BRAIN-SPINE-MUSCLE INTERFACES

BY: JACKSON POWELL
FOR: FUTURE JACKSON

root@/last_updated: January 17, 2024

Once you know the way broadly, you can see it in all things.
– The Book of Five Rings by Miyamoto Musashi

Contents

1 Prologue **1**
- 1.0.1 Purpose . . . 1
- 1.0.2 To Do: . . . 1
- 1.1 The Flagship Works . . . 2

I Electronics **3**

2 Background Information **4**
- 2.0.1 Overview . . . 4
- 2.1 How Electricity Works . . . 4
- 2.1.1 The misconception . . . 4
- 2.2 Laws and Devices . . . 5
- 2.2.1 Thevenin and Ideality . . . 6
- 2.2.2 LED Circuits and PNP . . . 8
- 2.3 Resistor Lattice Digression . . . 9

3 Capacitors, Inductors, Filters **15**
- 3.1 Overview . . . 15
- 3.2 Capacitor Behavior . . . 16
- 3.3 Filtration Conceptually . . . 17
- 3.3.1 Integration and Differentiation . . . 17
- 3.4 Frequency Dependence . . . 19
- 3.4.1 Imaginary Numbers Digression. . . . 20
- 3.5 Impedance . . . 21
- 3.6 Filtration Quantitatively . . . 22
- 3.6.1 Corner Frequency and Phase Shift . . . 23
- 3.6.2 Summarizing Thoughts . . . 23
- 3.7 Inductors . . . 24
- 3.8 Peak Detection . . . 25
- 3.9 Input and Output Resistance . . . 26

4 Opamps **28**
- 4.1 Introduction and Gold . . . 28
- 4.1.1 Golden Rules . . . 29
- 4.2 Simple Opamp Circuits . . . 29
- 4.3 Opamp Imperfections . . . 34

5 Transistors **36**
5.0.1 Transistor Overview . . . 36
5.1 Transistor Circuits . . . 36
5.1.1 A Follower . . . 36
5.1.2 Amplifiers . . . 39
5.1.3 Current Sources . . . 40
5.1.4 The Components of an OpAmp . . . 41

6 A Digital World **45**
6.0.1 Introduction . . . 45
6.1 Digital Logic, and FETs . . . 45
6.1.1 Digital to Analog Conversion . . . 47

7 Writing Hardware **48**
7.0.1 Introduction . . . 48
7.1 Building to a Computer . . . 52
7.1.1 Multiplexers . . . 52
7.1.2 Altogether, and Implications . . . 54

II Math and Models **55**
7.2 Perspective . . . 56
7.2.1 Further Insight . . . 56

8 Math Essentials **58**
8.1 Linear Algebra . . . 58
8.1.1 Cross Products, Dot Products, and Gradients . . . 58
8.2 Markov-chains . . . 58

9 Neuron Modeling **60**
9.1 Hodgkin-Huxley . . . 60
9.1.1 The Main Form . . . 60
9.1.2 Gating and Conductance . . . 60
9.2 Fitzhugh-Nagumo Reduction . . . 61
9.2.1 Why would we simplify this system? . . . 61
9.2.2 How to Reduce . . . 62
9.3 Diffusion . . . 64
9.3.1 Diffusion Equation Derivation . . . 64
9.3.2 Forward Euler's . . . 66
9.3.3 Backward Euler's . . . 67

10 Math in Electronics **69**
10.1 Modeling Circuits . . . 69
10.1.1 Error Correction Codes, Hamming . . . 70

11 Fluid Mechanics and Cell Motility **72**
11.1 Introduction . . . 72
11.2 Cell Motility . . . 72
11.3 Fluid Mechanics . . . 72

11.3.1 Conservation of Mass . . . 75
11.3.2 Stress Tensor . . . 76
11.3.3 Conservation of Momentum . . . 78
11.3.4 Navier-Stokes Equations . . . 79
11.3.5 Propulsion . . . 80
11.4 Literature . . . 81
11.4.1 CSF Flow . . . 81

12 Machine Learning **83**
12.1 Supervised Learning . . . 84
12.1.1 Various Models . . . 84
12.1.2 Neural Networks . . . 86
12.1.3 Linear Regression . . . 86
12.2 Unsupervised Learning . . . 87

III Physiology **88**
12.3 Perspective . . . 89

13 Biochemistry **90**
13.1 Membranes, Ions, and Potentials . . . 90
13.1.1 Action Potentials . . . 93
13.2 Ion Channels . . . 94
13.2.1 Voltage Gated Channels . . . 94
13.3 Current Literature . . . 95
13.3.1 Piezo . . . 98
13.3.2 An Aside: Quantum Biology. . . . 99

14 Muscle Physiology **101**
14.1 Skeletal Muscle . . . 101
14.1.1 A Cellular Level . . . 101
14.1.2 Generating Force . . . 102
14.1.3 Disorders Digression . . . 103
14.2 Smooth Muscle . . . 104
14.2.1 A Cellular Level . . . 104
14.3 Cardiac Muscle . . . 105
14.3.1 A Cellular Level . . . 105

15 Biomechanics of Movement **107**
15.1 Walking and Running . . . 107
15.1.1 Walking . . . 107
15.1.2 Running . . . 109
15.2 Muscles and Tendons Modeling . . . 111
15.2.1 Muscle Architecture and Modeling . . . 112
15.2.2 The Musculoskeletal System . . . 113
15.3 Quantifying Movement . . . 114
15.3.1 Transformations . . . 115
15.3.2 Computing the Moments . . . 115

16 Sensory Processing **117**
16.1 Various Senses . . . 118
16.1.1 Touch . . . 118
16.1.2 Temperature . . . 118
16.1.3 Pain . . . 118
16.1.4 Olfaction . . . 118
16.1.5 Ears in Hearing and More . . . 118
16.1.6 Vision . . . 120

17 The Central Nervous System **122**
17.1 Spinal Cord General Structure . . . 122
17.2 Autonomic Nervous System . . . 123
17.2.1 Sympathetic Nervous System . . . 123
17.2.2 Paraympathetic Nervous System . . . 124
17.3 Somatic Nervous System . . . 124
17.4 Acetylcholine, Nicotinic Receptors, and Muscarinic Receptors . . . 125
17.5 Cerebrospinal Fluid . . . 126
17.5.1 Structure Overview . . . 126
17.5.2 CSF Contacting Cells . . . 126
17.6 Myelin in Health and Regeneration . . . 127
17.6.1 Initiaion of Demyelination . . . 128
17.7 Central Pattern Generation . . . 128

IV Spinal Cord Injury **129**

18 The Injury Itself **130**
18.1 Cell Specific Responses . . . 130
18.1.1 Immune Response . . . 130
18.1.2 Neural Response . . . 131
18.1.3 IL-12 Sensing Neurons Promote Neuroprotection . . . 133

19 In The Clinic & Therapeutic Approaches **134**
19.1 Clinical Presentation . . . 134
19.2 Auxiliary Management . . . 135
19.3 Treatments . . . 136
19.3.1 Electrical Stimulation . . . 136
19.3.2 Biomaterials . . . 137
19.3.3 Drug Treatment . . . 137
19.3.4 Surgery . . . 137
19.3.5 Rehabilitation . . . 138
19.4 Complications After SCI . . . 138

20 Approaches to Researching SCI **140**
20.1 SCI Injury Models . . . 140
20.2 The Basso Scale . . . 141
20.3 Traumatic and Nontraumatic SCI . . . 142

21 Broad Coverage of Axonal Regeneration **144**
21.1 Intrinsic Factors . . . 144
21.2 Extrinsic Factors . . . 146
21.3 Recovery of Walking via Target Reconnection . . . 147
21.4 Neuronal Excitability and Regeneration . . . 148

22 Metabolism of Neurons **150**
22.1 Programming Mitochondrial Maintenance . . . 150
22.1.1 Mitochondrial Generation and Degredation . . . 150
22.1.2 Mitochondrial Trafficking . . . 151
22.1.3 Response to Stress . . . 151
22.1.4 Neuronal Injury and Regeneration . . . 152

23 Circuit Reorganization and Plasticity **153**
23.1 Reorganization After Injury . . . 153
23.1.1 Interneurons and KCC2 . . . 155
23.2 Non-coding RNA in Rewiring and Plasticity . . . 156
23.3 Synaptic Pruning . . . 157
23.3.1 Ube3a E3 Ligase Promotes Pruning in Flies . . . 157

24 Stem Cell and Tissue Engineering **159**
24.1 Cell Transplantation Therapy . . . 159
24.1.1 Long Distance Growth of Stem Cells . . . 160
24.1.2 3D Printed Spinal Scaffolds . . . 160

V Brain-Spine-Muscle Interfaces **162**
24.2 Perspective . . . 163

25 Interfacing with Nerves **164**
25.1 Overview and Historical Dealings . . . 164
25.2 Overview . . . 165
25.2.1 Electrostimulation . . . 165
25.2.2 Flexible Brain-Computer Interfaces . . . 166

26 Peripheral Nerve Stimulation **168**
26.0.1 Adaptive, Conductive, and Electrotherapeutic Scaffolds . . . 168
26.0.2 Vagus Nerve Modulates Circuits via Acetylcholine . . . 169
26.0.3 Hand Movement Recovery via Intrafascicular Stimulation . . . 170

27 Spinal Cord Stimulation **171**
27.0.1 Electronic Dura Mater . . . 172
27.0.2 Electronics with Shape Actuation . . . 173
27.0.3 SCI Specific Cells . . . 174
27.1 Clinical Applications . . . 174
27.1.1 Paresthesia and SCS . . . 175
27.1.2 Restoring Feedback in Phantom Limb . . . 176

28 Brain Stimulation **179**

28.1 Deep Brain Stimulation . . . 179
28.1.1 Technology Overview . . . 179
28.1.2 In the Clinic—Parkinson's and Essential Tremor . . . 180
28.1.3 The Surgery Itself . . . 180
28.2 Literature . . . 181
28.2.1 Recruitment of Neurons During ES . . . 181
28.2.2 Living Electrodes . . . 181

29 Muscle Stimulation **183**
29.0.1 Injectable Prosthesis for Muscle Rehab . . . 183

30 Reading Thoughts **185**
30.1 An Overview of Reading . . . 185
30.1.1 Reading Devices and Methods. . . . 185
30.1.2 What One is Actually Reading . . . 186
30.2 The Electroencephalogram (EEG) . . . 188
30.2.1 Therapeutic Applications . . . 188
30.2.2 Electro-convulsive Therapy (ECT). . . . 189
30.2.3 Artifacts . . . 190
30.3 The Flagship in Reading Thoughts . . . 191
30.3.1 The Works . . . 191
30.4 Literature . . . 193
30.4.1 Reading LFPs Without Surgery . . . 193
30.4.2 Neurograin . . . 194
30.4.3 Robotic Control with Eye. . . . 195
30.4.4 Computing via Organoids. . . . 195

31 Synthesis **196**
31.1 The Flagship Brain-Spine Interface . . . 196
31.1.1 WIMAGINE ECoG . . . 197
31.1.2 The Work . . . 198
31.2 The Flagship in Synthesis . . . 199
31.2.1 SCS-DBS Intersection . . . 200

VI Some Ideas **201**

32 Ramblings **202**
32.1 Idealized World . . . 202
32.2 Device Idea . . . 203
32.2.1 Motivation . . . 203
32.2.2 Framework . . . 203

VII Addendum **206**

33 Neurodegeneration **208**
33.1 Alzheimer's . . . 208
33.2 Huntington's . . . 220

33.3 Fragile X 230
33.4 Amyotrophic lateral sclerosis 234
33.5 Prion Diseases, and the Prion-like Behavior of Others 243

Chapter 1

Prologue

1.0.1 Purpose

The purpose of this book is to aggregate all of the content and tools I will need in forging the future I desire: building better brain-spine-muscle interfaces. Interestingly, I already find myself referencing the information here quite a bit! It has already paid back itself in full.

1.0.2 To Do:

1. More neuron regeneration models (SCN, etc.).
2. Wrap-up electronics intro (Thev., Nort., etc.).
3. Diodes section :(
4. Markov Models :(
5. Sensory processing :(
6. Better breakdown of cellular localization in the spinal cord.
7. Clinical outcomes section.
8. Giovanni age-dependent regeneration article.
9. Deepen neuron activity section with focus on Bradke, Munc-13 paper, alpha2delta2, and their L-type vgcc work
10. Sections for: Minassian, Rossignol, and the other titans.

1.1 The Flagship Works

Perhaps one of the most amazing privileges of taking courses with great, experienced professors is that they literally lived through the turning of the field. Dr. Nancy Bonini, a superb scientist, was largely at her maximum productivity through the 90s and early 2000s—exactly aligned with when the field of neurodegenerative diseases left its zygotic stage and began to pick up steam. Therefore, having her as a professor is wonderful, as she can recall exactly the implications of works as they were released—she is a historian and a scientist in one. Therefore, let us take a moment to notice the flagship works of my generation. We'll begin in 2023:

2023.

Lorach, H., ... Courtine, G. Walking naturally after spinal cord injury using a brain–spine interface. *Nature*. May 24, 2023. `https://doi.org/10.1038/s41586-023-06094-5`

> The paper that shook up the world. This paper is special in many ways. It's one of the few that got non-biologists, and those not interested in medicine talking about it. Perhaps the greatest benefit of this work is it put brain-computer interfaces on many people's radars.

Willet, F., ... Henderson, J. A high-performance speech neuroprosthesis. *Nature*. August 23, 2023. `https://doi.org/10.1038/s41586-023-06377-x`

Metzger, S., ... Chang, E. A high-performance neuroprosthesis for speech decoding and avatar control. *Nature*. August 23, 2023. `https://doi.org/10.1038/s41586-023-06443-4`

> These works marry two ever improving fields: machine learning and brain-computer interfaces. Too, there was no end to the headlines surrounding these works. They demonstrate the use of BCIs beyond neural repair, and instead in diseases that might otherwise be described as "without recourse". They highlights the ability to restore life to patients that seemingly have no recourse.

Milekovic, T., ... Courtine, G. A spinal cord neuroprosthesis for locomotor deficits due to Parkinson's disease. *Nature Medicine*. November 6, 2023. `https://doi.org/10.1038/s41591-023-02584-1`

> This is plausibly the most impressive work of the decade. When speaking with a neurosurgeon about it, he commented that it made him feel defeated, in a sense, because this work is so far beyond what many can hope to accomplish. The impressiveness is not in the results, per se, but rather in the absolutely incredible ability to pull this off. What we can look forward to is a bright, fascinating future.

Part I

Electronics

Like many of the vagabonds who live in the fields, stray horses seemed to him to be good-natured things. When you're through with them, they ask for nothing; they just go off quietly somewhere by themselves.
– Musashi by Eiji Yoshikawa

Chapter 2

Background Information

2.0.1 Overview

The bulk of this information comes from the Lab Electronics course at the University of Pennsylvania, taught to me by Professor Ashmanskas[1,2]. As a comment on notation: V represents a variable value, while V represents a unit (or at least, that's what it should be; sometimes I forget proper notation). Graphics like this were made using CircuitTikz, whose manual can be found here[3].

2.1 How Electricity Works

2.1.1 The misconception

One often represents electricity with the canonical metaphor of water flowing in a tube, equating the flow of water, which powers a wheel of some sort, as being the equivalent of electrons pushing through a wire. Incidentally, this simplifying schema illustrates a key misconception in the nature of electricity. This is easiest exemplified in the mode with which electricity reaches one's house from a power-plant, which before arriving will be subject to breaks in the circuit (transformers). In the traditional viewing of electricity, that is taught in early education, this is disconcerting as if electrons can not physically go from a power plant to the lights in your home, then how can their kinetic energy be transferred to them and turn them on?

Regarding the flow of *energy*: when a battery sits without wires attached, around it is an electric field. This field does not dissipate because no electrons flow from it. When wires are attached, charge accumulates on the surface of the wires. This causes a small electric field within the wires, but the drift velocity of electrons within the wire is quite slow—nowhere near the speed of light that you might expect electricity to flow. However, the flow of electrons within the wire is sufficient to drive an electric field which exists all around the wires. From this, we can determine the direction that energy will flow by taking the cross product of the electric and magnetic fields. In fact, if an lightbulb is attached, this means that energy flows from the battery to the bulb in all directions, not through the wires itself. This energy flow induces the vibration of electrons within the bulb's filament, thereby causing light. This means that a net flow of electrons is not required to power a bulb—but rather

[1] `https://www.hep.upenn.edu/Classes/Phys364_fall23/`
[2] `http://www.hep.upenn.edu/~ashmansk/`
[3] `https://texdoc.org/serve/circuitikzmanual.pdf/0`

only their vibrations. Thus is illustrated next:

Let us consider now what will happen with an AC circuit (120V AC outlets around your home), where the electromotive force flips with each cycle. In this case, both the electric and magnetic fields switch directions, meaning that their cross product will remain the same and, again, energy flows in all directions to power the lightbulb. Notably, the electrons do not move much (if at all) in this setup—but this is not a surprise, as it is not the electrons that carry the energy anyway. Now, it is still essential to recognize that it is the movement of electrons within the filament of a lightbulb that creates light. This is, indeed, from kinetic energy transferred from electrons bouncing against the metals lattice, dissipating energy in the form of light. The necessary distinction is that it is not electrons that flow all the way from the battery, but rather it is vibrations of those that were, and always will be, within the bulb itself. When you consider it like this, it is straightforward—as the electric field derived from the battery is what provides the electrons with enough kinetic energy to power the bulb.

Interestingly, comparisons to the "water flow" model fail dramatically in the traditional sense, but the Venturi Effect used to describe fluid flow actually succeeds. In adding a bulb you add a resistor, which is comparable to adding a part of a pipe with a smaller diameter. As water will flow faster in this section of the tube, so too will electrons. In order to maintain the same current as is through the rest of the tube, the drift velocity, V_d, must be higher. V_d is proportional to the electromotive force, E, meaning the force is highest within the bulb. Things like V_d are simplified into Ohm's Law ($V = IR$) and not often discussed.

Of course you may say: "well, then why do we use wires at all?" The answer is that wires are helpful in channeling the fields, thereby making them more efficient. But, we do not *need* wires, per se. Think of wireless charging, for example. Knowing all of this, in this work I will almost invariably describe the flow of current as electrons moving through a wire. This is because it is much easier to think of electricity in this way, hence the ubiquitous misconception.

Nuances in the Fields.

It is worth explicitly highlighting that the electric field that causes the actual flow is from charges along the wires, rather than the battery. This is notable because if this were not the case, the proximity of the bulb to the battery would dictate its brightness. One may wonder how this type of charge distribution can be established so rapidly, and the reason is that the distance an electron needs to travel in order to create such a distribution is subatomic in size—meaning that with movement at the speed of light, the time it takes to establish a surface charge is effectively zero.

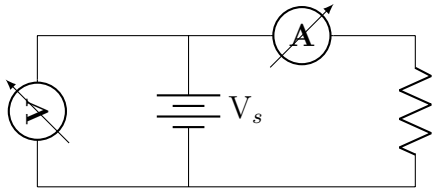
The whole idea is quite unintuitive, so it is appropriate to keep it smushed in the back of your mind, and to only draw it out when encountering things that are otherwise strange—like the aforementioned wireless charging, which should now be much more comprehensible.

2.2 Laws and Devices

The currents flowing in and out of a node will always be equal. The sum of the voltages over an entire circuit will always be equal. Kirchoff's Voltage Law (KVL) can be used to show that wires connected in parallel will have the same voltage across them, and the current flowing in and out of a node will always be the same, defined in Kirchoff's Current Law (KCL). In this way, we can predict the current

flowing through a circuit to be $V = IR$. The formula for power is given as: $VI = P$, which means that when one solves for voltage, and knowing that current is in units of charge/time and power in work/time, voltage is work per unit charge.

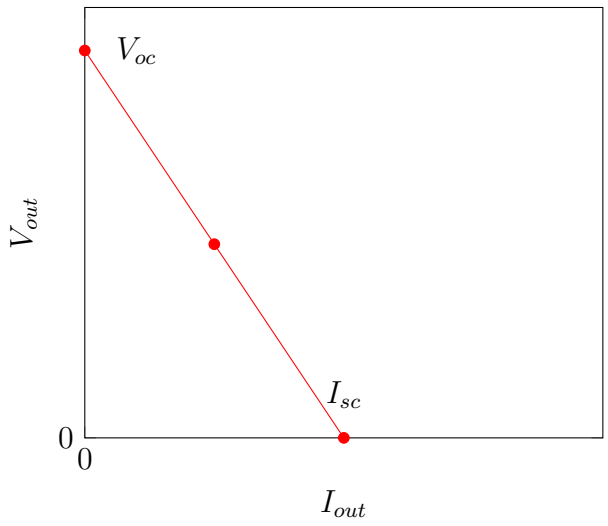
Voltage is measured using a voltmeter, a device in parallel with the load you are interested in measuring. An ammeter is used to measure current, which will be in series with the current you are trying to measure. This means that the voltmeter should have an extremely high resistance, so as to not draw any current, and an ammeter to have a low resistance, so as to not have any voltage drop. These are important considerations, as if the resistance of the load you are measuring is large (say, 1MΩ), it is possible that the voltmeter will have some non-negligible current flow through it. The same goes for if your circuit has very low resistance and you use an ammeter. To illustrate, one would measure the voltage and current coming from a battery as seen below:



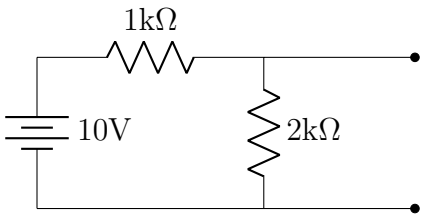
2.2.1 Thevenin and Ideality

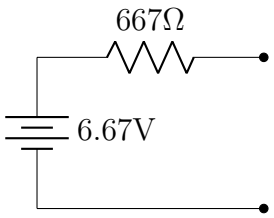
A Thevenin Circuit simplifies the circuit to have a single resistance, R_{th}, and a single voltage, V_{th}. R_{th} can be calculated via replacing all of the voltage sources with a wire, and disconnecting all of the current sources. This "short circuits" your circuit and leaves you with only resistors, which can be used to calculate R_{th} using the familiar resistor rules. One can also short circuit the terminals, and determine the current flow, giving us $R_{th} = V_{th}/I_{sc}$.

R_{th} can be measured in a circuit by varying the R_{load} added to a circuit. In this case, you will see the voltage supplied (and corresponding current) change. The slope of this change ($\Delta V/\Delta I$) will equal R_{th}. If you vary the load through the two terminals and measure the voltage across it, you will get a graph that looks something like this:



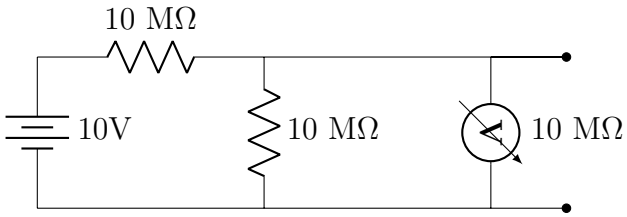
Again, the slope is what gives you R_{th}. I_{sc} is an important value which allows you to calculate V_{th}. The current that flows when you short circuit the load (I_{sc}), multiplied by R_{th}, gives you V_{th}.





The ideality of sources.

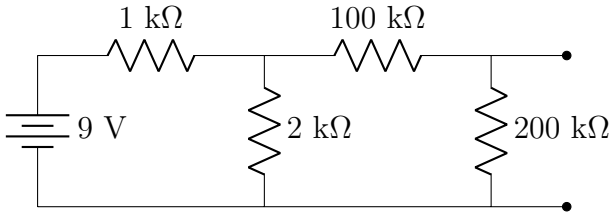
This coaxes us lightly into the topic of source ideality. Imagine all voltage or current sources as having a resistor in parallel with it, but inside of the component itself. An ideal battery, or voltage source, would be able to drive the same voltage, irrespective of the resistor/current. This would be like having a battery whose internal resistance is 0, causing the entirety of the voltage drop to occur on the circuit fragments outside of the battery. In the real world, batteries are not ideal. The canonical illustration of a batteries ideality is in trying to use a 9V battery to start your car. Naturally, the voltage dwindles as the current supplied increases. You can calculate the internal resistance of a battery by adding increasingly large loads to it, thereby giving you the batteries IV curve. It is worth considering this, as if your R_{load} is only $\approx 10 \times R_{th}$, then you may see drooping in the voltage supplied. Another way to state this is to make sure that the *input resistance* of your voltage source is much smaller than the *output resistance* of the upcoming circuit fragment you are attempting to drive. An example of when this fails is as follows:

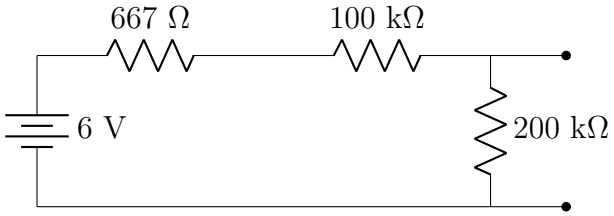


Because the circuit has a non-negligible resistance relative to the voltmeter, you should expect to read something inconsistent with using an ideal voltmeter. As we are adding a 10 MΩ voltmeter in parallel to our 10 MΩ resistor, we expect that the "R_{load}" in this case will now be 5 MΩ, so the voltage divider at V_{out} will now be $1/3 \times 20$V, or 6.67V.

This contrasts to a current source, whose desired internal resistance is ∞, as you will want no current to flow through it, and to flow entirely through the circuit fragments outside of the component. Once again, as the real world is not ideal, your goal in this case will be to have downstream circuit components whose input resistance is much smaller than the components output resistance.

Let us consider another example of input resistance vs. output resistance:

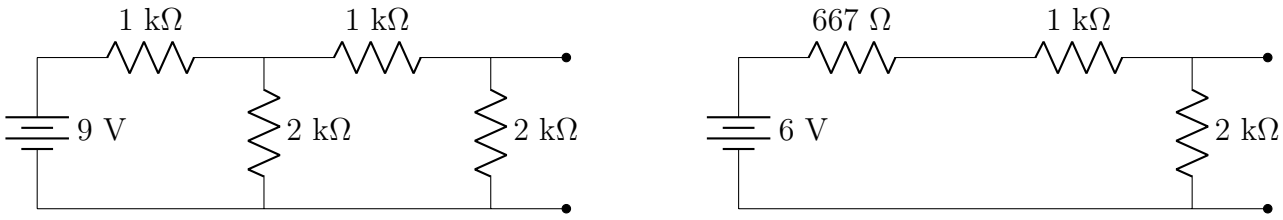




If we draw a black box around the first voltage divider, we can convert it to the circuit schematic on the right because 1 kΩ||2 kΩ = 667 kΩ just as before. We can find the V_{th} by first finding I_{sc}, when

we short the black box's load. So, I_{sc} = 9V/1kΩ = 0.9mA. We multiply I_{sc} (9 mA) by R_{th} (6.67 kΩ) to get a V_{th} of 6 V. If you were to compare the output resistance of the black box, 6.67 kΩ, to the input resistance of the upcoming voltage divider, 300 kΩ, you would find that it is much smaller. This naturally means that the entirety of the 6 V drop will occur over this part of the circuit.

Another way to think about this is as two successive voltage dividers, and qualitatively noting that the second's total resistance is much higher allows us to simplify things greatly. As a reminder, a voltage divider can be solved as $R_A I = R_A \times V_{total}/R_{total} = V_{total} \times R_A/(R_A + R_B)$. If you redraw as the Thevenin equivalent, the first voltage divider can effectively be ignored, because the voltage drop across this component will be minimal. Not by coincidence, since the first voltage divider is a 2/3 divider, and the second is a 2/3 divider, the voltage measured between our two terminals is 2/3 × 2/3 × 9V, or alternatively, 2/3 × 6V (V_{th}). This equivalency does not work when the input and output resistances of each fragment are comparable, here is an example:

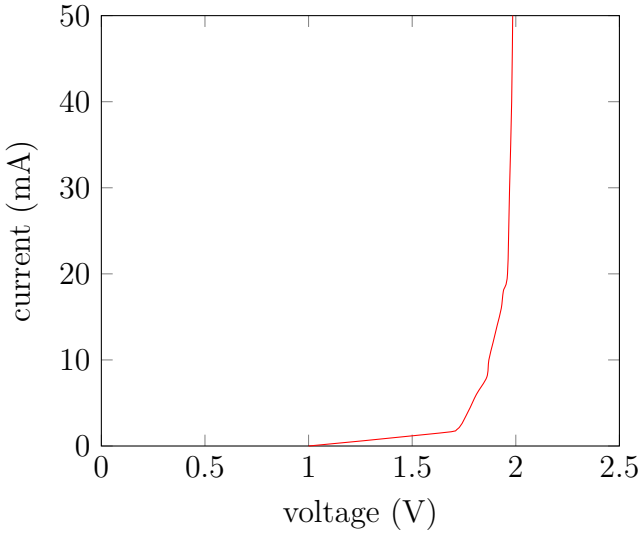


In this case, R_{th} will be the same, and R_{load} will be 3 kΩ. $V_{th} \times R_A/(R_A + R_B) = 3,000/3,667 \approx 5.45$V for the output of the first divider (i.e., between R_{th} and the 1 kΩ resistor). And then naturally, if you were to measure the voltage between the 1 kΩ and 2 kΩ resistors, it would be 2/3 × 5.45V ≈ 3.63V for the voltage at the second divider.

This kind of Thevenin analysis only works when you have a linear IV curve. When might you have a non-linear IV curve?

2.2.2 LED Circuits and PNP

The IV curve across a light emitting diode (LED) should look something like this:

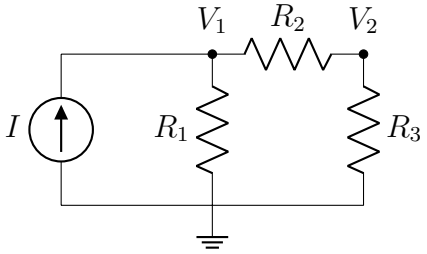


The IV curve for a diode, like an LED, is exponential in that the current slowly increases after the voltage across a diode hits some "threshold," after which the current rises exponentially with voltage. Why is this the case? A diode is a P-N junction bridged by some depletion zone. The P side of the diode contains positively charged elements that act as "holes" (a silly way to say there is an absent electron position). The N side contains elements whose outer layers are loosely filled with electrons (i.e., low ionization energy). Effectively, the P side is devoid of electrons, while the N side has many free to give. What does this mean with regard to current and voltage? It means that the "depletion zone" between the two requires electrons to be able to bridge the gap. This really can't happen unless they have a certain amount of energy, so increasing the voltage helps reach the "threshold" energy requires to pass the depletion zone (think of $P = IV$). Thus, as the electrons somewhat saturate the diode, you can theoretically pass an infinite current through it, as it will be effectively a short circuit.

2.3 Resistor Lattice Digression

One issue with modern BSIs is the usage of ECoGs, or EEGs, or other large measuring devices[4]. Too, they are almost universally hard electronics that require intense surgeries to implant. Therefore, if we could replace with soft electronics, we can cover considerable ground. For example, if one could drill a small hole into a patient's skull and spread over the cortex a fabric that contained electrodes, one could achieve a similar amount of readings with a minimally invasive surgery. It is probable that there will be electrodes small enough to accomplish this, but let's say there aren't. Another way that this could be solved is using a lattice of resistors, with probes at either corner. These corners can exit the brain and be the points at which a computer interfaces with them.

Don't be annoying, just go with the process.



Please, don't be annoying—the solution is trivial but we will be talking about methods you can generalize. This example is from the SPICE method[5]. Considering KCL at the nodes gives us these two equations:

$$\begin{aligned} I &= \frac{V_1}{R_1} + \frac{V_1 - V_2}{R_2} \\ \frac{V_1 - V_2}{R_2} &= \frac{V_2}{R_3} \end{aligned} \tag{2.1}$$

As it goes, we can reformat this so as to easily turn it into a matrix in the following way:

[4]This is expanded on in the later parts.

[5]Thank you Prof. Ashmanskas, `http://www.ecircuitcenter.com/SpiceTopics/Overview/Overview.htm`

$$
\begin{aligned}
I &= \frac{1}{R_1}V_1 + \frac{1}{R_2}\left(V_1 - V_2\right) \\
I &= \frac{1}{R_1}V_1 + \frac{1}{R_2}V_1 - \frac{1}{R_2}V_2 \\
I &= \left(\frac{1}{R_1} + \frac{1}{R_2}\right)V_1 - \frac{1}{R_2}V_2
\end{aligned}
\tag{2.2}
$$

$$
\begin{aligned}
\frac{V_1 - V_2}{R_2} &= \frac{V_2}{R_3} \\
0 &= -\frac{1}{R_2}V_1 + \frac{1}{R_2}V_2 + \frac{1}{R_3}V_2 \\
0 &= -\frac{1}{R_2}V_1 + \left(\frac{1}{R_2} + \frac{1}{R_3}\right)V_2
\end{aligned}
\tag{2.3}
$$

These sorts of equations will usually be simplified using conductance as below:

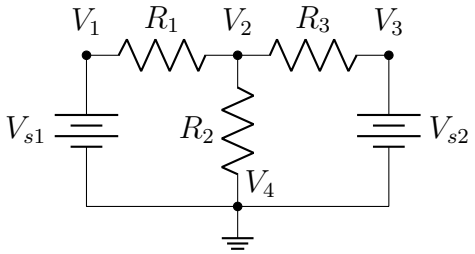
$$
\begin{aligned}
\left(G_1 + G_2\right)V_1 - G_2V_2 &= I \\
-G_2V_1 + \left(G_2 + G_3\right)V_2 &= 0
\end{aligned}
\tag{2.4}
$$

$$
\begin{bmatrix} G_1 + G_2 & -G_2 \\ -G_2 & G_2 + G_3 \end{bmatrix}
\begin{bmatrix} V_1 \\ V_2 \end{bmatrix} =
\begin{bmatrix} I \\ 0 \end{bmatrix}
\tag{2.5}
$$

So if $R_{1,2,3} = 100\Omega$, and $I = 1\text{mA}$, then:

$$
x = \text{A}^{-1}\text{B} = \begin{bmatrix} 67\text{mV} \\ 33\text{mV} \end{bmatrix}
\tag{2.6}
$$

This method is called *nodal analysis*. You'll find more trouble trying to use a voltage source rather than a current source, and in fixing this, a method of *modified nodal analysis* was invented. Let us take the below example[6]:



In this system, we have 4 nodes, but node 4 largely functions as a reference for the other 3 nodes. A small note to make this a tad clearer is that in thinking about the current flowing into a node, such as into V_1, you will view this as current going from V_2 to V_1, but because everything in electronics is backwards, this is calculated as $(V_1 - V_2)/R_1$, and the opposite for current flowing into V_2. Annoyance aside, using KCL, and then our definitions, we can gather these equations:

[6] `https://cheever.domains.swarthmore.edu/Ref/mna/MNA2.html`

$$
\begin{aligned}
I_1 + \frac{V_1 - V_2}{R_1} &= 0 \\
\frac{V_2 - V_1}{R_1} + \frac{V_2}{R_2} + \frac{V_2 - V_3}{R_3} &= 0 \\
I_2 + \frac{V_3 - V_2}{R_3} &= 0 \\
V_1 &= V_{s1} \\
V_3 &= V_{s2}
\end{aligned}
\tag{2.7}
$$

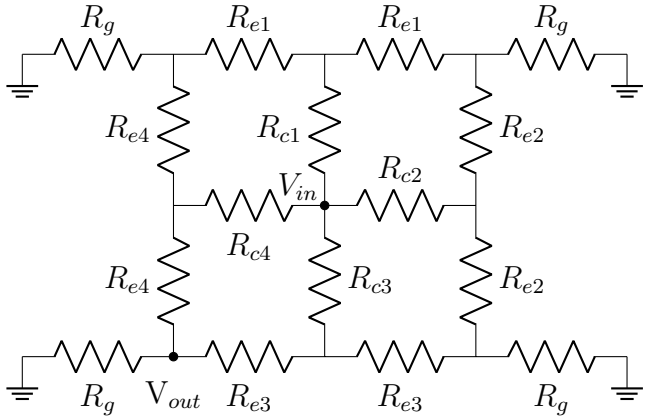
Which we convert to:

$$
\begin{aligned}
I_1 + G_1 V_1 - G_1 V_2 &= 0 \\
G_1 V_1 + \left(-G_1 + G_2 + G_3\right) V_2 - G_3 V_3 &= 0 \\
I_3 - G_3 V_2 + G_3 V_3 &= 0 \\
V_1 &= V_{s1} \\
V_3 &= V_{s2}
\end{aligned}
\tag{2.8}
$$

And then:

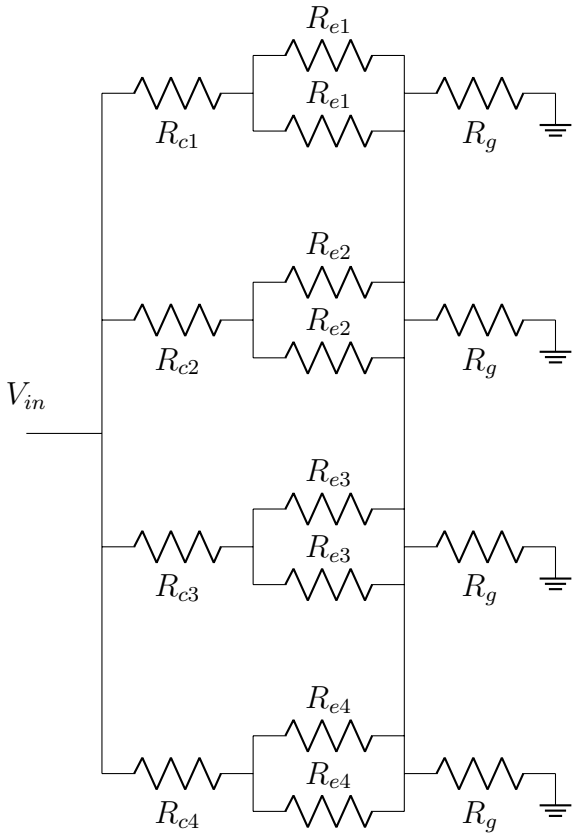
$$
\begin{bmatrix}
G_1 & -G_1 & 0 & 1 & 0 \\
-G_1 & G_1 + G_2 + G_3 & -G_3 & 0 & 0 \\
0 & -G_3 & G_3 & 0 & 1 \\
1 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0
\end{bmatrix}
\begin{bmatrix} V_1 \\ V_2 \\ V_3 \\ I_1 \\ I_3 \end{bmatrix}
=
\begin{bmatrix} 0 \\ 0 \\ 0 \\ V_{s1} \\ V_{s2} \end{bmatrix}
\tag{2.9}
$$

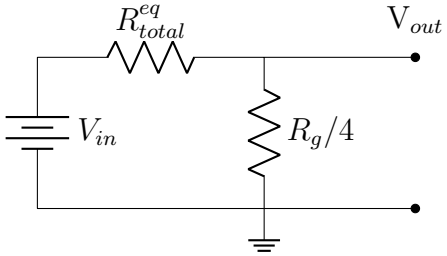
Now, I realize that one doesn't need to build a system of equations for most problems of small size, and they can be generally solved at a glance. Too, in many cases building a system is more hassle than it's worth. But, this will be important to keep in mind for the upcoming chapter **Modeling Circuits**. Let's look at some examples that are easier dealt with in the traditional way:



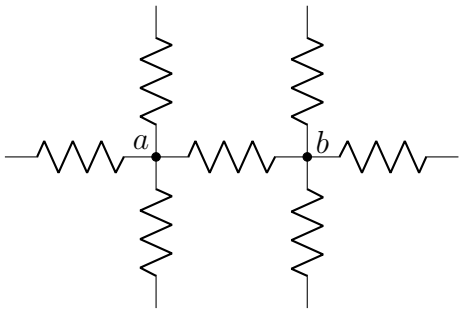
Firstly, apologies for the mess in labeling. The most difficult thing about this problem, in actuality, is how you look at it. Redrawing the circuit helps a great deal. I'll guide you through how one

might do this. In truth, the four grounded resistors are in parallel and can be simplified to one wire ($R_g||R_g||R_g||R_g$, or $R_g/4$). You can probably realize that the edge resistors, for example the pair of R_{e1}, are also in parallel and are in series with R_{c1}. Therefore, we have $(R_{e1}||R_{e1}) + R_{c1}$. I'll call this R_1^{eq}. Then you'll realize simply that all of the equivalent R are also in parallel. This gives us $R_1^{eq}||R_2^{eq}||R_3^{eq}||R_4^{eq}$, which I will call R_{total}^{eq}. Now, our circuit simplifies to:

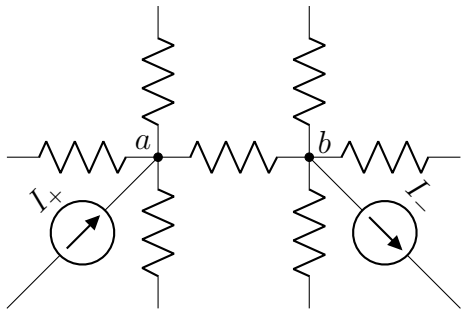




You may have had trouble at first glance because of all the "looped" sections of the circuit. It appears that current can flow any which way. Keep KVL in mind, and recall that current will only flow from high to low voltage. Therefore, you would never have current flowing from, say, R_{e2} to R_{e3}. A popular problem of the same flavor is the *infinite resistor grid*, which looks something like this:

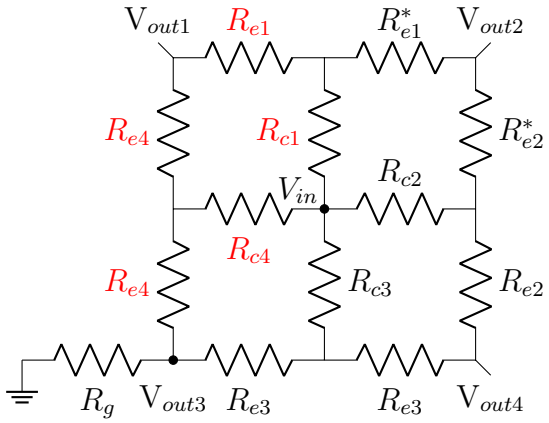


Imagine that the lattice of resistors continues out in every direction infinitely, in a grid-like fashion. The problem asks you to determine the equivalent resistance between points a and b. Considering the infinite nature, the more immediate method to solve would be to take some limit of parallel and series resistors. I actually think this is an important solution to keep in mind, as it can be generalized to other systems better. However, in this case the easiest solution is to use *superposition*. It goes as follows:

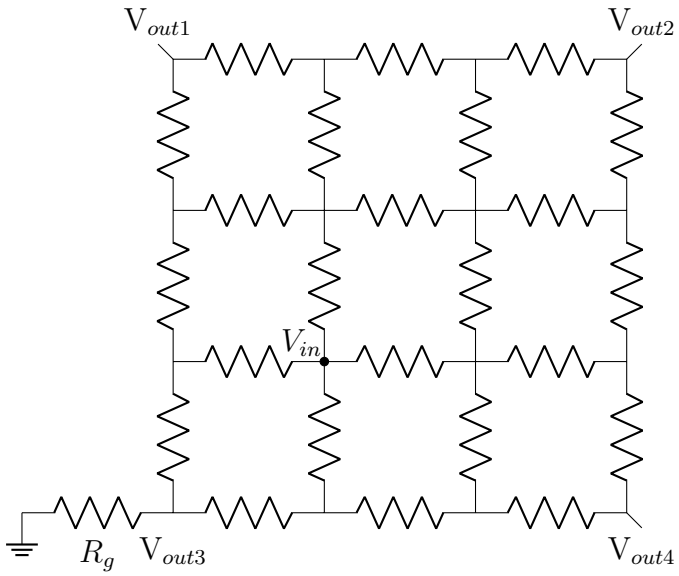


If you were to wire a positive current source to node a, you can immediately know that the current will be split equally in 4 ways, as the lattice is infinite. Then, if you apply a negative current source to node b, you can immediately know that it will draw $1/4I$ from each resistor as well. The total current flowing between nodes a and b will then be $1/2I$. Given that $V = IR$, and by superposition, we can know that $V_{ab} = (I/2)R$, so $R_{ab} = V_{ab}/I = (I/2)R/I = R/2$. The in-line math is kind of ugly, but I think you get the gist.

Another option of a similar flavor is:



In this case, we will think about what each V_{out} will read. One thing to note immediately is that the net current at V_{out2} will be 0. Therefore, R^*_{e1} and R^*_{e2} don't really matter, and can be neglected in our calculations. This once again is easiest to solve when re-visualized. The gist being that you have two quadrants (the R_{c1}, R_{c4}, which I colored to have red labels, and the R_{c2}, R_{c3} quadrants) where V_{in} is split into. The voltage drop across these quadrants being $3R||R$. We then add another R, for the one that connects to R_g, and voila, we are done. Let's call this R_q, so to find the equivalent resistance from V_{in} to V_{out3} it is $R_{q1}||R_{q2}$. I think you get the idea. The reason for going through these simple models is to prepare ourselves for some of the upcoming, more complicated ones. Let us level up now:


V$_{out1}$
V$_{out2}$
V_{in}
R_g
V$_{out3}$
V$_{out4}$

Chapter 3

Capacitors, Inductors, Filters

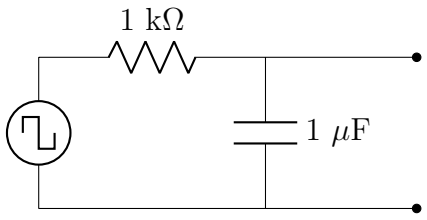
3.1 Overview

Getting a good intuition of capacitor behavior is an essential part of understanding electronics. A capacitor is a storer of charge, which allows it maintain a voltage for a period after the voltage source originally supplying the circuit is cut off. The basic structure of a capacitor is an anode and a cathode separated by a dialectric plate. Charge will be stored across it, and is released when the voltage supply dwindles. Once a capacitor is fully charged, there will be no current. However, if the voltage oscillates, the capacitor will follow V_{in}. Effectively, capacitors filter out DC voltage.

