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**
Glt_{Ph}	**Glutamate Transporter - *Pyrococcus horikoshii***
$GROMACS$	**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**
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**
TI	**Thermodynamic Integration**
$TICA$	**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

1.1 Physics in a test tube

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?

I posit that it is the diversity of components that makes biological questions so difficult to 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 skin. If you take a microscope you will notice the 3 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. 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 photon

The schrodinger wave function to find the structure of atoms, which can then be integrated further up to find their macroscopic organisations. More on this later.

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. Genes of hard spheres ?

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 creates the gene?

A slew of biological machinary that mostly take the form of proteins. These proetins are then coded for by the DNA in a strange loop.

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 nubmers of these things inside me all of the time

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

This makes sense if we think about it Somewhere on the scale between a single protein and a single cell this is what we consider "life". We have single unicellular organisms but we don't have uniproteomic organisms. So the fundamental length scale of life is somewhere between $10^{-10}m$ and $10^{-3}m$. This is the first loop in our strange loop.

After this things start to run away from me with my handful of GPUs and limited patience. Once we move from prokaryote to eukaryotes we have gone a few levels deeper. There is of course unicellular eukaryotes but how did we get from P to E? I'll have to leave that one for evolutionary cell biologists. Certainly there is something strangely loopy about the appropriation of cells by other cells. Then we have something more interesting, cellular collectivisation.

Cells clump together and act in unison to give us colonial organisms. (Self-similar colony morphogenesis by gram-negative rods as the experimental model of fractal growth by a cell population). Like any advanced economy cells .

1.3 Ion Channels: Natures laboratories to Learn Biophysics

1.4 Studying Cystic Fibrosis to Learn Biophysics

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, making every 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.

Ion Channels and Their Role in Developing Biophysics

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.

Beyond this simple equipment, due to the importance of the polarisation inside a cell there is a thriving field of electrophysiology with specialised techniques and equipment to measure and model the function of ion channels.

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.

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. 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 solve more problems and learn more about the universe.

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 get more and more data and more and more computer power we can approach more complete models. These in turn inform more powerful theoretical models these help direct the material efforts of experimental expertise .

Chapter 2

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

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. The motions of atoms and their 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. It is hoped that this section can be of use to both biologists and physicists, in order to teach the physicist what they need to know about the models they will be using to perform these simulations (and the many technical problems they will encounter) and to inform the biologist what the physicist is doing with all that computer time.

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 the potential energy of 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.

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

Note that the wave function $\Psi(\mathbf{x}, t)$ is still allowed to evolve in time.

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

$$H = T_n + T_e + V_{n-n} + V_{n_e} + V_{e-e} \tag{2.3}$$

Where T_n and T_e denote the kinetic energy of the nucleus and electrons respectively. While $V_{n-n}, V_{n-e}, V_{e-e}$ 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.

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

Here the z_i represent the charge 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 quite simple.

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

M_i represents the mass of the ith nucleon and m_l represents the mass of the lth electron. 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 apply the Born-Oppenheimer approximation to separate the wave functions and solve them separately.

Here, the M_i are the masses of the nuclei and the operator $\nabla = \frac{\partial}{\partial x} + \frac{\partial}{\partial y} + \frac{\partial}{\partial z}$

2.1.2 The Born-Oppenheimer approximation.

We now make use of the Born-Oppenheimer approximation [1]. This is motivated by the observation that electrons are 3-4 orders of magnitude lighter than the nucleus, and so we can assume that the electrons will respond instantaneously to any changes in the wave function of the nucleus. Thus, we can disregard T_e, U_{n-e} and U_{e-e}

This allows us to split the total wave function into two parts using a direct product. One term deals with the nuclei and one with the electrons in the system.

$$\Psi(R_i, t) = \psi_e(r_l, R_i)\psi_n(R_i, t) \tag{2.6}$$

2.2 Classical MD, Molecular Motions Without Quantum Mechanics

The Born-Oppenheimer approximation gives rise to Hartree-Fock methods which allow calculations of the organisation of electron clouds around small molecules. This lets us derive the energy profile of certain degrees of freedom within the molecule such as the energetics of stretching out a bond or twisting a dihedral angle.

However, even with these approximations simulating a large number of atoms is not computationally tractable. 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 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. So named because it effectively approximates the behaviour of the full quantum mechanical system.

