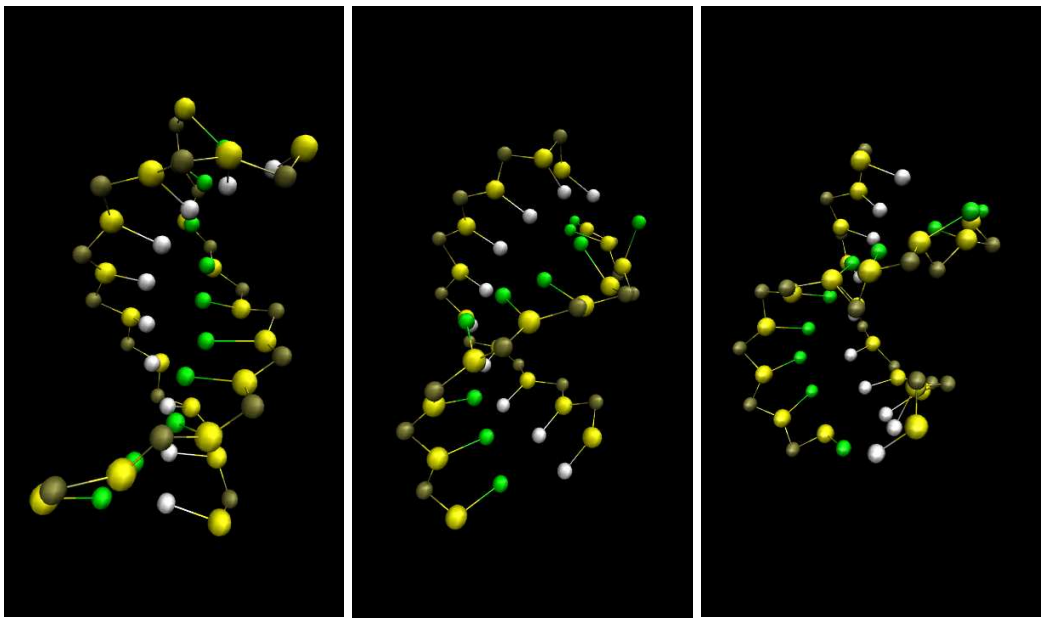


Figure 9.11: Comparison between the RMSD for the atomistic and CG system, over a time window of 1000ps.



(a) Initial configuration (b) (c)

Figure 9.12: Equilibrium fluctuation of the CG DNA model. The helical geometry is at least qualitatively reproduced.

Conclusions

10

Concluding remarks and suggestions for future work

The work described in this Thesis can be divided in two parts.

First, we have used Molecular Dynamics in order to gain a quantitative understanding of diffusive processes at the nanoscale, focusing on DNA in aqueous solution, using the widely adopted CHARMM potential in combination with the TIP3P-PME water model [36, 93].

In particular, the properties of model water in MD simulations in different simulation conditions were studied. Furthermore, the results obtained from the study of the water model were used for the simulation of diffusion coefficients of single-stranded DNA molecules.

As demonstrated by the good quantitative agreement between the computed diffusion values and the empirical values predicted by available models, the techniques implemented and tested in this thesis establish a reliable computational methodology for the study of fluid viscosities and the diffusive properties of small biomolecules. Such methods would be readily applied to other biomolecular systems of small diffusing organic

molecules, such as drugs.

The second part of this thesis focussed instead on the implementation and testing of a recently developed method for the parametrisation of coarse-grained force fields from atomistic simulations, called the Multiscale Coarse-Graining Method, or MSCG [17]. The idea behind the method is to parametrise coarse-grained interactions by matching the total force on the coarse grained sites, as predicted by the underlying fine grained all-atom model.

The method was implemented in an object-oriented software toolkit written in Python. Where possible, all subroutines dealing with numerically intensive tasks where re-written in C, in order to optimise performance. The toolkit was designed to be used with the LAMMPS molecular dynamics code by Sandia National Laboratories (USA) [25], but it can be expanded very easily in order to be interfaced with other softwares.