Capacitor discharge will be exponential decay in the form:

$$\begin{aligned} V_C &= V_S \times e^{-t/RC} \\ \tau &= RC \\ V_C &= \mathrm{V}_S \times e^{-t/\tau} \end{aligned} \tag{3.1}$$

This is quite similar to neurons, isn't it! Let's consider some examples to further our intuition. Let's say you have the following circuit:

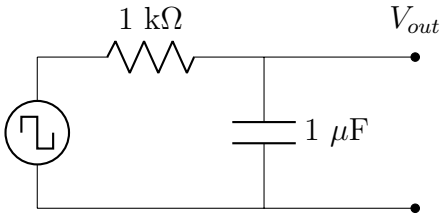


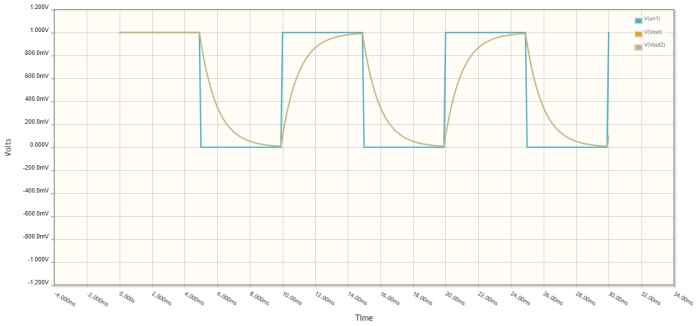
RC will be $1\mathrm{k}\Omega \times 1\mu\mathrm{F} = 1 \times 10^{-3}$, or $\tau = 1\mathrm{ms}$. What does this actually mean, though? It means that if you charge this capacitor up to 10V, then $V_C = V_S \times e^{-t/RC} \rightarrow 10V \times e^{-t/1\mathrm{ms}}$. So, 1 ms after voltage is removed, and or, one time constant, τ, after voltage is removed, the voltage at the capacitor will be $10 \times e^{-1} \approx 10 \times 0.367 = 3.67\mathrm{V}$. If you were using a square wave which oscillated between 0 and 10V with a very high frequency (100 kHz, for example), then the period would be 1/100kHz = 0.01 ms. Therefore, you would not expect the capacitor to ever fully discharge, and it would maintain a constant, high voltage. Perhaps you can intuitively predict that it would get stuck near a voltage of 5V, in the middle of our square wave input.

This circuit happens to be what is called a *low-pass filter*, which we will discuss in-depth later. But, you can see how it might get this moniker, as this very high frequency is not allowed to pass due to the capacitor's time constant. This was meant to serve as a basic intro to what a capacitor is. Now we can begin discussing the nuances.

3.2 Capacitor Behavior

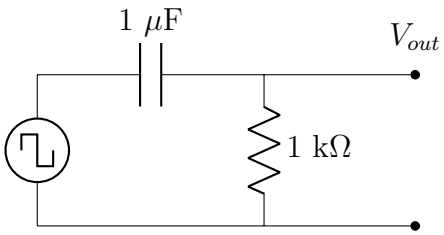
Let's look at the circuit below and its output (the same circuit as before)[7]. Only think about the shape of the output, no number. Only work on building up the conceptual understanding. This time our V_{in} (the blue square wave) is 100Hz.

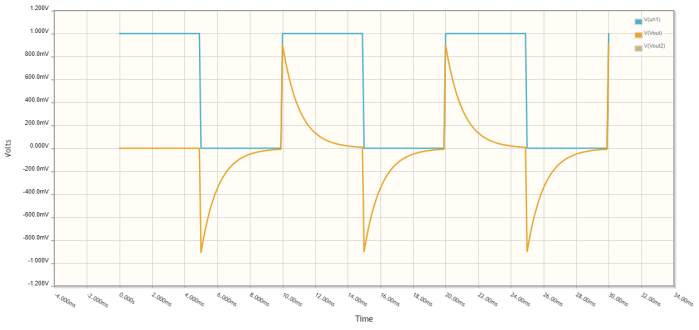




Think about how the capacitor will behave. At time $t = 0$, the capacitor is fully charged. No current will flow at this time, meaning it will act as an open circuit, and therefore have a "resistance" of ∞. You can think of the circuit as being like a voltage divider whose $R_2 = \infty$, and or $V_{out} = V_{in}$. At $t = (1/2)\omega$, the capacitor is able to discharge, and dissipates charge according to its decay curve. At $t = \omega$, the capacitor is subject to voltage again and begins to charge up. At first it acts like a wire, since it draws current proportional to the input voltage. As it becomes fully charged, it will again act like an open circuit.

Let's now consider swapping positions:





At $t = 0$, the capacitor is fully charged, so no current flows at all, and $V_{out} = 0$. This is expected, since capacitors are meant to filter out DC voltage. At $t = (1/2)\omega$, the capacitor shoots down to $-V_{pp}$. This is because V_{in} becomes negative relative to the voltage that the capacitor is held at. I.e., one plate of the capacitor is held at this V_{pp} value, so when V_{in} goes to 0, the voltage is negative between the plates of the capacitor, which draws negative current to the plate on the right side, making the voltage drop

[7] I really dislike needing to include .png images in works like this, but I believe I have no choice in this case.

across the 1kΩ resistor equal to $-IR$. It then undergoes its standard decay curve, following this rule. Again, the decay will end at 0 because capacitors filter out DC voltage.

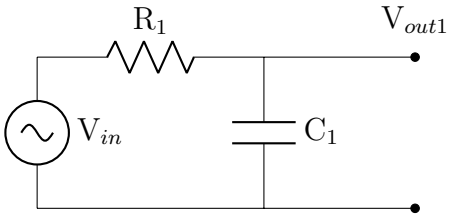
3.3 Filtration Conceptually

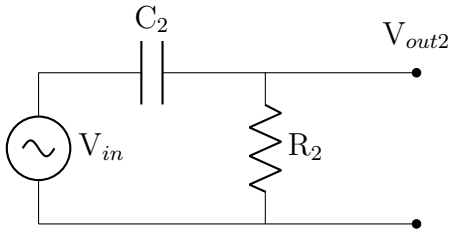
Firstly, what is filtration? Filtration refers to our filtering out of background signal, or any undesired signal, in order to isolate our signal of interest. For example, perhaps you are interested in a sine wave whose period is 1 ms, but this is superimposed on a bunch of other sine waves of variable periods. You can extract the 1ms sine wave with some clever circuits, and build your downstream circuit fragments based on it. It's like an analog version of the Fourier Transform!

Let's consider a specific example given by Professor Ashmanskas: imagine you are building a sensor that tests the water level in a pool. You don't want the pool to overflow when it rains, or get too low when the Summer comes and water evaporates—so you devise an automatic system to add or take out water as needed. One could simply add a sensor to the side of the pool, but it will be subject to the constant waves formed by people swimming, and its readings will be horribly variable and uninterpretable, as they do not reflect the so-called *steady-state* value of the water level. Each time someone cannon-balls in and splashes it, the sensor will think that the pool is greatly overflowing. Thus, one can devise a circuit that filters out all of these little fluctuations (this would be called a *low-pass filter*, for passing variations that occur on a long time-scale—low frequency).

3.3.1 Integration and Differentiation

Capacitors have this incredible ability to perform complex math, including taking derivatives or integrals of your wave form. Let's think about how this may occur using the circuit mentioned in the previous part. Firstly, recall that the current flowing through the resistor must equal the current flowing through the capacitor, and that $Q = CV$:





We can then solve for the voltage drop across R_1 as:

$$\begin{aligned} Q &= C_1 V_{out1} \\ \frac{d}{dt}Q &= \frac{d}{dt}C_1 V_{out1} \\ I &= C_1 \frac{dV_{out1}}{dt} \end{aligned} \tag{3.2}$$

$$\begin{aligned} IR_1 &= V_{in} - V_{out1} \\ \frac{dV_{out1}}{dt} &= \frac{1}{R_1 C_1}\left(V_{in} - V_{out1}\right) \end{aligned} \tag{3.3}$$

Therefore, if V_{out} is very small compared to V_{in}, you get:

$$\begin{aligned} \frac{dV_{out1}}{dt} &= \frac{1}{R_1 C_1} V_{in} \\ \int \frac{dV_{out1}}{dt} &= \int \frac{1}{R_1 C_1} V_{in} \\ V_{out1} &= \frac{1}{R_1 C_1} \int V_{in} \end{aligned} \tag{3.4}$$

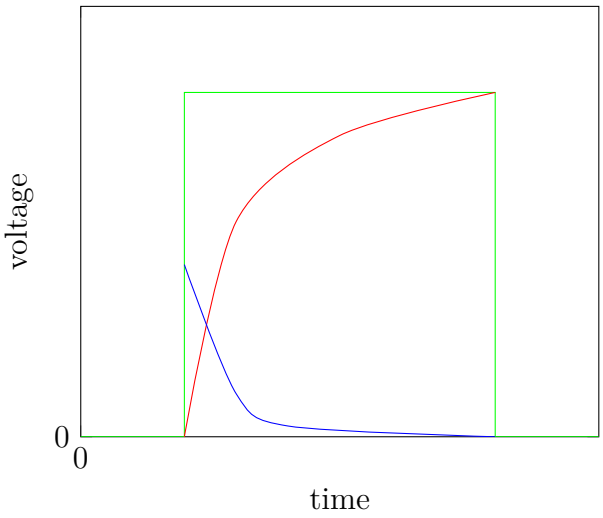
And or, that V_{out} integrates V_{in}. How will this change if we swap the positions of the resistor and capacitor in circuit 2? Once again, consider when V_{out} is much smaller than the input.

$$\begin{aligned} I &= C_2 \frac{d}{dt} (V_{in} - V_{out2}) \\ \frac{V_{out2}}{R_2} &= C_2 \frac{d}{dt} (V_{in} - V_{out2}) \\ \frac{V_{out2}}{R_2} &= C_2 \frac{dV_{in}}{dt} \\ V_{out2} &= R_2 C_2 \frac{dV_{in}}{dt} \end{aligned} \tag{3.5}$$

Thus, in this case V_{out} approximates the derivative of V_{in}.

Qualitative Thinking.

The best lead in to thinking about filtration, to me, is thinking simply about how the graphs look like when something integrates or differentiates a square wave, combined with our understanding of capacitor behavior from above. Let us not do any math, and think only qualitatively.

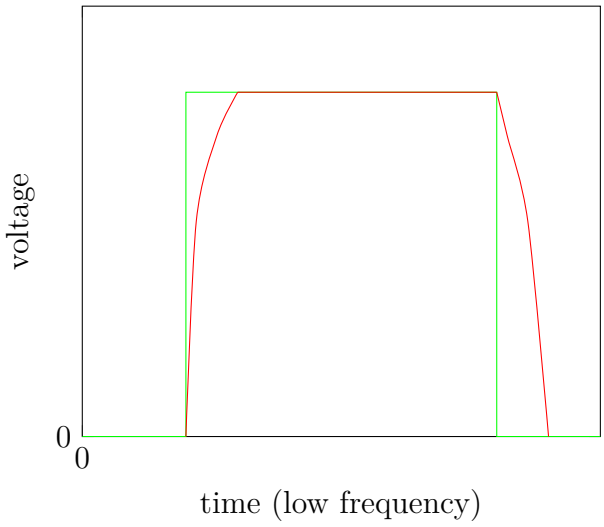


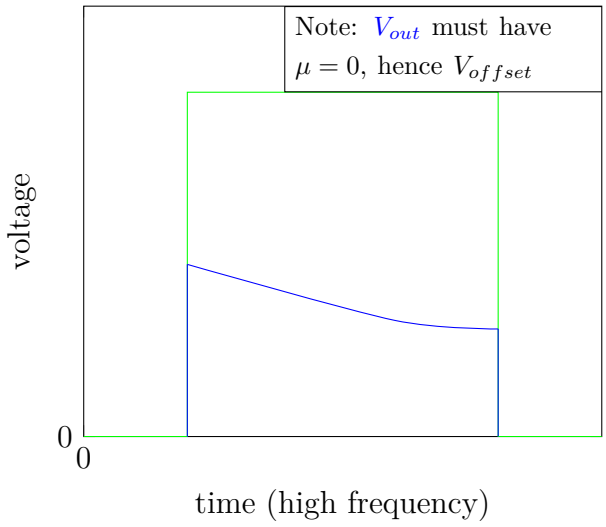
If the green waveform is our V_{in} from the previous example, then our red curve is similar to how V_{out1} may look, and the blue curve is similar to how V_{out2}. That is, the blue curve *kind of* differentiates

the green curve, because the relatively instant rise signifies a very positive derivative, marked by this blue spike. Similarly, the red curve *kind of* integrates the green curve, because the area under the green curve slowly accumulates, hence the overall positive slope of the red curve.

Something interesting you might notice, though, about the blue curve is that it seems centered around 0, while the red curve seems centered around the midway point of the green V_{in}. You can realize that the capacitor being before V_{out} means it'll filter out all DC voltage, thereby centering V_{out} about 0. Thus, you can expect there to be a symmetric, negative curve on the downward edge of the green curve, as mentioned and discussed in section **3.2**.

So if we say that the frequency is extremely slow, and the red curve is the voltage being measured at V_{out2}, how might the red curve look?





Similarly, if the frequency is very high, how might the blue curve look? Take a second to ponder the two graphs above and gather a bit of intuition on it. This should allow you to qualitatively state that an integrating circuit will be a *low-pass filter* (because V_{out} matches V_{in}), and a differentiating circuit will be a *high-pass filter* (except that there's an offset when the long-term average of $V_{in} \neq 0$). As in the left graph, as the frequency gets smaller and smaller, the red plot will more and more closely match the green plot. This is the basis of capacitor filtration, an essential tool in electronics!

3.4 Frequency Dependence

This section will primarily be quantitative. The important bits, though, are the qualitative understandings, and the final results of this quantitative section. What you find in the interim is likely not worth understanding fully. Let's begin by supposing we apply a cos wave to a resistor. The relationship between current and voltage is as follows:

$$\begin{aligned} f &= \frac{\omega}{2\pi} \\ V(t) &= V_{pp}\cos(\omega t) \\ V(t) &= I(t)\mathrm{R} \\ I(t) &= (V_{pp}/R)\cos(\omega t) \end{aligned} \tag{3.6}$$

Therefore, if you were to solve something like V_{out}/V_{in} your pesky cos terms will cancel. This simply means that V_{in} and V_{out} will be in phase, only changed by some scaling factor. Unfortunately, this is not so for capacitor equations. We can see this below:

$$\begin{aligned} Q &= CV \\ I(t) &= C\frac{d\left(V_{pp}\cos\left(\omega t\right)\right)}{dt} \\ I(t) &= -\omega C V_{pp}\sin\left(\omega t\right) \end{aligned} \tag{3.7}$$

This means that the current and the voltage will be 90° out of phase. Another way would be to write this as $A\cos\left(\omega t+\phi\right)$. Let us think about what will happen with our standard low-pass filter. Using KVL, we can state that:

$$\begin{aligned} V_{pp}\cos\left(\omega t\right) &= RC\frac{dv_C}{dt} + v_c \\ v_c &= A\cos\left(\omega t+\phi\right) \\ V_{pp}\cos\left(\omega t\right) &= RC\frac{d}{dt}A\cos\left(\omega t+\phi\right) + A\cos\left(\omega t+\phi\right) \end{aligned} \tag{3.8}$$

3.4.1 Imaginary Numbers Digression.

Maybe looked at the previous equation and thought it is solvable (probably?) but that it would be not worth your while to do so, and that there is likely a better way to go about it. Recall Euler's Relation[8]:

$$e^{j\omega t} = \cos(\omega t) + j\sin(\omega t) \tag{3.9}$$

One of the requirements of a linear circuit is superposition, and that is *kind of* the argument that allows us to use imaginary numbers. You can think that if you are using math that includes both a real and imaginary component, as long as you keep track of the real, your output will still be correct—but, don't think too hard about this. In the upcoming sections, I will use $\mathbf{v}_c$ to denote the complex number. Let us examine:

$$\begin{aligned} V_{pp}e^{j\omega t} &= RC\frac{d\mathbf{v}_c}{dt} + \mathbf{v}_c \\ \mathbf{v}_c &= Ae^{j\omega t} \\ V_{pp}e^{j\omega t} &= RC\frac{d}{dt}Ae^{j\omega t} + Ae^{j\omega t} \end{aligned} \tag{3.10}$$

We can differentiate, and then obtain an expression for A, and solve for the voltage at the capacitor as:

[8]Electrical Engineers use j instead of i for imaginary numbers. The claim is that it is easier to keep track of in the math. Other's claim it's because i is sometimes used to represent currents, particularly small ones.

$$
\begin{aligned}
V_{pp}e^{j\omega t} &= j\omega RCAe^{j\omega t} + Ae^{j\omega t} \\
V_{pp} &= j\omega RCA + A \\
V_{pp} &= A\left(1 + j\omega RC\right) \\
\frac{V_{pp}}{(1 + j\omega RC)} &= A \\
\mathbf{v}_c &= \frac{V_{pp}}{1 + j\omega RC}e^{j\omega t}
\end{aligned}
\tag{3.11}
$$

We now want to find the real component, which begins by rewriting the expression in its polar form:

$$
\begin{aligned}
\mathbf{v}_c &= \left(\frac{1}{\sqrt{1 + \omega^2R^2C^2}}e^{j\phi}\right)V_{pp}e^{j\omega t} \\
\phi &= \tan^{-1}(-\omega RC) \\
\mathbf{v}_c &= \frac{1}{\sqrt{1 + \omega^2R^2C^2}}V_{pp}e^{j(\omega t + \phi)}
\end{aligned}
\tag{3.12}
$$

From here we can simply take the real part and be on our way:

$$
v_c = \frac{V_{pp}}{\sqrt{1 + \omega^2R^2C^2}}\cos(\omega t + \phi) \tag{3.13}
$$

This is equivalent to saying:

$$
A_{out} = A_{in}\sin(\omega t + \phi) \tag{3.14}
$$

Where A is the amplitude at either time. This is one way to go about this problem. Another way, which will be discussed in the next section, is using impedance.

3.5 Impedance

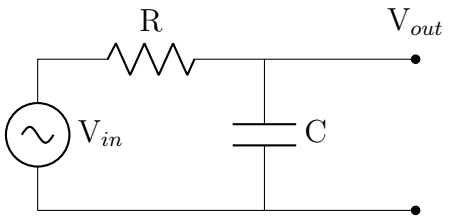
We mentioned earlier in equation (3.11) that the relationship between an input voltage and what is measured at a capacitor for a low-pass filter is:

$$
V_{in}\frac{1}{1 + j\omega RC} = v_c \tag{3.15}
$$

If we were to divide this fraction by $j\omega C$, and simplify using the representation Z_C, we would get:

$$
\begin{aligned}
V_{in}\frac{1/j\omega C}{1/j\omega C + R} &= v_c \\
V_{in}\frac{Z_C}{Z_C + R} &= v_c
\end{aligned}
\tag{3.16}
$$

This looks just like a voltage divider equation! The concept of *impedance* is used to summarize other circuit fragments, like capacitors or inductors, using some resistance equivalent Z. The impedance of a resistor, Z_R, is simply R. Z_C is $1/j\omega C$. Let's think of how this pertains to our low-pass filter.



If we measure at V_{out}, we can think of it like a voltage divider, giving us:

$$V_{out} = \frac{Z_C}{Z_C + R} V_{in} \tag{3.17}$$

You may wonder if $Z_C = 1/j\omega C$ has any real basis, or if it is simply an extraction from the above math. In reality, you can find it quite simply through:

$$\begin{aligned} I_c &= C\frac{dv_c}{dt} \\ I_c e^{j\omega t} &= CA\frac{d}{dt}e^{j\omega t} \\ I_c e^{j\omega t} &= j\omega CAe^{j\omega t} \\ I_c &= j\omega CA \\ I_c\frac{1}{j\omega C} &= A \end{aligned} \tag{3.18}$$

This is the equivalent of Ohm's law, where $1/j\omega C$ is the resistance. Understanding this, we can find our high-pass filter's equation to be:

$$V_{out} = \frac{R}{Z_C + R} V_{in} \tag{3.19}$$

3.6 Filtration Quantitatively

We are finally prepared to talk about filtration quantitatively! As was mentioned in the impedance discussion, resistor-capacitor (RC) circuits can be formulated as voltage dividers. We know that the impedance of a capacitor is $1/j\omega C$, which has some frequency dependence from ω (or, $2\pi f$). Therefore, we can get the sense that V_{out} may change depending on this frequency. With a bit of re-writing, we find:

$$\begin{aligned} \frac{V_{out}}{V_{in}} &= \frac{Z_C}{Z_C + R} \\ \frac{V_{out}}{V_{in}} &= \frac{1}{1 + j\omega RC} \\ \frac{V_{out}}{V_{in}} &= \frac{1}{\sqrt{1 + (2\pi fRC)^2}} \end{aligned} \tag{3.20}$$

We can see that as the frequency goes up, this converges to $1/\infty$, or $1/1$ as frequency goes down. Whereas, for a high pass filter:

$$\begin{aligned}\frac{V_{out}}{V_{in}} &= \frac{R}{Z_C + R} \\ \frac{V_{out}}{V_{in}} &= \frac{j\omega RC}{1 + j\omega RC} \\ \frac{V_{out}}{V_{in}} &= \frac{2\pi f RC}{\sqrt{1 + (2\pi f RC)^2}}\end{aligned} \tag{3.21}$$

Which converges to ∞/∞ as frequency goes up, and $0/1$ as frequency goes down.

3.6.1 Corner Frequency and Phase Shift

The way one quantifies this is with the "corner frequency," which you will more often hear as f_{3dB}. You'll notice that for both the high-pass and low-pass filter equations described above, the ratio of V_{out} to V_{in} is $1/\sqrt{2}$ when the frequency inputted is $f = 1/2\pi RC$. $1/\sqrt{2}$ corresponds to about 0.7, meaning that at this f_{3dB}, the output is about 70% the amplitude of the input. Thus, it is a good marker of how well your filter will work. If you are trying to pick a low pass filter that filters out anything at 1000 kHz and higher, you'll want to pick an RC circuit combination that is well below it.

Notably, filtration causes a phase shift. This shouldn't be surprising when you recall that a low-pass and high-pass filter also integrate and differentiate respectively. Let us take, for example, a sin input. If it is passed through a low-pass filter, would expect it to be integrated to $-\cos$. If passed to a high-pass, it would be differentiated to cos. Therefore, as $-\cos$ is $-90°$ ($\sin(x - \frac{\pi}{2}) = -\cos(x)$) relative to sin, and cos is $+90°$ ($\sin(x + \frac{\pi}{2}) = \cos(x)$), we would expect a low-pass filter to generate an output that lags V_{in}, while a high-pass will generate one which precedes V_{in}.

The filtration at f_{3dB} is 45°. I don't think this has a straightfoward math explanation, but conceptually it is when the real and imaginary parts are equal. Maybe don't think about this too hard.

3.6.2 Summarizing Thoughts

Let us look at these two circuits again, and see what we can tell just from a glance.

For the left circuit, since the current through the resistor and capacitor must be the same, we can say $Q = CV_{out}$, and thus $I = CV'_{out}$, that $V_{in} = IR = RCV'_{out}$, so V_{out} integrates V_{in}. Knowing that it integrates, we can intuit that it must be a low-pass filter, based on the graphs presented earlier. We can also know it is a low-pass filter if we recall that $Z_C = 1/j\omega C$, and that a voltage divider must look like $Z_C/R + Z_C$, which would give us $1/j\omega C + 1$. And, since it integrates, we know that $\int \sin = -\cos$, so there will be a $-90°$ phase shift at very high frequencies.

For the right circuit, since $V_{in} - V_{out} = V_C$, we know $CV_C' = I$, so $V_{out} = RCV_{in}'$. Therefor, it differentiates, which means it mus be a high-pass filter. Too, it must be of the form $R/Z_C + R$. And if it differentiates, then $\sin' = \cos$, so there must be a $+90°$ phase shift at very low frequencies.

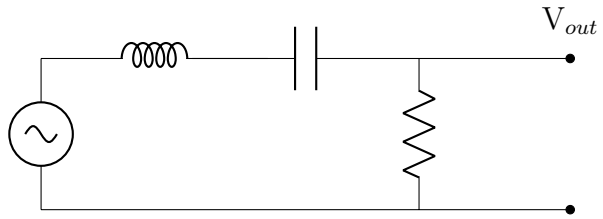
In other words, you can tell a lot just at a glance! This is without any math!

3.7 Inductors

An inductor stores energy in a magnetic field. An inductor is usually a tightly coiled bit of wire, and as you know a wire always has a magnetic field, so by coiling them and running current through it, you are summing these magnetic fields.

To conceptualize an inductor's behavior, as we will consider a water-wheel. An inductor is like a water-wheel in a pipe. As water flows through the pipe and reaches this large, heavy wheel, the speed of flow will be decreased as this wheel is pushed. Eventually, the wheel gets 'up to speed' so-to-speak, and its momentum carries it. The water's flow will not be impeded at this point. If you were to stop the flow of water, the momentum of the wheel would allow it to continue spinning, thereby continuing the water's movement for a time. If you were to have an LED in parallel with an inductor, you would expect that the LED would initially shine very brightly, as the inductor functions similarly to a large resistor its early stages. You'll see the LED dim as the inductor becomes fully charged, and now acts like a wire. If you remove the battery from the circuit, the LED will remain on for a time, as the inductor loses its charge. The rate at which the inductor dissipates its charge will depend on the total resistance of the circuit.

In fact, the magnetic field that forms as a reduce of current generates an electromotive force in the opposite direction, essentially opposing the current's flow. This field is held for some time constant, similar to a capacitor. An inductors time constant is $\tau = 2L/R$. The field is generated at at a rate of $V = L\frac{d}{dt}I$, compared to a capacitor's curve of $I = C\frac{d}{dt}V$. Therefore, you can imagine that an inductor may work opposite of a capacitor. Indeed, the impedance of an inductor is $j\omega L$, compared to a capacitors $1/j\omega C$. The f_{3dB} in this case is $R/2\pi L$. Inductors act like resistors, in that you can add them up in series like resistors, and treat them as you would with resistors in parallel. However, when we consider only impedance, everything functions like a resistor. That makes analyzing our below circuit much nicer.



This can be solved as $Z_2/Z_1 + Z_2$, where $Z_2 = R$, and $Z_1 = j\omega L + 1/j\omega C$. Let me try a bit of math here:

$$
\begin{aligned}
Z_1 &= j\omega L + \frac{1}{j\omega C} \\
Z_1 &= \frac{-\omega^2 LC}{j\omega C} + \frac{1}{j\omega C} \\
Z_1 &= \frac{1-\omega^2 LC}{j\omega C}
\end{aligned}
\tag{3.22}
$$

The impedance, Z_1, goes to zero when $\omega^2 LC = 1$. In other words, at some frequency, f_0, the impedance will be minimized. This is called the resonant frequency, and is found at $f_0 = 1/2\pi\sqrt{LC}$. You can think about what would happen when the inductor and capacitor are in parallel in the circuit above:

$$
\begin{aligned}
Z_1 &= j\omega L \,||\, \frac{1}{j\omega C} \\
Z_1 &= \frac{j\omega L / j\omega C}{j\omega L + 1/j\omega C} \\
Z_1 &= \frac{j\omega L}{1-\omega^2 LC}
\end{aligned}
\tag{3.23}
$$

Now, impedance will go to infinity at the described f_0—and you can filter out a specific frequency. It is useful to know how this filtering will affect the power dissipated by your circuit. At $\pm\Delta f = R/2\pi L$, the power halves, so one would call the bandwidth of this circuit to be $2\Delta f$. A halved power corresponds to an amplitude decrease of $1/\sqrt{2}$ (since $P = V^2/R$). Thus, at the two sides of the bandwidth, you'd expect to see an amplitude that is about 70% as large as at f_0.

Either or?

So why are capacitors much more widely used than inductors? To start, inductors are usually large and heavy. As it must be a coin of wires, usually wrapped around some magnetically-permeable metal, it doesn't easily fit into circuit boards. Too, inductors will continuously dissipate energy, while capacitors that are charged do not.

3.8 Peak Detection

An interesting application of a low-pass filter is peak detection. There may be times where you are only interested in extremes, and not the little fluctuations in between. For example, perhaps you're interested in neural activity. You probably don't care too much about individual peaks, but instead you want to get a measure of the overall activity. This is one such way to do so[9]:

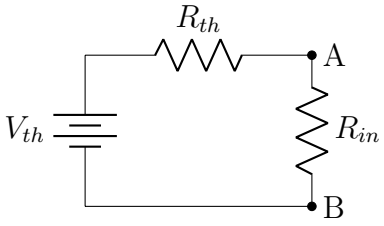
[9]These two examples are from Tom Hayes' Chapter 3N.

You can conceptually imagine that to create some form of peak detection, you want the RC decay value to be slow relative to the minute fluctuations, but fast relative to the longer term fluctuations you are interested in capturing.

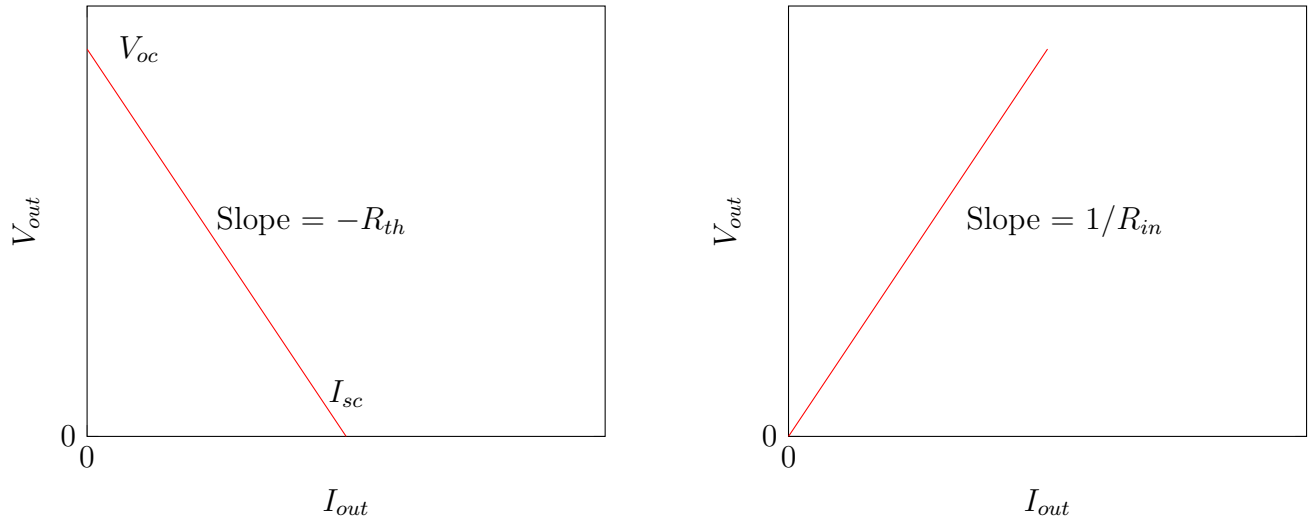
3.9 Input and Output Resistance

Now that we have a better understanding of impedance, let's discuss the idea of input and output resistance again. This concept is actually quite interesting and important, because when you are supplying a voltage to the body (say, in stimulating the spinal cord) you'll see impedance drifts, and phase shifts over time. Therefore, having a strong conceptual understanding of voltage sources & sinks, and input & output resistances is useful in quantifying this.

So for:



We get:



The left graph was presented in an earlier section. The important takeaway is that for a linear system, $R_{th} = -\Delta V_{\mathrm{AB}}/\Delta I_{\mathrm{A}}$. For a non-linear system, we can generalize this to include the impedance:

$$Z_{out} = -\frac{dV_{AB}}{dI_A} \tag{3.24}$$

Where V_{AB} is across the voltage source, and I_A is from the voltage source, and when we are varying the load resistance. Then, the input resistance of the load can be defined as:

$$\frac{1}{Z_{in}} = \frac{dI_A}{dV_{AB}} \tag{3.25}$$

For the I_A into the load and the V_{AB} across the load when we are varying the voltage source.
The next section should be diodes/LEDs, but I find those to be just so boring. So I'd rather skip to opamps.

Chapter 4

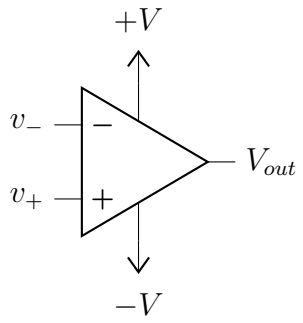
Opamps

Operational amplifiers (opamps) are one of the most ubiquitous tools in electronics, just behind classics like the resistor and capacitor. While they themselves are amazing, making CircuitTikz diagrams for them is a nightmare! So wish me luck in the coming parts!

As there are countless opamp applications, it can become muddled why you'd want a circuit to do such things. Thus, I'll do my best to think of interesting biological applications for the less obvious ones.

4.1 Introduction and Gold

Opamps are active components, in that they can take a circuit's power and increase it. They have the ability to amplify signal, act as a voltage clamp, do addition and subtraction, and even integration and differentiation. Let's discuss:



Inside of an opamp is a large network of transistors. It isn't worth trying to understand all the internals, but it is worth understanding all of the externals. An opamp has two inputs (v_- and v_+), and two power rails ($+V$ and $-V$). In general, an opamp's goal will be to minimize the difference between the two inputs, and it has some range of power to work with in doing so, which is limited by $+V$ and $-V$. It tries to minimize this input through its V_{out}. Therefore, one would often connect the v_- to V_{out} as a form of negative feedback. This will likely make more sense in the later sections. If there is no connection between V_{out} and either of the two inputs, your opamp will oscillate between the two power maximums set by $+V$ and $-V$. As in, if $v_- > v_+$, then $V_{out} = +V$. This is an opamp functioning as a *comparator*, in that it makes the somewhat binary comparison of which input is larger, and outputs either high or low voltage.

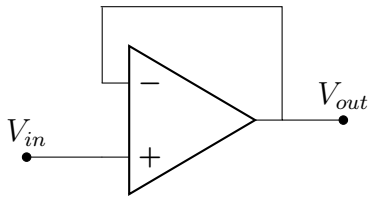
4.1.1 Golden Rules

In adding some form of negative feedback, and keeping your opamp from being saturating (i.e., clamping V_{out} to either of the two power rails)[10], an opamp will adhere to two essential rules:

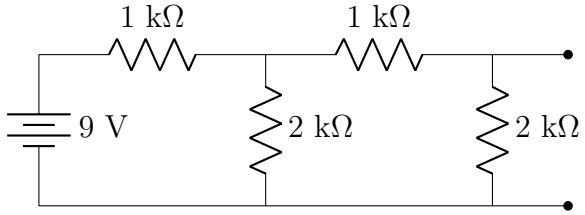
1. V_{out} does whatever needed to ensure $v_+ = v_-$.
2. v_+ and v_- draw no current.

4.2 Simple Opamp Circuits

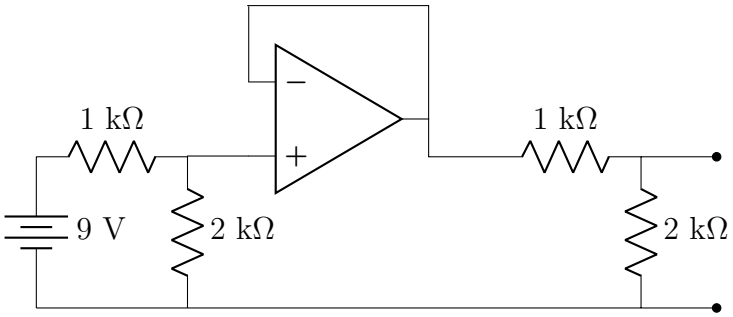
Implicit in rule two is that an opamp has immensely high resistance. That means that the input resistance seen by earlier circuit components is effectively an open circuit ($R = \infty$) for an ideal opamp. The simplest opamp circuit is seen as follows[11]:



This is called a *follower*, or *buffer*. In such diagrams, usually $-V$ and $+V$ are omitted. Based on our golden rules, v_- should match v_+, and our ciruit shows $v_+ = V_{in}$, so therefore V_{out} will be increased by the opamp until it reaches V_{in}, because v_- is tied to V_{out}. The point being, the circuit followers as $V_{in} = V_{out}$. You might say "*so it's just a wire then?*" In many ways, you are correct! But here is a nice application that shows why a follower would be useful:



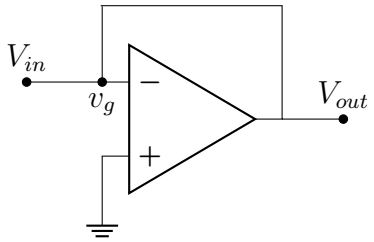
Recall this circuit from before. The important principle we learned was that because the R_{th} of the first divider was similar to the R_{eq} of the second, you could not treat this as $V_{in} \times \frac{2}{3} \times \frac{2}{3}$. However, if we do the following:



[10] These two specifications are sometimes called the 0^{th} rule.
[11] Get it?

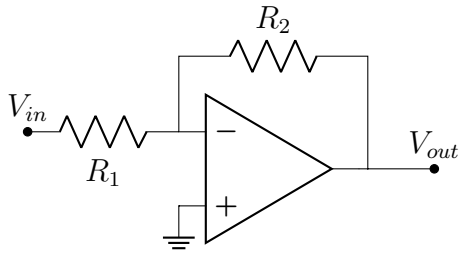
Now, the first voltage divider will "see" an input resistance of the opamp as being infinite. The opamp's output will be the voltage of the first voltage divider, and apply it to the second voltage divider. As such, we do get $V_{in} \times \frac{2}{3} \times \frac{2}{3}$.

v_- is commonly called the *inverting input*, because one can invert with it. An inverting follower would be seen as follows:



Intuitively, since v_+ is grounded (or, $V = 0$V), to make $v_+ = v_-$, v_- must equal $-V_{in}$, and or, $V_{out} = -V_{in}$. Because of this, this node is sometimes called a *virtual ground*, or v_g.

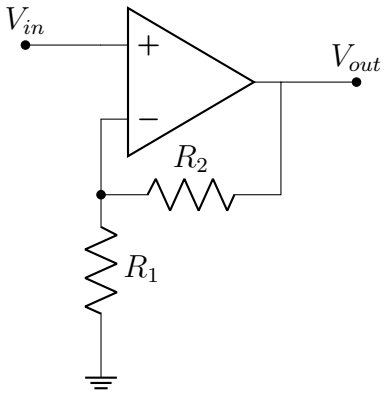
Given some of these ideas, we can imagine some special opamp uses. Like:



This is called an *inverting amplifier*. Sometimes you'll see circuits like this overcomplicated in explanation. In reality, you should just recognize that the current across R_1 must equal the current across R_2, since no current flows into v_-. Automatically, this means since $V_{in} = R_1 I$, and $V_{out} = -R_2 I$ (due to direction of flow), you'll find that:

$$V_{out} = -\frac{R_2}{R_1} V_{in} \tag{4.1}$$

A slightly more complex application is:



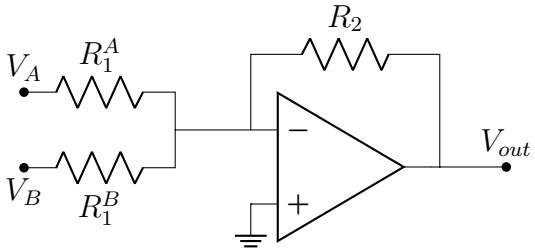
This is called a *non-inverting amplifier*. In this case, you can see that R_1 and R_2 form something of a voltage divider. That is, $V_- = R_1/(R_2 + R_1) \times V_{out}$. Knowing that an opamp always wants to ensure $v_- = v_+$, we can say:

$$\begin{aligned} V_{in} &= \frac{R_1}{R_2 + R_1} V_{out} \\ V_{in}\left(\frac{R_2}{R_1} + 1\right) &= V_{out} \end{aligned} \tag{4.2}$$

The output resistance of an ideal opamp should be 0. That is, ideally an opamp would be able to drive any voltage irrespective of the current. The input resistance of the inverting amplifier will be whatever value we have for R_1, while the input resistance for the non-inverting amplifier is ∞, as it is directly connected to the opamp's v_+.

Another thing you may wonder is if we can drive an infinite voltage at V_{out}. Recall that we are limited by whatever voltage is provided to the opamp's power rails, shown at the start of this section.

An expansion of the inverting amplifier is the *summing amplifier*:

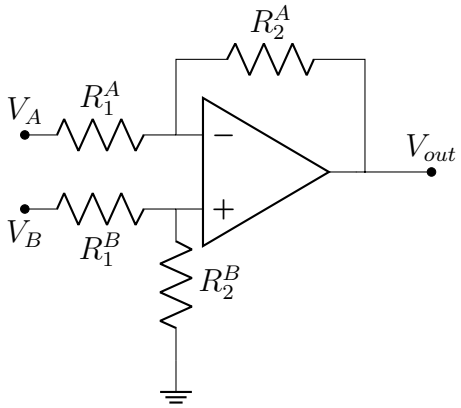


The interpretation is largely the same. Given that all of the currents through individual resistors will sum together through R_2:

$$\begin{aligned} V_{out} &= -R_2\left(\frac{1}{R_1^A}V_A + \frac{1}{R_1^B}V_B\right) \\ V_{out} &= -\frac{R_2}{R_1^A}V_A - \frac{R_2}{R_1^B}V_B \end{aligned} \tag{4.3}$$

Perhaps one application would be if you were interested in measuring the overall activity within the brain, and you are not so interested in discriminating individual electrodes in an EEG. Because the voltage changes measured by an EEG are so small, you'll need to amplify it. Thus, you throw them all together into a summing amplifier.

Next is a *differential amplifier* (not to be confused with derivative).

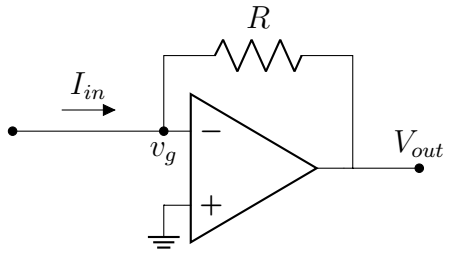


Knowing that $v_- = v_+$, and that $v_- = V_A - I_A R_1^A$ and $v_- = V_{out} + I_A R_2^A$:

$$\begin{aligned}
\frac{V_A - v_-}{R_1^A} &= \frac{v_- - V_{out}}{R_2^A} \\
R_2^A(V_A - v_-) &= R_1^A(v_- - V_{out}) \\
R_2^A V_A + R_1^A V_{out} &= R_1^A v_- + R_2^A v_- \\
\frac{R_2^A}{R_1^A + R_2^A} V_A + \frac{R_1^A}{R_1^A + R_2^A} V_{out} &= v_- \\
v_+ = v_- &= V_B \left(\frac{R_2^B}{R_1^B + R_2^B} \right) \\
\frac{R_2^A}{R_1^A + R_2^A} V_A + \frac{R_1^A}{R_1^A + R_2^A} V_{out} &= V_B \left(\frac{R_2^B}{R_1^B + R_2^B} \right) \\
R_2^A V_A + R_1^A V_{out} &= R_2^B V_B \\
R_1^A V_{out} &= R_2^B V_B - R_2^A V_A \\
V_{out} &= \frac{R_2^B}{R_1^A} V_B - \frac{R_2^A}{R_1^A} V_A
\end{aligned} \tag{4.4}$$

So that if $R_2^A = R_2^B$, you get $V_{out} = \frac{R_2}{R_1}(V_B - V_A)$. An important application of this is most physiological signals. Physiological signals are accompanied by an incredible amount of background noise, so it is useful to subtract an electrode placed on your location of interest with some other reference electrode. For example, if someone sustained damage to their left bicep, and doctors were using EMG recordings to measure recovery, an electrode should be placed on the left bicep and the right bicep. While the left bicep flexes, the right bicep can remain relaxed, and one can subtract the two signals to get the signal specific to the left bicep's flex. Because we are reading muscle depolarization through the skin, the measurable ΔV will be extremely small, hence the need to amplify it.

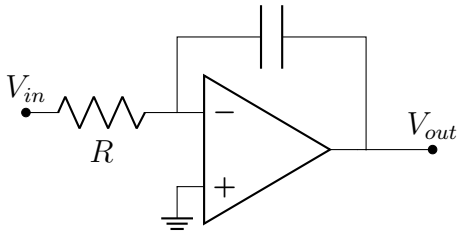
Next we will show the *current-to-voltage amplifier*. V_{out} will simply be $I_{in} \times R$:

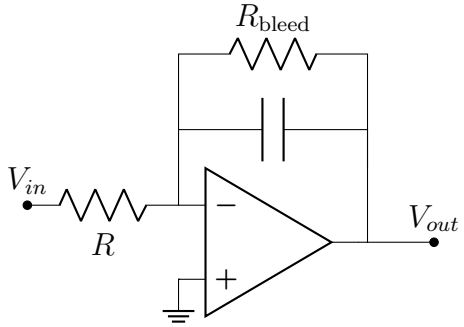


Using an opamp is particularly beneficial, in this case, because if your current source is weak, you will not further inhibit it by adding a resistor in its path. Whatever current source you use will essentially drive a short circuit, due to the opamp's virtual ground (v_g)—and current sources love short circuits.