This formulation gives us classical molecular dynamics where we try to match calculations made with the Born-Oppenheimer approximation.

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

We split up the molecular potential into several components dealing with the energies from covalent bonds, including bond stretching, twisting and pinching. As well as energies associated with the forces that atoms exert on each other when they are not bonded together. Namely and Coulomb forces due to electric charges on the atom and attractive Van Der Walls interactions and repulsion due to Pauli Exclusion the latter two forces are combined into one term we will analyse in detail U_{LJ}.

$$U_{CHARMM} = \underbrace{U_{LJ} + U_{Coulomb}}_{U_{non-bonded}} + \underbrace{U_{bonds} + U_{angles} + U_{dihedrals} + U_{impropers}}_{U_{bonded}+U_{Urey-Bradley}} \tag{2.7}$$

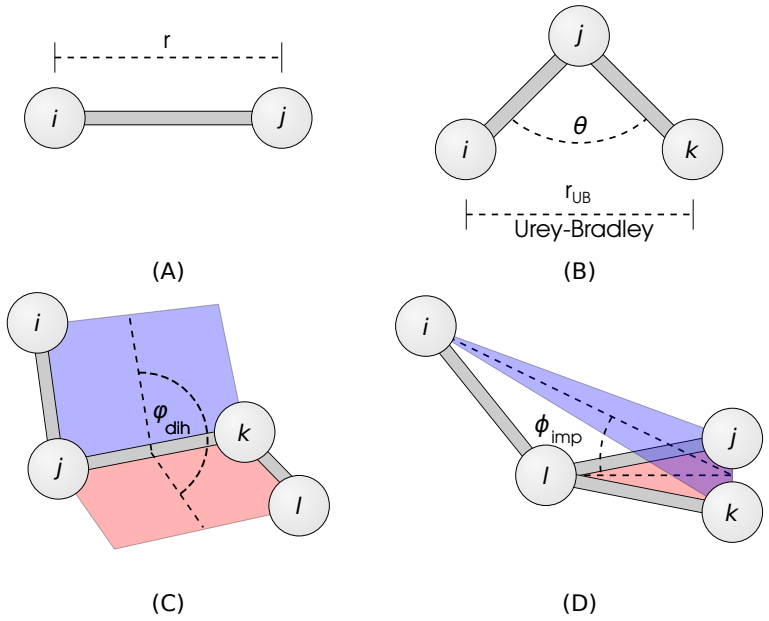

Figure 2.1: The Bonded Interactions Approximated In Classical Forcefields. (A) The energy of Bond Stretching is approximated as a harmonic oscillator with respect to their separation r. (B) Angles between neighbouring covalently bonded atoms are also approximated as a harmonic oscillator with respect 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. (D) The improper dihedral angles are again calculated with the use of two planes. Containing i, j and k and j, k and l respectively. The difference is that this parameter enforces planarity rather than flexibility.

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

$$U_{bonded} = \sum_{bonds} k_b(b - b_0)^2 + \sum_{angles} k_\theta(\theta - \theta_0)^2 + \sum_{dihedrals} k_\phi(1 + \cos(n\phi - \delta)) \\ \sum_{improper-dihedrals} k_\kappa(\kappa - \kappa_0)^2 + \sum_{Urey-Bradley} k_u(r_{UB} - r_{UB_0})^2 \tag{2.8}$$

Here, the k_i terms correspond to the strength of the harmonic restraint for that parameter. The 0 subscript denotes the equilibrium position for that parameter. Even though this formulation is quite simple, it has empirically been shown to accurately model the quantum energetics of bonded interactions at room temperature this can be seen in figure **??**.