We tested and validated the method by performing a one-point-per-molecule coarse graining of TIP3P-PME water. The resulting model was able to reproduce the fluid structure (its radial distribution function) in a nearly quantitative way.

Finally, we applied the MSCG method to a more demanding problem, namely the parametrisation of a 3-point-per-residue coarse-grained model of double-stranded DNA. DNA is probably one of the most difficult biomolecules to coarse-grain, for several reasons. Its nitrogenous bases are chemical groups with a planar structure, which is not well approximated by spherical CG sites. Furthermore, the DNA double helix is made of two complementary strands which are not covalently bonded, but interact instead through highly specific and directional hydrogen bonds between complementary base pairs (Adenine-Thymine and Cytosine-Guanine), a situation that is rather difficult to reproduce by means of spherical interaction sites and radial pairwise potentials. In synthesis, the complexity of DNA internal interactions suffers the great geometrical simplification of the coarse-graining.

As a consequence, the agreement of the obtained CG model with the atomistic structure was still not quantitative. In particular, the helical geometry was qualitatively preserved and the RMSD of the coarse-grained model was stable over the trajectory, but higher than its all-atom counterpart.

The parameter estimation is therefore promising but still improvable. In particular, the MSCG method relies on the quality of the phase-space sampling of the atomistic system's simulated trajectory, which is well-known for being difficult to achieve even on the longest simulation runs.

On the other hand, the MSCG method is very effective in systems of almost spherical molecules, where radial interactions are a good approximation and all the grid points of the discretised potential (outside the repulsive core region) can be sampled abundantly. The method described in Chapter 8 for the coarse-graining of water can be easily adapted to other fluid systems, such as CO_2 or methane, as long as all-atom models are available.

10.1 Evaluation of computational cost

The lower computational cost of CG simulations is due to two main factors: first, the number of atoms decreases; secondly, the greater mass of the resulting particles allows the use of larger timesteps without jeopardising the energy conservation.
It must also be noted that water is a rather unfavourable case, because even a 1-point model can only reduce the number of particles by a factor 3, compared to the factor 10 for the 3-point model of the DNA chain.
This is all the more relevant considering that the vast majority of atoms in a biomolecular simulation (around 90%) belong to water molecules, and in actual simulations, almost 80% of the CPU time is spent calculating pairwise interactions.
As a benchmark, we have clocked the required CPU time for 1000 iterations on a single core (AMD Turion TL-52 1900 Mhz). Unfortunately, when the original simulation were performed, we didn't systematically store information about CPU time for the all-atom parallel runs. However, LAMMPS is a software specifically designed to run on large clusters. Therefore, it's performance scales nearly linearly for a number of processors up to a few hundreds [25]. On Cranfield University's HPC supercomputer, equipped with Intel EM64T Xeon 51xx (Woodcrest) 3.0 GHz CPUs, an all-atom biomolecular simulation of a system of 15503 particles took 25893 seconds (a little more than 7 hours) to complete 10^6 iterations (roughly one nanosecond of simulated time), on 16 cores. As shown in Tab. 10.1 the computational cost of the benchmark CG simulation was much lower than its all-atom counterpart, roughly by a factor 20. Since the particles involved in the CG simulation are also considerably heavier, the timestep length can also be extended. The longest achievable timestep depends on the system being simulated, and it was found to be approximately four times longer than the all-atom simulation timestep, according to Zhou and coworkers [88]. The combination of reduced computational cost and extended timestep implies a possible 80-fold gain for the longest feasible simulation timescale. This can push the timescale limit from the current hundreds of nanoseconds to

TABLE 10.1: Comparison of required computation time for the calculation of 1000 timesteps, on 1 CPU, for the atomistic system and the corresponding CG representation of a cubic box of water of $40 \times 40 \times 40$ Angstrom. For comparison purposes, both simulations used the same timestep of 2fs.