You may ask why you would ever need to convert current to voltage? Or, when would you be able to read current as opposed to voltage? An example from Prof. Ashmanskas is in PET scanners—photons emitted from fluorodeoxyglucose (18F) hit the walls of the PET scanner, and excite atoms enough to release electrons, generating an extremely small current. Really all piezo-electric compounds can be measured in this way. You can imagine one mode of stimulation would be implanting some piezo-electric compound and inducing electron release through low-intensity ultrasound. In testing, you'll need to measure such electron release using a current-to-voltage amplifier.

The next circuits are both integrators.



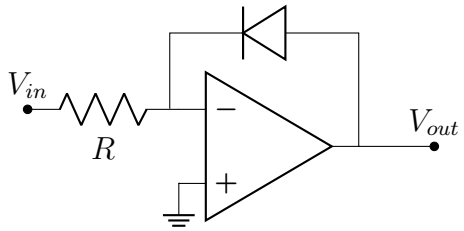


Both function as integrators (and low-pass filters), much like the low-pass filter discussion from long ago. That is, the current through the resistor is the same as through the capacitor, allowing us to use $I = CVdt$ to solve. Usefully, though, there is not a voltage limit as there is with the low-pass filter of the earlier parts. If you recall our opamp Golden Rules from section **(3.2)** you'll recall that one of the requirements is that there will be feedback. You may then realize that if V_{in}'s average over time is non-zero, there cannot be proper feedback in the left circuit, as the capacitor will block DC. Therefore, the second circuit uses a bleed resistor to allow for feedback when the capacitor is fully charged. This can be a bit dubious, as it is not expressly clear how this will affect your integration, but you can imagine that if your $R_{\text{bleed}}C$ value is large, it will drain charge from the capacitor over some long time scale. Intuitively, as the frequency applied to V_{in} increases, the amplification converges to R_{bleed}/R.

Another option is to use a switch to replace R_{bleed}. Therefore, whenever you want to begin integration you can flip the switch to zero the charge on the capacitor. Of course, the capacitor will saturate eventually, but you can "reset" it as desired.

As for applications, one interesting thing you can imagine is using an analog circuit to determine total muscle activity over time. Rather than digitally recording thousands of data points using an EMG and averaging them, you can use an integrator and record one voltage value at the end of your flexion—thus, giving some idea of total muscle output on a relative scale.

The last of this section will be the *logarithmic amplifier*:



I'll not explain the math so much, but you can think about the LED example from the very first section. A diode's current curve is exponential in nature, so the voltage will approach asymptotically toward the diode's diode-drop[12].

Biological applications of this are more obscure, but perhaps you can imagine a scenario where you'd really only care about the lower bounds of some stimulus. For example, if measuring the sub-threshold oscillations of a neuron, you may not want a linear curve—because once the neuron reaches it's threshold, the voltage rises dramatically. However, a logarithmic graph of the voltages between hyperpolarization and threshold might be quite useful, and you can configure it so that your circuit approaches its asymptote as the neuron approaches its threshold voltage.

In Summary.

Some of the circuits here are pretty neat. We learned some ways in which analog circuits can perform relatively complex functions, like multiplication, subtraction, addition, integrals and even logs.

4.3 Opamp Imperfections

It is important to understand the slight opamp imperfections that cause them to deviate from being a perfect voltage source, etc. The first two clear deviations from ideality are that the opamp's *input resistance* is not truly infinite, but is something on the order of $10^{10}\Omega$. Secondly, an opamp does not have infinite gain, A, that it can use to drive V_{out}. I.e., $V_{out} = A(v_+ - v_-)$ has some limit in that $A \neq \infty$.

Another difference is a slight offset that will always exist between v_+ and v_-, called the *offset voltage*. In modern opamps, this number might be immeasurably small, but in the canonical LM741 used in most electronics classrooms, it is close to 1mV. This is simply due to difficulty in creating two distinct nodes of exactly the same potential.

As mentioned, the opamp's input resistance cannot be infinite, meaning it will draw some small but non-zero current. This current is called the *bias current*, or I_{bias}, or I_{b+} and I_{b-} for individual inputs. I_{b+} and I_{b-} are slightly different, so I_{bias} typically refers to $1/2(I_{b+} + I_{b-})$. There is also a current due to the offset voltage, which is typically an order of magnitude or two below I_{bias}. I_{bias} for the LM741 is $\approx$ 100nA. The LM741's internal make-up is largely bipolar junction transistors (BJT), while higher-end opamps will use field effect transistors (FET), which allow for many orders of magnitude higher input resistance.

The LM741 is only capable of driving a small current of about 25mA. This is actually a very considerable point to keep in mind. Of course, this means an opamp cannot directly power any meaningfully

[12]I still haven't done a Part on diodes yet, so the term "diode-drop" is not yet explored.

high power devices without the help of some external transistors.

Notably, all transistors have some finite capacitance, so it implies that the inner workings of an opamp may demonstrate some of the same phase shift or filtration we saw in the capacitors section. This is indeed true, and is worth considering at exceptionally high frequencies. Fascinatingly, at such exceptionally high frequencies, where the phase shift extends beyond 180°, the negative feedback switches to positive feedback. This effectively switches your opamp to comparator mode as V_{out} oscillates from its V_+ to V_-. Intentionally, opamp designers include a low-pass filter within it to avoid this, reducing the gain, A, before you can reach 180°. This is called *frequency compensation*, and again is an intentional feature. This is potentially quite interesting, as the small, immensely fast oscillations at physiological levels could theoretically exceed the opamp's 180°. So, it is curious to consider if such situations would cause positive feedback.

Opamps also have a *slew rate* which defines some max $V_{out}dt$. That is, if the input is a square wave, the V_{out} cannot instantaneously follow it, and there will be some slight slope. For the LM741 it is $\approx 0.5\text{V}/\mu\text{s}$. The result is related to the internal capacitance and its charge time, making the effect non-linear and giving it some amplitude dependence. Driving a high amplitude and high frequency can therefore introduce considerable distortion.

NOTE: It would maybe be useful to make comments about the *output resistance*. Though, it is not immensely interesting.

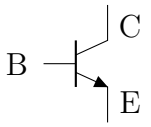
Chapter 5

Transistors

5.0.1 Transistor Overview

The prime purpose of a transistor is to use a small current to control a large one. A straightfoward example being using the turning of a key to start a car. Naturally, you wouldn't want a humongous current, capable of powering your car, to be connected to the keyhole—so you use this small current to control the flow of a larger one.

Bipolar junction transistors (BJTs) are 3-terminal devices either composed as NPN or PNP. The three terminals are the emitter (E), base (B), and collector (C). The large current you are interested in controlling flows from the collector to the emitter, and is operable by flow at the base. Though, the current at E will indeed be $I_C + I_B$, but you'd expect I_B to be many orders of magnitude below I_C. The relationship between I_C and I_B is called β, where $I_C = \beta I_B$.

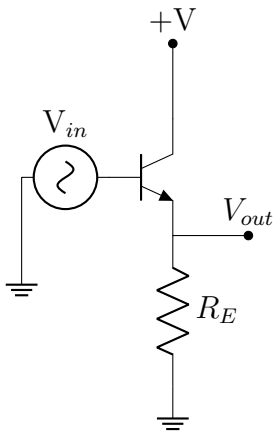


For an NPN transistor, it can either be off, where $I_B = 0$ (in this case, the voltage at the base will be below a diode drop), it can be active, where $I_B \neq 0$ and $I_C \leq \max$, or it can be saturated, where $I_B \neq 0$ and $I_C = \max$. You can probably imagine: active mode is useful if you'd like to make a transistor-based amplifier, while the saturation mode might be more useful in making a switch.

5.1 Transistor Circuits

5.1.1 A Follower

A follower, as with the opamp follower, is useful in building a high-input impendance, low output impedance element in your circuit. Again, this is useful as it will draw minimal current and preserve the voltage at that point in the circuit.



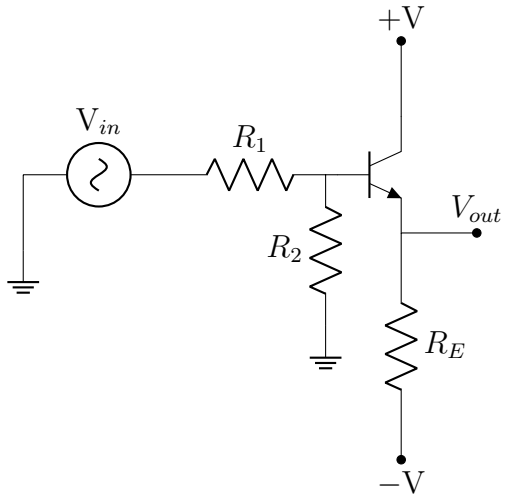
Alrighty, so in active mode, if this functions properly as a follower, we would expect that as V_{in} changes by ΔV, so too would V_{out}. Too, it will function very much like a diode—meaning V_{out} will feature a diode drop and cannot drop below 0. In later circuits, we'll show how to fix this by tying E to some negative voltage, −V. Thus, we get that: $\Delta I_B = \Delta I_E/(\beta + 1) = \Delta V/R_E(\beta + 1)$. Since $R_{in} = dV_{in}/dI_{in}$, we get:

$$R_{in} = (\beta + 1)R_E \tag{5.1}$$

Therefore, from the perspective of V_{in}, the load, which would be attached at V_{out} now appears to be much higher. More explicitly, it appears as:

$$R_{in} = (\beta + 1)(R_E||R_{load}) \tag{5.2}$$

Again, this is preferable for the source. As we explored in the very early voltage divider section, if you were to have the following circuit:



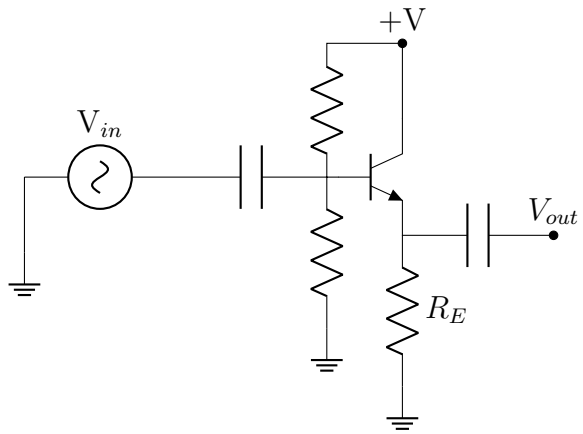
Now we have $R_2||(\beta+1)R_E$, which we would expect to be $\approx R_2$. Therefore, the voltage divider acts as it would if there were no load there at all. The output impedance will be that which exists at V_{out}. Since the current at ΔV_{out} varies with ΔV_{in}, and ΔI_{out} varies with $(\beta + 1)\Delta I_B$. So, $\Delta V_{out} = \Delta I_B(R_1||R_2)$ or $\Delta V_{out} = \Delta I_B R_{th}$. Since $R_{out} = dV_{out}/dI_{out}$, we get:

$$R_{out} = \frac{R_{th}}{\beta + 1} \tag{5.3}$$

Which we expect to be quite small. Therefore, the output resistance is quite small. Since technically R_{load} is again in parallel with R_E, it should actually be:

$$R_{out} = \frac{R_{th}}{\beta + 1} || R_E \tag{5.4}$$

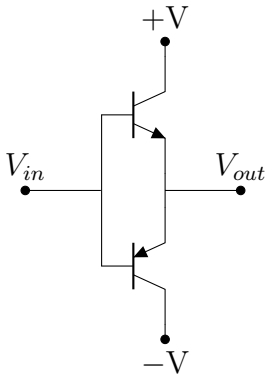
Which again we expect to be $\approx R_{th}/\beta + 1$, and or, a very small value. You could build a more complete BJT follower in this way:



I'll not walk through the entire idea, using explicit values, but you should be able to get the gist. The capacitors are used to remove the DC. The +V adds a DC-offset, so that the oscillations of V_{in} can be all positive, removing the restriction of keeping V_E above 0. The two resistors are used as a voltage divider, stemming from +V, allowing us to pick our DC offset. One of the largest takeaways from this is simply that it is much more complex, and annoying, than opamp-based followers. However, it should give insight into what might be within an opamp.

Push-Pull Follower.

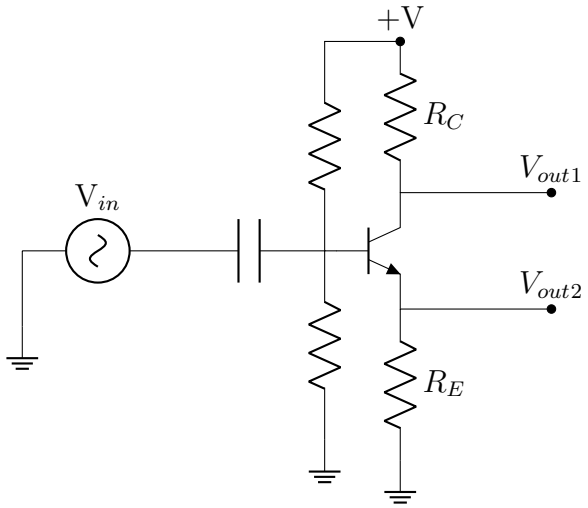
A neater way to avoid some of the drawbacks to the preceeding designs is to use what is called a *push-pull follower*. It takes advantage of using two BJTs. Where the first NPN handles the positive half, and the second, now a PNP, handles the negative half. It is seen as below:



The problem, of course, is that you will still have two diode drops in either direction. This is called *crossover distortion*. An intelligent way to avoid this crossover distortion would be to add some additional diodes around the base of the two transistors connected to a power supply. An easier way, virtually by cheating, would be to use an opamp whose feedback loop connects to the output of the push-pull circuit, allowing the opamp to "pre-compensate" for the two diode drops.

5.1.2 Amplifiers

The first iteration is the common emitter amplifier, seen as follows:



Let's again analyze solely by glancing at the circuit. We know that the voltage at V_E will be one diode drop (say, 0.6V) below V_B in the active mode. We know that the current through R_E must therefore be $(V_B - 0.6)/R_E$. We know that $I_C \approx I_E$, therefore $I_C = (V_B - 0.6)/R_E$. If $V_B = V_{in}$, this must mean that the voltage at C:

$$V_{out1} = +\mathrm{V} - (V_{in} - 0.6)\frac{R_C}{R_E} \tag{5.5}$$

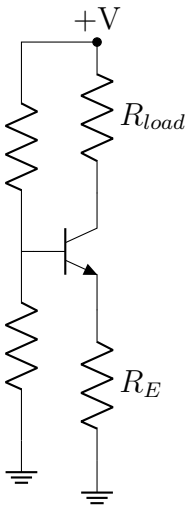
Using a capacitor to remove the DC offset will further make it function like the amplifier you desire. Something small to keep in mind is the limits of R_E. Let's suppose you were to take $R_E \to 0$. The truth of your equation is this:

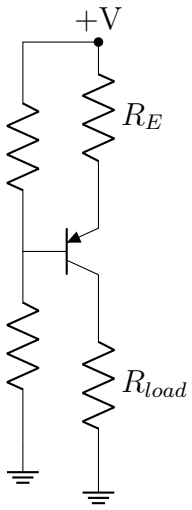
$$\begin{aligned} V_{out1} &= V_{out2}\frac{R_C}{R_E + r_e} \\ V_{out1} &= V_{out2}\frac{R_C}{r_e} \end{aligned} \tag{5.6}$$

Where r_e is the resistance of the diode, set approximately by $25\mathrm{mV}/I_C$. Therefore, your output will experience "barn roof distortion" due to the logarithmic nature of a diode's IV curve.

5.1.3 Current Sources

Let's take this circuit:





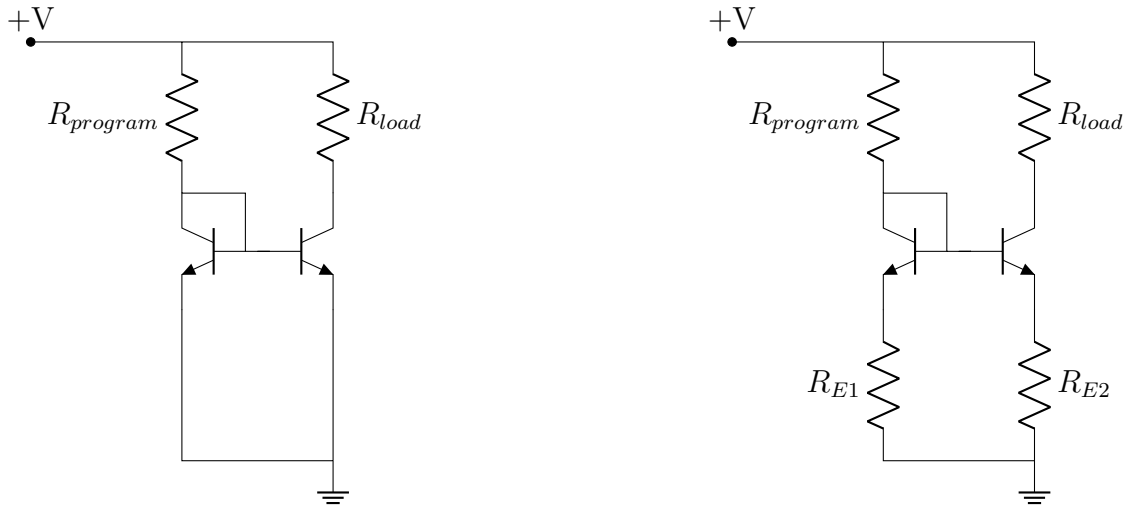
Recall that the current through the collector is set by the emitter. The voltage at the emitter is set by the base. V_E will be one diode drop below the base (for the leftward circuit, that is—for the rightward, it will be one diode drop above). So if the resistance at the emitter, R_E, is 1 kΩ, then we would want V_E to be 2 V, and V_B to be 2.6 V. Note that the transistors input resistance is $\beta(r_e+R_E) \approx 100k\Omega$.

Keeping a small R_E is important here, as the larger that R_E gets, the closer you will be to saturating your R_{load}. Let's try to consider an example. Covered much, much later in this book, we discuss a Courtine lab paper that used electronic dura mater. The load value appeared to vary all the way from 5kΩ to 100kΩ. If you wanted to supply a constant 2mA over this entire impedance range, you would need the following:

$$\begin{aligned} V_B &= 2.6\text{V} \\ R_E &= 1\text{k}\Omega \\ I_C &= 2\text{mA} \\ V_C &= +\text{V} - R_{load}I_C \\ V_C &= +\text{V} - (100k\Omega \times 1\text{mA}) \end{aligned} \tag{5.7}$$

It looks like you'll need a 100V supply to keep your transistor from saturating. That is... not good!

Current Mirror.



On the left circuit, your $R_{program}$ sets the current via the fact that the voltage drop between +V and ground happens only across $R_{program}$, so the current flowing through the emitter is $+V/R_{program}$. The same current will be pulled through R_{load}. Adding in the R_E on the rightward circuit gives a path to bleed and allows your programmed current to stay flat for longer. However, even in this case, R_{load} cannot exceed $R_{program}$.

So why would you want a current mirror? Here's an example from this video[13]. Amplifiers can be biased using current sources, which allows the biasing current to become very stable. This is important, because as voltage or temperature fluctuates, the current can be maintained. Adding a current source to every amplifier would require too much annoying circuit building, so one way to avoid this is using a current mirror.

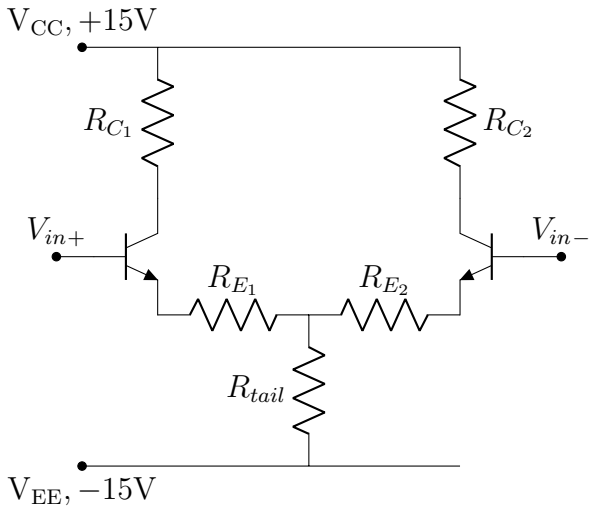
5.1.4 The Components of an OpAmp

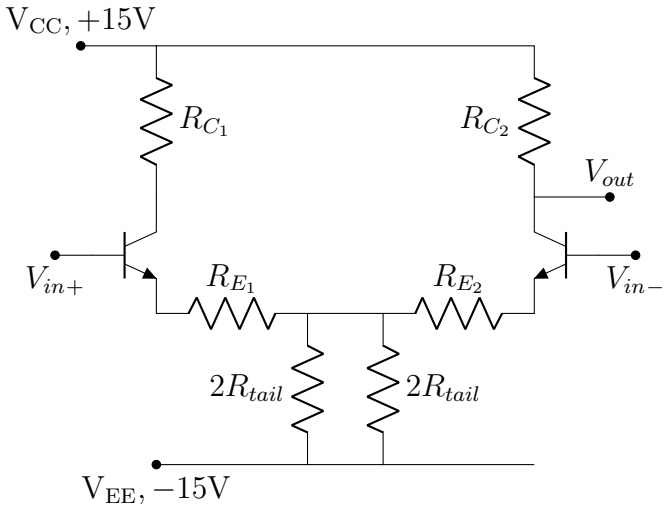
Opamps are transistor-based at heart. The key components of an opamp is that it should have a large, ideally infinite differential gain. This means that an opamp can ideally amplify any difference between V_+ and V_-. Conversely, an opamp should ideally have a 0 common mode gain. This means that when $V_+ = V_-$, $V_{out} \approx 0$. Other important components of an opamp include a high current supplied to V_{out}, accomplished by an ideally 0 output resistance. Let's consider how we can accomplish these things.

The first step is to generate a circuit with a decent differential gain and a somewhat low common-mode gain. Let's see what we can gather about the below circuit(s).

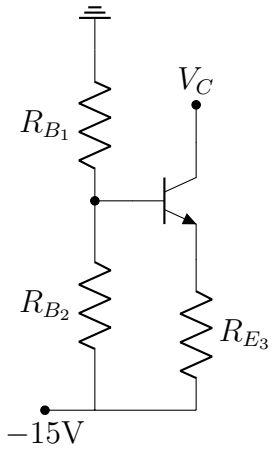
$$\begin{aligned} A_D &= \frac{R_C}{2(R_E + r_e)} \\ A_{CM} &= -\frac{R_C}{2R_{tail} + R_E + r_e} \end{aligned} \tag{5.8}$$

[13] https://www.youtube.com/watch?v=VnJHXQCPIvs&ab_channel=ALLABOUTELECTRONICS

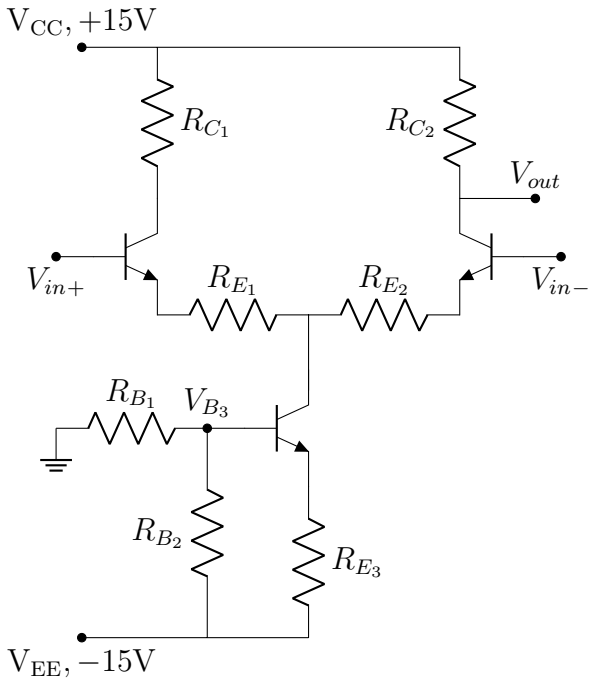




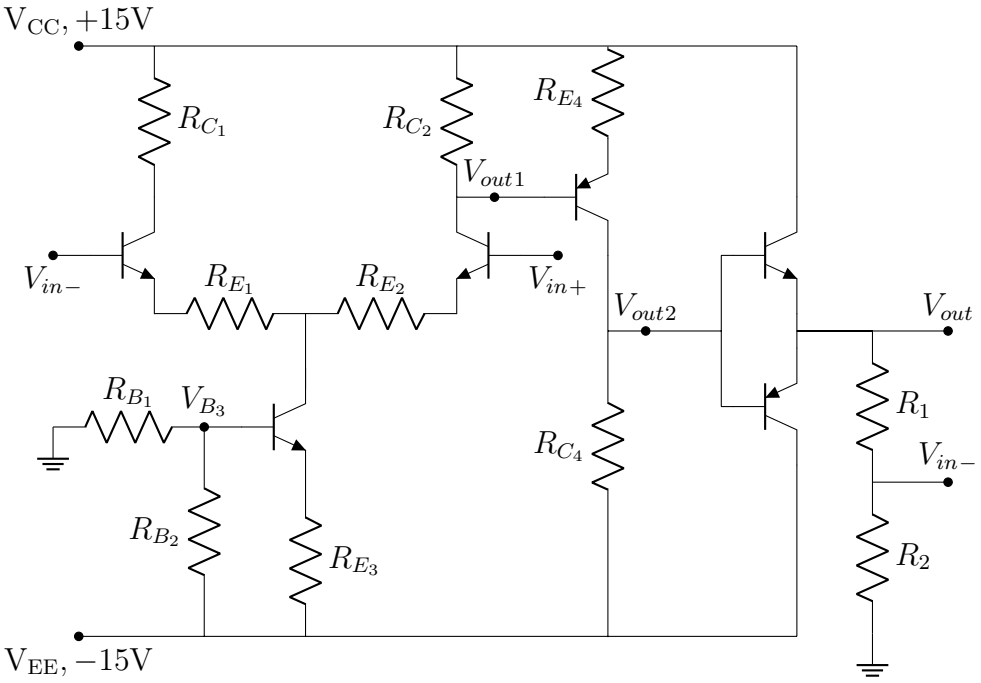
These two circuits are equivalent, but the right one is a bit easier to analyze. Considering equations **5.8**, you may want to pick an incredibly large R_{tail}, or better yet, to pick a downward circuit fragment with a very high R_{in}. Let's try:



In the quiescent state, we can easily determine the voltage at the base as $V_B = 15\text{V} \times R_{B_1}/(R_{B_1} + R_{B_2})$. Therefore, we can determine the voltage at the emitter as $V_E = V_B - 0.7\text{V}$. We can determine the voltage drop across R_{E_3} here as $15\text{V} - V_E$, and therefore easily determine I_{E_3}, and thus I_{C_3}, as $R_{E_3}/(15\text{V} - V_E)$. Let's connect this fragment to our circuit as below:



Because we know the quiescent I_{C_3}, we can know that $0.5I_{C_3}$ flows through R_{E_1} and R_{E_2}, allowing us to easily deduce what V_{out} is in the quiescent state. We might be tempted again to take the limit of $R_{E_{1,2}} \to 0$, but recall that we will introduce distortion this was as r_e becomes the only source of resistance in the system. If we assume the R_{in} into this third transistor fragment is extremely high, on the order of MΩs, then we can know that the differential gain is extremely high, while the common mode gain is extremely low. Let's add some extra circuit fragments:



We've now added a common emitter amplifier to the output of stage 1 (now called V_{out1}). It will amplify as $A = -R_{C_4}/(R_{E_4} + r_e)$. However, one consideration is to keep the voltage of V_{out1} low enough so that the transistor does not saturate. We take the output from this (V_{out2}) and connect it to a push-pull circuit—which is helpful in increasing the current supplied by the homemade opamp. However, it

introduces crossover distortion. This crossover distortion is eliminated when feedback is added, with V_{in-} connected to V_{out} as shown above. The whole circuit amplifies as: $(1 + \frac{R_1}{R_2})(V_{in+} - V_{in-}) = V_{out}$.

One slight issue you may encounter is an annoying fuzz on the output. This may be due to unintended high-requency gain. One way to avoid this is adding a very small capacitor between V_{out1} and V_{out2}. This effectively acts as a high-pass filter, removing it from amplification by the common-emitter amplifier.

Chapter 6

A Digital World

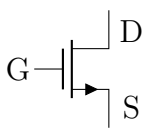
6.0.1 Introduction

What does it mean to digitize a signal? A digital signal is composed of discrete units—in electronics, that would typically be transistors. That is, transistors are either on or off, and if you wanted to "digitize" a voltage, you could use transistors, each supplying some small voltage, that sum together at V_{out}. But, why would you want to digitize a signal? Don't you lose some information by storing values in discrete increments? Perhaps, but digital signals are great for preventing signal degradation. For example, if you were to transmit an analog signal long distances, passing between different detectors and emitters, then this signal may degrade or become warped. However, if your signal is digitized to some binary value, equating to 5.0V on the dot, it won't degrade.

Since we are also biologists, a great question to ask is: is digital signals relevant in biological systems? The answer, of course, is yes. Neurons really are, for all intents and purposes, transistors. Neurons are digitizers. If you were to take a graded potential being received by mechanosensitive channels at the tip of the toe, do you think it could make it up to the brain without some degradation? How could you ensure the strength of the signal is not lost as it propagates up the spinal cord and onto the brain? Well, as you know, neurons are either (simplifying a bit here) on or off, meaning the "strength" is lost on an individual action potential, so therefore strength must be encoded in some other way—frequency. Higher frequency firing, more neurotransmitter release, and a stronger response in the brain—neural circuits are analog to digital converters in the body.

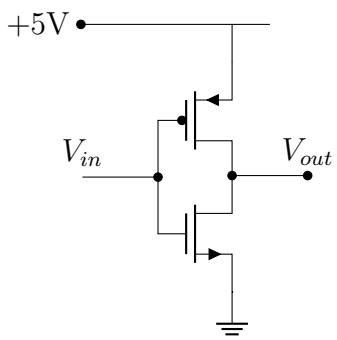
So, in electronics, how would you digitize a signal? A quite simple way, conceptually, would be to use a comparator, where a constant voltage (say, 1V) is connected to v_-. Therefore, when v_+ is above 1V, the circuit is `ON`, otherwise, it is `OFF`. You could imagine setting up a series of successive voltage dividers, each reducing the input voltage and connected to the v_+ of a series of comparators, which could be used to digitize a strength of signal by V_{in}. However, as mentioned, the true way a signal is digitized is by transistors, discussed next.

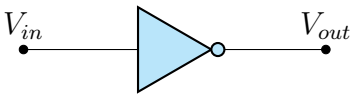
6.1 Digital Logic, and FETs



Modern electronics use field effect transistors (FETs), as opposed to BJT transistors. FETs have the

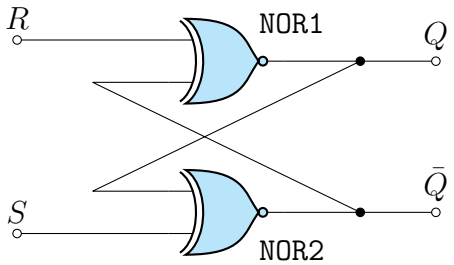
equivalent prongs, except that they're labeled gate, drain, and source as opposed to base, collector, and emitter. FETs, in large part, can be thought of as significantly more powerful BJTs. That is, while BJTs have some small current traveling from base to emitter, FETs have (essentially) zero current from gate to source. Rather than having something like a diode drop, FETs have a minimum voltage (a $V_{threshold}$) where they become active. Thus, if you were to make a push-pull out of FETs, you'd instead see a "diode drop" equivalent to the $V_{threshold}$—which is usually much higher than 0.7V, thereby meaning the crossover distortion observed will be worse. Traditionally, this $V_{threshold}$ is around 2.5V, but it varies from part to part. Therefore, the `ON` state (`HIGH`) is usually defined as when $V_{in} > 4.7$V creating a $V_{out} = 5$V, and `OFF` (`LOW`) as $V_{in} < 0.2$V creating a $V_{out} = 0$V. CMOS logic gates do not update instantaneously, but occur fast enough that it is considered instant. For the logic gates within one's computer, they update on the order of 10^{-1}ns. let's finally consider an example, a digital `NOT` gate is shown here:





When V_{in} is greater than $V_{threshold}$, the pMOS transistor (top one) will be off, the nMOS transistor (bottom one) will be on, and the circuit will be grounded. Conversely, when it is below $V_{threshold}$, the pMOS will be on, the nMOS will be off, and the circuit will be 5V.

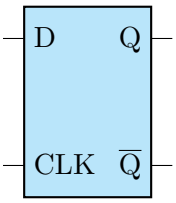
I'll spare you (and myself) the trouble of creating diagrams for tons of other types of gates, but you can surely imagine most complex configurations yielding `NAND`, `AND`, `NOR`, etc. However, I will like to highlight **flip flops**. Compounding together different logic gates can be incredibly powerful, and is sufficient to build complicated devices like adders. However, such devices only depend on the current inputs—**combinotorical logic**. To make a device which depends on previous states requires some form of memory—**sequential logic**. Let us consider the simplest form of memory, a 1-bit memory:



This is called a simple flip-flop and or an S-R latch. S and R stand for set and reset. Suppose $S = 0, R = 1$; as the first input of `NOR1` is 1, Q must equal 0. Therefore, the first input of `NOR2` is 0, so $\bar{Q} = 1$. If R is turned off, then the first input of `NOR1` goes to 0. However, $\bar{Q}$ is still 1, therefore Q remains 0. Now if S is turned on, then $\bar{Q}$ goes to 0, and Q goes to 1. And the logic continues. In short, turning $S = 1, R = 0$, you set $Q = 1$; by turning $S = 0, R = 1$, you reset $Q = 0$. It isn't immediately obvious why this is helpful, but you can imagine, as we will discuss later, that connecting

this sort of fragment to a clock (`clk`) can allow you to count up, and perform similar update-based operations.

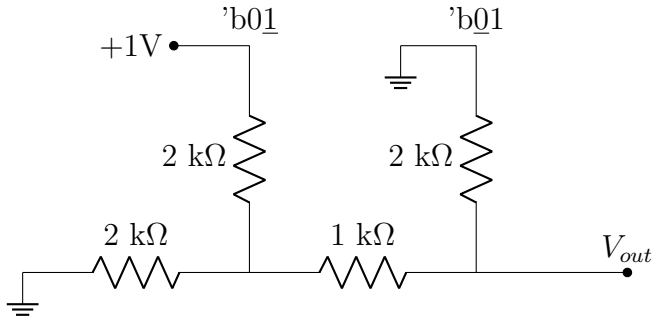
One of the most ubiquitous expansions of this circuit is called a **D-type flip flop** (DFF). This is composed primarily of `NAND` gates, and reacts to a clock. Instead of looking at the logic gate's lets just consider the top down view:



Essentially, Q updates to D at each `clk` cycle. Yada-yada, you can perhaps see how we can expand from here.

6.1.1 Digital to Analog Conversion

So, we know how to go from analog signals to digital signals by using some kind of comparator or transistor based circuit. But, how can we convert from digital to analog? One way is through an $R-2R$ ladder. Let us consider the simplest, 2-digit binary number below:



It's trivial to analyze this circuit, but let's still consider it anyway. The setup above corresponds to the binary number 01—the first digit is powered, the second is not. The voltage read at V_{out} can be solved by considering the second voltage divider as the load, thus:

$$\begin{aligned} R_{th} &= 2k\Omega||2k\Omega = 1k\Omega \\ I_{sc} &= 1\text{V}/2k\Omega = 0.5\text{mA} \\ V_{th} &= R_{th}I_{sc} = 250\text{mV} \end{aligned} \tag{6.1}$$

Therefore, we have taken this 01 number and transformed it into 250mV. Let us consider the case when the second digit is 1V, and the first is grounded. We can trivially see this as a voltage divider between $2k\Omega$ and $1k\Omega + 2k\Omega||2k\Omega$, meaning $V_{out} = 500\text{mV}$. Finally, let us consider 11. We know, by superposition, that V_{out} must be 750mV. Therefore, we have converted individual binary numbers to be 250mV—essentially putting it on a discrete, but continuous scale. This will work for any sized $R-2R$ ladder, and the formula is as follows:

$$\Delta\text{V} \times (\text{digit0} \cdot 2^0 + \text{digit1} \cdot 2^1 + \text{digit2} \cdot 2^2 + \text{digit3} \cdot 2^3 ...) \tag{6.2}$$

Where, in our above example we saw 'b01 $= (1 \cdot 2^0 + 0 \cdot 2^1) \times 250\text{mV}$, 'b10 $= (0 \cdot 2^0 + 1 \cdot 2^1) \times 250\text{mV}$, 'b11 $= (1 \cdot 2^0 + 1 \cdot 2^1) \times 250\text{mV}$.

Chapter 7

Writing Hardware

7.0.1 Introduction

Verilog[14] is a language used to describe electronics, and allows you to avoid the physical action of wiring. This is the reason for the designation **writing hardware**. In this way, you can pick your poison: debugging code, or debugging breadboards. Importantly, though, Verilog is capable of computation and writing data files that go beyond circuit descriptions. Thus, it is not a "markdown" language and is Turing Complete[15].

Tools like Field Programmable Gate Arrays (FPGAs) allow for this, as their internal composition is something of an array of transistors, which can be rewired through code in order to meet the demands of the programmer.

Creating Modules.

In Verilog, a circuit is called a `module`. Each module is defined between a `module` and `endmodule`, which can be named as shown in the example below. Different ports connect the module to things outside of the module.

```
1  `default_nettype none // this requires that all wires be explicitly named,
       which is helpful in preventing the compiler for erroneously assuming
       attachments if a name or module is not recognized
2
3  module example1(o, i1, i2);
4  // example1 is the name of our module, and o, i1, and i2 are our ports
5  // it is convention to list outputs first
6
7  output o; // this defines o as an output
8  input i1, i2; // this defines i1 and i2 as inputs
9
10 endmodule
```

Gates are also initialized like modules. The way to do this is with the built in primitives for AND and OR gates (`and` and `or` respectively). For example:

[14] A large part of the background information and general syntax comes from the YouTube channel: CompArchIllinois.
[15] Or at least, I think it is. I can never remember the exact definition of Turing Complete :)

```
module example2(o, i1, i2);

output o;
input i1, i2;
wire wire1, wire2; // this initializes two wires called wire1 and wire2

or or1(wire1, i1, i2);
// this makes an OR gate named or1 with inputs i1 and i2, and output called
    wire1
and and1(o, wire1, wire2);
// this makes a NOT gate named not1 with input i2, and output called wire2
// one of the outputs of the OR gate feeds into the AND gate (via wire1) in
    this example

endmodule
```

The order in which things are initialized do not matter. It is very important to not reuse wire or other variable names, as Verilog will read these as being connected irrespective of where they are intended to be. As mentioned, the code above uses modules built into Verilog, but you could make your own module in the following way:

```
module andgate(output o1, input i1, input i2);
    assign o1 = i1 & i2;
    // for OR you would use |, and for XOR you would use ^
endmodule;
```

Bus Notation.

Bus notation is used to simplify the pins used (in Verilog, this is called a vector). For example, a multiplexer or an adder will have many inputs, which would be inconvenient to initialize individually. Instead, we can use something like this:

```
module adder(c, a, b); // a, b, c are 3 bus inputs we will use
    output[3:0] c; // initializes 4 wires within our c bus
    input[3:0] a, b; // initializes 4 wires within our a and b buses
endmodule
```

Firstly, note that Verilog is 0 indexed, so `[3:0]` includes 4 wires. In a circuit schematic, busses are drawn as thicker wires with a slash through them and a number denoting the amount of wires in the bus. If we wanted to call individual wires from our busses into the `andgate` module we declared earlier, we could do it as:

```
andgate(c[0], a[0], b[0]); // this is fine if we are using a single AND
   gate
andgate and1 (c[0], a[0], b[0]); // this can be used to individually label
   different AND gates using the same module
```

And we can connect busses together, or wires together, using the assign command like before. For example:

```
wire wire3;
assign wire3 = c[2];
assign c[2:0] = a[2:0]; // wire3 will now be connected to bus a[2] through
   bus c[2]
```

Constants and Variables.

You can actually explicitly ascribe a bit value to a series of wires as seen below with `wire4`:

```
wire [2:0] wire4 = 3; // wire4 now holds 011
```

Verilog allows us to define constants using 3 parameters, defined as their size, method of encoding, and value. For example, 8'hd7 corresponds to a size of 8 bits, hexadecimal encoding, and the value d7 (equivalently, 11010111). You can use this in `boolean` comparisons, as below:

```
wire wire5;
'define CONST1 3'b011; // CONST1 is the name of the constant
wire5 = (a[2:0] == CONST1);
// This is a tad complicated. The gist is: all of the bit values in a[2:0]
   will be compared to CONST1 in a NXOR style statement. That will then be
   compared to all of the other bits in an AND style statement. So wire4
   will be on only if all wires in correspond to CONST1.
// The C++ equivalent would be something like:
//    (a[2] == 1'b0) && (a[1] == 1'b1) && (a[0] == 1'b1)
// Also, this could be completely wrong. I can't check any of this without
   an actual FPGA in front of me! So who knows!
```

This does bring up a worthwhile point, which is that everything you write in Verilog has a direct circuit component underlying it (I suppose the same is true for any program, but it's more... explicit with Verilog).

Logic.

Verilog has two types of logic it can perform—bitwise and logical. The expressions are: The operators are $\sim$ for bit-wise NOT, ! for logical NOT; & for bit-wise AND, && for logical AND; | for bit-wise OR, || for logical OR; and ^ (bit-wise XOR). As you'd imagine, a bitwise comparison goes bit by bit. For example:

```
wire [0:3] a = 0;
wire [0:3] b = ~a;
wire [0:3] c = !a;
```

Here, `a` = 0000, `b` = 1111, and `c` = 0001. While:

```
wire [0:3] a = 1;
wire [0:3] b = ~a;
wire [0:3] c = !a;
```

Now, a = 0001, b = 1110, and c = 0000.

Adder.

Verilog has the ability to make adders itself. For example (from Prof. Ashmanskas):

```
module add4bit (output [3:0] s, output cout,
    input [3:0] a, input [3:0] b, input cin);
    assign {cout,s} = a + b + cin;
endmodule
```

Where each of the inputs (a, b) and outputs (s) is 4 bits wide. Writing {cout,s} is then a 5 bit wide statement.

Maintaining a State.

This should go in the digital section, but I am throwing it into here for now.

If one were to implement *sequential logic*, we would need a source of memory, or a way to maintain a state. The simplest way to accomplish this is with two NOR gates, which functionally act as a 1 bit memory source. By feeding the output of each NOR gate into the input of the other, we make it so one can set the memory to 1, and the other can reset the memory to 0. The output of the reset we can call Q, the current state of the system.

Anyway, a D-type flip-flop similarly gives us the ability to story memory that updates on a clk cycle. See below:

```
module dflipflop (output wire q, input wire clk, input wire d);
    reg q_reg;
    always @ (posedge clk) q_reg <= d;
    assign q = q_reg;
endmodule
```

Here, q_reg updates to the state of d at the positive edge (posedge) of every clk cycle. reg is like a wire which can maintain a state (or rather it is like a flip-flop). We can add an additional layer of complexity with an enable feature, which allows q to update only when enable is on. For example, with the updated:

```
    always @ (posedge clk) if (enable) q_reg <= d;
```

This sort of thing is useful, for example, as a counter. You can imagine that d-type flip flops could be used to store the state and update on a clock cycle only if the previous bits are all on, except for the first bit, which is always enabled.

7.1 Building to a Computer

7.1.1 Multiplexers

A ternary operator is a step beyond regular logic. It allows us to switch between two values, depending on if some other value is higher or low. For example:

```
module multiplexer (output O, input I, input A, input B);
    assign O = (S ? A : B)
    // alternatively, we could write:
    // O[3:0] = (I ? A[3:0] : B[3:0])
    // for multi-digit outputs
endmodule
```

`O` here evaluates to `A` or `B` depending on if `S` (select) is 1 or 0. While we're here, I'll also mention that we can string together this style of conditional statement in the following way:

```
// suppose there is some input, A, and some set of values, first, second,
   third ...
assign value = A == 0 ? first  :
               A == 1 ? second :
               A == 2 ? third  : fourth ;
// this can be conceptualized as a series of if, else if, else statements
```

A super (!) cool trick we can use to define a single module with multiple uses would be as follows:

```
module multiplexer #(parameter N=1)(
    output [N-1:0] O,
    input S,
    input [N-1:0] A,
    input [N-1:0] B);
    assign O[N-1:0] = (S ? A[N-1:0] : B[N-1:0])
endmodule

wire [2:0] output, a, b;
multiplexer #(3) multi1 (output, select, a, b);
```

Here, you can see that the parameter `N` is variable, thus meaning `multiplexer` can be variably defined. Very neat!

Another neat trick is that, instead of feeding in inputs and outputs by their position in the arguments, you can do it explicitly by name. This avoids accidentally wiring things together that should not be, when keeping a consistent naming convention may be difficult. It is see (following from the previous example) as:

```
wire [2:0] output, a, b;
multiplexer #(.N(3)) multi1 (.O(output), .S(select), .A(a), .B(b));
```

Here, the order that you add them is irrelevant as the extension will properly map the two, allowing you to evade conventions.

Memory.

Extending from the multiplexers in the previous section, suppose you have a set of 256 vectors, each containing 16 bits. Using an 8 bit `select` key, you can choose any of these vectors and output them accordingly. This is a **read-only memory (ROM)**.

