Effect of Macromolecular Crowding on Diffusive Processes

ZAHEDEH BASHARDANESH
Abstract


Macromolecular crowding are innate to cellular environment. Understanding their effect on cellular components and processes is essential. This is often neglected in dilute experimental setup both in vitro and in silico.

In this thesis I have dealt with challenges in biomolecular simulations at two levels of modeling, Brownian Dynamics (BD) and Molecular Dynamics (MD).

Conventional BD simulations become inefficient since most of the computational time is spent propagating the particles towards each other before any reaction takes place. Event-driven algorithms have proven to be several orders of magnitude faster than conventional BD algorithms. However, the presence of diffusion-limited reactions in biochemical networks lead to multiple rebindings in case of a reversible reaction which deteriorates the efficiency of these types of algorithms. In this thesis, I modeled a reversible reaction coupled with diffusion in order to incorporate multiple rebindings. I implemented a Green's Function Reaction Dynamics (GFRD) algorithm by using the analytical solution of the reversible reaction diffusion equation. I show that the algorithm performance is independent of the number of rebindings.

Nevertheless, the gain in computational power still deteriorates when it comes to the simulation of crowded systems. However, given the effects of macromolecular crowding on diffusion coefficient and kinetic parameters are known, one can implicitly incorporate the effect of crowding into coarse-grain algorithms by choosing right parameters. Therefore, understanding the effect of crowding at atomistic resolution would be beneficial.

I studied the effect of high concentration of macromolecules on diffusive properties at atomistic level with MD simulations. The findings emphasize the effect of chemical interactions at atomistic level on mobility of macromolecules.

Simulating macromolecules in high concentration raised challenges for atomistic physical models. Current force fields lead to aggregation of proteins at high concentration. I probed scenarios based on weakening and strengthening protein-protein and protein-water interactions, respectively. Furthermore, I built a cytoplasmic model at atomistic level based on the data available on Escherichia coli cytoplasm. This model was simulated in time and space by MD simulation package, GROMACS. Through this model, it is possible to study structural and dynamical properties under cellular like environment at physiological concentration.

Zahedeh Bashardanesh, Department of Cell and Molecular Biology, Computational Biology and Bioinformatics, Box 596, Uppsala University, SE-751 24 Uppsala, Sweden.

© Zahedeh Bashardanesh 2019

ISSN 1651-6214
urn:nbn:se:uu:diva-395119 (http://urn.kb.se/resolve?urn=nbn:se:uu:diva-395119)
To Alex and Sophie
List of papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.


III  **Bashardanesh, Z.**, Haiyang Z., Elf, J. and van der Spoel, D. Rotational and translational diffusion of proteins as a function of concentration *Submitted*

IV  Bortot, L. O., **Bashardanesh, Z.** and van der Spoel, D. Making Soup: Preparing and Validating Molecular Simulations of the Bacterial Cytoplasm. *Submitted*

Reprints were made with permission from the publishers.
## Contents

1 Introduction .................................................................................................. 9  
  1.1 Macromolecular Crowding .............................................................. 9  
  1.2 Quantitative/Computational Biology ............................................ 11  
  1.3 Thesis outline ................................................................................. 12  

2 Molecular Dynamics ................................................................................. 13  
  2.1 Equation of Motion ........................................................................ 14  
  2.2 Potential Function .......................................................................... 15  
     2.2.1 Electrostatic Interactions ................................................. 15  
     2.2.2 Harmonic Potentials ........................................................ 15  
     2.2.3 Van der Waals Interactions .............................................. 16  
     2.2.4 Alternatives and Extensions ............................................ 17  
  2.3 Parameter Optimization ................................................................. 19  
  2.4 Force Calculation ........................................................................... 20  

3 Brownian Dynamics - Smoluchowski Approach .................................... 21  
  3.1 Deterministic Modelling ................................................................ 21  
  3.2 Stochastic Modelling ..................................................................... 22  
  3.3 Green’s Function Reaction Dynamics .......................................... 24  

4 Summary of papers .................................................................................... 27  
  4.1 Paper I: Efficient Green’s Function Reaction Dynamics (GFRD) simulations for diffusion-limited, reversible reactions 27  
  4.2 Paper II: Impact of Dispersion Coefficient on Simulations of Proteins and Organic Liquids ................................. 28  
  4.3 Paper III: Rotational and translational diffusion as a function of concentration ........................................................... 31  
  4.4 Paper IV: Making Soup: Preparing and Validating Molecular Simulations of the Bacterial Cytoplasm ........................ 31  

5 Concluding Remarks ................................................................................. 33  

6 Svensk Sammanfattning ............................................................................ 37  

7 Acknowledgments ..................................................................................... 39  

References ........................................................................................................ 41
1. Introduction

1.1 Macromolecular Crowding

Biology is the scientific field that studies living organisms. The most basic unit at which life happens is a cell. Cells are capable of performing complex functions to keep them alive. These cellular functions, such as protein synthesis, deoxyribonucleic acid (DNA) transcription and ribonucleic acid (RNA) translation, are outcomes of multiple physio-chemical processes at various temporal- and spatial scales. Studying cellular functions requires experimental setups with appropriate tools for the temporal- and spatial scale of the question.

Generally speaking, biological experiments are divided in two categories. In vivo refers to the experiments that are performed in living cells, and in vitro to those in which the components of living cells are isolated from their original environment. The results from in vitro experiments do not fully predict the outcomes of in vivo experiments. One reason is the different composition of the intracellular environment and the test tubes.

The intracellular environment is composed of wide range of biomolecular assemblies, from macromolecules such as nucleic acids, proteins and carbohydrates, to much smaller molecules such as metabolites, ions and water. The cell interior is estimated to be up to 40% macromolecules [153, 89, 33]. Water and small molecules are the remaining constituents. For an illustration of interior of cell see Fig. 1.1. Numerous attempts have been made to make the in vitro milieu resemble the cellular environment by addition of high concentrations of synthetic polymers or cell lysate [27, 95, 64]; however, to what extent in vivo measurements are reproducible with artificial supplements is still ambiguous [145, 96]. The effects of the intracellular environment on cellular processes is referred to as (macromolecular) crowding. Some of the problems addressed under crowding conditions are protein folding [133], protein-DNA binding [47] and dynamical properties such as diffusion [82, 144, 145].

Macromolecular Crowding and Structure

Much of the knowledge about biomolecular structures at atomistic resolution are derived from two experimental techniques, X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy [23, 17, 121]. It is not clear how the structure and stability of biomolecules would be altered under non-physiological conditions, but for advances in in vivo NMR see [120].
Figure 1.1. a) Graphic illustration of *Escherichia coli* by David S. Goodsell. Free image downloaded from the website http://pdb101.rcsb.org/sci-art/goodsell-gallery. b) All-atom representation of *E. coli* cytoplasm pertaining to about 1/5000 of its volume.

**Macromolecular Crowding and Dynamics**

Molecular transportation throughout the cell takes place mainly by two means; active transport or through random movements. An active transport is a process in which energy is expended to move substances from lower concentration to higher concentration. On the other hand, random movements of molecules are due to their thermal energy and collisions with neighboring molecules such as water molecules. The substance is transported without expense of energy, down the concentration gradient. This passive transport is often referred to as Brownian motion [14] and it plays an important role in the cellular processes taking place globally throughout the cell.

The presence of molecules in a wide range of sizes, shapes and electrostatic properties creates a heterogeneous environment that affects the diffusive processes [34, 75].

Diffusive properties of macromolecules are crucial in determining the kinetic properties of cellular processes, because an alteration in rates of diffusive processes could change the response of the biochemical network [127]. Knowledge about the alteration in the diffusional properties of biomolecules in living cells or cell-like environments is important in order to understand cellular biology at the molecular level.

**Excluded Volume and Non-Specific Interactions**

This crowding leads to a diminished volume available to biomolecules. Its effect has been studied for a long time using a range of experiments, theory, and simulations. For instance, macromolecular crowding has been imitated by artificial crowding agents *in vitro*. In these studies [151], the association rates
between biomolecules accelerate as the result of high concentration of artificial crowders and in turn, the chemical reaction equilibrium shifts towards the associated state. Proteins fold more compactly and the free energy of protein unfolding increases under these conditions as the excluded volume leaves less conformational space for unfolded proteins. For a review on these studies see [151]. These studies, to some extent, are consistent with predictions from simple theoretical models known as excluded-volume [152, 112], which takes the steric effects into account. However, it is not obvious how well non-reactive polymers can replace biomolecular entities that contain surfaces with positive, negative and neutral charges. Nevertheless, they provide a suitable model for steric repulsion.