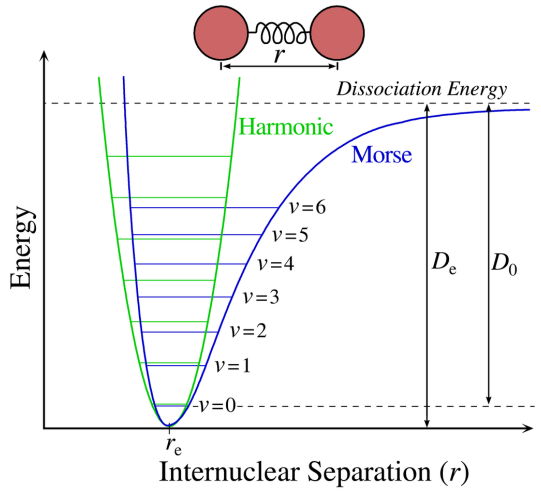

Figure 2.2: The Morse Potential Compared to a Harmonic Potential
The Morse potential was formulated to approximate the potential the potential energy surface of the separation of covalent bonds (blue). At low temperatures (the ground state, v=0) like those found in classical MD there is good agreement between the Morse potential and the harmonic oscillator (green). Credit Mark Somoza 2006.

In classical forcefields the non-bonded interactions are expressed using the Couloumb's law because the partial charges assigned to each atom and the Lennard-Jones potential to approximate the interactions arising from both Pauli exclusion and Van Der Walls Interactions.

$$U_{non-bonded} = \sum_{i>j} \epsilon \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6} - \sum_{i>j} \frac{q_i q_j}{r_{ij}} \tag{2.9}$$

The σ parameter denotes the location of the local minima in the Lennard-Jones potential. This is the 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 [2]

Conversely, the partial charges in a system have the greatest influence on the solvation energy.

By focussing on 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: AMBER, CHARMM, 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 and CHARMM or top down, in the case of GROMOS and OPLS. Bottom up forcefields take the results from quantum *ab initio* calculations and approximate them with the

functional form mentioned above. Conversely, top down forcefields take experimental measurable such as Osmotic pressure, solvation energy. This philosophy is closest to physics

2.2.2 The Process of Preparing a 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 modifications is rarely straight forward. How do you build a missing loop? What charge state is an amino acid most likely to take during the physiological context.

2.2.3 Controlling the Temperature and Pressure in a Simulation

2.2.4 Periodic Boundaries to Simulate a Realistic System Size

2.2.5 Short Comings of Classical MD

The short comings of classical molecular dynamics fall into two classes whose solutions stand opposed to one another. These are the accuracy of the chemical forcefields outlined above and the inability of modern computers to deliver enough samples of the energy landscape to collect sufficient statistics for rigorous conclusions. The issue is that, as the above physical formulation might indicate. The more accurate the forcefields, the more computationally expensive. And so the solutions to the two are constantly in tension with one another. In the next section we will explore the current efforts to bring solutions.

The Problem with Forcefields

These approximations 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 results of current.

The other context where polarisation is important to consider are on divalent ions. Here, the solvation energy is underestimated due to the consistent lack of polarisation, making investigations of these biologically imporatnt chemical species difficult.

However, for most situations, particularly those involving bulk water and protein motions Molecular Dynamics is proving to be an invaluable tool for investigating the properties of biological systems CITATION NEEDED.

There are several efforts to correct address these issues. These include the inclusion of the effects of polarisation, the most popular methods at the moment being adding a massless drude oscillator as an extra bead to most(?) atoms as in the CHARMM

drude forcefields, championed by the Mackerell lab and the use of forcefields such as AMOEBA which explicitly calculate the dipole and quadropole moments of each atom. 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 [3].

Ultimately, the functional form in equation **??** does not have sufficient degrees of freedom to address all possible chemical contexts and so careful consideration must always be given to whether the forcefield is being used in a faithful way to what it was intended to simulate. So long as the user is aware of the situations where classical forcefields fall short they can be a powerful tool for the study of molecular systems.

The Problem with Sampling

To physicists the sampling is the more intuitive. Collecting sufficient statistics about the system of interest is difficult and comes at both a computational and human cost. 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 is we struggle to sample the time scales of diffusion through an ion channel, despite the problem standing for decades.

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.

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 which closely 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 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 toaccelerate them with already established free energy methods. These have the potential to uncover new drug binding pockets and revolutionise our understanding of biomolecular systems.

2.2.6 Accelerating Simulations with Virtual Site Topologies

The discrete time step, Δt in equation **??**, 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, which usually runs to nanoseconds or milliseconds.

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 correlation time is on the order of 1 femtosecond, in classical simulations we are able to get away with using 2 femtoseonds by enforcing a rigid hydrogen bond length, i.e interpolating its position backward from the 2 fs timestep.

