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 [1]

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. This list of thanks is by nature incomplete. If it was you'd be reading about a conversation I had with a middle aged public service woman in a hostel north of San Francisco, but that has little to do with Cystic Fibrosis.

To my parents. You raised me with not only academic rigor in mind but also a respect for the arts 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 the exposure to the arts you gave me I'd have remained illiterate.

To Jeffry Setiadi. You got me started with these weird and wacky simulations. For your tutelage and patience, even from across the pacific ocean.

To Poker Chen. 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.

To 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.

To Renate Griffith, as well as Katelin Allen, Sharon Wong, Laura Fawcett and the rest of the miCF lab. For teaching me so much cell biology and helping me write manuscripts. 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 dedicated lab to teach me to do so.

To the patients and their families who participated in the studies designed by miCF. I hope there is some day a cure for CF and that our work is a small step toward that goal.

To Shafagh Waters herself, head of miCF. For your vision, your drive and all your advice. You brought me a truly fascinating PhD project and I benefited greatly from your mentorship. You really helped me figure out *how* to do this kind of research and your approach to science has made me a better researcher.

To 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 deep physical insight into biological systems to future research projects.

Of course I must also thank many friends for inspiration and motivation. Alon, Ollie, Chris, Zac, Josh, Markus, Deb, Harry, Ashwin, Calida, Hway, Nicko, David, Alex, Amy, Nick, Natasha, Aleksa and Frank.

Finally, Maddy, I love you. 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.

To all of you, thank you for your help along the way. You're all in my Loop [2] and I hope I'm in some of yours.

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 **xii**

List of Figures **xiv**

List of Tables **xv**

1 Introduction: Biology, The Hardest Science **4**

1.1 Thesis and Chapter Summary 4

1.2 What is Physics? 5

1.3 The Physics Inside your Cells 7

1.4 Using Ion Channels as Natural Laboratories to Learn Biophysics 8

1.5 Studying Cystic Fibrosis; Toward a Molecular Theory of Disease. . . . 11

1.6 The Future is Biological 13

2 From Protons to Proteins: Methods to simulate the inside of a cell. **15**

2.1 Quantum Mechanics is Not Tractable at the Scale of Biology. 15

2.1.1 A full quantum mechanical treatment 16

2.1.2 The Born-Oppenheimer approximation. 17

2.2 Classical MD; Molecular Motions Without Quantum Mechanics 18

2.2.1 Philosophy of Different Molecular Mechanics forcefields. 22

2.3 Periodic Boundaries to Simulate the Inside of a Cell 24

2.4 Controlling the Temperature and Pressure in a Simulation 27

2.4.1 Hot and Cold with the Nosé-Hoover Thermostat 28

2.4.2 Under Pressure with the Parinello-Rahman Barostat 30

2.5 The Process of Preparing an MD Simulation 31

2.6 Choosing an Appropriate Time Step 32

2.6.1 Verlet Leap-Frog Integration 33

2.7 Free Energy Calculations: Making Simulations More Useful 34

2.7.1 Umbrella Sampling 35

2.7.2 Metadynamics 37

2.8 Short Comings of Classical MD 39

2.8.1 The Problem with Forcefields 39

2.8.2 The Problem with Sampling 40

2.9 Conclusion 41

3 Review of the Molecular Cause of Cystic Fibrosis and Its Treatment **43**

3.1 Clinical outcomes of Cystic Fibrosis 43

3.2 CFTR Structure 44
3.3 CFTR is a Unique ABC Transporter 46
3.4 CFTR Structure and Function 46
3.4.1 The Gating Cycle 48
3.4.2 Anion Selectivity 48
3.5 Classes of Mutation Which Cause Cystic Fibrosis 49
3.6 CFTR Modulators 50
3.6.1 Correctors 50
3.6.2 Potentiators 50
3.6.3 Patients with rare mutations struggle to gain access to modulator therapy 51
3.7 Patient Derived Organoids as a Pre Clinical Model 51
3.7.1 Forskolin Induced Swelling 52
3.7.2 Cilia Beating Frequency 52
3.7.3 Electrophysiology 52
3.7.4 Western Blotting to Assess CFTR Trafficking 52

4 Molecular Dynamics and Functional Characterization of I37R-CFTR Lasso Mutation Provide Insights into Channel Gating Activity 53
4.1 Introduction 54
4.2 Results 56
4.2.1 I37R-CFTR Baseline Activity in Patient-Derived Rectal Biopsies and Intestinal Organoids 56
4.2.2 I37R-CFTR Functional Response to CFTR Modulator Monotherapy in Intestinal Organoids 58
4.2.3 I37R-CFTR Perturbs the Noose Structure of the Lasso Motif . . 61
4.2.4 I37R Mutation Strengthens Lasso Motif Interaction with the R-domain 62
4.3 Discussion 63
4.4 Method Details 66
4.5 Acknowledgements 70
4.6 Author Contributions 70
4.7 Delcaration of Interests 70

5 Molecular Dynamics and Theratyping in Airway and Gut Organoids Reveal R352Q-CFTR Conductance Defect 71

6 Unique S945L-CFTR defect Restored by CFTR Modulator Co-Therapy In Vitro Correlates with In Vivo Biomarkers Post-Therapy 72

7 Resolving a Conducting Conformation of CFTR Using Free Energy Calculations 73
7.1 Introduction 74
7.2 Results 74
7.2.1 The outer pore in the CryoEM structure of phosphorylated human CFTR is not sufficiently open to conduct anions 74

7.2.2 Using metadynamics and long simulations we can dilate the pore to discover a conducting conformation 74
7.2.3 Proving the discovered conformation is capable of ion conduction with umbrella sampling 74
7.2.4 This open conformation gives rise to a novel salt bridge 74
7.2.5 This open state can be used to study disease causing mutations such as R334W in the outer pore 74
7.3 Discussion 74
7.4 Conclusion 75
7.5 Methods Details 75

8 Concluding Remarks: Where next for Molecular Studies of CFTR. 76
8.1 Addressing Controversies Surrounding the Structure of CFTR 77
8.2 A Physics Motivated Model for the Molecular Modulation of Mutant CFTR 79
8.3 A Physics Motivated Approach to Precision Medicine in Cystic Fibrosis 79
8.4 Outstanding fundamental questions about CFTR Function 80
8.5 Grouping of Theratypes 80
8.6 Resolving Drug Action 80
References 82

Bibliography 83

List of Abbreviations

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

List of Figures

1 Excitory and Inhibitory Synapses by David Goodsell 2

2 The strange position that biological phenomenon occupy compared to the rest of physics 3

1.1 The Structure of DNA has Periodic and Aperiodic Elements 6

1.2 The Action Potential is a Solution to the Hodkin-Huxley Model 9

1.3 Different Ion Channels Ammenable to Molecular Simulation 10

1.4 A physical model for how to view protein evolution 12

2.1 The Bonded Interactions Calculated In Classical Forcefields 21

2.2 Comparison Between Potentials in Quantum and Classical Forcefields . 22

2.3 The Lennard-Jones Potential 23

2.4 An Example of a Solvated Biomolecular Environment Ready for Simulation 25

2.5 Particle Mesh Ewald Summation 26

2.6 Illustration of Umbrella Sampling 35

2.7 Illustration of Metadynamics 38

3.1 CF Clinical Progress 44

3.2 CFTR Structure 45

3.3 CFTR Structure 47

4.1 Graphical Abstract. Integration of in silico and in vitro experiments for personalised medicine 54

4.2 Characterization of I37R-CFTR residual function in rectal biopsies and intestinal organoids 57

4.3 Characterization of I37R-CFTR functional response to corrector or potentiator monotherapy in intestinal organoids 59

List of Tables

2.1 Timescales of Motions in a Molecular System 33

Foreword

I've never held a pipette

The more I wrote of this thesis the more I found myself writing down things I wish I'd known when I first started studying biophysics. I was writing for my past self. Hence, I think the audience for this thesis would be those with a loose grasp of third year undergraduate physics. The most difficult thing for this audience will not be the mathematics or technical content herein, but rather the breadth of biochemical pre requisites to understand the scope of the contents. I have not had time to write an introduction to molecular biology, so I recommend a physics based introduction to those concepts such as those found in [3]. On this note of pedagogy, some care has been taken to name certain authors to give the reader a kind of anchor to keep track of the literature. Similarly, when a technical concept is mentioned, I have searched for a useful review article, book or explainer. So please take citations as reading recommendations [4, 5].

When I first started simulating them, I didn't even know what a protein was, but for the past 4 years I have been captivated by the unending complexity in biological systems. For some examples, please take this opportunity treat yourself to the fabulous work of David Goodsell in figure 1 . I've found that the mindset for solving biological problems feels very different to the focus we cultivate in students when they study idealised problems in mathematics and physics. The problems are more specific and the required knowledge is broader. For example, plasma physicists may use the same mathematical tools to describe materials as diverse as the dense stellar core to the sparse intergalactic nebulae. These objects span 28 orders of magnitude in density [9]. Would that we were so lucky in biology, where we struggle to apply same physical models to deal with phenomena across a single order of magnitude.For context of the length scales of biology, see figure **??** This heterogeneity means we need many hands to solve problems in biology. Note that the publications arising from this thesis have many authors. Each researcher specialises, not unlike their cells, in a specific discipline but we all work together to answer different aspects of the same biological questions.

Why do we need so many specially trained people to do biological research? The complexity of biology is easy to observe. If you look at your hand, you will notice hair, pores, dry skin, dead skin, perhaps even tendons and muscles twitching beneath a faint web of ghostly blood vessels. If you were to pluck a single cell from anywhere in this hierarchy and place it under an electron microscope, you would find diverse

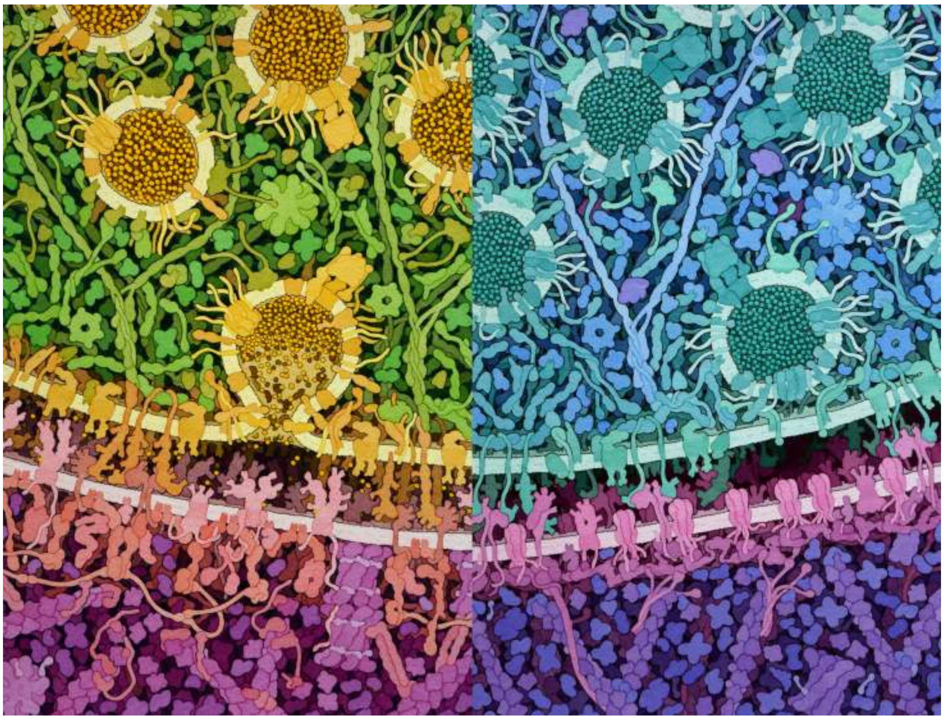

Figure 1: Excitory and Inhibitory Synapses by David Goodsell
David Goodsell is an artist who produces water colors cellular environments. Many of his works can be downloaded and used for free. This particular painting depicts some molecular processes in the brain, showing shows how excitory (Glutamate) or inhibitory (GABA) neurotransmitters are packaged by proteins into vesicles and then released into the synapse, where they bind to other protein receptors on the other side, allowing signals to propagate between neurons. Goodsell has produced many books and articles which would serve as a light, layperson friendly introduction to molecular biology [6–8].

structures, called organelles. The size, shape and function of the organelles would be different if the cell was taken from somewhere else in your body. Within and between each those organelles is a wet, salty dance of molecular machines called proteins, which we will study in detail. The length scales of this journey from your arm to a single protein spans 8 orders of magnitude. At each step along the way, there are thousands of experts studying specific phenomenon at that length scale. Only by working together can these experts try to understand the whole organism.

If the reader is anything like myself they will find the amount of required knowledge to study biology a substantial barrier to entry. It is quite difficult at first to figure out what questions to ask or even figure out which subfields to study to alleviate this confusion. These challenges can be tackled by cultivating a broad coalition of connections; speak to medical doctors, clinicians, molecular biologists, evolutionary biologists [4, 10], philosophers, biochemists[1], cell biologists, geneticists, bioinformaticians, theoretical chemists, computer scientists, neuroscientists, physicists, mathematicians, everyone. It will take

[1] These last two are in fact different subfields but like many in this list it'll take you some time to understand the subtle differences which define each one.

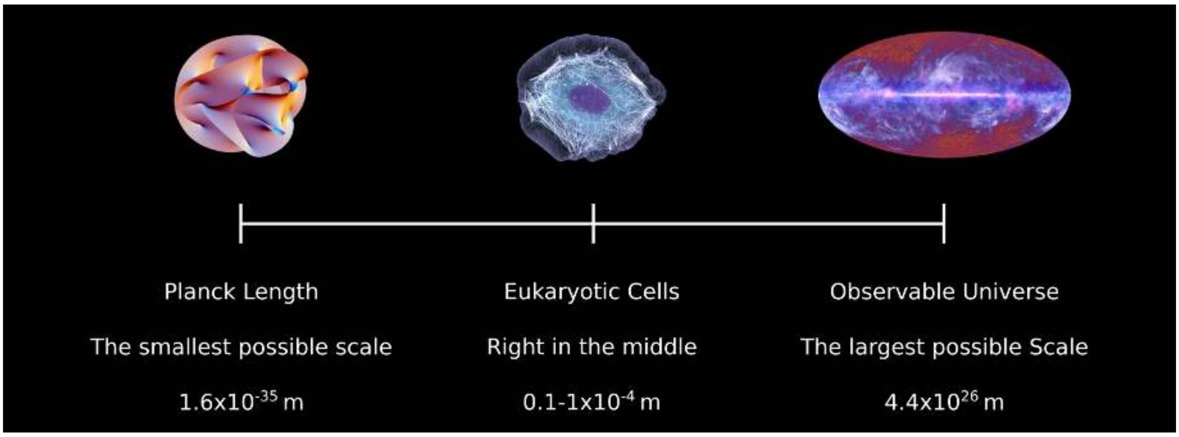

Figure 2: The strange position that biological phenomenon occupy compared to the rest of physics
It just so happens that if you plot the size of everything in the universe on a log scale, eukaryotic cells fall right in the middle. A physicist can talk about both ends of this scale, we should learn something about the middle too.

time but remain patient and you will find that a physics motivated approach can indeed explain and eventually predict outcomes in biological experiments. A broad world view awaits you and it's really quite fun.

If this is being read by a future trainee, I hope the physics focussed philosophy in the introduction of chapter 1 and the literature review of simulation techniques in chapter 2 can serve as a road map, but a physicist studying biology will be best served by nurturing a strong base in electrodynamics and statistical mechanics [11–13]. One particularly thorny issue is that the field is now progressing so quickly that I'm sure much of this thesis will be out of date by the time it's read by anybody I'd hope to train. But really that's part of the excitement. Shoot me an email if you want help miro.astore@gmail.com. Hopefully I'm still around.

The task head is daunting. We're not exactly sure how many proteins are expressed by the human genome, but it's estimated that there are upward of 20 thousand [14]. By contrast, this thesis represents an all consuming effort by a single PhD student expending decades worth of computation time to make incremental progress in the study of a single one. There is so much to do. Good luck, and bring a towel [15].

Chapter 1

Introduction: Biology, The Hardest Science

> *Whatever complexity means, most people agree that biological systems have it.*
>
> -Frauenfelder and Wolynes [16]

1.1 Thesis and Chapter Summary

This thesis seeks to apply a philosophy of molecular biophysics to demonstrate its capability to investigate pressing problems in biology and medicine. In particular, we will use molecular dynamics (MD) to look at how a specific gene, the Cystic Fibrosis Transmembrane Conductance Regulator (CFTR), to cause a disease (CF). These MD techniques will allow us to formulate a model of CFTR's misfunction which we hope will direct research efforts and allow more patients suffering from CF access life saving medication.

In this first short chapter we will quickly build a philosophy of how to look at biology through the lens of a physicist. We will outline what the goal of a physicist is, to create abstract formalisms which can be used to model the natural world. We will then observe what makes the construction of such formalisms so difficult in biology. We will walk through how a specific set of systems, namely ion channels have historically served as laboratories to help understand more complex biological systems, such as whole cells or organisms.

In more detail, chapter 2 will describe the chemical and numerical simulation techniques we have used to study the CFTR protein system, while chapter 3 gives an overview of the CFTR system itself, and also a set of *in vitro* assays which compliment our computational modelling. Chapters 4, 5, **??** and 7 demonstrate the details of how a diverse application of the simulation techniques in chapter 2 can be used to discover the unique modes of misfunction in CFTR. In combination with *in vitro* cellular techniques, these simulation results prove that these mutations can be rescued by existing drug regimens. Finally, chapter **??** ties together these results to argue for a physical model which elucidates the mechanism of action for cystic fibrosis drugs. Armed with this model we will work through the available literature and identify some priorities for future studies using molecular modelling for cystic fibrosis research.

We hope this small example can demonstrate the utility of physics expertise for the field of molecular medicine. We anticipate that such methods will only grow in power with improvements in computational and experimental techniques.

1.2 What is Physics?

When I was in high school I always described physics as “the study of how things move”. Although intuitive, this description does not shed light on the philosophy of doing physics which make it such a powerful tool for understanding the natural world. To create a predictive physical, theory one must first carefully define a naturally motivated formalism. Then, using mathematics, the implications of this formalism are built up to make predictions about measurable phenomena. Should the predictions from the formalism agree with experimental results, it validates the physical theory. This is what makes physics feel like the most ”fundamental” of the sciences

These formalisms can take a few forms. Examples include:

- The gas of hard spheres, which we use to derive the Boltzmann’s kinetic theory of gasses.

$$\frac{\partial f_1}{\partial t} + \frac{\mathbf{F}}{m} \cdot \nabla_{\mathbf{r}} + \mathbf{v} \cdot \nabla_{\mathbf{r}} f_1 = \int \int (f_1' f_2' - f_1 f_2) \tilde{\kappa} d\hat{\kappa} d\mathbf{v}_2 \tag{1.1}$$

- The interacting electric and magnetic vector fields in Maxwell’s laws of electromagnetism.

$$\partial_\alpha F^{\alpha\beta} = \frac{4\pi}{c} J^\beta \tag{1.2}$$

- The Riemannian manifolds which define the curvature of spacetime in Einstein’s theories of relativity.

$$G_{\mu\nu} + \Lambda g_{\mu\nu} = \kappa T_{\mu\nu} \tag{1.3}$$

- The complex probability waves which evolve according to Schröedinger’s equation, describing quantum mechanics.

$$i\hbar \frac{d}{dt} |\psi(t)\rangle = \hat{H} |\psi(t)\rangle \tag{1.4}$$

As biophysicists we would wish to find the most basic formalisms for biological phenomenon, so we can fully understand the function of an organism. The quest for such a formalism has an interesting origin. In 1944 Erwin Schröedinger wrote an essay titled “What is Life?” This remarkable work was written before the discovery of DNA’s structure or the maturation of information theory. Schröedinger uses first principals in thermodynamics and quantum mechanics to speculate at the nature of life at the

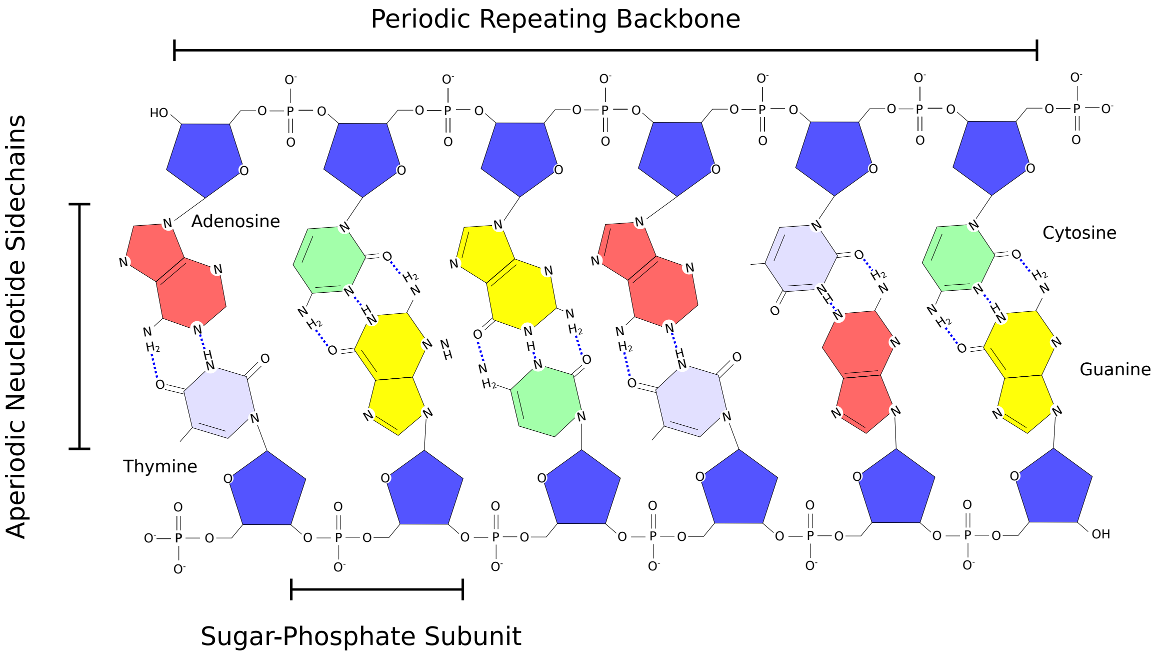

Figure 1.1: The Structure of DNA has Periodic and Aperiodic Elements Although not a crystal, the structure of DNA has some periodic and some aperiodic elements. This is reminiscient of Schröedinger's speculation that genetic information is chemically encoded in an aperiodic crystal or an aperiodic solid.

atomic level. The most remarkable thing about this essay is how much the author gets right.

He observes that since organisms exist at temperatures on the order of 10^2 Kelvin the phyical encoding of genetic information inside cells must be chemical in nature, as physical arrangements of atoms would be unstable at such temperatures without chemical bonds. Schröedinger posits the existence of what he calls an "aperiodic crystal". Although not a crystal in the rigorous sense, the double helix structure of DNA is not far off such an analogy. As figure 1.4 shows, the DNA double helix is a combination of periodic and aperiodic elements, conceptually reminiscent of what one would expect of an aperiodic crystal [17]. This allegory is an example of how physical principals can in fact be used to direct questions in fundamental biology. In fact, James Watson (one of the discoverers of DNA's structure) credited Schroëdinger's book as one of his influences in how he thought about genetics [18].

In this example, Schröedinger has naturally chosen a formalism of interacting atoms, mostly consistent with the statistical mechanics in equation 1.1, to arrive at his model of an aperiodic crystal. In this thesis, as chapter 2 will outline, we will use a more careful, but similar formalism to study the function of the CFTR protein. The goal of chapters 4, 5, **??** and 7 is to collect evidence in order to develop a physics inspired model of Cystic Fibrosis disease which we will analyse in chapter 8. The model we arrive at is more abstract than we are used to for physicists and the next section will explore why this is often the case for biological systems.

1.3 The Physics Inside your Cells

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

These might seem like odd questions but if you asked a physicist how much power it would take to ionise a gas or how long it will take a black hole to evaporate and they will have highly accurate models at the ready to answer easily.

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

A physical theory such as those in the list in the previous section may fail for one of three reasons. Either we do not have sufficient computational power to integrate the formalism to make predictions about a specific phenomenon, we lack sufficient data to specify reasonable initial conditions for integration by the formalism, or the predictions from the formalism disagrees with experimental measurements of a phenomena. The latter occurs when a theory is applied outside the energy or length which it is capable of describing[1]. Quantitative theories of biology, from a fundamental physics point of view, fall somewhere in the middle of the first two categories. Our current physical theories have sufficient accuracy at the energy and length scales of biology that we can accurately model every phenomenon inside a living being [20].

The difficulty of studying biology then, does not arise from complex interactions. As we will see in chapter 2 the interactions between atoms within living things is surprisingly simple and a formalism of interacting atoms is appropriate for modelling cellular functions. Rather, the complexity in biological problems arises from the sheer number of interactions we must consider. Inside cells we find proteins, lipids, solvents, salts each with their own properties. Biophysics distinguishes itself from more traditional physics as it considers systems that are highly heterogeneous and anisotropic. This makes it difficult to scale up formalisms using human tractable mathematical tools. The more heterogeneous the system the more complex the mathematics becomes and thus, the more we must rely on brute force computation of a lower level theory.

This is not to say we can simply solve all biological with brute force computation of low level theory. There is a rich field of biological mathematics which shows how elegant applications of mathematics can shed light on macroscopic biological systems. We will see this in the examples of the conduction of signals through a nerve.

This heterogeneity is perhaps why this thesis contains discussions of quantum mechanics all the way up to a patient lung capacity.

Although considerable success can be found in modelling biology with formalisms that do not include the explicit treatment of interacting atoms, such models must be tailor made in order to treat specific phenomenon [3]. The advantage of MD and the formalism of interacting atoms in chapter 2 is that they are the most accurate available to the biophysicist. The issue is the considerable computational load attached to them.

[1] An example of this would be attempting to use Newton's theories of gravity to predict the motion of an object around a super-massive black hole. For this we need results from Einstein's general relativity [19]

Thus, in order to move towards more predictive theories of biology it is necessary to develop layers of physical theories applicable in different contexts. One form of this from fundamentals approach is the simulation of every atom in a biological system. Although computationally expensive, this approach has been proven necessary when studying the molecular details of proteins, due to the heterogeneous nature of biological systems [21, 22].

1.4 Using Ion Channels as Natural Laboratories to Learn Biophysics

Ion channels are a special kind of protein which allow the passive of charged particles through a cell membrane. They are excellent laboratories for the study of biophysics for two reasons. Firstly, it is very easy to measure their activity with a technique called electrophysiology [23] [2]. Secondly, they are critical to the health and function of cells. As cell biology has advanced it has become clear that the level of polarisation (potential difference from the inside to the outside) in a cell is critical to its function. Changes to the polarisation regulate many chemical reactions inside the cell [24–27].

It is perhaps then not surprising but nonetheless remarkable that ion channels are the targets of 19% of clinically approved drugs [28]. However, there is much more work to be done as candidates drugs are often insufficiently selective for desired ion channel, leading to sometimes lethal drug side effects [29–31].

Historically, ion channels have served as a testing ground for biophysical models. The first interest in modelling their behaviour comes from the experiments of two biophysicists, Andrew Hodgkin and Alan Huxley. They threaded silver wires through the thick nerves of a giant squid and measured the current running through the nerve in response to electrical stimulation. What they found was intriguing. Signals would only propagate down the nerve when the input signal was of a sufficient voltage. They managed to match their experimental data with a model comprised of the following set of ordinary differential equations:

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

$$\begin{aligned}
\frac{dn}{dt} &= \alpha_n(V)(1-n) - \beta_n(V)n, \\
\frac{dm}{dt} &= \alpha_m(V)(1-m) - \beta_m(V)m, \\
\frac{dh}{dt} &= \alpha_h(V)(1-h) - \beta_h(V)h
\end{aligned} \tag{1.5}$$

[2] Electrophysiology is a field a physicist would best understand as using sophisticated setups involving precise oscilloscopes to measure the amount of current and voltage across a membrane. This can be on the scale of the whole cell all the way down to a single ion channel. A brief discussion of some different techniques can be found in chapter 3.

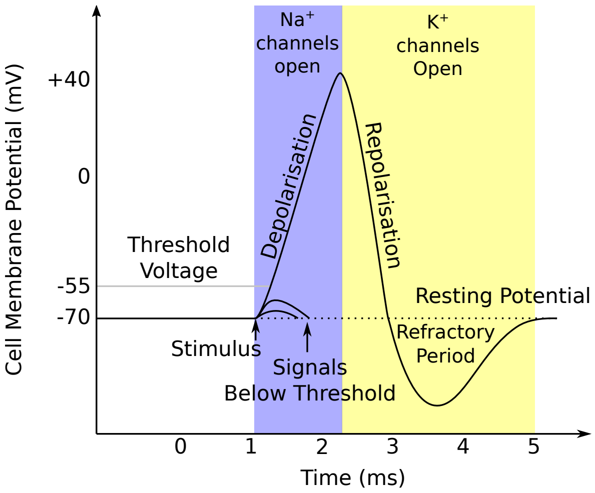

Figure 1.2: The Action Potential is a Solution to the Hodgkin-Huxley Model The shape of the action potential is a similar sight in many physiology textbooks. It was is in fact discovered as a result of the mathematical modelling of Hodgkin and Huxley hinting at the deep biophysics of ion channels won them the 1963 Noble Prize in medicine. This discovery is an excellent example of how deep theoretical insight can lead to predictable models of living systems [32–36].

Here, the n, m and $h \in [0,1]$ parameters are associated with potassium channel subunit activation, sodium channel subunit activation, and sodium channel subunit inactivation, respectively. C_m is the capacitance of the lipid membrane per unit area, and $\bar{g}_i$ is the maximal conductance allowed across the membrane, per unit area. The terms V_i denote either the total voltage or the contribution to the total from a specific charged species.

The solutions to the Hodgkin Huxley model allow us to mathematically discover and describe several important cellular functions. The model encodes the existence of a cell's resting potential and selective voltage gated ion channels. Even today, the molecular mechanisms behind these discoveries are used understand protein and cellular function [].

This is an example of the development of a mathematical formalism n equations **??** is not fundamental in the same way as equations found in physics theories but it is built for an express purpose, to model the propagation of signals through a nerve. This model shows how quantitative thinking can lead to insights in biology. The sheer complexity of biology demands this of us. We cannot create complete theories so we must find useful formalisms for small domains of the problem space.

In this thesis we aim to do something similar, by building up from fundamental physics outlined in 2 we will build a model for the disfunction of a single gene (CFTR) to understand a disease (CF). Again, we do not possess sufficient computational power to produce a complete physical model of CF, so we will have to settle for a qualitative model which we will outline in 8.

