Computer Modelling the Root Cause of Cystic Fibrosis

by

Miro Alexander Astore

A thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

School of Physics
Faculty of Science
The University of Sydney

2022

Declaration of Original contribution

of the dissertation submitted by

Miro Alexander Astore

This is to certify that to the best of my knowledge, the content of this thesis is my own work. This thesis has not been submitted for any degree or other purposes.

I certify that the intellectual content of this thesis is the product of my own work and that all the assistance received in preparing this thesis and sources have been acknowledged.

Miro Alexander Astore, Author

Date

Abstract

placeholder text

In loving memory of Madeline Jennifer Dell

"Fear cuts deeper than swords."

Arya Stark

Acknowledgments

Daniel Golestan, a wise man, once told me that to be given the opportunity to create this thesis was a gift. It was. It was a gift given to me by every friend, colleague, teacher, mentor and family member I've spent any time with. The list that follows of those to thank is not complete. If it was you'd be reading about a conversation I had with a middle aged woman in a hostel north of San Francisco, but that has little to do with Cystic Fibrosis.

To My parents raised me with not only academic rigor in mind but also a respect for aesthetics which has served me strangely well. I've never had a talent for the creative side of things compared to quantitative disciplines. But were it not for their demand for respect for the arts I'd have remained illiterate.

To Jeffry for his tutelage and patience, even across the pacific ocean. To have been your first mentee is an honor. You will go far.

To Poker, I am a better human being in every conceivable way for having known you. Your wisdom, intelligence and kindness are boundless. You have taught me an inordinate number of things. And yes, I do mean inordinate.

Nono and Nona I don't think you'll ever read this. I'm sad that you won't understand what I've done but I think you'd be proud if you did. Living in Condell park did more for me than you could know. Far from war torn Beirut or dirt poor Orria I'm sitting in a well lit office writing this with a full stomach and few worries. Sometimes this luck makes my head spin.

Thank you to Shafagh Waters for her vision, her drive and all her advice. You brought me a truly fascinating PhD project and I benefited greatly from your mentorship. Bridging the gap between cell biology and molecular physics is something that will happen more in the future and I'm lucky to have met such a driven lab to teach me to do so.

Serdar, a brilliant mind and a patient boss. Thank you for giving me the best possible experience at grad school I could have asked for. Your willingness to let me pursue self directed projects with a guided hand is a privilege during a PhD and I'm all the better for having gotten it from one of the best. I'm excited to carry some of your physical insight into biological systems to future research projects.

Maddy, I miss you every day. You couldn't have imagined what it was like to do this after losing you. I carry much of you with me and I wish I had more. I miss your intelligence, your warmth and your love.

You're all in my Loop and I hope I'm in yours in some way.

List of Publications

MA - Miro Alexander Astore
SK - Serdar Kuyucak

1. placeholdertext

Publication Authorship Attribution

In addition to the statements above, in cases where I am not the corresponding author of a published item, permission to include the published material has been granted by the corresponding author.

Miro Alexander Astore, Student — Date

As the supervisor for the candidature upon which this thesis is based, I can confirm that the authorship attribution statements above are correct.

Serdar Kuyucak, Supervisor — Date

Contents

List of Abbreviations

AMBER	**Assisted Model Building with Energy Refinement**
BAR	**Bennett-Acceptance-Ratio**
CF	**Cystic Fibrosis**
CFTR	**Cystic Fibrosis Transmembrane Conductance Regulator**
CHARMM	**Chemistry at Harvard Macromolecular Mechanics**
COM	**Centre of Mass**
CV	**Collective Variable**
FEP	**Free-Energy Perturbation**
gA	**Gramicidin A Ion Channel**
GROMACS	**GROningen MAchine for Chemical Simulations - MD program**
GROMOS	**GROningen MOlecular Simulation - MD program**
LJ	**Lenard-Jones Potential**
MBAR	**Multistate Bennett-Acceptance-Ratio**
MD	**Molecular Dynamics**
MetaD	**Meta Dynamics**
NAMD	**Nanoscale Molecular Dynamics - MD Program**
NBD	**Nucleotide Binding Domain**
NPT	**Constant number of Particles, Pressure and Temperature**
NVE	**Constant number of Particles, Volume and Energy**
NVT	**Constant number of Particles, Volume and Temperature**
OpenMM	**Open Molecular Mechanics - MD Program**
OPLS	**Optimised Potentials for Liquid Simulations**
PBC	**Periodic Boundary Condition**
PCA	**Principal Component Analysis**
PDB	**Protein Data Bank**
PMF	**Potential of Mean Force**
PME	**Particle Mesh Ewald - Long-range Electrostatics Method**
POPC	**1-palmitoyl-2-oleoyl-sn-glycero-3-phosphocholine**
POPE	**1-palmitoyl-2-oleoyl-sn-glycero-3-phosphoethanolamine**
RMSD	**Root-Mean-Square Deviation**
RC	**Reaction Coordinate**
TICA	**Time-lagged Indepenent Component Analysis**
US	**Umbrella Sampling**
VMD	**Visual Molecular Dynamics - MD Visualisation Program**
WHAM	**Weighted Histogram Analysis Method**

List of Figures

List of Tables

Chapter 1

Introduction

Whatever complexity means, most people agree that biological systems have it. -Frauenfelder and Wolynes [1]

1.1 The Physics Inside your Cells

Why can't I write down an equation will tell me how long I will live? Or how many hairs I will grow?

This might seem like an inane question but if you asked a physicist for the formula for how long it takes a radioactive material to decay or how long it will take an object to fall into a black hole they will be able to answer easily.

What makes the first set of questions so much more difficult to answer?

Our current physical theories have sufficient accuracy at the energy and length scales of biology that we can accuarately model every component inside a living being[2]. So why are living systems so hard to study?

I posit that it is the diversity of components that makes biological questions so difficult to both ask and answer. Biology distinguishes itself amongst scientific disciplines requiring the study of systems that are both complex and heterogeneous. In the study of more simple physical systems a simple analogy such as a mass on a spring or a gas of hard spheres can be extremely successful in explaining macroscopic phenomena. For biological systems there appears to be too much complexity for such analogies to have the same level of success. They may struggle to answer questions such as "If this gene mutates how will that affect lung function?" "If this drug were given at a higher dosage what would its effect be?" "What if we change this chemical moiety?" At the moment,

a trained chemist needs to go and answer these questions pipette in hand, the physicist with their notebook is hopeless.

It seems like a silly question but it seems important to ask why we can't just use a device similar to a harmonic oscillator or a perfect black body to speculate at useful answers for these quantitative questions. The answer is just as silly. If you look with your naked eye at your arm, you will notice hair, pores, dry skin, dead skin, perhaps even tendons and muscles under the all that. Under that top layer, there are two more layers to your skin with different functions and composition. If you were to take a single cell from any of those layers and stain it to distinguish features in an electron microscope you would notice all sorts of complex structures and the size and number of these structures would vary depending on where you took the cell from in the body. Within and between each those structures is a salty, wet dance of molecules large and small. This heterogeneity on length scales hints at the reasons behind biology's physical complexity. Plasma physics is often characterised by the density of the plasma studied. This parameter may span 28 orders of magnitude from a dense stellar core to the sparse intergalactic nebulae[3]. The same mathematical tools can be used to map any plasma in these energy scales. Would that we were so lucky in biology. We struggle to apply same physical models to deal with phenomena across a single order of magnitude.

Thus, in order to move towards more predictive theories of biology it is necessary to consider much more of the fundamental physical processes occurring within biological systems than simply searching for statistical trends. One form of this from fundamentals approach is the simulation of every atom in a biological system. Although computationally expensive, this approach appears necessary due to the heterogeneous nature of biological systems.

One of the things we're trying to do with molecular dynamics is fill in the gap left by the sequence-¿function paradigm which is internalised in current understandings of molecular biology. We usually talk about how the sequence of the gene defines its function because it gives the protein its structure but really there is a considerably larger amount of regulatory pressure exerted by the environment. This is what is missing from the sequence alone paradigm.

1.2 What is Physics?

Personally I have always given answers along the lines of "the study of the movement of energy within a system" or when I was in high school "The study of how things move". Although adequate for a layman these might obscure the fundamental structure within physics that make it such a powerful tool. It is the conception of some causal unit in a system and the ability to scale up the behaviour of that unit to make predictions about measurable phenomena.

This might take a few different forms at different scales, it's what makes physics feel like the most "fundamental" of the sciences.

Examples include:

Newton's laws of gravitation to explain the organisation of the solar system.

Einstein's theories employing Reimannian geometry to track the motions of galaxies and black holes.

The conception of atoms as hard spheres used to derive the macroscopic behaviour of gasses.

The schrodinger wave function to find the structure of atoms, which can then be integrated further up to find their macroscopic organisations. In chatper 2 I'll walk you through how we use these laws to model biomolecular dynamics.

The field of biophysics has an interesting origin. Erwin Schroedinger wrote an essay titled "What is Life?" This remarkable work, written in 1944 before the discovery of DNA or the maturation of information theory speculates from first principals in thermodynamics and quantum mechanics, the nature of life at the atomic level. The most remarkble thing about this essay is how much the author gets right. The observation that since organisms exist at high (?) temperatures the phyical encoding of their genome must be chemical in nature, as energy barriers lower than that of chemistry would be oblated by at physiological temperatures. Schroedinger posits the existence of what he calls an "aperiodic crystal". He was right, this crystal just happened to be one dimensional. This allegory is an example of how physcial principals can in fact be used to make testable, far reaching predictions about fundamental biology. The details may simply be more difficult since it is harder to capture the heterogeneous details inside cells.

Biological systems exhibit such a problem for the physicist because unlike the above problems it is extremely hard to pick out a fundamental unit to even begin our upwards journey. An evolutionary biologist might say to choose the "gene" but this is actually far too high in our spatial heirarchy already. Really, a gene is only meaningful to the dance of life if it has partners to dance with.

A coil of DNA in water doesn't really do much in solution except decay without machinary that can preserve, read, translate and replicate it. The gene is an emergent property, we have to go deeper.

So, what are the gene's partners?

A slew of biological machinary that mostly take the form of proteins. These proetins are a special case of chemistry, with many observable functions. Their sequence is coded by the DNA in something reminiscent of a strange loop [4].

This self referential loop is one of the reasons biology is so difficult. Since we know that this strange loop is kicked off by atomic interactions we will start there. As we are taking a physical, pragmatic approach here it would make sense to begin with the protein, after all, they stave off the march of entropy constantly trying to eat up all of your cells. It also just so happens that they are much easier to understand computationally since their motions are faster and more flexible.

The first level sub cellular organisation is perhaps the most intimdating first step for me personally after spending 4 years simulating a single protein. Glimpsing the complexity within a single one of these molecules has been one of the most existential experiences

of my life but the knowledge that there are astronomical numbers of these things inside me all of the time terrifies me.

It is hoped that illustrating the monumental task in both intellectual effort and resources of incrementally increasing the understanding of a single protein amongst the 23000 or so encoded in our genome will give the reader and understanding of how we might continue our quest to understand the molecular dance that plays within all of us.

After this things start to run away from me with my handful of GPUs and limited patience. So in this thesis we will only discuss single proteins.

1.3 Using Ion Channels as Natural Laboratories to Learn Biophysics

The physiological importane of ion channels became clear after the experiments of Hodkin and Huxley. These mathematicians took nerves from giant squid and measured the current running through the nerve in response to electrical stimulation. What they found was intruiging. Current would only flow when the input signal was of a sufficient voltage. The measurements and modelling they carried out gave an exciting set of results. They found that the cell had to maintain a constant electrochemical gradient, they discovered that the presence of voltage gated ion channels and cation selective ion channels[5]. Each of these features, motivated by mathematical modelling have been found to be critical to the functioning of the cell and fundamental to the foundation of molecular biophysics. The following set coupled ordinary differential equations were discovered by testing functions which fit the measurements taken from the squid axon.

$$I = C_m \frac{dV}{dt} + \bar{g}_K n^4 (V - V_K) + \bar{g}_{Na} m^3 h (V - V_{Na}) + \bar{g}_l (V - V_l),$$

$$\begin{aligned} \frac{dn}{dt} &= \alpha_n(V)(1-n) - \beta_n(V)n, \\ \frac{dm}{dt} &= \alpha_m(V)(1-m) - \beta_m(V)m, \\ \frac{dh}{dt} &= \alpha_h(V)(1-h) - \beta_h(V)h \end{aligned} \tag{1.1}$$

The α and $\beta \in [0, 1]$ parameters are the proportion of the sodium and potassium channel populations which are activated, respectively. This example shows how basic theoretical tools can be used to predict and discover physical phenomena in biological systems. The Hodgkin Huxley model proved the existence of a cell's resting potential, the possibility of voltage gated ion channels, and channels whose pores are selective for certain ions. Even today the molecular mechanisms behind some of these discoveries are debated. In this thesis we aim to do the same by building up from fundamental quantum mechanics in order to understand the motion of single proteins so we might speculate as to the function of the whole organism.