System	# Particles	CPU time (s)
All-Atom	5943	323.965
Coarse-Grained	1981	16.348

the milliseconds, which is for example the folding timescale of some small proteins [88].

10.2 Future work

Room for improvement in the modelling of DNA can be seen in two main respects.

First, our model did not include explicitly the torsional degrees of freedom of the DNA backbone, which were initially neglected because they involve energies and forces lower than angle-bending and bond-stretching interactions [13]. Their inclusion would stiffen the backbone and prevent it from fluctuating too far from the configuration predicted by the all-atom system, increasing the alignment of complementary bases and stabilising the base-pairing interaction.

Secondly, several authors have recently proposed various refinements that may improve the parameter estimation, for example using statistical Bayesian inference techniques, and including a more accurate treatment of Coulombic interactions [89, 91].

Therefore, the methodology described in this thesis, and the developed tools, can offer a very promising platform upon which future work on multiscale modelling of biomolecules can be built.

References

References

[1] S. Izvekov and G. A. Voth. *A multiscale coarse-graining method for biomolecular systems*. Journal of Physical Chemistry B **109**(7), 2469 (2005).

[2] A. Soper and M. Phillips. *A new determination of the structure of water at 25 C*. Chemical Physics **107**(1), 47 (1986).

[3] M. Laliberte. *Model for calculating the viscosity of aqueous solutions*. Journal of Chemical and Engineering Data **52**(2), 321 (2007).

[4] A. E. Nkodo, J. M. Garnier, B. Tinland, H. Ren, C. Desruisseaux, L. C. McCormick, G. Drouin, and G. W. Slater. *Diffusion coefficient of DNA molecules during free solution electrophoresis*. Electrophoresis **22**(12), 2424 (2001).

[5] H. L. Zhang and S. J. Han. *Viscosity and density of water + sodium chloride + potassium chloride solutions at 298.15K*. Journal of Chemical Engineering Data **41**(3), 516 (1996).

[6] F. Müller-Plathe. *Coarse-graining in polymer simulation: From the atomistic to the mesoscopic scale and back*. ChemPhysChem **3**(9), 754 (2002).

[7] J.-W. Chu, S. Izvekov, and G. A. Voth. *The multiscale challenge for biomolecular systems: coarse-grained modeling*. Molecular Simulation **32**(3), 211 (2006).

[8] E. Lyman and D. Zuckerman. *Ensemble-based convergence analysis of biomolecular trajectories*. Biophysical Journal **91**(3), 164 (2006).

[9] E. Lyman and D. Zuckerman. *On the structural convergence of biomolecular simulations by determination of the effective sample size*. Journal of Physical Chemistry B **111**(44), 12876 (2007).

[10] V. Tozzini. *Coarse-grained models for proteins*. Current Opinion in Structural Biology **15**(2), 144 (2005).

[11] S. Izvekov and G. A. Voth. *Modeling real dynamics in the coarse-grained representation of condensed phase systems*. Journal of Chemical Physics **125**(15), 151101 (2006).

[12] J. D. McCoy and J. G. Curro. *Mapping of explicit atom onto united atom potentials*. Macromolecules **31**(26), 9362 (1998).

[13] D. Reith, M. Pütz, and F. Müller-Plathe. *Deriving effective mesoscale potentials from atomistic simulations*. Journal of Computational Chemistry **24**(13), 1624 (2003).

[14] A. Soper. *Empirical monte-carlo simulation of fluid structure*. Chemical Physics **202**(2), 295 (1996).

[15] F. Ercolessi and J. B. Adams. *Interatomic potentials from first-principles calculations: The force-matching method*. Europhysics Letters **26**(8), 583 (1994).

[16] S. Izvekov and G. A. Voth. *Multiscale coarse graining of liquid-state systems*. Journal of Chemical Physics **123**(13), 134105 (2005).