In addition to the mesoscopic models ion channels spawned by hodgkin and huxley, there has been considerable interest in these systems from the early adopters of com-

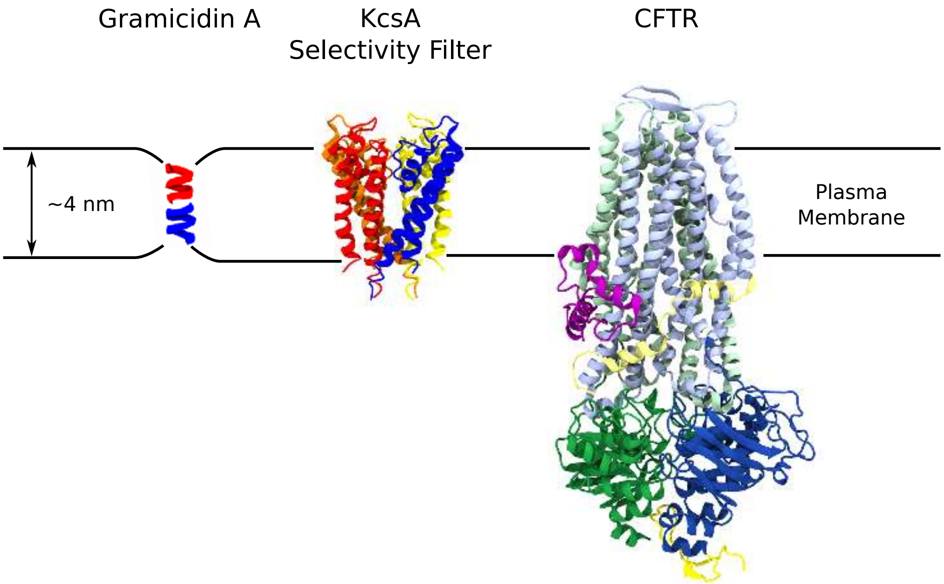

Figure 1.3: Different Ion Channels Ammenable to Molecular Simulation Gramicidin A was initially used as a toy model to test different *in silico* modelling techniques (PDB ID 1NT5) [48]. KcsA, is a bacterial potassium channel. This structure only comprises the pore domain, sometimes called the selectivity filter (PDB ID 1BL8) [49]. This structure drew interest because potassium channels are critical physiologically. Finally, the CFTR anion channel which this thesis is written about (PDB ID 6MSM) [50]. The fact that we have gone from simulating just a few nanoseconds of Gramicidin A to collecting almost half a millisecond of data in this thesis (a time span of 30 years) heralds an exciting future for computational biophysics [51]. What is even more exciting will be the next 30 years. At the time of writing, a computational engine named Anton 3 has come online. This purpose built computer could perform all the calculations in this thesis in less than a week [52]. That's not a week in parallel, that's in serial!

putational molecular biophysics. Biophysicists such as Martin Karplus, Benoît Roux, Shin Ho-Chung, Mark Sanson, Serdar Kuyucak and Toby Allen have devoted significant parts of their career to studying ion channels[37–42].

Early studies usually focussed on gramicidin A (gA) as a toy model for the diffusion of charged species. With the advances of this kind of modelling and experimetnal techniques the molecular details of the function of ion channels has become accessible to computational techniques.

The work on gA enabled careful studies of potassium channels [43–45]. Quite rapidly, the availability of protein structures and the maturation of these computational methods has enabled diverse studies of ion channels and other ion channel protein systems [46, 47].

Until recently, there have been a limited number of structural targets for biophysicists to work with. Initially, inquiries were limited to modelling gramicidin A, an antibacterial peptide which assembles into an ion channel in the cell walls of gram positive bacteria[53]. The mechanism of action for this little peptide is to simply destroy the bacterium's ability to hold an ion gradient. This causes the cell to dis regulate in all sorts of ways, such as the inability to produce ATP.

Gramicidin is one of the toy systems we use as biophysicists to develop theoretical methods. The others being the dialanine peptide, decalanine and sometimes ubiquitin [].

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

These factors have to allowed biophysicists sufficient data to build sufficiently accurate models of protein systems which generalise. Leading to a thriving field, analysing systems as diverse as photocells to gold nano particles CITATIONS NEEDED. 2

So by using ion channels as basic biophysical laboratories we can try to understand higher level protein physics []. In this thesis we will use that higher level understanding to develop a molecular theory of a diseae, Cystic Fibreosis. In future I hope we can use our understanding of protein physics to understand more complex diseases such as diabetes or neurodegenerative diseases. Maybe one day we can build this molecular theory of disease to do medicine with atomic precision.

1.5 Studying Cystic Fibrosis; Toward a Molecular Theory of Disease.

The sad truth of Cystic Fibrosis (CF) is that those afflicted are extremely unlucky. A single, change to the genome and their lungs fill with sticky mucus and become infected with bacteria, each breath becomes 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 [54]. Historically diseases have been diagnosed based on symptoms and not causes. As section **??** outlines, this is antithetical to how we would like to study physical systems. We would best understand the root cause of a disease so we can predict how to manipulate it. Discovering this root cause can be extremely difficult and often requires decades of clinical enquiry []. CF has the helpful characteristic of being a monogenic disease, so our molecular theory of this disease only needs to treat a single protein.

In this way, my motivations for studying the CFTR protein aren't solely focussed on treating disease. This problem is also an interesting opportunity to develop molecular theories of biology.

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 [55]:

$$W(\Delta G) \propto \exp\left(\left[-\frac{\Delta G - \Delta G_{opt}}{\sigma_{\Delta G}}\right]^4\right) + c \tag{1.6}$$

Here, W represents the evolutionary fitness of an organism, ΔG is the folding energy

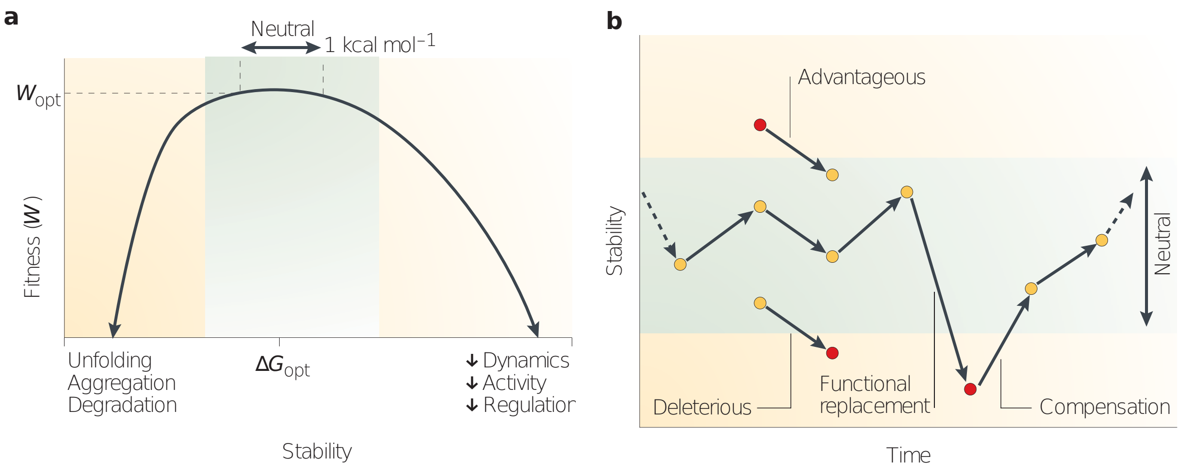

Figure 1.4: A physical model for how to view protein evolution
a)The stability of a protein ΔG is related to the fitness of an organism W by equation 1.6. This model gives rise to a peaked distribution which helps us understand how so many mutants can give rise to cystic fibrosis. It appears as though the CFTR gene has an exceptionally narrow $\sigma_{\Delta G}$ and so small changes to the sequence of the gene have a comparatively dramatic effect on the evolutionary fitness of the organism. b) The model in equation 1.6 Gives rise to a random walk through in sequence space for subsequent generations. In the case of CF it would appear as though *homo sapiens* are stuck at a specific snapshot in evolutionary time where CFTR may easily lose function. So, without gene therapies which can modify a gene sequence *in vivo* we must use a physics based model to somehow broaden the peak in equation 1.6.

of the protein with a given gene sequence and ΔG_{opt} is the folding energy of the protein in the average (fit) population. The parameter $\sigma_{\Delta G}$ controlls how broad this distribution will be and depends significantly on the protein physics of the gene. Figure **??** demonstrates the types of random walks of a gene through sequence space which this model predicts.

It just so happens that the CFTR gene exhibits an exceptionally small $\sigma_{\Delta G}$, so the band in the modified gaussian in figure **??** is narrow. This means that CFTR sits at the precipice of a daunting cliff in sequence space. This model is relevant to other diseases as well, such as sickle cell anemia, Alzheimer's disease, and Huntington's disease. So by taking small steps in sequence space and figuring out what factors have caused us to plunge down this cliff, we can try to understand how we might push the needle of protein stability back into the optimal zone.

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.

This theoretical model informs the conclusions of chapter 8.

Due to the array of disease causing mutations which occur across the cystic fibrosis protein, there is a large body of literature on its unique function [56]. This allows us a glance into its function and an opportunity to simultaneously perform basic biophysical research while directly assisting in furthering patient outcomes. This is the sort of inquiry which drives basic science forward, combining interesting experimental data into theoretical models to make testable predictions. The aim of this thesis is to build a model to make predictions about which kind of drugs will produce positive patient outcomes.

As we will see in chapter 3 the integration of basic biology into the treatment of cystic fibrosis has drastically improved patient outcomes. The opportunity of this thesis to shed light on the molecular details of this disease could lead to much greater patient outcomes.

1.6 The Future is Biological

We are on the cusp of developing biology from a descriptive to a predictive science [57–61]. This transition is driven by the combination of advanced of experimental techniques, rich datasets, strong theories and powerful computational engines. As an example of what will soon be possible. Biological systems has happened upon ingenious problems to some very difficult problems through evolution. It has much of the hard work for us and as we understand it better we can begin to apply its logic to our own problems [62].

Throughout science, the integration of experimental data with theoretical models leads to new and exciting research, this is particularly true in biology with its important applications in medicine, agriculture and increasingly, manufacturing [63, 64]. Wet lab biologists take advantage of experimental techniques which allow them to understand the dynamics and structure of living things from the top down. The finer the experimental instrument, the finer the detail they may resolve. Conversely, computational and theoretical biologists take a bottom up approach. We aim to take the granular details of a system, and integrate them upwards to model the macroscopic behaviour of that system. With more powerful computers and more detailed models we can make predictions about the behaviour of more complex systems. What is so exciting about the current era of biological research is that the domains of these two approaches are beginning to overlap, where they can synergize and drive further breakthroughs. As we discover more systems where this overlap can be found we will develop more sophisticated treatments for diseases and global problems [63].

The reason this has happened before in physics is two fold. The systems physicists usually study are are much more homogeneous. So it's much easier to integrate their formalisms upward. Once we understand the initial conditions and the formalism underlying the interactions in a system it is simply a question of whether or not we have the theoretical and computational capacity to predict the bulk behaviour of that system.

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

While previously, we were limited it functional data concerning ion channels we now have unprecedented resolution for the structure dynamics for the inside of a cell. The development of an array of experimental techniques[3] has allowed us to allow us to glimpse with unprecedented clarity, the salty dance of life inside our cells.

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

Armed with this philosophy we will delineate how to use the formal object of the Schröedinger wave equation to make approximations to atomic systems in order to create a biophysical model for macromolecular systems like proteins.

[3] Important biophysical techniques which one will encounter often in the literature are cryogenic electron microscopy (Cryo-EM) [65–67], electrophysiology [68], nuclear magnetic resonance (NMR) spectroscopy [69], confocal and fluorescence microscopy [70], X-ray Crystallography [71, 72], and genetic engineering (explanations of CRISPR-Cas9 or inverse PCR based techniques can be found in [73] and [74] respectively)

Chapter 2

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

> *Nature isn't classical, dammit, and if you want to make a simulation of nature, you'd better make it quantum mechanical, and by golly it's a wonderful problem, because it doesn't look so easy.*
>
> - Richard P. Feynman [75]

This chapter is written for somebody who has studied undergraduate physics and now wishes to model biological systems at the molecular level. Care is taken to dive deeper into the mathematical formulations of simulation methods than is conventionally given in introductory texts. Essentially, this is the understanding of simulation techniques I wish I had when I started studying them. An excellent overview which I would recommend as first reading for any new student can be found in an article by Braun et al. [76] followed by Gapsys et al. [77] for statistical rigour, and Pohorille et al. [78] for free energy calculations.

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

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

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

2.1.1 A full quantum mechanical treatment

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

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

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

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

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

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

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

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

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

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

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

Here the z_i represent the atomic number (and thus the charge) of the ith nucleus and q_e is the unit charge of the electron. The reason for the separate coordinates R_i and r_l

is to separate out the treatment of nuclei and electrons, which will be important once we apply the Born-Oppenheimer approximation.

Meanwhile, the kinetic energy terms are of the form

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

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

2.1.2 The Born-Oppenheimer approximation.

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

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

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

Fixing the nuclei in this way gives the "clamped nuclei" approximation [79]. To solve the wave function for the whole system Ψ_{tot}, we use an *ansatz* which decomposes the wave function with an electronic basis into two components: $(\psi_e)_k$ and $(\psi_n)_k$ which are the kth eigenfunction solutions to H_e and H_n, respectively

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

Note that there is an implied direct product between the wave functions $\psi_e(\mathbf{r}_l, \mathbf{R}_i)$ and $\psi_n(\mathbf{R}_i)$. When we substitute this expression into the full Schrödinger equation 2.1 we find the following expression for the kth nuclear eigenfunction [80]

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

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

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

Using the “adiabatic assumption” [80], the off-diagonal terms of C_{kj} can be set to 0 as they represent the interactions between the electrons and the nuclei. This completely decouples the wave function into two components

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

A further approximation is justified by ignoring the diagonal terms C_{kk} as they are 4 orders of magnitude smaller than the other terms in equation 2.8 [79].

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

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

The separation between atoms in a molecule is on the order of $10^{-10}m$, while the de Broglie wavelength of the nuclei at room temperature is on the order of $10^{-11}m$. Hence, we treat the nuclei as point particles at the scale of the full molecule. So, by rearranging equation 2.11 and taking the derivative with respect to time, we can see how to use Newton’s equations of motion to calculate the forces on the nuclei from the surrounding electric potential

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

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

2.2 Classical MD; Molecular Motions Without Quantum Mechanics

Following the Born-Oppenheimer approximation there are Hartree-Fock methods and density functional theory (DFT) which further simplify Schroëdinger’s equation. These more sophisticated physical methods allow us to simulate the organisation of electron clouds around small molecules, finding broad applications in chemistry and materials science [81]. These methods are known as *ab initio* MD.

However, even with these approximations, simulating a large number of atoms is still not computationally tractable. State of the art DFT methods can only simulate on the order of 10^3 atoms [82] and scales as $O(N^3)$ [83]. This is not sufficient to simulate proteins and their surrounding solvation environment where the molecular system is usually on the order of $10^4 - 10^6$ atoms. So, we must use another round of approximations to reach the spatial and time scales necessary to simulate biological molecules. We do this by creating a set of mathematical functions to simplify the calculations further. Here we use a set of virtual springs and other simple models for the energetic interactions between atoms. This creates what's known as an effective potential.

The CHARMM effective potential employed in all simulations in this thesis is similar to those found in all-atom classical molecular dynamics forcefields. The same functional forms are used in other forcefields such as AMBER, GROMOS and OPLS but with different parameters and design philosophies[84].

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

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

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

Interestingly, the bonded terms may be reasonably approximated by simple harmonic functions, with an exception we will discuuss shortly,

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

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

Over time several additions have been made to this formalism in order to reproduce experimental measurements of chemical systems [85]. We will discuss two of the major additions to the CHARMM formalism. The first is energy correction maps, abbreviated "CMAP corrections" concerning the dihedral parameters[86, 87]. These were motivated by the observation that MD simulations were not correctly reproducing protein secondary structure and resulted in errant Ramachandran distributions [86, 88]. Mathematically, these corrections take the form of a two dimensional grid which is used to interpolate results from *ab initio* quantum mechanics calculations for the torsion angles. Continued adjustments to $U_{dihedral}$ in the form of these CMAP corrections have resulted in substantial improvements to the CHARMM forcefield, particularly for disordered proteins [89].

There have also been additions of “non-bonded fix” or “NBFIX” corrections to the Lennard-Jones parameters in CHARMM. NBFIX parameters modify the Lennard-Jones parameters between *specific* atom types. For example, in order to reproduce experimental values of osmotic pressure for sodium chloride, the values of ϵ and σ are modified values *only* when calculating the Lennard-Jones potential between sodium and chloride atoms. These modifications fixed issues with certain charged species, such as sodium and chloride, which were associating too much in bulk solution or the overly stable salt bridges within proteins [89–92].

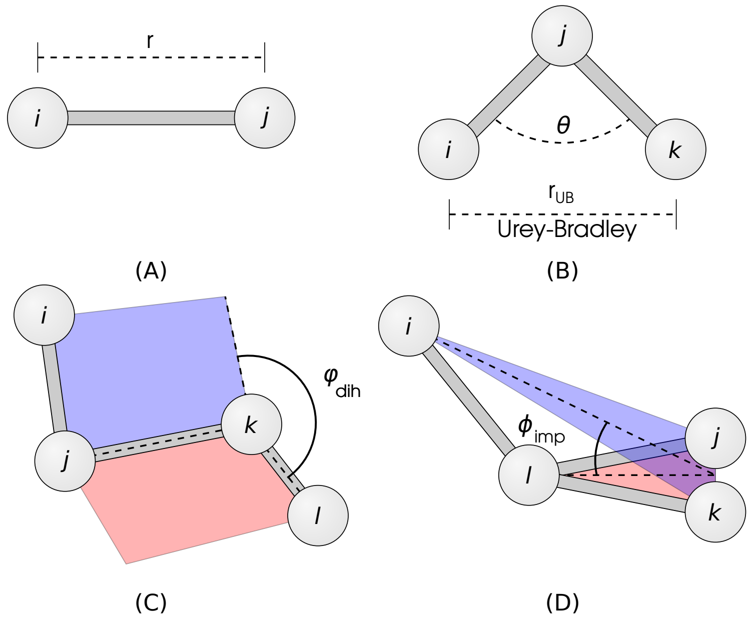

Figure 2.1: The Bonded Interactions Calculated In Classical Forcefields.
(A) The energy of Bond Stretching is approximated as a harmonic oscillator with respect to their separation r. (B) Angles between neighbouring covalently bonded atoms are also approximated as a harmonic oscillator with respect to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

Non Bonded Interactions

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

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

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

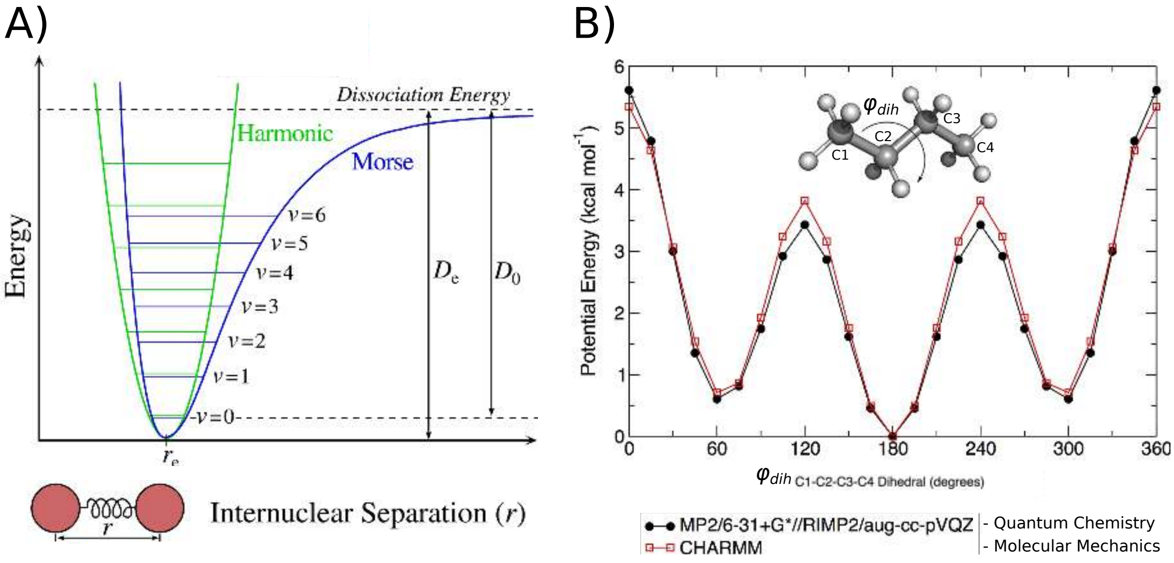

Figure 2.2: Comparison Between Potentials in Quantum and Classical Forcefields
A) The Morse potential was formulated to approximate the potential energy surface associated with the stretching of covalent bonds (blue). At low temperatures (the ground state, $v = 0$) like those found in classical MD there is good agreement between the Morse potential and the harmonic oscillator (green) (Credit, Mark Somoza 2006). B) Here the potential of the dihedral angle between the atoms C1,C2,C3 and C4 in a butane molecule is calculated using two methods: Quantum Chemical calculations and approximations using the functional form in equation 2.14 [84]. Note how the appropriate choices of k_φ, n and δ have closely approximated the results in the more accurate quantum mechanical calculations.

optimum distance that two atoms will rest against each other in the absence of other effects. The ϵ parameter denotes the depth of the potential well, or how stable the two atoms will be in the minimum energy configuration. This is very important for certain physical parameters such as osmotic pressure [90].

Meanwhile, the partial charge assignments q_i for each atom are very important in a biological context, for stability of protein conformations of salt bridges and the solvation energy of different molecules [93].

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

2.2.1 Philosophy of Different Molecular Mechanics forcefields.

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

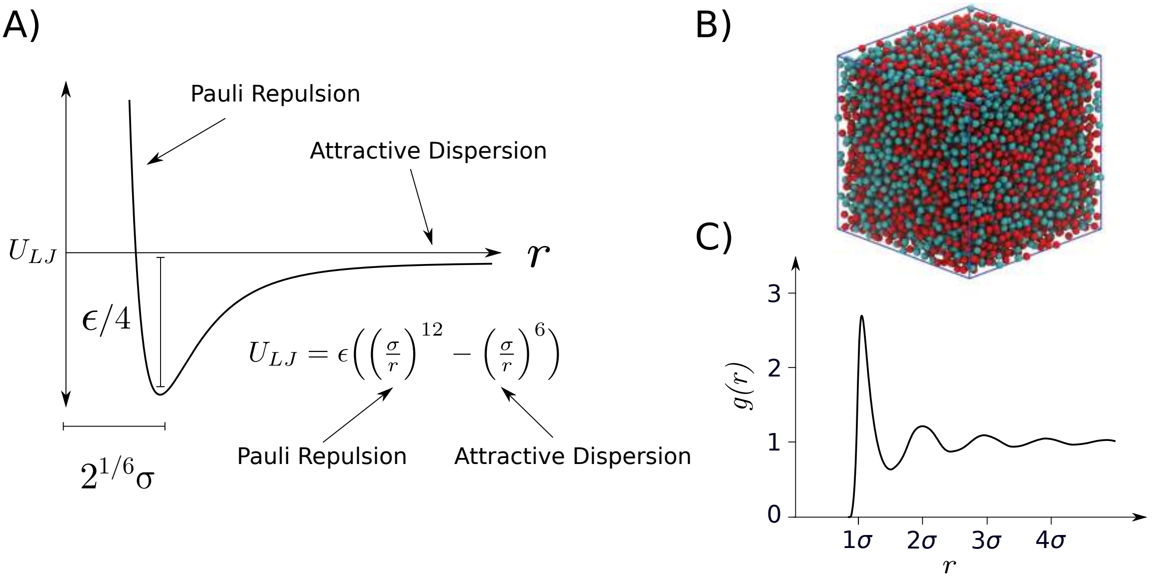

Figure 2.3: The Lennard-Jones Potential
A) The Lennard-Jones potential function has two regimes, the far region one dominated by attractive dispersion forces and the close region dominated by repulsion. In the case of atomic systems this is due to the Pauli exclusion principal. B) An example of a fluid modelled with Lennard-Jones particles [94]. C) The radial distribution function (g) for a Lennard-Jones fluid [95]. Note that the peaks in the distribution are spaced roughly 1 σ apart.

from *ab initio* MD calculations as an initial guess and approximate them by tweaking the parameters in the functional form in equation 2.13. Conversely, top down forcefields tweak these parameters with reference to experimental measurements. Ultimately, the results from *in silico* experiments must match those of wet lab experiments, so the development of all forcefields has elements of both philosophies. All forcefields assign the partial charges to atoms using the results of *ab initio* MD calculations. The rest of the parameters are derived with other methods such as attempting to match the known secondary structures of peptides [89]. Different forcefields have different methods of deriving their parameters. Below is a short summary of the philosophy for the four major forcefields taken from the review by Justin Lemkul [84].

- CHARMM: The most popular all atom forcefield. To build CHARMM, QM optimized geometries and molecular dipole moments are compared to those found from calssical MD simulations. Molecular degrees of freedom such as dihedral angles are also fit with QM energy profiles, an example can be seen in 2.2. Macroscopic experimental quantities are also used to validate the parameters in this forcefield, such as solvation energies, crystal geometries, heats of vaporization and conformational sampling of biomolecules [89]. One of the reasons for this forcefield's popularity is its considerable library of supported compounds, especially when combined with its generalisation module CGENFF [96].

- AMBER: An all atom forcefield which is built from the ground up from results from quantum mechanical calculations and results from spectroscopy. A version of this forcefield has become the favorite for the simulation of disordered proteins [97, 98]. There is also a generalised version of the AMBER forcefield known as GAFF [99, 100]. Comparisons between generalised forcefields can be found in [101].

- OPLS: An all atom forcefield. The OPLS forcefield takes the philosophy that, since many biomolecules share similar geometries to certain organic liquids, biomolecules can be accurately parameterised by creating a forcefield which correctly reproduces the experimental measurements for these species. Parameters are derived to accurately reproduce the liquid density of certain organic liquids. These parameters are then used as roots to construct larger biomolecules by drawing analogies between similar molecular geometries.

- GROMOS: A united atom forcefield, where hydrogen atoms are typically merged into the heavy atom they are bound to. Hence, they are not explicitly treated. Charge assignment is done with DFT. Interestingly, GROMOS uses a quartic form of the bond stretching term

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

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

2.3 Periodic Boundaries to Simulate the Inside of a Cell

Inside cells, proteins are immersed in a large solvation environment composed of water and salts [3]. An example can be seen in figure 2.4. However, simulating such a large environment is computationally expensive and truncating it with vacuum at the boundaries leads to water molecules aligning their dipole moments along the boundary and perturbing equilibrium dynamics. In order to avoid such artefacts, we have to replicate the large cellular environment somehow [102]. We could make a simulation box large enough to replicate the behavior of a bulk solvent, but even with a large simulation box we can still observe artifacts associated with the vacuum at the boundaries [77]. So, to avoid these boundary effects, we use periodic boundary conditions (PBCs), allowing atoms to move between images in the simulation box. This replicates the molecular system infinitely in every direction.

Using PBCs might remove vacuum from our molecular system but now we have a different problem. Effectively, with the PBCs, we have created a system with an infinite number of atoms. We have to somehow limit the number of computations we perform. We could simply truncate the calculation of interactions $U_{non-bonded}$ after a certain cutoff distance. This is not an issue for U_{LJ} because the $1/r^6$ and $1/r^{12}$ terms

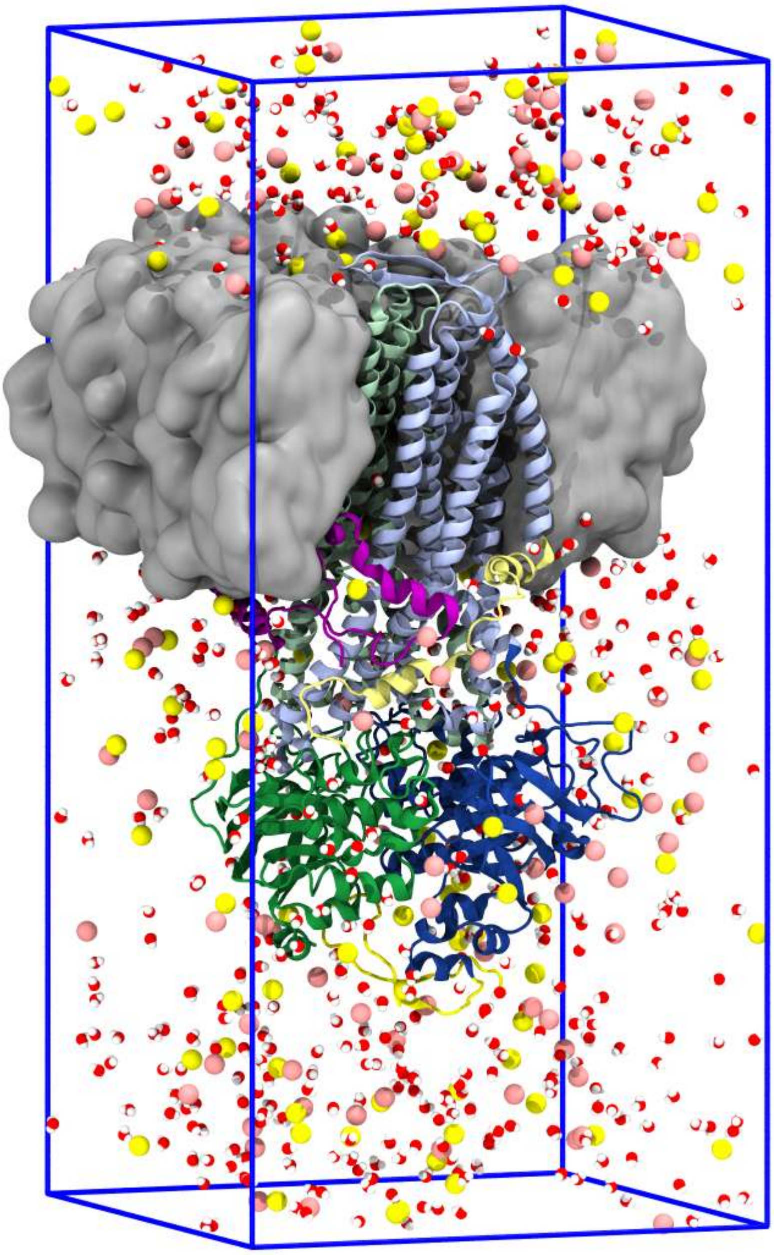

Figure 2.4: An Example of a Solvated Biomolecular Environment Ready for Simulation
This rendering shows a CFTR protein embedded in a lipid bilayer, immersed in a potassium chloride solvent. Half of the bilayer has been stripped away, as has much of the solvent so the protein is visible. The phospholipid lipid bilayer is colored grey, while the potassium chloride ions are colored pink and yellow, respectively. Water molecules are red and white. The blue box indicates the boundaries of the unit cell. This system contains roughly 190,000 atoms in total.