More recently, reactive agents, such as cell lysates or protein crowders are used in macromolecular crowding studies, but the results are at times contrary to those studies in which just the inert crowders are used [146]. Both steric repulsion and chemical interactions must be considered to understand the effects of macromolecular crowding [118].

Nonspecific interactions between proteins and crowders are also investigated by simulations [94, 42, 41]. For a review about the effects of macromolecular crowding on biomolecular properties in silico see [44].

The biggest bottleneck in understanding cellular processes under crowded condition is the limitation of the experimental techniques. There are rarely any experimental techniques that can give atomistic resolution of the interior of the cellular environment. Therefore, theoretical and computational tools are valuable resources to give insights and provide new hypotheses.

1.2 Quantitative/Computational Biology

Quantitative biology spans a wide range of problems, from evolution and ecology, to cellular biology and biochemical reactions. From a methodological point of view, this field also spans a wide range of mathematical tools [106], from mean-fields [99] to quantum mechanical descriptions of biochemical reactions [30].

Biological systems are intrinsically complex. Their overall behavior is the result of interplays of many molecular components in a complex network. Even though the completed human genome project [139, 29] has yielded the blueprints of all human genes, this is not sufficient to predict the complex interaction of various networks. Computational approaches are used to capture the features of these complex biological networks. As an example, see the review on mathematical models used in constructing gene regulatory networks [26].

Certainly, an analytical solution to mathematical models is preferable to numerical solutions because obtaining statistical properties then becomes easier. Unfortunately, reaching an analytical solution for complex systems such as
those in biology is not doable. For these complex problems only numerical approaches can provide solutions. There are efficient algorithms developed in a variety of packages [69, 70, 68] for nonlinear but simple systems. However, new algorithms and models are needed when the degrees of freedom in the problem grow or when stochasticity plays an important role, as happens in biological systems [109].

The generated output from quantitative models is often compared with the experimental output. One advantage of computational biology over experimental practice is that it is often cheaper and faster. The possibility of probing scenarios that are not feasible in an experimental setup is another advantage of computational studies. The comparison of generated data from simulations and experiments is a necessary step to either get insight into the underlying mechanism, or to test a new hypothesis by experiment, or to change/improve the underlying quantitative models.

An important step in setting up computational experiments is choosing the level of details of the quantitative model. As a rule of thumb, more detailed models have higher accuracy, at the expense of costly (sometimes close to impossible) calculations. However, the nature of the question is the determinant of the required level of details. For example, understanding the enthalpic contribution to the observed crowding effect cannot be resolved using coarse-grained models, such as simple Brownian Dynamics. On the other hand, understanding the effects of crowding on macroscopic behavior requires models that can resolve the system for a couple of hours of their life time which is currently impossible with atomistic level models, e.g., Molecular Dynamics. This is the classic trade-off between accuracy and efficiency. This trade-off is shown in this thesis between two models, Molecular Dynamics and Brownian Dynamics, that are often used in biomolecular simulations.

1.3 Thesis outline
The introductory chapter is meant to give the background on the biological problem that is the main focus of the work and to name the tools. In chapters 2 and 3, the tools are explained in more details. Chapter 2 deals with the numerical method known as Molecular Dynamics (MD) in solving atomistic movements of molecules. Chapter 3 deals with an analytical model (diffusion equation) and the numerical method Green’s Function Reaction Dynamics (GFRD) to solve the Brownian movements of molecules. Chapter 4 summarizes the scientific papers/manuscripts that have been produced during the thesis work. Chapter 5 finalizes the thesis with concluding remarks on the state-of-the-art.
2. Molecular Dynamics

In order to investigate problems pertaining to living organisms quantitatively, one needs a proper level tool from mathematics, statistics and computational sciences. All living organisms are composed of one or multiple cells that consume energy, grow, and multiply to keep the organism alive, the study of which is collectively called cellular biology. In turn, the fate of each single cell is governed by the interplay of its sub-components. These cellular processes occur at a wide range of temporal and spatial scales. Therefore, one needs to know in advance at which physical level the problem of interest is sufficiently and/or efficiently described. For example, questions regarding the dynamical properties of cellular component at a spatial scale equivalent to the size of the cell cannot be resolved with tools based on electron movement described by quantum mechanics. For a broad overview of biological processes at spatial scale, see Fig. 2.1. However, an atomic resolution investigation of biomolecules at a larger scale than the size of the molecules would be beneficial in understanding some dynamical properties. This has not been much practiced due to high computational costs until recently [45, 78, 150].

Any physical model is based on certain physical and mathematical assumptions and approximations. This practice of simplification allows us to solve larger scale problems. For example, the time-dependent Schrödinger equation could be used, in principle, to predict all properties of any molecule; however, as the number of particles increases the procedure ceases to be efficient if practical at all. Accordingly, Molecular Dynamics, the subject matter of this chapter, is the result of Born-Oppenheimer approximation [19] which assumes the motion of nuclei and electrons of an atom in a molecule are separable. Since the weight of nuclei is by far larger than that of the electrons, the movement of atoms are predictable by their nuclear movement and the electrons are treated in an average fashion, for instance through partial charges. This approximation allows us to build models of molecules and their dynamics based on simple physical laws, such as, Newton Equations of motion, Hooke’s law and Coulomb’s law.

Molecular dynamics (MD) simulation with biological relevance was developed in the 70s for systems with a couple of hundreds of atoms. It has advanced since then, to include e.g., studies of enzymatic reactions [147], protein folding [93] and protein dynamics in water [81]. The applicability of MD simulations to systems with hundreds of thousands of atoms would not be practical without novel algorithmic advances in energy calculations. These are now possible thanks to high performance computing [38] and parallelization [11, 63], or the use of graphical processing units [53, 117, 1]. MD
can now be applied to larger biological systems with explicit solvent [12], membrane embedded proteins [8] or macromolecular complexes such as the nucleosome [113], ribosomes [129] and highly complex systems in the presence of explicit water such as whole viruses [45, 78], fractions of *Mycoplasma genitalium* [150] and Paper IV.

In this chapter, the equation of motion is introduced, the potential function and the models behind its constituents are explained, and some notes on parametrization of functional forms and calculation of forces are provided.

### 2.1 Equation of Motion

Molecular Dynamics simulations are in fact numerical techniques to solve the $N$-body problem, $N$ being the number of atoms representing a (bio)molecular system. Molecular dynamics simulation yields a deterministic trajectory of the state of the (bio)molecular system. In this model, the system represented in 3D (or 2D) coordinates evolves through Newton’s equation of motion,

$$\vec{F} = m \frac{d^2 \vec{x}}{dt^2} \quad (2.1)$$

where $\vec{F}$ is the force exerted on the particle with mass $m$ at the position $\vec{x}$. Eq. (2.1) is not solvable analytically for systems with more than two particles like biomolecular systems. Numerical approximations are used instead to give a trajectory of positions at discrete time points of all particles, as opposed to closed-form analytical solutions. Given the initial position and velocity of all particles, one can use finite difference methods to solve Eq. (2.1) numerically.

There are multiple methods of numerical integration to solve ordinary differential equations. Examples are Euler integration, Leap-frog integration and Verlet integration. Verlet method is the most used one in MD simulations [3].

However, before the integration step in any MD algorithm, the left hand of Eq. (2.1), the force between each two atoms, has to be calculated. The calculation of force through energy is explained in the subsequent section.
2.2 Potential Function

To calculate the forces exerted on particles at each instant of time, the relation between force and energy is used,

\[
\vec{F} = -\frac{dU(\vec{x})}{d\vec{x}} \tag{2.2}
\]

\(U(\vec{x})\) is the potential energy at position \(\vec{x}\). The determination of potential energy for a complex system such as biomolecules is an active field of research [111]. Different energy functions can have different functional forms and different parameters, the determination of which are collectively called force-field development. Examples are AMBER [108, 142, 65, 86], CHARMM [66, 90, 18] and OPLS [58]. Below the most common functional form will be described. However, there are other functional forms that are more accurate, sometimes at the cost of larger calculations [111].