[17] W. G. Noid, J.-W. Chu, G. S. Ayton, V. Krishna, S. Izvekov, G. A. Voth, A. Das, and H. C. Andersen. *The multiscale coarse-graining method. I. A rigorous bridge between atomistic and coarse-grained models*. Journal of Chemical Physics **128**(24), 244114 (2008).

[18] W. Noid, P. Liu, Y. Wang, J. Chu, G. Ayton, S. Izvekov, H. Andersen, and G. Voth. *The multiscale coarse-graining method. II. Numerical implementation for coarse-grained molecular models.* Journal of Chemical Physics **128**(24) (2008).

[19] A. Arkhipov, P. L. Freddolino, K. Imada, K. Namba, and K. Schulten. *Coarse-grained molecular dynamics simulations of a rotating bacterial flagellum.* Biophysical Journal **91**(12), 4589 (2006).

[20] P. Freddolino, A. Arkhipov, S. Larson, A. McPherson, and K. Schulten. *Molecular dynamics simulations of the complete satellite tobacco mosaic virus*. Structure **14**(3), 437 (2006).

[21] F. Ercolessi. *A molecular dynamics primer* (1997). Available at http://www.fisica.uniud.it/ ercolessi/md/md/. Accessed on 14 August 2007, URL `http://www.fisica.uniud.it/~ercolessi/md/md/`.

[22] J. Gibbs. *Elementary principles of statistical mechanics* (Dover Publications, Inc., New York, 1960).

[23] M. Allen and D. Tildesley. *Computer simulation of liquids* (Oxford University Press, Oxford, 1987).

[24] D. Rapaport. *The Art of Molecular Dynamics* (Cambridge University Press, Cambridge, 1995).

[25] S. Plimpton. *Fast parallel algorithms for short-range molecular dynamics*. Journal of Computational Physics **117**, 1 (1995).

[26] J. Shewchuk. *An introduction to the conjugate gradient method without the agonizing pain* (1994). Available at www.cs.cmu.edu/ quake-papers/painless-conjugate-gradient.pdf. Accessed on 14 August 2007, URL `www.cs.cmu.edu/~quake-papers/painless-conjugate-gradient.pdf`.

[27] A. F. Voter, F. Montalenti, and T. C. Germann. *Extending the time scale in atomistic simulation of materials*. Annual Review of Materials Science **32**, 321 (2002).

[28] D. Hamelberg, J. Mongan, and J. A. McCammon. *Accelerated molecular dynamics: A promising and efficient simulation method for biomolecules*. Journal of Chemical Physics **120**(24), 11919 (2004).

[29] C. de Oliveira, D. Hamelberg, and J. A. McCammon. *On the application of accelerated molecular dynamics to liquid water simulations*. Journal of Physical Chemistry B **110**(45), 22695 (2006).

[30] J. Walecka. *Fundamentals of statistical mechanics - Manuscript and notes by Felix Bloch* (Imperial college Press, London, 1989).

[31] P. H. Hünenberger. *Calculation of the group-based pressure in molecular simulations. I. A general formulation including Ewald and particle-particle-particle-mesh electrostatics*. Journal of Chamical Physics **116**(6880), 1463057 (2002).

[32] P. Deuflhard and J. Hermans. *Computational Molecular Dynamics: Methods, Challenges, Ideas* (Springer Verlag, Berlin Heidelberg, 1999).

[33] A. White. *Intramolecular potentials of mixed systems: testing the Lorentz-Berthelot mixing rules with ab-initio calculations* (2000). DSTO Aeronautical and Maritime Research Laboratory, Melbourne (technical note).

[34] A. K. Al-Matar and D. Rockstraw. *A generating equation for mixing rules and two new mixing rules for interatomic potential energy parameters*. Journal of Computational Chemistry **5**(25), 660 (2003).

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

[36] A. D. MacKerell and N. K. Banavali. *All-atom empirical force field for nucleic acids: II. application to molecular dynamics simulations of DNA and RNA in solution*. Journal of Computational Chemistry **21**(2), 105 (2000).