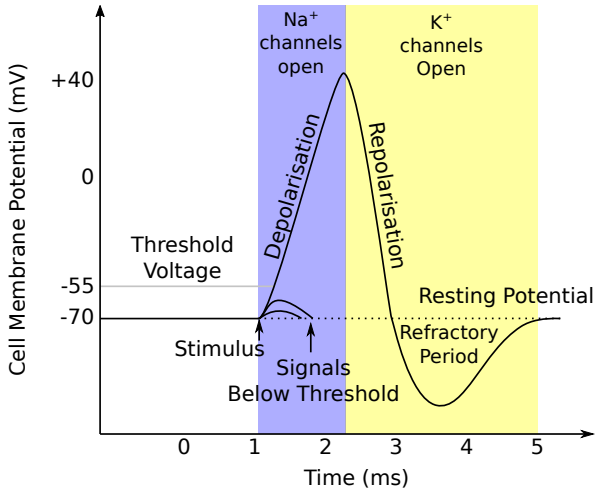

Figure 1.1: The Action Potential is a Solution to the Hodgkin-Huxley Model The shape of the action potential is a similar sight in many physiology textbooks. It was is in fact discovered as a result of the mathematical modelling of Hodgkin and Huxley hinting at the deep biophysics of ion channels won them the 1963 Noble Prize in medicine. This discovery is an excellent example of how deep theoretical insight can lead to predictable models of living systems [5][6][7][8][9][**hodgkin1952e**].

Similar to the above story, ion channels have always motivated the early pioneers of molecular biophysics. This is due to their ubiquity and importance in biological systems and the ease of measuring their activity with biochemical assays. One just needs an oscilloscope to measure their current. As cell biology has advanced it has become clear that the resting potential of a cell is critical to its function, regulating many chemical reactions inside it.

It's quite interesting that ion channels are the targets of comprise 19% of approved drugs.[10] An exciting implication of this is the finesse with which we can now manipulate these systems in order to regulate morphogenesis in addition to the internal signalling environment of cells []. CITE Mike LEVIN.

These factors have to allowed biophysicists sufficient data to build sufficiently accurate models of biomolecular systems which generalise to other systems. Leading to a thriving field, analysing systems as diverse as protocells to gold nano particles CITATIONS NEEDED.

The discovery of voltage gated channels and a resting potential are still subjects studied in cell biology today as they are critical to the cells' function [].

So by using ion channels as basic biophysical laboratories have been used consistently to understand higher level protein physics [], and now we can apply this molecular understanding to attempt to tackle other diseases as experimental techniques have improved such as diabetes, and neurodegenerative diseases. Having an atomistic understanding is allowing us to do personalised medicine with atomic precision.

1.4 Studying Cystic Fibrosis is a way to learn biophysics and Treat disease.

The sad truth of this debilitating disease is that those afflicted are extremely unlucky. A single, small change to the genome and their lungs fill with sticky mucus and become infected with bacteria, each breath cumbersome. Personally, I've not met somebody who has this disease. I have consistently wondered what perspective I'm missing by not suffering myself from such a condition or even knowing somebody with it. I'm not been trained in the ethics of studying medicine.

In this way, my motivations for studying this protein aren't solely focussed on treating disease. There is a perspective on protein evolution which states that the primary sequence of a particular gene contributes to the overall fitness of an organisms by a formula. []

It just so happens that the CFTR gene sits at the precipice of a daunting cliff in sequence space. So by taking small steps in sequence space and plunging down this cliff we can try to understand how we might push the ball back up the cliff and retain functionality.

Moreover, by learning the nuts and bolts of what goes wrong with CFTR we can start to think about where some of these cliffs might be in other places in the proteome, to gain function and avoid disease and debilitation.

The reality of disease pathogenesis being caused by so many different mutations means that there has been decades of investigation into the function of every domain in the protein.

Due to the array of disease causing mutations which occur accross the cystic fibrosis protein, there is a large body of literature on its unique function. This allows us a glance into its function and an opportunity to simultaneously perform basic biophysical research while directly assisting in furthering patient outcomes.

1.5 Well. We're in the future

Throughout science, the integration of experimental data with theoretical models leads to new and exciting research, this is particularly true in biology with its important applications in medicine, agriculture and manufacturing. Wet lab biologists take advantage of experimental techniques which allow them to understand the dynamics and structure of living things from the top down. The finer the experimental instrument, the finer the detail they may resolve. Conversely, computational and theoretical biologists take a bottom up approach, we aim to take the granular details of a system, and integrate them upwards to model the macroscopic behaviour of that system. With more powerful computers and more detailed models we can make predictions about the behaviour of more complex systems. What is so exciting about the current era of biological research is that the domains of these two approaches are beginning to overlap, where they can synergize and drive further breakthroughs. As we discover more

systems where this overlap can be found we will develop more sophisticated treatments for diseases and problems found around the world.

The reason this has happened before in physics is two fold. Physical systems are much more homogeneous. So it's much easier to integrate upwards in length scale. Once you understand the pairwise interaction between two components it's simply a question of having the theoretical and computational capacity to model the bulk behaviour of that system.

The difference with biological systems is that they have so many different components that finding an analytic or even computationally tractable solution is usually impossible. However, as we collect more data and build more powerful computers we can approach more complete models. These in turn inform more powerful theoretical models these help direct the material efforts of experimental expertise .

While previously we were limited it functional data concerning ion channels we now have unprecedented resolution for the structure dynamics for the inside of a cell. Advances in cryoEM, confocal microscopy, fluorescent microscopy and genetic engineering allow us to glimpse unprecendented information about the salty dance of life inside cells.

Alphafold is a good example. This new breakthrough builds on decades of inquiry from the structural biology community and advancements in AI to give high resolution protein structures. Now this result can be used to fill in the gaps of structural biology. Crucually, alphafold konws what it doesn't know. So we can tell where to direct the efforts of structural biology. Together these advances will fill more gaps in our knowledge of protein physics.

Chapter 2

From Protons to Proteins: Methods to simulate the inside of a cell.

Nature isn't classical, dammit, and if you want to make a simulation of nature, you'd better make it quantum mechanical, and by golly it's a wonderful problem, because it doesn't look so easy.- Richard P. Feynman

This chapter is written for somebody who has studied undergraduate physics and now wishes to model biological systems at the molecular level. Care is taken to dive deeper into the mathematical formulations of simulation methods than is conventionally given in introductory texts. Essentially, this is the understanding of simulation techniques I wish I had when I started studying them. An excellent overview which I would recommend as first reading for any new student can be found in an article by Braun et al. [11] followed by [12] for statistical rigour and [13] for free energy calculations.

2.1 Quantum Mechanics is Not Tractable at the Scale of Biology.

Living things are made of atoms and atoms themselves are composed of many particles, protons, neutrons and electrons. The motions these constituent particles are governed by quantum mechanics. Unfortunately, performing simulations for the number of atoms involved in proteins and other cellular components at quantum mechanical accuracy is impossible. Hence, we will show how to take the fundamental formulation of atomic interactions in the Schrödinger wave equation and apply approximations in order to produce a model which is capable of simulating macromolecular systems at biologically relevant timescales.

We will gradually integrate upwards, beginning with the interactions in a single atom we will work our way up to a complex macromolecular system with lipids, water, salts and of course, proteins. Ultimately this section rationalises the treatment of atoms as point charges in classical molecular dynamics simulations.

2.1.1 A full quantum mechanical treatment

Since we are dealing with atoms which are governed by quantum mechanics we must begin our journey upwards with the time dependent form of the Schrödinger wave equation.

$$i\hbar\frac{\partial}{\partial t}\Psi(\mathbf{x},t) = \big[-\frac{\hbar^2}{2m}\nabla^2 + V(\mathbf{x},t)\big]\Psi(\mathbf{x},t) \tag{2.1}$$

In quantum systems we treat all particles as waves hence the use of the wave function $\Psi(\mathbf{x},t)$. The complex amplitude of the wave function $|\Psi(\mathbf{x},t)|^2$ tells us the likelihood of detecting the particle at time t and at place $\mathbf{x}$. The term in the brackets correspond to $-\frac{\hbar^2}{2m}\nabla^2$, the kinetic energy of the particle with mass m while $V(\mathbf{x},t)$ is an externally applied potential on the system. Given that the left hand term $i\hbar\frac{\partial}{\partial t}\Psi(\mathbf{x},t)$ contains a gradient with respect to time, it governs how the wave function will evolve in time.

When the external potential V has no explicit dependence on time, this equation reduces to the familiar time independent form.

$$U\Psi(\mathbf{x},t) = \big[-\frac{\hbar^2}{2m}\nabla^2 + V(\mathbf{x})\big]\Psi(\mathbf{x},t) = H\Psi(\mathbf{x},t) \tag{2.2}$$

Here, U is an eigenvalue of the Hamiltonian operator H. Note that the wave function $\Psi(\mathbf{x},t)$ is still allowed to evolve in time.

In atomic systems there are two types of particles, nuclei which we will denote with the subscript n and electrons denoted by e. In order to treat these elements separately we decompose the Hamiltonian of the system into a few components.

$$H = \underbrace{T_n + U_{n-n}}_{H_n} + \underbrace{T_e + U_{e-e} + U_{n-e}}_{H_e} \tag{2.3}$$

Where T_n and T_e denote the kinetic energy of the nuclei and electrons respectively. While $U_{n-n}, U_{n-e}, U_{e-e}$ denote the potential energy for interactions between nuclei, between electrons and nuclei and between electrons respectively.

Since the potential terms all describe charged species, they follow coulomb's law and have the form.

$$U_{n-n} = \sum_{i>j}\frac{q_e^2 z_i z_j}{|\mathbf{R}_i - \mathbf{R}_j|}, \quad U_{n-e} = -\sum_{i,l}\frac{q_e^2 z_i}{|\mathbf{r}_l - \mathbf{R}_i|}, \quad U_{e-e} = \sum_{l>k}\frac{q_e^2}{|\mathbf{r}_l - \mathbf{r}_k|} \tag{2.4}$$

Here the z_i represent the atomic number (and thus the charge) of the ith nucleus and q_e is the unit charge of the electron. The reason for the separate coordinates R_i and r_l is to separate out the treatment of nuclei and electrons which will be important once we apply the Born-Oppenheimer approximation.

Meanwhile, the kinetic energy terms are of the form

$$T_n = -\sum_i \frac{\hbar^2}{2M_i}\nabla_i^2, \quad T_e = -\sum_l \frac{\hbar^2}{2m_e}\nabla_l^2 \tag{2.5}$$

M_i represents the mass of the ith nucleon and m_e represents the mass of an electron. The operator $\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}$. The separate subscripts i and l are due to the different coordinates which we use to denote the positions of the nuclei and the electrons. The reason for this will become clear when we derive the Born-Oppenheimer approximation to separate the wave functions and treat them separately.

2.1.2 The Born-Oppenheimer approximation.

In order to reach the Born-Oppenheimer approximation we start with the observation that electrons have a mass 3-4 orders of magnitude smaller than the nuclei. This motivates two simplifications. The "clamped nuclei assumption" where we solve the Schrödinger equation whilst nuclei are fixed in space and do not move. And a related assumption known as the "adiabatic assumption" which postulates that the electrons will respond instantaneously to any changes in the positions of the nuclei. Combining these physical approximations we derive the "Born-Oppenheimer approximation" for the Schrödinger equation which can be used to simplify calculations involving several atoms at once.