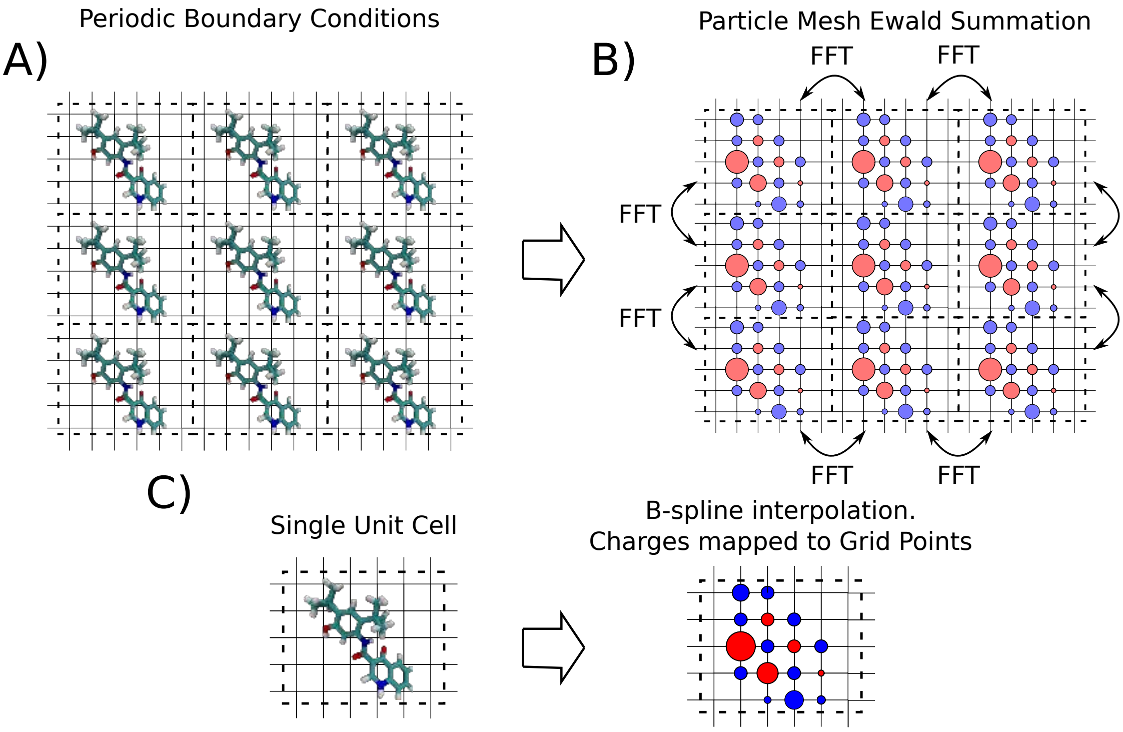
Periodic Boundary Conditions
A)
Particle Mesh Ewald Summation
B)
FFT
FFT
FFT
FFT
FFT
FFT
FFT
FFT
FFT
FFT
C)
Single Unit Cell
B-spline interpolation.
Charges mapped to Grid Points

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

in equation 2.3 decay very quickly for large r. Inaccuracies due to this approximation can be further ameliorated with the use of a smooth switching function [103, 104]. On the other hand, the $1/r$ dependence in $U_{coulomb}$ scales much more slowly so truncating it leads to a loss of accuracy in the results of the simulation [105–109]. So, we have to calculate the contributions of $U_{coulomb}$ in between all periodic images. Note that this periodicity requires that the unit cell is electrically neutral, else the contribution of potential energy from $U_{coulomb}$ will be infinite, leading to artefacts [110].

To calculate $U_{coulomb}$ in all periodic images and limit computational intensity of our calculations we use a clever scheme known as Particle-Mesh Ewald summation (PME). Interestingly, this scheme ends up scaling better than the pairwise summation in equation 2.15 might imply. The direct summation scales with computational complexity of $O(N^2)$ with the number of atoms while the infinite PME scheme scales as $O(N \log N)$ [111], though there are some further considerations for large systems on parallel architectures [112]. Even with these sophisticated algorithms, the calculation of electrostatic potential in the infinite system still represents the largest computational bottle-neck in classical MD [112].

For a detailed review of different Particle-Mesh Ewald summation methods and the mathematics behind the method see [113]. A brief outline of the Smooth Particle

Mesh Ewald summation is given below

1. An *ansatz* is used where $U_{coulomb}$ has Gaussian screening charges added to it and simultaneously subtracted away to create a smooth potential. The details can be found in [113].

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

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

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

2. The charges to be evaluated using U_{FT} are interpolated onto a grid using B-spline interpolation functions. This is the procedure demonstrated in figure 2.5.

3. The charge density functions for the charges on the grid are transformed into frequency space using Fast Fourier Transforms.

4. The Poisson equation is solved numerically in frequency space for these charges.

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

   Where $\tilde{U}$ is the component of U_{FT} we solve for in frequency space and $\tilde{\rho}$ is the Fourier transform of the smooth scalar function for the interpolated charge densities.

5. An inverse Fourier transform is calculated for the solution to $\tilde{U}$ to transform it back into real space.

6. The interactions in $U_{direct-sum}$ are evaluated using a simple pairwise summation.

7. Now that $U_{coulomb}$ is known at every position in the unit cell, we can move atoms according to the contributions from this potential using Newton's second law.

2.4 Controlling the Temperature and Pressure in a Simulation

Living things are very sensitive to their external environment. Enzymes only work in a narrow range of temperatures and cells burst apart in the absence of pressure[114, 115]. As such, to correctly understand biological systems we not only need to simulate the dynamics of the atoms inside them, but we must also make sure that the virtual environment in our simulations matches what is found inside cells or in the laboratory. Our simulations should seek to approximate the environment of an open topped test-tube sitting in a pressure and temperature controlled laboratory. To do this, we make use of some statistical ensembles chosen for their performance in regulating the thermodynamic quantities in a simulation and their computational expense.

2.4.1 Hot and Cold with the Nosé-Hoover Thermostat

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

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

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

Despite starting from the same Cartesian coordinates, randomly sampling velocities from the Maxwell-Boltzmann means that replicate simulations will immediately begin from different points in phase space. Their coordinates will quickly diverge, raising questions around how long one should run a simulation and how many replicates they should run in order to collect reliable statistics. According to Knapp et al. [116], a good rule of thumb is to simulate between 5 and 10 replicates depending on the availability of computing resources[1]. In this thesis we prioritised long time scales to sample slow motions, so 3 replicates with run times between 1 and 2 microseconds were produced for all systems.

After the initial choice of velocities, the temperature in a simulation is maintained by directly modulating the velocities of the atoms to maintain the target temperature T_0. There are many schemes which attempt this. We will discuss the Nosé-Hoover thermostat in detail here because it was used during the production runs in this thesis [117–119]. However, for the equilibration phase of the simulations, we used the Berendsen thermostat because it is faster at correcting large temperature differentials but does not produce the correct statistical ensemble [120, 121]. We also note that the field has since moved on to favor the Bussi thermostat, which is an extension of the Berendsen thermostat, as it works well in most contexts [76, 120].

The Nosé-Hoover thermostat is characterised by the use of an extra, massive particle coupled to an external bath. The use of a single bath has been associated with issues with ergodicity and so usually this particle is coupled to a chain of external baths. Usually, the software simulation package GROMACS uses a chain of 10 baths, $M = 10$ [119, 122, 123]. The Hamiltonian for the Molecular Dynamics system coupled to a chain of M external baths is then

[1] Recall that in the ergodic limit, a quantity f calculated from the ensemble average of replicates will be equal to the time average calculated from one replicate in the limit $t \to \infty$. That is

$$\langle f \rangle_i = \lim_{t\to\infty} \frac{1}{t} \int_0^t f(\tau) d\tau$$

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

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

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

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

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

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

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

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

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

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

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

The Nosé-Hoover thermostat, when chained infinitely, allows us to accurately produce what's known as an NVT ensemble, also called the canonical ensemble in the statistical mechanics literature [119]. Where the number of particles in the system (N) remains constant, the volume of the system remains constant (V) and the temperature remains constant (T). In a realistic environment, the pressure remains constant rather than volume, so we need another regulatory mechanism to modulate the volume of the system to regulate the pressure (P) and produce an NPT ensemble.

2.4.2 Under Pressure with the Parinello-Rahman Barostat

Pressure is critical to the function of living organisms. Membranes burst apart at low pressures [124] and at high pressures cellular function is disrupted [125]. In order to accurately reflect the atmospheric pressure at which living things thrive we have to accurately calculate and modulate it during our simulation.

In order to measure the pressure at the simulation walls, we follow the procedure in [126] by calculating a quantity known as the virial:

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

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

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

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

For this purpose, we apply the Parrinello-Rahman barostat [127, 128], using a procedure with a similar philosophy to the extended Hamiltonian used in the Nosé-Hoover thermostat. In this case, the system is coupled to an external pressure bath rather than an external temperature bath. First we define that the basis vectors for the periodic simulation box as $\underline{h} := [\mathbf{a}, \mathbf{b}, \mathbf{c}]$. When the box is scaled to change the volume these basis vectors are multiplied by a set of scalars $s_i := (\xi_i, \eta_i, \zeta_i) \in [0, 1]$. We perform a change of coordinates so that the contributions of the particles onto the boundaries is easily calculated from our equations so we express the atomic coordinates as

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

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

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

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

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

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

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

represents the stress tensor which acts across each of the faces of the unit cell. This system of equations can be solved numerically to control the pressure of the simulation system by modulating the length of the basis vectors $\mathbf{a}, \mathbf{b}$ and $\mathbf{c}$ contained in $\underline{h}$.

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

2.5 The Process of Preparing an MD Simulation

The process of taking a molecular structure and putting it in a cellular environment to simulate it at physiological temperatures is both an art and a science. It's a science because a biophysicist must be aware of the many tricks that structural biologists use to image a macromolecular complex. But it's an art because accounting for those tricks and making the necessary modifications is rarely straightforward. How do you build a missing loop? What charge state is an amino acid most likely to take in a physiological context? These questions must be carefully answered by analysing the literature about the system of interest. Once the protein structure has been built, the system is immersed in a water solvation bath alongside a concentration of salt ions, usually sodium chloride if the environment is thought at be extra cellular or potassium chloride if the environment is thought to be intracellular. Once the initial conditions have been decided, we need to make a few preparatory steps so the simulation collects reasonable results and doesn't hurtle along some unphysical trajectory. The steps so that the simulation remains realistic are fleshed out in some more detail in [76] but we produce a short summary below.

1. Minimisation: Here the atoms are moved down along U_{CHARMM} to resolve any clashes in the system which would cause LINCS or SHAKE to diverge. This is usually done with simple minimisation algorithms such as steepest descent or conjugate gradient descent [129]. This is usually done with a constant volume.

2. Relaxation: Harmonic restraints are placed on the heavy atoms (non hydrogen) in the system so that large conformational changes do not occur while the macromolecules are heated and settle into their solvation environment. This may be done under the NVT or NPT ensembles depending on the system. Sometimes the system is heated slowly from 0 Kelvin up to the desired thermal temperature in order to avoid large conformational changes which might result from different parts of the system heating at different rates.

3. Equilibration: Often after relaxation more simulation time is run so that the system can settle into local minima further. This process makes sure that the physically relevant local minima are being sampled once we move to production. This process could be run for a few nanoseconds or up to a microsecond. It depends on the system. This is usually done with the NPT ensemble.

4. Production: Here the NPT ensemble is applied and the system is allowed to evolve under Newton's equations of motion while data is collected for analysis.

2.6 Choosing an Appropriate Time Step

The discrete time step, Δt which is used to integrate our equations of motion is one of the most important determinants in the performance of the simulation. We would like Δt to be as large as possible, so that the minimum number of calculations are made to sample the desired time scale. In the case of proteins this usually runs between 10^{-6} and 10^{-3} s [97].

As you can see in table 2.1 the fastest motion in molecular systems is dictated by stretching of covalent bonds. Studies of the resonance of molecules by infrared spectroscopy determined that the O-H type bonds oscillate the fastest, with a resonance peak at 3600 cm^{-1}[130].

Due to Nyquist's theorem the largest Δt parameter we can choose *must* be less than half the speed of the fastest degree of freedom in the system [131]. However, empirically we have found that condensed matter systems require even shorter time steps to maintain their stability [129]. The Verlet leap-frog scheme used in most MD codes requires between 5 and 10 integration steps per period of the fastest harmonic mode in a system, to maintain stability [132, 133]. The choice of too large a time step means that the system will escape local free energy minima, accumulating kinetic energy and eventually "blow-up" [76]. 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

Motion	Timescale
Covalent Bond-stretching	$1-2\times10^{-15}$ s
Covalent Bond-angle bending	$5-10\times10^{-15}$ s
Sidechain Motions	$10^{-12}-10^{-6}$ s
Rigid Body Motions	$10^{-9}-1$ s
Ion Conduction	$10^{-9}-10^{-6}$ s
Protein Conformational Changes	$10^{-9}-10^{-3}$ s
Alpha Helix Formation	$10^{-9}-10^{-6}$ s
Beta Sheet Formation	$10^{-6}-10^{-3}$ s
Protein Folding	$10^{-6}-10$ s

Table 2.1: Timescales of Motions in a Molecular System
The time step of a simulation must be small enough to capture the motions in the fastest degree of freedom. In hydrogen-rich biomolecular systems the bottle neck can be found in the fast bond vibrations in lighter atoms. This stands in tension with the phenomena we are interested in on longer timescales such as protein folding. Sources: [129, 130, 133, 138–140]

to get away with using 2 femtoseconds with the use of specialised integration schemes such as SHAKE[134] and LINCS[135] to constrain the fast motion of hydrogen atoms. This allows us to use $\Delta t = 2$ fs in atomistic classical MD simulations.

The use of techniques such as hydrogen mass repartitioning [136], virtual site topologies [133] and multiple time step schemes [137] have also gained popularity in recent years in order to increase time steps further, up to $\Delta t = 5$fs.

2.6.1 Verlet Leap-Frog Integration

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

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

We can use this calculation of acceleration a_i of the ith atom to update the positions and velocities of the atoms in the molecular system with the following triplet of equations known as the leap-frog Verlet method [130]:

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

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

In MD, we are less concerned with the accuracy of a particular trajectory so much as collecting sufficient statistics to calculate macroscopic properties such as free energies or diffusion profiles. This means the choice of 4th order solvers such as the Runge-Kutta method would be inappropriate. Although they may use a large time step, they require 4 evaluations of U_{CHARMM} per iteration and are thus more expensive than any second order method. Hence, we prefer symplectic (energy preserving), 2nd order methods such as Verlet integration so the simulation remains stable after millions of time steps [137].

2.7 Free Energy Calculations: Making Simulations More Useful

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

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

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

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

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

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

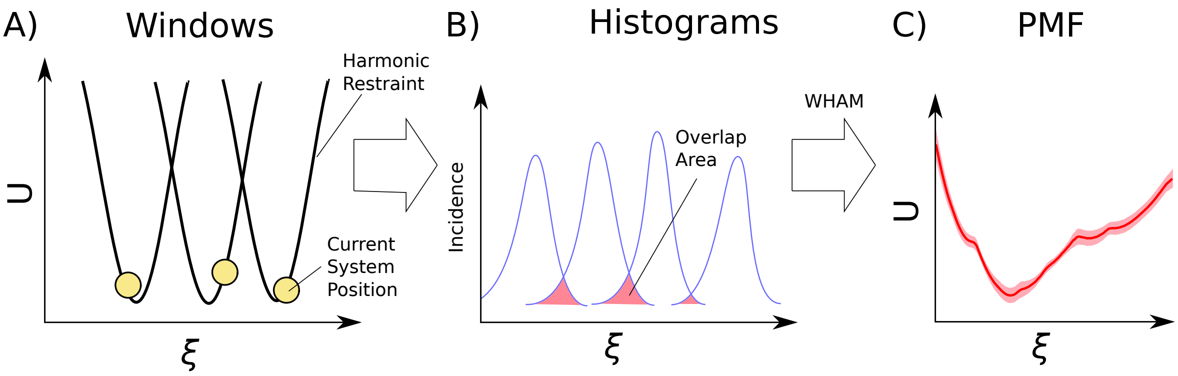

Figure 2.6: Illustration of Umbrella Sampling
A) Several simulations are repeated with only one change. A bias potential is added somewhere along the reaction coordinate ξ. B) The value of ξ is recorded in each of the windows and then graphed as histograms. C) The Overlap in neighbouring histograms is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

The functional form of U_{bias} depends on the free energy calculation being employed. There are two varieties of techniques, equilibrium and non-equilibrium methods. We will focus on the equilibrium methods in this work. We note that there is another set of methods called alchemical methods which modify the chemical composition of the system which we will not cover. An extensive, frequently updated overview of modern free energy methods can be found in [141].

2.7.1 Umbrella Sampling

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

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

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

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

Umbrella sampling is extremely useful for calculating all sorts of experimental quantities and physiologically relevant properties such as folding energies [142], lipid binding [143], ion conduction [144], and drug binding [145]. However, it is particularly sensitive to the choice of initial configuration and collective variable [143]. The former issue is particular to umbrella sampling because generally short runs are used since so many windows are spawned during the method, the simulations must be at equilibrium *before* the method is attempted, and then sufficient statistics must be collected to average over any conformational changes orthogonal to the collective variable. In these ways, care must be taken when using this method to not introduce systematic error into the calculation [146]. A deep knowledge of the molecular system under investigation can help alleviate some of these issues.

Weighted Histogram Average Method (WHAM)

There are a few candidates for calculating a PMF using the statistics calculated in umbrella sampling. The Weighted Histogram Average Method (WHAM) [147], Umbrella Integration (UI) [148] and the Multistate Bennett acceptance ratio (MBAR) [149] are all used. We will briefly outline the mathematical formulation and estimation of errors of the WHAM method as it is more popular [144]. Our explanation follows [150] which covers the topic in more detail.

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

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

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

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

where the weights, $\sum_{i=0}^N p_j = 1$ are chosen such that the statistical error is minimised accross the domain of ξ [151]. This means we solve an optimisation problem for the variance of the unbiased estimates

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

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

Essentially we are vertically shifting the estimates obtained in each of the histograms in order to minimise the error across the PMF. By convention, we can estimate the error in the PMF by splitting up the statistics collected for the distributions of $P_i^b(\xi)$ and constructing PMFs from these independent samples. For example, we might have collected 100 ns of data in each window, but we could construct 5 independent estimates for the PMF using 20 ns blocks of trajectories. The Standard Error of the Mean (SEM) can then be used to estimate the error across the surface [77]. By convention, an umbrella sampling calculation is said to have converged when the SEM from these independent samples has fallen below 1 kcal/mol across the PMF. Note that there may be sources of systematic error not captured by this criterion.

2.7.2 Metadynamics

Developed in the lab of Michele Parrinello [153], metadynamics has proven to be a popular method in many applications of computational science, not just in molecular dynamics simulations of biomolecular systems [154–156]. A pedagogical review of the method and its variants written by Pratyush Tiwary can be found in [152]. The method relies on a time dependent form of U_{bias} given by

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

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

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

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

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

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

where V_ξ is the volume of the phase space spanned by ξ. However, in practice the function $U_{bias}(\xi)$ is simply inspected at intervals for fluctuations about an average function [157].

There are many flavours of metadynamics. The most popular is well-tempered metadynamics which gradually reduces the Gaussian height B as the simulation progresses[158]. In theory, this guarantees convergence with vanishingly small error. However, this

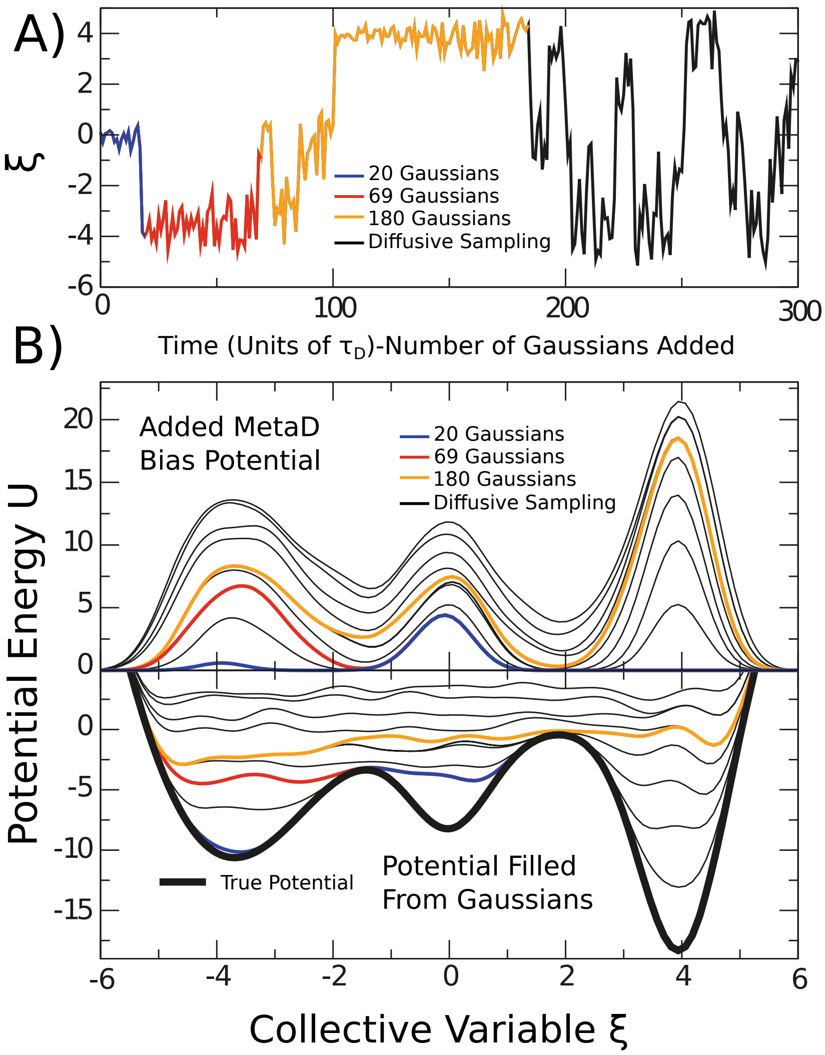

Figure 2.7: Illustration of Metadynamics
A) The trajectory of a collective variable ξ in a metadynamics simulation. The blue region denotes the trajectory up to the time where 20 Gaussians have been deposited. After which the basin at $\xi = 0$ has been filled. The red data points denote the time up until 69 Gaussians have been deposited, after which time the second basin at $\xi = -4$ has been adequately sampled. Finally the Orange data indicates that the 3rd basin at $\xi = 4$ has been adequately sampled after which which time the system begins to sample ξ diffusively and the simulation has converged. B) The upper panel demonstrates the repulsive potentials that have been added to the simulation in order to drive it to new, unexplored regions. Multiplying the values of this function by -1 will be the estimate of the free energy surface[2]. Note how in the lower panel the true potential energy surface is gradually filled by the added Gaussians functions. Source [152].

method requires an estimate of how long the simulation will take to converge. There is also infrequent metadynamics which can be used to estimate the diffusion profile along ξ [141, 159–161].

A useful feature of the Metadynamics method is that it can be linearly sped up with the number of parallel simulations. This is known as multiple walker metadynamics [162]. Since we are simply attempting to sample from the same potential U_{CHARMM} up to a desired time, until the criterion in equation 2.42 is met, we can speed this process up by running simulations in parallel and adding Gaussian to the same U_{bias} function.

Conceptually, metadynamics performs the same role as umbrella sampling and should in theory produce the same results in the same system with the same collective variable ξ. However, it has some specific contexts where it outperforms umbrella sampling. This method is more suited to an exploration of free energy space where it can intelligently explore regions which are poorly sampled, whereas umbrella sampling requires some foreknowledge of the surface being investigated in order to guess which parts of the landscape require more sampling. However, metadynamics can be very difficult to converge. A good indicator of such systematic errors are the presence of unphysically large barriers, indicating that there are orthogonal degrees of freedom not sampled along ξ which might correspond to a minimum energy pathway. These barriers will eventually come down in the infinite sampling limit but many barrier crossings will need to be observed. A discussion of how to solve these systematic errors can be found in [156]

2.8 Short Comings of Classical MD

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

2.8.1 The Problem with Forcefields

The approximations inherent in equation 2.13 are not without a cost to accuracy. In certain situations, many of which are biologically relevant, it has been shown that polarisation effects not captured by fixed partial charges play an important role in the dynamics of the system. This has been demonstrated in the literature for Gramicidin where polarisable forcefields are able to more accurately reproduce the experimental measurements of current [163].

The other context where polarisation is important to consider involve divalent ions such as calcium or magnesium. Here, the highly charged environment near a divalent ion will induce changes in the dipole moment of surrounding atoms. This is not possible in the fixed charge formulation in equation 2.13, making investigations of these biologically important chemical species difficult [164, 165].

However, for most situations, particularly those involving bulk water with a low concentration of solute, classical forcefields are sufficiently accurate to sample conformational motions of biomolecules [166]. Sadly, it should also be kept in mind that classical MD is not able to simulate any chemistry such as forming and breaking of bonds or a change in the protonation state of an amino acid. Such interactions require considerations of quantum mechanics which are computationally expensive [167].

There are several efforts to address some of the above issues. Some groups are trying to improve the accuracy of classical forcefields using machine learning and Bayesian inference [168]. But there are also attempts to move beyond the functional form of equation 2.13 by explicitly including the effects of polarisation. The most popular methods at the moment are adding a massless drude oscillator as an extra bead to atoms as in the CHARMM drude forcefields, championed by the Mackerell lab [169], and explicitly calculating the dipole and quadrupole moments of each atom as in the forcefield AMOEBA [170]. 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 [163, 170, 171].

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

2.8.2 The Problem with Sampling

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

The slow time step demanded in classical MD due to the fast motions of certain atomic groups such as hydrogen is fundamentally at odds with the time scales of many important biological processes such as drug binding or protein folding which occur on the time scale of milliseconds or seconds.

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

The success of simulations at the millisecond timescale by D.E Shaw research suggest that we are in reach of an exciting area in biological research [97, 172]. Enhanced sampling methods will be able to routinely reach motions that occur on these time scales and as software and hardware improve we will be able to push further, to simulate larger systems.

The advances we are seeing at the moment which I find exciting are the use of machine learning methods to tease out these degrees of freedom in order to accelerate them with already established enhanced sampling methods. That is, make a more careful choice of the collective variable ξ. This could be done with a variety of algorithms such as Time lagged independent component analysis (TICA) [173, 174] from Frank Noë, a variational approach to conformational dynamics (VACs) from Michelle Parinello's laboratory [175] or Reweighted autoencoded variational Bayes for enhanced sampling (RAVE) [176] from Pratyush Tiwary's laboratory. These algorithms have the potential to build on the above rigorous physics of simulations and revolutionise our understanding of biomolecular systems.

2.9 Conclusion

It is hoped that the preceding chapter can serve as a roadmap for any physicists interested in beginning to study the exciting field of computational biophysics. The information in each section should serve to help the reader understand the foundations of how simulations are performed. The next steps would be to learn more about the molecular biology and biochemistry of macromolecules so they can better understand the chaotic dance occurring inside cells. For recommended reading on this topic there is Schlick [130], Frauenfelder [71] and Phillips [3]. There is a lot to learn and the barrier to entry can seem daunting. The reader is encouraged to join mailing lists and other forums where the thriving computational chemistry and computational biophysics communities communicate. Send cold emails asking for help when you are stuck, remember to consult software manuals as well. Below is a list of software to get you started building and running simulations.

- Python [177]. A versatile programming language that will help with all parts of a computational workflow.
- Visualised Molecular Dynamics (VMD) [178]. Molecular visualisation software with a scripting language, great for building simulation systems, viewing trajectories and rendering figures.
- CHARMM-GUI [179]. A web server which offers a great starting point for constructing MD simulation systems and also offers configuration scripts for all the major MD software packages.
- MDANALYSIS [180, 181]. A python package with a diverse set of tools to analyse molecular dynamics trajectories and structures.
- Statistical Biophysics Blog A personal blog by respected computational biophysicist Daniel M. Zuckerman. Here you will find some simple theoretical exercises and some helpful personal advice, not just technical.

- MODELLER [182–184]. A python package which helps build homology models of protein structures or add missing parts to protein models.
- NAMD [185]. A powerful, scalable user friendly MD simulation package.
- GROMACS [123]. A less user-friendly MD simulation package than NAMD but offers some different features and (at the time of writing) faster performance.
- OPENMM [186]. A flexible MD simulation package that is optimised for usage on GPUs. Is controlled with a python API. A good choice for desktops and workstations.
- PLUMED [187]. A plugin for existing MD simulation packages which lets the user define their own collective variable ξ. Very useful for sophisticated analysis and free energy calculations.
- Protein Data Bank (PDB). An open source database of all published biomolecular structures at atomic resolution.
- Uniprot [188]. A database containing a wealth of biochemical data and annotations for proteins, very handy when starting a new project.
- ALPHAFOLD2 [61]. A highly accurate AI generated database of proteins. This was considered a watershed moment in the field when it was released.
- CHARMM forcefield [89]. The parameter files for the charmm forcefield formatted to be read by popular MD simulation packages.
- AMBER forcefield and simulation package [189, 190]. An MD simulation package and toolkit. Here you will also find instructions for how to extract and use the latest AMBER forcefield in other MD simulation packages.
- PARMED [191]. A python package for converting and manipulating MD file types.

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

Chapter 3

Review of the Molecular Cause of Cystic Fibrosis and Its Treatment

Because of what's inside me; Because of my genes.-Bob Flanagan [192].

3.1 Clinical outcomes of Cystic Fibrosis

Cystic Fibrosis (CF) is the most common fatal genetic condition in Caucasian populations. 165 000 people are estimated to be afflicted globally [193]. 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[194]. The symptoms of the disease are due to the inability of epithelial cells to regulate their salt content. .

When dehydrated the cilia on the epithelium collapse leaving them unable to clear the mucus that naturally lines the airway[195]. 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 require hours of physical therapy each day to help clear this mucus since their lungs are unable to. They must also inhale saline solutions in order to counteract the osmotic pressure in their epithelium. This helps draw more moisture out of the epithelial cells to allow the cilia to move.

CF patients struggle to intake nutrients due to the build up of mucus in the ducts of their pancreas and large intestines. This leads to CF related diabetes which afflicts roughly half of adults with CF [196]. Patients with CF related diabetes are often administered enzymes and must adhere to a specific diet.

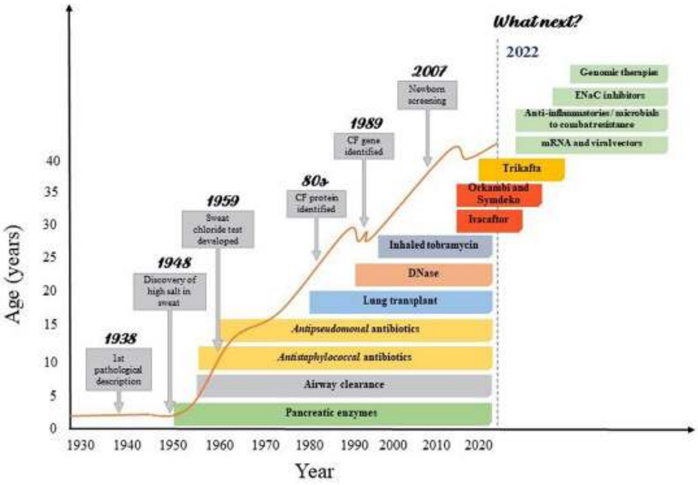

Figure 3.1: CF Clinical Progress
Life expectancy of CF patients correlates highly with translational research. Source [197]

3.2 CFTR Structure

CFTR is composed of one chain with pseudo-symmetric structure, the protein is well organised into 7 domains ??. In the order of their primary structure they are:

1. The Lasso motif (AA 1-68). Anchors into the membrane and serves as an interaction hub with protein partners such as syntaxin and filamin which are important in cellular trafficking [198–200] as well as WNK1 which plays a role in bicarbonate selectivity [201].