[37] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz Jr., D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman. *A second generation force field for the simulation of proteins, nucleic acids, and organic molecules*. Journal of the American Chemical Society **117**(19), 5179 (1995).

[38] H. Andersen. *Rattle: a velocity version of the shake algorithm for molecular dynamics calculations*. Journal of Computational Physics **52**, 23 (1983).

[39] G. Ciccotti and W. Hoover. *Molecular Dynamics Simulation of Statistical-Mechanical Systems* (North-Holland Physics Publishing, Amsterdam, 1986).

[40] K. Feenstra, B. Hess, and J. Berendsen. *Improving efficiency of large time-scale molecular dynamics simulations of hydrogen-rich systems*. Journal of Computational Chemistry **20**(8), 786 (1999).

[41] J.-P. Ryckaert, G. Ciccotti, and H. Berendsen. *Numerical integration of the cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes*. Journal of Computational Physics **23**, 327 (1977).

[42] D. Evans and G. Morriss. *Statistical mechanics of non-equilibrium liquids* (Academic Press, London, 1990).

[43] P. Ewald. *Die Berechnung optischer und elektrostatischer Gitterpotentiale*. Annalen der Physik **369**, 253 (1921).

[44] E. Madelung. *Das Elektrische Feld in Systemen mit regelmässig angeordneten Punktladungen*. Physik Zeitung **19**, 524 (1918).

[45] L. L. T. Darden, L. Perera and L. Pedersen. *New tricks for modelers from the crystallography toolkit: the particle-mesh Ewald algorithm and its use in nucleic acid simulations*. Structure **7**(3), R55 (1999).

[46] E. L. Pollock and J. Glosli. *Comments on P3M, FMM, and the Ewald method for large periodic Coulombic systems*. Computer Physics Communications **95**(2-3), 93 (1996).

[47] H. C. Andersen. *Molecular dynamics simulations at constant pressure and/or temperature*. The Journal of Chemical Physics **72**(4), 2384 (1980).

[48] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak. *Molecular dynamics with coupling to an external bath*. The Journal of Chemical Physics **81**(8), 3684 (1984).

[49] S. Nose. *A unified formulation of the constant temperature molecular dynamics methods*. Journal of Chemical Physics **81**(1), 511 (1984).

[50] W. Hoover. *Canonical dynamics: Equilibrium phase-space distributions*. Physical Review A **31**(3), 1695 (1985).

[51] W. Hoover. *Generalization of Nose's isothermal molecular dynamics: Non-Hamiltonian dynamics for the canonical ensemble*. Physical Review A **40**(5), 2814 (1989).

[52] G. Martyna and M. Klein. *Nose-Hoover chains: The canonical ensemble via continuous dynamics*. Journal of Chemical Physics **97** (1992).

[53] W. Hoover. *Nose-Hoover Nonequilibrium Dynamics and Statistical Mechanics*. Molecular Simulation **33**, 13 (2007).

[54] S. Melchionna, G. Ciccotti, and B. Hoolian. *Hoover NPT dynamics for systems varying in shape and size*. Molecular Physics **78**(3), 533 (1993).

[55] J. Kirkwood. *Statistical mechanics of fluid mixtures*. Journal of Chemical Physics **3**, 300 (1935).

[56] J. Norberg and L. Nilsson. *Stacking free energy profiles for all 16 natural ribodinucleoside monophosphates in aqueous solution*. Journal of the American Chemical Society **117**(44), 10832 (1995).

[57] J. Norberg and L. Nilsson. *Influence of adjacent bases on the stacking-unstacking process of single-stranded oligonucleotides*. Biopolymers **39**(6), 765 (1996).

[58] J. Norberg and L. Nilsson. *Solvent influence on base stacking*. Biophysical Journal **74**(1), 394 (1998).

[59] J. Norberg and L. Nilsson. *Molecular dynamics applied to nucleic acids*. Accounts of Chemical Research **35**(6), 465 (2002).