However, there are more involved strategies to account for these effects. Hydrogen Mass repartitioning and multiple time stepping have become popular but we will discuss Virtual Sites in detail as they were used for some simulations in this thesis.

NOTE: ADD BIOPHYS TABLE THAT SERDAR HAS IN THE COURSE NOTES

As you can see in table **??** the fastest motion in molecular systems is dicatted by the translation of hydrogen atoms. Virtual site topologies aim to remove the requirement for calculating these motions every time step by instead interpolating the positions of hydrogen atoms from the positions of surrounding heavy atoms.

2.2.7 Umbrella Sampling

Weighted Histogram Average Method

2.2.8 Metadynamics

Chapter 3

Review of the Molecular Cause of Cystic Fibrosis

3.1 Clinical outcomes of Cystic Fibrosis

Cystic Fibrosis (CF) is the most common fatal genetic condition in Caucasian populations. 90 000 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[4]. 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 mucis in their pancreas and large intestines. This leads to CF related diabetes which afflicts roughly half of adults with CF [**Kayani2018**]. 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 [].

3.2 CFTR Structure

CFTR is organised into 7 domains (FIGURE). In the order of their primary sequence they are; The Lasso motif, which anchors into the membrane and serves as an interaction hub with other protein partners such as syntaxin and filamin []. Transmembrane Domain 1 (TMD1) which forms half of the pore. Nucleotide Binding Domain 1 (NBD1) which binds ATP when the channel is in the open state. The Regulatory domain (R-domain) which, when phosphorylated allows the channel to open. Transmembrane domain 2 (TMD2) which forms the other half of the ion conducting pore. Nucleotide Binding Domain 2

CFTR belongs to a super family of proteins known as ATP Binding Cassette Transporters, many of these proteins perform active transport across cell membranes. The substrates they transport can vary, including lipids and drug molecules. Proteins in this family share a common motif known as Nucleotide Binding Domains (NBDs). These domains act as ATPases, accelerating the hydrolysis of ATP. The energy from hydrolysis is then transferred into the protein in order for it to pump its substrate against a concentration gradient.

3.3 CFTR classification and structure

The primary cause of the disease Cystic Fibrosis (CF) is the misfunction of a chloride channel, the Cystic Fibrosis Transmembrane Conductance Regulator (CFTR). This ion channel is a member of the ABCC subfamily of ABC transporters, designated ABCC7. This channel is unique amongst this family because it is not generally considered an active transporter but something of a low conductivity channel or a "weak pump"[**linsdell2018**].

CFTR is distinguished by a regulatory region known as the R-domain (residues 645-845) which links NBD1 to TMD2. This region acts to lock the channel in the closed state by wedging itself between the TMDs and dislodging when any one of 3 sites are phosphorylated [5]. In experimentally determined structures of human CFTR the secondary structure of a section of the R-domain but not at high enough resolution to determine the identity of individual sidechains [6][7]. Further secondary structure information can be found through experiments with NMR [8].

Previous computional studies of CFTR have been used homology models based on the phosphorylated zebra fish protein PDBID:5W81 [9]. These have yielded interesting results but the sequence similarity between human and zebrafish CFTR is only 55% []. For a protein structure where a single amino acid mutation leads to misfunction, more precision can only help. Additionally, the activity of CFTR modulators is not conserved in mutant zCFTR possibly because it has differnet kinetics to the human channel []. In order to do precision medicine we need precision structures.

An open state of the channel has been proposed by combining both the zebra fish homology model and the fully outward facing conformer of a bacterial ABC transporter Sav1866 [**Hoffmann2018**]. Although this model has several characteristics expected of the open channel, such as the critical R352-D993 salt bridge, it lacks a salt bridge between R104-E116. In experiments, tehse residues could be replaced by cysteines and the channel would still function. However, when reducing agents were added to the system the channel lost its ability to open fully. This indicates that in the oxidised environment the C104-C116 cysteins formed a disulfide bridge but its breaking upon exposure to reducing agents caused a loss of function in the channel. This indicates that in the WT channel R104-E116 form a stable salt bridge.

This salt bridge is clearly visible in the recent cryo-EM structure of ATP-bound human CFTR [6].

