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

Acknowledgments

In loving memory of Madeline Jennifer Dell

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 have spent any time with before this point. By now many of those categories overlap, which is a gift in itself. The list that follows 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. These are the people who for arbitrary reasons are being given explicit thanks for helping me with this thesis.

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.

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. 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 bringing me a truly fascinating PhD project and to her and the rest of her lab for bearing with two physicists who still don't really know any biology. Bridging the gap between cell biology and molecular physics is something the world will be doing more of in the future and I'm lucky to have met such a driven lab to help me 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 privelage during a PhD and I'm all the better for having gotten it from one of the best. Never did I lack for your time or your wisdom.

Maddy I miss you every day. You couldn't have imagined what it was like to do this after losing you. You gave me so much that helped me do this. 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.

"Fear cuts deeper than swords."

Arya Stark

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 oscilator 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 the skin. If you take a miscroscope 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

miscroscope 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 moelcules 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 intergallactice 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 occuring 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 heterogenous 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 prokaryotes 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 .

Biological strange loops would not seem to be as self similar as the clean nice logics in the strange loop of the Godelian knot. Why is this?

1.3 Why Cystic Fibrosis

The sad truth of this debiletating 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 perpsective I'm missing by not suffering myself from such a condition or even knowing somebody with it. I'm a relatively healthy well adjusted Male. I have not been trained in the ethics of studying medicine and my undergraduate professors were only concerned with what was morally acceptable when it came to mathematical theorems.

In this way, my motivations for studying this disease aren't wholly humanitarian. 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 debiletation..

Chapter 2

Theory and Methods

2.1 Approximating Quantum Mechanics with the Goal of Studying the Components of Living Things

Living things are made of atoms and the motion of atoms is governed by quantum mechanics. Unfortunately, quantum mechanical calculations for the number of atoms involved in proteins and other cellular components are computationally intractable. Hence, we will show how to take the fundamental formulation of atomic interactions in the Schrödinger wave equation and apply physically motivated approximations in order to simulate macromolecular systems at biologically relevant timescales.

For any quantum system we begin from the time dependent Schrödinger wave equation

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

When the external potential V is independent of time this equation reduces to the familiar time independent form.

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

The Hamiltonian of this system is given by $H = \nabla^2 + V$

For an atomistic system we begin by noting that there are two types of particles, nucleii which we denote with the subscript i and electrons denoted by e. In order to treat these elements seperately we decompose the Hamiltonian of the system into a few components.

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

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 clilia on the epithelium collapse leaving them unable to clear the mucus that naturally lines the airway[**boucher2006**]. 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"[1].

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 [2]. 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 [3][4]. Further secondary structure information can be found through experiments with NMR [5].

Previous computional studies of CFTR have been used homology models based on the phosphorylated zebra fish protein PDBID:5W81 [6]. 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 [**Hoffman2018**]. 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 [3].

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 [**Stratford2007**]. This provides a "kick" to provide the kinetic energy for the opening of the channel.

The NBD

3.5 Classes of Misfunction to CFTR

The 360 disease causing mutations to CFTR have been classified into 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 relavent, 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 trancription 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 traficked 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 misfunctions of the channel it makes sense to pursue CFTR as a drug target. Through high throughput *in vitro* secreening 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 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 compouds.

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 strcutures discovered high resolution electron density in the pocket in the shape of the drug compounds.

In combonination 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 Poteniators

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 experiments 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 efficatious even though it has lower affinity for CFTR binding (CITATION NEEDED).

These drugs are clinically efficatious [**VanGoor2014**] 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.

Chapter 4

Concluding Remarks