Finite State Machine.

A finite state (FSM) can be devised through a series of D-type flip-flops. First, you can assign a set of states (called `state`) via an index stores in D-type flip-flops. For example, in a simple case where you want to oscillate a traffic light from green (`green`, to yellow (`yellow1`, to red (`red`), and perhaps a fourth state for when a button is pressed by someone on a crosswalk (`yellow2`), you can use a 2-bit wide flip-flop module.

We can use a series of logical operations to determine what the next state will be. I.e., if the light is currently `green` (say, represented by `00`), then we can know that the next state (`nextstate`), updated on every `clk` cycle, can either be `yellow1` (`01`) or `yellow2` (`10`) or remain `green`. And, we can know that both yellow states will always transition to `red` (`11`) after. Whether or not we go from `00` to `00` or `01` can depend on the duration (employing a counter), and if it will go from `00` to `00` or `10` can be determined by some tertiary module tied to an external button. For example, this may look something like:

```
wire [1:0] state;
wire [1:0] nextstate;

module dflipflop #(parameter N=1)
    (output wire q,
    input wire clk,
    input wire d);
    reg q_reg;
    always @ (posedge clk) q_reg <= d;
    assign q = q_reg;
endmodule

dflipflop #(.N(2)) current_statedff
    (.q(state),
    .d(nextstate),
    .clock(clock),
    .enable(1),
    .reset(reset));

assign nextstate =
    (state == green && button)      ? yellow2 :
    (state == green && duration)    ? yellow1 :
    (state == yellow1 && duration)  ? red     :
    (state == yellow2 && duration)  ? red     :
    (state == red && duration)      ? green   :
                                      state   ; // keeps current state
```

A Microprocessor.

7.1.2 Altogether, and Implications

Let us first take a step back and recall what we "know" so far: transistors, namely CMOS transistors, can be added together to make digital logic gates, which take on either `HIGH` or `LOW` values. Such logic gates can be stacked together to form complex decisions, like `NAND`, `AND`, `OR`, etc. Mathematical functions, like adding and multiplying, can be accomplished with these relatively simple gates. These are examples of combinotorical logic, which depend only on the current state of the system. An extra layer of complexity, in the form of memory, is required to gain sequential logic. The simplest form of memory we have discussed so far, a 1-bit storage, is a flip-flop, or SR-latch, composed of two NOR gates. D-type flip-flops add additional complexity, are composed of AND gates, and update on a clock. We can accomplish quite a lot by simply stacking together D-type flip-flops with accompanying `enable` features, such as counters.

Now, we can use a multiplexer, which switches between two outputs given an input, to build greater memory. Stacking together multiplexers gives us a read only memory (ROM). To give us a random-access memory (RAM), which is editable, we instead stack together D-type flip-flops[16]. D-type flip-flops can also be stacked together to create a finite state machine (FSM). The combination of a RAM and an FSM provides us with a microprocessor.

A CPU's microprocessor is equipped with a few different functions. Memory is accessed through the `FETCH` function, which is based upon the Program Counter. The processor can then `DECODE` what is fetched, and make a decision for what function to perform. The CPU performs many of its math functions using the accumulator. The accumulator is able to compare numbers, and what it determines elicits the `JUMP` command, which may be conditional, to access new memory locations. The CPU can output this info to the world through the `OUT` feature.

Why FSM?

There's something very interesting about finite state machines. In a way, a computer is a complicated finite state machine, neural networks are a complicated finite state machine, your brain is a complicated finite state machine.

Note from future Jackson: I have no idea what I was getting at here. Let us both pretend this comment was, in some way, insightful (it wasn't).

[16] Today's world seems to use capacitors in place of D-type flip-flops in a way that is still unknown to me.

Part II

Math and Models

All truly strong people are kind.
– Vagabond by Takehiko Inoue

7.2 Perspective

Truly, I find it endlessly fascinating and miraculous that you can quantitatively describe physical or biological phenomena. Of course, it stands to reason, but its shine is not lost on me. I vividly recall being in my Human Physiology class and having our professor pose the question: *what will happen if two action potentials run into one another*? I knew that they'd terminate, so the result was not what drew my intrigue. What was miraculous is that, to answer this question, our professor pulled up MatLab and simulated it before our eyes. Indeed, the action potentials terminated.

Prior to this, I hadn't thought much about quantitative descriptions like this. Qualitatively, and or conceptually, it made sense to me that action potentials cannot propagate past one another. But naturally, not all answers can be intuited. After scrolling around through the amazing Keener and Sneyd *Mathematical Physiology* textbook, I realized the incredible breadth of phenoma that we have modeled. From membrane electrophysiology, enzyme kinetics, to disease spreading, to fluid dynamics of the circulatory system, to the movement of bacterium, to every other little itty-bitty thing you can become fascinated by—someone has build some differential equations for it.

If anyone ever reads this, besides me, I hope you too will be enthralled by both the brilliance and creativity required to describe something as complex as the biochemistry of a neuron in just a few equations.

7.2.1 Further Insight

I thought that I might include this, because it is indeed inspiring, and I figured you may want to know. On Wednesday Nov. 15th, 2023, I attended a super interesting lecture titled "A Language Whose Characters are Triangles" by Rob Phillips from Caltech. The name is quite ambiguous, and the talk spanned a wide range of topics. The overarching theme was modeling very different physical systems using the same range of principles.

One of the key points he made is that many journals are hyper-focused on structure right now. New articles are published every month with AlphaFold / Cryo-EM structures of proteins in Nature and Science. However, he makes the joking argument that these scientists are all neglecting the quarks, which are as key to the protein's structure as the atoms are. He uses this argument to suggest we don't need an ultra-structural understanding to model these systems. That is also the rough idea of the talk's title: that nearly all systems can be solved using similar geometries, and described some examples throughout history where different phenomena in physics were solved using basic geometric shapes. Then, he moves into the key topic: comparing the herding of wildebeest and the trafficking of actin inside a toxoplasma (a microorganism).

Dr. Phillips touched on the Navier-Stokes equations, and new iterations and applications of them. He showed that wildebeest herd, both in their "flow" and "density" according to the same modified Stokes equations as do the actin fibers inside a toxoplasma. Another interesting thing he mentioned to look out for in the future is: (1) our growing understanding of vectors and (2) greater modeling of non-reciprocal physics. Point (1) is because, especially recently, computer scientists have really advanced our use of vectors. I.e., all ML is based on parameter vectors, Amazon has preference vectors for each user, Myers-Brigg is a 4 dimensional vector representation of people, etc. He said he believes an increased ability to vectorize problems will help our understanding of them. Point (2) is because there are many phenomena where the influence of X on Y is not the same as Y on X. The example he

discussed was in wildabeest herding, or fish swimming in the water. A wildabeest (X) running in the middle of the pack as knowledge of the wildabeest in front of it, but not behind it (Y). Therefore, X can influence Y, but Y cannot influence X. No wildabeest are aware that the wildabeest behind them are keeping proper pace, yet somehow all of them do. The interesting notion here is that you can model the flow of a herd in a very similar way that you can model a fluid, but fluids work through hydrodynamic interactions, and the physics is reciprocal. Yet, here it is not. But somehow it works. Why?

I mention this all for the purpose of inspiring you to try modeling some phsyical systems. Look for some vectors. Classify behaviors. Run some simulations. See what insight you can get. Best of luck!

Chapter 8

Math Essentials

8.1 Linear Algebra

8.1.1 Cross Products, Dot Products, and Gradients

Recall that the cross product is a vector value whose direction is perpendicular to both of the two vectors crossed. The formula is:

$$\vec{A} \times \vec{B} = ||\vec{A}|| \; ||\vec{B}|| \sin\theta \vec{n} \tag{8.1}$$

Where $||\vec{A}||$ is the length of $\vec{A}$, θ is the angle between the two vectors, and $\vec{n}$ describes the unit vector between $\vec{A}$ and $\vec{B}$.

The dot product is a scalar value achieved by multiplying the positions of vectors together, such as below:

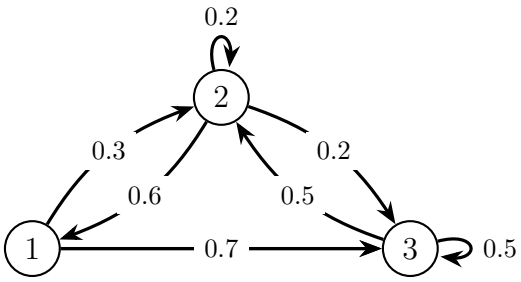
$$\mathrm{A} = \begin{bmatrix} a \\ b \end{bmatrix} \cdot \begin{bmatrix} x & y \end{bmatrix} = ax + by \tag{8.2}$$

A gradient, ∇, describes a derivative vector of a corresponding vector. For example, if vector v_1 exists in 3 dimensions, then:

$$\nabla \cdot \vec{v_1} = \begin{bmatrix} \frac{\partial}{\partial x} \\ \frac{\partial}{\partial y} \\ \frac{\partial}{\partial z} \end{bmatrix} \cdot \vec{v_1} \tag{8.3}$$

8.2 Markov-chains

Markov chains are useful in predicting the next state desired.


0.2
2
0.3
0.2
0.6
0.5
1
0.7
3
0.5

Chapter 9

Neuron Modeling

It would certainly be worth reviewing the electrophysiology contained in Part **III** before trying this, unless you already have a strong understanding of biochemistry.

9.1 Hodgkin-Huxley

9.1.1 The Main Form

The pair won the nobel prize for this model, which formed the basis of our understanding of action potentials. Beyond neurons, it was used in modeling pacemakers of the heart, and muscle cell depolarizations before better models existed. The basis is simply KCL:

$$C\dot{V} = I - I_{Na} - I_K - I_{Leak} \tag{9.1}$$

Because the equations can be found in nearly any textbook[17] or Wikipedia page, I will focus on some of the conceptual understanding I had issues with at first. The complete equation Hodgkin and Huxley arrived at is as follows:

$$C\dot{V} = I - \bar{g}_{Na}m^3h(V - E_{Na}) - \bar{g}_K n^4(V - E_K) - \bar{g}_L(V - E_L) \tag{9.2}$$

There are a few main points to make here. Firstly, this model considers only 3 currents. Na and K are self explanatory, but *Leak* represents the small amount of current that will always occur in cells due to the many routes of charged particles passing through the membrane. It is restorative, in that it pushes the membrane potential back to the resting voltage.

9.1.2 Gating and Conductance

The $\bar{g}$ represent the maximal conductance of these ions. Of course, you mak ask, "*shouldn't conductance be variable, depending on how many channels are open*?" Yes, which is what m, h, and n are

[17] Izhikevich, *Dynamical Systems Neuroscience*

for! These three variables are effectively kinetic fits of the opening and closing dynamics of sodium and potassium channels. Again, I will not mention these equations explicitly as they can be found anywhere. Conceptually, there are three things to know:

Firstly, m is an activation curve for sodium, and the power to the 3rd represents that there are three activation gates. h is an inactivation gate for sodium. Potassium has 4 activation gates, n, and no inactivation gate. Gating can be any number of things, for example, h could be a conformtional change that occurs in the channel after it has been open for $0.1ms$ that closes it again. Naturally, the gating for every channel will be different. Because the *Leak* current is an ensemble of many channel types, it will not have "gating" per se.

Secondly, m, h, and n all are between 0 and 1 and represent the **proportion of channels open**. For instance, if $n = 1$, then 100% of potassium channels will be open. This is why we multiply by the maximal conductance.

Thirdly, m, h, and n are dependent upon voltage, which affords them a time constant τ. This is the conceptually most difficult part. The experimental explanation may be beneficial in understanding. Hodgkin and Huxley realized that these three gating variables will converge to different values depending on the voltage. This makes sense, because we know potassium channels are voltage gated, we would expect the gating variable n to converge to around 1 as the voltage increases. But, the rate at which channels open and close is different. Therefore, their experiments were done to vary the voltage and determine how long it took the conductance of the channels to converge to some value. Does this make sense? In simplest terms: channels open and close at different rates, and that depends on the voltage.

What is the implication of this? Again, look up the exact equations if you are interested. Otherwise, trust the following: m has a time constant τ_m which is very small compared to τ_h and τ_n. Meaning, sodium channels will open the fastest in response to a voltage increase, causing depolarization of the cell. After some delay, sodium channel inactivation (h) and potassium channel activation (n) will kick in, causing repolarization and then hyperpolarization.

These are all derivatives.

One of the most difficult conceptual understandings I had was that $\dot{V}$, m, h, and n are all rates that depend on different time constants, which take voltage as their input. So, the derivative of voltage depends on the derivative of m, h, and n, which depend on voltage. The cyclic nature of this makes it strange, but still doable. Use the general form of derivative, $x_{i+2} = x_{i+1} + (x_{i+1} - x_i)/t$, follow the math, and you will survive.

9.2 Fitzhugh-Nagumo Reduction

9.2.1 Why would we simplify this system?

Reduction implies we are reducing the amount of variables. But why would we do this? The system is already incredibly generalized. We only consider two ion channels and are looking at a static neuron. How can we be accurate if we simplify this system any further?

Let's start by doing a simple thought experiment regarding the previous model:

$$C\dot{V} = I - \bar{g}_{Na}m^3h(V - E_{Na}) - \bar{g}_K n^4(V - E_K) - \bar{g}_L(V - E_L)$$

As mentioned, m, h, and n have their own time constants $\tau_{m,h,n}$. That means you'll need to do at least 6 calculations in order to determine $\dot{V}$, which, because it is a derivative, has its own time constant τ_v. Thus, the whole equation is 4^{th} dimensional with respect to time and requires at least 7 or so calculations per time step. If you'd like to simulate an action potential for around 10ms with a time step of 0.01ms, that means you'll perform around 7,000 calculations. Which is not so bad!

However, let's say you want to attempt a propagating action potential. Many people would model this on an infinitely long neuron/wire, but for the sake of this thought experiment let's say you're just interested in a 1 cm neuron/wire for 10 ms. To account for this spatial consideration, you'll need to add in another term besides I which receives current input from the previous segment of the neuron. So, this brings us up to at least 8,000 calculations.

You'd probably want to divide up the neuron into segments on the order of 1 μm. This multiplies our 8,000 calculations by an additional 100,000, giving us 800,000,000 to worry about. Still, this is not horrendous. But, this considers a 1D wire. Neurons are 3D dimensional. We are already considering a system that is 4^{th} dimensional with respect to time, and now we desire to consider 3^{rd} dimensional with respect to space. Imagine trying to calculate the flux through a $1000 \times 1000 \times 1000$ resolution box (i.e., perhaps μm^3 with good resolution). The surface area of this box is thus 6×10^6. Now extend this surface area to include the length of the wire and the area of the soma and dendrites, giving you thousands of millions of points to calculate per time iteration. And, we are still only considering two ion channels. Neurons have dozens and dozens of channels all with different gating kinetics. It does not consider things like lateral inhibition, birufcation, dendritic input, etc. I'll not bother telling you how many calculations we need to perform beyond this point—but it would be large.

9.2.2 How to Reduce

A note from future Jackson: I believe this section was written sometime in early 2023. After taking Mathematical Biology, I realize how stupidly simplified this is! It provides a good bit of conceptual understanding, but the actual path to creating the FHN model was much more complicated and elegant. FHN also provided the basis for many Dynamical Systems, extending far beyond neuroscience. It should be wholly respected.

What do we know about the time constants mentioned in the previous section? Roughly speaking, some are fast and some are slow. The upswing of an action potential is on a fast time constant, and the repolarization is on a slow time constant. We also know that the upswing portion is roughly a positive feedback loop, so as voltage increases, so should the derivative of voltage.

This helps us arrive at least at the following:

$$\dot{V} = V \times f(x) \tag{9.3}$$

Simply meaning that the derivative should scale with voltage in some way. We also know that there are at least two "equilibrium points" in a neuron. Meaning, when the neuron is at rest, the $\dot{V}$ will

be zero. And, when the neuron reaches the peak of the action potential, the same is true. This will allow us to immediately assume something interesting:

$$\dot{V} = V(V - V_{rest})(V_{max} - V) \tag{9.4}$$

We are already almost there. What we have just done is said that when either $V = V_{rest}$ or $V = V_{max}$, the $\dot{V}$ will not change. These are all on the aforementioned "fast" time scale, and as this is representative of the activation of the action potential, it is effectively a simplification of the sodium channel dynamics. This is also extremely easy to measure experimentally.

On our second time scale, the slow time scale, we have the inactivation/repolarization function. How will this look like? Just as with the first equation, we will want this curve to increase in magnitude with voltage. Because n represented the potassium channel activation in the previous segment, we can use that as our repolarization function here.

$$\dot{n} = V - \gamma n \tag{9.5}$$

What does this say? It says that our repolarization curve $\dot{n}$ will increase with respect to voltage. But, it will also decrease with respect to itself according to some scaling factor γ.

Now we have reduced our function down to two dimensions and can combine terms:

$$\begin{aligned} \dot{V} &= V(V - V_{rest})(V - V_{max}) - n \\ \dot{n} &= V - \gamma n \end{aligned} \tag{9.6}$$

But, we still want voltage to be affected by an injected current, so we can simply add this term back in. And it is also in this equation that we will add our spatial dependence to reach the following:

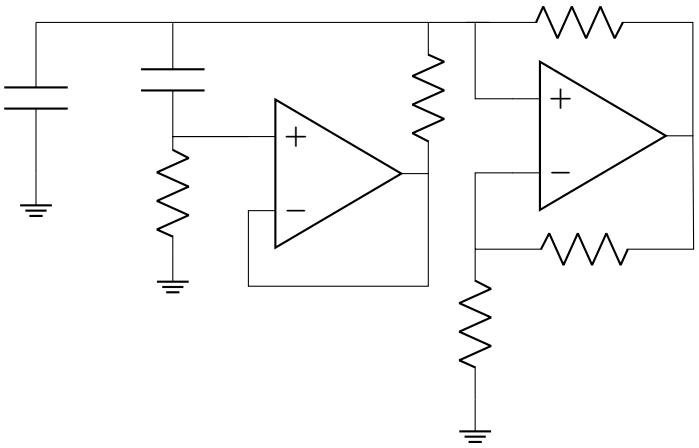
$$\dot{V} = I_{app} + [V(V - V_{rest})(V - V_{max}) - n] + D\frac{\partial V^2}{\partial x^2} \tag{9.7}$$

D is our spatial dependence, which represents the diffusion of charge around the neuron membrane. And that's it, for now!

Circuits digression.

A lot of the original work done by Fitzhugh and Nagumo used circuit equivalents in order to model neurons. One such example is as follows[18]:

[18]This circuit is adapted from *Mathematical Physiology*, by James Keener & James Sneyd (1998).



The rightward opamp functions as a Schmitt Trigger, and the entire thing is effectively an opamp oscillator with a second opamp in the middle. The purpose is to simulate an excitable system, like a neuron that is continually firing. Excitable systems systems are those that fire and have some refractory period before firing again (for example forest fires, or even your toilet).

9.3 Diffusion

Note that in the coming sections, $[f]$ is used to describe the concentration of some molecule f. But, this principal can be applied to anything, including the spreading of voltage across some surface.

9.3.1 Diffusion Equation Derivation

So, let's consider some particle that adheres to the arbitrary function $f(x)$, and has some probability density function of $\Psi(x,t)$ and experiences some noise of W. We can know that, regarding W:

$$\begin{aligned} dW &= 0 \\ dW^2 &= dt \end{aligned} \tag{9.8}$$

By the Itō lemma. And where $dx = \alpha dW$, where α is some function of $k_B T$ and the mobility coefficient, M, of the particle. We can say for an ensemble of particles that:

$$\langle f(x) \rangle (t) = \int \Psi(x,t) f(x) dx \tag{9.9}$$

If we are interested in the change of $f(x)$ over dx, we consider:

$$df = f[x + dx] - f[x] \tag{9.10}$$

And we Taylor Expand around df as:

$$\begin{aligned} df &= f' \alpha dW + \frac{1}{2} f'' \alpha^2 dW^2 + \mathcal{O} dW^3 \\ df &= f' \alpha dW + \frac{1}{2} f'' \alpha^2 dt + \mathcal{O} dt dW \end{aligned} \tag{9.11}$$

If we average over the ensemble, we get:

$$\frac{d\langle f\rangle}{dt} = \frac{1}{2}\langle f''\rangle\alpha^2 \tag{9.12}$$

This can be solved as:

$$\frac{\partial\Psi}{\partial t}f = \frac{1}{2}\frac{\partial^2\Psi}{\partial x^2}f\alpha^2 \tag{9.13}$$

And, because f never meant anything anyway, we are saying:

$$\frac{\partial\Psi}{\partial t} = D_c\frac{\partial^2\Psi}{\partial x^2} \tag{9.14}$$

Another approach.

Let's consider beginning this equation with the Master equation, which describes the probability (still using Ψ) of being at position r at time $t + \Delta t$ for one dimension as:

$$\Psi(r, t + \Delta t) = \Phi\Psi(r - \vec{e_i}, t) + (1 - \Phi)\Psi(r + \vec{e_i}, t) \tag{9.15}$$

Where Φ is the probability of moving left by a distance of $\vec{e_i}$ along your 1-dimensional lattice. In other words, $\vec{e_i}$ is the discrete length that you can travel along this lattice. Do you hate this notation? Me as well.

Let us know try to find the solution to this by Taylor expansion:

$$\begin{aligned}
\Psi + \Delta t\frac{\partial\Psi}{\partial t} &= \Phi\left(\Psi - \vec{e_i}\Psi' + \frac{\vec{e_i}^2}{2}\Psi''\right) + (1 - \Phi)\left(\Psi + \vec{e_i}\Psi' + \frac{\vec{e_i}^2}{2}\Psi''\right) \\
\Delta t\frac{\partial\Psi}{\partial t} &= -2\Phi\vec{e_i}\Psi' + \vec{e_i}\Psi' + \frac{\vec{e_i}^2}{2}\Psi'' \\
\frac{\partial\Psi}{\partial t} &= \frac{\vec{e_i}(1 - 2\Phi)}{\Delta t}\Psi' + \frac{\vec{e_i}^2}{2\Delta t}\Psi''
\end{aligned} \tag{9.16}$$

Which we can simplify to:

$$\frac{\partial\Psi}{\partial t} = \frac{\vec{e_i}(1 - 2\Phi)}{\Delta t}\frac{\partial\Psi}{\partial x} + D_c\frac{\partial^2\Psi}{\partial x^2} \tag{9.17}$$

We could assign the value of $\vec{e_i}(1 - 2\Phi)/\Delta t$ to be some constant. But, in general, $\Phi = 0.5$, signifying an equal probability of moving left or right, so this entire term is deleted, giving us the same as we got for equation **9.14**.

9.3.2 Forward Euler's

Diffusion is accomplished using some diffusion coefficient (D_c) multiplied by some measure of the proportion in one compartment verses another (often $\partial^2[f]/\partial x^2$). D_c can be tuned however desired. The important bit is the second derivative of $[f]$ with respect to space. This can be done using the general form of a second derivative, as written below:

$$f'' = \frac{f_{x+1} - 2f_x + f_{x-1}}{x^2} \tag{9.18}$$

One may wonder how one would solve for an edge case, as the general form of a second derivative requires three data points. There are some nuances, but in general the solution is simply the first derivative of the non-edge side. That is, since the second derivative is the difference in derivatives, that leaves simply the derivative of one side minus zero. This is like applying a closed end to your surface. You can ponder how to solve for an open end, if that ever arises.

This method has some slight issues in which the $[f]$ can occasionally go negative. The way in which this occurs is stated below (note that now $[f]$ is used instead of f to signify concentration at a value x and time t). We can first describe a simplified version of the problem:

$$\begin{aligned}
[f]_i &= f_0 \exp(-mt) \\
\frac{d[f]}{dt} &= \frac{[f]_{i+1} - [f]_i}{\Delta t} = -m[f]_i \\
[f]_{i+1} &= (1 - m\Delta t)[f]_i = [f]_i - m\Delta t[f]_i \\
[f]_i &= (1 - m\Delta t)[f]_{i-1} = (1 - m\Delta t)^2[f]_{i-2} \\
[f]_i &= (1 - m\Delta t)^i[f]_0
\end{aligned} \tag{9.19}$$

You can see easily, from this, that if Δt is too big, you will abandon the characteristic decay you'd expect from $f_0 \exp(-mt)$, and instead get some diverging oscillatory function. How this applies to our interest in diffusion is described below:

$$\begin{aligned}
\frac{[f]_x^{t+1} - [f]_x^t}{\Delta t} &= \frac{[f]_{x+1}^t - 2[f]_x^t + [f]_{x-1}^t}{x^2} \\
[f]_x^{t+1} - [f]_x^t &= \frac{\Delta t}{x^2}\left([f]_{x+1}^t - 2[f]_x^t + [f]_{x-1}^t\right) \\
[f]_x^{t+1} &= \frac{\Delta t}{x^2}[f]_{x+1}^t + \left(1 - 2\frac{\Delta t}{x^2}\right)[f]_x^t + \frac{\Delta t}{x^2}[f]_{x-1}^t
\end{aligned} \tag{9.20}$$

Therefore, if we want to ensure that the concentration is always positive, we are constrained by:

$$\begin{aligned}
1 - 2\frac{\Delta t}{x^2} &= 0 \\
\Delta t &< \frac{x^2}{2}
\end{aligned} \tag{9.21}$$

The relevance of this being that if one were interested in modeling on a very small Δx, then one would have to use a Δt that is not physiological, and thus waste a great deal of computing power in doing so. This can be avoided explicitly using some other methods, discussed next.

9.3.3 Backward Euler's

This form serves to solve the time-scale dilemma by swapping $[f]_{x+1}$, and can be used with any Δt. Let us consider the same example from above:

$$\begin{aligned} [f]_i &= f_0 \exp(-mt) \\ \frac{d[f]}{dt} &= \frac{[f]_{i+1} - [f]_i}{\Delta t} = -m[f]_{i+1} \\ [f]_{i+1} - [f]_i &= -m\Delta t[f]_{i+1} \\ [f]_{i+1} &= \frac{1}{1+m\Delta t}[f]_i \end{aligned} \tag{9.22}$$

Naturally, there is no longer a concern of the size of Δt. Though, one immediate concern is the difficulty of solving your equation for $[f]_{i+1}$. Getting back to the diffusion interest, we now have:

$$\begin{aligned} \frac{[f]_x^{t+1} - [f]_x^t}{\Delta t} &= \frac{[f]_{x+1}^{t+1} - 2[f]_x^{t+1} + [f]_{x-1}^{t+1}}{x^2} \\ [f]_x^{t+1} - [f]_x^t &= \frac{\Delta t}{x^2}\left([f]_{x+1}^{t+1} - 2[f]_x^t + [f]_{x-1}^{t-1}\right) \end{aligned} \tag{9.23}$$

Which simply replaces the previous $[f]_x^t$ with $[f]_x^{t+1}$. This leaves us with three unknowns (those being $[f]^{t+1}$ at $x-1, x, x+1$). We must use linear algebra to solve by first rewriting the left and right side as vectors and a matrix in the following way:

$$[\vec{f}]^{t+1} = \begin{bmatrix} [f]_0^{t+1} \\ [f]_1^{t+1} \\ \vdots \\ [f]_x^{t+1} \end{bmatrix}; [\vec{f}]^t = \begin{bmatrix} [f]_0^t \\ [f]_1^t \\ \vdots \\ [f]_x^t \end{bmatrix} \tag{9.24}$$

and

$$\mathrm{A} = \frac{\Delta t}{x^2}\begin{bmatrix} \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & -2 & 1 & & \vdots \\ \vdots & 1 & -2 & 1 & \vdots \\ \vdots & & 1 & -2 & 1 \\ \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix} \tag{9.25}$$

The corners $1,1$ and x,x were intentionally omitted, as what one desires to do with this is dependent on how they would prefer to treat their edges. As described before in the *Forward Euler's* method, one can use instead of the $(1,-2,1)$ pattern, simply $(\varnothing,-1,1)$ pattern, which signifies a closed edge. Therefore, together now we get:

$$\begin{aligned} [\vec{f}]^{t+1} - [\vec{f}]^t &= \mathrm{A}[\vec{f}]^{t+1} \\ \mathrm{I}[\vec{f}]^{t+1} - \mathrm{A}[\vec{f}]^{t+1} &= [\vec{f}]^t \\ [\vec{f}]^{t+1} &= (\mathrm{I}-\mathrm{A})^{-1}[\vec{f}]^t \end{aligned} \tag{9.26}$$

Where I is the identity matrix.

Remarks.

Foward Euler's is explicit, and will be preferred whenever the differential equations are non-stiff[19]. It is the more accurate of the two methods, and can be less computationally intensive if your decay rates are all slow. Backward Euler's is implicit, and can not be used to solve everything, but is doable in most cases. The extra computation required to solve the system of equations more than makes up for potential limitations in your Δt.

Ramblings.

I have this constant wonder if this method can be used to model the spread of voltage over a resistor lattice, by perfecting the D_c value. I am presuming it would work for a sufficiently large lattice, perhaps 100×100. This would be a good exercise to do sometime in the future.

[19] On approximately the same time scale.

Chapter 10

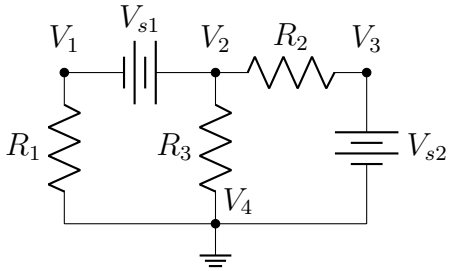
Math in Electronics

10.1 Modeling Circuits

The idea of this algorithm comes from[20] and is an expansion of what was discussed earlier in section **2.3**. Not by coincidence, both of the examples in the earlier section had symmetric matrices. The general way to extract equations from a circuit is in the following pattern:

$$\begin{bmatrix} \mathrm{G} & \mathrm{V} \\ \mathrm{V}^T & \mathrm{Z} \end{bmatrix} \tag{10.1}$$

G is essentially the way nodes are connected by resistors (represented as conductances). The diagonal represents the total number of resistors connected to a diagonal, and the other bits represents how nodes are connected by resistors. I will use the example provided in the footnote directly to illustrate:



There are 4 nodes, but as mentioned before, Node 4 is grounded, a reference, and thus not included in our calculations. This will mean our matrix G must be $n \times n$, where n is $N - 1$. Node 1 is connected only to R_1, resolving the $(0, 0)$[21] position of the matrix to be G_1. Node 2 is connected to R_2 and R_3, giving us $(1, 1) = G_2 + G_3$. Node 3 to only R_3, giving us $(2, 2) = G_3$. Node 2 and Node 3 are connected via R_2, meaning the $1, 2$ and $2, 1$ positions will have a G_2, though notably, it will be $-G_2$:

$$\mathrm{G} = \begin{bmatrix} G_1 & 0 & 0 \\ 0 & G_2 + G_3 & -G_2 \\ 0 & -G_2 & G_2 \end{bmatrix} \tag{10.2}$$

[20] `https://lpsa.swarthmore.edu/Systems/Electrical/mna/MNA3.html`

[21] You are a computer scientist in addition to a mathematician, so of course our matricies must be 0 indexed.

There are two voltage sources ($V = 2$), and 3 nodes, meaning the matrix must be $n \times m$, where $n = N - 1$ and $m = V$. You fill the matrix as if the positive terminal of the j^{th} voltage source is connected to the i^{th} node, point $(i, j) = 1$, or $(i, j) = -1$ for the negative terminal. So for the above, since the negative terminal of V_{s1} is connected to Node 1, $(0, 0) = -1$, and as the positive terminal connects to Node 2, $(1, 0) = 1$. Lastly, as Node 3 connects to the positive terminal of V_{s1}, $(2, 1) = 1$.

$$\mathrm{V} = \begin{bmatrix} -1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{10.3}$$

$$\mathrm{V}^T = \begin{bmatrix} -1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{10.4}$$

And lastly, Z for zero is an $n \times n$ matrix of zeros where $n = V$:

$$\mathrm{V}^T = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \tag{10.5}$$

Giving us an $\mathrm{A}x = \mathrm{B}$ of:

$$\begin{bmatrix} G_1 & 0 & 0 & -1 & 0 \\ 0 & G_2 + G_3 & -G_2 & 1 & 0 \\ 0 & -G_2 & G_2 & 0 & 1 \\ -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \\ V_3 \\ I_1 \\ I_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ V_{s1} \\ V_{s2} \end{bmatrix} \tag{10.6}$$

You may, rightfully, say "*Uh, who cares?*," since we already know that each row corresponds to a different equation. Well, there are a few reasons. Firstly, doing this algorithmically allows us to avoid accidentally underdetermining our matrix using the ol' eye-balling it technique. More importantly, though, the algorithmic approach allows us to solve this via code. Once I can write a bit more neatly, I'll likely upload some snippets here. Note from future Jackson: Lol, I just realized I did not, in fact, ever do that.

10.1.1 Error Correction Codes, Hamming

We know digital files are stored as 1s and 0s[22]. However, there must be a mode of corruption-prevention, lest any sort of damage, where it be digital (a 1 misread and stored as a 0) or physical (the scratching of a CD), would lead to loss of signal. A simple way to do this would simply be to store additional copies of every file, and allow your computer to compare the copies and select the proper result. Let us consider Hamming codes, the old-style technique for error correction.

Long ago, a so-called method of parity check was invented. Simply put, if you have an $n \times n$ array of binary numbers, a single additional bit was allocated to determine if the number of 1s is even or odd. Therefore, you can allocate this parity position for an $n \times n$ array such that the probability of error is unlikely to occur twice. This of course allows you to detect the existence of an error, but not its location or the ability to correct it.

[22] https://www.youtube.com/watch?v=X8jsijhllIA&ab_channel=3Blue1Brown

First, consider an array of bits, say 3×3. To exactly identify any errors in said bits, you might expect you'd need a 9-bit copy. However, imagine instead that you use a Sudoku-style arrangement, where 6 bits are stores on the edges of this array and determine the sum of the indices (summing the bits: $000 = 0$, $010 = 1$, $110 = 0$, etc.). Now, you can certainly identify an error, such that if you have a row $000 = 1$, you can know one of the bits has flipped. By comparing with the other sums, you can identify which of the three.

Now suppose that rather than applying these checks to the edges, you instead apply parity checks to certain subsets of a larger $n \times n$ matrix. How can we apply parity checks in clever ways such that they minimize memory and maximize our ability to check for errors? Consider the bits in the matrix below, where bolded bits are parity bits.

Correct:

x	**1**	**1**	0
1	1	0	0
0	0	1	1
1	0	0	1

Incorrect:

x	**1**	**0**	0
0	1	0	0
1	0	1	1
1	0	1	1

Incorrect:

x	**1**	**1**	0
1	1	0	0
0	0	1	1
1	0	0	1

Parity bit at position $(1, 2)$ ($P_{1,2}$) can check for parity in columns 2 and 4—which I will write as $(:, 2)$ and $(:, 4)$ now as in Python notation. $P_{1,3}$ can check for parity in $(:, 3)$ and $(:, 4)$. Therefore, you can immediately localize a single error to its correct column. That is, if $P_{1,2}$ and $P_{1,3}$ are both incorrect, you can know immediately that the error must be in column 4. Similarly, $P_{2,1}$ can check error in $(2, :)$ and $(4, :)$, and $P_{3,1}$ can check for error in $(3, :)$ and $(4, :)$. You can see clearly how the error can be localized to a single bit from the examples above. Naturally, though, there is complete breakdown when there is more than one error. You may ask: what if a parity bit gets corrupted? Consider the third column—only the parity bit itself changes. No other parities recognize an error. Therefore, the computer can localize the error to this bit, once again.

Position 0 is intentionally left as an x. No altered parity bits suggests either that there are no errors, or that the error is in position 0. To avoid questions with this 0th position, we can either discard it, or allocate it as a second tier of parity. One such way to do so will be to use it as a whole-number parity. That is, if this bit flips once, you can know there was one error. If it flips twice, you can know there are at least two errors. One would normally call the example above a $(15, 11)$ Hamming code, and adding the 0th block is called an extended Hamming code.

Now the question is, how do we scale? First, ask yourself: how many binary searches are required to fully parse a $2^n - 1$ binary number? Well, just now we saw that a 16-bit number requires 4 parity bits to be allocated. Therefore, a $2^8 - 1 = 256$ bit number requires 8 parity bits.

Chapter 11

Fluid Mechanics and Cell Motility

This either needs complete removal or a serious overhaul—I was not as diligent as I hoped I would be with fleshing out this section.

11.1 Introduction

The life of small things is hard. Brownian motion reorients E. coli a full 90° approximately once per second. The objects which they are trying to chase after and eat, too, undergo Brownian motion. A good question to ask yourself, then, is what do they do to make it worth it?

11.2 Cell Motility

I have an idea. I'm wondering if you can use cell motility models, with a bit of adaptation, and use it to model growth cone motility. I.e., model an ensemble of neurons after injury. The motivation being that the architecture of the neurons after injury is quite important to recovery. This idea, sprouting, is explored in-depth throughout this work, such as in a discussed Courtine review. Similarly, the movement of axons on a single cell level is naturally a form of cell motility.

There have been a number of papers that have gathered the 3D structure of axons after injury—so it is not impossible to verify models[23,24].

11.3 Fluid Mechanics

This actually might be an important section, in retrospect. For example, similar concepts are used to solve Maxwell's equations, etc., and to understand the current spread through an electric field in neural tissue.

There are two main open questions in the field of fluid mechanics. First, Newton's equations have been used for centuries and, in large part, model the world very well. However, we still haven't determined if these work completely, in all cases. Another open question is the origin of turbulence. Broadly speaking, we can define fluids as something that can be deformed, are continuous, and have no memory of its previous position. Complex fluids, like toothpaste, exhibit both fluid-like behaviors,

[23] https://www.sciencedirect.com/science/article/pii/S0896627321007753?via%3Dihub
[24] https://www.sciencedirect.com/science/article/pii/S2211124720308883?via%3Dihub

but also retain their shape when left alone.

Knudsen number is used to define how continuous a fluid is:

$$K_N = \frac{\lambda}{L} \tag{11.1}$$

Knowing a system is continuous allows us to define dynamic, average quantities like $\rho(x;t), p(x;t), \mathbf{u}(x;t)$. Our pressure value, $p(x;t)$, is useful in defining the resistance of a fluid to force. Pressure is isotropic, meaning it is independent of direction. Defining the velocity, $\mathbf{u}(x;t)$, may be done by quantizing some small segment of the fluid and following its trajectory. This is the Lagrangian description. The trajectory that your quantized box follows is called the *pathline*. Alternatively, the Eulerian description treats the fluid as an ensemble, defining $\mathbf{u}$ everywhere and producing a vector-field like array of streamlines. Conversion between the Lagrangian and Eulerian descriptions is:

$$\frac{Df}{Dt} = \frac{\partial f}{\partial t} + \mathbf{u} \cdot f \tag{11.2}$$

You can define movement of a fluid with three fundamental flows: translation, rotation (solid body), and extension. Translation and rotation both do not cause any deformation. The streamlines simply move in unison in some direction. Extension causes deformation, and have hyperbolic streamlines. A microfluidic chamber, in which 4 channels come together in a cross, is an example of extension flow. At the intersection, components will be stretched. Deformation like this will lead to friction.

The Reynolds Number.

A large part of this section will come from[25] and[26].

The Reynolds number, R_e, described the flow of a fluid and is used to characterize it as turbulent or laminar. Such turbulent or laminar flow is interesting and important in biology. For example, flow in the circulatory system is typically turbulent (unless we are discussing the capillaries). Areas of high turbulence are known to accumulate plaques more frequently. It is calculated as:

$$R_e = \frac{\rho L U}{\eta} \tag{11.3}$$

$$R_{e\omega} = \frac{\rho L^2 \omega}{\eta} \tag{11.4}$$

Where ρ described the density of the fluid, L describes the characteristic length, U describes the flow speed, and μ describes the fluid's viscosity. It is common to see these variables interchanged with others, such as describing U as v, or η as μ, or others. Another common thing you may see is the

[25] https://www.cambridge.org/us/universitypress/subjects/mathematics/fluid-dynamics-and-solid-mechanics/fluid-dynamics-cell-motility?format=HB&isbn=9781107174658

[26] https://pubs.aip.org/aapt/ajp/article-abstract/45/1/3/1043148/Life-at-low-Reynolds-number?redirectedFrom=fulltext

replacement of ρ/η with a different constant called the *kinematic viscosity*, usually represented by ν.

The Reynolds number can be interpreted as a ratio of the inertial force to the viscous force. An R_e less than 10^3 is considered to be viscous dominated and laminar. An R_e greater than 10^4 is considered inertia dominated and turbulent. A more apt description may be that the viscous nature of the fluid holds significantly less weight for something at a high R_e—meaning that a blue whale's motion, which swims at an R_e of around 10^7, is dictated nearly purely by inertia, as you would expect. In microbiology, R_e is $<< 1$, making it highly viscosity dependent.

A nice equation you can find is the following:

$$\frac{\mu^2}{\rho} = F_{tow} \tag{11.5}$$

F_{tow} is the force required to "tow" an object in some liquid when $R_e = 1$. That is, any object with $R_e = 1$ can be towed by this force. The purpose of this is to demonstrate that small R_e is correlated to small forces. Another interpretation of R_e is that if you have an object, like a micro organism, with a very small R_e, you could push it with all your might, and nearly the instant you stop pushing, it will immediately halt. This is, again, because it is completely inertia-independent.

Swimming itself is a cyclic process, usually described as n cycles. Reciprocal swimming is when one's motion is equal in either the forward or reverse direction, and where one ends up in the same state at the end of the cycle. Interestingly, at high R_e this cycle of swimming is acceptable. The example described in *Life at Low* R_e is a clam who opens its mouth to move and then closes it, shunting it forward and is carried by momentum. Of course, such momentum does not help at low R_e, so a swimmer must have multiple, non-reciprocal movements.

Lizard Digression.

A neat question you can ask is how exactly a large animal, like a lizard, are able to run on water? To do this, we'll make a number of huge simplifications. First, let's say the lizard's foot deflects / indents the water in an approximately cube-like manner with a length R. The density of the fluid is what we'll call ρ_f. The speed that they're running we'll call v, and the frequency of their steps to be ω. The force generated by stepping, also called the Shoke force, is due to the inertia of the fluid, and is summarized as:

$$F_{up} \approx R^3 \rho_f v \omega \tag{11.6}$$

The lizard is fighting the force of gravity pulling it down. If we take the lizard to be approximately cube like, with dimension L and density ρ_l, then:

$$\begin{aligned} F_{down} &\approx L^3 \rho_l g \\ R^3 \rho_f v \omega &\geq L^3 \rho_l g \end{aligned} \tag{11.7}$$

To simplify further, let's assume $R \approx L$, and that $\omega \approx v/L$. Now we can simplify to:

$$\begin{aligned} \rho_f \frac{v^2}{L} &\geq \rho_l g \\ v &\geq \left(\frac{\rho_l}{\rho_f} g L\right)^{\frac{1}{2}} \end{aligned} \tag{11.8}$$

For the sake of not thinking too hard, you can intuit that body's have physiological limits. Perhaps theoretically, an elephant can run on water if it moves its legs fast enough. However, you can probably see that this isn't feasible; muscles have limits. Let's say the physiological constraint conforms to the following equation, where K is a constant describing muscle output:

$$\begin{aligned} Fv &\leq \rho_l L^3 K \\ R^3 \rho_f v^2 \omega &\leq \rho_l L^3 K \\ v &\leq \left(\frac{\rho_l}{\rho_f} L K\right)^{\frac{1}{3}} \end{aligned} \tag{11.9}$$

Therefore, the result you'll see is that to run on water, you are constrained between equation (11.8) and equation (11.9). At some point the result of (11.8) exceeds that of (11.9), making is so the burden of running on water exceeds physiological limits. This example wasn't explained in significant depth, and we made some grand simplifications, but the central theme is the ability to map the real, moving world onto equations and make some statements about what is possible.

11.3.1 Conservation of Mass

Let's discuss a fluid at point x and time t. Its velocity is $\mathbf{u}(x;t)$, and its density $\rho(x;t)$. If we are interested in some space, ω, we can determine the mass of fluid in this space by:

$$\int_\omega \rho(x;t) d\omega \tag{11.10}$$

If we are interested in the rate of change of fluid in our space ω, we can determine it by:

$$\int_\omega \frac{\partial}{\partial t} \rho(x;t) d\omega \tag{11.11}$$

Knowing that any change in mass of the fluid is due only to fluid flowing in and out of our space ω, we can rewrite this integral in terms of fluid passing the space's boundaries, which we'll call S, and let's define a vector normal to this space as $\vec{n}$. In other terms:

$$-\int_{S,\partial\omega} \rho \mathbf{u} \cdot \vec{n} dS = \int_\omega \nabla x \cdot [\rho \mathbf{u}] d\omega \tag{11.12}$$

The conversion between the two is a result of Gauss' theorem. This also means that, because ω is arbitrary:

$$\frac{\partial}{\partial t} \rho(x;t) = -\nabla x \cdot [\rho(x;t) \mathbf{u}(x;t)] \tag{11.13}$$

This is like saying the total change in density is equal to the change in density and velocity at all points x. Which we can rewrite in a way that is clearer, demonstrating mass is conserved:

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{u}) = 0 \tag{11.14}$$

Where $\partial\rho/\partial t$ is the change in density and $\nabla \cdot (\rho \mathbf{u})$ is the advection term. For an incompressible fluid, where your density is fixed, you are left with:

$$\nabla \cdot \mathbf{u} = 0 \tag{11.15}$$

This, of course, is the first equation in the Navier-Stokes equations, discussed in section **11.3.4**.