3.4 The Gating Cycle

The conformational transition from inactive to active differs significantly in CFTR compared to other ABC transporters. The NBD domains are largely similar to other to those found in other ABC transporters, they dimerise in what is termed a head to tail configuration so both subunits contact both bound ATP molecules [] See FIGURE. Residue E1371 allows neucleophilic attack on the γ phosphate of the ATP bound to Walker B [10]. This provides a "kick" to provide the kinetic energy for the opening of the channel CITATION NEEDED.

3.5 Classes of Misfunction to CFTR

The 360 disease causing mutations to CFTR have been classified into 6 common classes based on the nature of the CF they cause, their reaction to CFTR modulators, and results *in vitro* assays. Ultimately I aim to show that at the atomic level these classes of mutations are less meaningful and as patient specific theratyping evolves these classes will become less relevant, serving as illustrative tools only to communicate at a higher level what is going wrong with the CFTR protein. The canonical classification is as follows:

- **Class I** No functional protein. Under these mutations no protein is transcribed due to either problems with the transcription of mRNA or a premature stop codon truncating protein synthesis early, meaning the resulting peptide is missing key domains.
- **Class II** Folding defect. These mutations cause the translated peptide to misfold into the incorrect tertiary structure. This can inhibit the protein's journey as it is trafficked to the cell membrane, its function while once it is there or its functional life time at the surface.
- **Class III** Impaired Gating. Here the mutation inhibits the ability of the protein to transition from the closed to the open state.
- **Class IV** Decreased Conductance. These mutations cause a barrier in the energy landscape of the CFTR chloride conductance pathway.
- **Class V** Less Protein Expressed.
- **Class VI** Decreased Lifetime

Although useful, in reality this paradigm struggles to reflect the fact that a mutation can belong to multiple categories to different levels due to different modes of pathogenesis. Through our molecular simulations we can see that in reality CFTR modulators are capable of treating several different mutations with very different molecular fingerprints.

FIGURE demonstrates how each of the canonical classes at the molecular level is broken down into many sub classes and a mutation might belong to one of many of these subclasses. Structural biology paradigms and *in silico* modelling can help classify mutations into these different classes. In combination with wet lab assays we can understand which classes of these molecular defects are most effectively treated with specific drug regimens. Our computational microscope is helping choose treatments for patients at the atomic level.

3.6 CFTR Modulators

Since CF is caused by malfunctions of the channel it makes sense to pursue CFTR as a drug target. Through high throughput *in vitro* screening several (GET NUMBER) compounds have been developed that aim to rescue the function of CFTR. These fall

into two classes. Correctors, which aid CFTR to fold into the correct state and potentiators which help the channel reach the fully open state once it has already folded correctly. Emerging evidence suggests that specific genetic defects may be optimally rescued by specific combinations and doses of both correctors and potentiator compounds. Recently, cryo-EM structures of these compounds in their bound state have been released. In addition to several *in vitro* biophysical experiments to determine the precise mechanism of action and binding site of these compounds.

3.6.1 Correctors

The mechanism of action for corrector compounds appears to be to bind to to a pocket between TMH1 and TMH3. Circular dichromism and fluorescence experiments found that an isolated construct of TMH3 and TMH4 were more likely to fold correctly in the presence of corrector compounds. Later cryo-EM structures discovered high resolution electron density in the pocket in the shape of the drug compounds [**fiedorczuk2022**].

In combination this is strong evidence for the precise mechanism of action for corrector compounds. Further work will aid in the creation of new compounds to refine our exploitation of this mechanism.

Mention that there are some interactions between correctors and NBD1.

3.6.2 Potentiators

There is more uncertainty surrounding the mechanism of potentiator drugs. Experiments clearly demonstrate that they act directly on CFTR in order to increase the likelihood that it occupies the open state. They bind to the protein with picomolar affinity. There are are cryo-EM structures which show the drugs bound to the TM8 hinge region []. *In vitro* experiments suggest at least two membrane facing binding pockets due to the drugs extreme hydrophobicity[]. The location of this second binding site is unknown. The difficulties arise with mutagenesis experiments. The dose-response curves in several studies show that when various sites are mutated the activity of the drug is lowered. This indicates additional binding sites not yet well defined.

GLPG1837 has not been approved in a clinical setting. *in vitro* experiments suggest that it is more efficacious even though it has lower affinity for CFTR binding (CITATION NEEDED). This would indicate that the highest affinity binding pocket does not produce the greatest modulation. More work is needed to resolve the mechanism which results in the clinical effectiveness of these drugs.

