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

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

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 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 a special case of chemistry, with many observable functions. Their sequence is coded by the DNA in something reminiscent of a strange loop [**Hoffstadter2008**].

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

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. So in this thesis we will only discuss single proteins.

1.3 Ion Channels: Natures laboratories to Teach Us Biophysics

The physiological importane of ion channels became clear after the experiments of Hodkin and Huxley. These mathematicians took nerves from fished 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[**hodgkin1952**]. 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.

$$\begin{aligned} I = C_m \frac{dV}{dt} + \bar{g}_K n^4 (V - V_K) + \bar{g}_{Na} m^3 h (V - V_{Na}) + \bar{g}_l (V - V_l), \\ \frac{dn}{dt} = \alpha_n(V)(1-n) - \beta_n(V) n, \\ \frac{dm}{dt} = \alpha_m(V)(1-m) - \beta_m(V) m, \\ \frac{dh}{dt} = \alpha_h(V)(1-h) - \beta_h(V) h \end{aligned} \tag{1.1}$$

The α and *beta* parameters are the proportion of the sodium and potassium channel populations which are activated, respecitively. This example shows how basic theoretical tools can be used to predict and discover physical phenemon in biological systems. The Hodkin Huxley model proved the existence of a cell's resting potential, the posibility 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.

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 biolgoy has advanced it has become clear that the resting potential of a cell is critical to its function, regulating many chemical reactions inside it.

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 .

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, 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. 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 cure diseases and create amazing things.

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 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 .

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.

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 $\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}$

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 sometimes referred to as molecular mechanics. The aim of the classical forcefields discussed here is to use *ab initio* MD as a target to approximate.

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}.

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