We begin the derivation by examining the time-independent form of the electronic Schrödinger wave equation where the nuclei are fixed at positions R_i.

$$H_e(\mathbf{r}_l, \mathbf{R}_i)\psi_e(\mathbf{r}_l, \mathbf{R}_i) = U_e(\mathbf{R}_i)\psi_e(\mathbf{r}_l, \mathbf{R}_i) \tag{2.6}$$

Fixing the nuclei in this way gives the "clamped nuclei" approximation [14]. To solve the wave function for the whole system Ψ_{tot} we use an *ansatz* which decomposes the wave function with an electronic basis into two components: $(\psi_e)_k$ and $(\psi_n)_k$ which are the kth eigenfunction solutions to H_e and H_n respectively.

$$\Psi_{tot}(\mathbf{r}_l, \mathbf{R}_i, t) = \sum_{k=0}^{\infty} \psi_e(\mathbf{r}_l, \mathbf{R}_i)_k \ \psi_n(\mathbf{R}_i)_k \tag{2.7}$$

Note that there is an implied direct product between the wave functions $\psi_e(\mathbf{r}_l, \mathbf{R}_i)$ and $\psi_n(\mathbf{R}_i)$. When we substitute this expression into the full Schrödinger equation 2.1 we find the following expression for the kth nuclear eigenfunction [15]

$$i\hbar\frac{\partial}{\partial t}\psi_n(\mathbf{R}_i)_k = \Big[-\sum_i \frac{\hbar^2}{2M_i}\nabla_i^2 + U_e(\mathbf{R}_i)_k \Big]\psi_n(\mathbf{R}_i)_k + \sum_j C_{kj} \ \psi_n(\mathbf{R}_i)_j \tag{2.8}$$

Where we have coupled the electronic wave functions to each other with the operator

$$C_{kj} = \int (\psi_e)_k^* \Big[\sum_i \frac{\hbar^2}{2M_i} \nabla_i^2 \Big] (\psi_e)_j d\mathbf{r} + \frac{1}{M_i} \sum_i \Big[\int (\psi_e)_k^* [-\hbar i \nabla_i] (\psi_e)_j d\mathbf{r} \Big] [-\hbar i \nabla_i] \quad (2.9)$$

Using the "adiabatic assumption" [15] the off-diagonal terms of C_{kj} can be set to 0 as they represent the interactions between the electrons and the nuclei. This completely decouples the wave function into two components

$$\Psi_{tot}(\mathbf{r}_l, \mathbf{R}_i, t) = \psi_e(\mathbf{r}_l, \mathbf{R}_i)_k \; \psi_n(\mathbf{R}_i, t)_k \quad (2.10)$$

A further approximation is justified by ignoring the diagonal terms C_{kk} as they are 4 orders of magnitude smaller than the other terms in equation 2.8 [14].

We now write the Born-Oppenheimer approximated wave equation for an atomic system.

$$i\hbar \frac{\partial}{\partial t} \psi_n(\mathbf{R}_i)_k = \big[- \sum_i \frac{\hbar^2}{2M_i} \nabla_i^2 + U_e(\mathbf{R}_i)_k \big] \psi_n(\mathbf{R_i})_k \quad (2.11)$$

By rearranging this equation and taking derivative we can see how to use Newton's equations of motion to calculate the forces on the nuclei from the surrounding electric potential

$$M_i \ddot{\mathbf{R}}_i(t) = -\nabla_i U_e(\mathbf{R_i}) \quad (2.12)$$

By choosing an appropriate time-step one can simply iteratively solve this equation of motion to understand the dynamics of an atomic system. The nuclei will move according to their relative positions to each other and the electron clouds will rearrange in response to that motion. There is no need to explicitly treat the electrons at all. This is sufficient accuracy to simulate the low energy motions of molecules such as the environment found in biological systems.

2.2 Classical MD, Molecular Motions Without Quantum Mechanics

The Born-Oppenheimer approximation gives rise to Hartree-Fock methods and density functional theory (DFT). These more sophisticated physical methods allow us to simulate the organisation of electron clouds around small molecules, finding broad applications in chemistry and materials science [16]. These methods are known as *ab initio* MD.

However, even with these approximations simulating a large number of atoms is still not computationally tractable. State of the art DFT methods can only simulate on the order of 10^3 atoms [17] and scales as $O(N^3)$ [18]. This is not sufficient to simulate

proteins and their surrounding solvation environment where the molecular system is usually on the order of $10^4 - 10^6$ atoms. So, we must use another round of approximations to reach the spatial and time scales necessary to simulate biological molecules. We do this by creating a set of mathematical functions to simplify the calculations further. Here we use a set of virtual springs and other simple models for the energetic interactions between atoms. This creates what's known as an effective potential.

The CHARMM effective potential employed in this work is similar to those found in all common all-atom molecular dynamics forcefields. The same functional forms are used in other forcefields such as AMBER, GROMOS and OPLS but with different parameters and design philosophies[19].

This formulation gives us classical molecular dynamics, sometimes referred to as molecular mechanics (MM). The aim of the classical forcefields discussed here is to use *ab initio* MD as an initial target for approximation and then refine the model to better match certain experimental quantities. This is discussed in detail in section 2.2.1.

We split up the molecular mechanics potential into several components dealing with the energies from covalent bonds, including bond stretching, twisting and bending as well as contributions associated with the forces that atoms exert on each other when they are not bonded together

$$U_{MM} = \underbrace{U_{LJ} + U_{coulomb}}_{U_{non-bonded}} + \underbrace{U_{bonds} + U_{angles} + U_{dihedrals} + U_{impropers}}_{U_{bonded}} \tag{2.13}$$

Interestingly, the bonded terms may all reasonably be approximated by harmonic springs

$$\begin{aligned} U_{bonded} = &\sum_{bonds} k_b(b - b_0)^2 + \sum_{angles} k_\theta(\theta - \theta_0)^2 + \sum_{Urey-Bradley} k_u(r_{UB} - r_{UB_0})^2 \\ &+ \sum_{dihedrals} k_\varphi(1 + \cos(n\varphi - \delta)) + \sum_{improper-dihedrals} k_\phi(\phi - \phi_0)^2. \end{aligned} \tag{2.14}$$

Here, the k_i terms correspond to the strength of the harmonic restraint for a parameter. The 0 subscript denotes the equilibrium position for that parameter. Even though this formulation is quite simple, it has empirically been shown to be a reasonable approximation for the potential energy functions of quantum mechanics in covalently bonded chemical species. Examples can be seen in figure 2.2.

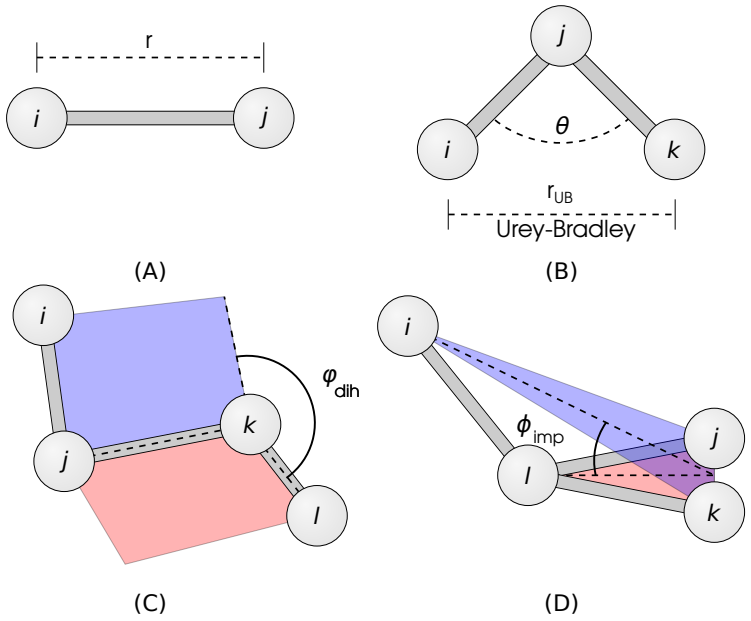

Figure 2.1: The Bonded Interactions Calculated In Classical Forcefields. (A) The energy of Bond Stretching is approximated as a harmonic oscillator with respect to their separation r. (B) Angles between neighbouring covalently bonded atoms are also approximated as a harmonic oscillator with respect to the angle θ. In some forcefields such as CHARMM there is a correction term for these angular interactions known as Urey Bradley forces. This is calculated using the separation between the non-bonded atoms i-k in the triplet with the parameter r_{UB}. (C) The dihedral angle between four atoms is calculated by constructing two planes. Each plane is constructed to contains three of the four atoms in the set. One plane encompasses atoms i, j and k here colored in blue and the other plane contains the j, k and l atoms colored in red. The dihedral angle is then calculated by taking the angle between these two planes along the line they intersect, the line formed by the j-k bond. (D) The improper dihedral angles enforce the planarity of a molecular configuration. A plane is constructed to contain the i, j and k (blue) atoms and another plane is constructed to contain the j, k and l atoms (red). The improper angle is then calculated as the angle between these two planes.

Non Bonded Interactions

The term $U_{non-bonded}$ captures interactions which arise when atoms are not covalently bound to each other. Namely, coulomb forces due to electric charges on the atom, attractive Van Der Walls interactions and repulsion due to Pauli Exclusion,

$$U_{non-bonded} = \underbrace{\sum_{i>j} \epsilon_{ij} \left(\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right)}_{U_{Lennard-Jones}} - \underbrace{\sum_{i>j} \frac{q_i q_j}{r_{ij}}}_{U_{coulomb}} . \tag{2.15}$$

Note how the repulsive Pauli Exclusion and attractive dispersion forces have been combined into one term known as the Lennard-Jones potential or U_{LJ}. The σ parameter denotes the location of the local minima in the Lennard-Jones potential. This is the

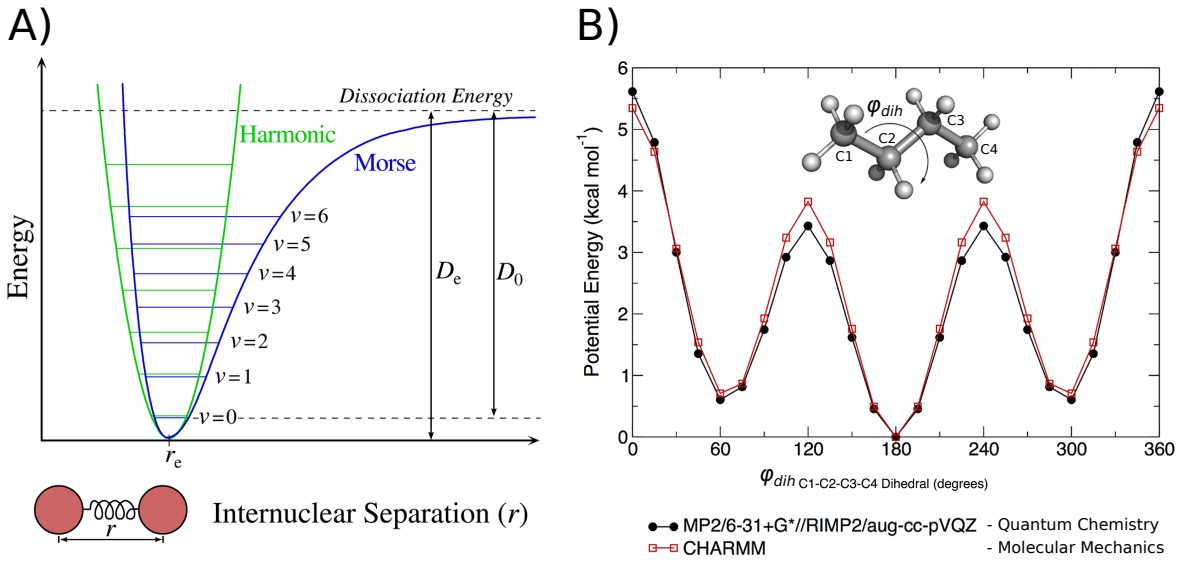

Figure 2.2: Comparison Between Potentials in Quantum and Classical Forcefields
A) The Morse potential was formulated to approximate the potential the potential energy surface associated with the stretching of covalent bonds (blue). At low temperatures (the ground state, $v = 0$) like those found in classical MD there is good agreement between the Morse potential and the harmonic oscillator (green). Credit Mark Somoza 2006 B) Here the potential of the dihedral angle between the atoms C1,C2,C3 and C4 in a butane molecule is calculated using two methods: Quantum Chemical calculations and approximations using the functional form in 2.14 [19]. Note how the appropriate choice of k_φ, n and δ have closely approximated the results the more accurate quantum mechanical calculations.

optimum distance that two atoms will rest against each other in the absence of other effects. The ϵ parameter denotes the depth of the potential well, or how stable the two atoms will be in the minimum energy configuration. This is very important for certain physical parameters such as osmotic pressure [20].