[60] M. J. Sippl, M. Ortner, M. Jaritz, P. Lackner, and H. Flöckner. *Helmholtz free energies of atom pair interactions in proteins*. Folding and Design **1**(4), 289 (1996).

[61] M. Doi and S. Edwards. *The Theory of polymer dynamics* (Calendon Press, Oxford, 1986).

[62] R. G. Larson. *Principles for coarse-graining polymer molecules in simulations of polymer fluid mechanics*. Molecular Physics **102**(4 PART III), 341 (2004).

[63] P. T. Underhill and P. S. Doyle. *On the coarse-graining of polymers into bead-spring chains*. Journal of Non-Newtonian Fluid Mechanics **122**(1-3), 3 (2004).

[64] P. Underhill and P. Doyle. *Development of bead-spring polymer models using the constant extension ensemble*. Journal of Rheology **49**(5), 963 (2005).

[65] P. Underhill and P. Doyle. *Alternative spring force law for bead-spring chain models of the worm-like chain*. Journal of Rheology **50**(4), 513 (2006).

[66] P. T. Underhill and P. S. Doyle. *Accuracy of bead-spring chains in strong flows*. Journal of Non-Newtonian Fluid Mechanics **145**(2-3), 109 (2007).

[67] L. J. Smith, X. Daura, and W. F. van Gunsteren. *Assessing equilibration and convergence in biomolecular simulations*. PROTEINS: Structure, Function, and Genetics **48**, 487 (2002).

[68] G. Torrie and J. Valleau. *Nonphysical sampling distribution in monte carlo free-energy estimation: Umbrella sampling*. Journal of Computational Physics **23**, 187 (1977).

[69] A. Ferrenberg and R. Swendsen. *New monte carlo technique for studying phase transitions*. Physical Review Letters **61**(23), 2635 (1988).

[70] S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, and P. A. Kollman. *The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method*. Journal of Computational Chemistry **13**(8), 1011 (1991).

[71] B. Roux. *The calculation of the potential of mean force using computer simulations*. Computer Physics Communications **91**, 275 (1995).

[72] M. Souaille and B. Roux. *Extension to the weighted histogram analysis method: Combining umbrella sampling with free energy calculations*. Computer Physics Communications **135**(1), 40 (2001).

[73] M. Shirts and J. Chodera. *Statistically optimal analysis of samples from multiple equilibrium states*. Journal of Chemical Physics **129**(12), 124105 (2008).

[74] C. Jarzynski. *Nonequilibrium equality for free energy differences*. Physical Review **78**(14), 2690 (1997).

[75] G. Hummer and A. Szabo. *Free energy reconstruction from single-molecule pulling experiment*. Proceedings of the National Academy of Sciences **98**(7), 3658 (2005).

[76] S. Park and K. Schulten. *Calculating potentials of mean force from steered molecular dynamics simulations*. Journal of Chemical Physics **120**(13), 5946 (2004).

[77] D. Beveridge and F. DiCapua. *Free energy via molecular simulation: application to chemical and biomolecular systems*. Annual Review of Biophysical Chemistry **18**, 431 (1989).

[78] F. Trovato and V. Tozzini. *Supercoiling and local denaturation of plasmids with a minimalist DNA model*. Journal of Physical Chemistry B **112**(42), 13197 (2008).

[79] S. C. Harvey and M. Prabhakaran. *Umbrella sampling: Avoiding possible artifacts and statistical biases*. Journal of Physical Chemistry **91**(18), 4799 (1987).

[80] M. Mezei. *Adaptive umbrella sampling: self-consistent determination of the non-boltzmann bias*. Journal of Computational Physics **68**(1) (1987).

[81] A. Ferrenberg and R. Swendsen. *Optimized monte carlo data analysis*. Physical Review Letters **63**(12), 1195 (1989).