11.3.2 Stress Tensor

Conservation of momentum can be solved using the forces applied to the system. This includes body forces ($\mathbf{F}(x;t)\delta V$ for some volume δV) and forces applied to a segment of the surface ($\tau(x;t)\delta S$ for some surface $\vec{n}\delta S$).

Body forces scale with $\mathcal{O}(\mathbf{F}L^3)$, and the canonical example is gravity. Forces applied to the surface include pressure, which scales with $\mathcal{O}(\mathbf{F}L^2)$. The force of friction (or viscosity), just as the friction of two solids, comes in two components: the normal force and the force which opposes movement laterally.

To think about this, consider some arbitrary object with a face. Regardless of the face, it has some normal vector $\vec{n}$, and let's say this surface has an area A. As mentioned before, body forces scale with $\mathcal{O}(\mathbf{F}L^3)$, while surface with $\mathcal{O}(\mathbf{F}L^2)$. This means that when $\lim_{V\to 0}$, body forces and surface forces must be equal. So let's think of τ as:

$$\tau(x;t) = \lim_{A\to 0} \frac{F}{A} \tag{11.16}$$

Let's first think of a cube whose faces are aligned with basis vectors $\{\vec{e_1}, \vec{e_2}, \vec{e_3}\}$. All traction vectors can be summarized using these components such that if we have vector τ acting in the direction e_3, we can describe it as:

$$\begin{aligned} \tau_{e_3} &= \sigma_{3,1}e_1 + \sigma_{3,2}e_2 + \sigma_{3,3}e_3 \\ \tau_i &= \sigma_{i,j}e_j \end{aligned} \tag{11.17}$$

This works nicely and succinctly when we are discussing a vector τ in the direction of a basis vector. Cauchy describes a more general case for some arbitrary tetrahedron. Forces applied to three of the faces on the tetrahedron can be described using $\{\sigma_{i,1}, \sigma_{i,2}, \sigma_{i,3}\}$, but the fourth face is misaligned with the bases. The 4^{th} face of the tetrahedron has the normal vector $\vec{n}$, which can be defined as $\vec{n} = \vec{n_1}e_1 + \vec{n_2}e_1 + \vec{n_3}e_1$. A force acting on this face, τ is not aligned with any basis vector, but we can cheat by splitting τ into its components, $\tau = \tau_{n_1} + \tau_{n_2} + \tau_{n_3}$.

I'm going to yada-yada through the proof a little bit, but the gist is that if you take Newton's laws and describe them for each normal, you get something along the lines of:

$$\sum \mathbf{F_{e_1}} = -\sigma_{1,1}n_1 - \sigma_{2,1}n_2 - \sigma_{3,1}n_3 + \tau_{n_1} + mg = ma \tag{11.18}$$

Since, as mentioned earlier, we are taking the limit as the volume goes to zero, we can say that:

$$\begin{aligned} -\sigma_{1,1}n_1 - \sigma_{2,1}n_2 - \sigma_{3,1}n_3 + \tau_{n_1} + 0 = 0 \\ \sigma_{1,1}n_1 + \sigma_{2,1}n_2 + \sigma_{3,1}n_3 = \tau_{n_1} \end{aligned} \tag{11.19}$$

And this holds true for all of the components of τ. This makes it more obvious that σ will be linear and homogenous. We can now write this as:

$$\begin{bmatrix} \tau_{n1} \\ \tau_{n2} \\ \tau_{n3} \end{bmatrix} = \begin{bmatrix} \sigma_{1,1} & \sigma_{2,1} & \sigma_{3,1} \\ \sigma_{1,2} & \sigma_{2,2} & \sigma_{3,2} \\ \sigma_{1,3} & \sigma_{2,3} & \sigma_{3,3} \end{bmatrix} \begin{bmatrix} n1 \\ n2 \\ n3 \end{bmatrix} \tag{11.20}$$

Simplified, this is:

$$\tau_n = \sigma^T n \tag{11.21}$$

Where the transpose of σ multiplied by the normal gives us the force. So in short, to find the force acting on a plane you take some stress tensor and multiply it by the normal of that plane.

One component you may be interested in calculating is the force normal to the plane. Let's call this σ_n. This can be accomplished via a linear projection. From linear algebra, this is:

$$\begin{aligned} \sigma_n &= \text{proj}_n \tau_n = \frac{\tau_n \cdot n}{||n||} \\ \sigma_n &= \tau_n \cdot n \end{aligned} \tag{11.22}$$

Because the length of a unit vector, n, must equal 1. A shear stress, on the other hand, is parallel to the surface of the plane. For now, I'll call this π_n. Because the traction vector, τ_n, is the vector addition of σ_n and π_n. In other words, for either the vector or the scalar:

$$\begin{aligned} \vec{\pi_n} &= \tau_n - \sigma_n n \\ \pi_n &= ||\tau_n - \sigma_n n|| \end{aligned} \tag{11.23}$$

$\tau(x,t)dS$ describes the force exerted on a surface, called a traction. One component will be the normal force. The normal force will always point in the direction opposite of τ, which can be written as $\tau(x,t,-\vec{n}) = -\tau(x,t,\vec{n})$. To characterize this, you need to know the magnitude, direction, and orientation of the surface.

Defining $\tau(x,t)dS$ can be done with $\tau(\vec{n})dS = -\tau(e_1)dS - \tau(e_2)dS... = -[\tau(e_x)dS] \cdot \vec{n}$, where e_x is a surface. Somehow this leads to, via some linear projection, $dS_i = dS\vec{n} \cdot e_i$ and then $\tau = \sigma \cdot \vec{n}$. σ is your Cauchy stress tensor, in units of Pascals, which usually comes in the form of a 3×3 matrix.

This matrix, σ, is inherently symmetric.

11.3.3 Conservation of Momentum

To achieve the hydrodynamical equation of motion, we equate the rate of change of momentum within ω and out of the boundaries with the sum of forces acting on ω and the boundaries. The rate of change of momentum, in this case, is:

$$\int_{\omega} \frac{\partial}{\partial t}[\rho(x;t)\mathbf{u}(x;t)]d\omega \tag{11.24}$$

We can describe the outflow of momentum from the space ω by:

$$\int_{\S} \rho(x;t)\mathbf{u}(x;t)\mathbf{u}(x;t)\cdot dS = -\int_{\omega} \nabla x \cdot [\rho\mathbf{u}\mathbf{u}]d\omega \tag{11.25}$$

The sum of forces acting on the fluid within ω are loosely defined with X, where X is the force per unit volume from external forces:

$$\int_{\omega} X(x;t)d\omega \tag{11.26}$$

And the force force experienced, using σ as the stress tensor, is:

$$\int_{S} \sigma(x;t)\cdot dS = \int_{\omega} \nabla x \cdot \sigma d\omega \tag{11.27}$$

These are all combined and simplified into:

$$\frac{\partial}{\partial t}[\rho\mathbf{u}] + \nabla x \cdot [\rho\mathbf{u}\mathbf{u}] = X + \nabla x \cdot \sigma \tag{11.28}$$

We are allowed to make this simplification because ω is a nonexistant, arbitrary space.

$$\begin{aligned} \frac{\partial}{\partial t}\int_{\omega} \rho\mathbf{u}d\omega = \int_{\omega} \mathbf{F}d\omega + \int_{\partial S} \sigma\cdot\vec{n}dS - \int_{\partial D} \rho\mathbf{u}(\mathbf{u}\cdot\vec{n})dS \\ \rho[\frac{\partial \mathbf{u}}{\partial t} + \mathbf{u}\cdot\nabla\mathbf{u}] = \rho g + \nabla\cdot\sigma \end{aligned} \tag{11.29}$$

At present, this cannot be solved because we do not know how σ behaves in the system. With an assumption, we can find the constitutive equation for σ:

$$\sigma = -pI + \mu(\nabla\mathbf{u} + {}^{T}\nabla\mathbf{u}) \tag{11.30}$$

Where $-pI$ is pressure multiplied by an identity matrix, giving us an isotropic term. The second term is viscosity, which was empirically determined by Newton. It was shown that the drag force experienced by an object scales with $\mathbf{u}$ and A, but inversely with h. Thus, we get:

$$\frac{F}{A} = \mu\frac{\mathbf{u}}{h} \tag{11.31}$$

μ is a scaling factor, which is called the dynamic velocity and has units of Pascal·seconds.

11.3.4 Navier-Stokes Equations

In knowing the velocity vectors at all points in a fluid allows you to describe the movement of the fluid over time. Three key assumptions in this are that the fluid be Newtonian (meaning applying a shear force does not affect it in any way), incompressible, and isothermal. This brings us to:

$$\begin{aligned} \nabla \cdot \mathbf{u} &= 0 \\ \rho \frac{\partial \mathbf{u}}{\partial t} &= -\nabla p + \mu \nabla^2 \mathbf{u} + \mathbf{F} \end{aligned} \tag{11.32}$$

$\nabla \cdot \mathbf{u} = 0$ simply states that the mass is conserved. When describing a 3 dimensional velocity, we'll consider (x, y, z). Thus:

$$\nabla \cdot \mathbf{u} = \frac{\partial \mathbf{u}}{\partial x} + \frac{\partial \mathbf{u}}{\partial y} + \frac{\partial \mathbf{u}}{\partial z} \tag{11.33}$$

Taking the gradient of $\mathbf{u}$ is used to solve the divergence of the field. A nonzero divergence would describe a source of fluid flowing into the system (or a sink).

The second equation is a rewriting of Newton's second law, $\sum F = ma$. Some would describe it as being a conservation of momentum equation. Let's see how:

$$\begin{aligned} F &= ma/V \\ F &= \rho a \\ F &= \rho \frac{d\mathbf{u}}{dt} \end{aligned} \tag{11.34}$$

Because we are considering individual points, rather than the ensemble, we divide by the volume V, giving us the density. Acceleration can be rewritten as the change in velocity over time, $d\mathbf{u}/dt$. Next we will consider the forces relevant to this system.

$$-\nabla p + \mu \nabla^2 \mathbf{u} + \sum F = \rho \frac{d\mathbf{u}}{dt} \tag{11.35}$$

The ∇p describes one of two internal forces. It accounts for forces due to a change in pressure (recall that $p = F/A$). Again, in a 3D space, $\nabla p = \partial p/\partial x + ...$, etc. The second term, $\mu \nabla^2 \mathbf{u}$, describes force of friction, and or viscosity.

Lastly, we can summarize all external forces ($\sum F$) within this big $\mathbf{F}$. We may see gravity as an external force, for example, in which case we may replace $\mathbf{F}$ with ρg. Therefore, the Navier-Stokes equations are essentially the reapplication of fundamental concepts in physics for fluids.

$\mathbf{u}$ we can define, perhaps overly complexly, as:

$$\mathbf{u}(r_s\, t) = \dot{r_s} + \mathrm{U}(t) + \Omega(t) \times r_s \tag{11.36}$$

Where r_s is defined as any point on the swimmer, and $\dot{r_s} = \partial r_s / \partial t$. U$(t)$ describes the instantaneous linear velocity, and $\Omega(t)$ the angular.

As it goes, this equation can be made dimensionless as:

$$-\nabla p + \mu \nabla^2 \mathbf{u} + R_e \sum F = R_{e\omega} \frac{d\mathbf{u}}{dt} \tag{11.37}$$

Given that both R_e and $R_{e\omega}$ are expected to be tremendously small for micro organisms, we can reduce the equations to the simplified Stokes equations. Importantly, this is simply the same as **removing the inertial terms**:

$$\begin{aligned} \nabla \cdot \mathbf{u} &= 0 \\ \nabla p &= \mu \nabla^2 \mathbf{u} \end{aligned} \tag{11.38}$$

Another slight confusion is that this entire equation seems to be zero. Maybe I am simply bad at math, but if $\nabla \cdot \mathbf{u} = 0$, then how does $\mu \nabla^2 \mathbf{u} \neq 0$?

11.3.5 Propulsion

Presently, I do not understand the theory behind this, but somehow everything that occurs at low R_e is linear. As it goes, force and torque are related by the following equations:

$$\begin{aligned} F &= AU + B\Omega \\ \tau &= CU + D\Omega \end{aligned} \tag{11.39}$$

We can use this to construct *propulsion* matrices like:

$$P = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \tag{11.40}$$

The proof is not shown here but interestingly $B = C$. To think of this matrix, let's consider dropping a cork-screw into a vat of syrup. You probably expect that the cork-screw will turn to some degree, but that there will still be some *slippage*. That is, it sinks at a rate faster than it spins. If something were to sink without spinning at all, you can imagine that the matrix would be completely diagonal, because, given our equation 11.39, that would mean the linear force, F, and the rotational force, τ, are completely decoupled. The propulsion efficiency is defined as (or proportional to) these off diagonal elements squared (i.e., B^2)—evidently, though, this is considered unimportant in the grand scheme of interactors.

Diffusion.

Diffusion coefficients describe how quickly an object moves around in a fluid. The formula, where a is the object's radius, is:

$$D = \frac{k_B T}{6\pi\eta a} \tag{11.41}$$

You may see the denominator rewritten as f, representing the friction coefficient of the object.

Fascinatingly, at very low R_e your movement inherently takes your surrounding environment with you. This actually has quite large consequences, and is called the *added mass phenomenon*. In short, if you move and carry a load of liquid along with you, it is as if you are propelling a much greater mass. It's estimated that:

$$\begin{aligned} \tau_{stir} &= \frac{l}{v_{stir}} \\ \tau_{diff} &= \frac{l^2}{D} \end{aligned} \tag{11.42}$$

Where τ is the time it takes you to move an object some distance l by either stirring or diffusion, and v_{stir} is the velocity of stirring performed. The ratio of such, i.e. lv/D (or LU/D), is called the Pèclet number (P_e). Maybe you can recall from earlier that R_e can be written as lv/ν (equivalently LU/ν), so you can see how the Reynolds number and the Pèclet number describe systems in a similar way. Stirring is only helpful if $lv/D > 1$, which gives us an indication of how much you can accomplish either by stirring or diffusion. When $P_e << 1$, diffusion dominates. For micro swimmers, like bacteria, stirring is virtually useless, and they're better off simply waiting for items to diffuse toward them. Another fascinating realization is that bacteria do not move to reach more nutrients—as we just stated that this would be useless. Instead, they travel to areas of a higher density, where the probability of diffusion is greater[27].

11.4 Literature

Here I'll discuss some of the literature surrounding fluid mechanics and neuroscience.

11.4.1 CSF Flow

A large topic is the flow of CSF inside the brain and spinal cord, which has implications in drug delivery and hydrocephalus. Other notable topics include the application of shear stress to neurons, which may activate mechanically gated channels.

Here I'll discuss[28]. The composition and flow of cerebrospinal fluid (CSF) is both a key diagnostic tool and often at the center of pathologies. For example, measuring CSF composition aids in diagnosis of diseases like Alzheimer's via monitoring amyloid presence[29]. Similarly, being able to model particle flow in the CSF is helpful in understanding drug delivery. Blockage of CSF flow is responsible for hydrocephalus, which on its face is easy to treat, but is frequently misdiagnosed as Alzheimer's or other neurodegenerative diseases[30].

[27] Okay—indeed there is something immensely fascinating here. Can the same be true for growth cones? I suppose... yes!

[28] https://www.cambridge.org/core/journals/journal-of-fluid-mechanics/article/onedimensional-model-for-the-pulsating-flow-of-cerebrospinal-fluid-in-the-spinal-canal/3710D122EC70FCB29DEF0AB196ADCCCD

[29] https://www.youtube.com/watch?v=zCmngDk9VDU

[30] https://www.alz.org/alzheimers-dementia/what-is-dementia/types-of-dementia/normal-pressure-hydrocephalus

The Sánchez group at UCSD has many papers on modeling CSF flow, so their works are a great resource for further learning. The specific focus of this work is on intracranial pressure (ICP), which can be clinically measured with the surgical implantation of sensors. The goal of this work is to allow for indirect measures, such as MRI, combined with computation to serve in place of invasive procedures. CSF is generated in the brain ventricles, travels through the CNS (including the spinal canal), and eventually exits through arachnoid villi, where it drains into the venus sinus. The venus sinus connects directly to the circulatory system, and thus the spaces housing CSF experience pressure fluctuations with the beats of the heart—they call this ICP pulsations. Many works have modeled the spinal canal and connected spaces as hollow—these simplifications limit the clinical relevance of the models. This work improves upon previous, as it considers the role of arachnoid trabeculae (web-like structures in the arachnoid space, which reduce the pressure).

In initial models, because the spinal canal's axial distance is significantly longer than its transverse, they use a slender-flow approximation, and state that the velocity and pressure is uniform for a given cross-section. The total volume of CSF is calculated to be 80cm^3, and with each ICP pulse $\approx 1\text{cm}^3$ of CSF either in or out of the spinal canal, which induces fluctuations in the cross sectional area of the spinal canal (some ΔA). They approximate this change to be about 1mm^2, and use a linear-elastic model to survey it. These simplifications allow the creation of a one-dimensional model. However, they comment that these initial models fail to capture a "phase lag" between pressure and flow changes that are observed by MRI—they attribute this to the exclusion of trabeculae from the space which would contribute to viscous pressure losses.

Therefore, the bulk of this work focuses on including the viscous pressure loss associated with trabeculae. They introduce a resistance coefficient which is dependent on the viscosity and pulse frequency, $\mathcal{R} = \nu/(k\omega)$, where k is a function of the trabeculae webbing. They also utilize the Womersley number, α, which can be written (using numbers from our class) as $= \sqrt{2\pi R_e St}$, or as $\sqrt{A_0\omega/\nu}$. An $\alpha << 1$ means the frequency of pressure pulses is slow enough that the velocity changes will be in-phase with pressure—whereas an $\alpha >> 1$ means it will be out of phase by $\approx 90°$ (from Wikipedia[31]). In fact, they include the pressure loss due to the Stokes layer (Stokes flow induced by the moving boundaries[32]) but state that this pressure loss is much smaller than is due to trabeculae. Therefore, future models may explore removing this term in order to further simplify.

To verify their model, they compare with actual MRI data from two patients. They vary the resistance coefficient, $\mathcal{R}$, and show that when $\mathcal{R} = 0$, no phase lag is observed—but one becomes present as $\mathcal{R}$ increases (i.e., the inclusion of trabeculae). Therefore, their updated one-dimensional model is able to account for clinically observed phase lags through the inclusion of increased resistance caused by trabeculae. Some interesting takeaways to me: 1) That you can make a clinically relevant, one-dimensional model at all! I'm surprised you can take the complex, 3D, curvilinear geometry of the spinal canal and reduce it down so much. 2) Learning of the Womersley number, α, is interesting and worth remembering for the future. 3) The fact that they verified the model with clinical data is very supportive. It's good to see them take it out of just theory and apply it.

[31] `https://en.wikipedia.org/wiki/Womersley_number`
[32] `https://en.wikipedia.org/wiki/Stokes_problem`

Chapter 12

Machine Learning

Yeah, whatever. I'll do it myself. I always find the cure to fret is learning.

So what is Machine Learning[33]? ML is a subset of AI which looks to find patterns and make predictions about data. ML has a few types:

1. Supervised learning, which uses input data which has an associated, given label.
2. Unsupervised learning, where data does not include labels and instead the learning's goal is to find patterns in the data.
3. Reinforcement learning, where an algorithm learns based on a system of rewards and punishments.

Ratings in learning can be binary, which is called *one-hot encoding* (i.e., you are either happy or not happy, 1 or 0), or there can be *ordinal data*, which has some order (i.e., a spectrum of feelings, say rated 1 to 5). These are qualitative pieces of data, and there can similarly be quantitative discrete or quantitative continuous types of data.

A computer can either predict a class, which is discrete, or can use a regression, to output a continuous value. A model will take a *feature vector* and output a classification.

Calculating Loss.

The first way to calculate loss is called L1 loss. It is simply calculated as:

$$L_1 = \sum \left(|y_{real} - y_{predicted}| \right) \tag{12.1}$$

The second way is called L2 loss. The function is quadratic, which increases penalty the further you get from the true value as:

$$L_2 = \sum \left(\left(y_{real} - y_{predicted} \right)^2 \right) \tag{12.2}$$

Finally, we can calculate binary cross-entropy loss. Without diving in to the formula, it is seen as:

[33] From: `https://www.youtube.com/watch?v=i_LwzRVP7bg&t=243s&ab_channel=freeCodeCamp.org`

$$L = -\frac{1}{N}\sum \left(y_{real} \cdot \log(y_{predicted}) + (1 - y_{real} \cdot \log(1 - y_{predicted})\right) \tag{12.3}$$

One could also compute by accuracy.

12.1 Supervised Learning

12.1.1 Various Models

KNN.

One model is called K-nearest neighbors (KNN). Suppose you have a set of binary classifications plotted on an X-Y graph. For example, you plot the relation between years spent in school and the money someone earns per year, and your binary classification is whether or not they own a home. You could compute the distance (on the plot) between nearby points. Suppose you take $k = 3$, to find the 3 nearest neighbors. If all of the nearest neighbors to a point is that one owns a home, you'd predict that such a point is also a home owner. If it is ambiguous, you'd predict via the minimum distance between each class and that point.

Naive Bayes.

Naive Bayes depends on conditional probability. So, recall that Bayes theorem goes as:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \tag{12.4}$$

Where this is the probability of A given B. Therefore, suppose the probability of getting a false positive is 0.05. Suppose the probability of getting a false negative is 0.01. And suppose the probability of having the disease at all is 0.1. Now, ask yourself: what is the probability of having the disease, given that you've tested positive?

We filled in 0.99 and 0.95 simply by knowing that they should sum to 1. That is, the probability that you test either positive or negative most definitely = 1. Now:

$$P(A|B) = \frac{0.99 \cdot 0.1}{0.99 \cdot 0.1 + 0.05 \cdot 0.9} = 0.6875 \tag{12.5}$$

We can use this for classifications. See below:

$$P(C_k|x) = \frac{P(x|C_k) \cdot P(C_k)}{P(x)} \tag{12.6}$$

This essentially asks: what is the probability of x fitting into some category C_k? $P(C_k|x)$ is considered the posterior. x is the feature vector. $P(x|C_k)$ is the likelihood. $P(C_k)$ is the prior. And finally, $P(x)$ is the evidence[34].

[34] I always forget what the Π symbol means. It's obvious from the below equations anyway, but simply because I know future me will go "*huh?*" I'll add the reminder that it is like summation notation, but with multiplication rather than summing. Thank me later.

$$
\begin{aligned}
P(C_k|x_1, x_2, ..., x_n) &= \frac{P(x_1, x_2, ..., x_n|C_k) \cdot P(C_k)}{P(x_1, x_2, ..., x_n)} \\
P(C_k|x_1, x_2, ..., x_n) &\propto P(x_1, x_2, ..., x_n|C_k) \cdot P(C_k) \\
P(C_k|x_1, x_2, ..., x_n) &\propto P(x_1|C_K) \cdot P(x_2|C_K) \cdot ... \cdot P(x_n|C_K) \\
P(C_k|x_1, x_2, ..., x_n) &\propto P(C_k)\Pi_{i=1}^{n} P(x_i|C_k)
\end{aligned} \tag{12.7}
$$

We removed the denominator because $P(x_1, x_2, ..., x_n)$ is the same for all classes, so it provides no class-specific info and thus can be neglected. The second assumption that we make is that all probabilities are independent, allowing us to multiply everything together.

We use a $\hat{y}$ as an *argmax*. I.e., we maximize $P(C_k)\Pi_{i=1}^{n} P(x_i|C_k)$. This is also called a Maximum A Posteriori (MAP).

Logistic Regression.

Normally, with a linear regression, we would be approximating a $\hat{y}$ value for some line. In this case, however, we will want to approximate a probability of being one class or the other. That is:

$$
\ln(\frac{p}{1-p}) = mx + b \tag{12.8}
$$

We take the ratio of p to $1 - p$ to ensure it remains in the range of $\{0, 1\}$. We use ln here to ensure the value never goes negative, as the slope can be positive or negative. Therefore, we can get:

$$
\begin{aligned}
\frac{p}{1-p} &= e^{mx+b} \\
p &= (1-p)\, e^{mx+b} \\
p &= \left(\frac{e^{mx+b}}{1+e^{mx+b}}\right) \cdot \frac{e^{-(mx+b)}}{e^{-(mx+b)}} \\
p &= \left(\frac{1}{1+e^{-(mx+b)}}\right)
\end{aligned} \tag{12.9}
$$

This is a sigmoid function, clearly, where the switch between 0 and 1 occurs between two classes. This is considered *simple logistic regression*, and a function that considers many classes is *multi logistic regression*.

SVM.

Try to learn how to find a good hyper plane later.

Support vector machines (SVM) looks to find a line (in 2D) that best separates data. In multi-dimensions we would be looking for a hyperplane that best separates our data. Broadly, how this works is by finding a plane which separates the data and maximizes the distance between points and the plane. This is the margins of our plane—we want to maximize margins. The vectors which mark the distance between our plane and nearby points are called *support vectors*. A key source of error

with SVM is outliers. That is, outliers will greatly change how our plane is drawn.

This method is also immediately off when considering clumps of data interspersed between each other—i.e., data which is not distinctly separable. This can be aided by projecting data onto a different reference plane. For example, one dimensional data, where some class of data is clumped around 0, and another class is clumped in two groups around 10 and -10, clearly cannot have a single line separating them. However, you can transform the data, perhaps by squaring it, and therefore easily find a line separating the two. This is called the *kernel trick*.

12.1.2 Neural Networks

Neural networks describe layers of neurons, which each have inputs and various weights, which converge and produce an output. Input and output layers can be seen as a series of linear models. An activation function prevents each layers from collapsing into a simple linear model.

You'll recall that the L2 loss function is hyperbolic. Our descent down to the minimum of this loss function is called *gradient descent*. The way to descent toward our minimum loss is as follows:

$$w_{new} = w_{old} + \alpha \cdot \nabla \tag{12.10}$$

Where w is our weight. I'm writing ∇ as the descending gradient, just because I always say `grad` in my head instead of `nabla`.

12.1.3 Linear Regression

In simple li near regression takes a value of x and gives a prediction for y, in accordance with some line of best fit. It's modeled with:

$$\begin{aligned} y &= b_0 + b_1 x \\ \text{MAE} &= \frac{\sum_i |y_i - \hat{y}_i|}{n} \end{aligned} \tag{12.11}$$

Where we try to minimize the mean absolute error (MAE). Sometimes we will want to minimize the sum of the squared errors (MSE), which increases the "penalty" for outliers, so-to-speak. Finally, one can also square root this error, for root mean squared error (RMSE). Linear regression requires assuming linearity, independence, normality, and homoscedasticity (equal variance).

With multiple linear regression, we may aim for something like:

$$y = b_0 + b_1 x_1 + b_2 x_2 + ... + b_n x_n \tag{12.12}$$

One can measure the strength of prediction by:

$$\begin{aligned} R^2 &= 1 - \frac{RSS}{TSS} \\ RSS &= \sum (y_i - \hat{y}_i)^2 \\ TSS &= \sum (y_i - \bar{y})^2 \end{aligned} \tag{12.13}$$

Where we are comparing the error associated with the mean to error associated with the fit. If the fit is quite good, then $R^2 \to 1$.

12.2 Unsupervised Learning

Now, we'll move on to unsupervised learning!

K-means Clustering.

K-means clustering attempts to compute k clusters within the data. k is a predefined number. If $k = 3$, you will chose 3 random points (centroids) on the data and compute the distance from all points to each centroids. We then determine the closest centroid for each point, classify all points by their centroid, then determine the average location (or mean) of each data point per centroid. We then use these means as our new centroids and repeat the process. This allows us to progressively nail down the center of each cluster.

This process, which iterates until some stable equilibrium is reached, is called *expectation-maximization*. Then, if a new point arises, you can quickly classify it by how close it is to each of the three centroids.

Principal Component Analysis.

This is a dimensionality reduction technique. A *principal component* refers to the component or direction with the greatest variance—which will tell us the greatest about our data set. It is projecting our data onto a new dimension with high variance. This is also the line that minimizes the residuals (from our current point and the newly projected point).

Part III

Physiology

Preoccupied with a single leaf, you won't see the tree. Preoccupied with a single tree, you'll miss the entire forest.
– Vagabond by Takehiko Inoue

12.3 Perspective

Chapter 13

Biochemistry

To be able to interface with the electrophysiology of neurons, one must fully understand it.

13.1 Membranes, Ions, and Potentials

This will be a general overview of some of the electrophysiology of neurons. This should probably precede the math and models section, but here we are anyway.

Of course, the ion composition of cells varies a bit, but a good enough approximation is as follows:

$$[Na^+]_i = 15mM, [Na^+]_e = 150mM$$
$$[K^+]_i = 120mM, [K^+]_e = 5mM$$
$$[Ca^{2+}]_i = 100nM, [Ca^{2+}]_e = 2.5mM$$
$$[Cl^-]_i = 15mM, [Cl^-]_e = 120mM$$

Where $[X^+]_i$ and $[X^+]_e$ refer to the intracellular and extracellular concentrations respectively. Building up a strong intuition for electrophysiology is quite important to approaching the field. Some of the absolutely key facts to recall include:

1. **The intracellular and extracellular spaces are electro-neutral.** If this weren't the case, cell membranes would be under constant force. This means the sum of positive and negative ions must be equal between the intracellular and extracellular spaces.

2. **$[Na^+]$ and $[Cl^-]$ are high outside, $[K^+]$ is high inside.** This will tell you which directions the ions flow during an action potential—when the gates open, Na^+ wants to enter, K^+ wants to exit[35]. In other words, $[Na^+]_i < [Na^+]_e$, $[K^+]_i > [K^+]_e$.

3. **The concentration of ions is approximately static.** The amount of ions that actually flow in an action potential and or through ion channels is quite small relative to the total amount.

4. **Intracellular Ca^{2+} is extremely regulated, and its concentration is small (around 100nM).** Ca^{2+} is the ion responsible for most cellular processes, so its tight regulation is essential. In other words, $[Ca^{2+}]_i << [Ca^{2+}]_e$.

[35] A slight nuance is that this does not preclude ions from flowing against their gradient. It can certainly still happen, but the net movement will be out if channels are open in either direction.

Expanding on the calcium regulation, much of the free $[Ca^{2+}]_i$ is stored within the endoplasmic reticulum (ER), whose concentration will be many orders of magnitude higher than the cytosol. This is important for cells like myocytes who require a large influx of calcium to contract. Rather than attaining all of this Ca^{2+} from the extracellular space, the cells can used Ca^{2+}-induced-Ca^{2+}-release in order to drive the opening of the ER.

Ramblings.

As a complete aside, it is hypothesized that our cells control ions in this way because we originate from the ocean. As the ocean is salty, cells learned how to pump out as much Na^+ as possible, and based their electrophyiology off of this potential. When evolution allowed us to exit the ocean, we maintained this machinery and keep sodium in the extracellular space.

Direction of Flow.

Knowing the general concentrations of ions is a very imporant way to intuit ion behavior under different conditions. However, the Nernst Potential offers a way to quantify it as:

$$E_X = \frac{RT}{zF} \ln \frac{[X]_e}{[X]_i} \tag{13.1}$$

Where z is the charge of the ion. This allows you to solve the electromotive force, E, for your ion of interest. Knowing the E values for your ions allows you to determine the overall membrane potential using the Goldman–Hodgkin–Katz flux equation:

$$v = \frac{g_{n_1} E_{n_1} + g_{n_2} E_{n_2} + ... + g_{n_k} E_{n_k}}{g_{n_1} + g_{n_2} + ... + g_{n_k}} \tag{13.2}$$

The GHK formula is essentially like taking a weighted average of each ion's (n_i) membrane potential, giving us the total membrane voltage. The conductance, g, is the opposite of resistance (convered in more depth at section **9.1.2**). In this case, you can think of conductance as being a measure of the membrane's permeability to an ion. That is, a neuron with a lot of Na^+ channels will have a very high total sodium conductance. Conductance is measured in terms of its maximum, meaning that although sodium channels will open and close, the value g_{Na} refers to when all of them are open. When calculating membrane voltage, generally we only use sodium and potassium (though, the GHK formula technically requests using all charged molecules). The reason being because the conductance of all of the other ions is essentially negligible in comparison.

In most cases, the GHK equation is simplified to only include Na^+ and K^+. The reason is because these two ions have the highest conductances, by far, making them hold the most weight. For example, even though cells may contain HCO_3^-, there aren't tons and tons of ion channels for it like there are for sodium and potassium—thus it does not contribute to v to nearly the same degree. If this concept is confusing, simply consider V = IR. If there is no current, there is no voltage. Therefore, we can use the equation below as a good enough approximation of the resting membrane potential:

$$v = \frac{g_{Na} E_{Na} + g_K E_K}{g_{Na} + g_K} \tag{13.3}$$

Now, let's assume a membrane is at rest around -55mV and that the sodium concentrations are what we listed above (150mM and 15mM). We can use the Nernst equation as follows:

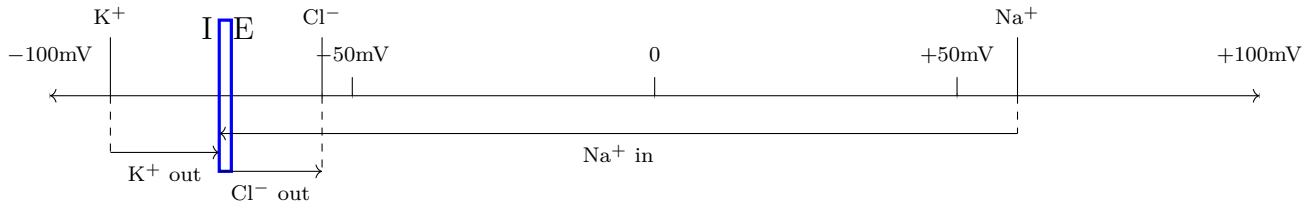
$$E_{Na} = \frac{(8.314)(310.15)}{(96,485)(1)} \ln \frac{150}{15} \rightarrow 0.062\text{V} \tag{13.4}$$

Let's assume E_K and E_{Cl} are around -52mV and -85mV respectively. Here is both a quantitative and a conceptual way to think of the direction of flow. Because 0.062V (62mV) > the resting -55mV, we can know Na^+ will (on average) flow into the cell if sodium channels open. However, if the cell were to massively depolarize all the way to 100mV, we would actually expect Na^+ to leave the cell, despite its gradient seemingly telling you otherwise. This is, of course, because ions will be subject to both their chemical and electrical gradients. To simplify:

$$\begin{aligned} v > E_K &\longrightarrow \text{outward current} \\ v < E_{Cl} &\longrightarrow \text{inward current} \\ v < E_{Na} &\longrightarrow \text{inward current} \end{aligned} \tag{13.5}$$

As a definition, recall that current (in electronics / physics) describes the flow of *positive*. Flux describes the actual flow of a molecules—so therefore, for a positively charged molecule, the current and the flux are in the same direction, and opposite for a negatively charged molecule. This means that, as the currents for K^+ is outward and Na^+ is inward, their fluxes will be the same (since they are positive). Since Cl^- is negative, and the current describes positive charge as flowing inward, it means Cl^- will have an outward flux[36].

Let us consider now more qualitatively / visually below. Note that this is not to scale, at all. The blue rectangle represents the membrane, which for this example is between E_K and E_{Cl}. Note that the right side is the extracellular space (labelled E), and the left side is the intracellular (labeled I) (this makes sense because the membrane voltage is measured as the intracellular with respect to the extracellular). Plotted are the E_x values we have.



For positive ions, drawn is an arrow from their E_x toward the membrane. That is, the arrow for E_{Na} is going from right to left, signifying sodium wants to go from the extracellular side to the intracellular. For K^+, you can see the arrow is going from left to right, meaning K^+ wants to go from the intracellular side to extracellular (i.e., out). Chloride ions, being negative, means the arrow will be switched. Thus drawn is an arrow from the membrane to E_{Cl}. In this example, Cl^- clearly wants to go out. This isn't meant to exactly replicate physiology, but rather meant more as a procedure to think about ion flux. Another minor benefit of this qualitative approach is you can see the actual magnitude of the potential. That is, in this example the arrows for Cl^- and K^+ are much smaller than

[36]Do you also find this confusing? I find this to be very easy to mess up—which is why I prefer thinking a bit more conceptually. Even when typing this up, I confused myself for a moment :)

for Na^+. Therefore, we are closer to potassium and chloride's equilibrium potentials (they're happier, so-to-speak). Sodium, on the other hand, wants to rush in very badly—and if sodium ion channels open, it will (hence the start of action potentials is via sodium influx!).

Of course, these values agree with our quantitative assessment. In short, flux will be outward for K^+ and Cl^- (again, because current describes the flow of positive charge), but inward for Na^+.

13.1.1 Action Potentials

Overview.

When considering an action potential, likely you'll want to focus on the flux of Na^+, K^+, and Ca^{2+}. The standard description is that some kind of stimuli causes an influx of positive ions into the neuron's dendrites or soma, which leads to opening of voltage gated sodium channels in the initial segment of the axon. Sodium flows inward, which causes this segment of the axon to greatly depolarize (shooting up from ≈ -55mV to $\approx +20$mV). When the membrane reaches these more positive voltages, voltage gated calcium channels will begin to open and sodium channels will begin to close. More slowly (i.e., delayed in respect to sodium channel opening) in the depolarization process, potassium channels will open which allow for the efflux of potassium ions, causing the membrane to repolarize. Similarly, the depolarization will cause the next set of sodium channels further along the axon to open, causing the depolarization to propagate down the axon. After the action potential is complete, active transport can somewhat quickly restore the ionic composition to exactly what it was before. This process is remarkably energy intense, which is touched upon in other sections (**22.1**).

A Common Misconception.

The concentration of ions really does not change during an action potential. It is only a negligibly small amount of ions that are required to change the membrane voltage of the neuron. So, for example, $[Na^+]_e$ really only goes from 120mM to 119.99...mM during an action potential. Therefore, the membrane potentials that you calculate here are roughly true at any point in an action potential. What changes is the membrane voltage of the neuron, causing channels to open and close, which is what leads to the characteristic peaking of an action potential.

This idea is quite important. Many assume that ions flowing in and out substantially impacts their concentrations, and that is what leads to the in and outflowing of ions. It is often said that ion channels always pass ions "with their concentration gradient." While "transporters" pass ions or molecules against their gradient. But this is not true! Firstly, you can imagine that if channels are unidirectional, there will always be some stochastic passage against a concentration gradient. But more explicitly, many channels solely exist to passively pass ions against this gradient. For example: there are many "inward rectifying potassium channels," meaning they pass potassium ions into the cell (a famous example being the Kir family of channels). Importantly, the concentration of K^+ outside of the cell will *never* exceed the concentration inside the cell[37]. Therefore, any and all inward pointing K^+ channels will inherently be against the chemical concentration gradient.

You may be slightly puzzled about why such a small concentration of ions is able to generate an action potential. This is because the voltage generated is hyper-localized to the membrane. The lipid com-

[37]Pretend you are not a good physiologist and are unaware of the existence of endolymph.

position of the membrane is also very strictly controlled. The hydrophobic heads of the lipids on the outside monolayer of the membrane (thus, those facing the extracellular space) are most often neutral in charge—still polar, though. Those on the inner monolayer (the cytoplasmic side) have a higher proportion of negatively charged heads. This helps create a partial membrane voltage and is incredibly important for ion behavior, as well as how transmembrane proteins orient themselves in the membrane.

Let's motivate this a little bit: the capacitance of a cell (C_m) is around $1\mu\text{F/cm}^2$, shown here[38]. If the radius of the cell is 10μm. Given that the voltage swings by around 100mV during an action potential, we can determine that:

$$\begin{aligned} C_{total} &= C_m \times 4\pi r^2 \\ q &= C_{total}\Delta V \\ q &= 1.3 \times 10^{-12}\text{C} \end{aligned} \tag{13.6}$$

Dividing this number by Faraday's constant, F, gives you the moles of charge, m. You can divide the moles by the volume to approximate ΔC (i.e., the change in total charge):

$$\Delta C = \frac{\text{m}}{4\pi r^3/3} \rightarrow 3.1 \times 10^{-7}\text{mM} \tag{13.7}$$

Don't forget to convert the volume into units of liters.

3.1×10^{-7}mM is a very small number, many orders of magnitude below the concentrations of either sodium or potassium. Therefore, in theory, to generate an action potential you'd require a nearly 0% change in ion concentration. However, what this actually describes is the *net* movement. That is, if the concentration of sodium inside the cell increases by x when sodium channels open, the concentration of K^+ must decrease by no more than $x - 3.1 \times 10^{-7}$mM to depolarization the cell by 100mV. Thus, by this logic, x can be nearly any value—and perhaps the change in sodium and potassium is actually quite large? To verify that the general idea of this exercise is true, I simulated an action potential using the Hodgkin-Huxley model to find some approximations of ion currents, and found that the sodium concentration changes by ≈ 0.002mM. This is only a small fraction of a percentage of the total concentration, so the principle holds.

13.2 Ion Channels

This section probably should also be deleted, as I ended up not being as successful as I thought in building it out. There is some interesting info in here, but I'm not sure if it justifies a whole section.

This section will be filled in better throughout the semester, through more Biochem courses.

13.2.1 Voltage Gated Channels

Voltage Sensing.

You may ask yourself "how does an ion channel know when the voltage is positive or negative?" Structure informs us of almost every puzzle in biology, and this case is no different! Let us consider

[38] https://pubmed.ncbi.nlm.nih.gov/19571202/

one potassium channel called KvAP. KvAP is hypothesized to be voltage gated due to the presence arginine side chains (as positively charged amino acid)[39]. Because of the positive charge on arginine, you can imagine that if the outer membrane voltage becomes especially negative, it will pull the arginines outward, almost like opening the lid of the channel.

Some notes:

To clean up later. 1) How membrane capacitance is measured[40]. Currents are often corrected for by capacitance in order to account for changes in conductance (i.e., channel expression).
2) Gating current refers to the current through a ion channel induced in the channel's opening (and or, while conformational changes are occuring). This sort of thing is frequently measured with channel inhibitors, therefore allowing current through only during these "phase changes." For example, one could: administer tetrodotoxin, and perform a voltage step from ≈ -75mV to 0mV, thus capturing the opening of sodium channels. Two currents will be seen, the capacitance current and the gating current. One can determine the capacitance current of the membrane by performing a voltage step from ≈ -75mV to -100mV. Thus, no voltage gated channels will be opened, but you'd see a current induced by the membrane's capacitance.

13.3 Current Literature

Octopus Sensory Receptors.

Chemotactile receptors (CR) allow cephalopods to taste via touching[41]. The key question addressed by this paper is how ion channels diverge and become specific to certain functions.

To this point, quite a bit is known about nicotinic acetylcholine receptors in the context of humans. This includes their basic function in neuromuscular junctions, structural information, where acetylcholine binds, etc. However, not much is known about octopus chemotactile receptors, nor how they arose.

Many molecules remain insoluble in the ocean, thereby requiring physical contact to sense them. This diverges from how terrestrial species detect airborne chemicals. It was known that CRs were evolutionarily related to mammalian nicotinic receptors and are both members of the cys loop family of channels, which have a large extracellular domain and underlie fast transmission. Under disruptive selection, signified by the ratio of nonsynonymous to synonymous mutations, these receptors diverge and become specialized for different molecules. Those presented in this paper have highly hydrophobic binding pockets, making them specialized for analyzing greasy materials, namely terpenes. This paper focused specifically on chemotactile receptor for terpenes 1 (CRT1) (which was the best characterized CR to this point) in order to compare it with the $\alpha 7$ nicotinic receptor (which it shared the most sequence homology to).

Their paper used two primary routes of investigation: structural studies and electrochemical studies. Using cryo-EM, it was found that CRT1 has a symmetric, homopentameric structure, (similar to $\alpha 7$ receptors). Interestingly, in both CRT1 and $\alpha 7$ are glumatate residues which are key to their calcium

[39] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1253646/
[40] https://pubmed.ncbi.nlm.nih.gov/21713564/
[41] https://www.nature.com/articles/s41586-023-05822-1

permeability's, but not their permeability of other ions. Naturally, upon identifying such conserved features, they looked to find different sites in order to address how CRT1 became specialized for ocean life. Before doing so, they applied different terpenes and found that the channel was consistently activated, but not when ACh was applied. Conversely, standard $\alpha 7$ receptors were activated by ACh but none of the terpenes. To identify the binding site of the terpenes, they tested costunolide.

Amazingly, upon structural analysis they found the channel was open but no costunolide was bound—in fact, the detergent they used to extract and purify the protein was bound, called diosgenin. As such, they suspected similar oily compounds could tightly bind to CRT1. This is the same site that ACh would bind to the $\alpha 7$ channel. The binding sites differ in that one loop, the C-loop, is shorter in CRT1, and is missing aromatic and cystine residues. Such residues are needed for ACh binding, and their absence makes the binding site flatter and more hydrophobic.

Given this surprising result with diosgenin, they tested other oily compounds and too found strong responses. In amputated octopus limbs, nerve responses were generated when in contact with such compounds. Octopus limbs have a high degree of autonomy, and react strongly to stimuli even after amputation, including moving physically away from such stimuli.

Bacteria Deliver Channels to Plants.

Bacteria use a type III secretion system, meaning it allows them to insert bacterial proteins into their host. One well conserved effector family is called AvrE, and is associated with 'water soaking' and increased pathogenesis. Water soaking refers to dark, sunken in spots often seen on plants where bacteria are multiplying. AvrE family proteins have been challening to study due to their size and high toxicity—as well as limited sequence similarity to other known proteins.