Meanwhile, the partial assignments q_i each atom are very important in a biological context, for stability of protein conformations of salt bridges and the solvation energy of different molecules[21].

By focussing on adjusting the charges of an atom to fit the solvation energy of a molecule and adjusting the Lennard-Jones parameters to fit the osmotic pressure measurements, we can these two physical parameters we can isolate and improve the non-bonded parameters.

2.2.1 Philosophy of Different Molecular Mechanics forcefields.

At the time of writing, the four popular forcefields for the simulation of biomolecules are CHARMM, AMBER, GROMOS and OPLS. Each of these have a slightly different philosophy in their formulation. They may be bottom up, as in the case of AMBER

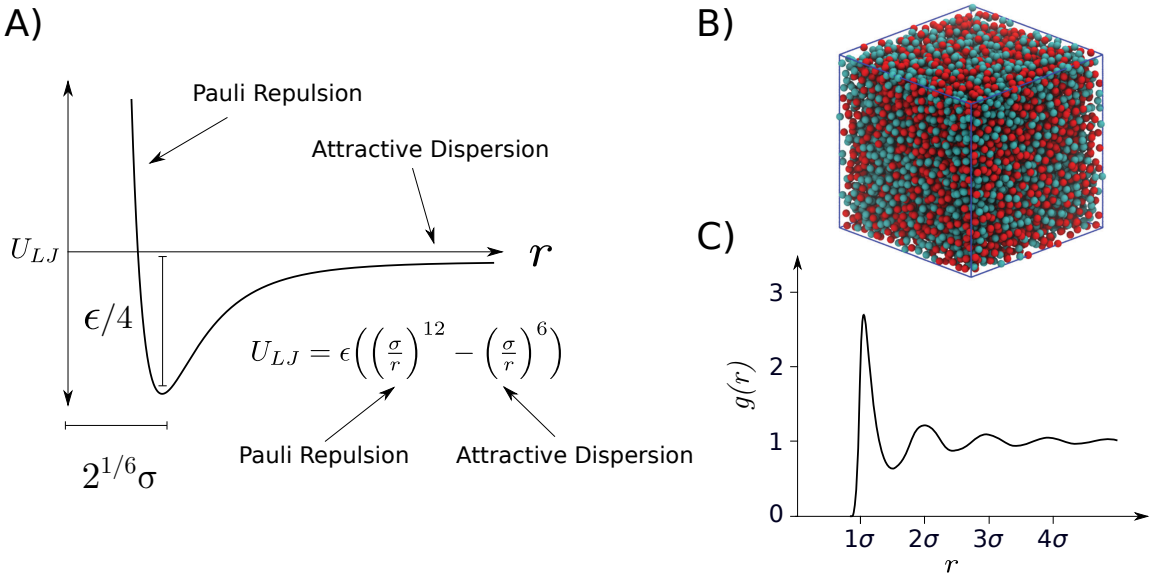

Figure 2.3: The Lennard-Jones Potential
A) The Lennard-Jones potential function has two regimes, the far region one dominated by attractive dispersion forces and the close region dominated by repulsion. In the case of atomic systems this is due to the Pauli exclusion principal. B) An example of a fluid modelled with Lennard-Jones particles [22]. C) The radial distribution function (g) for a Lennard-Jones fluid [23]. Note that the peaks in the distribution are spaced roughly 1 σ apart.

and CHARMM or top down, in the case of OPLS. Bottom up forcefields take the results from *ab initio* MD calculations as an initial guess and approximate them by tweaking the parameters in the functional form in equation 2.13. Conversely, top down forcefields tweak these parameters with reference to experimental measurable. Ultimately, the results from *in silico* experiments must match those of wet lab experiments so the development of all forcefields has elements of both philosophies. All forcefields assign the partial charges to atoms using the results of *ab initio* MD calculations. The rest of the parameters parameters are derived with other methods such as attempting to match the known secondary structure of peptides[24]. Different forcefields have different methods of deriving their parameters. Below is a short summary of the philosophy for the 4 major forcefields taken from the review by Justin Lemkul [19].

- CHARMM: The most popular all atom forcefield. To build CHARMM, QM optimized geometries and molecular dipole moments are compared to those found found from calssical MD simulations. Molecular degrees of freedom such as dihedral angles are also fit with QM energy profiles, an example can be seen in 2.2. Macroscopic experimental quantities are also used to validate the parameters in this forcefield, such as solvation energies, crystal geometries, heats of vaporization and conformational sampling of biomolecules[24]. One of the reasons for this forcefield's popularity is its considerably library of supported compounds, especially when combined with its generalisation module CGENFF [25].
- AMBER: An all atom forcefield which is built from the ground up from results

from quantum mechanical calculations and results from spectroscopy. A version forcefield has become the favorite for the simulation of disordered proteins [26][27].

There is also a generalised version of the AMBER forcefield known as GAFF [28][29]. Comparisons between generalised forcefields can be found in [30].

- OPLS: An all atom forcefield. The OPLS forcefield takes the philosophy that, since many biomolecules share similar geometries to certain organic liquids, biomolecules can be accurately parameterised by creating a forcefield which correctly reproduces the experimental measurements for these species. Parameters are derived to accurately reproduce the liquid density of certain organic liquids. These parameters are then used as roots to construct larger biomolecules by drawing analogies between similar molecular geometries.
- GROMOS: A united atom forcefield, where hydrogen atoms are typically merged into the heavy atom they are bound to. Hence, they are not explicitly treated. Charge assignment is done with DFT. Interestingly, GROMOS uses a quartic form of the bond stretching term

$$U_b = \frac{1}{4}k_b(b^2 - b_0^2)^2 \tag{2.16}$$

The parameters are adjusted for agreement with experimental parameters such as solvation energies, liquid densities. GROMOS forcefields are popular in some contexts because of its ability to reproduce partition coefficients between polar and non-polar media, a similar chemical context to what is found at the interface between a membrane and bulk water.

2.3 Periodic Boundaries to Simulate the Inside of a Cell

Inside cells, proteins are immersed in a large solvation environment composed of water and salts [31]. An example can be seen in figure 2.4. However, simulating such a large environment is computationally expensive and truncating it with vacuum at the boundaries leads to water molecules aligning their dipole moments along the boundary and perturbing equilibrium dynamics. In order to avoid artefacts we have to replicate the large cellular environment somehow[32]. We could make a simulation box large enough to replicate the behavior of a bulk solvent, but even with a large simulation box we can still observe artifacts associated with the vacuum at the boundaries [12]. So, to avoid these boundary effects we use periodic boundary conditions (PBCs), allowing atoms to move between images in the simulation box. This replicates the molecular system infinitely in every direction.

Using PBCs might remove vacuum from our molecular system but now we have a different problem. Effectively, with the PBCs, we have created a system with an infinite number of atoms. We have to somehow limit the number of computations we perform. We could simply truncate the calculation of interactions $U_{non-bonded}$ after a

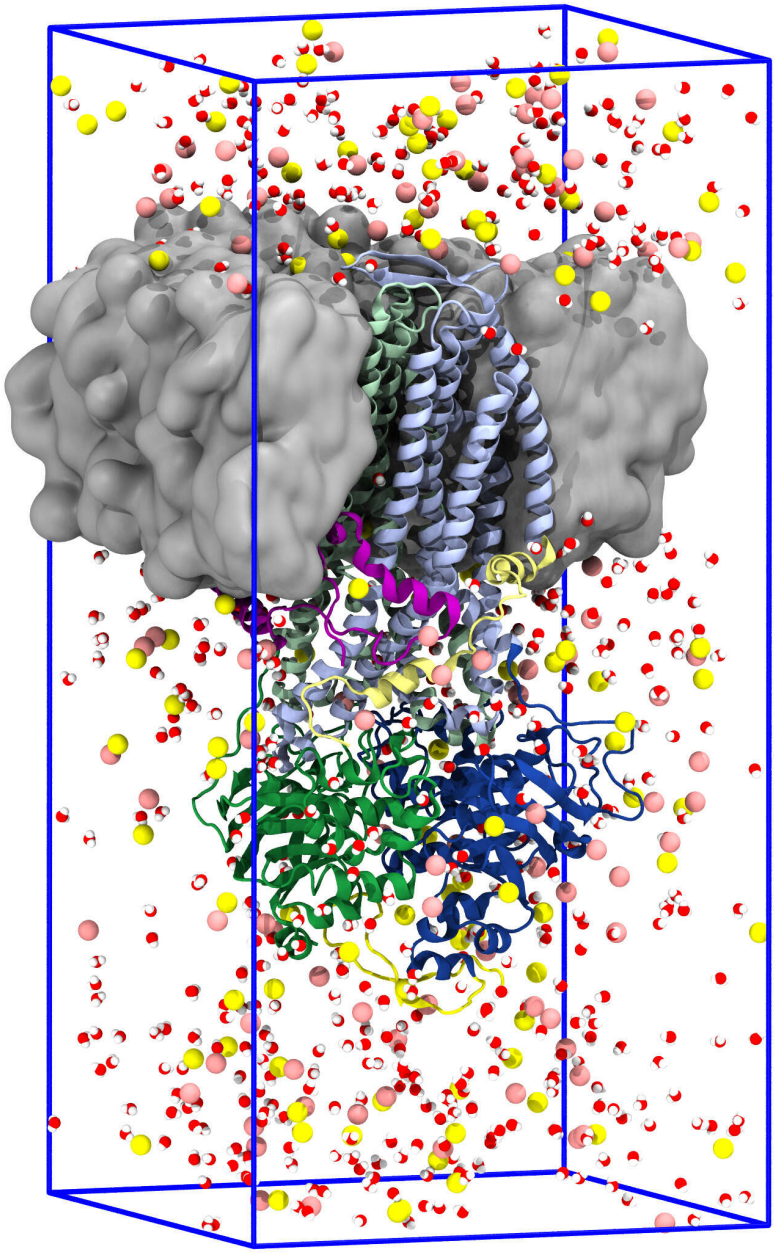

Figure 2.4: An Example of a Solvated Biomolecular Environment Ready for Simulation

This rendering shows a CFTR protein embedded in a lipid bilayer, immersed in a potassium chloride solvent. Half of the bilayer has been stripped away, as has much of the solvent so the protein is visible. The phospholipid lipid bilayer is colored grey, while the potassium chloride ions are colored pink and yellow respectively. Water molecules are red and white. The blue box indicates the boundaries of the unit cell. This system contains roughly 190 000 atoms in total.