2. Transmembrane Domain 1 (TMD1 AA 69-376). This domain forms half of the chloride conducting pore and importantly, TM1 and TM6 in this domain form the extracellular end of the pore for anion permeation [202, 203].

3. Nucleotide Binding Domain 1 (NBD1 AA 377-629). One of the ATP binding sites, this domain has a dense concentration of disease causing mutations, including the most common mutation $\Delta F508$ [204].

4. Regulatory Domain (R-domain AA 630-855). A disordered domain containing up to 11 phosphorylation sites[205]. In the inactivated conformation a helical segment of this domain wedges between the TMDs. Upon binding of PKA and phosphorylation the wedge relocates to a location just below the R-domain. The identity of a fragment of the R-domain is analysed in detail in chapter **??**. The kinetics of this domain is important to the overall function of CFTR [205, 206].

5. Transmembrane Domain 2 (TMD2 AA 856-1168). This domain forms the other half of the chloride conducting pore. There is ongoing controversy over the structure and function of TM8 the function of CFTR [207, 208].

6. Nucleotide Binding Domain 2 (NBD2 AA 1169 - 1450). Home to the conserved Q-loop, which plays an important role in the binding of ATP in ABC transporters [209–211].

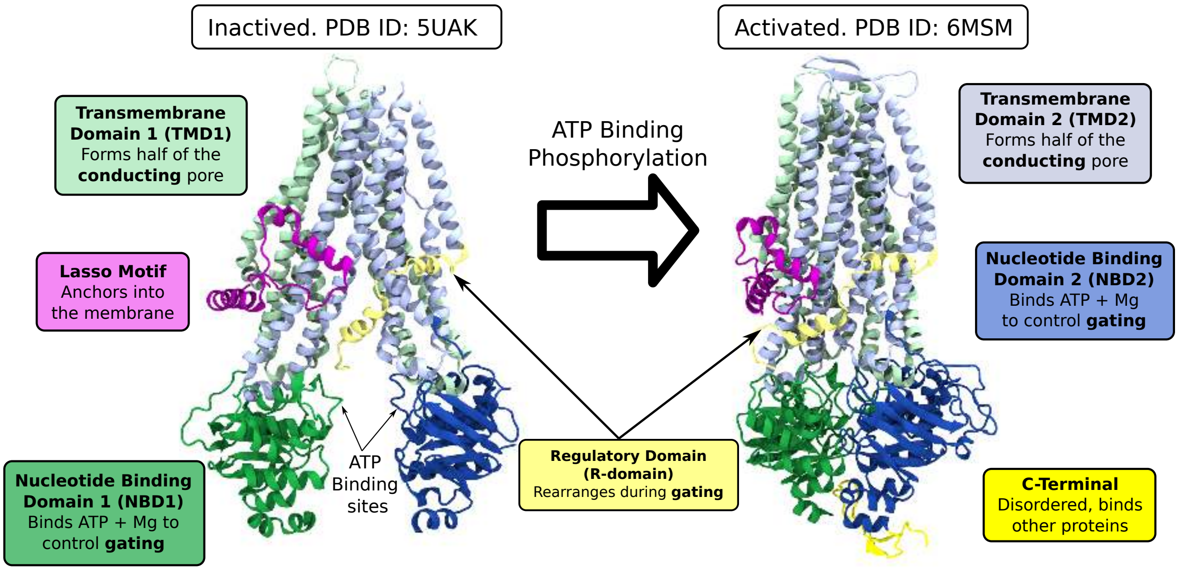
Inactived. PDB ID: 5UAK
Activated. PDB ID: 6MSM
Transmembrane Domain 1 (TMD1) Forms half of the conducting pore
ATP Binding Phosphorylation
Transmembrane Domain 2 (TMD2) Forms half of the conducting pore
Lasso Motif Anchors into the membrane
Nucleotide Binding Domain 2 (NBD2) Binds ATP + Mg to control gating
ATP Binding sites
Regulatory Domain (R-domain) Rearranges during gating
Nucleotide Binding Domain 1 (NBD1) Binds ATP + Mg to control gating
C-Terminal Disordered, binds other proteins

Figure 3.2: CFTR Structure
There are currently two resolved human structures. The inactivated state is neither phosphorylated nor bound to ATP. Observe how the NBDs are far apart and the TMDs are not parallel, forcing a constriction which does not allow the passage of ions. By contrast, the activated structure is abound to ATP at both sites, bringing the TMDs into a parallel configuration where they form a pore. There are unresolved questions as to whether CFTR may conduct chloride in this conformation which we will analyse in chapter 7.

7. C-terminus (NBD2 AA 1451 - 1480). This structure is natively disordered but it serves as an interaction hub in WT-CFTR, anchoring CFTR to other proteins through its PDZ binding domain [212, 213].

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 is a Unique ABC Transporter

ATP-Binding Cassette (ABC) transporters are an intriguing super family of proteins. On the whole, they transport substrates by using a combination of phosphorylation energy from ATP hydrolysis. These can be diverse substrates such as lipids or small molecules. Their structural diversity can be seen in figure ?? reflecting their array of functions.

CFTR is unique, as it is not a transporter, but rather an anion *channel*. The kinetic energy of the ATP is not used to translocate substrate across the membrane but rather simply used in the regulation of the gating cycle. Chloride, bicarbonate and other anions are able to *passively* diffuse through the channel. This evolutionary misappropriation of a transporter to a "leaky channel" is perhaps the reason so many mutations can create a non-functional protein [215].

3.4 CFTR Structure and Function

The primary cause of the disease Cystic Fibrosis (CF) is the malfunction 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"[215].

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 [205]. In experimentally determined structures of human CFTR the secondary structure of a section of the R-domain but not at high enough resolution to

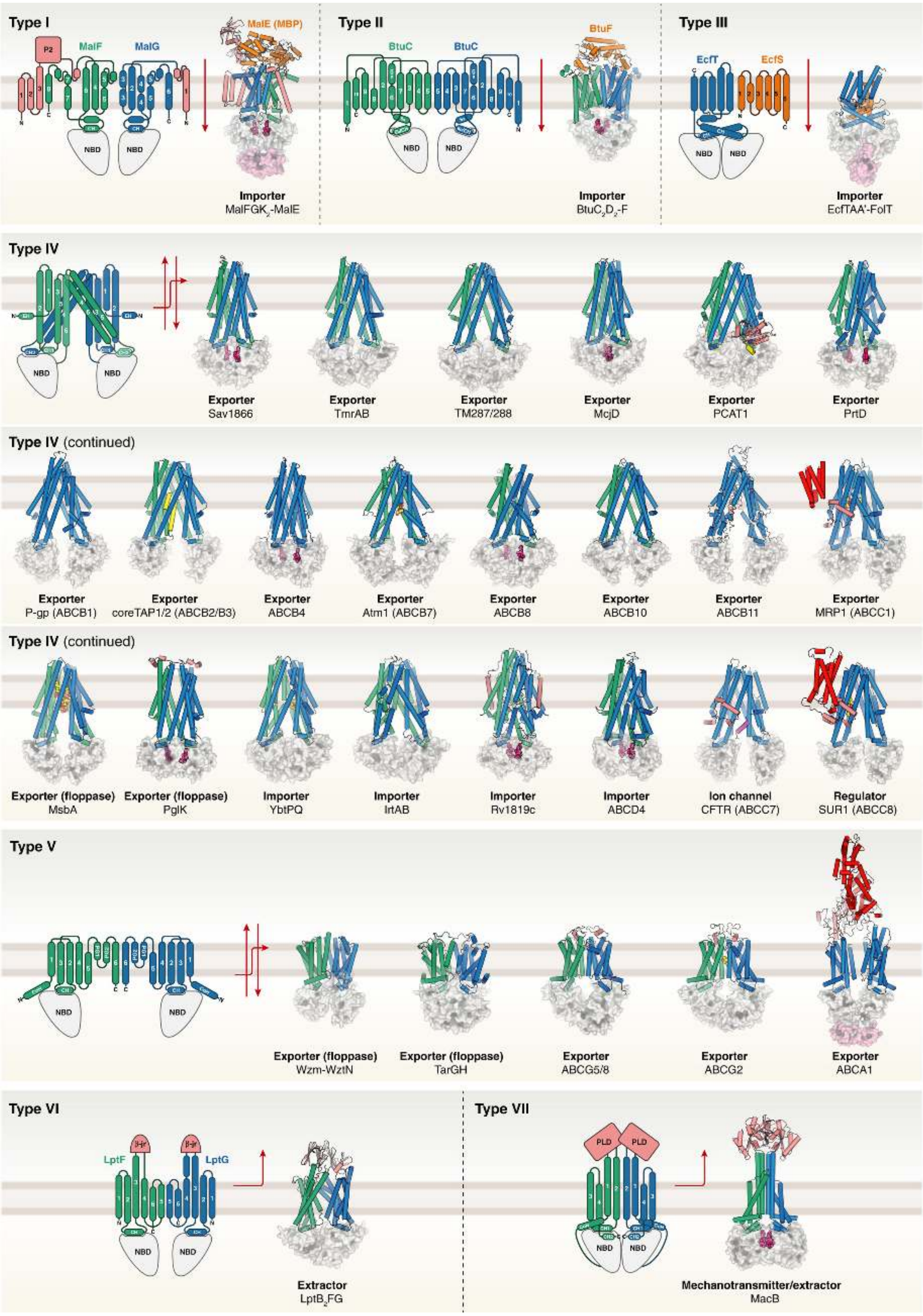

Figure 3.3: CFTR Structure
The structural diversity of ABC transporters. Structures are classified based on the organisation of their TMDs source [214].

determine the identity of individual side chains [216, 217]. Further secondary structure information can be found through experiments with NMR [218].

Previous computational studies of CFTR have been used homology models based on the phosphorylated zebra fish protein PDBID:5W81 [219]. These have yielded interesting results but the sequence similarity between human and zebra fish CFTR is only 55% []. For a protein structure where a single amino acid mutation leads to malfunction, more precision can only help. Additionally, the activity of CFTR modulators is not conserved in mutant zCFTR possibly because it has different 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 [220]. 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, these 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 cysteines 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 [216].

3.4.1 The Gating Cycle

The conformational transition from inactive to active differs significantly in CFTR compared to other ABC transporters. The NBDs 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 nucleophilic attack by surrounding water on the γ phosphate of the ATP bound to Walker B [221]. The hydrolysis of ATP is the event which causes the channel to gate back to the closed conformation [].

3.4.2 Anion Selectivity

CFTR is weakly selective for specific anions. 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?). WNK1 is known to influence the selectivity of the channel https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6889609/. The permeation of bicarbonate is very important physiologically because if a mutation permeates bicarbonate it means there is a high likelihood the patient will be pancreatic sufficient.

Compared to cation channels like Gramicidin and KcsA, CFTR is only weakly selective, permeating a large set of anions with varying radii and geometries. Supposedly it is more permeant to lyotropic (low solvation energy anions) rather than cosmotropic anions (high solvation energy anions) indicating that dehydration of the anion is likely during conductance (CITATION NEEDED). The radius of hydrated chloride ions is 1.7A[222] so even with this larger pore partial dehydration must take place.

3.5 Classes of Mutation Which Cause Cystic Fibrosis

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. We will break down this paradigm into more molecular detail in chapter **??**

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 potentiators 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 [223].

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.

3.6.2 Potentiators

There is more uncertainty surrounding the mechanism of potentiators 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 [224] 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 Patients with rare mutations struggle to gain access to modulator therapy

3.7 Patient Derived Organoids as a Pre Clinical Model

The basic unit of living things are cells. In the medical field there is growing capability to discern the functioning of an individual patient's cells. These can be used as models for testing what might produce the best clinical outcomes for the patient. For the case of Cystic Fibrosis researchers have begun to take samples of epithelial stem cells of patients with the disease and grow those samples into tissues which mimic the function of the entire organ[225]. This is possible in the epithelium due to a population of adult stem cells which maintain the ability to differentiate into a variety of cell types (a property known as pluripotency).

Primarily, these epithelial stem cells are taken from the nasal passages of patients or from rectal biopsies. These samples can then be grown into organoid models of the lung or gut respectively.

Adult stem cells in the epithelium are preferable because other sources of stem cells such as induced pluripotent stem cells (iPSCs) require complex, time consuming protocols to grow into fully developed organoids. Already the differentiation and expansion of epithelial samples takes a month [].

In the case of CF this technology allows the construction of a scalable, patient specific platform where a patient's own tissues can be tested to determine the best treatment for them. These pre-clinical models will allow more patients in the heterogeneous set of disease causing mutations to access modulators. This has given rise to an exciting prospect of a practice known as theratyping, enabling clinicians to make a personalised choice of which drugs which drug regimen will best serve a patient [226–229]. This thesis demosntrates that integration of *in silico* simulations into this process can further the capabilities of these pre-clinical models.

The response of the patient's epithelium is characterised using a few *in vitro* assays which we will briefly discuss below.

3.7.1 Forskolin Induced Swelling

Forskolin Induced Swelling (FIS) assays have been used to characterise the patient specific response of a patient's organoids to a drug regimen [230]. When epithelial cells are exposed to a chemical known as Forskolin they begin to rapidly produce cyclic AMP (cAMP, a precursor to ATP in a cell) []. This allows the down stream activation of CFTR ion channels, causing the organoids themselves to swell. This swelling allows cell biologists to easily quantify the activity of CFTR within a patient, in a variety of conditions.

3.7.2 Cilia Beating Frequency

The lungs are covered in cilia which fluctuate or "beat" in order to clear mucus and other particles in the respiratory tract [231, 232]. One of the most characteristic symptoms of patients with Cystic Fibrosis is the build up of mucus around the epithelium. This causes cilia to collapse, meaning they cannot move []. By measuring .

3.7.3 Electrophysiology

Since CFTR is an ion channel, measuring its electrical activity is a direct way to assess its function. For single channel studies this is done with a patch clamp. However, this does not give an assessment of the whole epithelium. Often the whole organoid is used as a patch and put in an Ussing chamber. By blocking other ion channels such as ENAC a clear picture of CFTR function can be measured in order to create a pre-clinical model for a specific patient.

3.7.4 Western Blotting to Assess CFTR Trafficking

The above methods may be able to properly quantify gating class mutations but they will struggle to assess the amount of CFTR at the cell surface. For this we employ .

Chapter 4

Molecular Dynamics and Functional Characterization of I37R-CFTR Lasso Mutation Provide Insights into Channel Gating Activity

My name is Benjamin John Goodwin

-Benjamin John Goodwin (personal communication)

Abstract

Characterization of I37R, a mutation located in the lasso motif of the CFTR chloride channel, was conducted by theratyping several CFTR modulators from both potentiator and corrector classes. Intestinal current measurements in rectal biopsies, forskolin-induced swelling (FIS) in intestinal organoids, and short circuit current measurements in organoid-derived monolayers from an individual with I37R/F508del CFTR genotype demonstrated that the I37R-CFTR results in a residual function defect amenable to treatment with potentiators and type III, but not type I, correctors. Molecular dynamics of I37R using an extended model of the phosphorylated, ATP-bound human CFTR identified an altered lasso motif conformation which results in an unfavorable strengthening of the interactions between the lasso motif, the regulatory (R) domain, and the transmembrane domain 2 (TMD2). Structural and functional characterization of the I37R-CFTR mutation increases understanding of CFTR channel regulation and provides a potential pathway to expand drug access to CF patients with ultra-rare genotypes.

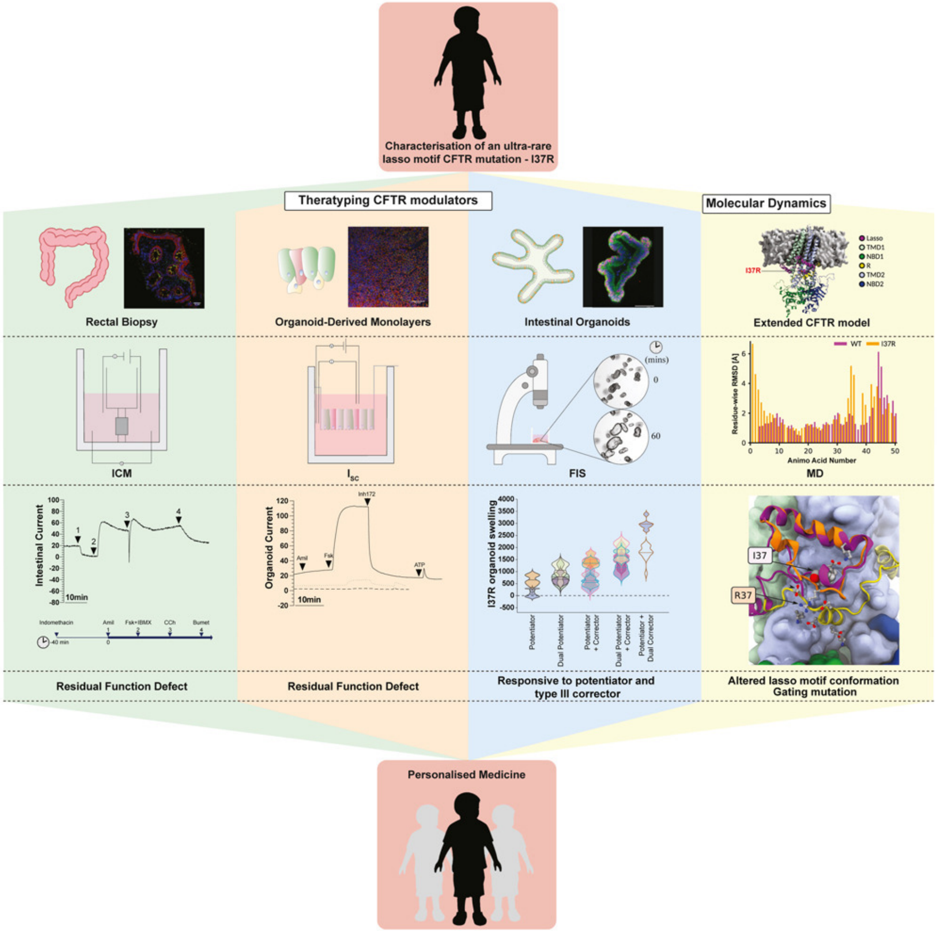

Figure 4.1: Graphical Abstract. Integration of in silico and in vitro experiments for personalised medicine

4.1 Introduction

Cystic fibrosis (CF) is a life-limiting genetic disease resulting from mutations in the CF transmembrane conductance regulator (CFTR) gene [233]. CFTR—the only member of the ABC transporter family known to be an ion channel—consists of two transmembrane domains (TMD1 and TMD2) which form an anion-selective pore, two highly conserved nucleotide-binding domains (NBD1 and NBD2) with ATP-binding pockets and a newly described N-terminal lasso motif [217, 234]. In addition, CFTR has a unique, disordered regulatory (R) domain which contains protein kinase A (PKA) phosphorylation sites. For the CFTR channel to open and close (gate), cAMP-dependent PKA phosphorylation of the R domain first activates the CFTR ([235]). Then, ATP-binding induces the dimerization of the two NBDs which opens the channel pore and ATP hydrolysis closes the pore.

The lasso motif (amino acids (aa) M1-L69), which is partially embedded in the bilayer and interacts with the R domain, was recently resolved following advancements in cryo-electron microscopy (cryo-EM) of the CFTR structure [50, 236]. The first 40 amino acids of the lasso motif, which include lasso helix 1 (Lh1, aa V11–R29), form a circular "noose" structure [**hoffmann2018**]. The noose structure wraps around the transmembrane helices (TM2, TM6 of TMD1 and TM10, TM11 of TMD2) and is held in place by hydrophobic interactions with L15, F16, F17, T20, L24, and Y28. The C-terminal end of the lasso, which includes the lasso helix 2 (Lh2, aa A46–L61), is tucked under the elbow helix (aa I70–R75) [**hoffmann2018**]. Variable disease severity and heterogeneous clinical presentation have been reported for the 78 CFTR variants identified so far in the lasso motif (CFTR1 and CFTR2 databases, Table S1). Evidently, the lasso motif has a multifunctional role in CFTR regulation with variants impacting folding, gating, and stability of the CFTR protein [200, 237–240].

CFTR modulators, small molecules which directly target CFTR dysfunction, are now available to certain individuals with CF. Currently, two classes are approved; (1) potentiators, which open the channel pore such as ivacaftor (VX-770) and (2) correctors, which assist CFTR protein folding and delivery to the cell membrane. Type I correctors (lumacaftor/VX-809, tezacaftor/VX-661) stabilize the NBD1-TMD1 and/or NBD1-TMD2 interface by binding directly to TMD1 [241, 242] or NBD1 which improves the interaction between NBD1 and the intracellular loops [243, 244]. Type II correctors (C4) stabilize NBD2 and its interface with other CFTR domains while type III correctors (elexacaftor/VX-445) directly stabilize NBD1 [245]. Combination therapies of corrector(s) and a potentiator (Orkambi, Symdeko/Symkevi, Trikafta/Kaftrio) have been approved for CF individuals with F508del, the most common CFTR mutation, as well as several specific residual function mutations. Most recently, Trikafta/Kaftrio has been approved for patients with a single F508del mutation in combination with a minimal function mutation, broadening the population of patients with CF eligible for treatment with CFTR modulator therapy.

Mounting evidence has shown that in vitro functional studies in patient-derived cell models successfully predict clinical benefit of available CFTR modulators for individuals bearing ultra-rare mutations [246–248]. In individuals with CF, adult stem cells are usually collected by taking either airway brushings or rectal biopsies. Single Lgr5+ stem cells, derived from crypts within a patient's intestinal epithelium, can be expanded in culture medium and differentiated into organized multicellular structures complete with the donor patient's genetic mutation(s), thus representing the individual patient [249]. Stem cell models can be used for personalized drug screening to theratype and characterize rare CFTR mutations [246, 250, 251]. Determining the functional response of rare, uncharacterized CFTR mutations to modulator agents with known CFTR correction mechanisms enables characterization of CFTR structural defects and enhances our understanding of CFTR function.

I37R-CFTR is a novel missense mutation in the lasso motif, detected in an Australian male child diagnosed through newborn screening with elevated immunoreactive trypsinogen, raised sweat chloride (>60 mmol/L), and CFTR Sanger sequencing identifying c.1521-1523del (F508del) and c.110C > T (I37R) mutations (Table S2). We used functional studies and molecular dynamics (MD) simulations to characterize the functional and structural defects of I37R-CFTR. CFTR function was assessed using intestinal current measurements (ICM) in rectal biopsies, forskolin-induced swelling (FIS) assays in intestinal organoids, and short circuit current measurements (Isc) in I37R/F508del organoid-derived monolayers, respectively. The potentiators VX-770 (approved), GLPG1837 (phase II clinical trials), and genistein (a natural food component with potentiator activity [252]) were tested as monotherapies, dual potentiator therapies, or in combination with correctors (VX-809, VX-661, and VX-445). We compared this to our laboratory reference intestinal organoids. For MD simulations, we modeled and examined the structural defect of the I37R mutation on an extended cryo-EM structure of ATP-bound, phosphorylated human CFTR (PDB ID code 6MSM) [50].

4.2 Results

4.2.1 I37R-CFTR Baseline Activity in Patient-Derived Rectal Biopsies and Intestinal Organoids

Intestinal current measurements (ICM) were performed on I37R/F508del and reference CF (F508del/F508del, G551D/F508del) and non-CF (wild-type: WT/WT) rectal biopsies using a standard protocol [253, 254] (Figure 1A). Following stimulation with a forskolin (fsk) and IBMX cocktail, rectal biopsies from the I37R/F508del CF participant elicited cAMP-dependent currents of 45.8 ± 3.8 μA/cm^2—an appreciable 50% of WT-CFTR activity ($p < 0.05$; Figure ??A, Table S3). This response was at least 4-fold higher than those of the reference CF biopsies, although statistical significance was not reached.

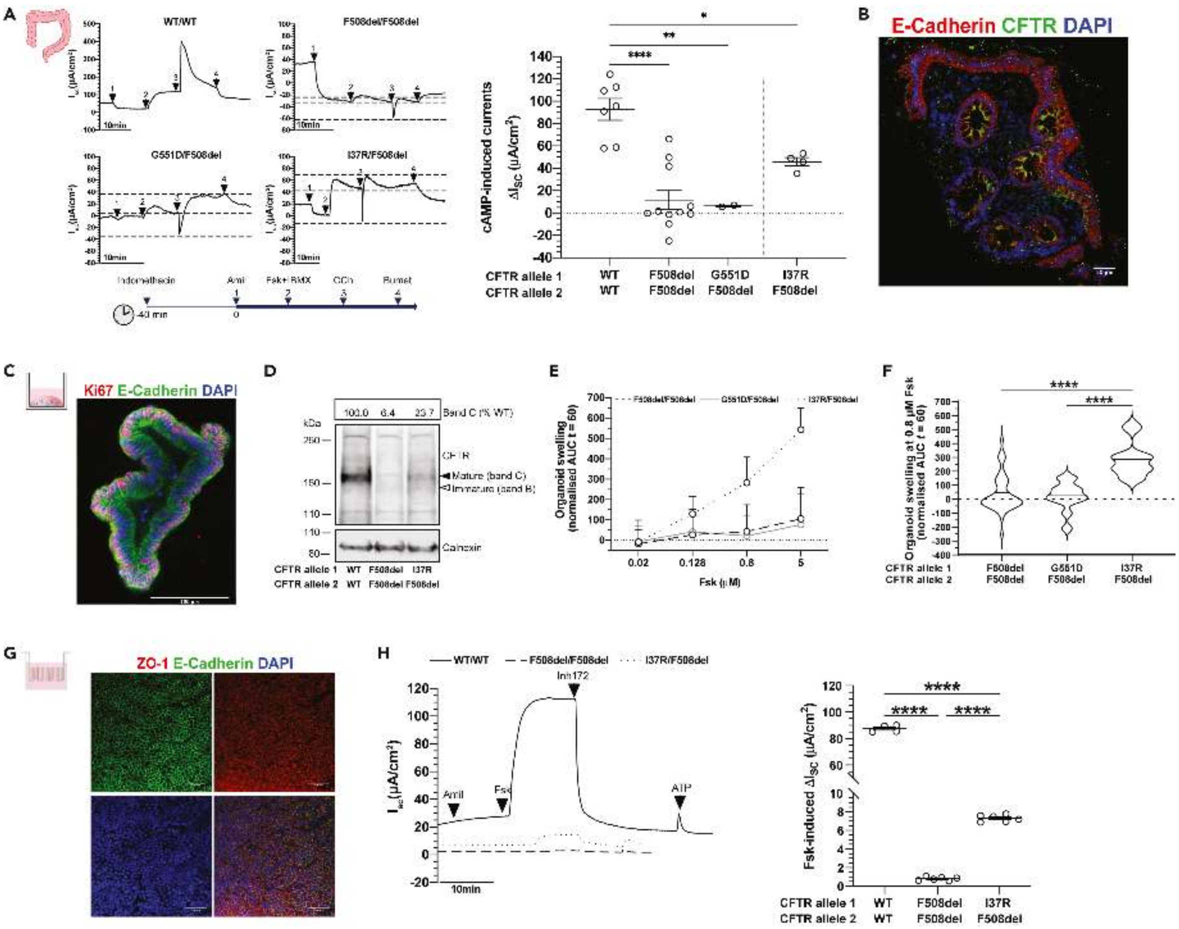

Figure 4.2: Characterization of I37R-CFTR residual function in rectal biopsies and intestinal organoids
(A) Representative Ussing chamber recordings of intestinal current measurements (ICM) in rectal biopsies from WT-CFTR control participants and participants with CF. Dot plots of cAMP-induced current (ΔIsc-Fsk + IBMX) in participants with WT/WT (n = 2), F508del/F508del (n = 3), G551D/F508del (n = 1), and I37R/F508del (n = 1) CFTR genotypes. Experiments were performed in the presence of 10 μM indomethacin. Arrows indicate the addition of compounds: 100 μM apical amiloride (1. Amil), apical and basal addition of 10 μM forskolin +100 μM IBMX cocktail (2.Fsk + IBMX), 100 μM basal carbachol (3.CCh), and 100 μM basal bumetanide (4.Bumet). The Isc at the time CCh was added (middle horizontal dotted line), and the maximum (top dotted lines) and minimum (bottom dotted lines) Isc induced are indicated. Each dot represents an individual replicate.

(B) Immunofluorescence staining of CFTR (green), e-cadherin (red), and DAPI (blue) in a rectal biopsy derived from an I37R/F508del participant. 63x/1.4 oil immersion objective. Scale bar = 50 μm.

(C) Immunofluorescence staining of e-cadherin (green), Ki67 (red), and DAPI (blue) in intestinal organoids derived from an I37R/F508del participant. 20x/0.75 dry objective. Scale bar = 100 μm.

(D) Western blot in WT/WT, F508del/F508del, and I37R/F508del intestinal organoids. CFTR maturation was calculated by measuring the level of mature mutant CFTR (Band C) as a percentage of mature CFTR from WT organoids (% normal CFTR). All data were normalized to the calnexin loading control. B and C represents the mature, complex-glycosylated CFTR. B and B represents the immature, core-glycosylated CFTR. See Figure S9 for uncropped Western blot images.

(E and F) Forskolin-induced swelling (FIS) assay in organoids from participants with F508del/F508del (n = 5), G551D/F508del (n = 2), and I37R/F508del (n = 1) CFTR genotypes. Organoids were stimulated with forskolin (fsk) concentrations ranging from 0.02 to 5 μM.(E) FIS expressed as the means ± standard deviation (SD) of the area under the curve (AUC) calculated from t = 0 (baseline) to t = 60.(F) FIS of organoids at 0.8μM fsk at baseline represent residual CFTR function. Data represented as violin plots with mean to show distribution.

(G) Immunofluorescence staining of e-cadherin (green), ZO-1 (red), and DAPI (blue) in organoid-derived monolayers from a CF participant. 20x/0.75 dry objective. Scale bars = 50 μm.

(H) Representative Ussing chamber recordings of short circuit current in organoid-derived monolayers from a WT-CFTR control participant and participants with CF. Dot plots of fsk-induced current (ΔIsc-Fsk) in participants with WT/WT (n = 1), F508del/F508del (n = 1), and I37R/F508del (n = 1) CFTR genotypes. Experiments were performed in the presence of 10 μM indomethacin. Arrows indicate the addition of compounds: 100 μM apical amiloride, 5 μM basal fsk, 30 μM apical CFTR inhibitor CFTRinh-172, and 100 μM apical ATP. Each dot represents an individual replicate. Data in (A) and (H) represented as mean ± standard error of the mean (SEM). One-way analysis of variance (ANOVA) was used to determine statistical differences. * $p < 0.05$, ** $p < 0.01$, **** $p < 0.0001$.