This work identified AvrE proteins, namely the DspE channel, as folding into a β-barrel, which has a porin-like structure via AlphaFold2 analysis, which was verified through cryo-EM. Prior to this work, nothing was known beyond that these proteins are likely membrane bound. Therefore, their hydrophillic interior and hydrophobic exteriors gave the first indication of pore formation at all.

They determined the channels are permeable to ions and generate currents in either direction, with a reversal potential around-25mV. Notably, canonical inhibitors did not block the channel's permeability, and replacing the extracellular ion compositions did not alter the current noticeably. The reversal potential was mildly affected when Cl^- was removed, suggesting a potential selectivity for anions.

In doing their anion experiments, they noticed oocytes swelling. Expression of AvrE proteins in *Xenopus* oocytes allows for water permeability and oocyte thus swelling, as evidenced by altering the solution's osmolarity. Because the ion's pore is so large (15Å), it seemed plausible that the channel would be permeable to other molecules larger than water. Indeed, dyles like fluorescein, but not large proteins like GFP, were able to pass through it.

In mutations deleting a large amount of the β-barrel, currents are abolished. Deletion of leucine residues around the base of the channel similarly abolish conductance. Because surface biotinylation did not detect the protein, it is suspected that these residues are necessary for membrane inclusion. Moving forward, they looked for object that might be able to fit into this wide pore. They specifically selected polyamidoamine dendrimers (PAMAM), because their characteristic length can be altered. Indeed, PAMAM G1, which has a length of ≈ 22Å, inhibited DspE 71% and AvrE 93%, as evidenced

by limited membrane current. Baseline swelling was also inhibited, as well as fluorescein uptake.

Inhibition of AvrE was associated with decreased virulence in plants.

Using FRET for VRACs

Here we discuss[42]. The suprachiasmatic nucleus (SCN), a bilateral nucleus in the hypothalamus, is responsible for maintaining the circadian rhythm. It is known that the SCN produces changes in repeat firing that varies with the time of day. It is thought that subthreshold, leak currents are responsible for changes in firing rate. Research has been done on increased K^+ conductance, but this paper postulates a greater role of subthreshold Na^+ conductance. This role of Na^+ was proposed first in *Drosophila*, so this work helps to verify conservation in mammals. They determined that sodium conductance has a greater role in the daytime than night, and that their affect depends on K^+ currents and their affect on the input resistance of the membrane.

The authors focused on three different neuron-types—NMS, GRP, and VIP-expressing cells. VIP-labeled cells were recorded from via removal and slicing either during the day or night. VIP cells were found to fire more frequently during the day than night, and their R_{in} was higher during the day than night. The depolarization (ΔV) associated with action potentials was similar at night than during the day, though. Interestingly, they found that sodium currents associated with leak channels (i.e., in the presence of TTX) was variable, but consistent across day and night. Though, when Na^+ is removed and replaced with $NMDG^+$, the resting membrane voltage was much more hyperpolarized, and this effect was observed to be stronger during the day than night. This led them to conclude Na^+ leak channels are particularly useful in maintaining day time resting membrane voltages[43].

In comparing NMS+ and GRP+ cells in the same tests, they found that NMS+ cells show higher firing rate, higher input resistance, and higher resting membrane potential during the day than night. This was not so for GRP+ cells. Na^+ current density was shown to be similar in both the day and night for NMS+ and GRP+ cells, however removing Na^+ current demonstrated a greater hyperpolarizing shift in the daytime than nighttime for both cell types. This, to them, suggested that Na^+ leak currents too maintain a higher resting voltage during the day than night across these cells types.

Ryanodine in Chemobrain

Here we discuss[44]. Chemotherapy can lead to something called "chemobrain"—a cognitive dysfunction in cancer treatment that is currently not understood. PET and MRI scans have shown structural and functional changes in the brain associated with prolonged treatment. This work determined that improper modifications to RyR2 channels leads to leakiness, thereby inducing dysfunction in metabolic signaling paths, leading to cognitive decline.

RyR channels are the largest known ion channels—on the order of 5,000 amino acids! RyR channels have 3 isoforms, with RyR2 being the most common in the brain. Using a mouse breast cancer model, their initial tests showed that treatment with doxorubicin leads to increased RyR channel opening time and open probability from signal unit recordings. They developed an orally admissible drug called S107 that is said to stabilize RyR channels. Use of this drug in conjunction with doxorubicin

[42] `https://pubmed.ncbi.nlm.nih.gov/37339878/`

[43] Is this surprising, or is this something you learn in introductory neuroscience? That's the question.

[44] `https://pubmed.ncbi.nlm.nih.gov/37756377/`

results in WT levels of openness. Via PET scans, they determined a reduction in glucose metabolism in doxorubicin treated mice, which was recovered in S107 treated mice. However, it's worth noting that these phenotypes seem quite weak (i.e., on the order of a couple percent difference). Similar effects were observed with a second chemoterapeutic treatment course called MTX + 5-FU.

RyR immunoprecipitated from whole brain lysates were compared for post-translational modifications. In order of their Figure, RyR2 shows increased phosphorylation, oxidation, nitrosylation, and decreased calstabin-2 binding. Calstabin binding is known to stablize RyR channel structure, and treatment of with S107 recovered this binding, but none of the other post-translational modifications. Through behavioral tests, like the Moris water maze, they found clear cognitive impairments in doxorubicin treated mice that were rescued with S107. In the novel object discrimination test, S107 actually seemed to improve scores above WT.

13.3.1 Piezo

Piezo channels are near and dear to my heart, as we study them extensively. Piezo channels have been implicated in nearly every cell of various organisms. There's no shortage of mechanisms hinging on Piezo channels, including neuron regeneration. Here, I'll discuss some current literature pertaining to Piezo channels.

Piezo in B Cells.

Adding yet another function to Piezo, they play a crucial role in B cell responses[45]. B cells are activated when antigens bind to B cell receptors (BCRs), primarily through surface expression via antigen-presenting cells (APCs). This induces BCR clustering, phosphorylation, and formation of a signalosome. Importantly, such responses involve significant membrane and matrix reorganization. This is marked by the formation of an "immune synapse," where BCRs cluster with other important membrane proteins at the interface of the APC. Because B cells seem preferentially activated by antigens on the surface of cells, rather than soluble antigens, it is plausible that some physical interaction is beneficial to B cell activation.

Some background work showed B cells express Piezo1 mRNA, but not Piezo2 mRNA. To explore a potential Piezo1 role, they first physically poked B cells using a micropipette and saw, via a calcium indicator, that Ca^{2+} influx occurs. In administering the Piezo1 inhibitor GsMTx4, such Ca^{2+} influx was abolished. They also imaged B cells laid out on a glass sheet. When flattening out on the glass, calcium influx occurred. To test if this was Piezo-dependent, they applied oligomerization-blocker 1 (OB1), which is said to block Piezo channels. In this case, again calcium influx and cell flattening was decreased. It's not clear to me why they used two different Piezo blockers for these experiments. Using an interesting tool called Flipper-TR, they were able to fluorescently measure membrane tension, and found that indeed tension increases when B cells flatten on glass.

To test if Piezo is required for B cell response, they bound anti-Ig to a polymer substrate and found that BCRs cluster around the interface between the substrate and the B cell. This clustering was abolished in the presence of the OB1 Piezo inhibitor. This same result was observed when using RNAi knockdown of Piezo channels.

[45] `https://www.science.org/doi/10.1126/scisignal.abq5096`

13.3.2 An Aside: Quantum Biology.

Quantum biology is an emerging field[46]. Because biology requires the aggregation of hundreds of trillions of atom, it remains unclear if quantum effects will ever be biologically relevant. So far, there are some examples, but much remain under explored or speculative in nature. All are familiar with the double-slit experiment, where individual atoms exhibit wave-like behavior. But, what of small molecules or proteins? We know that electrons can tunnel, but what of hydrogen atoms, and or, the moving of proteins in a biological context? We would expect that the larger an object, the less likely it were to tunnel. Therefore, if we took a hydrogen-transferring enzyme which appears to undergo tunneling, we could replace this hydrogen with a deuterium (a heavier isotope) and expect to see the transfer efficiency decrease—indeed, this was done for alcohol dehydrogenase and other comparable enzymes. It's not clear if this is explicitly "used" in biology, as these effects are not monstrous, and tunneling is not necessarily something that can be harnessed, but rather is a side-effect of any such small scale process. Much of this work, particularly in relation to metabolism, was done by Britton Chance of the University of Pennsylvania.

In the olfactory bulb, it has been long known that GPCRs are combinotorically activated by odorants. However, it remains unclear exactly how this occurs. The immediate inclinciation would be structural: depending on the structure of the odorant, it will bind to different GPCRs. Strangely, different odorants which have quite different structures can smell similarly, implying they activate a different array of GPCRs. An alternative hypothesis is that these GPCRs are vibrational sensors, and the molecular vibrations of distinct structures are still similar. Depending on the vibrations, the distance between the odorant and the walls of the receptors changes, allowing increased likelihood of tunneling, and therefore activates GPCRs. By the mechanism mentioned above, we would expect odorants with regular hydrogens would smell differently than those with heavy hydrogens. A *Drosophila* paper showed that, in a maze, flies were able to distinguish between these isotopes and can be trained to avoid a deuterated odorant. However, clearly there may be other effects, for example in the differences in masses may have obvious receptor-binding effects. Many conventional biologists are quite opposed to this idea.

A purely quantum effect is the spin of an electron, which can either be "up" or "down". Some are attempting to control and read electron spin, hopefully acting as a new form of electronics and or digital logic (spintronics, bad name). Interestingly, somehow different biological molecules are able to selectively allow electron current of certain spins only. Its thought that some free radical reactions might by sensitive to magnetic fields. Some even speculate that this is how birds seem to have an intuition about which way the poles are. An even more drastic theory is that these birds utilize *quantum entanglement*. It's proposed that, because electrons must occupy independent states, measuring that an electron is spin-up immediately means the corresponding electron is spin-down—thus, they are "entangled." It's suggesting that photons entering a bird's retina may generate a pair of entangled, free radicals that are shifted in state by the global magnetic field and fields of small proteins within the bird. It's thought that this may occur through the Cryptochrome protein, which generates radical pairs.

Quantum computing, using the multiple states of quantum particles, is thought to be the next wave—as theoretically, quantum bits can store many states and do multiple computations in parallel. However, a big drawback is quantum, states are quite delicate. To use a quantum computer, one must cool

[46] `https://www.youtube.com/watch?v=bLeEsYDlXJk&t=622s`

it to very low temperatures, lest the digits will be susceptible to decoherence. Remarkably, biological systems seem capable of maintaining these states. In photosynthesis, photons generate "packets of energy" called exitons that are said to hop between enzymes until they reach the photosynthesis reaction centers. Presumably, this would be a quite inefficient process which would lose energy as it randomly hops from location to location. It was proposed that exitons may instead exist in a state of quantum coherence, exploring many possible routes at once before actually moving—a mechanism similar to the ideal quantum computer.

Finally, the part you likely care about most (note that this is highly, highly speculative in nature): some speculate that the brain, and consciousness, are derived from quantum theory. Roger Penrose suspected that conscious thought arises from wave-function collapse (when one makes a measurement, they measure a singular value, while we know particles exist as smeared waves)—where all possibilities collapse into a single state, something a computer is not capable of. Electrons on microtubules may be at the root of this.

Chapter 14

Muscle Physiology

A few things: First, as it turns out, there is some incorrect info in the Silverthorn book [i.e., Dr. Delp comments that skeletal muscles have a force-stretch graph similar to cardiac muscle], especially pertaining to muscle physics. Second, it would be wise to include some more cardiac info since loss of MAP is a key cause of death after SCI.

A large part of this information comes from Dee Silverthorn's *Human Physiology* textbook—arguably the best textbook of all time.

14.1 Skeletal Muscle

14.1.1 A Cellular Level

Structure.

Skeletal muscles are composed fasicles, which are composed of muscle fibers, which are composed of myofibrils. An individual muscle fascicle is around $10 - 100\mu$m in diameter and can range from $1 - 30$cm in length. Muscles use a silly nomenclature, in which "sarco" is added to words. For example, the whole structure sits within the sarcoplasm (cytoplasm), fibers are intertwined with the the sarcoplasmic reticulum (equivalent to endoplasmic reticulum) and surrounded by the sarcolemma (equivalent to the cell membrane). An individual myofibril is made up of overlapping actin and myosin segments, held together, to some degree, by titin. The myosin heads are the canonical structure you imagine, which bud off of the end of a chain like leaves on a branch and bind to actin. Undeniably, the most uninteresting part of muscle physiology is as follows: The "centerline" of the myosin networks is called the M line, while the "centerline" of the actin network is called a Z disk. On a microscope image, the Z disks can actually be seen as horizontal stripes. Titin branches off from the Z disks to bind to the ends of myosin chains, providing both some elasticity and support for myosin. Actin chains are centered around a line of nebulin, which too provides structural support and organization. The I (isotropic[47]) band is considered to be the unbound part of the actin structure, centered around the Z disks. The H zone is the unbound part of the myosin structure, centered around the M line. The A (anisotropic) band is the entire length of the myosin fibers, thereby encapsulating both the bound part of the actin-myosin complex and the H zone, meaning it too is centered around the M

[47]Note that the isotropic and anisotropic naming comes from their initial discovery and should not be used to make assumptions about chirality, etc.

line. Therefore, when contraction occurs, the size of the H zone and I band decreases, while the size of the A band stays the same.

Contraction-relaxation.

Myosin heads desire to bind to actin, but are blocked by tropomyosin. Ca^{2+} can bind to troponin, bound to tropomyosin, to cause conformational changes resulting in the exposure of the actin to the myosin heads. In doing so, binding can occur, followed by myosin's power stroke. The energy for the stroke comes from hydrolyzing ATP, which had already occured by the time myosin attached to the actin. The powerstroke allows the release of the ADP and P_i. ATP then can bind to the empty active site of myosin, which causes the release of the head and prepares it for another cycle. Notably, when ATP does not bind to myosin, the muscle will be stuck in the rigor state.

In skeletal muscles, the source of Ca^{2+} is a combination of the extracellular Ca^{2+} flowing inward, and further release from the sarcoplasmic reticulum. The story goes as follows: a motor neuron releases acetylcholine onto the motor end plate (an area which a high density of sodium channels). This causes a depolarization, which propagates down the muscle fiber. Structures called T-tubulues sink lower into the tissue, allowing for more direct access to the inner processes (visually, these look similar to gyri in the brain or the crypts of the intestinal wall). The T-tubules are lined with dihyropyridine (DHP) channels, an L-type VGCC (specifically $Ca_v1.1$). DHP and ryanodine receptors (RyR) can be mechanically coupled, where influx of Ca^{2+} in through DHP mechanically opens RyR channels of the sarcoplasmic reticulum (a large store of Ca^{2+}). The free calcium is lowered through things like active pumping out of the sarcoplasm.

A steady supply of ATP is needed to maintain pumping, and it is said that at any given time, there is 8 or so twitches worth of ATP within the muscle fiber. Therefore, frequent production and alternate stores are required for continuous movement. One such storage is phosphocreatine, whose phosphate group can be quickly transferred to ADP through creatine kinase. Muscles therefore contain high levels of this enzyme, and **testing for it in the bloodstream can be a good proxy for muscle damage**.

Notably, it is very difficult to fully deplete a muscle of its ATP. Other forms of fatigue begin before this can possibly occur, which include CNS or PNS feedback. An example of this may be that acetylcholine is not synthesized fast enough to continually stimulate muscle fibers. Continual simulation of muscle fibers, beyond what is allowable under normal conditions, **will fully deplete ATP levels and therefore cause damage to muscles**. Another consideration is the continuous use of ATP may result in P_i buildup in the sarcoplasm, making release of ADP + P_i from myosin less likely to occur. Too, this opens the possibility of calcium phosphate forming, which can be quite damaging if it crystallizes further. There are also ion concentration changes to consider, and continued stimulation can result it tetanus.

14.1.2 Generating Force

The ratio of motor neurons to muscle fiber dictates the relationship between fine control and force generation. The nervous system is digitized, and signals are frequency encoded (see section **16**). This stands to reason for the neuromuscular system as well. You might ask yourself: how can it be that my fingers can delicately hold a pencil in my hand, but also forcefully grip the grocery bags. Twitch experiments have shown that successive stimulation, resulting in muscle twitching, can cause a superposition of muscle contraction (and thus force generated). Thus, the frequency of neuromuscular firing helps

answer our question—higher frequency of firing results in stronger contraction. The fastest twitching results in "fused tetanus" where the force generated by muscles smooths out and becomes constant.

An additional mode of force control is by size. Dictated by their ability to hydrolyze ATP faster, fast-twitch fibers are able to contract faster. Similarly, they're able to undergo contraction cycles more quickly due to their ability to pump Ca^{2+} out of the cytosol (and or into the sarcoplasmic reticulum) faster. ATP is generated in fast twitch fibers primarily through anaerobic glycolysis—which lowers the pH in the process, contributing to fatigue. Slow twitch fibers, which rely more on oxidative phosphorylation, are able to avoid this fast fatigue. Slow twitch fibers harbor more mitochondria and consume more oxygen. This oxygen demand is met by such muscles being smaller (which allows oxygen to diffuse into it more easily) and more myoglobin present in slow twitch muscles, giving them their characteristic red color. Fast twitch muscles, on the other hand, are larger and white. Interestingly, the smaller, slow-twitch fibers are recruited first in muscle movement, and then larger, fast-twitch fibers are recruited. This size-dependent recruitment is called the *size principle*. Aiding in this, smaller fibers have lower activation thresholds than do large ones.

Importantly, external muscle stimulation through external muscle stimulation bypasses this size principle, and rather typically activates larger muscles first. Because these fibers are more prone to fatigue, this can be a poor way to artificially generate muscle movement.

An EMG is one way to estimate muscle force generation. To do this, one would take EMG signal, rectify it, and then perform low-pass filtration or peak detection.

NOTE TO SELF: Add section on building EMG.

A Power Thought Experiment.

For some sense of scale: muscle contraction depends on ATP, and a single molecule of ATP generates forces on a pico-newton scale. If a single myosin head's contraction moves the fiber on the order of nanometers, then the work performed by one molecule of ATP is on the order of pN-nm. Consider briefly how many molecules of ATP, and actin-myosin interactions are required to do any basic task. The answer: quite a few.

14.1.3 Disorders Digression

There are many ways one can lose control of their muscles. Nerve damage, for example, will halt the release of acetylcholine onto the motor plate. In a similar manner, botulism is a result of the botulinum toxin blocking release of acetylcholine, resembling the effect of nerve damage. In McArdle's disease, muscles simply cannot convert glycogen to glucose-6-phosphase, causing the energy supply to be limited.

Muscular Dystrophy.

Muscular dystrophy is an umbrella term[48]. There are more than 30 types, and in modern times the lines between muscular dystrophy and other disorders is more blurred. Duchenne muscular dystrophy (DMD) is the canonical type, which presents in children. Early signs include weakness in the shoulders or hips—i.e., the larger muscles. In DMD, the protein dystrophin is absent, which would normally

[48] https://www.youtube.com/watch?v=S6gPsYVmIEI&ab_channel=MayoClinic

attach actin to the cell membrane. This can result in membrane permeability, calcium influx, and thus activation of digestive enzymes that breakdown muscle fibers.

Treatment of muscular dystrophy is centered on symptom management. For DMD, steroid usage seems to help. It is thought that muscle breakdown leads to immune responses, which can be aided by steroids. But, there is no cure, and eventually these steroids become less effective.

Multiple Sclerosis.

Multiple sclerosis (MS) is a CNS disorder, thought to be autoimmune in origin, and resulting in degradation of the myelin. It can be identified by plaques, and or, scar-looking lesions in the brain.

MS is rapidly developing. Over the course of days or weeks, one might slowly find numbness in a limb. Some light, tingling feeling is a typical first symptom. A clinician would follow-up with an MRI of the brain and spinal cord, looking for lesions within the myelin. The scarred regions are characterized in 4 ways: (1) an *active lesion*, which is acute and marked by inflammation and continual de- and remyelination, (2) an *inactive lesion*, which is chronic and marked by demyelination and limited inflammation, (3) a *mixed lesion*, where the center is primarily demyelinated, but the borders include microglia aiding in remyelination, and (4) a *remyelinated lesion* where a significant degree of healing has occurred.

The furthest downstream causes are classified as (1) a reduction in myelin during the active lesion, (2) total oligodendrocyte loss during the mixed and inactive lesions, and (3) poor differentiation of progenitor cells into remyelinating oligodendrocytes.

14.2 Smooth Muscle

Smooth muscle is much more variable than skeletal, differing by location, contraction pattern, required inputs, and structure.

14.2.1 A Cellular Level

Structure.

Smooth muscle is not considered to have sarcomeres, despite it having the same basic structural components of skeletal muscles. Smooth muscle contains much more aactin than does skeletal muscle, and notably does not contain troponin like skeletal muscle does. Smooth muscle networks are connected through intermediate filaments, which usually attach to dense bodies within the cytoplasm. Actin also attaches to dense bodies, maintaining the actin-myosin network within the cell as well. Smooth muscles do not have T-tubules like skeletal muscles. A comparable structure may be calveolae, which do indent into the membrane and seem spatially associated with the sarcoplasmic reticulum. The autonomic nervous system stimulates fibers through neurotransmitter release from varicosities, or bulbous stores of the chemical. The neurons may innervate the muscle fibers, allowing multiple muscle cells to be stimulated at once, or they may be release to a few fibers, which are connected through gap junctions and stimulate the nearby ones, causing a propagating wave to stimulate others. The first case describes a multi-subunit muscle, and the second a single subunit.

Contraction-Relaxation.

Initiation of contraction begins the same as skeletal muscle, in that calcium enters and the concentration is further driven up by sarcoplasmic calcium release. Though, in this case Ca^{2+} may enter either through gap junctions or membrane ion channels. As such, there are many more modes of entry than in skeletal muscle. For example, voltage-gated Ca^{2+} channels may open, but there are also ligand gated channels or stretch-activated channels, adding extra layers of possible regulation. Intracellular differences arise beginning from release from the SR. Firstly, it is no longer a mechanically gated RyR channel which allows its release. The release mechanism is now Ca^{2+} activated RyR release (commonly called Ca^{2+}-induced-Ca^{2+}-release (CICR)), and the IP_3 path. GPCRs activate phospholipase C, driving IP_3 production which binds to SR channels and causes them to open. The IP_3 path is usually considered the greatest way to drive up intracellular Ca^{2+} (or at least, that is what computational biologists seem to think). When Ca^{2+} is available, it binds to calmodulin (CaM), which then binds to the myosin light chain kinase (MLCK). MLCK phosphorylates myosin to increase myosin ATPase activity. Importantly, once contraction occurs, it stays stiff until released by a different mechanism. Because after contraction, no work is being done in the stiff state, smooth muscle is able to stay contracted for long periods. This explains why sphincters in the body are able to stay closed all the time, while one's bicep fatigues after carrying groceries for just a little while. Relaxation begins when Ca^{2+} is either pumped out of the cell through a Ca^{2+}ATPase pump, or sodium transporter. This causes CaM to unbind, myosin light chain phosphatase (MLCP) to dephosphorylate myosin, and the myosin heads to release from actin. Interestingly, diacylglycerol (DAG), another product of the IP_3 path, inhibits MLCP and thereby enhances muscle contraction.

The calcium stored in the SR is maintained in a number of ways. One example being the protein STIM1 responding to lower Ca^{2+} levels within it, moving toward the cell membrane, and activating store operated Ca^{2+} channels, such as Orai1.

14.3 Cardiac Muscle

14.3.1 A Cellular Level

Structure.

Cardiac cells contain sarcomeres, but the fibers themselves are smaller and often mono-nucleated. Cell edges are called *intercalated disks*, held together by desmosomes and permeated by gap junctions. The role of gap junctions is obvious but essential, and key to the rhythmic firing of the heart. Despite not requiring nervous system input, cardiac muscles have large T-tubules, and a significant amount of mitochondria.

Contraction-Relaxation.

Initiation of cardiac muscle contraction requires extracellular calcium, and generation can be graded. The entrance of Ca^{2+} powers the opening of non-mechanical RyR channels, as a method of CICR. Because of this, the sarcoplasmic reticulum is smaller in the myocardium, because some of the Ca^{2+} required for contraction comes from extracellular entry in about a 10:1 ratio of sarcoplasmic reticulum:extracellular origin. Ca^{2+} is removed using the Na^{+}-Ca^{2+} exchanger (NCX).

The electrophysiological profile is marked by Na^{+} entry but maintained by Ca^{2+} in a plateau phase.

The fall occurs via K^+ entry. The refractory period is long due to this plateau phase, which prevents Na^+ channels from resetting.

Pacemakers.

Hyperpolarization cyclic nucleotide (HCN) gated channels induce an inward hyperpolarization (I_h, or I_f) current. Within the hyperpolarization range, HCN channels open and cause steady depolarization rhythmically. HCN channels are typically permeable to multiple cations, and in this case Na^+ is the biggest player. When the membrane potential is high enough, VGCCs will open and cause depolarization required to propagate the signal between myocytes. At peak, slow opening K^+ channels open and repolarize the membrane. Notably, voltage gated Na channels do not play a role in myocyte depolarization.

Expectedly, as β-blockers will inhibit a rise in cAMP, HCN channels will be less active. This decreases heart rate, among other things.

Chapter 15

Biomechanics of Movement

Much of this will come from Dr. Scott Delp's online Biomechanics course[49]. Secondarily, it will come from Dr. Lawrence Rome's Systems Physiology course.

Introduction.

The study of movement is the study of one's physical abilities, the physical abnormalities that arise from disease, the loads places on the body during activity, and the ways in which we can lighten these loads and correct movement faults to restore one's normalcy. The design of external robotic prosthesis, implanted joint replacements, the reading and interpreting of EMG signals, and more are all at the core of biomechanics.

Broadly speaking, movement is produced first by neural commands initiating muscle-tendon dynamics, which produce forces that are integrated to produce velocities and movement.

15.1 Walking and Running

15.1.1 Walking

Walking Definitions.

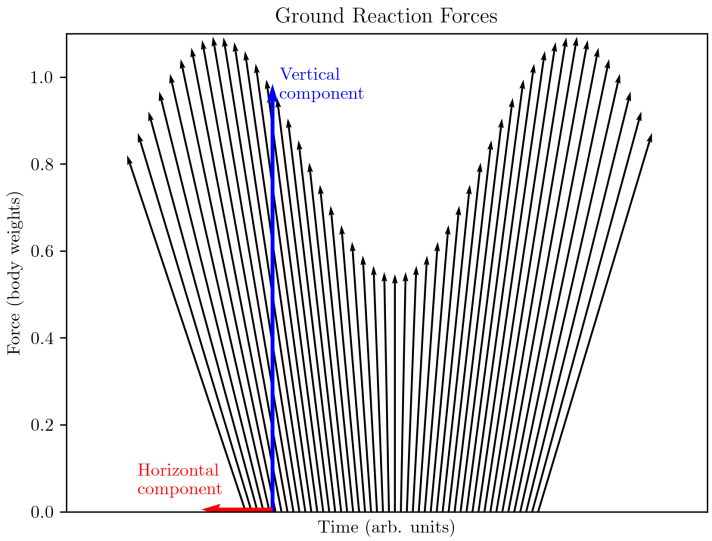
To begin, the walking "gait cycle" is composed of a *stance phase* (beginning when the heel touches the ground and ending when the toes lift off) and a *swing phase* (while the foot is off the ground). The stance phase is longer than the swing phase, making up approximately 60% of the gait cycle (which makes sense, as there are times in walking where both feet are on the ground (called *double support*). Naturally, as one's walking speed increases, the length of the stance phase decreases, and so too does the period of double support (which transitions eventually from walking to running).

The *cadence* is the frequency of steps, the *step length* is the distance traveled by a single foot, the *stride length* is the total distance per cycle (2× the step length), the step width is the distance between feet (perpendicular to the direction of walking), and the *toe-out angle* is the angle of the foot with respect to the direction of walking. The direction of walking is sometimes referred to as the "line of progression." One's step width is a good proxy for balance troubles.

[49] `https://www.youtube.com/@biomechanicsofmovement3388`

Walking Forces and Energetics.

The "ground reaction forces" (GFR) is the reaction force the ground exerts in response to one's walking. That is, you take a step and push down on the ground, and the ground pushes back up on you. The ground reaction forces begin when the heel makes contact with the ground, and increase as your weight is loaded onto the ground. In this loading phase, the ground reaction forces will be slightly backward pointed, as your heel makes contact with the ground at an angle. When the foot is flat on the ground this force will be vertical, called the midstance, and is lower due to the double support. At this point, GFRs are lower than one's bodyweight. Then, in moving toward push-off, they begin to ramp again and have a forward trajectory as your toes push off the ground again at an angle. Finally, the GFR scale down again at toe-off, the point at which your toes leave the ground. See a sketch below:



You can take this vector of ground reaction forces and separate it into components. The horizontal component modulates the speed of forward movement, and the vertical component how much one is held up. Plotting these forces individually reveals how loaded our joints are. You find that the vertical component actually exceeds body weight during the loading and push-off phases of walking, but is below body weight in the midstance phase. Similarly, we see that the horizontal component is negative (a deceleration) in the loading phase and positive in the push-off phase (an acceleration).

Walking is extremely energy efficient. Given Newton's equations:

$$\begin{aligned} \sum \mathbf{F}_v &= ma_v \\ \sum \mathbf{F}_h &= ma_h \\ K_E &= \frac{1}{2}mv^2 \end{aligned} \tag{15.1}$$

We can approximate the kinetic energy of walking as:

$$K_E = \frac{1}{2}m\left(\frac{\int F_h}{m}\right)^2 \tag{15.2}$$

And it is easy to see from measuring the horizontal and vertical force components that there is a cyclic back and forth between kinetic and gravitational potential energy—that is, a great amount of gravitational potential energy is converted to forward movement, helping to make walking quite energy efficient.

Walking Models.

The simplest conceivable model is considering the limbs to be inverted pendulums. In 1980, Thomas McMahon developed the ballistic walking model. In this model, we consider the stance limb to be an inverted pendulum that is fully extended (i.e., it is a rigid rod). The swinging limb is modeled as a double pendulum (i.e., one about the hip, one about the knee). The next major innovation came in 2010 with the Dynamic Walking Model by Kuo and Donelan, which added feet and allowed for step-like movements.

15.1.2 Running

Walking vs. Running.

Running, unlike walking, has the addition of a *flight phase* where neither foot is on the ground. The ground reaction forces in running do not ride, fall, and then rise again as they do in walking. Instead, trend toward a single peak, which is much higher than that of walking. The exact shape of the running GRF curve varies slightly in how one runs, and their terrain. For example, if one strikes with their heel first, the horizontal component of the GRF will begin negative, and cause the total GRF to have a fast rise and then a slight dip (when the horizontal component switches from negative (heel impact) to zero (midstance) to positive (toe-off)). The speed of the rise is modulated similarly, and by how stiff of a surface the impact occurs on. This is discussed further in a bit.

Interestingly, while walking has kinetic and potential energy out of phase, they are actually in phase for running. Rather than "storing" energy as potential energy, instead it is cyclically stored as elastic energy in our muscles and tendons.

We can model running as a simple, one spring system. Such a system would have a natural oscillation frequency ($f = \sqrt{k/m}$). This implies that runners have some natural frequency, and that they can modulate their stride length more so than frequency.

Consider again the simple inverted pendulum model. In a literal sense, the switch from walking to running occurs when the centripetal force exceeds that of gravity. That is, there is a velocity where the mass atop the pendulum escapes the downward pull of gravity. This occurs at $v^2/l > g$, where l is the length of the pendulum.

The gait transition from walking to running is driven by energetics. That is, there is some point at which increasing speed via walking has a higher energetic cost than converting to running. The energetic cost per unit distance is called the *cost of locomotion*, and the cost per hour is simply one's metabolism. Walking speed naturally settles to one's minimum cost of locomotion (which happens to be around 1m/s).

Spring Running Model.

A large problem in biomechanics is injury prevention. The advent of injury in running is largely due to how much we load our musculoskeletal system. Faster running imposes a higher load. Running speed is simply stride frequency (ω) times stride length (L). A slower running speed, say around 2m/s, may maximize the load at 2× body weight. This increases with the faster one runs. Rather than modulating the peak force to prevent injury, it would be better to modulate the slope of the load's rise. That is, if you decreased the maximum force, you would also be slowing down running speed, so this provides a beneficial alternative. The way one's foot lands on the ground affects this rise—landing on the front of the foot gives a shallower load slope than landing on the heel. However, one question to address is how a track can be optimized to minimize injury but maximize running speed.

The simplest running model would include a single spring representing the limb. An updated version was proposed by McMahon, which included a spring representing the ground landed on and a spring plus dashpot for the limb (giving 2 degrees of freedom total). It is governed by the equations:

$$\begin{aligned} F &= k\Delta x \\ F &= b\dot{x} \end{aligned} \tag{15.3}$$

Where b is the damping coefficient of the dashpot. At the junction between the limb (spring and dashpot) and the track (spring), the forces must be balanced, meaning:

$$\begin{aligned} F_{\text{spring}_\text{t}} &= F_{\text{spring}_\text{l}} + F_{\text{dashpot}_\text{l}} \\ k_t\Delta x_1 &= k_l\Delta x_2 + b_l\dot{x_2} \\ k_t\Delta x_t &= k_l\Delta(x_l - x_t) + b_l(\dot{x_l} - \dot{x_t}) \end{aligned} \tag{15.4}$$

Where x_t is at the interface of the track and the limb, and x_l is the top of the limb. We can suppose that the limb supports some mass at point x_l, giving us a total force for the mass of:

$$\begin{aligned} \sum \mathbf{F}_{\text{mass}} &= m\ddot{x_l} \\ m\ddot{x_l} &= k_l\Delta(x_l - x_t) + b_l(\dot{x_l} - \dot{x_t}) \end{aligned} \tag{15.5}$$

We neglect explicitly adding gravity because, in this one example, it would simply be a constant offset to the forces.

It makes intuitive sense that a harder track would provide runners with a greater speed, as less compliance allows one to exert a greater force on the ground. However, by analyzing these kinematics, you can find that there is a region of track stiffness where one's speed actually increases (as computed by the time one's foot spends in contact with the ground). Indeed, this increases stride frequency, and thus speed. Further, a compliant track also modulates stride length.

Upon finding the ideal tuned region, McMahon convinced the Harvard athletics department to build a track with this ideal compliance. Such a track improved runner speed and decreased injury rate.

Some Conclusions.

We discussed the transition from walking to running, and the ground reaction forces of running and walking. This implies a connection between leg length and beginning to run, as well as the forces required in running. Therefore, we can wonder if larger animals are capable of running. Indeed, large animals like elephants do not seem able to run. Further, this implies quite large animals like dinosaurs definitely could not run.

15.2 Muscles and Tendons Modeling

Muscles.

An important thing to recall is that muscles pull, not push. The *active* force-length properties can be plotted, and you'll find that maximum force correlates with the amount of crossover between actin and myosin—which occurs at some optimized sarcomere length. There is also a passive contribution, which is due to the structural proteins like titin. When stretched beyond normal limits, proteins like titin act like springs. Therefore, summing the active and passive contributions shows a force-length curve that is almost bimodal (up to a point, where the muscle would simply snap).

Muscle also has a force-velocity curve. An elongating muscle (eccentric) has a negative velocity, and a shortening muscle (concentric) has a positive velocity. The amount of force you can generate depends on this velocity. At peak velocity, when shortening, the force you can generate is significantly lower than the isometric force you can generate (when length is constant). When the muscle is lengthening, the maximum force you can generate is higher than this isometric force. Think of lifting a dumbbell during a bicep curl[50]. One can lift a light dumbbell quite quickly, and a heavy dumbbell much more slowly (where the muscle can exert a greater force). When the dumbbell is too heavy to lift, one can still resist its falling, and when it gets even heavier, you can extend your bicep slowly (a negative velocity) where force is maximized.

Mechanical power is the product of force and velocity. There is some maximum mechanical power generated, which is where we naturally attempt to operate. For example, when shifting the gears on your bike, you settle to some region that maximizes power[51].

Tendons.

Tendons too have a force-length curve. It is unimodal and nonlinear, where the elongation of a tendon recruits more tendon fibers, eventually reaching maximal extension, at which point some begin to give, and the tendon eventually snaps altogether. A tendon's properties are given by:

$$\begin{aligned} \text{Tendon stress} = \sigma^T &= \frac{F^T}{A^T} \\ \text{Tendon strain} = \epsilon^T &= \frac{l^T - l_S^T}{l_S^T} = \frac{\Delta l^T}{l_S^T} \end{aligned} \tag{15.6}$$

Where the stress in a tendon is the force it experienced over the area, assuming the force is uniform. The strain is the stretch of the tendon over it's slack length (where force begins to develop in the

[50]Example from Dr. Delp.
[51]Example from Dr. Delp.

tendon) divided by it's slack length.

15.2.1 Muscle Architecture and Modeling

Two components of a muscle are its pennation angle and its physiologic cross-sectional area (PCSA). A muscle can either be parallel (where its PCSA is perpendicular to the muscle) or pennate (where its PCSA is "featered," or angeled relative to the muscle fibers). The section option allows for more muscle fibers to act in conjunction. This degree of pennation varies greatly throughout the body. The deltoid, for example, has multiple pennate groups together (multi-pennate), whereas some muscles in the quad are strictly long, parallel fibers. The length of the tendon also varies greatly. For example, the calves (specifically the gastrocnemius) have short, pennate fibers and long tendons.

We can approximate the force a muscle can generate as:

$$\begin{aligned} F_0^M &= \text{PCSA} \cdot \sigma_0^M \\ \text{PCSA} &= \frac{V_M}{l_0^M} \\ l_0^M &= n \cdot l_0^S \end{aligned} \tag{15.7}$$

Where F_0^M is the max force a muscle can generate. σ_0^M is the peak stress of a muscle (specific tension of a muscle). The PCSA can be determined by the volume of a muscle (V_M) divided by its optimal muscle fiber length (l_0^M). The optimal fiber length is computed by the number of sarcomeres (n) times the optimal sarcomere length (l_0^S).

In humans, the optimal sarcomere length is approximately 2.7μm. Therefore, one could simply count the number of sarcomeres—but this would be extremely time consuming. Instead, one can dissect out the muscle fiber and image to find the length of sarcomeres. Using this method, we instead calculate l_0^M as $l^M \cdot l_0^S / l^S$, where l^M is the fiber length and l^S is the measured sarcomere length.

Therefore, the force-length curve has a strong effect on a muscle's operable range. That is, 20 sarcomeres in series result in a longer muscle fiber with a wider force-length curve. 10 in series results in a shorter force-length curve. They both, however, have the same maximal force—because the number of sarcomeres in parallel did not change. This similarly modulates the shortening velocity of the muscle. More fibers in series can result in faster shortening. Again, though, peak force does not change. Increasing pennation will increase the maximum force that a muscle can generate, due to an increase in PCSA. This will come at the expense of its operable range, which will be shorter. Further, it will have a slower peak velocity.

A Dimensionless Muscle-Tendon Model.

The force in muscles, F^M, can be computed with parameters $a, \tilde{l}^M, \tilde{v}^M$, and or its activation, normalized length, and normalized velocity respectively. Such values are normalized as l^M/l_0^M and $v^M/v_{\max}^M$. Activation is also normalized between 0 and 1, where 1 is a maximally active muscle.

$$\begin{aligned} F^M &= f(a, \tilde{l}^M, \tilde{v}^M) \\ &= F_0^M \cdot \left(a \cdot f^l \left(\tilde{l}^M \right) \cdot f^v \left(\tilde{v}^M \right) + f^{PE} \left(\tilde{l}^M \right) \right) \end{aligned} \tag{15.8}$$

Where F_0^M is the maximum force for the muscle. The first term is the active term, and the second is the passive (which is independent of activation).

The above equations are dimensionless, and thus can be scaled to different systems. Notably, if simulating an actual muscle contraction, one would also have to consider the dynamic tendon length. In a static case, this is not complex. However, in an orchestra of many muscles working together, with variably compliant tendons, it becomes much more complicated. Generally, we would scale the tendon's properties to the size and strength of a muscle—that is, because a stronger muscle will require a larger tendon.

Modeling the properties of all muscle-tendon interactions in the body therefore requires the: active muscle force-length curve, the passive muscle force-length curve, the active force-velocity curve, and the tendon force-length curve. The complex-specific values we need are the optimal fiber length, the peak isometric force, the pennation angle, and the tendon slack length.

15.2.2 The Musculoskeletal System

The *moment* is the force generated about point 0. Suppose we take the force $\vec{F}$ acting from point e_1 to e_2. The space from point 0 to e_1 transverse by vector $\vec{R}$. The moment is then defined as the cross product $\vec{M}_0 = \vec{R} \times \vec{F}$. The *moment arm*, r, is the perpendicular distance from the line of action to point 0. For example, the moment arm in the $\hat{z}$ direction is given by:

$$r = \frac{M_{0,z}}{|\vec{F}|} = \frac{\vec{R} \times \vec{F}}{|\vec{F}|} \cdot \hat{z} \tag{15.9}$$

These moment arms can be approximated or imaged via MRI (that is, identifying the point of a joint and it's distance to the tendon).

To take an example: suppose we are statically holding a dumbbell. the location of point 0 is somewhere in our elbow. We can easily compute the torque generated on our arm as dW, where d is the length from 0 to the dumbbell, and W is its weight. We can know that the muscle must exert some force F^M to support this weight, and the sum of the moments must equal zero in this static case. Thus:

$$\begin{aligned} \sum \mathbf{M}_0 = rF^M - Wd = 0 \\ F^M = \frac{Wd}{r} \end{aligned} \tag{15.10}$$

Again, where r is a distance. We can also show that:

$$r = \frac{\Delta l^{MT}}{\Delta \theta_{\text{rad}}} \tag{15.11}$$

Where l^{MT} is the muscle tendon length and θ_{rad} is the joint angle. Intuitively, this implies that muscles with larger moment arms undergo large stretches as the joint angle changes. You can think about the implications of this for the force-length curve of a muscle—a muscle with a short force-length curve should therefore not have an extremely long moment, as it may stretch too much.

Because the moment arm r is likely to be significantly smaller than the distance d (let's say they are a factor of $10\times$ off), we can know that the force generated by the bicep must be $10\times$ the weight of the dumbbell. This signals the mechanical advantage of the muscle, which may be quite low.

To do this calculation requires assuming the muscle acts in 2D and that the joint is effectively frictionless—both of these assumptions are good (joints have extremely low friction). It also requires assuming muscle contraction is from one point e_1 to another point e_2, which may bring rouble in some instances. For static moments in 3D, we can consider each direction individually. That is, the moment in the $\hat{x}, \hat{y}, \hat{z}$ direction's contributions.

Forces in Muscles.

The forces generated therefore are not constant, and depend on the muscles force-length curve which changes with joint angle. Similar.y, the moment arm changes. Therefore, the peak moment occurs at the product of the peak moment arm and peak force. Thus, you can measure the moment generated my a muscle-tendon-joint complex, but the peak of this measure is not necessarily when the muscle is exerting its maximum force.

15.3 Quantifying Movement

First, to begin, the key reason for quantifying movement is providing a basis or source of data to reverse calculate some of the values noted above.

Optical motion capture.

The gist of how one would capture movement is as follows:

A camera exists at some point 0 in the $(\vec{e}_x, \vec{e}_y, \vec{e}_z)$ space. Typically, the $\hat{z}$ component is referred to as the principal axis, and or, the one down the lens of the camera. The object one wants to capture is at some point e_1. Notably, the distance from point e_1 to the camera is generally quite difficult to calculate, so a second camera is helpful for this. Similarly, some objects in the world should remain fixed for calibrating the cameras.

The cameras will track markers placed on the body, and use this formation to describe how the musculoskeletal system is moving. However, a large source of error is called *soft-tissue error*—the markers sit atop skin (or clothes), which may shift in harsh movement, and thus not give exact measures of the underlying joints. Higher quality data can be achieved by attaching markers directly to the bone (via screwing in or some other form of clamp), but this may not be doable in many cases.

Force Plates.

A force plate quantifies the force generated on a plate in the ground, and is used to track the ground reaction forces of movement. These devices may employ piezoelectrics, where a force is converted into electricity. However, the placement of the reading device will not be exactly aligned with the center of pressure directed by the foot.

We can solve for the center of pressure by finding the location of the normal force (F_N, which has only a $\hat{z}$ component). Therefore, if we measure the moment M at location $(0, 0, 0)$ and center of pressure

at point r, we can use the following equations:

$$\begin{aligned} M &= r \times F + F_N \\ M_x &= r_y F_z - r_z F_y + F_{N_x} \\ M_y &= r_z F_x - r_x F_z + F_{N_y} \\ M_z &= r_x F_y - r_y F_x + F_{N_z} \end{aligned} \tag{15.12}$$

$$\begin{aligned} M_x &= r_y F_z \\ M_y &= -r_x F_z \\ M_z &= r_x F_y - r_y F_x + F_{N_z} \end{aligned} \tag{15.13}$$

$$\begin{aligned} r_x &= -\frac{M_y}{F_x} \\ r_y &= -\frac{M_x}{F_z} \\ F_{N_z} = F_N &= M_Z - r_x F_y - r_y F_x \end{aligned} \tag{15.14}$$

15.3.1 Transformations

This is basic linear algebra, but I suppose it is worth reviewing regardless.