2.2.1 Electrostatic Interactions

In classical point-charge force fields, the electrostatic interaction (Fig. 2.2) is calculated as the sum of interactions between pairs of atoms modeled as point charges through Coulomb’s law,

\[
V_{ij}(r_{ij}) = \frac{1}{4\pi\varepsilon_0} \frac{q_iq_j}{r_{ij}} \tag{2.3}
\]

where \(q_i\) and \(q_j\) are the partial charges of atoms \(i\) and \(j\), \(\varepsilon_0\) the dielectric constant, and \(r_{ij}\) the distance between atom \(i\) and \(j\).

2.2.2 Harmonic Potentials

When atoms share an electron, i.e., they are covalently bonded, the charge distribution around the nucleus cannot be assumed to be symmetrical anymore.
Furthermore, the distance between charges are very small (less than a couple of Ångström). Therefore, simple electrostatic models cannot precisely measure the contribution of the energies of the bonded atoms to the total energy of the systems. Consequently, their electrostatic energy contribution is excluded from the calculation of the total energy of the system. In these instances, i.e., covalently bonded atoms (Fig. 2.3), the potential energy for inter-atomic interaction between two and three atoms at respectively one and two covalent bonds apart are simply modeled by Hooke’s law. The potential energy pertaining to stretching a covalent bond is given by:

\[ v(l) = \frac{k_b}{2} (l - l_0)^2 \]  

(2.4)

where \( k_b \) is the stretching constant of the bond and \( l_0 \) the reference bond length.

The potential energy pertaining to bending a two-covalent bond is given by:

\[ v(\theta) = \frac{k_a}{2} (\theta - \theta_0)^2 \]  

(2.5)

where \( k_a \) is the bending constant of the angle and \( \theta_0 \) the reference angle.

2.2.3 Van der Waals Interactions

Van der Waals interactions are a combination of attractive and repulsive forces between non-bonded atoms (Fig. 2.4). The nature of the energy resulting from these forces shows that it vanishes at a very large distance and is very repulsive at a short distance. The energy reaches a minimum at a distance that is called equilibrium distance. In quantum mechanics terms, the repulsion is due to the overlap of the electron clouds of two atoms. It has been shown that the repulsive potential is accurately represented by an exponential function for some noble gases [21]. Attractive forces at intermediate and larger distances are due to the instantaneous created dipole that induces dipole-dipole moment in the neighboring atoms. These forces are called London dispersion forces, whose potential is the inverse of the sixth power of the distance between two atoms. [79].
The number of van der Waals interactions in a biomolecular system consisting of $N$ atoms scale as $N^2$. This prohibits the use of quantum mechanical models to measure the contribution of these interactions to the potential energy. The most famous model of van der Waals potential function is the Lennard-Jones (12-6) potential [80], in which the attractive part is modeled as a function of $r^{-6}$ and the repulsive part as a function of $r^{-12}$

$$U_{vdw} = 4\varepsilon_{ij} \left( \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left( \frac{\sigma_{ij}}{r_{ij}} \right)^6 \right).$$  \hspace{1cm} (2.6)

Here, $\varepsilon_{ij}$ is the well depth and $\sigma_{ij}$ is the distance at which the energy between atom $i$ and $j$ is zero.

2.2.4 Alternatives and Extensions

Most of the time, the functional forms used in the force fields are chosen such that their derivative is easily obtainable and the computational demand of their calculations is not large. However, this leads to the classic compromise between computational efficiency and accuracy. Below, alternative functions to, or the extension of, the functional forms mentioned above are introduced and discussed briefly.

**Harmonic Potentials**

The harmonic potential functions (Eqs. (2.4) and (2.5)) for a typical bond work reasonably well for biomolecules under normal conditions. However, this model is a simple approximation of more accurate inter-atomic potentials such as the Morse potential [98]. The Morse potential describes the potential energy even when the bond deviates to its bond breaking point and dissociation. The Morse potential has the following form:

$$v(l) = D_e \left( 1 - \exp \left( -a (l - l_0) \right) \right)^2$$  \hspace{1cm} (2.7)
where $D_e$ is the depth of potential energy minimum, $a = \omega \sqrt{\mu/2D_e}$, $\mu$ is the reduced mass, and $\omega$ is the frequency of the bond vibration. $\omega$ is related to the stretching constant of the bond, $k$, by $\omega = \sqrt{k/\mu}$, and $l_0$ is the reference value of the bond. The Morse potential describes the inter-atomic interaction correctly, but it is not often used in force field calculation. It has a high computational cost for calculation of exponential function and it also has an extra parameter compared to the harmonic model. To approximate the Morse potential more accurately one can add cubic, quadratic, and higher terms to the simple harmonic potentials for a bond-stretching model (Eq. (2.4)). This inclusion of an angle-bending model (Eq. (2.5)) can also give better accuracy of force fields, but at the expense of more parameter estimations.

Torsional Term

The torsional energy contributed from the rotation around one bond (see Fig. 2.5) is represented by,

$$v(\phi) = k_t (1 + \cos (m\phi - \gamma))$$

(2.8)

where $k_t$ is the force constant of torsion, $\phi$ is the current torsional angle, $m$ is the multiplicity (the number of energy minima as the bond is rotated 360°), and $\gamma$ is the phase angle (0 or $\pi$).

Polarized Charges

Point-charge models, as described previously, are broadly used in many biomolecular force fields, despite the importance of the polarization effect in condensed phases or biomolecular systems. Polarization effect refers to the fact that the charge density around the nucleus is not fixed—it's distribution varies depending on the environment. For example, in the simple case of a water molecule, the dipole moment of water varies from 1.9 $D$ in the gas phase [83] to 2.1 $D$ in a water cluster [54] and to 2.9 $D$ in bulk water [9]. In a more complex system such as protein ligand binding, the explicit polarization treatment of electrostatic charges yields a reliable prediction of binding free energy [62]. However, in most force fields, the electrostatic properties of atoms are treated
by placing a point charge at the site of the nucleus where its magnitude is
determined beforehand and kept constant.

**Van der Waals Potentials**
Other alternatives to van der Waals potential include 7-14LJ, 6-10LJ and the
Buckingham potential. For more details, see [79].

Putting all the functional forms together the expression below needs to be
calculated in order to calculate Eq. (2.2),

$$ U = \sum_{bonds} \frac{k_{bm}}{2} (l_m - l_{m,0})^2 + \sum_{angles} \frac{k_{an}}{2} (\theta_n - \theta_{n,0})^2 $$

$$ + \sum_{i} \sum_{j \neq i} \left( 4 \varepsilon_{ij} \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left( \frac{\sigma_{ij}}{r_{ij}} \right)^6 \right) + \frac{q_i q_j}{4 \pi \varepsilon_0 r_{ij}} \right) $$

where, $m$ and $n$ enumerate the number of bonds and angles, respectively,
and $i$ and $j$ enumerate the number of atoms $N$. In the next section the determina-
tion of the parameters that are introduced above will be explained. Parameter-
ization is of utmost importance in developing force fields after choice of func-
tional forms.

2.3 Parameter Optimization
The main two challenges in developing force fields are the choice of func-
tional forms (discussed in previous section) and accurate parameters. First,
the functional forms need to be computationally efficient and realistic enough
to capture meaningful physical behavior. Seconds, correct parametrization is
crucial in force-field development, which depends on abundant and reliable
data from experiments and/or QM measurements, and efficient, reliable opti-
mization algorithms. Roughly speaking, the parametrization protocol includes
the choice of target data, the order of parameters to be determined, and opti-
mization methods. Each force field contains a large number of parameters,
some of them are more crucial than others in reproducing certain properties.
These properties include structural properties such as conformational energies
and vibrational frequencies, and thermodynamic properties such as densities
and heat of vaporization.

Due to the important role of the non-bonded parameters in thermodynamic
properties and their role in biomolecular structure and interactions, their proper
optimization is essential for any reliable biomolecular force field. Non-bonded
parameters include those in van der Waals interactions and electrostatic interactions.

One approach is to perform the parametrization by trial and error, i.e., the parameters are refined gradually to give better fits to the targeted data. Another approach is to define a penalty function and perform least-squares fitting [84].

Because different force fields have different functional forms and hence parameters and/or different parametrization protocols, it is not possible to use one sets of parameters for all force fields, i.e., the parameters are not transferable among different force fields.