Co-activation with carbachol (CCh) resulted in a biphasic response in the I37R/F508del biopsies, characteristic of residual CFTR chloride channel function in the CF colon [254, 255]. The initial negative Isc peak indicates apical potassium secretion reached 9.4 ± 2.5 $\mu A/cm^2$. Following this, the CCh-induced positive Isc indicates the increase of apical chloride secretion reached 15.78 ± 2.07 $\mu A/cm_2$. This biphasic response was similarly observed in the G551D/F508del biopsies (25.77 ± 2.16 $\mu A/cm^2$) but was diminished in the F508del/F508del biopsies (-2.28 ± 1.65 $\mu A/cm^2$). These findings are in accordance with the localization of CFTR protein at the plasma membrane (mature complex-glycosylated CFTR) of the I37R/F508del rectal biopsies, as demonstrated by immunofluorescence staining (green; Figure 1B).

Next, CFTR protein expression and maturation was assessed in I37R/F508del, reference F508del/F508del, and WT/WT organoids using Western blot (Figures 1C and 1D). The expression of complex-glycosylated C band in I37R/F508del organoids was 23.7% that of the WT/WT organoids, considerably higher than the 6.4% detected from F508del/F508del organoids (Figure 1D). CFTR activity was then evaluated in I37R/F508del and CF reference intestinal organoids using a fsk-induced swelling (FIS) assay at four fsk concentrations between 0.02 and 5 μM (Figure 1E). FIS of I37R/F508del intestinal organoids at 0.8 μM fsk—the optimal concentration for baseline assessment of CFTR activity [256]—was 282.9 ± 36.0 (Figures 1E and 1F). This exceeded the baseline FIS of the reference intestinal organoids by at least 7-fold (F508del/F508del: AUC = 42.8 ± 19.4; G551D/F508del: AUC = 21.3 ± 29.4).
The morphological difference between WT (pre-swollen) and CF organoids [257], means comparing CFTR activity between CF and healthy CFTR function by FIS assay cannot be achieved [256, 258]. In order to compare I37R/F508del to wild-type CFTR activity, organoid-derived monolayers were created (Figure 1G) and CFTR ion transport was performed [259]. Fsk-stimulated CFTR-dependent currents were 9-fold higher in I37R/F508del monolayers than those of reference F508del/F508del monolayers (7.3 ± 0.2 vs 0.8 ± 0.1 $\mu A/cm^2$; $p < 0.0001$), but 12-fold lower than WT/WT monolayers (87.5 ± 1.3 $\mu A/cm^2$; $p < 0.0001$) (Figure 1H). This is consistent with the FIS assay results demonstrating high baseline CFTR activity in I37R/F508del intestinal organoids.

4.2.2 I37R-CFTR Functional Response to CFTR Modulator Monotherapy in Intestinal Organoids

We investigated the functional response of I37R/F508del organoids to single potentiators: VX-770, GLPG1837 (G1837), and genistein (Gen). Treatment with VX-770 minimally increased FIS of I37R/F508del organoids by AUC of 59.7 above baseline at 0.128 μM fsk (Figures 2A–2C and S1)—the optimal concentration for in vitro assessment of CFTR modulator response to predict clinical effect [256]. G1837 and Gen both significantly increased FIS, albeit with different efficacies (655.8 and 256.8, respectively; Figures 2A–2C and S1). None of the potentiator treatments increased FIS in F508del/F508del organoids, indicating no improvement in CFTR activity in response to potentiator therapy (Figure 2C). Only G1837 significantly increased FIS in the G551D/F508del organoids (210.4 ± 57.5; p < 0.01). In comparison to G551D/F508del organoids, G1837 was 3-fold more efficacious in the I37R/F508del organoids (p < 0.0001).

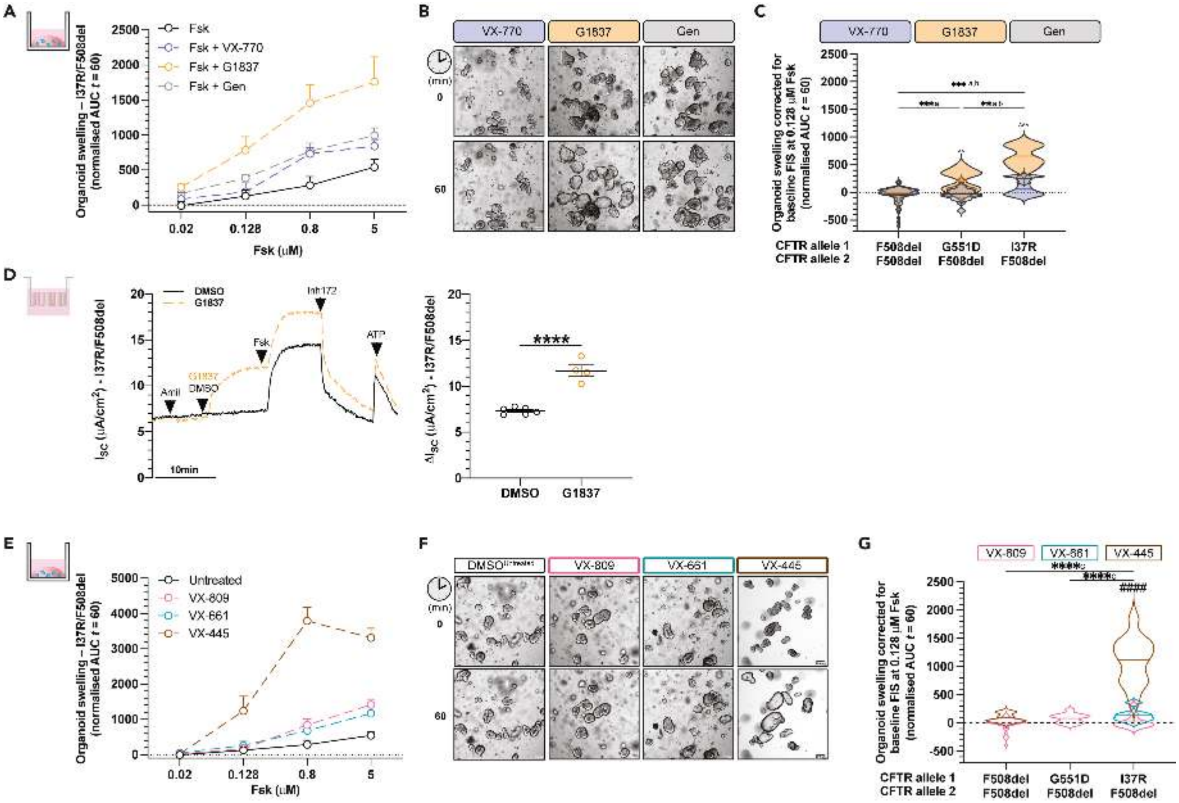

Figure 4.3: Characterization of I37R-CFTR functional response to corrector or potentiator monotherapy in intestinal organoids

Forskolin-induced swelling (FIS) assay in organoids from participants with F508del/F508del (n = 5), G551D/F508del (n = 2), and I37R/F508del (n = 1) CFTR genotypes. Organoids were incubated overnight with 0.03% DMSO (untreated) or 3 μM VX-809 or 3 μM VX-661 or 3 μM VX-445. After 24 h, organoids were stimulated with fsk concentrations ranging from 0.02 to 5 μM, either alone or in combination with potentiator monotherapy (3 μM VX-770 or 3 μM G1837 or 50 μM Gen).

(A) FIS of I37R/F508del organoids stimulated with VX-770, GLPG1837 (G1837), or genistein (Gen) monotherapy, expressed as the means ± standard deviation (SD) of the area under the curve (AUC) calculated from t = 0 (baseline) to t = 60 min.

(B) Representative brightfield images of I37R/F508del organoids at baseline (t = 0) and after 1 h of stimulation (t = 60) at 0.128 μM fsk. Scale bars = 100 μm.

(C) FIS of organoids at 0.128μuM fsk following stimulation with VX-770, GLPG1837 (G1837), or genistein (Gen) monotherapy. Data corrected for baseline FIS and represented as violin plots with mean to show distribution.

(D) Representative Ussing chamber recordings of short circuit current in I37R/F508del organoid-derived monolayers. Dot plots of total currents stimulated by DMSO or G1837 plus fsk. Experiments were performed in the presence of 10 M indomethacin. Arrows indicate the addition of compounds: 100 M apical amiloride, apical addition of either vehicle control 0.01% DMSO or 10 μM G1837, 5 μM basal fsk, 30 μM apical CFTR inhibitor CFTRinh-172, and 100 μM apical ATP. Each dot represents an individual replicate. Data represented as mean ± standard error of the mean (SEM).
(E) FIS of I37R/F508del organoids pre-incubated with corrector (VX-809 or VX-661 or VX-445) for 24 h, expressed as the means ± standard deviation (SD) of the area under the curve (AUC) calculated from t = 0 (baseline) to t = 60 min.
(F) Representative brightfield images of I37R/F508del organoids at baseline (t = 0) and after 1 h of stimulation (t = 60) at 0.128 μM fsk. Scale bars = 100 μm.
(G) FIS of organoids at 0.128μM fsk following incubation with corrector (VX-809 or VX-661 or VX-445) for 24 h. Data corrected for baseline FIS and represented as violin plots with mean to show distribution. One-way analysis of variance (ANOVA) was used to determine statistical differences except in (D) where unpaired t test was used. **p ¡ 0.01, ***p ¡ 0.001, and ****p ¡ 0.0001. aP for G1837, bP for Gen and cP for VX-445 of I37R/F508del, $\hat{P}$ for G1837 vs VX-770, or Gen and #P for VX-445 vs VX-809 or VX-661.
Because G1837 demonstrated the greatest restoration of CFTR activity in I37R/F508del organoids, we evaluated G1837 treatment of I37R/F508del organoid-derived monolayers. G1837 led to a significant 1.5-fold increase in fsk-stimulated currents (ΔIsc: 4.4 μA/cm^2; $p < 0.0001$) (Figure 2D). This is consistent with the FIS of I37R/F508del organoids, indicating that I37R-CFTR responds to potentiator agents.
Given the I37R/F508del high residual CFTR activity and its localization at the epithelial cell surface, we hypothesized that the I37R-CFTR mutation has minimal impact on CFTR protein folding or maturation. Treatment of I37R/F508del organoids with type I corrector agents (VX-809 or VX-661) did not significantly increase FIS above baseline (Figures 2E–2G and S1). In contrast, treatment of I37R/F508del organoids with a type III corrector agent (VX-445) significantly increased FIS by AUC of 1112.5 above baseline, greater than those in the F508del/F508del organoids (42.5). VX-445 has been shown to act as both a corrector and potentiator for certain CFTR mutations [260–262]. Acute treatment of I37R/F508del organoids with VX-445 did not improve potentiation of CFTR (Figure S1). This supports the observation that VX-445-stimulated rescue of CFTR in I37R/F508del organoids acts by a correction mechanism improving I37R mild folding and processing defects. I37R-CFTR functional response to CFTR modulator co-therapies in intestinal organoids.
Combination treatments of CFTR modulators are used to treat patients bearing CFTR mutations with multiple functional defects such as F508del and patients who are heterozygous for CFTR mutations. We investigated the effect of combinations of potentiators. Dual potentiator combinations increased FIS of I37R/F508del organoids to a greater extent than the respective single potentiators (Figure 3A) and had a synergistic effect, where the FIS was greater than the sum of the respective single potentiators (Table S4). Despite G1837 + Gen having greater efficacy than the other dual potentiator combinations, the magnitude of response was not statistically different between the different combinations of dual potentiators (Figure 3A).

Co-therapy with a corrector (VX-809 or VX-661) and dual potentiators significantly ($p < 0.01$) increased FIS of I37R/F508del organoids compared to co-therapy of a corrector with VX-770 or Gen, but not G1837 (Figure 3B). VX-809/G1837 + Gen co-therapy had the greatest efficacy, increasing FIS 1904.0 above baseline. In contrast, corrector/VX-770 + Gen co-therapy had the least efficacy. This trend was consistent with that of the dual potentiators synergistic effect.

Dual correctors (VX-445+VX-661) increased FIS in I37R/F508del organoids by AUC of 1856.6 above baseline, which corresponds with the level of rescue achieved by the most effective corrector/dual potentiator co-therapy (VX-809/G1837 + Gen). The triple combination therapy with dual correctors and a potentiator further increased FIS in I37R/F508del organoids by AUC of 3101.6 above baseline. It is therefore the most effective modulator combination tested in this study.

4.2.3 I37R-CFTR Perturbs the Noose Structure of the Lasso Motif

We next characterized the structural defect of I37R-CFTR using MD simulations. The primary structure of the lasso motif (M1-L69) is conserved across 230 vertebrate species (Figure S2, Table S5). The lasso motif formed a noose structure that rested against TMD2 (Figure 4A). Amino acids V12-R29 were embedded in the plasma membrane while the rest of the lasso motif resided in the cytosol. The noose structure was maintained by a salt bridge formed between K26 and D36 (Figure 4B). I37 was positioned in the center of this noose, within a hydrophobic pocket formed by amino acids from the lasso, TMD2, and the poorly resolved R domain in the cytosol (Figure 4C).

Mutation of the evolutionarily conserved, non-polar and uncharged isoleucine (I) of I37 to a positively charged arginine (R) introduced an unstable lone charge into the hydrophobic pocket within the lasso motif noose. We hypothesized that this likely results in the rotation of the R37 side chain out of the hydrophobic pocket, and possible coordination with negative charges in the nearby R domain.

To identify a reasonable conformation of the mutant lasso motif, the WT 6MSM model was mutated to R37 and three 2 μs simulations were performed at physiological temperature (310 K). The R37 side chain rotated out of the hydrophobic pocket in only one of the three simulations. The difference between the root-mean-square deviation (RMSD) of the noose structure of I37R-CFTR compared to the WT was on average 2.8 Å at the amino acids M1-L6, and 1.8 Å at L34-S50 (Figure 4D). To confirm this observation, repeat simulations were performed at 350 K (40°C above physiological temperature), a temperature shown to accelerate the potential conformational transitions of proteins [263]. In these higher temperature simulations, the root-mean-square fluctuation (RMSF) of the region around amino acid 37 doubled in two out of three simulations, compared to WT-CFTR at 310K (Figure S3). This confirmed the destabilization of the lasso motif by I37R-CFTR. All WT-CFTR domains and the surrounding bilayer remained stable at the elevated temperature (Figures S4 and S5).

4.2.4 I37R Mutation Strengthens Lasso Motif Interaction with the R-domain

In the 6MSM structure, the R domain is largely unresolved with two exceptions: the first (Q637) and last (T845) amino acids that adjoin neighboring domains, and the backbone atoms of a 17 amino acid segment. This latter segment consists of an eight amino acid disordered coil followed by a nine amino acid alpha-helix [50]. The alpha-helix was separated by approximately 10Å (1 nm) minimum C-alpha distance to I37 in the lasso motif. This suggested a likely interaction between this segment of the R domain and I37, which necessitated partial modeling of the R domain (Figure 5A).

Modeling of these 17 unidentified amino acids was performed by creating 24 different in silico models of this segment based on the 6MSM structure. In each model, a unique 17 amino acid sequence was determined with a sliding window of one amino acid, starting backwards from amino acid T842 due to the alpha-helix's 20 Å proximity to T845. The 17 amino acids were then connected to T845 with the missing linking amino acids. The structural stability of all 24 modeled segments was tested by performing up to 300 ns simulations for each model and comparing the backbone RMSD measurements against 6MSM (Figures 5B and S6). The model with the lowest RMSD (3 Å) and thus the highest stability was attained when L818-F834 was assigned to the unidentified 17 amino acids, of which the alpha-helix maps to E826-F834 (Figures 5B and S6). This assignment was corroborated by NMR measurements of the isolated R domain in solution, where the same segment retained partial helicity [**baker2007**]. Predictions of the structure of human CFTR by Alphafold2 also aligned with this assignment of primary structure to the unidentified amino acids (Figure S7) [61]. Several favorable interactions between this R domain model and other parts of the CFTR protein further supported this assignment (Figures 4C and 5D). Two hydrophobic amino acids (L829 and F833) contributed to the hydrophobic pocket that stabilized the lasso motif around I37. The negatively charged E831 formed a salt bridge with positively charged K968 in TMD2. Together, these interactions secured the R domain alpha-helix into position throughout an extended 2 μs simulation, resulting in a smaller minimum C-alpha distance to the lasso motif of 8.9 ± 0.2 Å compared to the 10 Å in the 6MSM cryo-EM structure.

The reoriented R in position 37 in the I37R mutant protein, which pointed out of the hydrophobic pocket, rearranged the salt bridge network supporting the lasso motif by breaking the evolutionarily conserved salt bridge K26–D36. Two new salt bridges were formed, one with the negatively charged E823 and another with E826 of the R domain (Figure 5C). Furthermore, the E831–K968 salt bridge between the R and TMD2 domains in the WT was exchanged for a D828–K1080 salt bridge in I37R-CFTR (Figure 5C). The backbone motions required to accommodate these new charge interactions also perturbed parts of the lasso motif (Figure S3) and R domain. The lasso N-terminus shifted its position towards the R domain and reduced the minimum C-alpha distance between them by 3.5 Å (Figure 5D). The overall result was a tighter coupling between the lasso and the R domain which is anticipated to inhibit the R domain movements required for channel gating.

4.3 Discussion

We have described the functional and structural defects of I37R, a novel CF-causing mutation in the segment of the CFTR lasso motif which interacts with the R domain. These were compared to reference CFTR mutations which have known functional defects, either a CFTR folding/maturation (F508del/F508del) or a gating (G551D/F508del) defect. First, ICM performed in I37R/F508del rectal biopsies identified I37R confers high residual activity (50% of WT-CFTR activity). High baseline CFTR activity was similarly observed in FIS of I37R/F508del intestinal organoids and Isc measurements in organoid-derived monolayers. Given we and others showed that F508del is a severe mutation which contributes little functional CFTR [264], this suggests that I37R mutation produces CFTR protein which localizes to the epithelial cell surface. These observations are consistent with the patient's mild CF clinical phenotypes (pancreatic sufficient with faecal elastase $>$ 500 μg/g, FEV1 z-score -0.11, 99% predicted).

We also characterized the response of I37R-CFTR to modulators (potentiators and correctors) in I37R/F508del intestinal organoids and organoid-derived monolayers. I37R was responsive to potentiators which improve CFTR gating function and a newly approved corrector (VX-445). Among the three potentiator agents tested, the response to VX-770 was minimal. The reason for the lack of efficacy of VX-770 is not known, because molecular modeling studies propose that VX-770 shares the same mechanism of action and binding sites with G1837 [208, 265]. Both VX-770 and G1837 are proposed to potentiate CFTR by increasing channel open probability (Po) through stabilization of the open-pore conformation, independent of NBD dimerization and ATP hydrolysis which normally controls channel gating [266, 267]. However, the differing potentiator efficacies are not a new observation. G1837 was previously shown to be more potent and effective than VX-770 in human bronchial epithelial cells from a G551D/F508del and a R334W/F508del CF participant [268, 269]. Similar observations were reported in heterologous HEK293 cells expressing Class III (G551D, G178R, and S549N) and Class IV (R117H) CFTR mutants [268, 269]. We conclude that perhaps G1837 has additional binding sites or actions distinct from VX-770, which in the case of I37R-CFTR, results in significant potentiation of the CFTR channel.

We further showed that dual potentiator combinations exerted synergistic restoration of CFTR activity in I37R/F508del organoids. This synergistic restoration is not exclusive to I37R-CFTR, because similar findings have been reported for other CFTR mutations responsive to potentiators [270–273]. Synergism is commonly achieved when potentiators have distinct binding sites and mechanisms of actions. One potentiator could induce allosteric interactions that favor the activity of the other potentiator [274]. The potentiator synergy observed in our dual potentiator combinations supports our hypothesis that G1837 may have additional binding sites or mechanisms of action to VX-770. While VX-770 has been shown to provide clinical benefit to patients with responsive mutations [275–277], it does not restore the Po of gating defect mutants (G551D-CFTR) to full WT-CFTR activity [266]., 2009). This opens the possibility that using another potentiator with a different mechanism of action could complement VX-770 activity and increase CFTR activity beyond that of VX-770 monotherapy. While VX-770 and G1837 act independently of NBD dimerization and ATP hydrolysis [266, 267], genistein promotes ATP-dependent gating of CFTR by binding to the NBD1/2 interface and inhibiting ATP hydrolysis [278]. Genistein has been demonstrated to increase VX-770-potentiated CFTR activity in intestinal organoids, even when VX-770 was used at near-saturating concentrations [270]. Our observations reiterate and expand on these findings to suggest that potentiators with different mechanisms of action could provide synergistic restoration of CFTR activity to responsive CFTR mutations compared to potentiator monotherapy.

Chronic treatment with type III corrector VX-445 rescued CFTR activity in I37R/F508del organoids, while neither type I correctors (VX-809 or VX-661) rescued activity. This response is attributed to the I37R and not the F508del mutation in the I37R/F508del organoids, because VX-445 did not restore CFTR activity in F508del/F508del organoids. While VX-445 has been shown to have partial potentiator activity [260–262], VX-445 did not potentiate CFTR activity in I37R/F508del organoids when administered acutely. This is the first study to interrogate the potentiator action of VX-445 in intestinal organoids; however, previous studies have been performed in donor-derived bronchial and nasal epithelial cells and immortalized cell lines. The higher correction efficacy of VX-445 when compared with VX-809/VX-661 has previously been shown, although this is likely to be dependent on the CFTR variant [279–281]. For instance, direct binding of VX-445 to NBD1 to stabilize and prevent the domain unfolding may make it more effective in correcting CFTR mutations that impact NBD1 function (such as F508del located in NBD1).

The lack of I37R-CFTR correction by VX-809 or VX-661 could be attributed to the dependency of these modulators binding to and stabilizing the TMD1. TMD1 function is modulated by interaction with lasso helix 2 (Lh2, aa A46–L61) as deletion of Lh2 from the WT CFTR was shown to completely abrogate VX-809-mediated CFTR maturation [240]. MD studies showed that VX-809 occupancy at the TMD1 binding site causes the Lh2 to move, such that the network of salt bridges in Lh2 holds TMD1 (CL1) and TMD2 (CL4) in the correct orientation [245, 282]. This then allows for allosteric coupling between NBD1 and TMD1 or 2, which is important for cooperative domain folding of CFTR. In support of this, mutation of critical amino acids at the binding pocket of VX-809 on CFTR, or those involved in the architecture of this site, were shown to diminish the sensitivity to VX-809 correction. L53V and F87L mutations, which are located in the vicinity of the VX-809 binding site in the TMD1, were shown to prevent VX-809 correction in F508del HEK283 cells [282]. Considering the above and because I37 is only a few amino acids away from the Lh2, it is plausible that the local conformational changes associated with the I37R mutation which we have identified in our study (Figure 4D) may disrupt the allosteric coupling between NBD1 and TMD1 or 2, preventing correction with type I correctors.

CFTR missense mutations in the lasso motif are not well characterized. This is because most of these mutations are rare, with an allele frequency of less than 0.01% in the CF population (Table S1). The only characterized missense mutations in the region of the lasso motif where I37 resides—between Lh1 (amino acid 19–29) and Lh2 (amino acid 46–61)—are R31C and R31L [204, 239]. Experimental studies in heterologous COS-7 cells showed both mutations cause a mild processing defect and accelerated CFTR internalization. Individuals heterozygous for these CFTR mutations are reported to have a mild disease phenotype with pancreatic sufficiency [239]. One individual with the R31C/F508del CFTR genotype was reported to have a normal sweat chloride level (25 mmol/L) and nasal potential difference [283]. CFTR2 classifies R31C as a non-CF disease causing mutation. Notably, mild disease phenotypes (mild pulmonary symptoms, pancreatic sufficiency) are reported for several other lasso motif missense mutations including P5L, E56K, and P67L (Table S1), as was found for the I37R/F508del participant in this study. This suggests that perhaps lasso motif mutations do not significantly impact the overall CFTR structure and function given its short length (69 of 1480 amino acids, 4.7%). It is also plausible that the role of the lasso motif could be compensated for by other CFTR domains.

To better understand the functional defect of I37R-CFTR, we used MD simulations to model the structural features of I37R and how they are altered relative to WT-CFTR. The amino acids 34–39 were shown to interact with the R domain in the phosphorylated, ATP-bound CFTR structure [50]. This interaction was absent in the closed conformation of CFTR [217], suggesting that the short region of amino acids 34–39 interacts with the R domain to regulate CFTR channel gating. We found that the disruption of the evolutionarily conserved K26-D36 salt bridge in I37R-CFTR brings the lasso motif closer to the R domain. We also found that the I37R side chain rotates out of its hydrophobic pocket to form interactions with negatively charged E823 and E826 on the R domain. We speculate that R37 clamps the lasso motif to the R domain, preventing the dynamic movement of the two domains necessary for a normal CFTR opening and closing cycle, thus causing a gating defect. This supports our functional observations, wherein I37R-CFTR demonstrated significant responsiveness to potentiator agents which are known to increase channel opening time. Furthermore, in the I37R-CFTR model, conformational changes in the lasso motif were also evident but were limited to short regions (M1-L6, L34-S50), indicating that the overall architecture of the CFTR protein remains largely intact. Additionally, our simulations did not show any change to the pore architecture of CFTR (Figure S8).

The simulated structure in this work is of CFTR in its active state [50]. Because of this, we believe the pathogenic interactions discovered in this study have a significant contribution to the deleterious effects of the I37R mutation. However, the enhanced lasso motif-R domain interactions should be interpreted in the context of the μs timescales reachable by unbiased simulations. The lasso domain is known to exhibit conformational flexibility during both folding and functional stages of CFTR [284], which take place on timescales longer than is currently feasible to study in atomistic simulations. Therefore, there may be pathogenic interactions in I37R-CFTR in addition to the ones captured by the simulation of this particular CFTR structure.

The I37R/F508del participant in this study will only meet the Therapeutic Goods Administration (Australia) requirements for treatment with Trikafta/Kaftrio triple combination therapy once he turns 12 years old given the single copy of the F508del mutation. He is not eligible for single potentiator therapy or corrector/potentiator combinations of lumacaftor/ivacaftor or tezacaftor/ivacaftor. This emphasizes the importance of characterizing the structural and functional defects of ultra-rare CFTR mutations together with the assessment of in vitro response to modulator drugs in patient-derived cell models to build the case for access to treatment with available modulators through precision medicine health technology assessment pathways. Furthermore, when multiple CFTR

modulators are available to patients with CF, determining the best modulator for patients with a rare mutation not investigated in a clinical trial may be supported using in vitro personalized cell models. Limitations of the study
Organoids often lack specialized cell types and fail to recapitulate the complexity of native organs [285]. For example, mesenchymal, endothelial, and microbiome are absent from intestinal organoids. Integration of such features remains technically challenging and their absence may impact drug response. Another important drawback of organoid systems is the heterogeneity in their size when seeded for FIS assay. As the size of organoids increases, diffusion-dependent drug supply becomes less efficient. This may in turn impact the accuracy of outcome of drug assay. Reducing this variability will be essential to fully capitalize on the potential of organoids in drug screening. Another limitation of the organoid systems is the variability in the magnitude of FIS response in intestinal organoids across different CF laboratories. This is due to the dependence of organoids on media that is developed in-house with many locally produced media factors [248, 256]. This limitation can be resolved by the creation of reference donor organoids which are made available and used internationally between CF laboratories.

4.4 Method Details

Intestinal Current Measurement Superficial rectal mucosa samples (2 – 4 per donor) were freshly obtained using biopsy forceps (CK Surgitech NBF53-11023230) and placed in cold RPMI1640 media (Sigma R5886) with 5% FBS. Intestinal current measurements were performed under voltage-clamp conditions using VCC MC8 Ussing chambers (Physiologic Instruments, San Diego, CA) [286–288]. Biopsy tissues were bathed in Ringer solution containing (mM) 145 NaCl, 3.3 K_2HPO_4, 0.4 K_2HPO_4, 10 D-Glucose, 10 $NaHCO_3$, 1.2 $MgCl_2$ and 1.2 $CaCl_2$. Ringer solutions were continuously gassed with 95% O2-5% CO2 and maintained at 37°C. 10 μM indomethacin was added to both apical and basal chambers, and tissues were stabilised for 40 min. Tissues were then treated with pharmacological compounds (in order): 100 μM amiloride (apical) to inhibit epithelial sodium channel (ENaC)-mediated Na+ flux, 10 μM forskolin + 100 μM IBMX cocktail (apical and basal) to induce cAMP activation of CFTR, 100 μM carbachol (basal) to increase intracellular Ca2+ levels and activate basolateral Ca2+-dependent K+ channels and 100 μM bumetanide (basal) to inhibit basolateral Na+/K+/2Cl- (NKCC) co-transporter.

Forskolin-Induced Swelling Assay Passage 3-15 organoids were seeded in 96-well plates, in 4 μl 70% matrigel droplet per well containing $\tilde{2}$5–30 organoids. The next day, organoids were incubated with 1.84 μM calcein green (Thermo Fisher Scientific C3100MP) for at least 30 min prior to addition of fsk at 0.02, 0.128, 0.8 or 5 μM concentrations, to determine cell viability. For CFTR potentiation, a single potentiator (3 μM VX-770 or 3 μM G1837 or 50 μM Gen) or dual potentiators (VX-770+G1837 or VX-770+Gen or G1837+Gen) was added together with fsk. Time-lapse images of organoid swelling were acquired at 10-min intervals for 60 min at 37°C using Zeiss Axio Observer Z.1 inverted microscope (Carl Zeiss, Jena, Germany), on an EC Plan-Neofluar 5x/0.16 M27 dry objective. Organoids were pre-incubated with 3 μM VX-809 or 3 μM VX-661 or 3 μM VX-445 or 3 μM VX-445+18 μM VX-661 for 24 h prior to

FIS for CFTR correction where indicated. Three wells were used per condition and each participant's FIS experiment was repeated 3 to 4 times.

Quantification of Forskolin-Induced Swelling Organoid swelling was quantified using a custom-built script. A segmentation strategy implemented using ImageJ/Fiji was performed on brightfield images. The raw image was processed with a gaussian blur (s=1.3) to reduce noise. After the directionality and magnitude of the local gradient was identified, pixels were classified as either 'Background', 'Ridge', 'Valley', 'Rising' or 'Falling' dependent on their neighbouring pixels along the previously calculated local directionality. Clean-up filters were applied that remove noise and small objects, such as ridges that only touched background pixels, and erosions to decrease rising and falling edges to better approximate object boundaries ('Peaks'). A size exclusion was applied that would discriminate debris in the sample preparation from organoids of interest. This segmentation strategy was used to identify area covered by organoid at each time point. The total surface area of organoid at 10-min intervals over 60 min post-fsk stimulation were calculated and normalized against t=0 to render the relative amount of swelling from t=0. The area under the curve, AUC (calculated increase in organoid surface area from t=0 to t=60; baseline=100%) was then calculated using GraphPad Prism software.