Suppose we have two reference frames, which are defined by markers on the body. Let us call these reference frames A_0 with coordinates a_x, a_y, a_z, and B_0 with coordinates b_x, b_y, b_z. We will want to know what a point in the A_0 space is in reference to B_0. The vector mapping the origins of A_0 and B_0 can be called $\vec{P}^{A \to B}$, or:

$$\vec{P}^{A \to B} = \begin{bmatrix} \vec{P}_x \\ \vec{P}_y \\ \vec{P}_z \end{bmatrix}_A \tag{15.15}$$

To rotate the frame of reference, we use the rotation matrix $R^{A \to B}$, a 3×3 matrix.
The complete transformation matrix, then, is:

$$\vec{R}^{A \to B} = \begin{bmatrix} & R^{A \to B} & & P^{A \to B} \\ & & & \\ 0 & 0 & 0 & 1 \end{bmatrix}_A \tag{15.16}$$

15.3.2 Computing the Moments

To compute these moments, we will need ground reaction forces and motion data. Let us begin at the foot.

At the foot, we will consider two points, let us call them (x_t, y_t) and (x_a, y_a) for the toes and ankle respectively. We'll make the simplifying assumption that the foot is massless and does not accelerate, and that the forces on the ground are conveyed only through the toes. We can know at the sum of the forces at point $t_{x,y}$ and point $a_{x,y}$ must be equal, because the foot is not accelerating. Therefore:

$$\begin{aligned} \sum \mathbf{F}_x &= F_{x_t} - F_{x_a} = 0 \\ \sum \mathbf{F}_y &= F_{y_t} - F_{y_a} = 0 \end{aligned} \tag{15.17}$$

We will calculate the moment about the ankle, which has some torque τ_a. If the ankle is at height h above the ground, and the length of the foot is l, we can know that the moment:

$$\sum \mathbf{M} = F_{x_t} h - F_{y_t} l - \tau_a = 0 \tag{15.18}$$

The signs are a bit funny and arbitrary—it depends how we define direction. Here, F_{x_t} and F_{y_t} will be known from our measured ground reaction forces. Thus, we can construct the solvable system of equations:

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} F_{x_t} \\ F_{y_t} \\ \tau_a \end{bmatrix} = \begin{bmatrix} F_{x_a} \\ F_{y_a} \\ F_{x_t} h - F_{y_t} l \end{bmatrix} \tag{15.19}$$

Now, let us move on to the shank. Again at the ankle we have forces F_{x_a}, F_{y_a}, τ_a. At the knee we have F_{x_k}, F_{y_k}, τ_k. The center of mass (COM) of the shank is important, now. We can find the location of it by using the ankle as our reference point, where the length to the center of the shank is r_s. Therefore, the angle between the shank and the x plane, θ_s can be used to find the location in the $x - y$ referential as:

$$\begin{aligned} x_{s,\mathrm{COM}} &= r_s \cos\theta_s \\ y_{s,\mathrm{COM}} &= r_s \sin\theta_s \end{aligned} \tag{15.20}$$

We can easily take the derivative and second derivative to find the velocity and acceleration of each point, giving these equations for the forces:

$$\begin{aligned} \sum \mathbf{F}_{(x,y)} &= m_s(\ddot{x}, \ddot{y})_a \\ F_{x_a} - F_{x_k} &= -m_s r_s \left(\ddot{\theta}_s \sin\theta_s + \dot{\theta}_s^2 \cos\theta_s\right) \\ F_{y_a} - F_{y_k} - m_s g &= -m_s r_s \left(\ddot{\theta}_s \cos\theta_s + \dot{\theta}_s^2 \sin\theta_s\right) \end{aligned} \tag{15.21}$$

Giving:

$$\begin{bmatrix} 1 & -1 & 0 & \\ 0 & 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} F_{x_a} \\ F_{x_k} \\ F_{y_a} \\ F_{y_k} \end{bmatrix} = \begin{bmatrix} -m_s r_s \left(\ddot{\theta}_s \sin\theta_s + \dot{\theta}_s^2 \cos\theta_s\right) \\ m_s \left(r_s \left(\ddot{\theta}_s \cos\theta_s + \dot{\theta}_s^2 \sin\theta_s\right) + g\right) \end{bmatrix} \tag{15.22}$$

Blah, blah, you can add a bunch more etc.—will fill in later.

Chapter 16

Sensory Processing

Somatic senses are picked up by receptors on *primary sensory neurons*, whose cell bodies reside in the dorsal root ganglia and project to CNS interneurons. Such interneurons are considered *secondary sensory neurons* and they location depends on function. Fast traveling senses, like proprioception, are sent to the medulla before reaching interneurons (cross the body's midline in the brain), while slower traveling senses, like temperature, synapse onto interneurons upon enterring the spinal cord (immediately crossing the midline). Neurons of the thalamus are considered *tertiary sensory neurons* and project to the somatosensory cortex.

Convergence and Lateral Inhibition.

Sensory neurons have some range of stimuli that will activate it. The easiest examples are touch receptors. That is, individual sensory neurons cover some range on your skin. However, many primary sensory neurons may converge onto a single secondary sensory neuron. Therefore, when multiple stimuli are initiated within this range, it is perceived as one touch (because only a single downstream neuron, the secondary sensory neuron, is activated by it, despite their being multiple signals inputted). A high degree of convergence (i.e., many primary neurons synapsing onto a single secondary neurons) means there is a wide receptive field.

Lateral inhibition is when a sensory neuron becomes activated (usually a secondary sensory neuron) and inhibits the surrounding neurons from transmitting signals. That is, imagine a pin prick: the neurons at the center of the pin prick inhibits the response of the surrounding neurons, thereby allowing your body to precisely locate this fine stimuli. This helps discriminate between stimuli even better, as well as in precise perception and edge detection of stimuli.

Adaptation and Frequency Encoding.

Adaptation refers to how a signal alters in response to a stimuli. Tonic and phasic receptors differ in how they response to stimuli. Tonic sensory neurons are those that continually fire in response to some stimuli. Phasic are those that respond when the stimuli changes (a phase change). Phasic receptors are therefore considered fast adapting.

Frequency coding is the conversion of some graded potential into some corresponding frequency of action potential firing. That is, a long, strong stimuli manifests as many action potentials, resulting in heightened neurotransmitter release. Conversely, a quick, weak stimuli may not initiate an action

potential at all, and if it does it will be only one, resulting in minimal neurotransmitter release. In a sense, this helps digitize neural signals (i.e., they become binary 1s or 0s by neurons either firing or not firing).

16.1 Various Senses

16.1.1 Touch

16.1.2 Temperature

16.1.3 Pain

Nociceptive neurons respond to a variety of stimuli that can cause damage to cells. Afferent pain signals are carried either by myelinated Aδ and unmyelinated C fibers. Aδ carried pain is faster and perceived as sharp, local pain. C fiber carried pain is less sharp and less localized.

Transient receptor potential (TRP) channels are the most canonical class of ion channels, and are responsible for a great deal of pain sensation. Vanilloid ($TRPV_1$) channels respond to heat, while TRPM8 respond to cold. An immediate response to cell damage is lysing, causing marked increase in K^+ in the surrounding areas. Other released chemicals include histamine and prostaglandins. Upon activation, nociceptive neurons can trigger two responses. The first being reflexes that are integrated at the spinal cord, and the second being somatosensory cortex integration[52].

As an aside, the idea of referred pain is: pain is perceived as being from different body parts because they share the same neural paths, but does not necessarily originate from that part of the body. The canonical example is a heart attack percieved as pain in the left arm.

16.1.4 Olfaction

Information is transmitted directly from the olfactory epithelium to your brain. This occurs first through the olfactory bulb (where secondary sensory neurons are present), to the olfactory tract, and to the olfactory cortex.

Primary olfactory neurons contain GPCRs (one type per neuron). In total there are ≈ 350 olfactory GPCRs (350 per parent, which theoretically means up to 700 total, if you assume mutations cause functional divergence). As you should know, GPCRs use secondary messengers like cAMP. cAMP can be used to open ion channels and thus allow for depolarization. Different primary sensory neurons, which express the same GPCR, synapse onto the same secondary sensory neuron in the olfactory bulb. The way that different chemicals bind to different GPCRs generates different arrangements of neuron firing, so they can be combinatorically encoded.

16.1.5 Ears in Hearing and More

The Auditory Path.

The outside of the ear is called the pinna, and is helpful in capturing sound. Sound travels down the ear canal until it reaches the eardrum (tympanic membrane). Beyond the tympanic membrane

[52] This ended up being much less interesting to write than I thought it would be.

is considered the middle ear, where the malleus (hammer) vibrates the incus (anvil), which then vibrates the stapes (stirrup). These three bones, called ossicles, help amplify sound / vibrations. This middle ear area is connected to the throat through the Eustachian tube (which allows pressure to be maintained between the spaces of the ear). The stapes is connected to the oval window, which marks the beginning of the inner ear and or cochlea.

The cochlea has 3 layers: the vestibular duct, the cochlear duct, and the tympanic duct. The oval window interfaces with the vestibular duct. Vibrations propagate through the cochlea and cause movement in the basilar membrane (between the cochlear and tympanic ducts). The vibrations are dissipated through the tympanic duct, which connects to the round window. A key component of this system is that the cochlear duct houses endolymph, while the other two house perilymph. Perilymph is standard, and has a composition similar to most interstitial / extracellular fluid. Endolymph's composition is more similar to the intracellular fluid. Finally, at the apex of the cochlea is what's called the helicotrema. On a microscale, the cochlear duct houses the organ of corti, which is where the cells that detect sound are housed. Between the tectorial (top) and the basilar (bottom) membranes are layers of hair cells which are used to sense sound by the movement of such membranes.

Sound travels down the ear canal, vibrate the tympanic membrane, and moves the ossicles. The ossicles amplify sound onto the oval window, which transmits it into the cochlea and activate hair cells. Inner hair cells, of which there are $\approx$ 3,500, transmit information to the brain by depolarizing nearby sensory neurons. Outer hair cells are helpful in amplifying the response of inner hair cells by physically magnifying the membrane displacement. There are 3 rows of outer hair cells, and there is $\approx 3,500$ outer hair cells per row too—so there's about $3\times$ as many outer haircells as there are inner.

As the basilar membrane moves up or down, hair cells move and deflect. Because they're tied to the tectorial membrane above, their movement causes pulling on the tips of hair cells, which opens mechanically gated ion channels. When deflected in a way that increases the angle between the stereocilia and the cells (i.e., generating more force on the mechanically gated channels), they open and cause the cell to depolarize. Conversely, when the angle lessens (lessening the force on mechanically gated channels), they hyperpolarize. Stereocilia are linked together via tip-links, allowing for the pulling of one to pull, and thereby open, many others. Recall that the endolymph is high in potassium, so it is potassium that enters and causes depolarization.

We can hear a range from about 20 to 20,000Hz, depending on where the vibrations occur in our ear. This is called tonotopic arrangement[53]. The innermost segment of the cochlea (near the apex / helicotrema) senses longer wavelengths, while the outermost part senses higher frequencies. The basilar membrane is also much thinner and wider closer toward the helicotrema, while it is much thicker and narrower closer to the entrance. This allows certain frequencies to preferentially excited different segments along the cochlea, and is key to our perception of sound. Note that this scale is not linear—it is logarithmic (think of decibels).

Balance and Such.

recall that it is indeed endolymph in this system, as it connects directly to the cochlear duct. They detect rotational acceleration (form equals function) via ampulae, which contain cristae, which contain

[53]There's a super great lecture by Dr. Ian Shipsey. He went deaf, got a cochlear implant, and had his hearing restored. He discusses the science and limitations of cochlear implants. Link: `https://www.youtube.com/watch?v=aecxK2gFBRY&ab_channel=OxfordUniversityPhysicsSociety`

cupula. They detect rotational acceleration via deflection of the cupula, which contains hair cells and thus deflect and respond similar to inner / out cells.

The utricle and saccule contain macula, which detect linear acceleration and gravity. They do so via these little crystal structures called otoliths, which are slightly denser than the endolymph and move around in response to motion. Their sinking or deflection of the membrane is particularly helpful in detecting things like gravity.

Many years ago, NASA sought out deaf employees for their early space programs—specifically looking for those with damaged vestibular systems. Those selected were typically deaf from childhood spinal meningitis, which left their hair cells damaged. Therefore, the hair cells in the vestibular systems made them incapable of feeling motion sickness. This allowed scientists to test the effects of high g on their bodies, without them feeling sick. They are colloquially known as the Deaf 11.

16.1.6 Vision

The cornea is the outermost layer, the pupil changes in size to let in more or less light, the iris is the muscle which opens and closes the pupil, the lens is adjusted in order to focus light onto the back of the eye, the retina is at the back of the eye and where light rays hit, the macula / fovea is where the highest density of cones are. The aqueous humor is a fluid surrounding the lens, and which provides nutrients to the lens (because the lens does not have any blood vessels).

The right visual field of both eyes going to the left visual cortex, and the opposite for the left visual field. This information is transmitted down the optic nerves, intersects at the optic chiasm, continues down the optic tracts, is trafficked through the lateral geniculate nucleus [probably don't need to know that] of the thalamus, and onto the visual cortex in the occipital lobe.

As briefly mentioned earlier, the macula is the central region where cones are, and at the center of that is the fovea. Rods are more densely populated in the outer regions of the retina. Where the optic nerve leaves the retina, there is a blind spot. This is intuitive, because there cannot be rods / cones right at the "exit" of the eye (recall that the photoreceptors are technically beneath the nerves). At the kind of "base" of the retina, i.e., the innermost layer, is the pigment epithelium. This is useful in absorbing light to prevent scattering and in providing nutrients to the photoreceptors. The next layer is the photoreceptors, i.e., the rods and cones (which differ in both structure and function—rods literally look like rods, and cones literally look like cones). Photoreceptors transmit to interneurons (bipolar cells), which then transmit to retinal ganglion cells. Such transmission is under considerable control by horizontal cells (which are at the synaptic junction between photoreceptors and bipolar cells) and amacrine cells (at the junction between bipolar and retinal ganglia). Such cells are helpful in lateral inhibition.

To understand how photoreceptors operate, first consider rods at rest: cyclic guanine monophosphate (cGMP) levels are high. cGMP binds to cyclic-nucleotide gated (CNG) channels, allowing for influx of positive ions, maintaining a relatively high membrane potential. This leads to the tonic release of neurotransmitters. The structure used to sense light is called rhodopsin, a GPCR in the photoreceptor membrane. It is composed of opsin, a protein, and retinal, a compound derived from Vitamin A. Retinal is able to change conformations (from *cis* to *trans*) upon being hit by light, which alters opsin's structure too. When "activated by light," opsin decreases production of cGMP through the g-protein

transducin. The key oddity here is that neurotransmitters are released in the absence of light, but not when light is present. The difference between rods and cones, in this regard, is that cones express different proteins (pigments) other than rhodopsin that respond preferentially to different wavelengths of light.

Humans can detect a very small amount of photons and so too can we in bright light. This is achieved in a number of ways, including that the pupil can adapt (let in more or less light), we can rely more so on cones (high light levels) or on rods (low light levels). Rods in particular can converge on a single retinal ganglion, which is helpful in "amplifying" the response to a light stimuli. Further aiding in our visual perception is on and off center reginal ganglia, two modes of detecting light (or the absence of it). On center describes when light levels are high at some central point, so on center ganglia are active, while off center are inhibited. Off center describes when surrounding light levels are high, but not necessarily at the center. Thus, on center ganglia are inhibited, while off center ganglia are active. The difference lies in how they respond to the neurotransmitter released by photoreceptors (glutamate). Because photoreceptors are depolarized in the absence of light, these off center ganglia fire in response to glutamate, while on center ganglia are inhibited by glutamate. Thus, when a photoreceptor becomes hyperpolarized in the presence of light, on center ganglia can be active due to loss of glutamate, and off center become inactive for the same reason. Off center cells are, in a sense, detecting darkness, and or the lack of stimuli. For example, walking through a doorway into a dark room from a well lit room. You can clearly detect that the upcoming room is dark, because off center ganglia are active.

Chapter 17

The Central Nervous System

The Meninges.

The layers of the meninges are: the dura mater, the subdural space, the arachnoid membrane, the subarachnoid space, and the pia mater, which all sits beneath the skull. The dura mater is the thickest of the three. The subarachoid space has a spongy texture and houses the flowing cerebrospinal fluid. The bottom layer, and the thinnest, is the pia mater. The pia mater is surprisingly tough—if you've ever done brain dissections before, it feels similar to a thin film around the brain. A key role of the meninges is to cushion the brain.

READ THIS: `https://journals.lww.com/onsonline/fulltext/2016/06000/anatomy_of_the_spinal_meninges.10.aspx`

17.1 Spinal Cord General Structure

A cross section of the spinal cord would reveal meninges just like the brain and skull; an outer layer of dura, then arachnoid, and finally pia mater. Too, there is an outer layer of white matter followed by a grey matter interior, which is centered around a "central canal." The spinal cord has 33 vertebrae, beginning with cervical (C_1 to C_7), then thoracic (T_1 to T_{12}), then lumbar (L_1 to L_5), then sacral (S_1 to S_5), and finally one coccyx (Co_1 to Co_4). The sacral, and especially the coccygeal, vertebrae are fused, so you may see them depicted as a single unit. Adding slightly to the confusion is that there are only 31 pairs of spinal nerves. These are C_1 - C_8, T_1 - T_{12}, L_1 - L_5, S_1 - S_5, and Co_1. **Note: The difference in the nerve location and vertebrae labels is actually quite important, as it is not always that a fracture at some vertebrae leads to nerve damage at the same site**—this is expanded on in later chapters. The spinal nerves exit the spinal cord on either side, i.e., either dorsal or ventral roots. Dorsal entry neurons carry sensory information to the CNS, while ventral exit zones carry information from the CNS to the muscles. Though, importantly, their axons can not necessarily be found on either side, this will depend on where crossing over occurs. The secton of grey matter which connects to the dorsal root is called the dorsal horn, and the same is true for the ventral horn. The lateral horn is in between the dorsal and ventral horns.

Note to self: It might be fun to add a more detailed breakdown at some point, such as the general arrangement of all the nuclei in the spinal cord.

Another note to self: It would be good to integrate more clinical outcomes into this. For example, what happens in the case of SCI at various vertebrae, or in vagotomy, etc.

17.2 Autonomic Nervous System

Pathways of the autonomic nervous system require two neurons, one that originates in the CNS and terminates at a ganglion, and a post-ganglionic neuron which terminates at the tissue of interest. While the circuit is often considered to have only two neurons, in fact many pre-ganglionic neurons synapse onto many post-ganglionic neurons, which means one neuron can affect many target tissues. Adding to this is the mode of release. The connection between neuron and target tissue is called the *neuroeffector junction*. At this junction, transmitters are secreted indirectly from bulbous varicosities into the interstitial space, meaning neurotransmitters can diffuse over a larger area. Fascinatingly, these neurotransmitters are actually often synthesized within the axon / varicosities of neurons.

17.2.1 Sympathetic Nervous System

Pre-ganglionic neurons of the sympathetic nervous system originate in the hypothalamus or reticular formation and extend down to the mid-sections of the spinal cord, from neuron pairs T_1 to L_2. Sympathetic ganglia are directly beside the spinal cord in a long chain (often called the sympathetic chain). Generally speaking, neurons closer to the T_1 section correspond to organs higher in the body, such as the heart or lungs, while neurons closer to the L_2 section corresponds to those lower in the body, like one's reproductive organs[54].

Pre-ganglionic neurons release acetylcholine onto nicotinic receptors, and post-ganglionic neurons release norepinephrine onto adrenergic receptors. As all adrenergic receptors are GPCRs, their cellular response will be slower than ion channel transmission and occur at the protein-level primarily. Too, they will occur through secondary messenger, for example through the IP_3 path or cAMP. Norepinepherine is a tyrosine derivative and, like other catecholamines, can be broken down either by monoamine oxidase (MOA) in the neuron's mitochondria, requiring re-uptake, or in the liver by catechol-o-methyltransferase (COMT).

Dysregulation of the sympathetic nervous system most often manifests as cardiovascular pathologies, such as low blood pressure. Though, incontinence or impotence may also occur. On a cellular level, a process called *denervation hypersistivity* may also occur, where in response to decreased sympathetic neurotransmitters, the body increases expression adrenergic receptors. This is easily measured as increased response from small doses of norepinephrine.

Adrenal Medulla Digression.

It's worth noting that the adrenal medulla secretes epinephrine in a hormone-like manner. However, the adrenal medulla is often considered a collection of pre-ganglionic neurons, meaning epinepherine acts globally like a hormone, but is also a neurotransmitter.

Let us look at the different adrenergic receptors[55]:

[54] Memorizing which nerves respond to which segments is likely not a good use of your time at this stage..., although it might have some clinical use to know which organs may be affected after SCI.

[55] This is ripped directly from Silverthorn.

Adrenergic Receptors			
Receptor	Location	Sensitivity	Effect
α_1	Most tissues	NE > E	Increase Ca^{2+}
α_2	Gut and pancreas	NE > E	Decrease cAMP
β_1	Heart and kidney	NE = E	Increase cAMP
β_2	Select smooth muscle	NE < E	Increase cAMP
β_3	Adipose	NE > E	Increase cAMP

17.2.2 Paraympathetic Nervous System

Parasympathetic neurons originate from the uppermost and lowermost sections of the spinal cord, either leaving directly from the cranial nerves or from the S_2 to S_4 nerves. Ganglia in the parasympathetic nervous system are close to the target tissue, meaning post-ganglionic neurons are very short compared to the pre-ganglionic ones. One of the key paths being from the vagus nerve (a cranial nerve), which carries the majority of parasympathetic signals to organs including the heart, lungs, liver, stomach, intestines, and pancreas. Like the sympathetic nervous system, pre-ganglionic neurons release acetylcholine onto nicotinic receptors. Notably, parasympathetic post-ganglionic neurons also secrete acetylcholine, but onto muscarinic receptors instead.

Cholingergic Receptors		
Receptor	Location	Effect
N_N	Postganglionic neurons	Opens cation channels
N_M	Skeletal muscle	Opens cation channels
M_1, M_3, M_5	Target tissues	Increase Ca^{2+}
M_2, M_4	Target tissues	Decrease cAMP, open K^+ channels

Exceptions.

There are some exceptions to these rules, such as some sympathetic postganglionic neurons that terminate on sweat glands secrete norepinepherine rather than acetylcholine. These are called "sympathetic cholinergic neurons." There are also neurons which secrete none of these, are are called "nonadrenergic, noncholingeric neurons." There are also neurons which secrete multiple types of neurotransmitters.

17.3 Somatic Nervous System

The somatic motor division is exclusively excitatory, and neurons project from the CNS all the way to muscles. Inhibition can occur at the neuronal level (i.e., motor neurons are inibited) but not at the muscle level. Cell bodies of motor neurons are either within the ventral horn of the spinal cord or within the brain. Close to the muscle itself, the axons branches and connects to the muscle at the neuromuscular junction (NMJ), which includes both the target muscle and Schwann cell projections. The "synapse" equivalent is called the *motor end plate*, which is an area on the muscle with a high density of nicotinic receptors. The ECM contains acetylcholinesterase, responsible for breaking down acetylcholine. Notably, the nicotinic receptors on muscle cells are considered N_M, a slightly different

version than is found on neurons (N_N). This difference has proven to be curical in examples such as α-bungarotoxin, which binds only to N_M[56].

17.4 Acetylcholine, Nicotinic Receptors, and Muscarinic Receptors

Acetylcholine (which I'll call ACh for this section) is unquestionably one of the most important molecules in the body. It is sythesized from choline (a hydroxyl azanium) and acetyl-CoA. Typical usage goes as follows: ACh is released from vesicles and binds to cholinergic receptors. Upon unbinding, acetylcholinesterase (AChE) breaks it down into acetate and choline. Choline is brought back into the pre-synaptic neuron through sodium cotransporters so that it may be reused. It is then re-combined with acetyl-CoA and repackaged and vesicles.

Funnily enough, nicotinic receptors (nAChRs) have been described as the most well understood membrane receptor[57]. Nicotinic receptors are fast opening non-specific cationic channels. They are expressed throughout much of the major structures in the brain, peripheral nervous system, and skeletal muscle. As such, drugs targeting these receptors are of considerable interest in many neurological disorders. The binding site for ACh is composed of aromatic amino acids, namely W and Y. This paper describes continual exposure to drugs treating nAChRs as causing their eventual desensitization, while exposure to nicotine increases their expression greatly.

Both of these are of key interest in treating Alzheimer's disease, as loss of cholinergic synapses is one of its features (including significant reduction of both muscarinic and nicotinic receptors). Interestingly, Alzheimer's shows varying loss of nicotinic receptor subtypes across the brain. There is no shortage of drugs currently in development for reversing this.

Myasthena Gravis Digression.

Myasthena Gravis, an autoimmune disorder, is an example of an actylcholine dysregulation caused muscle disorder. As a digression within a digression, autoimmune disorders most often target endocrine organs, and the belief is that they are to protect against mutations causing hypersecretion[58]. Upon first glance, you may think that Myasthena Gravis evades this generality, but in fact it does not, as it is well associated with patients that dually have a thymoma (tumor of the thymus). Immune cells begin attacking cells with ACh recepetors, tagging them with antibodies. Therefore, Myasthena Gravis presents itself usually in the weaker muscles, such as those that control the eyes, as they have less ACh receptors and therefore a diminished response to signals. Thus, patients may have drooped eyelids (ptosis), double vision (diplopia), or trouble following moving objects with their eyes. The disease worsens with increased activity, as this causes more antibodies to be released onto ACh receptors—but improves on rest. Fascinatingly, men and women "get" the disease at different times (women typically under 40, while men over 60).

Antibodies may also target other genes (such as MUSK or LRP4) which are involved in ACh receptor localization, or other forms of regulation.

[56] Dee Silverthorn quite possibly wrote the greatest textbook known to humanity, didn't she?

[57] `https://www.nature.com/articles/nrd2927`

[58] `https://www.cell.com/immunity/pdf/S1074-7613(20)30180-1.pdf`

17.5 Cerebrospinal Fluid

17.5.1 Structure Overview

Cerebrospinal fluid is derived from the choroid plexus, which exists in two main places in the brain (toward the center of the brain in and in the brainstem) along the ventricles. The ventricles, importantly, is derived from the hollow space of the neural tube. The choroid plexus is derived from ependymal cells. It is composed primarily of transport epithelium with many capillaries, which should be unsurprising since a key role it has is to transport ions, which draws out fluid into in the brain / spinal cord. After being secreted into ventricles, cerebrospinal fluid (CSF) flows into the subarachnoid space and eventually out of the arachnoid villi into the venous sinus. CSF is composed similar to the extracellular space. When CSF flow is obstructed, it can cause hydrocephalus.

When patients have hydrocephalus, the surgical intervention is fairly simple: you take a tube and connect it from the brain ventricles into the stomach to help drain the CSF. Interestingly, hydrocephalus is becoming a growing interest in neuroscience because there is some evidence that hydrocephalus can present like a neurodegenerative disease in older patients. I.e., it can cause people to have fading memory or general cognitive decline—so it is hypothesized that some patients may be misdiagnosed as having diseases like Alzheimer's when in reality they have hydrocephalus.

17.5.2 CSF Contacting Cells

This topic may appear niche, but it is indeed of considerable interest! To start, epidural electrical stimulation (EES), which is covered in considerable detail later, does not stimulate specific neurons but rather depolarizes the entire CSF. Therefore, the cells in direct contact with the CSF may be the key modulators in restoration and regeneration.

To begin, let us discuss this paper[59]. CSF contacting neurons (CSF-cNs) are described as being chemosensitive, mechanosensitive, and may even act as stem cells. As mechanosensors, they inhibit motor neurons (i.e., in normal walking they are used to prevent over-compression of the spinal cord). Canonically, they are ciliated and function in tandem with the Reissner fiber to sense spinal curvature. The Reissner fiber is a protein fiber that extends along the spinal cord, and its breakdown is a known cause of scoliosis in model organisms[60]. CSF-cNs cells modulate excitatory interneurons and reticulospinal neurons. Because of their location, CSF-cNs cells are very difficult to study. They are seated in the ependymal layer of the central canal. The Review cited above discusses their investigation primarily in zebrafish. An important advancement in their study was in discovering the marker polycystic kidney disease 2-like 1 (PDK2L1), which labels all CSF-cNs.

Given their hair-like structures that extend into the apical side of the central canal, their discoverors compared them to hair cells of the ear (though, these hairs are disordered, unlike inner or outer cells). Interestingly, though, these cells are GABAergic, and thus inhibit rather than excite. In spinal curvature tests, to the surprise of the experimenters, only those that were on the concaved side of the spinal cord responded, meaning they response to compression. Restated, those cells on the left side of the body respond to bending of the spinal cord toward the left side.

[59] `https://www.nature.com/articles/s41583-023-00723-8`

[60] Very interestingly, no studies have been done to show the existence or lack of a Reissner fiber in humans.

Remarkably: PDK2L1 is an ion channel with a quite high conductance (hundreds of pS) and the high membrane resistance of CSF-cNs neurons (on the order of GΩ)... the opening of a single PDK2L1 is able to depolarize a cell! It was demonstrated that mechanical force opens PDK2L1 channels, and that in the absence of such channels, CSF-cNs do not respond. Another responder to mechanical force seems to be acid-sensing ion channels (ASICs), which is indirectly affected when an ASICs modulator is altered by force. An unanswered question was what PDK2L1 is actually responding to, as the response did not correlate with the calculated CSF flow. Instead, evidence suggests it is coupled to the Reissner fiber. Structural analysis revealed vesicles on both the apical and basolateral sides, suggesting they can secrete both neurotransmitters, namely GABA, onto neurons or peptides into the CSF. Notably, after SCI it is said that CSF-cNs secrete monoamines. While neurotransmitter containing vesicles have been shown to release after a single depolarization, the root of apically released vesicles is not known.

CSF-cNs-connected neurons ascend ispilaterally a couple hundred microns. Those that are particularly rostral may connect directly to the midbrain. Driving CSF-cNs also modulate activation of the central pattern generator, suggesting neural connection there as well. Their connection to V2a spinal interneurons is not especially conclusive yet, but is quite a promising possibility. Similar connections have been shown with reticulospinal neurons and propriospinal neurons.

Fasinatingly, despite CSF-cNs being inhibitory, optogenetic stimulation of them can lead to locomotion in zebrafish—indicating they have a multiplexed role. This data seems quite contradictory and unclear, as contrastingly, somatostatin expressing cells (specifically Sst1.1) are said to synapse onto excitatory interneurons and inhibit them. Sst knockout zebrafish exhibited longer durations of spontaneous locomotion. Their ability to promote locomotion seems much less supported than their ability to inhibit it.

As a lasting note, CSF-cNs show characteristics of immature neurons, and indeed seem to be able to differentiate into astrocytres or oligodendrocytes. Similarly, they seem to be able to recover after injury, and this was shown in a SCI model with EES.

17.6 Myelin in Health and Regeneration

Here we'll discuss the role of microglia in regulating myelin[61]. Microglia are the macrophages of the CNS, and are derived from erythromyeloid progenitor cells. During embryogenesis, they enter the CNS. Notably, they have a long lifespan in non-pathogenic environments. Microglia respond to CNS damage via receptors for damage-associated molecular patterns (DAMPs) or pathogen-associated molecular patterns (PAMPs) and proliferate. Interestingly, microglia phagocytose oligodendrocyte progenitor cells before they mature, preventing myelination during development which helps "organize" myelin wraps depending on neural activity. After injury, microglia are activated to clear debris and force an environment in which oligodendrocyte progenitors can migrate and re-wrap neurons. Perhaps more fascinatingly, upon aging, microglia lose the ability to prune myelin sheaths, resulting in hypermyelination that in turn becomes degenerative.

Discussing such matters in the conext of degenerative diseases is some of the most helpful. For example, multiple sclerosis is touched on a bit in section **14.1.3**, and Alzheimer's is discussed in depth in the

[61] https://www.nature.com/articles/s41577-023-00907-4

Addendum (Part **VII**, which is supposed to be a secret). Of course, in MS the topic of myelin is core, but demyelination is a common feature of many degenerative diseases, including AD. Fascinatingly, changes in the myelin / white matter structures can be seen decades before AD symptoms arise. Such regions are called white matter hyperintensities (WMHs) and can be seen by MRI—and splotches in the parietal lobe are especially predictive of AD. Though, it is presently unclear if axonal changes precede such myelin changes.

17.6.1 Initiaion of Demyelination

The factors leading to demyelination are largely unknown. In the case of MS, it is thought to be an immune response which triggers myelin degradation. However, it is also hypothesized that transcriptional profile alterations lead to oligodendrocyte instability, perhaps predisposing them to degradation. Similarly, hypermyelination precedes demyelination in many cases, and is thought to be triggered by an autophagy pathway[62]. In a mouse AD model, preceding a period of degeneration was months of oligodendrocyte growth, in which myelin sheaths were thicker than wildtype neurons. Follow-up work showed both de- and remyelination occur throughout the pathology's progression, but in the end demyelination prevails. Inhibiting demyelination via promoting oligodendrogenesis through inhibition of oligodendrocyte muscarinic receptor 1s seem to improve the cognitive impairments associated with AD. If interested, muscarinic receptors are discussed throughout section **17.2**. These authors suggest monitoring lipid profiles as an early diagnostic tool of AD.

17.7 Central Pattern Generation

Defining the central pattern generator is still somewhat contested. Computational models have been used to explore its existence[63], and for now we should take a few unifying assumptions: **(1)** the central pattern generator exists and is used to generate rhythmic moving, such as walking, **(2)** in many cases it is initiated by the CNS, but aside from that is largely devoid from CNS input, and **(3)** the CNS can work in tandem to compute integrated moving that requires coordination and balance.

[62] This is actually interesting. I wonder if this is linked to the broad endocrinology hypothesis regarding quickly growing cells

[63] `https://www.nature.com/articles/s41598-021-91714-1`

Part IV

Spinal Cord Injury

Maybe it sounds like I'm just spouting moral platitudes. But from a vagabond like me, it's not that. I can't begin to tell you how lonely I feel when I come across a beautiful view, then suddenly realize there's no one to enjoy it with me.
– Musashi by Eiji Yoshikawa

Chapter 18

The Injury Itself

Spinal cord injury is composed of the primary injury, prototypically, but not restrictively, due to some kind of high impact action. This is usually unpredictable and highly variable. Secondary injury, resulting from inflammation, oxidative stress, and other biological responses is much more predictable and potentially lends itself better to therapeutic intervention.

The lesion's composition is categorized in three ways: **(1)** the non-neural core, **(2)** the astrocytic scar around the core, and **(3)** the spare reactive neural tissue. In the mix of immune cell influx and scar formation, no neural cells can survive at the center of the lesion. On a neuronal level, the rostral end retracts in a process of Wallerian degeneration. The caudal end dies away. Growth from the cell body is limited both by the damaged cell's biochemistry and by the physical barriers which now present themselves in front of the axons. The physical barriers that immediately succeed injury are often called damaged axon-glia complexes (AGCs). Discussed further later, immune cell influx causes astrocytes to form a scar, meant to save the spare surrounding neural tissue, which is composed of both glia and neurons.

The traditional aim of treatment is to bridge the corticospinal tract with distant neurons through a therapeutic combination of inhibiting anti-regenerative and promoting regenerative factors. As I have commented many times, getting neurons to regenerate alone is insufficient in many cases, as reformation of the correct synapses will not necessarily follow. Forcing axon regeneration alone is, incidentally, not too hard—one can pump neurons full of metabolites or simply implant stem cells. The issue being that they do not know where to grow to. One possible route to solving this is remodeling neural circuits using interneurons to bridge these connections. There are also attempts to use biomaterials to simulate a pro-regenerative environment, hopefully enhancing plasticity of the circuits.

18.1 Cell Specific Responses

The discussion, for the moment, will mostly use information gathered from[64].

18.1.1 Immune Response

As SCI breaks the spine-blood barrier, influx of immune cells can cause further damage. Evidently, the nature of the immune response being helpful or harmful is still largely contested.

[64] https://www.nature.com/articles/s41392-023-01477-6

Neutrophils.

Neutrophils compose part of the immediate response to injury, which are recruited by cytokines and chemokines secreted by cells damaged in the primary injury. They essentially initiate the secondary injury, and reach their peak around 1 DPI. Like most cells, the role of neutrophils cannot be characterize as solely pro- or anti-regenerative. While a high influx of neutrophils is associated with poor patient outcomes, so too are neutrophils associated with guiding macrophages to damaged tissue, suggestive of better recovery.

Microglia.

In mouse SCI models, it seems that there are two peaks of microglial activity. The time course is remarkably long and disparate, reported 7 DPI and 60 DPI. Microglia can either promote inflammation, thereby worsening the secondary injury (called the M1 phenotype) or decrease inflammation, and promote repair (called M2). It is likely that this response depends on the subtype of microglia, which varies depending on the environment. Regardless, it is true that the earlier one treats SCI, the more likely one is to avoid negative microgial effects. Fascinatingly, in a neonatal setting microglia are able to heal SCI almost entirely through their role secreting fibrinogen, which is able to connect damaged axons back together.

Macrophages.

Macrophages are considered to be the dominant immune cell located around the injury site. Microglia, conversely, are scattered around the borders of the injury. Depending on the type of glial scar that is formed, different types of macrophages have been found. Macrophages mediate the corralling[65] of cells around the injury site. The phagocyotitic abilities of macrophages are of key importance, as loose fragments of cells must be removed, and microglia are incapable of keeping up such a high demand for removal. As a large part of this includes the destroyed mylein of oligodendrocytes, macrophages uptake great amounts of lipids. This can result in the formation of lipid droplets, which causes macrophages to become "foamy." This foamy phenotype impairs further repair.

Lymphocytes.

The adaptive immunity is fairly universally regarded as harmful to regeneration (with some exceptions, of course). T cells further break down the spine-blood barrier and increase immune cell invasion. Evidently, T cell entry is also a major sourse of neuropathic pain in SCI patients. In general, it seems established that the overall immune response after injury impairs further regeneration, and a good example of this can be found here[66]. NOTE: It would be good to read this entire article.

18.1.2 Neural Response

Apoptosis of neurons initiates at around 4 h AI, but peaks at only 8 h AI[67].

[65] Corralling is a term used to describe the formation of a barrier around the injury, preventing further injury. It is composed astrocytes, and other cells, and is important in repair.

[66] `https://www.science.org/doi/full/10.1126/science.abd5926`

[67] It is worth noting that these times are likely quite inaccurate, or very injury-type specific. But, they do give a good indication of the approximate timeframe—such as that the majority of this apoptosis occurs within the first day or so after injury.

Interneurons.

Fascinatingly, it was shown that the ability of neonatal mice to fully recover from SCI was due, in part, to interneurons maintaining excitatory conditions. In adult mice, these interneurons switch to inhibitory after SCI, which dampens signals to motor neurons. A paper investigating this can be found here[68]. One must wonder, can you electronically mimic the excitatory interneurons in fully grown mice? Similar approaches have been done therapeutically, such as with potassium-chloride cotransporter-2 (KCC2) agonist CLP290[69] which seems to dampen the overexcited, inhibitory interneurons. Perhaps one could simply use something like DBS (or in this case, DSS?) on these interneurons.

Astrocytes.

Astrocytes, being the dominant supportive cell, plays an essential role in SCI. After injury, astrocytes form a physical barrier that is supposedly intended to limit the secondary injury. The old perspective was that astrocytes, like the fibrotic scar, impair recovery as a physical barrier. It is now more accepts that the astrocytic scar (or border) is a necessary mechanism of limiting inflammation. This occurs after astrocytes become activated, and are helpful in the initial stages but later form a glial scar, impairing regeneration. Astrocytes may either be activated by inflammation, causing them to be neurotoxic (called A1 cells, containing a C3 marker) or by ischemia, causing them to be neuroprotective (A2 cells, lacking a C3 marker). The first transformation occuring through the NF-κB path, and the second through STAT3.

As a brief digression, let us discuss this recent work which identified a molecular switch in astrocyte neurotoxicity[70]. The group then asked if one could inhibit the production of neurotoxic or promote neuroprotective astrocytes. The work used a mouse optic nerve crush model and performed single cell RNAseq. Interestingly, after injury the neuroprotective cells were found in a more proliferative state. Because cAMP is a strong regulator of the cell cycle, they tested and found that soluble andenyl cyclase (sAC) is a regulator of the astrocyte cell cycle. Specifically, deletion of sAC increases the proportion of astrocytes escaping the G1 phase. sAC is said to act upstream of p21, which induces cell-cycle arrest. Knockdown of sAC in astrocytes increased retinal ganglion death after injury, suggesting knockdown increased the proportion of neurotoxic astrocytes.

Microglia are said to be the greatest contributor to activating astrocytes through release of signaling molecules. Another important factor is type 1 collagen upregulation, which results in astrocytic adhesion through cadherin, causing activation and eventual scar formation.

The astrocyte scar is surprisingly thin, only a few layers of cells. Though, its importance is not to be underestimated. When ablated, mice with SCI were worse off by almost every metric. A cornerstone paper on the topic seems to be here[71]. While scar tissue is primarily astrocytic in origin, it is worth mentioning that pericyte derived scar tissue (sometimes called the fibrotic scar) too play a role. Their positive roles include boosting tissue integrity, but so too do they seem to block axon regeneration as a physical barrier.

[68] https://www.nature.com/articles/s41593-022-01067-9
[69] https://www.sciencedirect.com/science/article/pii/S009286741830730X
[70] https://www.nature.com/articles/s41586-023-06935-3
[71] https://www.nature.com/articles/nature17623

Oligodendrocytes.

Oligodendrocytes reportedly begin apoptosis around 1 DPI and it peaks around 8 DPI. Oligodendrocyte precursor cells (OPCs) are may differentiate into oligodendrocytes or Schwann cells after SCI. OPCs have been show to remyelinate neurons after SCI, but fascinatingly, it has been shown that locomotor recovery after SCI does not necessarily require remyelination by oligodendrocytes[72]. Though, plenty of other research suggests it is required—so it is likely context dependent.

18.1.3 IL-12 Sensing Neurons Promote Neuroprotection

Here I'll comment on this article[73], which does not directly describe spinal cord injury, but still broadly covers neural damage, inflammation, and the immune response.

Multiple sclerosis (commented on in section **14.1.3**) is characterized by inflammatory lesions, induced by type-1 immune response. The type-1 immune response is typically responsible for protection against bacterial or viral infections. Interleukin-12 (IL-12) is a necessary activator of type-1 immunity. With IL-23, a complex relationship is formed resulting in the regulation of inflammation. Though, in the CNS, IL-12 has been known to reduce inflammation and protect against worsening MS. The protective benefits of IL-12 is the topic of this work.

Using the *CMV* promoter, driving *Cre* expression in all tissues, this group generated global knockdown of the IL-12R subunit B2, dubbed $Il12rb2^{del/del}$. Such mice were shown to be suseptible to autoimmune encephalomyelitis. They postulated that this was due to altered T cell activity, since T cells are the primary respondents in autoimmune encephalomyelitis. However, after generating T cell-specific knockdowns—they found that the mice were still suspetible to the disease. The same was true for NK cell-specific knockdown.

In order to determine which cells express the IL-12 receptor, they used both immunostaining for β-gal in $Il12rb2^{LacZ/LacZ}$ mouse lines, as well as RNAscope to detect *Il12rb2* mRNA. They identified the IL-12 receptor (IL-12R) as being expressed in natural killer (NK) and T cells, in addition to oligodendrocytes and neurons. Specifically, they found expression in NeuN-positive and $Rbfox3/Map2$-positive cells (neurons) and myelin forming and mature oligodendrocytes. They verified this expression profile in deceased patients who had multiple sclerosis. Ablation of IL-12R in neurons and glia by $Nestin^{Cre}$ worsened the response to experimentally induced encephalomyelitis. This phenotype was marked by increased $CD4^+$ T cell infiltration into the CNS and accompanying microglial activation. Finally, they demonstrated neurons are the key actors by performing glia-specific knockdown, and demonstrating a lack of inflammation.

To understand the transcriptional response to IL-12, they performed RNAseq showed that IL-12R induced a strong response particularly in granule cells, excitatory neurons, cholinergic neurons, microglia, molecular layer interneurons (MLIs), and MOL1 and MOl2. Particularly in grandule cells, responses include increased expression of genes involved in neuroprotection and decreased transcription of genes involved in immune cell recruitment. In short, they revealed a novel role of IL-12 in neurons, rather than immune cells, which mediates a neuroprotective response.

[72] `https://www.nature.com/articles/s41467-018-05473-1`
[73] `https://www.nature.com/articles/s41593-023-01435-z`

Chapter 19

In The Clinic & Therapeutic Approaches

19.1 Clinical Presentation

Death rates from SCI are still as high as 20% in some countries. There is a strong age dependence to this, as the probability of walking again after SCI in those older than 50 is much lower than those under.

Diagnosing Injury.

Immediate diagnosis is done through scanning, such as X-ray or CT scans, combined with general neurological exams, including voluntary or involuntary motor control tests. An alternative technique in neural evaluation, if the patient is not responsive, is electrophysiological recordings (either through EEG or EMG). X-rays are often used to immediately see large fractures, and follow-up CT scans for investigating the more possible hairline fractures. CT angiography may be used to investigate vascular destruction. More soft-tissue damage follow-up is done through MRIs. Interestingly, a trade-off exists in an MRI vs. immediate surgery. An MRI may be used to detect non-obvious issues, like disc herniations away from the primary injury site—which, if not fixed, will cause further degeneration. However, doing an MRI delays one's ability to decompress the spine.

Classifying Injury.