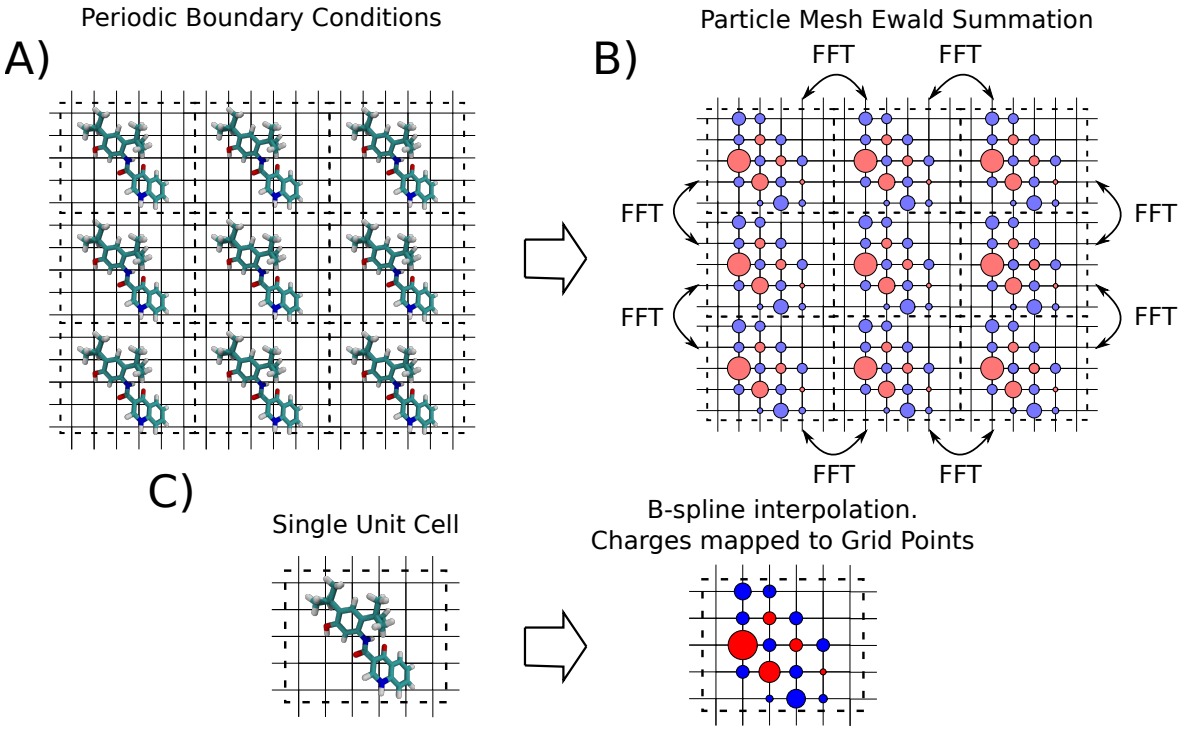

Figure 2.5: Particle Mesh Ewald Summation
A)The molecular system is repeated infinitely along all axes, when atoms reach the edge of the simulation box they are allowed wrap around to the other side of the box. B) The charges in the infinite periodic system are approximated onto a regular grid. Then the potential in the infinite system is calculated via a Fast Fourier Transform (FFT). C) A more detailed view of the charge mapping procedure. The point charges in the system are interpolated onto the grid using B-spline interpolation.

certain cutoff distance. This is not an issue for U_{LJ} because the $1/r^6$ and $1/r^{12}$ terms in equation 2.3 decay very quickly for large r. Inaccuracies due to this approximation can be further ameliorated with the use of a smooth switching function [33][34]. On the other hand, the $1/r$ dependence in $U_{coulomb}$ scales much more slowly so truncating it leads to a loss of accuracy in the results of the simulation.[35][36][37][38][39]. So, we have to calculate the contributions of $U_{coulomb}$ in between all periodic images. Note that this periodicity requires that the unit cell is electrically neutral, else the contribution of potential energy from $U_{coulomb}$ will be infinite, leading to artifacts [40].

To calculate $U_{coulomb}$ in all periodic images and limit computational intensively of our calculations we use a clever scheme known as Particle Mesh Ewald summation (PME). Interestingly, this scheme ends up scaling better than the pairwise summation in equation 2.15 might imply. The direct summation scales with computational complexity of $O(N^2)$ with the number of atoms while the infinite PME scheme scales as $O(N \log N)$ [41], though there are some further considerations for large systems on parallel architectures [42]. Even with these sophisticated algorithms, the calculation of electrostatic potential in the infinite system still represents the largest computational bottle-neck in classical MD [42].

For a detailed review of different Particle Mesh Ewald Summation methods and the

mathematics behind the method see [43]. A brief outline of the Smooth Particle Mesh Ewald summation is given below

1. An *ansatz* is used where $U_{coulomb}$ has Gaussian screening charges added to it and simultaneously subtracted away to create a smooth potential. The details can be found in [43].

$$U_{coulomb} = U_{screening-charges} + U_{coulomb} - U_{screening-charges} \tag{2.17}$$

   Terms in these equations are then rearranged such that one term is evaluated with a Fourier transform and one term is evaluated using a direct sum.

$$U_{coulomb} = U_{FT} + U_{direct-sum} \tag{2.18}$$

2. The charges to be evaluated using U_{FT} are interpolated onto a grid a grid using B-spline interpolation functions. This is the procedure demonstrated in figure 2.5.
3. The charge density functions for the charges on the grid are transformed into frequency space using Fast Fourier Transforms.
4. The Poisson equation is solved numerically in frequency space for these charges.

$$\nabla^2 \tilde{U} = 4\pi\tilde{\rho}(\mathbf{k}) \tag{2.19}$$

   Where $\tilde{U}$ is the component of U_{FT} we solve for in frequency space and $\tilde{\rho}$ is the Fourier transform of the smooth scalar function for the interpolated charge densities.
5. An inverse Fourier transform is calculated to for the solution to $\tilde{U}$ to transform to the Poisson equation back into real space.
6. The interactions in $U_{direct-sum}$ are evaluated using a simple pairwise summation.
7. Now that $U_{coulomb}$ is known at every position in the unit cell we can move atoms according to the contributions from this potential using Newton's second law.

2.4 Controlling the Temperature and Pressure in a Simulation

Living things are very sensitive to their external environment. Enzymes only work in a narrow range of temperatures and cells burst apart in the absence of pressure[44][45]. As such, to correctly understand biological systems we not only need to simulate the dynamics of the atoms inside them, but we must also make sure that the virtual environment in our simulations matches what is found inside cells or in the laboratory. Our simulations should seek to approximate the environment of an open topped test-tube sitting in a pressure and temperature controlled laboratory. To do this, we make use of some statistical ensembles chosen for their performance in regulating the thermodynamic quantities in a simulation and their computational expense.

2.4.1 Hot and Cold with the Nosé-Hoover Thermostat

Recall that the temperature of a system is a direct function of the velocity of its constituent particles. So by regulating the ensemble of velocities we can control the temperature. We begin a simulation by choosing the velocities of the atoms within the system from a Maxwell-Boltzmann distribution.

$$f(v_i) = \left(\frac{m_i}{2\pi k_B T}\right)^{3/2} \exp\left(-\frac{m_i v_i^2}{2k_B T}\right) \tag{2.20}$$

Where $f(v_i)$ is the proportion of particles with velocity v_i, k_B is Boltzmann's constant. Note that $i = 1, ..., N_{df} = 3N$ as we choose a velocity component for x, y and z separately.

Despite starting from the same Cartesian coordinates, randomly sampling velocities from the Maxwell-Boltzmann means that replicate simulations will immediately begin from different points in phase space. Their coordinates will quickly diverge, raising questions around how long one should run a simulation and how many replicates they should run in order to collect reliable statistics. According to Knapp et al.[46], a good rule of thumb is to simulate between 5 and 10 replicates depending on the availability of computing resources[1]. In this thesis we prioritised long time scales to sample slow motions, so 3 replicates with runtimes between 1 and 2 microseconds were produced for all systems.

After the initial choice of velocities, the temperature in a simulation is maintained by directly modulating the velocities of the atoms to maintain the target temperature T_0. There are many schemes which attempt this. We will discuss the Nosé-Hoover thermostat in detail because it was during the production runs in this thesis[47][48][49]. However, for the equilibration phase of the simulation we used the Berendsen thermostat because it is faster at correcting large temperature differentials but does not produce the correct statistical ensemble [50][51]. We also note that the field has since moved on to favor the Bussi thermostat which is an extension of the Berendsen thermostat as it works well in most contexts [50][11].

The Nosé-Hoover thermostat is characterised by the use of an extra, massive particle coupled to an external bath. The use of a single bath has been associated with issues with ergodicity and so usually this particle is coupled to a chain of external baths. Usually, the software simulation package GROMACS uses a chain of 10 baths, $M = 10$[49][52][53]. The Hamiltonian for the Molecular Dynamics System Coupled to a chain of M external baths is then

[1] Recall that in the ergodic limit, a quantity f will be equal when calculated in the ensemble average, between replicates, as when equal to when calculated using the time average, in one replicate.

$$\langle f \rangle_i = \lim_{t \to \infty} f$$

$$H_{NH}(\mathbf{x}, \mathbf{p}, \eta_1, ..., \eta_M, p_{\eta_1}, ..., p_{\eta_M}) = H_{MM} + \sum_j^M \frac{p_{\eta_j}^2}{2Q_j} + k_B T N_{df} \eta_1 + k_B T \sum_{j=2}^M \eta_j \quad (2.21)$$

Here, usually $N_{df} := 3N$ unless there are constraints placed within the system to freeze atoms. η denotes the 1 dimensional coordinate of the thermostat particle with mass Q, while H_{MM} is the Hamiltonian of the unregulated molecular mechanics system

$$H_{MM}(\mathbf{x}, \mathbf{p}) = \underbrace{\sum_i^N \frac{\mathbf{p}_i^2}{2m_i}}_{E_{kinetic}} + U_{MM}(\mathbf{x}). \quad (2.22)$$

By using Hamilton's equations of motion, H_{NH} evolves by

$$\begin{aligned}
\dot{\mathbf{x}}_i &= \mathbf{p}_i / m_i \\
\dot{\mathbf{p}}_i &= \mathbf{F}_i - \mathbf{p}_i \frac{p_{\eta_1}}{Q_1} \\
\dot{\eta_j} &= p_{\eta_j} / Q_j \\
\dot{p_{\eta_1}} &= \left[\sum_i^N \frac{\mathbf{p}_i^2}{m_i} - N_{df} k_B T \right] - p_{\eta_1} \frac{p_{\eta_2}}{Q_2} \\
&\vdots \\
\dot{p_{\eta_j}} &= \left[\frac{p_{\eta_{j-1}}^2}{Q_{j-1}} - k_b T \right] - p_{\eta_j} \frac{p_{\eta_{j+1}}}{Q_{j+1}} \\
&\vdots \\
\dot{p_{\eta_M}} &= \left[\frac{p_{\eta_{M-1}}^2}{Q_{M-1}} - k_b T \right]
\end{aligned} \quad (2.23)$$

Where $\mathbf{F}_i$ is the force vector on the ith particle. It may be calculated from U_{MM} using Newton's second law.

The parameters Q_j are chosen by the user to control the coupling strength of th baths to each other. We usually choose

$$Q_j = \frac{\tau_{NH} T_0}{4\pi^2} \qquad \forall j \quad (2.24)$$

where τ_{NH} is the time interval between when the thermostat parameters are updated. This means that whenever the simulation is not at a time step that is a multiple of τ_{NH} we can just evaluate U_{MM} as normal but every interval of τ_{NH} we rescale the velocities according to the equations of motion in 2.23 to match the correct temperature T.

Remember that we can always calculate the temperature using the instantaneous velocities in the simulation using

$$T = \frac{2E_{kinetic}}{3Nk_B} = \frac{\sum_i m_i v_i^2}{3Nk_B} \tag{2.25}$$

The Nosé-Hoover thermostat, when chained infinitely, allows us to use accurately produce what's known as an NVT ensemble, also called the canonical ensemble in the statistical mechanics literature [49]. Where the number of particles in the system (N) remains constant, the volume of the system remains constant (V) and the temperature remains constant (T). In a realistic environment, the pressure also remains constant, so we need another regulatory mechanism to modulate the volume of the system to regulate the pressure (P) and produce an NPT ensemble.

2.4.2 Under Pressure with the Parinello-Rahman Barostat

Pressure is critical to the function of living organisms. Membranes burst apart at low pressures[54] and at high pressures cellular function is disrupted[55]. In order to accurately reflect the atmospheric pressure at which living things thrive we have to accurately calculate and modulate it during our simulation.

In order to measure the pressure at the simulation walls we follow the procedure in [56] by calculating a quantity known as the virial:

$$W(\mathbf{x}) = \sum_{i}^{N-1} \sum_{j>i}^{N} \mathbf{r}_{ij} \cdot \mathbf{F}_{ij}. \tag{2.26}$$

Where $\mathbf{r}_{ij} = \mathbf{x}_i - \mathbf{x}_j$ is the Cartesian distance between the ith and jth atoms, while $\mathbf{F}_{ij}$ is the force extorted on atom j by atom i.