Quantification of CFTR-Mediated Ion Transport in Organoid-Derived Monolayers Short circuit current (Isc) measurements were performed under voltage-clamp conditions using VCC MC8 Ussing chambers (Physiologic Instruments, San Diego, CA). Cells were bathed in 20 mM HEPES buffered-Ringer solution containing (mM): 120 NaCl, 0.8 K_2HPO_4, 5 D-Glucose, 1.2 $MgCl_2$ and 1.2 $CaCl_2$. Ringer solutions were continuously gassed with 95% O2-5% CO2 and maintained at 37°C. 10 μM indomethacin was added to both apical and basal chambers and cells were stabilised for 15 min. Cells were then treated with pharmacological compounds (in order): 100 μM amiloride (apical) to inhibit epithelial sodium channel (ENaC)-mediated Na+ flux, vehicle control 0.01% DMSO or 10 μM G1837 (apical) to potentiate cAMP-activated currents, 5 μM forskolin (basal) to induce cAMP activation of CFTR, 30 μM CFTRinh-172 (apical) to inhibit CFTR-specific currents and 100 μM ATP (apical) to activate calcium-activated chloride currents. Isc in response to forskolin was considered as baseline activity (ΔIsc-Fsk) and Isc in response to forskolin and potentiator (ΔIsc-Fsk+Pot) was used as the measure of modulator response.

Immunofluorescence A rectal biopsy from a I37R/F508del participant was embedded in Tissue-Tek Optimal Cutting Temperature (OCT) compound (Sakura Finetek, CA) and snap frozen prior to storage at -80°C. The frozen biopsy was cut into 4 μm slice sections, and the sections were fixed in ice-cold methanol for 15 min. Intestinal organoids cultured from the I37R/F508del participant and organoid-derived monolayers cultured from a F508del/F508del participant were fixed in 4% paraformaldehyde and ice-cold methanol respectively for 15 min. Fixed samples were blocked using IF buffer (0.1% BSA, 0.2% Triton and 0.05% Tween 20 in PBS) with 10% normal goat serum (Sigma G9023) for 1 h at room temperature before incubation in primary antibodies overnight at 4°C. The biopsy section was stained with CFTR (1:50, Abcam ab2784) and E-cadherin (1:100, Cell Signalling 3195) antibodies. Intestinal organoids were stained with Ki67 (1:250, Abcam ab15580) and E-cadherin (1:250, Life Technologies 13-1700) antibodies. Organoid-derived monolayers were stained with ZO-1 (1:250, Life Technologies 61-7300) and E-cadherin (1:250, Life Technologies 13-1700) antibodies. On the following day, samples were washed with IF buffer 3 times, 5 min each and incubated with Alexa Fluor conjugated secondary antibodies (1:500, Life Technologies A-11029, A-21329) for 1 h at room temperature. Samples were mounted with Vectashield hardset antifade mounting medium containing DAPI (Vector Laboratories H-1500). Images were acquired using Leica TCS SP8 DLS confocal microscope (Leica Microsystems, Wetzlar, Germany), either on a 63x/1.4 or a 20x/0.75 objective. Images were processed using ImageJ (National Institutes of Health, Bethesda, MD).

Western Blotting Intestinal organoids were lysed with TNI lysis buffer (0.5% gepal CA-630, 50 mM Tris pH 7.5, 250 mM NaCl, 1 mM EDTA) [289] containing protease inhibitor cocktail (Roche 04693159001) on ice for 30 min. Lysates were then sonicated using the Bioruptor Pico (Diagenode, Liège, Belgium) at 4°C for 20 cycles of 30 sec on and 30 sec off. Lysates were spun down at 14,000 rpm at 4°C for 20 min and protein concentrations were determined using the BCA Protein Assay Kit (Thermo Fisher Scientific 23225). Lysates (100 μg per sample) were separated using NuPAGE 3 – 8% Tris-Acetate gels (Thermo Fisher Scientific EA0375BOX) at 100 V for 30 min, followed by 150 V until separation was complete. Proteins were transferred onto a nitrocellulose membrane using wet transfer at 20 V for 1 h at RT. The membrane was then incubated in 5% non-fat dry milk in phosphate-buffered saline containing 0.1% Tween (PBST) for 1 h at RT. CFTR bands were detected using anti-CFTR antibody 596 (1:500; University of North Carolina, Chapel Hill and Cystic Fibrosis Foundation) incubated at 4°C overnight. Protein bands were visualised using ECL Select detection reagent (Cytiva RPN2235) on the ImageQuant LAS 4000 (GE Healthcare, Chicago, IL). Calnexin was used as the loading control, detected using anti-calnexin antibody (1:1000; Cell Signalling Technology 2679). Protein band densitometry was performed using ImageJ (National Institutes of Health, Bethesda, MD). CFTR maturation in I37R/F508del and F508del/F508del organoids were estimated by measuring the level of mature mutant CFTR (band C) as a percentage of mature CFTR from WT organoids (% normal CFTR) [**vangoor2014**].

In Silico System Composition A 1-palmitoyl-2-oleoyl-sn-glycero-3-phosphocholine (POPC) bilayer was generated using the VMD membrane builder plugin (Humphrey et al., 1996) in which a model based on the phosphorylated human CFTR channel (PDB ID: 6MSM) was embedded [50]. The system was solvated with TIP3P water and neutralised with 0.15 M of potassium chloride ions [290]. The WT-CFTR system included 236 POPC molecules, 128 potassium ions, 140 chloride ions and 44503 water molecules.

Extended 6MSM Structure: Modelling the Unidentified Section of the R Domain The 6MSM structure was extended in order to resolve a previously unassigned section in the R domain. The R domain is 227 residues long (F630-H856) and is largely disordered [291]. In the 6MSM structure, the sidechains of 17 residues of this domain are labelled "UNKNOWN", due to inadequate electron density in the region. The first 8 residues are unstructured while the next 9 residues form an alpha helix. The distance between the end of the helix and the first visible residue in TMD2 (T845) is 20 Å [50]. Using VMD's autopsf plugin [178] we populated the side chains of the unknown section. Modeller 9.19 was then used to link the R domain to TMD2 at T845 [182]. 24 possible primary structure alignments of this region were simulated. The Root Mean Squared Deviation (RMSD) of the backbone alpha carbon atoms of the extended section with respect to the 6MSM structure was calculated over 300 ns of MD simulations. The most stable alignment was chosen from the lowest RMSD compared to the 6MSM structure. The most stable configuration was capped with the neutral forms of the C and N termini and incorporated into our CFTR model. Four other missing loops namely residues 410-434, 890-899, 1174-1201, 1452-1480 were reconstructed using Modeller 9.19, based on visual analysis and the lowest discrete optimised protein energy (DOPE) score [183]. The N and C termini of the CFTR model were capped with the physiological, charged termini.

Molecular Dynamics Simulation Protocols The 6MSM structure carries an engineered mutation to avoid the hydrolysis of the bound ATP, giving it a longer lifetime in the open conformation (E1371Q). This mutation was corrected to match the WT-CFTR sequence using the mutator plugin of VMD. The I37R missense mutation was constructed in the same way. GROMACS v2019.3 with the CHARMM36m forcefield was used for all MD simulations [89, 123]. Minimisation via a steepest descent algorithm was performed until all forces were below 24 kcal/mol/Å. This was followed by relaxation simulations of all heavy atoms in the system starting with a restraint of 10 kcal/mol/Å^2 and then halving this restraint every 200 ps in 15 iterations. Relaxation and production were run with 1 and 2 fs time steps, respectively. Relaxation was followed by 5 ns of equilibration. During relaxation, a Berendsen thermostat and barostat were applied, and for production a Nosé-Hoover and Parrinello-Rahman thermostat and barostat were applied respectively [121, 128, 292]. To maintain the area per lipid (APL) properties of the POPC membrane at experimental values during production runs, pressure coupling was applied in the z-direction normal to the membrane bilayer while the x-y dimensions of the cubic simulation volume was fixed [293]. While semi-isotropic pressure coupling better replicates membrane environments [294], this constant area approach was adopted to circumvent an issue with GROMACS 2019.3reb (https://gitlab.com/gromacs/gromacs/-/issues/2867). Production runs were extended up to 2 μs at 310 K with three replicates for all simple MD simulations. The last 1 μs of the longest simulations for each system were selected for further analysis. This was the longest time feasible to simulate with available computational resources. All RMSDs were calculated using the positions of alpha carbons with reference to the 6MSM experimental structure [50]. Analysis scripts were written in python using the MDAnalysis library [180, 181]. Bilayer thickness and area per lipid were calculated with the FATSLiM software package [295].

4.5 Acknowledgements

We thank the study participants and their families for their contributions. We also thank Sydney Children's Hospital's (SCH) Randwick respiratory department especially Leanne Plush, Amanda Thompson, Roxanne Strachan, and Rhonda Bell for the organization and collection of participant biospecimens. SAW is supported by an Australian National Health and Medical Research Council grant NHMRC_APP1188987. MA acknowledges support of a top-up scholarship from Cystic Fibrosis in Australia. MA and KA are supported by Australian Government Research Training Program Scholarship. We acknowledge for the generous provision of the L-Wnt3A cell line. Computations were performed on the Gadi HPC at the National Computational Infrastructure Center in Canberra and Artemis at the Sydney Informatics Hub in The University of Sydney. We thank Dr John R. Riordan (University of North Carolina– Chapel Hill) and Cystic Fibrosis Foundation for providing anti-CFTR antibody #596. Support statement: This work was supported in part by an Australian National Health and Medical Research Council grant (NHMRC_APP1188987), a Rebecca L. Cooper Foundation project grant, a Cystic Fibrosis Australia-The David Millar Giles Innovation Grant, Sydney Children Hospital Network Foundation, and Luminesce Alliance Research grants.

4.6 Author Contributions

Conception and design: SAW and AJ. Recruitment and consent: LF and SAW. Collection of rectal biopsies: CYO and LF. Ion transport assay: NTA. Culturing of organoids: NTA, SLW, and SAW. FIS microscopy: IS, KA, and SLW. FIS scripts: MC and RW. FIS analysis: NTA and SLW. Immunofluorescence microscopy: SLW. Western blot: SLW. Molecular Dynamics: MA, PC, RG, and SK. CFTR sequence alignment: AC. Figure preparation: SLW, MA, NTA, AC, KA, and SAW. Writing – original draft: SLW, MA, and SAW. Review and editing: SAW, KA, RG, SK, and LF with intellectual input from all other authors. Supervision: SAW and SK.

4.7 Delcaration of Interests

SAW is the recipient of a Vertex Innovation Grant (2018) and a TSANZ/Vertex Research Award (2020). Both are unrelated and outside of the submitted manuscript. AJ has received consulting fees from Vertex on projects unrelated to this study. CYO has acted as consultant and is on advisory boards for Vertex pharmaceuticals. These works are unrelated to this project and manuscript. All other authors declare no conflict of interest.

Chapter 5

Molecular Dynamics and Theratyping in Airway and Gut Organoids Reveal R352Q-CFTR Conductance Defect

Cells have a mind of their own

-Shafagh Waters (personal communication)

Abstract

A significant challenge to making targeted CFTR modulator therapies accessible to all individuals with cystic fibrosis (CF) are many mutations in the CFTR gene that can cause CF, most of which remain uncharacterized. Here, we characterized the structural and functional defects of the rare CFTR mutation R352Q – with potential role contributing to intrapore chloride ion permeation – in patient-derived cell models of the airway and gut. CFTR function in differentiated nasal epithelial cultures and matched intestinal organoids was assessed using ion transport assay and forskolin-induced swelling (FIS) assay respectively. CFTR potentiators (VX-770, GLPG1837 and VX-445) and correctors (VX-809, VX-445 +/- VX-661) were tested. Data from R352Q-CFTR were compared to that of twenty participants with mutations with known impact on CFTR function. R352Q-CFTR has residual CFTR function which was restored to functional CFTR activity by CFTR potentiators but not the corrector. Molecular dynamics (MD) simulations of R352Q-CFTR were carried out which indicated the presence of a chloride conductance defect, with little evidence supporting a gating defect. The combination approach of in vitro patient-derived cell models and in silico MD simulations to characterize rare CFTR mutations can improve the specificity and sensitivity of modulator response predictions and aid in their translational use for CF precision medicine.

Chapter 6

Unique S945L-CFTR defect Restored by CFTR Modulator Co-Therapy In Vitro Correlates with In Vivo Biomarkers Post-Therapy

Eventually you'll realise you're just two days from anywhere.
-David Goodwin (personal communication)

chap:s945l

Chapter 7

Resolving a Conducting Conformation of CFTR Using Free Energy Calculations

You wanna fight?
- Doctor Zachary Shadrach Cohen Picker. Layperson. (personal communication on the 423 bus)

Abstract

The misfunction of the CFTR gene causes Cystic Fibrosis. This protein conducts both chloride and bicarbonate. Understanding how CFTR conducts ions is critical to ongoing drug discovery efforts to treat Cystic Fibrosis. Existing structures of CFTR raise unresolved qustions as to how CFTR conducts ions as they exhibit a constriction smaller than the ions themselves. This indicates that there must be some level of conformational change for ions to pass through this structure. Here we present innovative simulation techniques combining principal component analysis of 8 microseconds of protein simulations and new advances in free energy calculations to resolve the full conduction pathway in human CFTR. We also propose experimental single ion channel electrophysiology techniques to experimetnally test whether this conformation indeed exists.

The findings of this study demonstrate that computational power and protein forcefields are now sufficiently developed to take experimental protein structures as a starting point. We can use them to explore the conformational neighbourhood around a given structure to discover more physiologically relevant protein conformations.

7.1 Introduction

7.2 Results

7.2.1 The outer pore in the CryoEM structure of phosphorylated human CFTR is not sufficiently open to conduct anions

7.2.2 Using metadynamics and long simulations we can dilate the pore to discover a conducting conformation

7.2.3 Proving the discovered conformation is capable of ion conduction with umbrella sampling

7.2.4 This open conformation gives rise to a novel salt bridge

7.2.5 This open state can be used to study disease causing mutations such as R334W in the outer pore

Mutagenesis studies of the R334 amino acid noticed that many different mutations appear to result in ephys readings hwich would indicate a loss of function, including K which seems surprising [296–298].

7.3 Discussion

The present study diverges in an important way from existing simulation investigations of protein conformational changes in the literature. Present studies have focussed on recreating intermediate free energy landscapes between *known* endpoints [46, 299]. Such approaches are critical to the development of molecular techniques in order to understand the energetics and kinetics of protein systems. However, these studies are inherently limited to the availability of high quality experimental 3d structures. . Protein forcefields may have their issues but they are now of sufficient quality that they can now be used, alongside considerable computer power to investigate parts of the conformational landscape which have critical functional roles but are not covered by experimental structures. This is akin to the development of an unsupervised vs. supervised machine learning algorithms. Each approach is powerful but has its own domain of applicability and drawbacks.

I predict that as free energy calculations are increasingly used to study protein systems we will see a delineation between *untargetted* and *targetted* MD methods.

As shown by the results in the present study, the difficulty in converging a free energy landscape with collective variables derived *ab initio* from long classical MD simulations can be difficult as the CVs are very likely going to be suboptimal. Machine learning techniques, more sophisticated than the simple PCA algorithm used here would likely do a much better job of choosing quickly converging CVs.

The conformational changes investigated in this study push against the lipid bilayer.

This means that the kinetics and energetics of the transitions we have discovered will be highly dependent on the composition of lipids used in the study. It is well understood that the bilayer composition plays an important role in CFTR regulation and the clinical implications of this are an active area of research [300, 301]. It would therefore be an interesting study to repeat similar free energy calculations with different bilayer compositions to understand how they might regulate such conformational changes.

Understanding the open structure of CFTR has important implications for the drug discovery efforts to treat Cystic Fibrosis and sheds light on other important clinical questions. Specificlaly, selectivity of bicarbonate has been found to play an important role in pancreatic sufficiency of patients. The elucidation of basic CFTR function using simulations heralds an exciting new era of Cystic Fibrosis research. We are performing molecular medicine with atomic precision.

The predictions of a stable salt bridge in section 7.2.4 fill a recent gap in the literature. The elegant study on the R117H mutation from Simon and Csnady's group [302] discovered that a long standing conclusion that R117 made a connection with E1126 was incorrect and in fact R117 makes a stable hydrogen bond with E1124. This study did not closely investigate the role of E1126, observing that the E1126P mutant had slower closing kinetics but was unable to definitely explain why, suggesting that ECL4 might move slower in this mutation. This leaves the partner of E1126 unknown. One study investigating the blockage of CFTR by zinc postulated at an interaction between R334 and E1126 [303]. Here, the researchers tested the inhibition of chloride conduction in the presence of zinc in R334C-CFTR. They found strong evidence that R334C-CFTR was blocked by Zinc ions, as no current was recorded in the presence of zinc. Because zinc has a +2 charge they suspected that a nearby negative amino acid might might play a role in binding the zinc cation. Subsequent experiments They found that the mutant R334C/E1126A-CFTR was no longer inhibited by zinc ions. This is consistent with our findings that R334 and E1126 may indeed form a salt bridge, coming closer together in the conducting conformation compared to how far they are in the cryoEM structure.

Previously it would have been very difficult to discover this interaction experimentally because R334 has plays such an important role in the conductance and selectivity of the channel. With the use of the atomic resolution offered by MD simulations we have been able to fill this gap in the experimental literature, demonstrating the power of *in silico* methods for studying protein dynamics.

7.4 Conclusion

7.5 Methods Details

Chapter 8

Concluding Remarks: Where next for Molecular Studies of CFTR.

We have more problems than hands.

- Eduardo Perozo (personal communication)

As figure ?? demonstrates, basic science discoveries concerning CF have had a direct effect on the life expectancy of patients. This proves the utility . The work in this thesis is a small example of how abstract physical models, such as those outlined in 2, can be applied to help real patients in a community such as allowing patients in Sydney Children's hospital to access medications which could add decades to their life span. We are entering an exciting era of biophysical research where advances in theoretical methods, computing power and experimental techniques are beginning to drive advances in each other at a frenetic pace. An example can be seen in the development of Alphafold. The maturation of cryoEM allowed the discovery of new protein folds which Alphafold's machine learning algorithms could then learn from. Now that the algorithm had these folds in hand it could predict entire proteomes. This now means that structural biologists can use the predictions of alphafold to solve even more structures more quickly. Eventually, more and more of the experimental work currently involved in biology will move onto the silicon chip, while experimental techniques will advance in other areas. Similarly, the theoretical model argued for in this thesis will eventually allow for patient assessments to be made *in silico*

The preceeding chapters have given technical and molecular details about how Cystic Fibrosis is caused by rare mutations and further demonstrated that these mutations may be treated by existing small molecule drugs. The unique molecular fingerprint of each mutation indicates that in order to deliver better outcomes to patients a more personalised approach is necessary to the choice of medication. For this personalisation to be possible there are some basic questions about CFTRs structure and function that remain to be answered. Some of these questions simulations are uniquely placed to answer.

Additionally, I'd like to propose a theoretical model for the action of these drugs which would appear to suggest that more patients should be given access to existing medications and inform the design of future generations of CFTR modulators.

As such, the areas which deserve the most attention for molecular studies of CFTR are to address the controversies surrounding its structure and elucidate the action of drug binding. Together these will allow a clearer understanding of CFTRs function in health and misfunction in disease, enableing more targetted drug development. Current generation modulators have been identified using high throughput screening, the new generation of high computational power, artificial intelligence and structural data will allow a more targetted, perhaps even mutation specific approach to the design of new modulators.

8.1 Addressing Controversies Surrounding the Structure of CFTR

The most recent structure of human CFTR in a phosphorylated environment has some interesting features which have lead to some controversies in the literature. After spending significant time researching them I have conducted an extensive literature review in order to learn more about of these concerns. The released structure of activated, human CFTR has two features that have caused some in the CF field to suggest issues with this structure. Firstly, this structure is not sufficiently open to conduct chloride ions. Chloride ions have a diameter of $1.7\mathring{A}$ while the structure has a constriction of $1.1\mathring{A}$[**Zhang2018**]. So, there must be some level of conformational changes, even if chloride were to move through the channel completely dehydrated. This becomes even more of an issue when considers the experimental evidence where much larger anionic species such as bicarbonate and glutathione were shown to permeate through the channel[304]. This suggests that there is a much larger conformation which has not been observed experimentally or in simulations. This was the motivation for chapter 7 of this thesis. Some studies have been performed in order to study the possible permeation paths of chloride but they have not addressed the pressing question of how larger ions might permeate the channel. Bicarbonate in particular is of great physiological importance as there is a high correlation between the channel's ability to permeate bicarbonate and the pancreatic sufficiency of a patient carrying the mutation. In light of this, structural knowledge of a fully open conformation of CFTR is critical to a personalised approach to the treatment of CFTR.

The second and harder to resolve controversy concern the role of TM8. This transmembrane helix has an unusual bend in the middle of the plasma membrane. This is not something seen before in ABC transporters of this type. So it has led to some open questions as to how this bend might contribute to the function of the channel *or* how it might be an artifact of the imaging process. For the former case, the structural biologists in the Chen lab proposed a mechanism whereby the upper hinge of TM8 swings 55^o during the transition to the open state. This mechanism would give justification of the pathogenesis of certain mutaions such as L927P.

The arguments for the bend in the helix appear to be unphysical. In cryoEM structures, we can observe that the bent conformation is stabilised by salt bridges R347-D924 and E873-R933. The former bond has been well studied experimentally and was expected in the 3d structure. Additionally, all hydrogen bonds along the in the bent helix are . Been observed to be stable in MD [305] . is energetically stable.

However, there are still some unanswered questions for the discrepancy between the solved human and solved chicken structures. Certain salt bridges are not present in the latter structure and so single channel electrophysiology experiments may be able to resolve these issues. For example, in 6MSM, the human structure of CFTR there is a salt bridge between amino acids 933 and D873 which is not present in the chCFTR structures. This bond is also present in the zCFTR structure 5W81. If a charge swapped mutant such as R933E/E873R restores WT-like gating behaviour to the channel it would be strong evidence for the unwound conformation of TM8.

The two proposed conformations also have vastly different ion permeation pathways, and so blockers engineered to target one conformatoin over the other would also go a long way to answering these questions. The available evidence strongly favors the R334 pathway between TM1 and TM6. Such as experiments to demonstrate the blockage of current with zinc have shown that mutations to R334 strongly suggest that chloride permeates along this route. Additionally, trhere are several disease causing mutations in the region surrounding R334, such as R334W, R117H, E116K, D110H, I336K[204]. This permeation route also explains the rationale behind the gain of function mutation F337A []. Determining which model is correct has wide implications for creating the next generation of mutation targeted potentiator class drugs.

[306] also concluded that the chloride opening lined TM6. Which is more consistent with the 6MSM model than the chicken structure.

The chicken structure has undermised NBDs, this is inconsistent with biochemical assays that dimerisation is tightly coupled to the dimerisation of NBDs[307, 308].

Mutagenesis studies of the outer pore would strongly support the Chen structure over the chicken streucture. In particular, Paul Linsdell performed very careful experiments to measure blockage chloride blockage and found that several residues play an important role in the permeation of chloride in the outer pore. In the chicken structure these residues are occluded and far from the chloride permeation pathway. However, in the Chen structure they appear to play an important role in the permeation of chloride. This will be discussed in detail in 7.

A study assessing the accuracy of Alphafold's predictions of transmembrane protein structures found an intersting result. When alphafold made predictions that involved the use of templates it predicts the unwound conformation of TM8. On the other hand, when templates are removed from alphafolds predictions it predicts a straight TM8 conformation, very similar to that found in chCFTR. The authors of this study suggested the reason for the discrepancy was due to the use of detergents in the deteremination of the structure of hCFTR. However, careful reading of cryo-EM literature reveals no examples where the use of detergents has resulted in such drastic conformational changes. One of the few examples where both detergents and native-like nanodisks were used to determine the structure of a protein are the determinations of the structure of TRPV1. These studies revealed no difference to the backbone helices but did suggest important information about the importance of different interactions with lipids[309]. A lot of work has been performed in order to create detergents which reflect a native lipid environment and these are the species that were used in the determination of the human CFTR structure (albeit at a higher than optimal concentration) [**Zhang2018**, 309, 310].

The authors of the chCFTR paper suggested that the different expression systems used in the two studies could be the reason for the discrepancy between the two systems, due it their different post translational processing apparatus. Although both groups used mammalian cell lines the chCFTR paper used hamster? cells [311] and the Chen lab used HEK293S cells. The chicken structure also underwent significant mutations in order to be locked open and imaged. The regulatory insertion was deleted in order to aid in the purification of the protein.

Figure ?? shows the large diversity of structures in type IV ABC transporters, many of which also exhibit bends within transmembrane helices[214]. Although none of these structures exhibit such a bend in TM8 specifically, the

?? gives strong evidence that TM8 must line the conduction pore. Due to its proximity to F337 and other pore lining amino acids. This is more consistent with the Chen structure than the chicken structure where 337 is far from the conduction pore.

8.2 A Physics Motivated Model for the Molecular Modulation of Mutant CFTR

In chapter ??, ??, ?? and 7 we have analysed a disease causing mutation in detail in order to understand *how* they cause CFTR to misfunction. What we have found is a large diversity of molecular phenotypes which may cause disease. What is thus remarkable is that the *in vitro* component of these papers all demonstrate that these mutations are responding to the same drugs, albeit with differing efficacies.

The above model in figure ?? gives a rational, physical basis for the wide range of molecular phenotypes that CFTR modulators appear capable of treating. This same model would appear to argue that for most missense mutations we would expect them to respond to some sort of CFTR modulator.

8.3 A Physics Motivated Approach to Precision Medicine in Cystic Fibrosis

In recent years there have been a slew of rare CF-causing genotypes discovered in populations with low rates of Cystic Fibrosis compared to Caucasians. These genotypes from Asia and the Middle East are often ultra-rare leading to poor outcomes for these patients, particular lb when local health care is sub standard [].

The ongoing discovery of rare mutations highlights the importance of this personalised approach to the treatment of CF. The process for developing potentiator class drugs was by studying the G551D mutation. Drugs that were found to restore function for this rare mutation are now widely used by sufferers of cystic fibrosis []. The study of rare mutations N1303K which currently do not respond to drugs may lead to the discovery of more effective compounds to treat cystic fibrosis such as. Thus, the approach to the treatment of Cystic Fibrosis is intersectional, as more rare mutations are discovered and treated the better the outcomes for all patients with Cystic Fibrosis will be. Each rare mutation sheds light on the function of CFTR and lets us understand and treat the root cause of the disease better.

This model makes clear what the pressing questions are in molecular cystic fibrosis research. These are three, primarily. Firstly and fundamentally, there is the task of finding the molecular details of the functional landscape in figure ??. Secondly, there is the elucidation of molecular misfunction of mutations. We must find *where* in the functional landscape these mutations are causing issues, which will also tell us how. This will allow us to meaningfully group mutations into molecular theratypes. This leads us to the final task: Finding drugs to treat each of these theratypes. My personal opinion for the direction of each of these tasks is layed out in the subsequent sections. Should the above model prove successful, such an approach approach to personalised medicine could be considered when studying other monogenic diseases such as Muscular Dystrophy, Sickle Cell Anemia and Huntington's disease []. Once single genes are understood the understanding could be built outward to encompass more complex diseases which involve the interactions between many genes such as diabetes and cancer. The future of personalised medicine is bright and it is possible that much of its breakthroughs will come from the molecular level.

8.4 Outstanding fundamental questions about CFTR Function

Firstly, there are controversies surrounding the structure of this protein. Primarily determining resolving the physiologically relevant conformations of TM8 and the conduction pathway of ions through CFTR.

Secondly, there are questions about how tightly the coupling of hydrolysis of ATP in the NBDs is to the conduction of ions.

8.5 Grouping of Theratypes

As outlined in figure **??** earlier, the molecular fingerprint of a mutation can be quite complex and ongoing work is needed to meaningfully group these mutations into more clinically meaningful categories. My prediction is that the conventional 7 classes will be more and more finely defined in order to choose CFTR modulators which are more specific to a patient's genotype and epithelial phenotype.

8.6 Resolving Drug Action

Closely related to the above two categories is the mechanism of action for existing drugs and the development of new drugs. The reason we were able to discover potentiator class drugs is through the study of a rare mutation. High throughput screening of small molecules in restoring the gating class mutation G551D led to the discovery of gating class drugs. In this way we can see how the study of rare CF can lead to better outcomes for all sufferers of the disease. This is especially pertinent as more rare genotypes are discovered in non-Caucasian populations such as in Asia and the Middle East.

Additionally, it should be obvious from this work that the action of these drugs is are highly dependent on the molecular function of CFTR. These small molecule drugs *select* for a physiologically present conformation, so we would wish to design molecules which select for conformations which deliver the most clinic benefit. This is non-trivial and I believe combination of careful molecular experiments, such as those of from the laboratories of Tzyh-Chang Hwang, Christine E. Bear, László Csanády and Paul Linsdell [215, 312, 313], and molecular simulations such as those found in this thesis and studies from the labs of John Paul Monron and Isabelle Callebaut [**hoffmann2018**].

An appendix of uncollected thoughts.

- Considering that the passive immune system consists of more inanimate layers of the body such as the skin would it not also make sees that social elements of our behaviour form part of our immune system as well? Think about the visceral reaction we have toward fecal matter or someone vomiting. These are neural queues to change our behaviour. Could our pandemic response, social distancing and vaccine development then be viewed as part of our adaptive immunity.
- V(D)J recombination is a process where the body somatically shuffles parts other genome in order to create variable antibodies which target pathogenic substrates with high affinity and specificity. It's amazing and potentially has important implications for protein physics and protein bioinformatics.
- There are considerable efforts to develop drug discovery mechanisms for specific cell types [314] as this has the potential to reduce the off target side effects and increase therapeutic efficacy.
- The megaplate experiment[315] is an illustrative example of a Strange Loop [2].

Bibliography

(1) Martin, G. R. R., *A Game of Thrones: Book 1*; Voyager GB: London, 1997.

(2) Hofstadter, D. R., *I Am a Strange Loop*; BasicBooks: New York, NY, 2007.

(3) Phillips, R.; Kondev, J.; Theriot, J.; Garcia, H.; Kondev, J., *Physical Biology of the Cell*, Second edition, first issued in paperback; Garland Science: Boca Raton London New York, 2012.

(4) Dawkins, R., *The Selfish Gene*, New ed; Oxford University Press: Oxford ; New York, 1989.

(5) Hofstadter, D., *Godel, Escher, Bach: An Eternal Golden Braid*, 1st edition; Basic Books: New York, 1999.

(6) Goodsell, D. S., *The Machinery of Life*, 2nd ed. 2009 edition; Copernicus: New York, 2009.

(7) Goodsell, D. S., *Atomic Evidence: Seeing the Molecular Basis of Life*, Softcover reprint of the original 1st ed. 2016 edition; Copernicus: 2018.

(8) Goodsell, D. S.; Olson, A. J.; Forli, S. *Trends in Biochemical Sciences* **2020**, *45*, 472–483, DOI: 10.1016/j.tibs.2020.02.010.

(9) Chen, F., *Introduction to Plasma Physics and Controlled Fusion*, 2018.

(10) Dawkins, R., *The Extended Phenotype: The Long Reach of the Gene*, Reprint edition; Oxford University Press, USA: Place of publication not identified, 2016.

(11) Griffiths, D. J., *Introduction to Electrodynamics*, 4th edition; Cambridge University Press: Cambridge, 2017.