These drugs are clinically efficacious [11] on several mutants with some curious exceptions like N1303K. I suggest the following mechanism for their action. I suspect a similar analogy exists for the action of the correctors. WT-CFTR exhibits a natural landscape with kinetic barriers in the transition between the closed and open states. A gating class mutation to CFTR will introduce a kinetic barrier in the pathway of this conformational transition. What these drugs do is reduce a barrier in the existing conformational landscape of CFTR. This compensates for the barriers introduced by

the mutation.

This provides a rationale for why it appears possible for diverse range of molecular defects to be treatable by these small molecules. In our work we've found that the atomic nature of the defects introduced by each mutation varies widely, what is interesting is that experiments in *ex vivo* models have shown that these drugs treat a variety of different defects. The classification of classes of defect is outdated, really there are as many classes as there are mutations.

3.6.3 Annion Selectivity

CFTR is weakly selective for specific annions. F337 is the most important amino acid for selectivity. Bicarbonate (HCO_3^- is known to have roughly 26% the permeability of chloride through the channel. Note that Fluoride has even higher conductance through CFTR, likely due to its small size and high solvation energy (does this indicate hydrated conductance?).

Compared to cation channels like gramicidin and KcsA CFTR is only weakly selective, permeating a large set of annions with varying radii and geometries. Supposedly it is more permeant to lyotropic (low solvation energy annions) rather than cosmotropic annions (high solvation energy annions) indicating that dehydration of the annion is likely during conductance (CITATION NEEDED). The radius of hydrated chloride ions is 3.2A[12] so even with this larger pore partial dehydration must take place.

Chapter 4

Concluding Remarks

Bibliography

(1) Born, M.; Oppenheimer, R. *Annalen der Physik* **1927**, *389*, 457–484, DOI: 10.1002/andp.19273892002.

(2) Yoo, J.; Aksimentiev, A. *Physical Chemistry Chemical Physics* **2018**, *20*, 8432–8449, DOI: 10.1039/c7cp08185e.

(3) Ngo, V.; Li, H.; MacKerell, A. D.; Allen, T. W.; Roux, B.; Noskov, S. *Journal of Chemical Theory and Computation* **2021**, *17*, 1726–1741, DOI: 10.1021/acs.jctc.0c00968.

(4) Boucher, R. C. Airway Surface Dehydration in Cystic Fibrosis: Pathogenesis and Therapy, 2007, DOI: 10.1146/annurev.med.58.071905.105316.

(5) Mihályi, C.; Iordanov, I.; Töröcsik, B.; Csanády, L. *Proceedings of the National Academy of Sciences* **2020**, 202007910, DOI: 10.1073/pnas.2007910117.

(6) Zhang, Z.; Liu, F.; Chen, J. *Proceedings of the National Academy of Sciences of the United States of America* **2018**, *115*, 12757–12762, DOI: 10.1073/pnas.1815287115.

(7) Zhang, Z.; Chen, J. *Cell* **2016**, *167*, 1586–1597.e9, DOI: 10.1016/j.cell.2016.11.014.

(8) Baker, J. M.; Hudson, R. P.; Kanelis, V.; Choy, W. Y.; Thibodeau, P. H.; Thomas, P. J.; Forman-Kay, J. D. *Nature Structural and Molecular Biology* **2007**, *14*, 738–745, DOI: 10.1038/nsmb1278.

(9) Zhang, Z.; Liu, F.; Chen, J. *Cell* **2017**, *170*, 483–491.e8, DOI: 10.1016/j.cell.2017.06.041.

(10) Stratford, F. L.; Ramjeesingh, M.; Cheung, J. C.; Huan, L. J.; Bear, C. E. *Biochemical Journal* **2007**, *401*, 581–586, DOI: 10.1042/BJ20060968.

(11) Van Goor, F.; Yu, H.; Burton, B.; Hoffman, B. J. *Journal of Cystic Fibrosis* **2014**, *13*, 29–36, DOI: 10.1016/j.jcf.2013.06.008.

(12) Yang, K.-L.; Yiacoumi, S.; Tsouris, C. *The Journal of Chemical Physics* **2002**, *117*, 8499–8507, DOI: 10.1063/1.1511726.