2.4 Force Calculation

To calculate force through potential function (Eqs. (2.2) and (2.9)), one needs the spatial derivative of the latter equation. The potential functions described above have rather simple analytical derivation which makes them favorable to calculate. As is clear from Eq. (2.9), the calculation of energy for bonded terms scale as $N$, the number of particles. However, it scales as $N^2$ for non-bonded terms. The magnitude of non-bonded interactions decreases as $1/r^{-1}$ for electrostatic interactions, and as $1/r^{-6}$ for van der Waals interactions. Therefore, one solution is to calculate the non-bonded interactions only for atoms falling within a certain distance $r_c$, known as the cutoff distance. This solution is not free from artifacts created by truncating a smooth function abruptly and solutions for it have been proposed [134]. In fact, the effect of long-range interactions on the biomolecular systems have been researched since 1993. For example, for electrostatic interactions, a method widely used is the Particle Mesh Ewald (PME) sum [31, 38]. The magnitude of van der Waals interactions falls off rapidly as $r^{-6}$, since their magnitudes are all negative (unlike electrostatic interaction) the effect of neglecting them is quite large. For long-range van der Waals interactions the addition of an analytical correction known as dispersion correction is proposed to account for those interactions beyond the cutoff distance [3]. The method of PME has been proposed and implemented for van der Waals interactions in the GROMACS software [148].
3. Brownian Dynamics - Smoluchowski Approach

As was mentioned in the introduction, the dimension of a problem determines the choice of research tool. Molecular Dynamics (MD) is an appropriate tool when the question has an atomistic nature (Fig. 2.1) and a sub-microsecond temporal scale. However, when the problem at hand is to understand the interactions and reactions of several species with hundreds to thousands of molecules for a duration of hours, MD simulations are not appropriate tools.

A common mathematical tool to study (bio)chemical kinetics of this type of problem is differential equations. One of the earliest examples of using differential equations goes back to 1850 [149], when the conversion of sucrose to glucose and fructose was described by using ordinary differential equations (ODE). However, incorporation of the spatial gradient of the molecular species to the model is necessary when the problem of interest is subject to spatial non-homogeneity e.g., due to slow diffusion or compartmentalization in the interior of cells. For this, partial differential equations (PDE) are used to describe the changes in molecular species in a continuous manner with respect to time and space. These models are called reaction-diffusion systems. The diffusion term describes the transportation of matter and the reaction term models the transformation process. Reaction-diffusion systems have been used to describe a wide range of problems in physics, chemistry and biology [136]. An example is two populations of prey and predator and their spread in the environment. The populations can be modelled by two sets of coupled reaction-diffusion equations, describing the birth and decay of each species by natural causes and their interactions. The spread of each species is captured by a diffusion term with a predefined diffusion coefficient for each species, in a model known as the Lotka-Volterra [87, 141] model or prey-predator model [88, 20]. Another example of reaction-diffusion systems to explain biological phenomena is in pattern formation such as stripes and spots (also known as Turing patterns [132]). These models, together with their analytical solutions [123], are widely used to explain Morphogenesis i.e., embryonic development of organisms.

3.1 Deterministic Modelling

One way of deriving the diffusive part of the reaction-diffusion equation is through Fick’s first law. This law asserts that the flux is linearly related to the
gradient of concentration,

\[ j = -D \frac{\partial c}{\partial x} \]  \hspace{1cm} (3.1)

and the conservation of mass that it results in (expanded just in one dimension here):

\[ \frac{\partial c}{\partial t} = -\frac{\partial j}{\partial x}. \]  \hspace{1cm} (3.2)

Combining Eqs. (3.1) and (3.2) leads to the diffusion equation,

\[ \frac{\partial c}{\partial t} = D \frac{\partial^2 c}{\partial x^2}. \]  \hspace{1cm} (3.3)

The reaction equation is derived through the law of mass action [55], which states that the rate of a chemical reaction is proportional to the masses of reactants.

\[ \frac{\partial c}{\partial t} = -k_d c \]  \hspace{1cm} (3.4)

where, \( k_d \) is the rate at which the reactant degrades. The so-called reaction diffusion equation is therefore written as:

\[ \frac{\partial c}{\partial t} = D \frac{\partial^2 c}{\partial x^2} - k_d c. \]  \hspace{1cm} (3.5)

In the macroscopic picture of (bio)chemical reactions above, it is assumed that the reagents of reactions have high concentrations. However, not all molecular species appear in high copy numbers in inter/intra cellular reactions. An example is gene expression, which involves low copy number of transcription factors, polymerase and mRNAs. Therefore, a macroscopic approach is not a suitable analytical tool to explain and predict these processes. An alternative and accurate way of modeling cellular processes with a low copy number of molecules and with reactions happening at random times is stochastic modelling [51, 36].

### 3.2 Stochastic Modelling

The chemical master equation (CME) is a widely used stochastic model to explain and predict the molecular copy number of a species that changes through chemical reactions. CME is a set of Ordinary Differential Equations (ODEs) that govern the evolution of probability distribution functions of certain number of molecules of each species at a certain time [136]. Except in a few special cases [46, 50], the CME is not solvable analytically.

Stochastic simulation algorithms (SSAs) such as the one by Gillespie [49] are the most well-known and well-extended algorithms to simulate analytically non-solvable systems. An essential assumption in CME is a well-stirred
environment where the diffusion is fast enough to create a homogeneous environment in between reactions. Some cellular processes involve molecular species with slow diffusion that create a non-homogeneous environment. In such cases, CME is not an appropriate model. Cellular processes with reactions happening stochastically in a non-uniform environment are primarily simulated by two particle-based models, Brownian Dynamics (BD) and Reaction Diffusion Master Equation (RDME).

RDME dates back to 70s [76, 48, 102] and can be interpreted as a spatial extension of the CME. In RDME, the space is divided into sub-volumes with a mesh size small enough to guarantee the assumption of the uniform distribution of particles within each voxel. Therefore, the same procedure as for CME can be applied to each sub-volume for the reactions in addition to the diffusion of particles to adjacent voxels [40, 59, 143]. The state of the system is tracked by the copy number of each species in each sub-volume. In spite of the progress in speeding up [32] and adjusting RDME to different geometries and unstructured meshes [35], seeking the smallest mesh size to validate the approach is still controversial [39, 37, 71, 60]. It has been shown [37] that by decreasing the mesh size relative to molecule size, no bimolecular reactions can take place. Instead, the molecules simply diffuse and are transformed by monomolecular reactions.

In contrast to RDME, the BD algorithm is a particle-based off-lattice model. Each particle in the system is characterized by its size and species and their positions are tracked individually. The particles undergo a random walk which is governed by the diffusion equation. In most of the BD algorithm developments, the diffusing steps are sampled from a Gaussian distribution by choosing a fixed time step.

As in RDME, the monomolecular reactions in BD algorithms are modeled as a Poisson process. However, the bimolecular reactions are not explicitly modeled in most of the algorithms [5, 110, 7]. For a discussion of different implementations of bimolecular reactions see [125]. Another drawback with time-driven BD algorithms is the large computation time due to small time steps. Increasing the time steps is a way of improving the efficiency; however, this could result in missing reactions between particles.

To make sure no reaction will be missed, a regime of BD algorithms known as event-driven algorithms has been developed. In these algorithms, the next time step is calculated such that an interesting event happens, e.g., monomolecular or bimolecular reactions. One example of these algorithms is Green’s Function Reaction Dynamics (GFRD) [138, 137]. GFRD uses the exact solution of the Smoluchowski equation to combine the diffusion in space and the reaction between particles into a single step. This allows GFRD to take large jumps in space and time to propagate the particles in systems with low concentration. Although the cost to compute every iteration is high, these algorithms are up to 5 orders of magnitude faster than time-driven algorithms [137]. In the GFRD algorithm, the calculated time to the next reaction is also compared
with the time step for the next diffusive step. Therefore, a cut-off distance is introduced for calculating the diffusion of particles. Introducing the cut-off distance makes the algorithm inaccurate to a certain extent, because there is a non-zero probability that the particle jumps to a distance beyond the cut-off distance. In a later version of the GFRD algorithm, eGFRD, this issue is resolved by introducing an absorbing boundary condition, see [103, 127].