(12) Reif, F., *Fundamentals of Statistical and Thermal Physics*, 1st edition; Waveland Press, Inc.: 2009.

(13) Zuckerman, D. M., *Statistical Physics of Biomolecules: An Introduction*, 1st edition; CRC Press: Boca Raton, 2010.

(14) Salzberg, S. L. *BMC Biology* **2018**, *16*, 94, DOI: 10.1186/s12915-018-0564-x.

(15) Adams, D., *The Hitch Hiker's Guide to the Galaxy*; Pan Original; Pan Books: London, 1979.

(16) Frauenfelder, H.; Wolynes, P. G. *Physics Today* **1994**, *47*, 58–64, DOI: 10.1063/1.881414.

(17) Varn, D. P.; Crutchfield, J. P. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **2016**, *374*, 20150067, DOI: 10.1098/rsta.2015.0067.

(18) Watson, J. D., *Avoid Boring People: Lessons from a Life in Science*; Vintage Books: 2010.

(19) Picker, Z. The Gravity of Particle Physics: Dark Matter, Black Holes, and Axions, Ph.D. Thesis, U. Sydney (main), 2022.

(20) Carroll, S. M. *Journal of Consciousness Studies* **2021**, *28*, 16–31.

(21) Moy, G.; Corry, B.; Kuyucak, S.; Chung, S.-H. *Biophysical Journal* **2000**, *78*, 2349–2363, DOI: 10.1016/S0006-3495(00)76780-4.

(22) Corry, B.; Kuyucak, S.; Chung, S. H. *Biophysical Journal* **2000**, *78*, 2364–2381, DOI: 10.1016/S0006-3495(00)76781-6.

(23) Hille, B., *Ion Channels of Excitable Membranes*; Sinauer Associates: 2001.

(24) Catterall, W. A. *Cold Spring Harbor Perspectives in Biology* **2011**, *3*, a003947, DOI: 10.1101/cshperspect.a003947.

(25) Muthuswamy, S. K.; Xue, B. *Annual review of cell and developmental biology* **2012**, *28*, 599–625, DOI: 10.1146/annurev-cellbio-092910-154244.

(26) Levin, M. *The Journal of Physiology* **2014**, *592*, 2295–2305, DOI: 10.1113/jphysiol.2014.271940.

(27) Levin, M. *Molecular Biology of the Cell* **2014**, *25*, 3835–3850, DOI: 10.1091/mbc.E13-12-0708.

(28) Santos, R.; Ursu, O.; Gaulton, A.; Bento, A. P.; Donadi, R. S.; Bologa, C. G.; Karlsson, A.; Al-Lazikani, B.; Hersey, A.; Oprea, T. I.; Overington, J. P. *Nature Reviews. Drug Discovery* **2017**, *16*, 19–34, DOI: 10.1038/nrd.2016.230.

(29) Stansfeld, P. J.; Sutcliffe, M. J.; Mitcheson, J. S. *Expert Opinion on Drug Metabolism & Toxicology* **2006**, *2*, 81–94, DOI: 10.1517/17425255.2.1.81.

(30) Kaczorowski, G. J.; McManus, O. B.; Priest, B. T.; Garcia, M. L. *The Journal of General Physiology* **2008**, *131*, 399–405, DOI: 10.1085/jgp.200709946.

(31) Waszkielewicz, A.; Gunia, A; Szkaradek, N; Słoczyńska, K; Krupińska, S; Marona, H *Current Medicinal Chemistry* **2013**, *20*, 1241–1285, DOI: 10.2174/0929867311320100005.

(32) Hodgkin, A. L.; Huxley, A. F. *The Journal of Physiology* **1952**, *117*, 500–544, DOI: 10.1113/jphysiol.1952.sp004764.

(33) Hodgkin, A. L.; Huxley, A. F.; Katz, B. *The Journal of Physiology* **1952**, *116*, 424–448.

(34) Hodgkin, A. L.; Huxley, A. F. *The Journal of Physiology* **1952**, *116*, 497–506.

(35) Hodgkin, A. L.; Huxley, A. F. *The Journal of Physiology* **1952**, *116*, 473–496.

(36) Hodgkin, A. L.; Huxley, A. F. *The Journal of Physiology* **1952**, *116*, 449–472.

(37) Sansom, M. S. *Progress in Biophysics and Molecular Biology* **1991**, *55*, 139–235, DOI: 10.1016/0079-6107(91)90004-C.

(38) Roux, B.; Karplus, M. *The Journal of Physical Chemistry* **1991**, *95*, 4856–4868, DOI: 10.1021/j100165a049.

(39) Allen, T. W.; Baştuğ, T.; Kuyucak, S.; Chung, S.-H. *Biophysical Journal* **2003**, *84*, 2159–2168, DOI: 10.1016/S0006-3495(03)75022-X.

(40) Allen, T. W.; Andersen, O. S.; Roux, B. *Proceedings of the National Academy of Sciences* **2004**, *101*, 117–122, DOI: 10.1073/pnas.2635314100.

(41) Chung, S.-H.; Kuyucak, S. *European Biophysics Journal* **2002**, *31*, 283–293, DOI: 10.1007/s00249-002-0216-4.

(42) Tieleman, D. P.; C. Biggin, P.; R. Smith, G.; S. P. Sansom, M. *Quarterly Reviews of Biophysics* **2001**, *34*, 473–561, DOI: 10.1017/S0033583501003729.

(43) Rashid, M. H.; Heinzelmann, G.; Huq, R.; Tajhya, R. B.; Chang, S. C.; Chhabra, S.; Pennington, M. W.; Beeton, C.; Norton, R. S.; Kuyucak, S. *PLOS ONE* **2013**, *8*, e78712, DOI: 10.1371/journal.pone.0078712.

(44) Li, J.; Shen, R.; Reddy, B.; Perozo, E.; Roux, B. *Science Advances* **2021**, *7*, eabd6203, DOI: 10.1126/sciadv.abd6203.

(45) Vandenberg, J.; Lau, C.; Flood, E.; Hunter, M.; Ng, C.-A.; Bouwer, J.; Stewart, A.; Perozo, E.; Allen, T. *Structural Basis for Rapid Voltage Dependent Inactivation of HERG Potassium Channels*; Preprint; In Review, 2021, DOI: 10.21203/rs.3.rs-1105661/v1.

(46) Lev, B.; Allen, T. W. *Journal of Computational Chemistry* **2020**, *41*, 387–401, DOI: 10.1002/jcc.26102.

(47) Chen, I.; Pant, S.; Wu, Q.; Cater, R. J.; Sobti, M.; Vandenberg, R. J.; Stewart, A. G.; Tajkhorshid, E.; Font, J.; Ryan, R. M. *Nature* **2021**, *591*, 327–331, DOI: 10.1038/s41586-021-03240-9.

(48) Sham, S. S.; Shobana, S.; Townsley, L. E.; Jordan, J. B.; Fernandez, J. Q.; Andersen, O. S.; Greathouse, D. V.; Hinton, J. F. *Biochemistry* **2003**, *42*, 1401–1409, DOI: 10.1021/bi0204286.

(49) Doyle, D. A.; Cabral, J. M.; Pfuetzner, R. A.; Kuo, A.; Gulbis, J. M.; Cohen, S. L.; Chait, B. T.; MacKinnon, R. *Science* **1998**, *280*, 69–77, DOI: 10.1126/science.280.5360.69.

(50) 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.

(51) Roux, B.; Karplus, M. *Journal of the American Chemical Society* **1993**, *115*, 3250–3262, DOI: 10.1021/ja00061a025.

(52) Russell, J. Anton 3 Is a 'Fire-Breathing' Molecular Simulation Beast, https://www.hpcwire.com/2021/09/01/anton-3-is-a-fire-breathing-molecular-simulation-beast/, 2021.

(53) Liou, J.-W.; Hung, Y.-J.; Yang, C.-H.; Chen, Y.-C. *PLoS ONE* **2015**, *10*, e0117065, DOI: 10.1371/journal.pone.0117065.

(54) Foucault, M., *The Birth of the Clinic: An Archaeology of Medical Perception*, Reprint edition; Vintage: New York, 1994.

(55) DePristo, M. A.; Weinreich, D. M.; Hartl, D. L. *Nature Reviews Genetics* **2005**, *6*, 678–687, DOI: 10.1038/nrg1672.

(56) Csanády, L.; Vergani, P.; Gadsby, D. C. *Physiological Reviews* **2019**, *99*, 707–738, DOI: 10.1152/physrev.00007.2018.

(57) Kochanski, Z. *Philosophy of Science* **1973**, *40*, 29–50, DOI: 10.1086/288494.

(58) Liu, E. T. *Cell* **2005**, *121*, 505–506, DOI: 10.1016/j.cell.2005.04.021.

(59) Mogilner, A. *Molecular Biology of the Cell* **2016**, *27*, 3377–3378, DOI: 10.1091/mbc.E16-09-0673.

(60) Covert, M. W.; Gillies, T. E.; Kudo, T.; Agmon, E. *Cell Systems* **2021**, *12*, 488–496, DOI: 10.1016/j.cels.2021.05.014.

(61) Jumper, J. et al. *Nature* **2021**, *596*, 583–589, DOI: 10.1038/s41586-021-03819-2.

(62) Benyus, J. M., *Biomimicry: Innovation Inspired by Nature*, Nachdr.; Perennial: New York, NY, 2009.

(63) Anonymous *The Economist* **2019**, *431*, 11.

(64) Scown, C. D.; Keasling, J. D. *Nature Biotechnology* **2022**, *40*, 304–307, DOI: 10.1038/s41587-022-01248-8.

(65) Cheng, Y.; Grigorieff, N.; Penczek, P. A.; Walz, T. *Cell* **2015**, *161*, 438–449, DOI: 10.1016/j.cell.2015.03.050.

(66) Callaway, E. *Nature* **2015**, *525*, 172–174, DOI: 10.1038/525172a.

(67) Callaway, E. *Nature* **2020**, *578*, 201–201, DOI: 10.1038/d41586-020-00341-9.

(68) Aidley, D. J.; Stanfield, P. R., *Ion Channels: Molecules in Action*; Cambridge University Press: 1996.

(69) Marion, D. *Molecular & Cellular Proteomics* **2013**, *12*, 3006–3025, DOI: 10.1074/mcp.O113.030239.

(70) Sanderson, M. J.; Smith, I.; Parker, I.; Bootman, M. D. *Cold Spring Harbor protocols* **2014**, *2014*, pdb.top071795, DOI: 10.1101/pdb.top071795.

(71) Frauenfelder, H.; Chan, S. S.; Chan, W. S.; Austin, R. H., *The Physics of Proteins: An Introduction to Biological Physics and Molecular Biophysics*; Biological and Medical Physics, Biomedical Engineering; Springer New York: New York, NY, 2010, DOI: 10.1007/978-1-4419-1044-8.

(72) Drenth, J., *Principles of Protein X-Ray Crystallography*, 3rd edition; Springer: New York Heidelberg, 2006.

(73) Silva, D.; Santos, G.; Barroca, M.; Collins, T. *Methods in Molecular Biology (Clifton, N.J.)* **2017**, *1620*, 87–100, DOI: 10.1007/978-1-4939-7060-5_5.

(74) *Nature Cell Biology* **2019**, *21*, 1463–1463, DOI: 10.1038/s41556-019-0434-y.

(75) Feynman, R. P. *International Journal of Theoretical Physics* **1982**, *21*, 467–488, DOI: 10.1007/BF02650179.

(76) Braun, E.; Gilmer, J.; Mayes, H. B.; Mobley, D. L.; Monroe, J. I.; Prasad, S.; Zuckerman, D. M. *Living journal of computational molecular science* **2019**, *1*, 5957, DOI: 10.33011/livecoms.1.1.5957.

(77) Gapsys, V.; de Groot, B. L. *eLife* **2020**, *9*, ed. by Faraldo-Gómez, J. D.; Grossfield, A., e57589, DOI: 10.7554/eLife.57589.

(78) Pohorille, A.; Jarzynski, C.; Chipot, C. *Journal of Physical Chemistry B* **2010**, *114*, 10235–10253, DOI: 10.1021/jp102971x.

(79) Sherrill, C. D., 7.

(80) *Dynamics of Molecular Collisions: Part B*; Miller, W. H., Ed.; Springer US: Boston, MA, 1976, DOI: 10.1007/978-1-4757-0644-4.

(81) van Mourik, T.; Bühl, M.; Gaigeot, M.-P. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **2014**, *372*, 20120488, DOI: 10.1098/rsta.2012.0488.

(82) Luo, Z.; Qin, X.; Wan, L.; Hu, W.; Yang, J. *Frontiers in Chemistry* **2020**, *8*.

(83) Kresse, G.; Furthmüller, J. *Computational Materials Science* **1996**, *6*, 15–50, DOI: 10.1016/0927-0256(96)00008-0.

(84) Lemkul, J. A. In *Progress in Molecular Biology and Translational Science*, Strodel, B., Barz, B., Eds.; Computational Approaches for Understanding Dynamical Systems: Protein Folding and Assembly, Vol. 170; Academic Press: 2020, pp 1–71, DOI: 10.1016/bs.pmbts.2019.12.009.

(85) Mackerell Jr., A. D. *Journal of Computational Chemistry* **2004**, *25*, 1584–1604, DOI: 10.1002/jcc.20082.

(86) MacKerell, A. D.; Feig, M.; Brooks, C. L. *Journal of the American Chemical Society* **2004**, *126*, 698–699, DOI: 10.1021/ja036959e.

(87) Mackerell, A. D.; Feig, M.; Brooks, C. L. *Journal of Computational Chemistry* **2004**, *25*, 1400–1415, DOI: 10.1002/jcc.20065.

(88) Ramachandran, G. N.; Ramakrishnan, C.; Sasisekharan, V. *Journal of Molecular Biology* **1963**, *7*, 95–99, DOI: 10.1016/S0022-2836(63)80023-6.

(89) Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; de Groot, B. L.; Grubmüller, H.; MacKerell Jr, A. D. *Nature Methods* **2016**, *14*, 71–73, DOI: 10.1038/nmeth.4067.

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

(91) Tolmachev, D. A.; Boyko, O. S.; Lukasheva, N. V.; Martinez-Seara, H.; Karttunen, M. *Journal of Chemical Theory and Computation* **2020**, *16*, 677–687, DOI: 10.1021/acs.jctc.9b00813.

(92) Savelyev, A.; MacKerell, A. D. *The Journal of Physical Chemistry B* **2015**, *119*, 4428–4440, DOI: 10.1021/acs.jpcb.5b00683.

(93) Jämbeck, J. P. M.; Mocci, F.; Lyubartsev, A. P.; Laaksonen, A. *Journal of Computational Chemistry* **2013**, *34*, 187–197, DOI: 10.1002/jcc.23117.

(94) Chari, S. S. N.; Dasgupta, C.; Maiti, P. K. *Soft Matter* **2019**, *15*, 7275–7285, DOI: 10.1039/C9SM00962K.

(95) Morsali, A.; Goharshadi, E. K.; Ali Mansoori, G.; Abbaspour, M. *Chemical Physics* **2005**, *310*, 11–15, DOI: 10.1016/j.chemphys.2004.09.027.

(96) Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; Mackerell, A. D. *Journal of Computational Chemistry* **2010**, *31*, 671–690, DOI: 10.1002/jcc.21367.

(97) Robustelli, P.; Ibanez-de-Opakua, A.; Campbell-Bezat, C.; Giordanetto, F.; Becker, S.; Zweckstetter, M.; Pan, A. C.; Shaw, D. E. *Journal of the American Chemical Society* **2022**, *144*, 2501–2510, DOI: 10.1021/jacs.1c07591.

(98) Robustelli, P.; Piana, S.; Shaw, D. E. *Proceedings of the National Academy of Sciences of the United States of America* **2018**, *115*, E4758–E4766, DOI: 10.1073/pnas.1800690115.

(99) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. *Journal of Computational Chemistry* **2004**, *25*, 1157–1174, DOI: 10.1002/jcc.20035.

(100) Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A. *Journal of Molecular Graphics and Modelling* **2006**, *25*, 247–260, DOI: 10.1016/j.jmgm.2005.12.005.

(101) Zhu, S. *Journal of Chemical Information and Modeling* **2019**, *59*, 4239–4247, DOI: 10.1021/acs.jcim.9b00552.

(102) Ross, G. A.; Rustenburg, A. S.; Grinaway, P. B.; Fass, J.; Chodera, J. D. *The journal of physical chemistry. B* **2018**, *122*, 5466–5486, DOI: 10.1021/acs.jpcb.7b11734.

(103) Klauda, J. B.; Wu, X.; Pastor, R. W.; Brooks, B. R. *The Journal of Physical Chemistry B* **2007**, *111*, 4393–4400, DOI: 10.1021/jp068767m.

(104) Venable, R. M.; Chen, L. E.; Pastor, R. W. *The Journal of Physical Chemistry B* **2009**, *113*, 5855–5862, DOI: 10.1021/jp900843x.

(105) Auffinger, P.; Beveridge, D. L. *Chemical Physics Letters* **1995**, *234*, 413–415, DOI: 10.1016/0009-2614(95)00065-C.

(106) Perera, L.; Essmann, U.; Berkowitz, M. L. *The Journal of Chemical Physics* **1995**, *102*, 450–456, DOI: 10.1063/1.469422.

(107) Roberts, J. E.; Schnitker, J. *The Journal of Chemical Physics* **1994**, *101*, 5024–5031, DOI: 10.1063/1.467425.

(108) Del Buono, G. S.; Figueirido, F. E.; Levy, R. M. *Chemical Physics Letters* **1996**, *263*, 521–529, DOI: 10.1016/S0009-2614(96)01234-1.

(109) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. *The Journal of Chemical Physics* **1995**, *103*, 8577–8593, DOI: 10.1063/1.470117.

(110) Hub, J. S.; de Groot, B. L.; Grubmüller, H.; Groenhof, G. *Journal of Chemical Theory and Computation* **2014**, *10*, 381–390, DOI: 10.1021/ct400626b.

(111) Darden, T.; York, D.; Pedersen, L. *The Journal of Chemical Physics* **1993**, *98*, 10089–10092, DOI: 10.1063/1.464397.

(112) Hardy, D. J.; Wu, Z.; Phillips, J. C.; Stone, J. E.; Skeel, R. D.; Schulten, K. *Journal of Chemical Theory and Computation* **2015**, *11*, 766–779, DOI: 10.1021/ct5009075.

(113) Shan, Y.; Klepeis, J. L.; Eastwood, M. P.; Dror, R. O.; Shaw, D. E. *The Journal of Chemical Physics* **2005**, *122*, 054101, DOI: 10.1063/1.1839571.

(114) Peterson, M. E.; Daniel, R. M.; Danson, M. J.; Eisenthal, R. *Biochemical Journal* **2007**, *402*, 331–337, DOI: 10.1042/BJ20061143.

(115) Song, X.; Li, Y. *Journal of Food Process Engineering* **2012**, *35*, 915–922, DOI: 10.1111/j.1745-4530.2011.00641.x.

(116) Knapp, B.; Ospina, L.; Deane, C. M. *Journal of Chemical Theory and Computation* **2018**, *14*, 6127–6138, DOI: 10.1021/acs.jctc.8b00391.

(117) Nosé, S. *Molecular Physics* **1984**, *52*, 255–268, DOI: 10.1080/00268978400101201.

(118) Hoover, W. G. *Physical Review A* **1985**, *31*, 1695–1697, DOI: 10.1103/PhysRevA.31.1695.

(119) Martyna, G. J.; Klein, M. L.; Tuckerman, M. *The Journal of Chemical Physics* **1992**, *97*, 2635–2643, DOI: 10.1063/1.463940.

(120) Bussi, G.; Donadio, D.; Parrinello, M. *Journal of Chemical Physics* **2007**, *126*, 014101–014101, DOI: 10.1063/1.2408420.

(121) Berendsen, H. J.; Postma, J. P.; Van Gunsteren, W. F.; Dinola, A.; Haak, J. R. *The Journal of Chemical Physics* **1984**, *81*, 3684–3690, DOI: 10.1063/1.448118.

(122) Martyna, G. J.; Tuckerman, M. E.; Tobias, D. J.; Klein, M. L. *Molecular Physics* **1996**, *87*, 1117–1157, DOI: 10.1080/00268979600100761.

(123) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindah, E. *SoftwareX* **2015**, *1–2*, 19–25, DOI: 10.1016/j.softx.2015.06.001.

(124) Karal, M. A. S.; Ahamed, M. K.; Ahmed, M.; Mahbub, Z. B. *RSC Advances* **2021**, *11*, 29598–29619, DOI: 10.1039/D1RA04647K.

(125) Macdonald, A. G. In *Cell Physiology Source Book (Third Edition)*, Sperelakis, N., Ed.; Academic Press: San Diego, 2001, pp 1003–1023, DOI: 10.1016/B978-012656976-6/50151-7.

(126) Allen, M; Tildesley, D, *Computer Simulation of Liquids*; Clarendon Press Oxford: 1991.

(127) Parrinello, M.; Rahman, A. *Physical Review Letters* **1980**, *45*, 1196–1199, DOI: 10.1103/PhysRevLett.45.1196.

(128) Parrinello, M.; Rahman, A. *Journal of Applied Physics* **1981**, *52*, 7182–7190, DOI: 10.1063/1.328693.

(129) Leach, A., *Molecular Modelling Principles and Applications*; Prentice Hall: 2001.

(130) Schlick, T., *Molecular Modeling and Simulation: An Interdisciplinary Guide*, 2nd ed; Interdisciplinary Applied Mathematics v. 21; Springer: New York, 2010.

(131) Shannon, C. *Proceedings of the IRE* **1949**, *37*, 10–21, DOI: 10.1109/JRPROC.1949.232969.

(132) Mazur, A. K. *Journal of Computational Physics* **1997**, *136*, 354–365, DOI: 10.1006/jcph.1997.5740.

(133) Feenstra, K. A.; Hess, B.; Berendsen, H. J. *Journal of Computational Chemistry* **1999**, *20*, 786–798, DOI: 10.1002/(SICI)1096-987X(199906)20:8<786::AID-JCC5>3.0.CO;2-B.

(134) Andersen, H. C. *Journal of Computational Physics* **1983**, *52*, 24–34, DOI: 10.1016/0021-9991(83)90014-1.

(135) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. *Journal of Computational Chemistry* **1997**, *18*, 1463–1472, DOI: 10.1002/(SICI)1096-987X(199709)18:12<1463::AID-JCC4>3.0.CO;2-H.

(136) Balusek, C.; Hwang, H.; Lau, C. H.; Lundquist, K.; Hazel, A.; Pavlova, A.; Lynch, D. L.; Reggio, P. H.; Wang, Y.; Gumbart, J. C. *Journal of chemical theory and computation* **2019**, *15*, 4673–4686, DOI: 10.1021/acs.jctc.9b00160.

(137) Streett, W.; Tildesley, D.; Saville, G. *Molecular Physics* **1978**, *35*, 639–648, DOI: 10.1080/00268977800100471.

(138) *Advances in Chemical Physics: Proteins: A Theoretical Perspective of Dynamics, Structure, and Thermodynamics*; Brooks, C. L., Karplus, M., Pettitt, B. M., Eds.; Advances in Chemical Physics; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 1988, DOI: 10.1002/9780470141205.

(139) Flood, E.; Boiteux, C.; Lev, B.; Vorobyov, I.; Allen, T. W. *Chemical Reviews* **2019**, *119*, 7737–7832, DOI: 10.1021/acs.chemrev.8b00630.

(140) Werner, T.; Morris, M. B.; Dastmalchi, S.; Church, W. B. *Advanced Drug Delivery Reviews* **2012**, *64*, 323–343, DOI: 10.1016/j.addr.2011.11.011.

(141) Hénin, J.; Lelièvre, T.; Shirts, M. R.; Valsson, O.; Delemotte, L. Enhanced Sampling Methods for Molecular Dynamics Simulations, 2022.

(142) Meshkin, H.; Zhu, F. *Journal of Chemical Theory and Computation* **2017**, *13*, 2086–2097, DOI: 10.1021/acs.jctc.6b01171.

(143) Domański, J.; Hedger, G.; Best, R. B.; Stansfeld, P. J.; Sansom, M. S. P. *The Journal of Physical Chemistry B* **2017**, *121*, 3364–3375, DOI: 10.1021/acs.jpcb.6b08445.

(144) Zhu, F.; Hummer, G. *Journal of Chemical Theory and Computation* **2012**, *8*, 3759–3768, DOI: 10.1021/ct2009279.

(145) Subramanian, N.; Schumann-Gillett, A.; Mark, A. E.; O'Mara, M. L. *Journal of Chemical Information and Modeling* **2019**, *59*, 2287–2298, DOI: 10.1021/acs.jcim.8b00624.

(146) You, W.; Tang, Z.; Chang, C.-e. A. *Journal of Chemical Theory and Computation* **2019**, *15*, 2433–2443, DOI: 10.1021/acs.jctc.8b01142.

(147) Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A. *Journal of Computational Chemistry* **1992**, *13*, 1011–1021, DOI: 10.1002/jcc.540130812.

(148) Kästner, J.; Thiel, W. *The Journal of Chemical Physics* **2005**, *123*, 144104, DOI: 10.1063/1.2052648.

(149) Kim, I.; Allen, T. W. *The Journal of Chemical Physics* **2012**, *136*, 164103, DOI: 10.1063/1.3701766.

(150) Kästner, J. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2011**, *1*, 932–942, DOI: 10.1002/wcms.66.

(151) Chen, P. C.; Kuyucak, S. *Biophysical Journal* **2011**, *100*, 2466–2474, DOI: 10.1016/j.bpj.2011.03.052.

(152) Bussi, G.; Laio, A.; Tiwary, P. In *Handbook of Materials Modeling*, Andreoni, W., Yip, S., Eds.; Springer International Publishing: Cham, 2020, pp 565–595, DOI: 10.1007/978-3-319-44677-6_49.

(153) Laio, A.; Parrinello, M. *Proceedings of the National Academy of Sciences of the United States of America* **2002**, *99*, 12562–12566, DOI: 10.1073/pnas.202427399.

(154) Cheng, T.; Goddard, W. A.; An, Q.; Xiao, H.; Merinov, B.; Morozov, S. *Physical Chemistry Chemical Physics* **2017**, *19*, 2666–2673, DOI: 10.1039/C6CP08055C.

(155) Giberti, F.; Salvalaglio, M.; Parrinello, M. *IUCrJ* **2015**, *2*, 256–266, DOI: 10.1107/S2052252514027626.

(156) Bussi, G.; Laio, A. *Nature Reviews Physics* **2020**, *2*, 200–212, DOI: 10.1038/s42254-020-0153-0.

(157) Sun, R.; Dama, J. F.; Tan, J. S.; Rose, J. P.; Voth, G. A. *Journal of Chemical Theory and Computation* **2016**, *12*, 5157–5169, DOI: 10.1021/acs.jctc.6b00206.

(158) Barducci, A.; Bussi, G.; Parrinello, M. *Physical Review Letters* **2008**, *100*, 020603, DOI: 10.1103/PhysRevLett.100.020603.

(159) Tiwary, P.; Parrinello, M. *Physical Review Letters* **2013**, *111*, 230602, DOI: 10.1103/PhysRevLett.111.230602.

(160) Tiwary, P.; Berne, B. J. *The Journal of Chemical Physics* **2016**, *144*, 134103, DOI: 10.1063/1.4944577.

(161) Salvalaglio, M.; Tiwary, P.; Parrinello, M. *Journal of Chemical Theory and Computation* **2014**, *10*, 1420–1425, DOI: 10.1021/ct500040r.

(162) Raiteri, P.; Laio, A.; Gervasio, F. L.; Micheletti, C.; Parrinello, M. *The Journal of Physical Chemistry B* **2006**, *110*, 3533–3539, DOI: 10.1021/jp054359r.

(163) 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.

(164) Mamatkulov, S.; Fyta, M.; Netz, R. R. *The Journal of Chemical Physics* **2013**, *138*, 024505, DOI: 10.1063/1.4772808.

(165) Bergonzo, C.; Hall, K. B.; Cheatham, T. E. *Journal of Chemical Theory and Computation* **2016**, *12*, 3382–3389, DOI: 10.1021/acs.jctc.6b00173.

(166) Hollingsworth, S. A.; Dror, R. O. *Neuron* **2018**, *99*, 1129–1143, DOI: 10.1016/j.neuron.2018.08.011.

(167) Melo, M. C. R.; Bernardi, R. C.; Rudack, T.; Scheurer, M.; Riplinger, C.; Phillips, J. C.; Maia, J. D. C.; Rocha, G. B.; Ribeiro, J. V.; Stone, J. E.; Neese, F.; Schulten, K.; Luthey-Schulten, Z. *Nature methods* **2018**, *15*, 351–354, DOI: 10.1038/nmeth.4638.

(168) Nerenberg, P. S.; Head-Gordon, T. *Current Opinion in Structural Biology* **2018**, *49*, 129–138, DOI: 10.1016/j.sbi.2018.02.002.

(169) Lin, F.-Y.; Huang, J.; Pandey, P.; Rupakheti, C.; Li, J.; Roux, B.; MacKerell, A. D. *Journal of chemical theory and computation* **2020**, *16*, 3221–3239, DOI: 10.1021/acs.jctc.0c00057.

(170) Shi, Y.; Xia, Z.; Zhang, J.; Best, R.; Wu, C.; Ponder, J. W.; Ren, P. *Journal of Chemical Theory and Computation* **2013**, *9*, 4046–4063, DOI: 10.1021/ct4003702.

(171) Li, H.; Chowdhary, J.; Huang, L.; He, X.; MacKerell, A. D.; Roux, B. *Journal of Chemical Theory and Computation* **2017**, *13*, 4535–4552, DOI: 10.1021/acs.jctc.7b00262.

(172) Lindorff-Larsen, K.; Maragakis, P.; Piana, S.; Shaw, D. E. *The Journal of Physical Chemistry B* **2016**, *120*, 8313–8320, DOI: 10.1021/acs.jpcb.6b02024.

(173) TICA Theory, http://docs.markovmodel.org/lecture_tica.html.

(174) Schultze, S.; Grubmüller, H. *Journal of Chemical Theory and Computation* **2021**, *17*, 5766–5776, DOI: 10.1021/acs.jctc.1c00273.

(175) Brotzakis, Z. F.; Limongelli, V.; Parrinello, M. *Journal of Chemical Theory and Computation* **2019**, *15*, 743–750, DOI: 10.1021/acs.jctc.8b00934.

(176) Ribeiro, J. M. L.; Bravo, P.; Wang, Y.; Tiwary, P. *The Journal of Chemical Physics* **2018**, *149*, 072301, DOI: 10.1063/1.5025487.

(177) Van Rossum, G.; Drake Jr, F. L., *Python Reference Manual*; Centrum voor Wiskunde en Informatica Amsterdam: 1995.

(178) Humphrey, W.; Dalke, A.; Schulten, K. *Journal of Molecular Graphics* **1996**, *14*, 33–38, DOI: 10.1016/0263-7855(96)00018-5.

(179) Mallajosyula, S. S.; Jo, S.; Im, W.; Mackerell, A. D. *Methods in Molecular Biology* **2015**, *1273*, 407–429, DOI: 10.1007/978-1-4939-2343-4_25.