This is then substituted into the equation

$$P(\mathbf{x}) = \frac{Nk_BT + \langle W \rangle_i}{V} \tag{2.27}$$

And so using equation 2.27 we can see that we can modulate the volume V of the simulation in order to control the pressure throughout the simulation.

To modulate the pressure we apply the Parrinello-Rahman barostat [57][58], with a procedure with a similar philosophy to the extended Hamiltonian used in the Nosé-Hoover thermostat. In this way the system is coupled to an external pressure bath rather than an external temperature bath. First we define that the basis vectors for the periodic simulation box to be $\underline{h} := [\mathbf{a}, \mathbf{b}, \mathbf{c}]$. When the box is scaled to change the volume these basis vectors are multiplied by a set of scalars $s_i := (\xi_i,\ \eta_i, \zeta_i) \in [0, 1]$. We perform a change of coordinates so that the contributions of the particles onto the boundaries is easily calculated from our equations so we express the atomic coordinates as

$$\begin{aligned} \mathbf{x}_i &= \xi_i \mathbf{a} + \eta_i \mathbf{b} + \zeta_i \mathbf{c} \\ &= \underline{h}\mathbf{s}_i \end{aligned} \tag{2.28}$$

Defining $\underline{G} := \underline{h}^T \underline{h}$. The Lagrangian for the scaling system then becomes

$$L = \frac{1}{2}\sum_{i}^{N} m_i \dot{\mathbf{s}}_i^T \underline{G} \dot{\mathbf{s}}_i - \sum_i \sum_{j>i} \phi(\mathbf{r_{ij}}) + \frac{1}{2} M \mathrm{Tr}(\dot{\underline{h}}^T \dot{\underline{h}}) - P_{ext} V \tag{2.29}$$

Where $\phi(\mathbf{r}_{ij}$ is the pairwise potential between two atoms in U_{MM}, while M is a constant of proportionality associated with the kinetic energy derived from the movement the particles undergo as they scale. It has units of mass. P_{ext} is our target, externally applied pressure. This Lagrangian allows us to derive the equations of motion

$$\begin{aligned} \ddot{\mathbf{s}}_i &= -\sum_{j\neq i} \frac{1}{m_i \mathbf{r}_{ij}} \frac{d\phi(\mathbf{r}_{ij})}{dr_{ij}} (\mathbf{s}_i - \mathbf{s}_j) - G^{-1}\dot{G}\dot{\mathbf{s}}_i \\ \ddot{\mathbf{h}} &= \frac{1}{M}(\mathbf{Y} - P_{ext})\underline{\sigma} \end{aligned} \tag{2.30}$$

The matrix $\underline{\sigma} := V(\mathbf{h}^T)^{-1} = V[\mathbf{b}\times\mathbf{c}, \mathbf{c}\times\mathbf{a}, \mathbf{a}\times\mathbf{b}]$ contains information about the size and orientation of the simulation box, while

$$\mathbf{Y} = \frac{1}{V}\sum_i m_i (\underline{h}\dot{\mathbf{s}}_i)(\underline{h}\dot{\mathbf{s}}_i)^T + \sum_i \sum_{j>i} \frac{1}{r_{ij}} \frac{d\phi(\mathbf{r}_{ij})}{dr_{ij}} \mathbf{r}_{ij}\mathbf{r}_{ij}^T \tag{2.31}$$

represents the stress tensor which acts across each of the faces of the unit cell.

This system of equations can be solved numerically to control the pressure of the simulation system by modulating the length of the basis vectors $\mathbf{a}, \mathbf{b}$ and $\mathbf{c}$ contained in $\underline{h}$.

Together with the Nosé-Hoover thermostat, the Parrinello-Rahman barostat produces NPT, also called the isothermal-isobaric ensemble in the statistical mechanics literature. The combination of these two methods is thus appropriate for simulating a cellular environment.

2.5 The Process of Preparing an MD Simulation

The process of taking a molecular structure and putting it in a cellular environment to simulate it at physiological temperatures is both an art and a science. It's a science because a biophysicist must be aware of the many tricks that structural biologists use to image a macromolecular complex. But it's an art because accounting for those tricks and making the necessary modifications is rarely straight forward. How do you build a missing loop? What charge state is an amino acid most likely to take during a physiological context. These questions must be carefully answered by analysing the literature about the system of interest. Once the protein structure has been built, the system is immersed in a water solvation bath alongside a concentration of salt ions, usually sodium chloride if the environment is thought at be extra cellular or potassium chloride if the environment is thought to be intracellular. Once the initial conditions have been decided, we need to make a few preparatory steps so the simulation collects

reasonable results and doesn't hurtle along some unphysical trajectory. The steps so that the simulation remains realistic are fleshed out in some more detail in [11] but we produce a short summary below.

1. Minimisation: Here the atoms are moved down along U_{MM} to resolve any clashes in the system which would cause LINCS or SHAKE to diverge. This is usually done with simple minimistation algorithms such as steepest descent or conjugate gradient descent[59]. This is usually done with a constant volume.
2. Relaxation: Harmonic restraints are placed on the heavy atoms (non hydrogen) in the system so that large conformational changes do not occur while the macromolecules are heated and settles into their solvation environment. This may be done under the NVT or NPT ensembles depending on the system. Sometimes the system is heated slowly from 0 Kelvin up to the desired thermal temperature in order to avoid large conformational changes which might result from different parts of the system heating at different rates.
3. Equilibration: Often after relaxation more simulation time is run so that the system can settle into local minima father. This process makes sure that the physically relevant local minima are being sampled once we move to production. This process could be run for a few nanoseconds or up to a microsecond. It depends on the system. This is usually done with the NPT ensemble.
4. Production: Here the NPT ensemble is applied and the system is allowed to evolve under Newton's equations of motion while data is collected for analysis.

2.6 Choosing an Appropriate Time Step

The discrete time step, Δt which is used to integrate our equations of motion is one of the most important determinants in the performance of the simulation. We would like Δt to be as large as possible, so that the minimum number of calculations are made to sample the desired time scale. In the case of proteins this usually runs between 10^{-6} and 10^{-3}s[26].

As you can see in table 2.1 the fastest motion in molecular systems is dictated by stretching of covalent bonds. Studies of the resonance of molecules by infrared spectroscopy determined that the O-H type bonds oscillate the fastest, with a resonance peak at 3600cm^{-1}[60].

Due to Nyquist's theorem the largest Δt parameter we can choose *must* be less than half the speed of the fastest degree of freedom in the system [61]. However, empirically we have found that condensed matter systems require even shorter time steps to maintain their stability [59]. The Verlet leap-frog scheme used in most MD codes requires between 5 and 10 integration steps per period of the fastest harmonic mode in a system, to maintain stability [62][63]. The choice of too large a time step means that the system will escape local free energy minima, accumulating kinetic energy and eventually "blow-up" [11]. In the case of biomolecular systems we are challenged by the fact that they are so hydrogen-rich. Since hydrogen is so light, its motion is much faster compared to the other molecular motions involving heavier, slower moving atoms. Its

Motion	Timescale
Covalent Bond-stretching	$1 - 2 \times 10^{-15}$s
Covalent Bond-angle bending	$5 - 10 \times 10^{-15}$s
Sidechain Motions	$10^{-12} - 10^{-6}$s
Rigid Body Motions	$10^{-9} - 1$s
Ion Conduction	$10^{-9} - 10^{-6}$s
Protein Conformational Changes	$10^{-9} - 10^{-3}$s
Alpha Helix Formation	$10^{-9} - 10^{-6}$s
Beta Sheet Formation	$10^{-6} - 10^{-3}$s
Protein Folding	$10^{-6} - 10$ s

Table 2.1: Timescales of Motions in a Molecular System
The time step of a simulation must be small enough to capture the motions in the fastest degree of freedom. In hydrogen-rich biomolecular systems the bottle neck can be found in the fast bond vibrations in lighter atoms. This stands in tension with the phenomena we are interested in on longer timescales such as protein folding. Sources: [59][60][68][69][70] [63]

correlation time is on the order of 1 femtosecond, in classical simulations we are able to get away with using 2 femtoseconds with the use of specialised integration schemes such as SHAKE[64] and LINCS[65] to constrain the fast motion of hydrogen atoms. Allowing us to use $\Delta t = 2$fs in during atomistic classical MD simulations.

The use of techniques such as hydrogen mass repartitioning [66], virtual site topologies [63] and multiple time step schemes[67] have also gained popularity in recent years in order to increase time steps further, up to $\Delta t = 5$fs.

2.6.1 Verlet Leap-Frog Integration

To produce molecular trajectories we can use the potential U_{MM} which we calculated with equation 2.13 and calculate the forces exerted on the atoms in the system. By Newton's 2nd law have

$$\mathbf{a}(\mathbf{x})_i = \frac{d^2\mathbf{x}_i(t)}{dt^2} = -\frac{1}{M_i}\nabla_i U_{MM}(\mathbf{x}_i). \tag{2.32}$$

We can use this calculation of acceleration a_i of the ith atom to update the positions and velocities of the atoms in the molecular system with the following triplet of equations known as the leap-frog Verlet method [60]:

$$\begin{aligned} \mathbf{v}_i^{n+1/2} &= \mathbf{v}_i^{n-1/2} + \Delta t\ \mathbf{a}_i^n \\ \mathbf{x}_i^{n+1} &= \mathbf{x}_i^n + \Delta t\ \mathbf{v}_i^{n+1/2} \\ \mathbf{v}_i^{n+1} &= \mathbf{v}_i^{n+1/2} + \frac{\Delta t}{2}\mathbf{a}_i^{n+1} \end{aligned} \tag{2.33}$$

Note that $v_i^{n-1/2}$ will have been calculated during the previous time step and a_i^{n+1} may be calculated by the updated positions found by calculating $\mathbf{x}_i^{n+1}$.

In MD, we are less concerned with the accuracy of a particular trajectory so much as collecting sufficient statistics to calculate macroscopic properties such as free energies or diffusion profiles. This means the choice 4th order solvers such as the Runge-Kutta method would be inappropriate. Although they may use a large time step they require 4 evaluations of U_{MM} per iteration and are thus expensive more expensive than any second order method. Hence, we prefer symplectic (energy preserving), 2nd order methods such as Verlet integration so the simulation remains stable after millions of time steps[67].

2.7 Free Energy Calculations: Making Simulations More Useful

The above work sets out how to perform what is known as unbiased MD simulations. These are powerful tools but as will be discussed in section The Problem with Sampling if one only relies on unbiased simulations they will quickly exceed the available computer power. Imagine there is an event our system undergoes that we know from experimental evidence our system must exist but is slow. Examples of this include the passage of an ion through a channel and the binding of a drug. We *could* calculate the Gibbs free energy of a given molecular configuration $\mathbf{x}_0$ using

$$G(\mathbf{x}_0) = -\frac{1}{k_B T} \ln(P_u(\mathbf{x}_0)). \tag{2.34}$$

Where $P^u(\mathbf{x}_0)$ represents the probability of obtaining state $\mathbf{x}_0$, estimated from an unbiased simulation. From here on a subscript of u indicates an quantity obtained from an unbiased simulation and a subscript b represents a quantity from a biased simulation.

Equation 2.34 shows how there is exponentially poor sampling in regions with high U_{MM}. So it is clear that we will not collect sufficient statistics for a good estimate with available computer power. So, we must be clever in how we direct our available resources. This means intelligently sampling sections of the molecular phase space which are of interest to us physically, but are not reached in our unbiased simulations. A technique that is used extensively throughout this thesis is the addition of a biased potential to the molecular potential U_{MM} calculated for the purposes of unbiased simulations. This will drive the simulation to regions of interest.

$$U'_{MM} = U_{MM} + U_{bias}(\xi) \tag{2.35}$$

Note how the U_{biased} term is explicitly dependent on a parameter ξ. This parameter is known by many names, an order parameter, a collective variable or a reaction coordinate. Each of these names has its origin in a different subfield but they all refer to the progress toward a target state. This could be a phase transition from a liquid to a gas, the progress of a chemical reaction or more likely in our case, the distance toward a target molecular configuration.