[82] R. Rajamani, K. Naidoo, and J. Gao. *Implementation of an adaptive umbrella sampling method for the calculation of multidimensional potential of mean force of chemical reactions in solution*. Journal of Computational Chemistry **24**(14), 1775 (2003).

[83] C. Bennett. *Efficient estimation of free energy differences from monte carlo data*. Journal of Computational Physics **22**(2), 245 (1976).

[84] A. Kong, P. McCullagh, X.-L. Meng, D. Nicolae, and Z. Tan. *A theory of statistical models for monte carlo integration*. Journal of the Royal Statistical Society B **65**(3), 585 (2003).

[85] Z. Tan. *On a likelihood approach for monte carlo integration*. Journal of the American Statistical Association **99**(468), 1027 (2004).

[86] D. Minh. *Multidimensional potential of mean force from biased experiments along a single coordinate*. Journal of Physical Chemistry B **111**, 4137,4140 (2007).

[87] D. Minh and A. Adib. *Optimized free energies from bidirectional single-molecule force spectroscopy*. Physical Review Letters **100**(18), 180602 (2008).

[88] J. Zhou, I. Thorpe, S. Izvekov, and G. Voth. *Coarse-grained peptide modeling using a systematic multiscale approach*. Biophysical Journal **92**(12), 4289 (2007).

[89] P. Liu, Q. Shi, H. Daumé, and G. A. Voth. *A bayesian statistics approach to multi-scale coarse graining*. Journal of Chemical Physics **129**(21), 214114 (2008).

[90] S. Izvekov, J. M. J. Swanson, and G. A. Voth. *Coarse-graining in interaction space: A systematic approach for replacing long-range electrostatics with short-range potentials*. Journal of Physical Chemistry B **112**(15), 4711 (2008).

[91] Q. Shi, P. Liu, and G. A. Voth. *Coarse-graining in interaction space: An analytical approximation for the effective short-ranged electrostatics*. The Journal of Physical Chemistry B **112**(50), 16230 (2008).

[92] M. R. Pear and J. H. Weiner. *Brownian dynamics study of a polymer chain of linked rigid bodies*. Journal of Chemical Physics **71**(1), 212 (1979).

[93] D. J. Price and C. L. Brooks. *A modified TIP3P water potential for simulation with Ewald summation*. Journal of Chemical Physics **121**(20), 10096 (2004).

[94] W. Jorgensen, J. Chandrasekhar, J. Madura, R. Impey, and M. Klein. *Comparison of simple potential functions for simulating liquid water*. Journal of Chemical Physics **79**(1), 920 (1983).

[95] B. Guillot. *A reappraisal of what we have learnt during three decades of computer simulations on water*. Journal of Molecular Liquids **101**(1-3), 219 (2002).

[96] D. R. Nutt and J. C. Smith. *Molecular dynamics simulations of proteins: Can the explicit water model be varied?* Journal of Chemical Theory and Computation **3**(4), 1550 (2007).

[97] C. Vega, J. Abascal, M. Conde, and J. Aragones. *What ice can teach us about water interactions: a critical comparison of the performance of different water models*. Faraday Discussions **141**, 251 (2009).

[98] I. C. Yeh and G. Hummer. *Diffusion and electrophoretic mobility of single-stranded RNAfrom molecular dynamics simulations*. Biophysical Journal **86**(2), 681 (2004).

[99] S. Nose. *Constant-temperature molecular dynamics*. Journal of Physics: Condensed Matter **2**(S), SA115 (1990).

[100] M. Green. *Markoff random processes and the statistical mechanics of time-dependent phenomena 2. Irreversible processes in fluids*. Journal of Chemical Physics **22**(3), 398 (1954).

[101] R. Kubo. *Statistical-mechanical theory of irreversible processes 1. General theory and simple applications to magnetic and conduction problems*. Journal of the Physical Society of Japan **12**(6), 570 (1957).

[102] D. Nevins and F. Spera. *Accurate computation of shear viscosity from equilibrium molecular dynamics simulations*. Molecular Simulation **33**(15), 1261 (2007).