(180) Michaud-Agrawal, N.; Denning, E. J.; Woolf, T. B.; Beckstein, O. *Journal of Computational Chemistry* **2011**, *32*, 2319–2327, DOI: 10.1002/jcc.21787.

(181) Gowers, R. J.; Linke, M.; Barnoud, J.; Reddy, T. J. E.; Melo, M. N.; Seyler, S. L.; Domański, J.; Dotson, D. L.; Buchoux, S.; Kenney, I. M.; Beckstein, O. In *Proceedings of the 15th Python in Science Conference*, ed. by Benthall, S.; Rostrup, S., 2016, pp 98–105.

(182) Sali, A.; Blundell, T. L. *Journal of Molecular Biology* **1993**, *234*, 779–815, DOI: 10.1006/jmbi.1993.1626.

(183) Shen, M.-y.; Sali, A. *Protein Science* **2006**, *15*, 2507–2524, DOI: 10.1110/ps.062416606.

(184) Webb, B.; Sali, A. *Current Protocols in Bioinformatics* **2016**, *54*, DOI: 10.1002/cpbi.3.

(185) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kalé, L.; Schulten, K. *Journal of Computational Chemistry* **2005**, *26*, 1781–1802, DOI: 10.1002/jcc.20289.

(186) Eastman, P.; Swails, J.; Chodera, J. D.; McGibbon, R. T.; Zhao, Y.; Beauchamp, K. A.; Wang, L. P.; Simmonett, A. C.; Harrigan, M. P.; Stern, C. D.; Wiewiora, R. P.; Brooks, B. R.; Pande, V. S. *PLoS Computational Biology* **2017**, *13*, ed. by Gentleman, R., e1005659–e1005659, DOI: 10.1371/journal.pcbi.1005659.

(187) Tribello, G. A.; Bonomi, M.; Branduardi, D.; Camilloni, C.; Bussi, G. *Computer Physics Communications* **2014**, *185*, 604–613, DOI: 10.1016/j.cpc.2013.09.018.

(188) The UniProt Consortium et al. *Nucleic Acids Research* **2021**, *49*, D480–D489, DOI: 10.1093/nar/gkaa1100.

(189) Amber22, 2022.

(190) Ponder, J. W.; Case, D. A. In *Advances in Protein Chemistry*; Protein Simulations, Vol. 66; Academic Press: 2003, pp 27–85, DOI: 10.1016/S0065-3233(03)66002-X.

(191) Shirts, M. R.; Klein, C.; Swails, J. M.; Yin, J.; Gilson, M. K.; Mobley, D. L.; Case, D. A.; Zhong, E. D. *Journal of Computer-Aided Molecular Design* **2017**, *31*, 147–161, DOI: 10.1007/s10822-016-9977-1.

(192) Dick, K. Sick: The Life &amp; Death of Bob Flanagan, Supermasochist, Documentary, 1997.

(193) Guo, J.; Garratt, A.; Hill, A. *Journal of Cystic Fibrosis* **2022**, *21*, 456–462, DOI: 10.1016/j.jcf.2022.01.009.

(194) McBennett, K. A.; Davis, P. B.; Konstan, M. W. *Pediatric pulmonology* **2022**, *57*, S5–S12, DOI: 10.1002/ppul.25733.

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

(196) Kayani, K.; Mohammed, R.; Mohiaddin, H. Cystic Fibrosis-Related Diabetes, 2018, DOI: 10.3389/fendo.2018.00020.

(197) E Garcia, L. d. C.; Petry, L. M.; Germani, P. A. V. D. S.; Xavier, L. F.; de Barros, P. B.; Meneses, A. d. S.; Prestes, L. M.; Bittencourt, L. B.; Pieta, M. P.; Friedrich, F.; Pinto, L. A. *Frontiers in Pediatrics* **2022**, *10*, 881470, DOI: 10.3389/fped.2022.881470.

(198) Cormet-Boyaka, E.; Di, A.; Chang, S. Y.; Naren, A. P.; Tousson, A.; Nelson, D. J.; Kirk, K. L. *Proceedings of the National Academy of Sciences of the United States of America* **2002**, *99*, 12477–12482, DOI: 10.1073/pnas.192203899.

(199) Naren, A. P.; Quick, M. W.; Collawn, J. F.; Nelson, D. J.; Kirk, K. L. *Proceedings of the National Academy of Sciences of the United States of America* **1998**, *95*, 10972–10977, DOI: 10.1073/pnas.95.18.10972.

(200) Thelin, W. R.; Chen, Y.; Gentzsch, M.; Kreda, S. M.; Sallee, J. L.; Scarlett, C. O.; Borchers, C. H.; Jacobson, K.; Stutts, M. J.; Milgram, S. L. *Journal of Clinical Investigation* **2007**, *117*, 364–374, DOI: 10.1172/JCI30376.

(201) Kim, Y.; Jun, I.; Shin, D. H.; Yoon, J. G.; Piao, H.; Jung, J.; Park, H. W.; Cheng, M. H.; Bahar, I.; Whitcomb, D. C.; Lee, M. G. *Cellular and Molecular Gastroenterology and Hepatology* **2019**, *9*, 79–103, DOI: 10.1016/j.jcmgh.2019.09.003.

(202) Linsdell, P. *Experimental Physiology* **2006**, *91*, 123–129, DOI: 10.1113/expphysiol.2005.031757.

(203) Linsdell, P.; Irving, C. L.; Cowley, E. A. *Journal of Biological Chemistry* **2022**, *298*, DOI: 10.1016/j.jbc.2022.101659.

(204) The Hospital for Sick Children *The Clinical and Functional TRanslation of CFTR (CFTR2)*; tech. rep.; 2020.

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

(206) Ostedgaard, L. S.; Baldursson, O.; Vermeer, D. W.; Welsh, M. J.; Robertson, A. D. *Proceedings of the National Academy of Sciences* **2000**, *97*, 5657–5662, DOI: 10.1073/pnas.100588797.

(207) Hegedűs, T.; Geisler, M.; Lukács, G. L.; Farkas, B. *Cellular and molecular life sciences: CMLS* **2022**, *79*, 73, DOI: 10.1007/s00018-021-04112-1.

(208) Liu, F.; Zhang, Z.; Levit, A.; Levring, J.; Touhara, K. K.; Shoichet, B. K.; Chen, J. *Science (New York, N.Y.)* **2019**, *364*, 1184–1188, DOI: 10.1126/science.aaw7611.

(209) Ivey, G.; Youker, R. T. *PLoS ONE* **2020**, *15*, e0227668–e0227668, DOI: 10.1371/journal.pone.0227668.

(210) Zolnerciks, J. K.; Akkaya, B. G.; Snippe, M.; Chiba, P.; Seelig, A.; Linton, K. J. *FASEB journal: official publication of the Federation of American Societies for Experimental Biology* **2014**, *28*, 4335–4346, DOI: 10.1096/fj.13-245639.

(211) Dong, Q.; Ernst, S. E.; Ostedgaard, L. S.; Shah, V. S.; Ver Heul, A. R.; Welsh, M. J.; Randak, C. O. *The Journal of Biological Chemistry* **2015**, *290*, 14140–14153, DOI: 10.1074/jbc.M114.611616.

(212) Moyer, B. D.; Demon, J.; Karlson, K. H.; Reynolds, D.; Wang, S.; Mickle, J. E.; Milewski, M.; Cutting, G. R.; Guggino, W. B.; Li, M.; Stanton, B. A. *Journal of Clinical Investigation* **1999**, *104*, 1353–1361, DOI: 10.1172/JCI7453.

(213) Cushing, P. R.; Fellows, A.; Villone, D.; Boisguérin, P.; Madden, D. R. *Biochemistry* **2008**, *47*, 10084–10098, DOI: 10.1021/bi8003928.

(214) Thomas, C. et al. *FEBS Letters* **2020**, *594*, 3767–3775, DOI: 10.1002/1873-3468.13935.

(215) Linsdell, P. *Channels (Austin, Tex.)* **2018**, *12*, 284–290, DOI: 10.1080/19336950.2018.1502585.

(216) Zhang, X. C.; Yang, H.; Liu, Z.; Sun, F. *Biophysics Reports* **2018**, *4*, 300–319, DOI: 10.1007/s41048-018-0074-y.

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

(218) 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.

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

(220) Hoffmann, B.; Elbahnsi, A.; Lehn, P.; Décout, J. L.; Pietrucci, F.; Mornon, J. P.; Callebaut, I. *Cellular and Molecular Life Sciences* **2018**, *75*, 3829–3855, DOI: 10.1007/s00018-018-2835-7.

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

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

(223) Fiedorczuk, K.; Chen, J. *Cell* **2022**, *185*, 158–168.e11, DOI: 10.1016/j.cell.2021.12.009.

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

(225) de Poel, E.; Lefferts, J. W.; Beekman, J. M. *Journal of Cystic Fibrosis: Official Journal of the European Cystic Fibrosis Society* **2020**, *19 Suppl 1*, S60–S64, DOI: 10.1016/j.jcf.2019.11.002.

(226) Clancy, J. P. et al. *Journal of Cystic Fibrosis: Official Journal of the European Cystic Fibrosis Society* **2019**, *18*, 22–34, DOI: 10.1016/j.jcf.2018.05.004.

(227) Wong, S. L.; Awatade, N. T.; Astore, M. A.; Allan, K. M.; Carnell, M. J.; Slapetova, I.; Chen, P.-c.; Capraro, A.; Fawcett, L. K.; Whan, R. M.; Griffith, R.; Ooi, C. Y.; Kuyucak, S.; Jaffe, A.; Waters, S. A. *iScience* **2022**, *25*, 103710, DOI: 10.1016/j.isci.2021.103710.

(228) Wong, S. L. et al. *American journal of respiratory cell and molecular biology* **2022**, DOI: 10.1165/rcmb.2021-0337OC.

(229) Ciciriello, F.; Bijvelds, M. J. C.; Alghisi, F.; Meijsen, K. F.; Cristiani, L.; Sorio, C.; Melotti, P.; Fiocchi, A. G.; Lucidi, V.; De Jonge, H. R. *Journal of personalized medicine* **2022**, *12*, DOI: 10.3390/jpm12040632.

(230) Dekkers, J. F. et al. *Nature Medicine* **2013**, *19*, 939–945, DOI: 10.1038/nm.3201.

(231) Mitchison, T. J.; Mitchison, H. M. *Nature* **2010**, *463*, 308–309, DOI: 10.1038/463308a.

(232) Bustamante-Marin, X. M.; Ostrowski, L. E. *Cold Spring Harbor Perspectives in Biology* **2017**, *9*, a028241, DOI: 10.1101/cshperspect.a028241.

(233) Ratjen, F.; Bell, S. C.; Rowe, S. M.; Goss, C. H.; Quittner, A. L.; Bush, A. *Nature reviews. Disease primers* **2015**, *1*, 15010, DOI: 10.1038/NRDP.2015.10.

(234) Hwang, T. C.; Kirk, K. L. *Cold Spring Harbor Perspectives in Medicine* **2013**, *3*, a009498–a009498, DOI: 10.1101/cshperspect.a009498.

(235) Gadsby, D. C.; Nairn, A. C. *Trends in Biochemical Sciences* **1994**, *19*, 513–518, DOI: 10.1016/0968-0004(94)90141-4.

(236) Liu, F.; Zhang, Z.; Csanády, L.; Gadsby, D. C.; Chen, J. *Cell* **2017**, *169*, 85–95.e8, DOI: 10.1016/j.cell.2017.02.024.

(237) Fu, J.; Ji, H. L.; Naren, A. P.; Kirk, K. L. *Journal of Physiology* **2001**, *536*, 459–470, DOI: 10.1111/J.1469-7793.2001.0459C.XD.

(238) Gené, G. G.; Llobet, A.; Larriba, S.; de Semir, D.; Martínez, I.; Escalada, A.; Solsona, C.; Casals, T.; Aran, J. M. J. *Human Mutation* **2008**, *29*, 738–749, DOI: 10.1002/humu.20721.

(239) Jurkuvenaite, A.; Varga, K.; Nowotarski, K.; Kirk, K. L.; Sorscher, E. J.; Li, Y.; Clancy, J. P.; Bebok, Z.; Collawn, J. F. *Journal of Biological Chemistry* **2006**, *281*, 3329–3334, DOI: 10.1074/jbc.M508131200.

(240) Sabusap, C. M.; Joshi, D.; Simhaev, L.; Oliver, K. E.; Senderowitz, H.; van Willigen, M.; Braakman, I.; Rab, A.; Sorscher, E. J.; Hong, J. S. *The Journal of Biological Chemistry* **2021 Jan-Jun**, *296*, 100598, DOI: 10.1016/J.JBC.2021.100598.

(241) Loo, T. W.; Bartlett, M. C.; Clarke, D. M. *Biochemical Pharmacology* **2013**, *86*, 612–619, DOI: 10.1016/J.BCP.2013.06.028.

(242) Ren, H. Y.; Grove, D. E.; De La Rosa, O.; Houck, S. A.; Sopha, P.; Van Goor, F.; Hoffman, B. J.; Cyr, D. M. *Molecular Biology of the Cell* **2013**, *24*, 3016–3024, DOI: 10.1091/MBC.E13-05-0240.

(243) Loo, T. W.; Clarke, D. M. *Biochemical Pharmacology* **2017**, *136*, 24–31, DOI: 10.1016/J.BCP.2017.03.020.

(244) Hudson, R. P.; Dawson, J. E.; Chong, P. A.; Yang, Z.; Millen, L.; Thomas, P. J.; Brouillette, C. G.; Forman-Kay, J. D. *Molecular Pharmacology* **2017**, *92*, 124–135, DOI: 10.1124/MOL.117.108373.

(245) Okiyoneda, T.; Veit, G.; Dekkers, J. F.; Bagdany, M.; Soya, N.; Xu, H.; Roldan, A.; Verkman, A. S.; Kurth, M.; Simon, A.; Hegedus, T.; Beekman, J. M.; Lukacs, G. L. *Nature Chemical Biology* **2013**, *9*, 444–454, DOI: 10.1038/NCHEMBIO.1253.

(246) Berkers, G. et al. *Cell Reports* **2019**, *26*, 1701–1708.e3, DOI: 10.1016/J.CELREP.2019.01.068.

(247) McCarthy, C.; Brewington, J. J.; Harkness, B.; Clancy, J. P.; Trapnell, B. C. *The European Respiratory Journal* **2018**, *51*, 1702457, DOI: 10.1183/13993003.02457-2017.

(248) Ramalho, A. S.; Förstová, E.; Vonk, A. M.; Ferrante, M.; Verfailli, C.; Dupont, L.; Boon, M.; Proesmans, M.; Beekma, J. M.; Sarouk, I.; Cordero, C. V.; Vermeule, F.; De Boeck, K. *The European Respiratory Journal* **2021**, *57*, 1902426, DOI: 10.1183/13993003.02426-2019.

(249) Sato, T.; Vries, R. G.; Snippert, H. J.; Van De Wetering, M.; Barker, N.; Stange, D. E.; Van Es, J. H.; Abo, A.; Kujala, P.; Peters, P. J.; Clevers, H. *Nature* **2009**, *459*, 262–265, DOI: 10.1038/NATURE07935.

(250) Awatade, N. T.; Wong, S. L.; Hewson, C. K.; Fawcett, L. K.; Kicic, A.; Jaffe, A.; Waters, S. A. *Frontiers in Pharmacology* **2018**, *9*, DOI: 10.3389/FPHAR.2018.01429.

(251) Pollard, B. S.; Pollard, H. B. *Pediatric Pulmonology* **2018**, *53*, S12–S29, DOI: 10.1002/PPUL.24118.

(252) Dey, I.; Shah, K.; Bradbury, N. A. *Journal of Genetic Syndromes & Gene Therapy* **2016**, *07*, DOI: 10.4172/2157-7412.1000284.

(253) Clancy, J. P. et al. *PLoS ONE* **2013**, *8*, DOI: 10.1371/JOURNAL.PONE.0073905.

(254) Graeber, S. Y.; Hug, M. J.; Sommerburg, O.; Hirtz, S.; Hentschel, J.; Heinzmann, A.; Dopfer, C.; Schulz, A.; Mainz, J. G.; Tümmler, B.; Mall, M. A. *American Journal of Respiratory and Critical Care Medicine* **2015**, *192*, 1252–1255, DOI: 10.1164/RCCM.201507-1271LE.

(255) Veeze, H. J.; Halley, D. J.; Bijman, J.; De Jongste, J. C.; De Jonge, H. R.; Sinaasappel, M. *Journal of Clinical Investigation* **1994**, *93*, 461–466, DOI: 10.1172/JCI116993.

(256) Dekkers, J. F.; Van Mourik, P.; Vonk, A. M.; Kruisselbrink, E.; Berkers, G.; de Winter-de Groot, K. M.; Janssens, H. M.; Bronsveld, I.; van der Ent, C. K.; de Jonge, H. R.; Beekman, J. M. *Journal of Cystic Fibrosis* **2016**, *15*, 568–578, DOI: 10.1016/j.jcf.2016.04.007.

(257) Cuyx, S.; Ramalho, A. S.; Corthout, N.; Fieuws, S.; Fürstová, E.; Arnauts, K.; Ferrante, M.; Verfaillie, C.; Munck, S.; Boon, M.; Proesmans, M.; Dupont, L.; De Boeck, K.; Vermeulen, F. *Thorax* **2021**, *76*, 1146–1149, DOI: 10.1136/THORAXJNL-2020-216368.

(258) Van Mourik, P.; Beekman, J. M.; Van Der Ent, C. K. *European Respiratory Journal* **2019**, *54*, DOI: 10.1183/13993003.02379-2018.

(259) Zomer-van Ommen, D. D.; de Poel, E.; Kruisselbrink, E.; Oppelaar, H.; Vonk, A. M.; Janssens, H. M.; van der Ent, C. K.; Hagemeijer, M. C.; Beekman, J. M. *Journal of Cystic Fibrosis* **2018**, *17*, 316–324, DOI: 10.1016/J.JCF.2018.02.007.

(260) Laselva, O.; Bartlett, C.; Gunawardena, T. N. A.; Ouyang, H.; Eckford, P. D. W.; Moraes, T. J.; Bear, C. E.; Gonska, T. *The European Respiratory Journal* **2021**, *57*, 2002774, DOI: 10.1183/13993003.02774-2020.

(261) Shaughnessy, C. A.; Zeitlin, P. L.; Bratcher, P. E. *Scientific Reports* **2021**, *11*, DOI: 10.1038/S41598-021-99184-1.

(262) Veit, G.; Velkov, T.; Xu, H.; Vadeboncoeur, N.; Bilodeau, L.; Matouk, E.; Lukacs, G. L. *Journal of Personalized Medicine* **2021**, *11*, DOI: 10.3390/JPM11070643.

(263) Beckerman, M. **2015**, 61–94, DOI: 10.1007/978-3-319-22117-5_3.

(264) Van Goor, F.; Hadida, S.; Grootenhuis, P. D.; Burton, B.; Stack, J. H.; Straley, K. S.; Decker, C. J.; Miller, M.; McCartney, J.; Olson, E. R.; Wine, J. J.; Frizzell, R. A.; Ashlock, M.; Negulescu, P. A. *Proceedings of the National Academy of Sciences of the United States of America* **2011**, *108*, 18843–18848, DOI: 10.1073/PNAS.1105787108.

(265) Yeh, H. I.; Qiu, L.; Sohma, Y.; Conrath, K.; Zou, X.; Hwang, T. C. *Journal of General Physiology* **2019**, *151*, 912–928, DOI: 10.1085/jgp.201912360.

(266) Van Goor, F. et al. *Proceedings of the National Academy of Sciences of the United States of America* **2009**, *106*, 18825–18830, DOI: 10.1073/PNAS.0904709106.

(267) Yeh, H. I.; Sohma, Y.; Conrath, K.; Hwang, T. C. *Journal of General Physiology* **2017**, *149*, 1105–1118, DOI: 10.1085/JGP.201711886.

(268) Gees, M. et al. *Frontiers in Pharmacology* **2018**, *9*, DOI: 10.3389/FPHAR.2018.01221.

(269) Van Der Plas, S. E. et al. *Journal of Medicinal Chemistry* **2018**, *61*, 1425–1435, DOI: 10.1021/ACS.JMEDCHEM.7B01288.

(270) Dekkers, J. F. et al. *Science Translational Medicine* **2016**, *8*, DOI: 10.1126/SCITRANSLMED.AAD8278.

(271) Phuan, P. W.; Son, J. H.; Tan, J. A.; Li, C.; Musante, I.; Zlock, L.; Nielson, D. W.; Finkbeiner, W. E.; Kurth, M. J.; Galietta, L. J.; Haggie, P. M.; Verkman, A. S. *Journal of Cystic Fibrosis* **2018**, *17*, 595–606, DOI: 10.1016/J.JCF.2018.05.010.

(272) Phuan, P. W.; Tan, J. A.; Rivera, A. A.; Zlock, L.; Nielson, D. W.; Finkbeiner, W. E.; Haggie, P. M.; Verkman, A. S. *Scientific Reports* **2019**, *9*, DOI: 10.1038/S41598-019-54158-2.

(273) Veit, G.; Da Fonte, D. F.; Avramescu, R. G.; Premchandar, A.; Bagdany, M.; Xu, H.; Bensinger, D.; Stubba, D.; Schmidt, B.; Matouk, E.; Lukacs, G. L. *Journal of Cystic Fibrosis: Official Journal of the European Cystic Fibrosis Society* **2020**, *19*, 236–244, DOI: 10.1016/j.jcf.2019.10.011.

(274) Nussinov, R.; Tsai, C. J. *Cell* **2013**, *153*, 293–305, DOI: 10.1016/J.CELL.2013.03.034.

(275) Berkers, G. et al. *Journal of Cystic Fibrosis* **2020**, *19*, 955–961, DOI: 10.1016/J.JCF.2020.04.014.

(276) McKone, E. F.; Borowitz, D.; Drevinek, P.; Griese, M.; Konstan, M. W.; Wainwright, C.; Ratjen, F.; Sermet-Gaudelus, I.; Plant, B.; Munck, A.; Jiang, Y.; Gilmartin, G.; Davies, J. C. *The Lancet Respiratory Medicine* **2014**, *2*, 902–910, DOI: 10.1016/S2213-2600(14)70218-8.

(277) Volkova, N.; Moy, K.; Evans, J.; Campbell, D.; Tian, S.; Simard, C.; Higgins, M.; Konstan, M. W.; Sawicki, G. S.; Elbert, A.; Charman, S. C.; Marshall, B. C.; Bilton, D. *Journal of Cystic Fibrosis* **2020**, *19*, 68–79, DOI: 10.1016/J.JCF.2019.05.015.

(278) Sohma, Y.; Yu, Y.-C.; Hwang, T.-C. *Current Pharmaceutical Design* **2013**, *19*, 3521–3528, DOI: 10.2174/13816128113199990320.

(279) Keating, D. et al. *New England Journal of Medicine* **2018**, *379*, 1612–1620, DOI: 10.1056/NEJMOA1807120.

(280) Veit, G.; Roldan, A.; Hancock, M. A.; da Fonte, D. F.; Xu, H.; Hussein, M.; Frenkiel, S.; Matouk, E.; Velkov, T.; Lukacs, G. L. *JCI Insight* **2020**, *5*, DOI: 10.1172/JCI.INSIGHT.139983.

(281) Veit, G.; Vaccarin, C.; Lukacs, G. L. *Journal of Cystic Fibrosis* **2021**, *20*, 895–898, DOI: 10.1016/J.JCF.2021.03.011.

(282) Baatallah, N.; Elbahnsi, A.; Mornon, J. P.; Chevalier, B.; Pranke, I.; Servel, N.; Zelli, R.; Décout, J. L.; Edelman, A.; Sermet-Gaudelus, I.; Callebaut, I.; Hinzpeter, A. *Cellular and Molecular Life Sciences* **2021**, *78*, 7813–7829, DOI: 10.1007/S00018-021-03994-5.

(283) Werlin, S. et al. *Journal of pediatric gastroenterology and nutrition* **2015**, *60*, 675–679, DOI: 10.1097/MPG.0000000000000623.

(284) Kleizen, B.; van Willigen, M.; Mijnders, M.; Peters, F.; Grudniewska, M.; Hillenaar, T.; Thomas, A.; Kooijman, L.; Peters, K. W.; Frizzell, R.; van der Sluijs, P.; Braakman, I. *Journal of Molecular Biology* **2021**, *433*, DOI: 10.1016/J.JMB.2021.166955.

(285) Clevers, H. *Cell* **2016**, *165*, 1586–1597, DOI: 10.1016/J.CELL.2016.05.082.

(286) De Jonge, H. R.; Ballmann, M.; Veeze, H.; Bronsveld, I.; Stanke, F.; Tümmler, B.; Sinaasappel, M. *Journal of Cystic Fibrosis* **2004**, *3*, 159–163, DOI: 10.1016/J.JCF.2004.05.034.

(287) Derichs, N.; Sanz, J.; Von Kanel, T.; Stolpe, C.; Zapf, A.; Tümmler, B.; Gallati, S.; Ballmann, M. *Thorax* **2010**, *65*, 594–599, DOI: 10.1136/THX.2009.125088.

(288) Li, H.; Sheppard, D. N.; Hug, M. J. *Journal of Cystic Fibrosis* **2004**, *3*, 123–126, DOI: 10.1016/J.JCF.2004.05.026.

(289) Pankow, S.; Bamberger, C.; Calzolari, D.; Martínez-Bartolomé, S.; Lavallée-Adam, M.; Balch, W. E.; Yates, J. R. *Nature* **2015**, *528*, 510–516, DOI: 10.1038/NATURE15729.

(290) Mark, P.; Nilsson, L. *Journal of Physical Chemistry A* **2001**, *105*, 9954–9960, DOI: 10.1021/jp003020w.

(291) Bozoky, Z.; Krzeminski, M.; Muhandiram, R.; Birtley, J. R.; Al-Zahrani, A.; Thomas, P. J.; Frizzell, R. A.; Ford, R. C.; Forman-Kay, J. D. *Proceedings of the National Academy of Sciences of the United States of America* **2013**, *110*, E4427–E4436, DOI: 10.1073/pnas.1315104110.

(292) Nosé, S.; Klein, M. L. *Molecular Physics* **1983**, *50*, 1055–1076, DOI: 10.1080/00268978300102851.

(293) Klauda, J. B.; Venable, R. M.; Freites, J. A.; O'Connor, J. W.; Tobias, D. J.; Mondragon-Ramirez, C.; Vorobyov, I.; MacKerell, A. D.; Pastor, R. W. *Journal of Physical Chemistry B* **2010**, *114*, 7830–7843, DOI: 10.1021/JP101759Q.

(294) Pandit, S. A.; Scott, H. L. *Soft Matter* **2009**, *4*, 1–82, DOI: 10.1002/9783527623372.CH1.

(295) Buchoux, S. *Bioinformatics (Oxford, England)* **2017**, *33*, 133–134, DOI: 10.1093/BIOINFORMATICS/BTW563.

(296) Ge, N.; Muise, C. N.; Gong, X.; Linsdell, P. *Journal of Biological Chemistry* **2004**, *279*, 55283–55289, DOI: 10.1074/jbc.M411935200.

(297) Gong, X.; Linsdell, P. *Archives of Biochemistry and Biophysics* **2004**, *426*, 78–82, DOI: 10.1016/j.abb.2004.03.033.

(298) Linsdell, P. *Biochimica et Biophysica Acta (BBA) - Biomembranes* **2021**, *1863*, 183558, DOI: 10.1016/j.bbamem.2021.183558.

(299) Bergh, C.; Heusser, S. A.; Howard, R.; Lindahl, E. *eLife* **2021**, *10*, ed. by Allen, T. W.; Faraldo-Gómez, J. D.; Poitevin, F., e68369, DOI: 10.7554/eLife.68369.

(300) Cui, G.; Cottrill, K. A.; Strickland, K. M.; Imhoff, B. R.; McCarty, N. A. *Biophysical Journal* **2020**, *118*, 588a, DOI: 10.1016/j.bpj.2019.11.3187.

(301) Cottrill, K. A.; Farinha, C. M.; McCarty, N. A. *Communications Biology* **2020**, *3*, 179–179, DOI: 10.1038/s42003-020-0909-1.

(302) Simon, M. A.; Csanády, L. *eLife* **2021**, *10*, ed. by Jara-Oseguera, A.; Aldrich, R. W.; Jara-Oseguera, A.; Hwang, T.-C., e74693, DOI: 10.7554/eLife.74693.

(303) Wang, G.; Linsley, R.; Norimatsu, Y. *The FEBS Journal* **2016**, *283*, 2458–2475, DOI: 10.1111/febs.13752.

(304) Kogan, I.; Ramjeesingh, M.; Li, C.; Kidd, J. F.; Wang, Y.; Leslie, E. M.; Cole, S. P.; Bear, C. E. *The EMBO Journal* **2003**, *22*, 1981–1989, DOI: 10.1093/emboj/cdg194.

(305) Corradi, V.; Gu, R. X.; Vergani, P.; Tieleman, D. P. *Biophysical Journal* **2018**, *114*, 1751–1754, DOI: 10.1016/j.bpj.2018.03.003.

(306) Gao, X.; Hwang, T. C. *Proceedings of the National Academy of Sciences of the United States of America* **2015**, *112*, 2461–2466, DOI: 10.1073/pnas.1420676112.

(307) Vergani, P.; Lockless, S. W.; Nairn, A. C.; Gadsby, D. C. *Nature* **2005**, *433*, 876–880, DOI: 10.1038/nature03313.

(308) Yeh, H.-I.; Yu, Y.-C.; Kuo, P.-L.; Tsai, C.-K.; Huang, H.-T.; Hwang, T.-C. *The Journal of Physiology* **2021**, *599*, 4625–4642, DOI: 10.1113/JP281933.

(309) Gao, Y.; Cao, E.; Julius, D.; Cheng, Y. *Nature* **2016**, *534*, 347–351, DOI: 10.1038/nature17964.

(310) Kampjut, D.; Steiner, J.; Sazanov, L. A. *iScience* **2021**, *24*, 102139, DOI: 10.1016/j.isci.2021.102139.

(311) Aleksandrov, L. A.; Jensen, T. J.; Cui, L.; Kousouros, J. N.; He, L.; Aleksandrov, A. A.; Riordan, J. R. *Protein Expression and Purification* **2015**, *116*, 159–166, DOI: 10.1016/j.pep.2015.09.018.

(312) Csanády, L.; Töröcsik, B. *eLife* **2019**, *8*, DOI: 10.7554/eLife.46450.001.

(313) Zhang, J.; Hwang, T.-C. *The Journal of General Physiology* **2017**, *149*, 355–372, DOI: 10.1085/jgp.201611664.

(314) Yu, H.; Yang, Z.; Li, F.; Xu, L.; Sun, Y. *Drug Delivery* **2020**, *27*, 1425–1437, DOI: 10.1080/10717544.2020.1831103.

(315) Baym, M.; Lieberman, T. D.; Kelsic, E. D.; Chait, R.; Gross, R.; Yelin, I.; Kishony, R. *Science (New York, N.Y.)* **2016**, *353*, 1147–1151, DOI: 10.1126/science.aag0822.