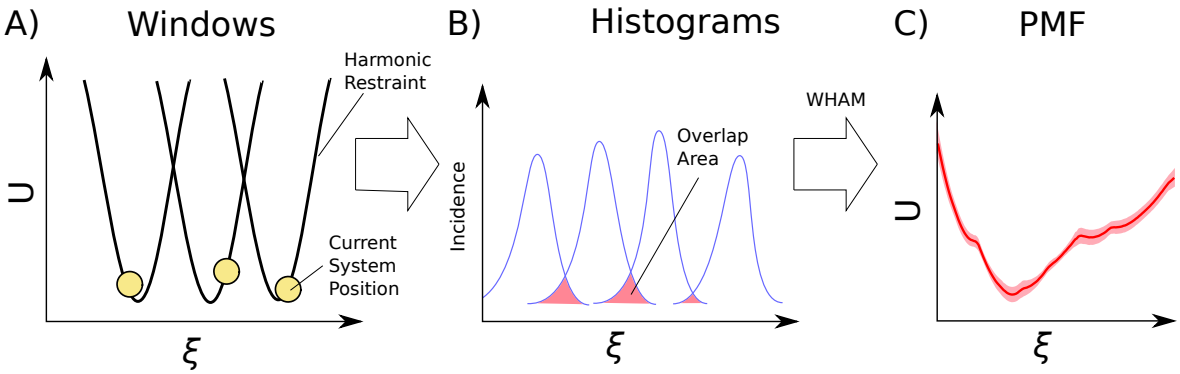

Figure 2.6: Illustration of Umbrella Sampling
A) Several simulations are repeated with only one change. A bias potential is added somewhere along the reaction coordinate ξ. B) The value of ξ is recorded in each of the windows and then graphed as histograms. C) The Overlap in neighbouring histograms is integrated via the WHAM method to calculate the Potential of Mean Force. This gives us the energy landscape. Fluctuations in the overlap in the data can be used to estimate the error for the PMF.

The functional form of U_{bias} depends on the Free energy technique being employed. There are two varieties of techniques, equilibrium and non-equilibrium methods. We will focus on the equilibrium methods in this work. We note that there is another set of methods called alchemical methods which modify the chemical composition of the system which we will not cover.

2.7.1 Umbrella Sampling

This is possibly the most popular method of calculating the potential of a system along a reaction coordinate. The conceptual philosophy for the method is demonstrated in figure 2.6. The molecular system is replicated in several "windows" and a harmonic U_{biased} is added at several points along the collective variable ξ. Statistics are then collected in order to calculate the potential of mean force (PMF) and thus the energy landscape along between those windows.

The functional form of U_{bias} in umbrella sampling is then separated into N windows. With the nth window having the biasing function:

$$U_b^n = \frac{k_\xi^n}{2}(\xi(\mathbf{x}) - \xi_0^n)^2 \tag{2.36}$$

Where ξ_0^n is the equilibrium position of the restraint. k_ξ^n is the strength of the harmonic restraint in the nth window. Typically this is the same in all windows. The more overlap between adjacent windows the more those windows are attracted to each other and the steeper the gradient of the free energy surface (FES) must be pushing those windows together. Conversely, when there is less overlap between adjacent windows it indicates the presence of a barrier in the energy landscape between those windows.

Umbrella sampling is extremely useful for calculating all sorts of experimental quanti-

ties and physiologically relevant properties such as folding energies [71], lipid binding [72], ion conduction [73], and drug binding [74]. However, it is particularly sensitive to the choice of initial configuration and collective variable [72]. The former issue is particular to umbrella sampling because generally short runs are used since so many windows are spawned during the method, the simulations must be at equilibrium *before* the method is attempted, and then sufficient statistics must be collected to average over any conformational changes orthogonal to the collective variable. In these ways, care must be taken when using this method to not introduce systematic error into the calculation [75]. A deep knowledge of the molecular system under investigation can help alleviate some of these issues.

Weighted Histogram Average Method (WHAM)

There a few candidates for calculating the PMF using the statistics calculated in umbrella sampling. The Weighted Histogram Average Method (WHAM)[76], Umbrella Integration (UI)[77] and the Multistate Bennett acceptance ratio (MBAR)[78] are all used. We will briefly outline the mathematical formulation and estimation of errors of the WHAM method as it is more popular[79]. Our explanation follows [80] which covers the topic in more detail.

The method begins by dividing the sampled region into a set of K histograms with K being greater than the number of biased windows N. The whole PMF can be estimated (poorly) from the ith biased window using

$$P_u^i(\xi) = P_b^i(\xi)\exp(\beta U_b^i(\xi))\langle\exp(-\beta U_b^i(\xi))\rangle. \tag{2.37}$$

Where the $P(\xi)$ functions represent the probability density function calculated from samples collected at point ξ. The samples collected in each histogram then gives us an estimate of the PMF according to equation 2.34. The estimates from each histogram are then are combined in a weighted sum using

$$P_u(\xi) = \sum_i^K p_i(\xi)P_u^i(\xi) \tag{2.38}$$

where the weights, $\sum_{i=0}^N p_j = 1$ are chosen such that the statistical error is minimised accross the domain of ξ[81]. This means we solve an optimisation problem for the variance of the unbiased estimates

$$\frac{\partial \mathrm{Var}(P_u)}{\partial p_i} \tag{2.39}$$

to give the best estimate of the PMF given the samples that have been in each $P_b^i(\xi)$.

Essentially we are vertically shifting the estimates obtained in each of the histograms in order to minimise the error across the PMF. By convention, we can estimate the error in the PMF by splitting up the statistics collected for the distributions of $P_i^b(\xi)$

and constructing PMFs from these independent samples. For example, we might have collected 100ns of data in each window, but we could construct 5 independent estimates for the PMF using 20ns blocks of trajectories. The Standard Error of the Mean (SEM) can then be used to estimate the error across the surface [12]. By convention, an umbrella sampling calculation is said to have converged when the SEM from these independent samples has fallen below 1kcal/mol across the PMF. Note that there may be sources of systematic error not captured by this criterion.

2.7.2 Metadynamics

Metadynamics has proven to be a popular method in many applications of computational science, not just in molecular dynamics of biological systems [83][84][85]. The method relies on a time dependent form of U_{bias} given by

$$U_{bias}(\xi, t) = \sum_{\substack{t'=\tau_D, 2\tau_D, \ldots \\ t'<t}} B \ \exp\left(- \frac{(\xi(t) - \xi(t'))^2}{2\sigma^2} \right) \tag{2.40}$$

This means the simulation is dropping virtual, repulsive Gaussian potentials at positions along ξ in order to encourage the simulation to sample regions of ξ it has not visited already. The process is illustrated in figure 2.7. The thermodynamic assumption of this method is that the deposition is done slow enough that the system remains at equilibrium, so a small Gaussian height should be chosen. Usually, the Gaussian widths σ are chosen to be the size of the variance in the unbiased measurements of ξ.

The FES estimate at time t from this method is simply the sum of all the Gaussians we have added into the system inverted:

$$U_u(\xi, t) = \frac{1}{t_c - t} \int_{t_c}^{t} U_b(\xi, t) dt \tag{2.41}$$

Formally, convergence is reached when the observed probability density is uniform across ξ. That is:

$$P_b(\xi) = \frac{1}{V_\xi} \tag{2.42}$$

where V_ξ is the volume of the phase space spanned by ξ. However, in practice the function $U_{bias}(\xi)$ is simply inspected at intervals for fluctuations about an average function [86].

There are many flavours of metadynamics. The most popular is well-tempered metadynamics which gradually reduces the Gaussian height B as the simulation progresses[87]. In theory, this guarantees convergence with vanishingly small error. However, this method requires an estimate of how long the simulation will take to converge. There is also infrequent metadynamics which can be used to estimate the diffusion profile along ξ [88][89][90].

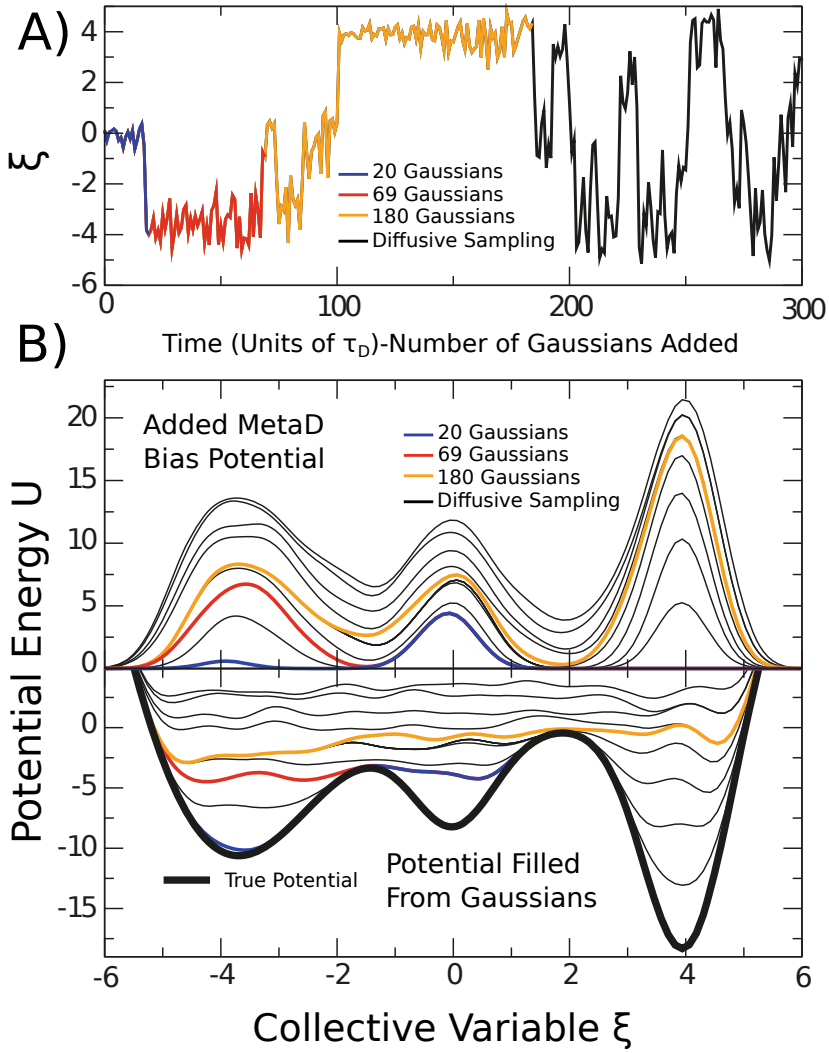

Figure 2.7: Illustration of Metadynamics
A) The trajectory of a collective variable ξ in a metadynamics simulation. The blue region denotes the trajectory up to the time where 20 Gaussians have been deposited. After which the basin at $\xi = 0$ has been filled. The red data points denote the time up until 69 Gaussians have been deposited, after which time the second basin at $\xi = -4$ has been adequately sampled. Finally the Orange data indicates that the 3rd basin at $\xi = 4$ has been adequately sampled after which which time the system begins to sample ξ diffusively and the simulation has converged. B) The upper pannel demonstrates the repulsive potentials that have been added to the simulation in order to drive it to new, unexplored regions. Multiplying the values of this function by -1 will be the estimate of the free energy surface. Note how in the lower panel the true potential the true potential energy surface is gradually filled by the added Gaussians functions. Source[82]

A useful feature of the Metadynamics method is that it can be linearly sped up with the number of parallel simulations. This is known as multiple walker metadynamics[91]. Since we are simply attempting to sample from the same potential U_{MM} up to a desired time, until the criterion in equation 2.42 is met, we can speed this process up by running simulations in parallel and adding Gaussian to the same U_{bias} function.