3.3 Green’s Function Reaction Dynamics

GFRD is an event-driven BD algorithm to simulate reaction diffusion systems at the particle level. A reaction diffusion system is composed of many particles that diffuse in space and react upon contact with each other. An analytical solution is not obtainable for a reaction-diffusion system with more than two particles. The original idea of GFRD was to divide the many-body system into one- and two-body problems and find an exact analytical solution for each system. The use of Green’s function solutions to set up an event-driven algorithm allows GFRD to take large jumps in space and time at each time step. Below the modeling behind GFRD is explained in more detail.

Biochemical reactions: monomolecular or bimolecular reactions

A cellular process can be depicted schematically with three basic reactions below (as many as needed):

\[ \emptyset \rightarrow X \] \hspace{1cm} (3.6)
\[ X \rightarrow ... \] \hspace{1cm} (3.7)
\[ X + Y \rightarrow ... \] \hspace{1cm} (3.8)

The monomolecular and bimolecular reactions above are reaction channels that cells use to transform one set of chemical species to another at a specific rate. Here the dynamics and kinetics of monomolecular and bimolecular reactions are modelled without going into the details of atomistic resolution in MD. In fact, molecules are considered spheres with radii to model their sizes. The spheres are reactive symmetrically.

Molecular transportation is modeled as a random walk

The diffusion equation can be derived from a random walk model. Here, we consider the random walk in one dimension, without loss of generality. We derive the equation for evolution of \( p(x,t) \), the probability density that the particle is at position \( x \) at time \( t \). We assume that the random walk is an unbiased walk, which means the probability of going either right or left is equal \( (p = q = 1/2) \). We assume that a particle takes a jump of length \( \Delta x \) in one unit of time, \( \Delta t \). If we look at the probability of being at position \( x \) in
one unit of time in the future, $t + \Delta t$, this happens either if the particle is at position $x + \Delta x$ at time $t$ and jumps to the left. Or if it jumps to the right when the particle is at position $x - \Delta x$ at time $t$. Given the Markov property of such processes, one can sum over the probabilities as follows:

$$p(x, t + \Delta t) = \frac{1}{2} p(x - \Delta x) + \frac{1}{2} p(x + \Delta x).$$

Subtracting $p(x, t)$ from both sides of Eq. (3.9) and dividing both sides by $\Delta t$ and with a slight manipulation, it results in:

$$\frac{p(x, t + \Delta t) - p(x, t)}{\Delta t} = \frac{(\Delta x)^2}{2\Delta t} \left( \frac{p(x - \Delta x) - 2p(x, t) + p(x + \Delta x)}{(\Delta x)^2} \right).$$

The equation above becomes the diffusion equation at infinitesimal values for $\Delta t$ and $\Delta x$:

$$\partial_t p(x, t) = D \partial_x^2 p(x, t)$$

where $\frac{(\Delta x)^2}{2\Delta t}$ is replaced with diffusion coefficient $D$.

**Monomolecular reactions are modeled as Poisson processes**

The transformation of one species to another or the degradation of species in Eq. (3.7) is independent of molecular diffusion or any spatial gradient, unless the species is involved in localized biochemical reactions such as those happening on the membrane. Moreover, the transformation is assumed to be instantaneous. For such description, a good approximation is a Poisson process with the predefined rate as the average lifetime of the corresponding species. Therefore, the waiting time for the next monomolecular reaction is distributed exponentially.

**Bimolecular reactions are modeled with boundary conditions**

In contrast to monomolecular reactions, bimolecular reactions are not easily separable from their diffusive dynamics when the particles are in close vicinity of each other. Let’s consider an isolated pair of molecules, $A$ and $B$, that diffuse with diffusion coefficients $D_A$ and $D_B$ and associate at the contact surface $\sigma = \sigma_A + \sigma_B$:

$$A + B \xrightarrow{k_a} C$$

where $\sigma_A$ and $\sigma_B$ are the radii of particles $A$ and $B$, respectively. Here, $k_a$ is the association rate at contact and $C$ the product. The probability distribution function $p(\vec{r}_A, \vec{r}_B, t)$ of the position of the particles, $\vec{r}_A$ and $\vec{r}_B$ at time $t$, is governed by the diffusion equation below,

$$\partial_t p(\vec{r}_A, \vec{r}_B, t|\vec{r}_{A0}, \vec{r}_{B0}, t_0) = \left( D_A \Delta \vec{r}_A + D_B \Delta \vec{r}_B \right) p(\vec{r}_A, \vec{r}_B, t|\vec{r}_{A0}, \vec{r}_{B0}, t_0),$$

where $\sigma_A$ and $\sigma_B$ are the radii of particles $A$ and $B$, respectively. Here, $k_a$ is the association rate at contact and $C$ the product. The probability distribution function $p(\vec{r}_A, \vec{r}_B, t)$ of the position of the particles, $\vec{r}_A$ and $\vec{r}_B$ at time $t$, is governed by the diffusion equation below,

$$\partial_t p(\vec{r}_A, \vec{r}_B, t|\vec{r}_{A0}, \vec{r}_{B0}, t_0) = \left( D_A \Delta \vec{r}_A + D_B \Delta \vec{r}_B \right) p(\vec{r}_A, \vec{r}_B, t|\vec{r}_{A0}, \vec{r}_{B0}, t_0),$$

where $\sigma_A$ and $\sigma_B$ are the radii of particles $A$ and $B$, respectively. Here, $k_a$ is the association rate at contact and $C$ the product. The probability distribution function $p(\vec{r}_A, \vec{r}_B, t)$ of the position of the particles, $\vec{r}_A$ and $\vec{r}_B$ at time $t$, is governed by the diffusion equation below,
where $\vec{r}_A$ and $\vec{r}_B$ are the initial positions of the particles at time $t_0$ and $\Delta$ is the Laplacian operator. In an attempt to solve the Eq. (3.13), one needs to simplify it by changing the position vectors $\vec{r}_A$ and $\vec{r}_B$ to the inter-particle distance vector $\vec{r}$ and the center of diffusion vector $\vec{R}$ through the following expression:

$$\vec{R} = \alpha \vec{r}_A + \beta \vec{r}_B,$$

$$\vec{r} = \vec{r}_B - \vec{r}_A$$

where, $\alpha$ and $\beta$ are chosen such that $p(\vec{r}_A, \vec{r}_B, t | \vec{r}_A^0, \vec{r}_B^0, t)$ can be written as $p(\vec{R}, \vec{r}, t | \vec{R}_0, \vec{r}_0, t)$ and further factorized in two independent processes with

$$p(\vec{R}, \vec{r}, t | \vec{R}_0, \vec{r}_0, t) = p^R(\vec{R}, t | \vec{R}_0, t_0)p^r(\vec{r}, t | \vec{r}_0, t_0).$$

The rule above imposes conditions on the positive $\alpha$ and $\beta$ as follows:

$$\begin{align*}
\alpha & = \frac{D_B}{D_A}, \\
\beta & = \frac{D_A}{D_B}, \\
\alpha^2 D_A + \beta^2 D_B & = D_R.
\end{align*}$$

Here, $D_R$ is the diffusion coefficient of the center of diffusion vector. Applying the changes above we obtain the diffusion equation for the center of diffusion

$$\partial_t p(R, t | R_0, t_0) = D_R \Delta p(\vec{R}, t | \vec{R}_0, t_0)$$

and the diffusion equation for the inter-particle distance vector

$$\partial_t p(\vec{r}, t | \vec{r}_0, t_0) = (D_A + D_B) \Delta p(\vec{r}, t | \vec{r}_0, t_0), \quad \sigma \leq r.$$  

The solution to Eq. (3.16) with a point source initial condition and absorbing boundary condition at infinity is the Gaussian solution.

The solution to Eq. (3.17) is derived in [25] with the following initial and boundary conditions:

$$p(\vec{r}, t_0 | \vec{r}_0, t_0) = \delta(\vec{r} - \vec{r}_0)$$

$$4\pi \sigma^2 D \partial_r p(\vec{r}, t | \vec{r}_0, t_0) |_{|\vec{r}|=\sigma} = k_a p(\vec{r}, t | \vec{r}_0, t_0) |_{|\vec{r}|=\sigma}$$

where the bimolecular reaction—association—is incorporated into the boundary condition, here $k_a$ is the intrinsic association rate. This problem is an extension of the Smoluchowski problem with a reactive boundary condition, see [28].