[103] J. Kestin, H. Ezzat Khalifa, and R. Correia. *Tables of the dynamic and kinematic viscosity of aqueous NaCl solutions in the temperature range 20-150 C and the pressure range 0.1-35 MPa*. Journal of Physical and Chemical Reference Data **10**, 71 (1981).

[104] S. Feller, R. Pastor, A. Rojnuckarin, S. Bogusz, and B. Brooks. *Effect of electrostatic force truncation on interfacial and transport properties of water*. Journal of Physical Chemistry **100**(42), 17011 (1996).

[105] M. Gonzalez and J. Abascal. *The shear viscosity of rigid water models*. The Journal of Chemical Physics **132**, 096101 (2010).

[106] J. Abascal and C. Vega. *A general purpose model for the condensed phases of water: Tip4p/2005*. Journal of Chemical Physics **123**(23), 234505 (2010).

[107] M. Murphy, I. Rasnik, W. Cheng, T. Lohman, and T. Ha. *Probing single-stranded DNA conformational flexibility using fluorescence spectroscopy*. Biophysical Journal **86**(4), 2530 (2004).

[108] M. Feig and B. M. Pettitt. *Sodium and chlorine ions as part of the DNA solvation shell*. Biophysical Journal **77**(4), 1769 (1999).

[109] B. Dünweg and K. Kremer. *Molecular dynamics simulation of a polymer chain in solution*. Journal of Chemical Physics **99**(9), 6983 (1993).

[110] G. Seisenberger, M. U. Ried, T. Endrebeta, H. Buning, M. Hallek, and C. Brauchle. *Real-Time Single-Molecule Imaging of the Infection Pathway of an Adeno-Associated Virus*. Science **294**(5548), 1929 (2001).

[111] D. Trebotich, G. H. Miller, P. Colella, D. T. Graves, D. F. Martin, and P. O. Schwartz. *A tightly coupled particle-fluid model for DNA-laden flows in complex microscale geometries*. In *Computational Fluid and Solid Mechanics 2005*, pp. 1018–1022 (2005).

[112] S. Chandrasekhar. *Stochastic problems in physics and astronomy*. Reviews of Modern Physics **15**(1), 1 (1943).

[113] P. Baldi. *Calcolo delle Probabilitá e Statistica, 2a edizione (in Italian)* (McGraw-Hill, 1998).

[114] L. Devroye. *Nonuniform random variate generation* (Springer-Verlag, New York, 1986).

[115] F. Panneton, P. L'Ecuyer, and M. Matsumoto. *Improved long-period generators based on linear recurrences modulo 2*. ACM Transactions on Mathematical Software **32**(1), 1 (2006).

[116] M. Schena. *DNA Microarrays: a practical approach* (Oxford University Press, Oxford, 2000).

[117] T. Knotts, N. Rathore, D. Schwartz, and J. D. Pablo. *A coarse grain model for DNA*. Journal of Chemical Physics **126**(8) (2007).

[118] T. Macke and D. Case. *Modeling unusual nucleic acid structures*. In N. Leontes and J. SantaLucia, Jr, eds., *Molecular Modeling of Nucleic Acids* (American Chemical Society, 1998).

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

[120] A. P. Lyubartsev and A. Laaksonen. *Calculation of effective interaction potentials from radial distribution functions: A reverse monte carlo approach*. Physical Review E **52**(4), 3730 (1995).

[121] V. Rühle, C. Junghans, A. Lukyanov, K. Kremer, and D. Andrienko. *Versatile object-oriented toolkit for coarse-graining applications*. Journal of Chemical Theory and Computation **5**(12), 3211 (2009).

[122] S. Jain, I. Saha Dalal, N. Orichella, J. Shum, and R. Larson. *Do bending and torsional potentials affect the unraveling dynamics of flexible polymer chains in extensional or shear flows?* Chemical Engineering Science **64**(22), 4566 (2009).