Conceptually, metadynamics performs the same role as umbrella sampling and should in theory produce the same results in the same system with the same collective variable ξ. However, it has some specific contexts where it outperforms umbrella sampling. This method is more suited to an exploration of free energy space where it can intelligently explore regions which are poorly sampled, whereas umbrella sampling requires some foreknowledge of the surface being investigated in order to guess which parts of the landscape require more sampling. However, metadynamics can be very difficult converge. A good indicator of such systematic errors are the presence of unphysically large barriers, indicating that there are orthogonal degrees of freedom not sampled along ξ which might correspond to a minimum energy pathway. These barriers will eventually come down in the infinite sampling limit but many barrier crossings will need to be observed. A discussion of how to solve these systematic errors can be found in [85]

2.8 Short Comings of Classical MD

The short comings of classical molecular dynamics fall into two classes. First, These is the accuracy of the chemical forcefields outlined in section 2.2.1. Second, there is the inability of modern computers to deliver enough samples of the energy landscape to collect sufficient statistics to make rigorous conclusions. The issue is that, as the physical formulation in section 2.1.2 might indicate, the more accurate the forcefields, the more computationally our expensive calculations become. And so the solutions to these two problems are diametrically opposed. In the this short section we will explore the current efforts to find solutions to both problems.

The Problem with Forcefields

The approximations inherent in equation 2.13 are not without a cost to accuracy. In certain situations, many of which are biologically relevant, it has been shown that quantum effects such as polarisation play an important role in the dynamics of the system. This has been demonstrated in the literature for Gramicidin where polarisable forcefields are able to more accurately reproduce the experimental measurements of current[92].

The other context where polarisation is important to consider involve divalent ions such as calcium or magnesium. Here, the highly charged environment near a divalent ion will induce changes in the dipole moment of surrounding atoms. This is not possible in the fixed charge model in 2.13, making investigations of these biologically important chemical species difficult[93][94].

However, for most situations, particularly those involving bulk water with a low concentration of solute, classical forcefields are sufficiently accurate to sample conformational motions of biomolecules[95]. Sadly, it should also be kept in mind that classical MD

is not able to simulate any chemistry such as forming and breaking or a change in the protonation state of an amino acid. Such interactions require considerations of quantum mechanics which are computationally expensive[96].

There are several efforts to address some of the above issues. Some groups are trying to improve the accuracy of classical forcefields using machine learning and Bayesian inference[97]. But there are also attempts to move beyond the functional form of equation 2.13 by explicitly including the effects of polarisation, the most popular methods at the moment being adding a massless drude oscillator as an extra bead to atoms as in the CHARMM drude forcefields, championed by the Mackerell lab[98] and the use of forcefields such as AMOEBA which explicitly calculate the dipole and quadrupole moments of each atom[99]. These both substantially increase computational cost but have displayed much better agreement with experiments in biological systems where classical forcefields have been shown to fail [92][100][101].

Ultimately, the functional form in equation 2.13 used by classical forcefields does not have sufficient degrees of freedom to address all possible chemical contexts. Careful consideration must always be given to whether the forcefield is being used in a faithful way to the situations it was intended to accurately represent. So long as the user is aware of the situations where a given forcefield falls short, classical forcefields can be a powerful tool for the study of molecular systems.

The Problem with Sampling

Collecting sufficient statistics about the system of interest is often computationally infeasible. Even though computers have sped up exponentially for the last 50 years we are still orders of magnitude from being able to reach the time scales of many biological processes, as displayed in table 2.1.

The slow time step demanded in classical MD due to the fast motions of certain atomic groups such as hydrogen is fundamentally at odds with the time scales of many important biological processes such as drug binding or protein folding which occur on the time scale of milliseconds or seconds. One possible solution to this issue is to completely rework the formulation of the chemical forcefield, removing fastest degrees of freedom entirely with course grained models which unite several atoms into a single bead such as MARTINI or Gō at substantial cost to accuracy[102][103][104].

Methods are now emerging which intelligently drive the simulation toward regions unexplored in the collective variable space by unbiased simulations. For some time the field has used steered methods or adaptive sampling methods such as Umbrella Sampling or Metadynamics to drive the simulation toward sections of the energy landscape which are under sampled. These methods universally rely on a choice of collective variable ξ which corresponds to a slow degree of freedom. Such a choice is not usually simple. In the case of ion channels one may rationally choose the placement of the ion along the conduction pathway as the collective variable but the choice is less obvious in the case of more global conformational changes.

The success of simulations at the millisecond timescale by D.E Shaw research suggest that we are in reach of an exciting area in biological research [105]. Enhanced sampling

methods will be able to routinely reach motions that occur on these time scales and as software and hardware improve we will be able to push further for larger systems. This indicates that the enhanced sampling approach holds great promise.

The advances we are seeing at the moment which I find exciting are the use of machine learning methods to tease out these degrees of freedom in order to accelerate them with already established free energy methods. This could be done with a variety of schemes such as TICA[106][107], VACs[108] or RAVE[109]. These have the potential to build on the above rigorous physics of simulations and revolutionise our understanding of biomolecular systems.

2.9 Conclusion

It is hoped that the preceding chapter can serve as a roadmap for any physicists interested in beginning to study the exciting field of computational biophysics. The information in each section should serve to help the reader understand the foundations of how simulations are performed. The next steps would be to learn more about the molecular biology and biochemistry of macromolecules so they can better understand the wet, messy dance occurring inside cells. For recommended reading on this topic there is Schlick [60], Frauenfelder [110] and Phillips [111]. There is a lot to learn and the barrier to entry can seem daunting. The reader is encouraged to join mailing lists and other forums where the thriving computational chemistry and computational biophysics communities communicate. Send cold emails asking for help when you are stuck, remember to consult software manuals as well. Below is a list of software to get you started building and running simulations.

- Python[112]. A versatile programming language that will help with all parts of a computational workflow.
- Visualised Molecular Dynamics (VMD)[113]. Molecular visualisation software with a scripting language, great for building simulation systems, viewing trajectories and rendering figures.
- CHARMM-GUI[114]. A web server which offers a great starting point for constructing MD simulation systems and also offers configuration scripts for all the major MD software packages.
- MDANALYSIS[115][116]. A python package with a diverse set of tools to analyse molecular dynamics trajectories and structures.
- MODELLER[117][118][119]. A python package which helps build homology models of protein structures or add missing parts to protein models.
- NAMD[120]. A powerful, scalable user friendly MD simulation package.
- GROMACS[53]. A less user-friendly MD simulation package than NAMD but offers some different features and (at the time of writing) faster performance.
- OPENMM[121]. An MD simulation package that is optimised for usage on GPUs. Is controlled with a python API. A good choice for desktops and workstations.

- PLUMED[122]. A plugin for existing MD simulation packages which lets the user define their own collective variable ξ. Very useful for sophisticated analysis and free energy calculations.
- Protein Data Bank (PDB). An open source database of all published biomolecular structures at atomic resolution.
- Uniprot[123]. A database containing a wealth of biochemical data and annotations for proteins, very handy when starting a new project.
- ALPHAFOLD2[124]. A highly accurate AI generated database of proteins. This was considered a watershed moment in the field when it was released.
- CHARMM forcefield[24]. The parameter files for the charmm forcefield formatted to be read by popular MD simulation packages.
- AMBER forcefield[24]. An MD simulation and instructions for how to extract and use the latest AMBER forcefield in other MD simulation packages.
- PARMED[125]. A python package for converting and manipulating MD file types.

No individual who has studied a specific discipline has the skills necessary to pick up biomolecular simulation software and begin using it. Physicists lack the understanding of the biology and the chemistry involved in the biological systems, while chemists and biologists will lack an understanding of the deep mathematics that has gone into producing highly accurate simulations of molecular systems. It will take time to get used to an interdisciplinary way of thinking. The reader is also encouraged to seek out members of other wet lab disciplines such as cell biologists and protein biochemists and learn what the important problems are in biology.

Chapter 3

Review of the Molecular Cause of Cystic Fibrosis and Its Treatment

Because of what's inside me; Because of my genes;-Bob Flanagan, "Why."

3.1 Clinical outcomes of Cystic Fibrosis

Cystic Fibrosis (CF) is the most common fatal genetic condition in Caucasian populations. 165 000 people are afflicted globally. Even with decades of research there is no known cure for CF. With the average life expectancy of patients falling below 50 even in countries with developed health care systems such as the USA and Australia[][]. The cause is from a build up of salts inside epithelial cells. This causes the surface of the epithelium to dehydrate. When dehydrated the cilia on the epithelium collapse leaving them unable to clear the mucus that naturally lines the airway[126]. The dehydration mentioned earlier causes the mucus to thicken. This buildup has two pathogenic functions. Firstly it inhibits the normal function of the organ, as mucus fills ducts that would normally pass nutrients in the pancreas or absorb gasses in the lungs. Secondly, the stationary mucus allows bacterial infection, this can further degrade lung function and remains one of the most troublesome chronic complications in CF patients.

Much of the clinical research into CF has been managing the movement of this mucus and the populations of bacterium in it. Patients often require to hours of physical therapy to help clear this mucus since their lungs are unable to. They must also inhale saline solutions in order to counteract the osmotic pressure in their epithelium. This helps draw more moisture out of the epithelial cells to allow the cilia to move.

CF patients struggle to intake nutrients due to the build up of mucus in their pancreas and large intestines. This leads to CF related diabetes which afflicts roughly half of adults with CF [127]. Patients with CF related diabetes are often administered enzymes and must adhere to a specific diet. A strict diet is particularly important when a patient is taking CFTR modulators because many compounds found in food have interactions with these drugs [].

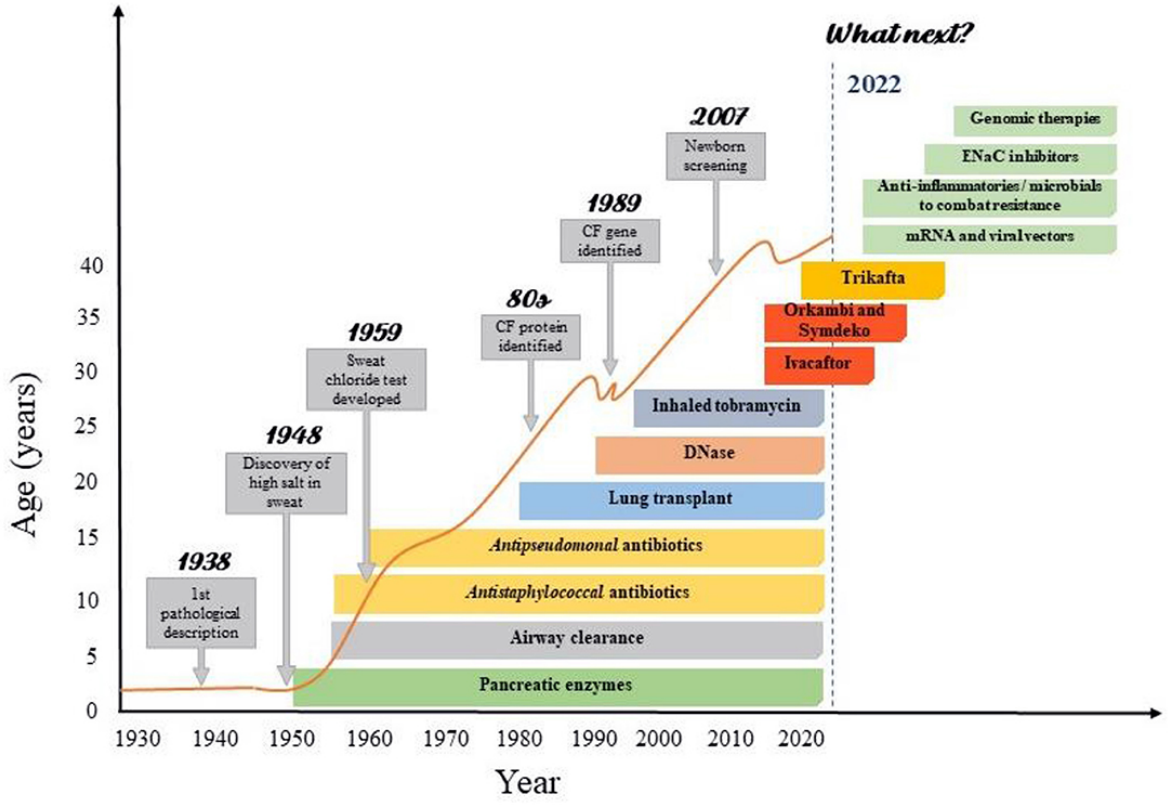

Figure 3.1: CF Clinical Progress
Life expectancy of CF patients correlates highly with translational research. Source [**garcia2022**]