GFRD is slow compared to RDME specifically when it comes to the diffusion-limited reactions that are characteristic of many biological processes. After each dissociation event, the particles stay in the vicinity of each other due to low diffusivity. Then they re-associate before diffusion moves the particles apart. This problem is addressed in Paper I.
4. Summary of papers

4.1 Paper I: Efficient Green’s Function Reaction Dynamics (GFRD) simulations for diffusion-limited, reversible reactions

The Green’s function reaction dynamics (GFRD) algorithm [138, 137] was proposed to resolve the inefficiency of Brownian Dynamics (BD) algorithms by taking larger time steps, as explained in Chapter 3, and to apply Smoluchowski theory for bimolecular reactions. Another advantage of the GFRD algorithm is that it obeys the detailed balance principle when it comes to reversible reactions. In GFRD, a pair of particles can associate when they reach each other’s contact surface and the products of the transformation of a single particle are put at contact surface. However, the efficiency of GFRD deteriorates when the reaction system involves diffusion limited reactions. These reactions include molecular species that have a slower diffusion than their association rate. This results in multiple re-bindings after each dissociation until diffusion separates the particles from each other. This is shown schematically in Fig. 4.1. In GFRD, many calculations should be carried out for sampling time to the next association and dissociation events. However, in reversible GFRD (rGFRD) the time to separation is sampled once. In this paper we propose a model for a single particle that can dissociate and re-associate reversibly until the diffusion moves the product away to a distance where the probability of rebinding is low.

Consider the schematic reversible reaction

\[ C \underset{k_d}{\overset{k_a}{\rightleftharpoons}} A + B \]  

(4.1)

together with transformation or decay reactions

\[ A \xrightarrow{\gamma_A} \ldots, B \xrightarrow{\gamma_B} \ldots, C \xrightarrow{\gamma_C} \ldots \]  

(4.2)

Here, \( k_a \) is the intrinsic association rate, \( k_d \) the intrinsic dissociation rate, and \( \gamma_A, \gamma_B \) and \( \gamma_C \) are transformation or decay rates of particles \( A, B \) and \( C \), respectively. The probability distribution function of inter-particle distance \( r \) between particles \( A \) and \( B \) satisfies the equation

\[ \partial_t p(r,t) = D \Delta_r p(r,t) - \gamma_a p(r,t) \]  

(4.3)
where $D$ is the sum of diffusion coefficients of particles $A$ and $B$ and $\gamma = \gamma_A + \gamma_B$. The distance $r$ is restricted to the shell

$$\sigma \leq r \leq b$$

(4.4)

where $\sigma$, the sum of the radii of particles $A$ and $B$ is the distance at which a reaction can take place and $b$ is an arbitrarily chosen distance with the absorbing boundary condition

$$p(b, t) = 0$$

(4.5)

at which the separation of $A$ and $B$ occurs. The reaction at the contact surface is modelled as a flux at the inner boundary sphere through

$$4\pi\sigma^2 D \frac{\partial p(r, t)}{\partial r}|_{r=\sigma} = k_a p(r, t)|_{r=\sigma} - k_d Q(t)$$

(4.6)

The probability of being bound as $C$ at time $t$ is denoted by $Q(t)$ and evolves as

$$\frac{dQ(t)}{dt} = -k_d Q(t) - \gamma C Q(t) + k_a p(r, t)|_{r=\sigma}$$

(4.7)

We derive the analytical solution of the equation above in the rGFRD algorithm to speed up the algorithm. Fig. 4.2 shows the CPU time for both implementations of GFRD and rGFRD for different values of re-bindings, $\mathcal{N}_d$. The relation between $k_a$ in Eq. 4.1, $D$ in Eq. 4.3 and $\mathcal{N}_d$ is $\mathcal{N}_d = 1 + \frac{k_a}{k_D}$, $k_D = 4\pi\sigma D$. We have shown that by increasing the number of re-associations, the execution time of GFRD algorithm increases linearly, however, it remains constant for rGFRD algorithm (Fig. 4.2).

4.2 Paper II: Impact of Dispersion Coefficient on Simulations of Proteins and Organic Liquids

Biological environments are densely packed with macromolecules. It is often difficult to get detailed insight into how the crowding affects what we know about cellular processes from experimental means. An alternative is to use computational techniques. The structure and dynamics of biomolecules can be tracked at high resolution at both spatial and temporal scales. Molecular Dynamics (MD) has the ability to get atomistic resolution of biomolecular dynamics and structure. However, MD simulations of concentrated models of proteins have shown non-physical coagulation of them [105, 2, 107]. The protein-protein interactions are too strong when simulated with standard force fields. One explanation is that the force fields have often been initially parameterized for simple and small-model systems, or proteins in diluted solutions. One way to resolve this issue is to modify the protein-protein or protein-water interactions. Strengthening the protein-water Lennard-Jones interactions improves the properties of disordered proteins [15]. Another explanation of the
Figure 4.1. Comparison of the time sampling in GFRD and rGFRD. rGFRD samples a longer time step than GFRD when it comes to reversible reactions. In rGFRD, the idea is to integrate over the short-lived time steps of all re-association events after each dissociation.

non-physical behavior of concentrated models suggests that the dispersion interactions may have been overestimated due to short cutoffs while parameterizing force fields [97]. In this work, we tested the latter explanation by systematically reducing the dispersion coefficient in the 12-6 Lennard-Jones (LJ) potential. The LJ potential is given by:

\[ U_{vdW}(r_{ij}) = 4\varepsilon_{ij} \left( \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left( \frac{\sigma_{ij}}{r_{ij}} \right)^6 \right) = \frac{C_{12}}{r_{ij}^{12}} - \frac{C_6}{r_{ij}^6} \] (4.8)

where \( r_{ij} \) is the distance between atoms \( i \) and \( j \), and \( \sigma_{ij} \) is the distance at which the energy is zero, \( \varepsilon_{ij} \) is the depth of the potential well. \( C_{12} \) and \( C_6 \) are repulsion and dispersion coefficients, respectively. They are derived from the distance and well-depth parameters through

\[ C_{12} = 4\varepsilon_{ij}\sigma_{ij}^{12}, \quad C_6 = 4\varepsilon_{ij}\sigma_{ij}^6. \] (4.9)

Here, we reduced the \( C_6 \) for non-water atom types by 0, 1, 2 and 4 \% to study the protein solutions and neat liquids. For neat liquids simulations, two general force fields were chosen: the generalized AMBER force field (GAFF) and the CHARMM general force field. The LJ long-range interactions were calculated with particle mesh Ewald (PME). For protein solutions, we tested two families
Figure 4.2. Comparison of the execution time for GFRD and rGFRD. Average execution time of 10 simulations is shown for a reversible reaction $A + B \rightleftharpoons C$ for a wide range of $N_d$ using the original GFRD and the rGFRD. The parameters of the system are $k_D = 0.2 \mu m^3 s^{-1}$, $k_a = (N_d - 1) k_D \mu m^3 s^{-1}$, $k_d = k_a/((2V_R) s^{-1}$ and $R = 1 \mu m$.

of force fields: AMBER ff99SB-ILDN [142, 65, 86] and CHARMM27[90, 18]. We used two methods for long-range LJ interaction calculations: cut-off (for AMBER ff99SB-ILDN) and PME (for both AMBER ff99SB-ILDN and CHARMM27). We also simulated the protein systems with scaled AMBER force fields designed for crowded systems to compare them with ours.

We found that the generalized force fields, AMBER and CHARMM, reproduce the experimental densities for neat liquids with little or no corrections. Reducing the dispersion coefficient led to underestimation of densities. For biomolecular force fields, the highest degree of reduction of long-range forces resulted in less coagulation. However, the scaled AMBER force field designed by Best [15] still performed better. Since reducing the long-range attractive forces even more may lead to unfolding and destabilizing the biomolecules structure, this solution is not favorable. Furthermore, since biomolecules and liquids force fields share atom types, a different solution is desirable for designing a force field for crowded system.
4.3 Paper III: Rotational and translational diffusion as a function of concentration

The interior of cells is densely packed (chapter 1). Macromolecular crowding affects the dynamical properties of proteins as well as the structural and kinetic properties [145, 33, 152, 118, 112]. It has been shown that the translational diffusion coefficient of Green Fluorescent Protein in E. coli measured in vivo is 10% of its in vitro value [34, 75]. However, an atomistic insight into the problem has not been tractable by an experimental techniques so far. In this work, we used computational techniques to simulate a simple model of crowded environment at atomistic level by increasing the concentration of proteins in a systematic way. We used Molecular Dynamics to simulate the systems for one microsecond with a force field designed for crowded systems [15]. The study was done for three proteins: Ubiquitin [140], 2kim [6], and 2k57 [115]. We focused on the dynamical properties of proteins, i.e., the translational and rotational diffusion. The translational diffusion coefficient and rotational correlation time (global tumbling) were calculated. These results were compared with hard sphere (HS) models that took into account the excluded volume effect. The full atomistic model of the concentrated proteins probed the effect of electrostatic forces in addition to steric forces. We found that the reduction in rotational diffusion has a larger spread and it depends on the proteins in our study. However, the reduction in translational diffusion is similar for different proteins. The difference between our simulations (MD) and HS models are clear for the proteins in our study which suggests that the excluded volume effect explains the reduction in translational diffusion to some extent. The effect of electrostatic interactions could be seen by formation of transient clusters in our study. This cannot be explained with simple HS models. Moreover, inclusion of explicit water and also study of rotational diffusion is not possible with HS models.

4.4 Paper IV: Making Soup: Preparing and Validating Molecular Simulations of the Bacterial Cytoplasm

In this work we took a step forward towards whole-cell simulation. We built a model of E. coli at atomistic resolution that can be used for resolving its dynamics with MD simulations. This model is available as a Python script on github repository https://github.com/dspoel/soup. We used this model to set up a simulation of the system that represents 1/5000 of E. coli volume. Incorporating a full list of proteins, nucleic acids and metabolites in a full atomistic model is prohibitive in terms of computational resources. Therefore, we selected the most abundant non-ribosomal proteins; these account for 50% of the total non-ribosomal proteins numbers. To account for nucleic acids, we chose RNAs, because 74% of the dry weight of non-ribosomal RNAs is
composed of tRNA, i.e., 5% of the total protein weight corresponds to the total RNA weight. tRNA_{Phe} was chosen as representative of tRNAs in our model because of the availability of the crystallographic structure [22].

The metabolite fraction of our cytoplasm model was built based on data that the number of metabolite molecules in the cytoplasm of \textit{E. coli} is 42.86 times greater than the number of proteins [13]. We considered the most abundant metabolite as representative of each metabolite class, i.e., amino acids, nucleotides, central carbon intermediates and, cofactors.

The number of water molecules to be added to the system depends on the total biomolecular mass and concentration. We chose 30% biomolecular concentration and calculated the number of water molecules to reach this value. At the end, we added 0.15 M ionic strength to our model to include cations, like Mg\(^{2+}\) and K\(^{+}\) and anions Cl\(^{-}\). We used the Molecular Dynamics (MD) package, GROMACS 2018, to simulate the cytoplasm model in three replica for 1\(\mu\)s for each replica. Additionally, we simulated each protein under a dilute condition for 200\(ns\) for each replica.

We validated our cytoplasm model by comparison of diffusion coefficient reduction with experimental results. Translational diffusion coefficient drops by around 85\% for the proteins and 93\% for the tRNA in our study. These findings are in agreement with experimental results showing that the diffusion coefficient of Green Fluorescent Protein (GFP) in \textit{E. coli} decreased 90\% when compared to \textit{in vitro} measurements [34, 75]. Our results show also that the drop in diffusion coefficient is independent of the size of the molecules. We found that higher drop of diffusion coefficient of tRNA is because of aggregation of ATP\(^{-3}\) and FBP\(^{-4}\) around Mg\(^{2+}\) that is used as cation for tRNA molecules. Therefore, we treat tRNA as an outlier and it was excluded from the rest of the analysis. Furthermore, we suggest to use fully protonated ATP\(^{-3}\) and FBP\(^{-4}\) in the model to avoid the aggregation problem.

In our study we found that most of the proteins are more flexible in cytoplasm model than in the dilute condition. This is in agreement with previous computational study [42] and experiments [96, 146] showing that non-specific interactions between proteins and their native environment leads to proteins being less stable.
5. Concluding Remarks

Atomistic models of biological macromolecules provide the highest level of detail. However, for investigating macromolecular behavior in a cellular environment, this is not sufficient unless the whole entity of the biological organism is modeled at an atomistic level. Whole cell modeling is not a new topic in biological studies. A mathematical model for single bacterial cell growth was proposed in late 70’s [122]. In this model, the concentration of the bacterial components was modelled for signal transduction, genetic regulatory networks and metabolic networks through reaction rate equations (kinetic models).

From a hypothetical whole-cell computational model, one can predict cellular phenotypes based on their genome. Beside basic scientific purposes, whole cell modelling may have medical applications. The effect of new drugs can be investigated on the entire cell behavior rather than on a single target.

Cellular environments are complex in many ways. Firstly, cellular processes occur at spatial and temporal scales that span over many orders of magnitudes. For example, macromolecular atomistic details have 0.1 nanometer resolution whereas whole biological processes happen at a few hundred nanometers to a hundreds of micrometers. From a temporal point of view, chemical reactions happen on the order of femtoseconds whereas cell division time for E. coli is 20 minutes. Secondly, cellular environments contain components at different level of concentrations and structural complexity. These range from macromolecules such as proteins, nucleic acids, and carbohydrates, to ions and water. All have different shapes and sizes in addition to different abundance. Another complexity pertains to cellular processes. Over the life of the cell, the different processes that take place are modulated in a complex way. Most often the cellular pathways are studied separately.

From a technical point of view, whole cell modelling poses grand challenges. In recent years, there has been an abundance of high-throughput data from various studies on cellular modules, e.g., transcriptomics, proteomics and metabolomics. However, acquiring homogeneous knowledge across different cellular pathways through a technology that spans all cellular functions is still a big challenge. Furthermore, some cellular processes are more studied than others. Signal transduction, metabolic network and regulatory networks are far more investigated than cell differentiation and apoptosis [61]. Therefore, data curation is one of the grand challenges on the way towards building a whole cell model. Another difficulty is the discrepancy between models among different cellular pathways. For instance, most often flux balance analysis (FBA) [104] is used in metabolism models, but ordinary differential equations (ODEs) in modeling signal transduction pathways [74]. Even
though there is a possibility to model all these pathways with a unique model e.g., ODEs, the large temporal difference in various cellular processes leads to stiff ODE’s that are not easily integrable.

Therefore, integrating the best model and all knowledge known for each pathway into a unified model are the greatest challenges on modelling the whole cell. For more details on principles and challenges in whole-cell modelling see [91, 73, 52, 126].

In spite of the challenges, there are examples of whole-cell models of different organisms at different levels of approximation. A project known as E-cell started as in 1999 [131] to build a platform for a whole-cell simulation. Through this software a hypothetical cell with a minimal number of genes sufficient for its survival was built based on differential equations. This hypothetical cell is derived from Mycoplasma genitalium bacterium. M. genitalium has only $\sim 480$ genes, almost one order of magnitude fewer than E. coli. Its genetic sequence has been published (http://www.tigr.org/), which makes it a good candidate for whole-cell modeling. Another whole-cell computational model of the bacterium M. genitalium at molecular level can be found in [72]. Other examples of organisms or cell type models are the computational one of erythrocytes [130], a data-driven model of E. coli [24], and those are found in [16, 110, 94, 4]. Furthermore, a model of cytoplasm at atomistic level has been built based on M. genitalium [41]. This model integrates genomic and metabolic data with physical and biochemical principles. In Paper IV, we have constructed an atomistic level of cytoplasm that pertains to 1/5000 of the volume of E. coli cytoplasm.

Constructing atomistic resolution models of cellular environment, or cellular types are extremely challenging due to several factors. The atomistic structure of all cellular components is not available. Globular protein structures are either experimentally resolved or they can be modeled with reasonable accuracy through homology modelling [116]. However, membrane proteins or disordered proteins are more difficult targets to address experimentally. Likewise, nucleic acids and cellular membranes have complex structures that are difficult to resolve. For a review on this issue see [67].

Macromolecular models reach 0.1 nm spatial resolution when modeled at atomistic level. Constructing a full atomistic model of the whole cell at this resolution is conceivable, but it would be prohibitively costly to investigate the dynamics of the model with e.g., Molecular Dynamics simulations. From a simple extrapolation in time, it has been estimated that the simulation of an E. coli bacterium would be possible in 25 years from now for one nanosecond with $10^{11}$ atoms [135]. However, no cellular processes happen at the nanosecond scale.

More recently macromolecular crowding has been investigated for simple systems [42, 57, 56, 100]; see Paper III. Many of these studies on macromolecular crowding show a trend in which compact native states are favored over extended states, due to entropic effects. However, the effects of non-
specific protein-protein interactions could have an opposing outcome depending on the type of crowding agents [114]. These non-specific interactions should not be neglected in the study of the diffusive properties of macromolecules.

The attempts in systematically studying simple models for crowding has advantages that bring us to the next point in the challenges around atomistic models, namely accurate physical models. The accuracy of MD simulations depends on the force-field models. In general, the force fields have been improved in reproducing experimental observations over time [77, 85, 10]. However, even state-of-the-art additive force fields lead to aggregation of biomolecules at cellular concentration levels [105, 2, 107]. This problem has been addressed in several studies [15, 107, 97] including Paper II. For recent reviews on force-field development see [101, 111].

Therefore, more accurate physical models and increased computational power would resolve the current challenges to some extent. For the time being, approximative models are of great value for shedding some light on cellular processes. In spite of the high concentration of macromolecules and nucleic acids in the cellular environment, the major component is water molecules. Therefore, using implicit water with reduced dielectric constant to model the effect of cellular conditions would significantly reduce the computational costs. For more thorough discussions on modeling biomolecular simulations from atomistic resolution towards coarse grained models see [43].

The highest level of approximation for cellular processes is achieved through kinetic models (ODEs). These are the simplest models and therefore computational power is gained at the expense of losing details, including spatial detail. For a review on the importance of space in simulation of cellular processes see [128]. Another feature of cellular processes that is not captured with ODEs or even PDEs is the stochastic nature of certain cellular processes [92, 109]. From the other hand, one key feature of cellular processes that is captured with ODEs and PDEs is the reactions. Reactions in terms of bond-breaking processes are not modeled in MD simulations. In order to incorporate biochemical reaction into the simulations, one needs to study the problem at the quantum mechanics (QM) level or through quantum mechanics/molecular mechanics (QM/MM) [119]. However, the increase in degrees of freedom increases the computational costs yet again.

Reaction Diffusion Master Equation (RDME) and Brownian Dynamics (BD) are among formalisms that incorporate key features of cellular processes such as reactions, stochasticity, and space. Furthermore, they can simulate processes at the cellular time scale. The advantage of RDME over BD is the event-driven scheme of the algorithm which makes it faster; however BD algorithms retain spatial information more accurately. For more discussion on this regime of algorithms see [51]. Green’s Function Reaction Dynamics (GFRD) algorithm is yet another type of BD algorithm that takes advantage of an event-driven scheme to speed up the simulations [138, 137]. It is extended
to 1-D, 2-D and 3-D processes [124]. In Paper I, its efficiency is improved for diffusion-limited reactions. In spite of large computational overhead, GFRD algorithms are up to 5 orders of magnitude faster than conventional BD algorithms; however, they are still slower than the RDME type of algorithms. Nevertheless, when accurate spatial information and kinetic parameters are crucial to the investigation, the GFRD algorithm is a better candidate. For more details on the choice of parameters for stochastic simulation algorithms see [125].

Macromolecular crowding investigation with BD algorithms poses a problem which is the high number of molecular collisions. From the other hand, BD algorithms are appropriate tools for processes that take place on large temporal scales, e.g., minutes to hours. This is prohibitive with e.g., MD simulations. In these instances, a lower level models, e.g., MD can be used to parametrize diffusion coefficient and/or kinetic parameters under crowding condition. These parameters can be used for higher level models, e.g., BD algorithms to see the effect of crowding on cellular processes.
6. Svensk Sammanfattning


I denna avhandling används två beräkningsmetoder, Molecular Dynamics (MD) och Brownian Dynamics (BD) simuleringer. MD-tekniker kan användas för att studera biomolekylers dynamik och interaktioner på atomnivå genom att lösa Newton-ekvationen numeriskt. BD-tekniker löser biomolekylär dynamik och reaktioner på ett förenklat sätt, i det här fallet genom att beskriva hela proteiner som en sfär.


Vinsten i beräkningskraften försämras emellertid när det gäller simulerande av trånga system, men med tanke på effekterna av de makromolekylära n på
diffusionskoefficient och kinetiska parametrar är kända, kan man implicit integrera effekten av trängsel i BD-algoritmer genom att välja rätt parametrar. Därför är det fördelaktigt att förstå effekten av trängsel på atomnivå.

Vi studerade effekten av trängsel på diffusiva egenskaper på atomnivå med MD-simuleringar. Våra resultat i artikel III betonar effekten av kemiska interaktioner på atomnivå på makromolekylers rörlighet. En utmaning var att nya mot atomistiska fysiska modeller s.k. kraftfält behövs ta fram för simulering av makromolekyler i höga koncentrationer, då befintliga kraftfält leder till aggregering av proteiner vid hög koncentration. I artikel II undersökte vi scenarier baserade på försvagning och förstärkning av växelverkan mellan proteiner respektive mellan proteiner och vatten. I artikel IV byggde vi en cytoplasmisk modell baserad på tillgängliga data på cytoplasma från Escherichia coli, på atomnivå. Denna modell kan beräknas i såväl tid som rum med MD-simuleringspaket GROMACS. Genom denna modell är det möjligt att studera strukturella och dynamiska egenskaper i cellmiljö vid en fysiologisk relevant koncentration.
7. Acknowledgments

This thesis wouldn’t be possible without the support of so many people during the last many years.

I would like to thank my supervisors and advisors. David van der Spoel, without your support, patience and scientific skills, I would have been lost. Per Lötstedt, thank you for helping me to save three years of my work, without your help my work would have been lost. Sorin Tanase, thank you for all the tutoring. I would also like to thank you, Elisabeth Larsson, for being a great mentor to me, being there all the time and advising me patiently.

I would like to thank you Alex, for being extremely patient during all these years. I always found right directions in your neat view to the problems. I hope you recognize me not being a PhD student.

I would like to thank you, my little Sophie, without you the last stretch of time wouldn’t have been as delightful.

I would like to thank my friends and colleagues in David’s group, Madeleine, Mohammad, Leandro, Nina, Natasha and Daniel. Thanks for sharing your experiences in science and life with me.

I would like to thank my friends I met during my “residency” at BMC that made my life happy, Alexandre, Beat, Miha, Paul, Sudarsan, Livia, Rezvan, Xana, Prune, Cedric, Zhen, Hugo, Sebastian, Aleksandra, Klev, Gurkan, Petar, Klaudia, Mirtha and Meike.

Madeleine, thank you for informing me on Swedish life style and upbringing of kids. I admire your steady, calm and consistent attitude.

Silvana, thank you for all the discussions on books, movies, friendships and relationships. I admire your point of view.

Sofia, thank you for always directing me towards all positive points in life. I admire your optimism.

Katrin, thank you for all the walkings during “mammeledighet”. I admire your calm attitude as well as all your energy for biking all the time.

Jevgenia, you are my very first friend in Uppsala since I started my PhD. You’ve seen all my ups and downs, I promise from now it will be more ups than downs. I admire your thorough and deep viewpoint.

Shabnam, thank you for inspiring me with your attitude. I admire your energy.

Showgy, you’ve talked to me through the difficulties and never stopped believing in me. I am so grateful for all your time, advice and support. It means a lot to me.

And you Anja, we have shared a lot during last couple of years. I don’t remember how we started to have a lot in common and subsequently discussed
a lot. You helped me a lot in many ways, I know you know it too, without your inspiring talks I wouldn’t have made it.

I would like also to thank my friends I got to know since I came to Sweden. I wouldn’t have made the first two years in Sweden without having you here in Uppsala; Fatemeh, Amin and Ershad.

I would like to thank my family for never doubting me and always asking when I am finishing my PhD, patiently. The question I had no answer to until recently. Thank you Maman, Baba, Arnavaz, Ziba and Sajjad.
References


[62] Christian Hensen, Johannes C Hermann, Kwangho Nam, Shuhua Ma, Jiali Gao, and Hans-Dieter Höltje. A combined qm/mm approach to protein- ligand


2013.


A doctoral dissertation from the Faculty of Science and Technology, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology. (Prior to January, 2005, the series was published under the title “Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology”.)