You will often hear injuries as being complete or incomplete. Another helpful distinction is discomplete, where the injury is considered clincially complete, but one can still observe connections through electrophysiology. The main method of classifying injury and tracking progress of patients is called the American Spinal Injury Association (ASIA) impairment scale. ASIA scores are broken into the following categories:

1. Grade A: Complete impairment, where there is no motor or sensory information being transmitted below the injury site[74].
2. Grade B: Incomplete impairment, where there is no motor information being transmitted, but some sensory information is preserved.
3. Grade C: There is some motor activity preserved, but more than half of the key muscles are too weak to move against gravity (Grade 3 muscles).

[74] It is not clear to me if this is measured by EMG or movement.

4. Grade D: A fair amount of motor activity is preserved, where at least half of the key muscles are above muscle Grade 3.

5. Grade E: There is no impairment at all!

The most common and complete method of classification is the International Standards for Neurological Classification of Spinal Cord Injury (ISNCSCI). The ISNCSCI uses the ASIA impairment scale, with ASIA motor and sensory scores. It is recommended that assessment be performed immediately upon hospital admission[75], with follow-ups in the future to assess improvement. Another important metric is quality of life (QOL) assessment.

Importantly, the injury site is often designated by the vertebrae that was fractured, but symptoms are due to the nerve pair that is damaged, which may be at a different location than the primary site of bone damage[76]. This discrepancy seems to be exacerbated the more caudal you go. Injury in the cervical portions can lead to severe bradycardia and hypotension, due to dysregulation of brain-heart communication, particularly regarding baroreceptor feedback. Too, damage to the vagus nerve can occur here, leading to dysregulation of most organs. Injury in the thoracic part may have widespread affects on the symathetic nervous system due to damage both of the spinal cord nerves and the nearby ganglia. Interestingly, an unconsidered byproduct of lower thoracic SCI is damage to motor signals to the legs. The main focus of such being that one loses their ability to walk, but accompanying this is reduced venous return, as veins rely on muscle movement to get blood back to the heart. Dampening of CNS-cardiovascular system communication seems to be one of the primary indicators of poor prognosis.

In evaluating the severity of the injury, a *spinal shock*, marked by temporary paralysis, may muddy the waters. While one may temporarily lose their reflexes, it can sometimes be regained soon later. However, the ability to define this state, and its duration, remains problematic. *Neurogenic shock* manifests similarly, but the cause is hypotension after SCI. This may be caused by hypovolaemia from blood loss, or pooling of blood due to reduced venous return. This occurs most often in SCI above T_6, as it is these sympathetic nervous which maintain vascular tone.

A few named pathologies exist, such as Central Cord Syndrome. This is the most common incomplete SCI. Often, this occurs in elderly patients who fall and already had some form of spondylosis[77]. It is marked by more damage to upper extremeties and possible incontinence. Brown-Séquard Syndrome occurs from penetrating SCI, such as a stab wound. It is usually characterized by sensory loss.

19.2 Auxiliary Management

Haemodynamics.

A steady blood supply to the damaged spinal cord is an essential component of treatment. Tool, system hypotension is a common symptom both from nerve damage and the surgery itself.

[75] This is interesting, as in other sections I have detailed that various forms of shock can cause misunderstanding in the initial phases of hospital stay.

[76] `https://www.nature.com/articles/nrdp201718`

[77] Weathering of the vertebrae.

Methylprednisolone sodium succinate.

Evidently, this is controversial! Methylprednisolone sodium succinate (MPSS) has the traditional role of reducing inflammation. The controversy is due to mixed results. Some large clinical trials showed no benefit, while others suggest there was a considerable improvement in ASIA scores, particularly if administered within 8h of injury. The adverse side effects seemed to be minimal, so many surgeons routinely administered it. Follow-up investigations suggest that negative effects are more consistent than previously thought, so the updated guidelines are not to administer MPSS.

19.3 Treatments

This section will cover some of the basic, in a broad overview manner. Complex research, such as the discussion of stem cells and the research surrounding them, will be covered in chapter **24**.

19.3.1 Electrical Stimulation

This will, of course, be covered in excruciating detail in later sections (and or, Parts, like Part **V**). Fascinatingly, electrical stimulation has been used in conjunction with physical therapy in the past with good resuts[78]. The reasons may be that this promotes stem cell differentiation[79], or disrupts inhibitory interneuron signaling. The optimal electrical application for differentiation has been explored extensively[80]. One may wonder if the benefits of BSI are in the interface itself, or simply the stimulation. Combining stem cell implantation and electronics is, likely, the future.

A slightly different rose by the same name is functional electrical stimulation (FES). Many trials have shown improvement in patients treated with either external stimulation or internal stimulation.

Light.

Light stimulation feels like a footnote in the electrical modulation story, to me. Though, if one wanted to control different neurons or enzymes on an alternate time course, optogenetic activation may be an option. The obvious issue being that one does not have genetic access to patients, and therefore would need to design (likely very complicated) targeted therapeutics.

Sound.

Another footnote is ultrasound simulation. In this case, it will be low intensity focused ultrasound. Some approaches have seen altered gene expression, but perhaps a more promising one is modulating mechanosensitive channels as was shown here[81]. Notably, this paper found that many mechanically activated channels are affected in ultrasound, including Piezo, and many of the Trp family proteins.

Magnetics.

I would be extremely curious to know if the magnetic field itself has any unique properties beyond its manipulation of the electric field. Still, too, the story is the same. Some seem to enhance channel

[78] https://www.nejm.org/doi/pdf/10.1056/NEJMoa1803588?articleTools=true
[79] https://www.mdpi.com/2073-4409/11/5/846
[80] https://www.frontiersin.org/articles/10.3389/fbioe.2021.591838/full
[81] https://www.nature.com/articles/s41467-022-28040-1

activation, while others expression. Interestingly, transcranial magnetic stimulation (TMS) has been used as a treatment with some success. Incomplete spinal cord injury has seen improvements from TMS.

19.3.2 Biomaterials

The overall goal in the use of biomaterials is to block a worsened immune response, scar formation, and promote neuron activity. Adding promise to stem cell implantation is the use of biomaterials that enhance proper network reformation[82]. Theoretically, a perfect biomaterial could be a substrate preferable for neuron growth, contain molecules that inhibit the immune response, neurotrophic factors that enhance stem cell differentiation and recruitment, and ion channel agonists. Notably, to date there have been no major publications where a "cocktail" like this has been successful. These sorts of things are usually made from hydrogels, collagens, or select inorganic fibers.

An open question is how one could leverage biomaterials to help clear damaged parts of neurons/cells that would normally be cleared by phagocytosis. Perhaps, one could add materials that are easily oxygenated to dampen the blow of ROS.

To date, implantation of biomaterials have been relatively lackluster in treating patients. While some regeneration scaffolds have proven to improve some neurological function, no patient has regained motor function.

19.3.3 Drug Treatment

Drug treatment primarily follows the same paths, being reduction of inflammation and neuroprotection. Methylprednisolone (MP) is the only drug approved to treat SCI and works through reducing inflammation. Notably, some side effects have been observed and therefore MP has fallen out of favor for treatment.

19.3.4 Surgery

I think you'll find that there is a disappointing lack of options—signaling the primitive nature of neurosurgery!

Decompression.

Anyway, surgical intervention aims to restabilize the spinal cord as quickly as possible, particularly through decompression. This begins with re-aligning the spine, usually with some kind of hardware to hold bone in place. Early surgery seems indicative of shorter ICU stays and better neurological recovery. The first day or so post SCI is the critical time. Interestingly, even after decompression, the pressure within the spine remains high due to fluid build-up within the dura matter. This makes blood reperfusion more difficult, leading to more problems.

[82] `https://pubs.rsc.org/en/content/articlepdf/2022/bm/d1bm01744f`

Dura Matter Manipulation.

While durotomy is often a complication of surgery due to progressive CSF leakage after operation, in this case it can be helpful to lessen spinal pressure, which there evidently is a long tradition of[83]. Duroplasty is a more modern and sophisticated alternative, and can allow opening of the dura matter without as much risk[84].

Myelotomy.

Incision directly into the spinal cord itself, myelotomy, has also been done with some success. The belief is that it helps drain some of the harmful dying tissue. There seems to be time dependence in this, where if performed too late after injury it will simply reinvigorate inflammation.

19.3.5 Rehabilitation

As you would intuit, exercise is the most common technique, as it preserves muscle mass and promotes circuit reorganization. Another rehabilitative technique is pumping in a significant amount of oxygen, as ischemia occurs after injury. BSIs have also become more popular. Fascinatingly, decoding of handwriting has been used to generate text[85]. Though, these seem to require deep access to the brain.

19.4 Complications After SCI

The multiplexed complications of SCI are, in reality, unending, as any combination of nerve damage can result in any combination of bodily dysfunction. We can discuss some of the standouts below, but note that this covers just a fraction of a fraction.

Syringomyelia.

Syringomyelia, a cyst forming within the spinal cord, is relatively rare. The fluid filled sac can be large and span far beyond the injury site itself. The symptoms of which may not show for months-years later, and progressively worsen. Cystic cavities are not uncommon, but syringomyelia differs in its size and reach. Symptomatic patients usually undergo a second decompressive surgery, or attempts to connect this cystic space to a drainage site.

Neuropathic Arthropathy.

After SCI, a patient may become numbed, leading to unnoticed injuries. One such result of this is arthropathy (a joint disease). Over time, a patient's joints may slowly degrade, leading to deformity—which may not present for over a decade after injury.

Spasticity.

Spasticity is a common side effect of chronic SCI. Spasticity can lead to further complications, such as microfractures.

[83] https://www.sciencedirect.com/science/article/pii/0020138388901325
[84] https://www.liebertpub.com/doi/full/10.1089/neu.2014.3668
[85] https://www.nature.com/articles/s41586-021-03506-2

Cardiovascular.

As mentioned in previous sections, often the clearest symptom of SCI is in significant changes to the cardiovascular system do to sympathetic nervous system impairment. This manifests often as systemic hypotension. In most cases, this seems to resolve itself within a matter of months.

Autonomic Dysreflexia.

This condition is an immediate response to *secondary* injury, and most often presents when the SCI is above T_6, and some other injury occurs in some peripheral organ, which causes a sympathetic spinal reflex. For example, injury to the gut may cause vasoconstriction to be overstimulated in the absence of spinal cord feedback. Too, it can result in overcompensation by the parasympathetic nervous system, leading to a swift drop in blood pressure. Notably, this can occur either in the acute or chronic stages of SCI. It is most often resolved via resolving this secondary injury.

Respiratory.

Damage to nerves innervating respiratory muscles, such as the phrenic nerve which controls the diaphragm, can lead to long term reduced lung capacity. Damage to the respiratory system also has big consequences on the effectiveness of rehabilitation. It is said that respiratory complications are one of the leading causes of death after chronic SCI.

Genitourinary and Gastrointestinal.

Dysregulation of the urinary or digestive systems is often the largest psychological consequence of SCI (barring, of course, large locomotor deficits). Quality of life is greatly impacted by this complication.

Neurogenic Heterotopic Ossification.

Ectopic bone formation can occur on the larger joints[86]. It is not known the exact cause, but physical therapy tends to be the best mitigator.

[86]Funnily enough, this review sites that 10-53% of people with chronic SCI form ectopic bone. Only a 43% margin.

Chapter 20

Approaches to Researching SCI

The first ever modeling and examination of traumatic spinal cord injury dates back to over 100 years ago. Dr. Alfred Allen, at the University of Pennsylvania, devised dog models of spinal cord injury and performed their accompany histological investigations[87]. I believe firms in finding the fundamental source, and appreciating where leaves meet branches, branches meet the trunk, and the trunk meets its roots (section **25.1**).

NOTE: It would be good to include some of the other approaches to researching neural regeneration in general—such sciatic nerve crush, optic nerve crush, *Drosophila* axon injury, etc.

20.1 SCI Injury Models

The bulk of this information comes from[88], and focuses on injuries applied to mice—but similar principles can be done in other mammals. Interestingly, most SCI models focus on thoracic level injury, while in humans cervical is more common.

Contusion.

Contusion is the quick application of a force. Models include dropping a weight on the spinal cord or using an air-gun. One such is called the *NYU (MASCIS) impactor*, where a piece of the bone is removed (laminectomy) and a weight is dropped directly onto the exposed nerves. Severity can be adjusted by the height from which the item was dropped. The most recent iteration is the *MASCIS III*, where an electromagnetic weight is dropped at the push of a button and includes digital measuring to ensure replication. One concern is that the weight dropping may bounce, causing multiple impacts.

Another option is called the *infinite horizon (IH) impactor*. Rather than dropping a weight, a machine applies a set force. A laminectomy is still required[89], but one can avoid the bounce effect. The *Ohio State University (OSU) impactor* is essentially the same, but uses an electromagnetic force applier. A more recent development is in the air gun. This requires drilling through the vertebrae, exposing the dura, where air pressure will be applied.

[87] `https://jamanetwork.com/journals/jama/article-abstract/448138`

[88] `https://www.nature.com/articles/sc201491`

[89] I suppose repeatability is considered paramount, but I find that these extensive controls may impede investigation. At a certain point, one must wonder if this is an SCI or an axotomy model.

Compression.

Compression occurs over a longer time frame. One interesting application of such models is in first applying contusion, followed by a compression model. This replicates the effects of increased pressure within the spinal cord after SCI, which occurs particularly in the absence of decompression surgery.

A common method is using an aneurysm clip and applying it to an exposed section of the spine, clamping down with some force for a few minutes. The force applied is in the range of 20-50g[90]. An interesting drawback is the velocity of the clip's closing is variable. In the right usage, it can be used to cut off blood supply to simulate ischemia and reperfusion after SCI. Another option is to insert a balloon on a catheter down the spinal cord and inflating it. Another method is called SC-STRAPPER[91]. A string is inserted and wrapped around the spinal cord, and its benefits are that it is relatively noninvasive, and compression is applied to the entire circumference of the spinal cord.

Distraction.

Distraction is done by stretching the spine. This is more commonly done in larger animals, and is difficult to develop a reproducable model. The Harrington distractor requires laminectomy and the addition of hooks to the upper and lower sides of the removed vertebrae. Motor steps are then applied to pull the spinal cord apart. However, as this occurs slowly and is variable in application, it has drawbacks. The UBC multimechanism device uses a wedge applied between the vertebrae to abruptly apply force. Interestingly, their device can also be used to induce contusion or slipped disc injuries—so learning it provides you with considerable versatility, particularly if dislocated vertebrae is your area of study. The singular wedge applies unidirectional injury, and so the UTA distractor aims to improve this by using clips attached externally (i.e., without laminectomy) and pull apart by a motor.

Transection.

Transection requires surgical severing of the spinal cord. Transection models may target the ventral, dorsal, or subsections of the spinal cord. Such models are most useful for studying regeneration, as clean transections are rare clinical pathologies. Interestingly, some animals have impressive, natural recovery of locomotion due to spinal cord circuitry post SCI (i.e., more robust central pattern generation or spinal reflex systems). Therefore, it is often that one must re-transect after locomotor recovery.

20.2 The Basso Scale

The Basso scale[92] has a few variations, such as the Basso, Beattie, Bresnahan (BBB) scale. It is a mode of characterizing injury after SCI in mouse models, specifically in the hindlimbs. You will often see Basso scale comparisons between treatment groups, or as a means to tract recovery over time after SCI. The scale is as follows (directly from the paper):

- Rating 0: No ankle movement.
- Rating 1: Slight ankle movement.

[90] Perhaps I am naive, but this seems remarkably small to me. When you consider that human SCI may result from a car accident, a few grams seems inconsequential.

[91] Take a look at Fig. 1: `https://www.sciencedirect.com/science/article/pii/S0165027008000368`

[92] `https://pubmed.ncbi.nlm.nih.gov/16689667/`

- Rating 2: Extensive ankle movement.
- Rating 3: Plantar placing of the paw, without stepping, OR consistent or frequent dorsal stepping (without plantar stepping).
- Rating 4: Occasional plantar stepping.
- Rating 5: Frequent or consistent plantar stepping, without coordination, OR some coordination with paws rotated at contact and lift off.
- Rating 6: Frequent or consistent plantar stepping, some coordination, and parallel paws at contact, OR mostly coordinated, paws rotated at contact and lift off.
- Rating 7: Frequent or consistent plantar stepping, mostly coordinated, parallel paws at contact, rotated at lift off, OR mostly coordinated with, parallel at contact and lift off, severe trunk instability.
- Rating 8: Frequent or consistent plantar stepping, mostly coordinated, parallel paws at contact and lift off, mild instability, OR mostly coordinated with paws rotated at contact and lift off, normal trunk stability, and tail up or down.
- Rating 9: Frequent or consistent plantar stepping, mostly coordinated, parallel paws, normal trunk stability, and tail always up.

Slight is defined as less than half joint excursion. *Extensive* is defined as more than half joint excursion. *Plantar stepping* is defined as placing both the thumb and the last toe touching the ground. *Consistent* is defined as without missing more than 5 steps. *Frequent* is defined as more than half the time. *Occasional* is defined as less than half the time moving forward. *Coordination* is defined as proper alternations between feet, and hindlimb and forelimb movements occurring at proper times, and the mouse must move at a consistent, non-slow pace. Some coordination meaning less than half the movements are coordinated, while most is more than half. *Severe trunk instability* is defined as an extreme waddle and collapsing during the test.

It is almost certain that I misworded some of those (as in accidentally wrote most instead of some a few times). But, I think you get the overall gist! Essentially, above 5 seems to have some degree of uncoordinated walking. Around 7-9 seem to have regained functional walking.

20.3 Traumatic and Nontraumatic SCI

Here we'll focus on neuroimaging and the information from[93]. A large focus of this review is the parallels between traumatic SCI and neurodegenerative diseases like degenerative cervical myelopathy (DCM). Traumatic spinal cord injury (tSCI) refers to impact or immediate stress applied to the spinal cord, while nontraumatic occurs over long periods through continuous compression or other. DCM is the most common example of nontraumatic SCI, and can occur through things like disc bulging. Naturally, months or years may go by without symptoms while degeneration is underway.

Animal models indicate considerable overlap in traumatic and nontraumatic SCI, but verification of these processes in humans is harder because metal implants immediately into tSCI patients makes MRIs impossible. Conventional MRI, too, is often not powerful enough to analyze microstructual changes, like demyelination, that are canonical to degeneration. A better method is diffusion tensor

[93] `https://www.nature.com/articles/s41582-019-0270-5`

imaging (DTI), which captures the diffusion of water through the tissue. It relies on the assumption that water diffuses along the white matter tracts. Core to the benefit of DTI and improved imaging techniques is identifying markers which can inform our understanding of improved recovery after SCI treatment. For example, clinical trials can be better motivated by quantifiable, noninvasive metrics like DTI.

Quantifying DTI.

DTI is devised by 3 eigenvectors, let's call them: $\vec{V}_a, \vec{V}_{r1}, \vec{V}_{r2}$. You would expect $\vec{V}_a$ to be the longest of the three as it runs axially, and or with the axons. Its eigenvalue, let's call it λ_a, corresponds to the diffusivity in the axial direction (parallel to the myelin tracts). The average of λ_{r1} and λ_{r2} describes the radial diffusivity, and or that which is perpendicular to the myelin tracts. The average of all 3 eigenvalues is the mean diffusivity. A scenario when $\lambda_a \neq 0$ and $\lambda_{r1} = \lambda_{r2} = 0$ would be described as an *anisotropic* diffusivity, because there is complete directional dependence of diffusion. $\lambda_a = \lambda_{r1} = \lambda_{r2}$ would be *isotropic*, because there is no directional dependence. The idea is that the diffusion of water gives some key insights into the microstructure of the injury, as things like ECM proteins and membrane composition will have some affect on such water diffusion. The axons themselves are naturally the largest contributor to this directional dependence, and you can imagine that in their absence water is much more likely to diffuse in all sorts of directions. The degree of anisotrophy is a strong way to measure degeneration, then. There are some limitations, of course, as diffusion is only a proxy for degeneration. Ulterior factors, like inflammation, may too limit diffusion and obscure the contribution by degeneration or demyelination. Similarly, DTI is not as effective at imaging lower spinal cord segements, and is primarily used in cervical injury models.

DTI in Degeneration and SCI.

Of considerable study is the parallels between SCI's secondary injury and degenerative diseases. After tSCI, retrograde and anterograde SCI initiate. Interestingly, in the first 2 years following SCI, tissue volume decreased by $\approx 14\%$, and $\approx 30\%$ in 5 years. Such a decrease was greater in the grey than in the white matter, which is similarly observed in DCM. Interestingly, no difference in diffusivity within the grey matter is observed after SCI, but it is believed this may be due to volume effects in the region.

Without going into the great detail on the vector changes associated with DCM and tSCI, the two show similar decreases in anisotrophy. Notably, DTI is able to resolve changes in asymptomatic DCM patients, making it a possibly useful tool in predicting worsening or improving outcomes of SCI that are not detected functionally. Remotely increased radial diffusivity was observed in both tSCI and DCM, suggesting that the local injury site / stenosis can have distant effects in both. However, DCM showed increased axial diffusivity above and below the site of stenosis, differing from tSCI. The authors suggest that this may be because compression at these sites increased cell density and decreased the extracellular space available for diffusion. Notably, though, the increase in radial diffusivity significantly outweighed axial increase, so it is still believed that demyelination is the dominant process in spinal stenosis.

Really, the takeaway from this is that cell death patterns in tSCI and degenerative diseases like DCM are comparable, but better investigations and imaging techniques will be needed to fully resolve this.

Chapter 21

Broad Coverage of Axonal Regeneration

Let's discuss[94] and[95]. During development, axons are guided by their growth cones to achieve functional connections at distant targets. A quintessential barrier to recovery after spinal cord injury is the central nervous system's inability to regain such functional connections. In general, neuron regeneration (or neuroregeneration) refers to axonal regeneration of an injured neuron. However, *sprouting* is another form of regeneration in which uninjured neurons may grow to replace the damaged synapses. Similarly, the use of stem cells can too be considered a form of regeneration.

While the CNS is relatively un-plastic after injury, some endogenous plasticity after spinal cord injury can and does occur, but in large part results in spasticity and neuropathic pain after SCI. An in-depth discussion on circuit reorganization is covered later (chapter **23**)—so for the moment, let's consider endogenous plasticity to be negligible or harmful after spinal cord injury.

21.1 Intrinsic Factors

Growth Cone Manipulation.

The physical interactions and the cell matrix, which typically relate to growth cone extension, are very important in growth. Deletion of non-muscle myosin II has shown enhanced growth. Deletion of profilin 1, which binds to actin and is thought to enhance its polymerization and use in filopodia, halts regeneration. Stabilizing the growth cone, or trending microtubule dynamics toward polymerization, is key to axonal extension. Doublecortin-like kinases (DCLKs) promote regeneration and are thought to modulate microtubule dynamics. Epothilone B and D are microtubule stabilizers, which have been shown to promote recovery after SCI. Epothilone B is also an anti-cancer drug as it inhibits mitosis, and similarly seemed to reduce glial scarring likely by inhibiting glial mitosis. Other similar, anti-cancer microtubule stabilizing drugs, like taxol, have shown similar outcomes. Importantly, other studies have called these findings into question, and some have shown quite negative outcomes in their usage. Similarly, it is unclear if their benefit is neuronal in origin or if their role in scar formation is sufficient to drive recovery. However, given that they are already used in cancer treatment entices researchers to find a use for them in treating SCI. To this point, though, their conclusive function is unclear.

[94] https://www.nature.com/articles/s41582-019-0280-3
[95] https://www.nature.com/articles/s41580-022-00562-y

Anterograde transport, including integrins and their RAB carriers, are shown to support regeneration by supplying the growth cone with necessities. However, selective transport is an essential focus, as RABs are likely to bring both pro and anti-regenerative factors to this site.

RhoA GTPase, a protein localized to the growth cone and functions in microtubule activity through Rho kinase (ROCK), has also been studied in an attempt to maintain the growth cone. Inhibition of RhoA or ROCK has promoted regeneration—specifically in acute SCI. If it is helpful in treating chronic SCI remains unclear.

Promising Proteins.

Target of rapamycin (mTOR) is all too familiar now, as it has been identified as an active translator in the axon and growth cone. mTOR phosphorylates 4EBP1, thereby inhibiting it, preventing it from repressing translation. Phosphatase and tensin homolgue (PTEN) is a negative regulator of mTOR. PTEN is likely the most famous regeneration blocker—and for good reason! Inhibition of PTEN even a year after SCI has been shown to promote enhanced recovery. Though, the regeneration is still typically insufficient for meaningful and complete functional recovery. Currently, developing therapeutics against PTEN are marred by worries that it will produce malignancies within the tissues, as PTEN is also an essential tumor suppressor. Interestingly, PTEN inhibition alone strikes somewhat variable results. Co-inhibition of PTEN with suppressor of cytokine signaling 3 (SOCS3) (mentioned in the next section) provided more consistent, enhanced growth—especially in the critical corticospinal tract.

Other proteins, like KCC2 (**23.1.1**), are discussed in-depth in other sections.

Transcriptional Possibilities.

Many transcription factors can be used to boost regeneration. One such is the cAMP-responsive element-binding protein (CREB) is necessary for axon growth in development. CREB activation by TTK21 promotes recovery after SCI. Another is signal transducer and activator of transcription (STAT3). Determining the biochemical barriers to CREB or STAT3 activation has been done through investigating histone deacetylases (HDACs) and DNA methyltransferases. It was shown that in peripheral nerves, where regeneration occurs naturally, histones were acetylated at specific, pro-regenerative sites. TTK21's activation of CREB is done through CREB-binding protein (CBP) acetyltransferase.

Fascinatingly, constitutive STAT3 activation is insufficient to promote regeneration. The reason being SOCS3 is a negative regulator of the JAK-STAT transcriptional pathway. While STAT3 may be activated by therapeutics or intrinsic growth factors, SOCS3 knockdown is required to attain its benefits. Another key family of transcription factors within this umbrella are Krüppel-like factors (KLFs). Many members of the KLF family have been found to block neurite outgrowth, with KLF4 being the strongest. Conversely, overexpression of other KLFs, like KLF6, enhance regeneration. Co-Knockdown of KLF4 and SOC3S further promoted growth. KLF4 is thought to bind with phosphorylated STAT3 and inhibit it, while KLF6 seems to further activate it.

Minorly related is the discovery of the dual leucine zipper-bearing kinase (DLK). DLK is necessary for retrograde transport of STAT3, suggesting it is used in injury signaling. Interestingly, evidence for the DLK pathway's injury response in glia has emerged. In this case, DLK knockout worsens the injury site as astrocytes do not properly respond. The same can be said for SOCS3-STAT3 and PTEN-mTOR, demonstrating a correlation between the intrinsic factors dictating neuronal and glial

injury responses.

21.2 Extrinsic Factors

Growth Factors.

A canonical problem is the lack of growth factors that normally guide axonal growth in development. Such canonical examples include neve growth factor (NGF), brain-derived neurotrophic factor (BDNF), neurotrophin 3 (NT3), and glia-derived neurotrophic factor (GDNF). Naturally, growth factors have different, neuron subtype-specific effects. NGF modulates nociceptive neurons, NT3 and GDNF modulates corticospinal neurons, BDNF modulates corticospinal and reticulospinal neurons, etc. Notably, matured neurons may not be recepetive to such growth factors, and one mode of treatment includes supplementing both the growth factors and their accompanying receptors. Other options exist, such as in sensitizing the neurons to growth factors therapeutically.

A key finding in this department is that the insulin-like growth factor 1 (IGF1) promotes regeneration. This may be partly due to its ability to promote mTOR activity, but the many potential outcomes of IGF1 overexpression are unknown. IGF1 injection has been used to treat autism in the past with relatively safe and promising outcomes, however the fear of tumor production remains a barrier to its use in treating SCI.

Growth Cone Navigation.

The cell-specific responses surrounding neurons is covered in-depth in section **18.1**. Many of the modifiers exist in the ECM, such as secretion of collagen protein (Cthrc1), a pro-regenerative protein [96]. Conversely, WNT signaling molecules accumulates around the injury site and repels axon regeneration and sprouting.

Other non-neuronal cells, like astrocytes and fibroblasts, secrete chondroitin sulfate proteoglycans (CPSGs). CPSGs are composed of a protein core with chondroitin sulfate sugar exteriors, whose inherent variability makes their study difficult. After injury, CPSGs too accumulate around the injury site and limit growth. They can do so through the protein tyrosine phosphatase σ (PTPσ) receptor or Nogo, which may trigger RhoA.

Administration of chondroitinase ABC (ChABC) has been shown to promote plasticity and growth by degrading CPSGs. ChABC administration in non-human primates promoted hand dexterity recovery without negative side effects. ChABC has poor stability and is difficult to produce in large quantities, making its clinical applications somewhat limited. Too, there is some evidence of their benefit, such as through perineuronal nets which may provide a supportive structure to neurons in the grey matter.

Myelin Inhibition.

Interestingly, many myelin-associated proteins are inhibitory, called myelin-associated inhibitory (MAI) proteins. These proteins bind to the Nogo receptor complex, and blocking this interaction, such as by NogoA antibodies, promotes growth, especially in the acute phase. Examples include myelin-

[96] https://www.cell.com/developmental-cell/pdf/S1534-5807(20)30984-9.pdf

associated glycoprotein (MAG) and oligodendrocyte myelin glycoprotein (OMGP), which both utilize a NOGO to RhoA-ROCK pathway to inhibit regeneration.

21.3 Recovery of Walking via Target Reconnection

Reversion to a state of biological development too is a core route of investigation commented on many times in this work. An important investigation by Anderson, Courtine, and Sofroniew showed that one could achieve robust regeneration past an SCI injury site through a cocktail of regenerative factors[97]. As proprioceptive neurons have been found to be possible relay neurons, these researchers targeted such neurons and were able to achieve growth across the scar. This required the use of growth factors applied to the neurons, fibroblasts, and attractive molecules downstream of the injury site. Release was done in a spatially and temporally controlled manner. Fascinatingly, **no functional recovery was found**. Therefore, while neurons were able to bridge this gap, they were unable to form useful synapses. Authors postulate this may be due to inactivity or lack of myelin, among other things. In a similar vein, the grafting of stem cells is often successful in allowing for axonal growth across the injury site, but retains the inability to restore function.

Follow-up from Anderson and Courtine labs was successful[98]. This perfectly highlights that functional recovery requires more than getting neurons to regenerate—it also requires getting them to regenerate in the right direction and make the right connections. That is a prime focus of this work. Commented in a later section (section **27.0.3**) the Courtine lab characterized the cells responsible for stimulation-induced recovery after SCI. Briefly, this work found that a set of interneurons were able to relay signals past the injury site and down to the lumbar spinal cord, where walking-initiating neurons are. Interestingly, these neurons are not required for walking in the absence of injury. This paper is a partial follow-up to this work.

First, to better characterize neurons projecting to the lumbar spinal cord, they injected an AAV containing GFP into the lumbar region, which was then expressed in neurons projecting there. They identified cells they call $\text{SC}^{Vsx2::Hoxa7::Zfhx3 \rightarrow lumbar}$. The first part describes their transcriptional profile, and the arrow to lumbar means they project to the lumbar spinal cord. Mice that exhibited recovery after spinal cord injury showed marked transcriptional changes in such cells, including upregulation of genes involved in dendritic connections and synaptic growth.

To "ablate" these cells, they used Vsx2^{Cre}, which blocked differentiation into this cell type. They found that there was a marked decrease in projections to the lumbar spinal cord, taking this to mean these neuron make up a significant portion of the axonal density there. They also found that Vsx2 expressing cells connect to many places in the spinal cord, including forming short and long projections. Those that project specificallt to the lumbar region are marked by expression of Zfhx3. They used similar methods to identify cells of the ventral gigantocellular nucleus (vGi) as projecting to these $\text{SC}^{Vsx2::Hoxa7::Zfhx3 \rightarrow lumbar}$ neurons.

With this better understanding of the neurons, they employed the following plan:

1. Reactivate the intrinsic growth capacity of neurons using growth factors like insulin-like growth factor 1.

[97] https://www.nature.com/articles/s41586-018-0467-6
[98] https://www.science.org/doi/10.1126/science.adi6412

2. Restore a favorable growth substrate using fibroblast and epidermal growth factors.
3. Guide the axons back to their desired targets using favorable biomaterials infused with growth factors and attractive agents.

This indeed allowed for growth of these neurons beyond the injury site. However, this did not allow for functional recovery. Thus, they concluded that functional recovery can only be achieved via meaningful connection re-formation. Instead, they tried replacing biomaterials with lentiviruses driving expression of growth factors, which indeed allowed for functional recovery. Via their previously mentioned AAV method, transcriptional and immunohistochemical classification, they verified that this was indeed due to Vsx2 neuron growth. They verified that motor neurons had regained CNS control via stimulating neurons of the vGi and identifying corresponding leg movements. They verified, via expression of toxin receptors and corresponding toxin administration in Vsx2 neurons, that these neurons are required for recovery.

There are two small criticisms of this work that have been floated around the field. The first is that "regeneration past the scar" is not inherently true if the scar has been heavily conditioned by biomaterials. Though, this can be achieved in therapeutics as well, so it is not an untenable drawback. A more worrisome one is that conditioning with growth factors, particularly permanently through lentiviruses, may introduce further pathologies. Namely, it could result in over-proliferation, and or, cancer. Regardless, it is an amazine work.

21.4 Neuronal Excitability and Regeneration

This is quite a difficult topic to cover. There is almost no end to the amount of conflicting information. Presumably, this is because neuronal activity dictates many, many underlying processes. There is a growing hypothesis, put forth largely by the Bradke group, that neuron activity is largely anti-regenerative. The basic principle is that being in "growth mode" or being in "active mode" are somewhat mutually exclusive. Further, other works have found that ablating individual synapses, but maintaining some, limits growth. Conversely, eliminating all synapses evokes a more strong regeneration response. The idea here is that a neuron may attempt to preserve its remaining function by not entering a so-called regeneration mode. The implications of this are not straightforward, as naturally many forms of CNS nerve damage, namely SCI, result in complete axonal severing—yet, of course, such neurons do not regenerate properly. On the other hand, axonal sprouting can and does occur from non-injured neurons. Thus, in my eyes, the whole theory is a bit confusing and may well be subtype-specific.

Furthermore, inhibition of certain calcium channel components, like the $\alpha 2\delta 2$ subunit of voltage gated calcium channels, promote axon regeneration. This subunit mediates influx through N and P/Q-type channels. Thus, it's thought that these channels may underlie activity-dependent halted regeneration. One method of inhibiting neuron activity is through pregabalin, which limits spasms. Further evidence by administration of gabapentinoids enhancing regeneration shows promise in using anti-convulsants to treat SCI. This work fits into the larger hypothesis of neuronal activity limiting axon regeneration. Further work demonstrates that active vesicle machinery, like the protein Munc13, limit axon regeneration.

However, a strange wrinkle arises when considering the Bradke group's work on L-type calcium channels. They found that L-type channels also inhibit axon regeneration in mouse DRG neurons in a

work from 2010[99]. Our lab, on the other hand, found that L-type VGCCs are necessary and sufficient for axon regeneration in the *Drosophila* PNS. My hypothesis would be that such channels may be mediating local influx in DRG neurons, whereas it is global in *Drosophila* PNS neurons. There also may be differences in timing and frequency of firing. Unfortunately, the Bradke group did not follow-up on this work, and therefore we have very little to analyze regarding it. The skeptic in me also wonders if they didn't do follow-up work because their results were conflicting, and or, because a phenotype was difficult to nail down. Regardless, the conflict warrants further study.

[99] `https://pubmed.ncbi.nlm.nih.gov/20579880/`

Chapter 22

Metabolism of Neurons

22.1 Programming Mitochondrial Maintenance

This is one of my favorite Reviews ever, so I am pleased to write about it[100]! A key component of this work is discussing the unique structural features of neurons that lead to trouble trafficking mitochondria and meeting injury needs after injury or in degeneration. An incredible amount of energy is consumed by the brain (20% of total body consumption), and approximately 1,000,000 molecules of ATP are used per action potential in an effort to restore ionic concentrations alone. Around 20,000 molecules of ATP are used per synaptic vesicle recycles, and its predicted that the concentration of ATP in the presynaptic terminal is around 2mM.

22.1.1 Mitochondrial Generation and Degredation

Notably, mitochondria are not spontaneously generated, but rather undergo fission to create more—which primarily occurs in the soma where such machinery is located. Thus, irrespective of how much mitochondria are present, the key is that it must be properly trafficked. Amazingly, it can take days for mitochondria to reach the tip of an axon. Late endosomes carrying RNA granules and ribosomes may dock on mobile mitochondria and serve as sites for mitochondrial protein synthesis. Peroxisome proliferator-activated recepetor-gamma coactivator (PGC)-1α is a key transcription coactivator of nuclear-encoded mitochondrial genes. ALS and Parkinson's disease models both show decreased distal translation of mitochondrial proteins and build-up of the corresponding mRNA, suggesting the energy demands are not met in these diseases.

Mitochondrial fission is mediated by dynamin-related protein 1 (DRP1), a cytosolic GTPase. DRP1 acts with mitochondrial adaptors, including fission protein 1 (FIS1), mitochondrial fission factor (MFF), and mitochondrial dynamics proteins (MiD), to induce fission. While larger mitochondria are better at energy production, smaller mitochondria are more easily trafficked to distant sites. Notably, fission can lead either to biogenesis or to mitophagy, depending on where the fission occurs and which adaptors mediate it. FIS1 specifically tends more toward mitophagy, while MFF more toward biogenesis.

Mitochondrial fusion is mediated by factors like optic atrophy 1 (OPA1) and mitofusion 1 and 2 (MFN1 and MFN2). Both are GTPases and aid in inner membrane and outer membrane fusion respectively. Fusion is useful in limiting mitochondrial stress, as combining the contents of two dilutes

[100] https://www.sciencedirect.com/science/article/pii/S0896627322002513

harmful agents and aids in the increased production of energy. Mutations in OPA1 are commonly linked to Parkinson's disease.

Fixing damaged components of the mitochondria may occur through proteases lining the inner membrane. Alternatively, mitochondria frequently bud off vesicles, called mitochondrial-derived vesicles (MDVs), that contain damaged proteins or reactive oxygen species. Such budding is achived through mitochondrial Rho GTPase (MIRO) and DRP1, and are then targeted to lysosomes. PTEN-induced kinase 1 (PINK1) is also able to enhance degredation through mitophagy, along with Parkin (PARK2). Such pathways are dysfunctional in Parkinson's disease as well. PINK1, on the outer membrane, recruits Parkin, which recruits degradative machinery. Interestingly, axonal mitophagy is incredibly rare. Typically, mitophagy occurs only after a damaged mitochondrion is trafficked to the soma, and in the case of Parkin or PINK1 mutations, damaged mitochondria build up in the soma rather than the axon.

22.1.2 Mitochondrial Trafficking

Trafficking is necessary for both the removal of damaged mitochondria and the addition of them to energy depleted sites. The speed of trafficking is quite variable, and is capable of both stalling and changing directions. Such movement occurs along the polarized microtubules within axons, especially by motor proteins like the kinesin-1 family (KIF5) for anterograde and dynein for retrograde transport. On the outer membrane, the MIRO-trafficking kinesin-binding proteins (TRAK) complex connects to KIF5 motors. It seems that there is some functional redundancy in the MIRO-TRAK complex with other outer membrane proteins, but in large part it seems to be a strong regulator of mitochondrial placement.

In the case of injury, movement of the mitochondria to the damage site is a key component of repair. One way to enhance promote movement is by deleting syntaphilin (SNPH), a mitochondrial anchoring protein. SNPH is capable of competing with the MIRO-TRAK complex's binding with KIF5. It is thought that p21-activated kinase 5 (PAK5) inhibits SNPH by phosphorylating it, enhancing mitochondrial mobility. Overexpression of MIRO1 was found to increase mitochondrial trafficking to the distal axon segments, leading to larger growth cones. MIRO1 is mediated by neuronal activity, as Ca^{2+} influx binds with MIRO and disrupts complex formation. Though, this blocking seems transient and seems to require the presence of SNPH. This mechanism seems particularly important for halting mitochondrial movement around the Nodes of Ranvier, where a higher energy demand exists to restore membrane potential.

Mitochondrial trafficking is adaptable. For example, high concentrations of glucose are capable of arresting its movement through glycosylation of TRAK. Guidance of mitochondria through the complex branches of an axon require such intracellular cues, some of which are discussed in the next section.

22.1.3 Response to Stress

Stress induced AMP-activated protein kinase (AMPK) phosphorylates PGC-1α to upregulate mitochondrial RNA. ATP depletion has been shown to activate the AMPK-PAK3 stress pathway, which uses myosin 6 to recruit mitochondria and SNPH to anchor it to surrounding F-actin. Therefore, exceptionally active synapses may experience mitochondrial buildup proportional to their firing. As it goes, stress induced anterograde trafficking in a highly branched axon may result in mitochondrial build-up at the branch sites, which is achieved too by SNPH. Parallelly, AMPK is capable of

blocking retrograde transport of mitochondria, resulting in their accumulation at distal, stressed sites. Counteractive to trafficking, reactive oxygen species may result in increasing intracellular Ca^{2+}, which arrests mitochondrial movement. As mitochondria are one of the main sources of such species, this can become a self-fulfilling prophecy of positive feedback.

Neuronal Degeneration.

Most neurodegenerative diseases are accompanied by mitochondrial dysfunction. Build up of protein aggregates, for example, hinder trafficking. Too, mtDNA mutations can regularly cause neurodevelopmental impairments.

PINK1/Parkin, while traditionally known for its role in Parkinson's, defines many mitochondria-related pathologies. Its function is thought to be in degrading MIRO through ubiquitination, causing local mitochondrial arrest and eventual mitophagy. Conversely, a mechanism of remobilization where mitochondrial vesicles containing SNPH are released with the idea of enhancing retrograde transport after stress. This mechanism is activated in pre-symptomatic ALS models and Alzheimer's disease. It seems that early disease causes mitophagy primarily in the soma, while later and or higher levels of stress may induce it within the axon. The site of mitochondrial degredation is therefore an indicator of disease progression.

22.1.4 Neuronal Injury and Regeneration

While degeneration is progressive and slow acting, injury is acute. Repair is inherently a highly energy intense process. The injury itself can cause mitochondrial structural damage, furthering the energy crisis. Preserving mitochondrial function is a mode of limiting Wallerian degeneration after injury. Axotomy quickly depolarizes mitochondria, and inhibiting SNPH serves as an important way to remobilize and restore their function. In certain injuries, the AKT path becomes activated which promotes PAK5 (mentioned above as inhibiting SNPH), serving as an energy-boosting mode of neuronal repair. Another such method is via ARMCX1, a protein that mobilizes mitochondria through MIRO1. Histone deacetylase 6 (HDAC6) is known to deacetylate the MIRO1 encoding gene, leading to lessened mobility. Conversely, hypoxia inducible factor 1α (HIF-1α) activates hypoxia upregulated mitochondrial movement regulator (HUMMR), which promotes MIRO1/2.

After injury resulting in limited blood supply, such as in a stroke, metabolism occurs through glycolysis, resulting in increased lactate production. However, after the blood supply is regained, repurfusion injury can overload the system with oxygen, resulting in production of reactive oxygen species, thereby further harming mitochondria.

Delivery of mitochondria to damaged neurons through vesicles or tunneling nanotubes (TNTs) is thought to be one mode of enhancing recovery. It has been shown that neural stem cells are capable of transferring mitochondria to damaged cells, and microglia use TNT to do so. Mitochondrial transplantation has been done, however the efficacy of such is heavily debated. Though, it is well supported that complex networks of metabolite shuttles exist in the nervous system. The astrocyte-neuron lactate shuttle (ANLS) is a prime example, thought to allow astrocytes to fuel neural oxidative phosphorylation. Glial regulation of neuronal metabolism too occurs, such as through release of sirtuin 2 (SIRT2) which promotes metabolism in neurons.

Chapter 23

Circuit Reorganization and Plasticity

We will cover circuit reorganization in considerable depth, not necessarily in the context of spinal cord injury, in this section.

23.1 Reorganization After Injury

Circuit reorganization naturally and with neuromodulation is covered in-depth here[101]. Part of this paper is framed on the fact that spinal cord circuits are much more complex than previously thought. That is, some computation is done at the level of the spinal cord. Too, the paper attempts to motivate investigating the specific neurons involved in recovery, and that previous investigations have not used the resolution necessary to investigate such matters. One constraint on the field is that, while the brain's cellular profile has been extensively investigated, detailed profiling of spinal cord cells is less explored—particularly in the injury condition.

Circuit reorganization can occur in many ways, both above and below the injury site. Examples seen in our own lab include damage to axons causing retraction and thus sprouting of the axon in seemingly random directions far above the injury site. Too, below the injury site, axons that once filled this space will fade away, and uninjured axons may grow into this now empty space. Interestingly, growth into this empty space has been shown to augment hand function after recovery[102].

Of course, when damaged neurons regrow in their previous locations it can cause functional recovery. However, fascinatingly, even when all corticospinal tract neurons are damaged, motor cortex signals can be sent through alternative, surviving paths. Here are two brainstem examples: (1) After SCI, cortical neurons can synapse onto the reticular formation and recover walking. (2) Activation of the mesencephalic locomotor region (MLR) after SCI can recover walking. This region is evolutionrily ancient and unable to generate walking on its own, but is able to recruit reticulospinal neurons and generate walking.

In a similar vein, otherwise non-essential neurons, like propriospinal neurons in the cervical region, can be reprogrammed to improve hand dexterity. While one may think that information must be conveyed directly from the motor cortex to the muscles on interest, in this case propriospinal neurons relay information past the injury site. KCC2 downregulation leads to inhibition of such circuits. This

[101] `https://www.nature.com/articles/s41593-022-01196-1`

[102] This is all very interesting when it comes to BSMIs—as one usually focuses on or around the injury site, or the dorsal root exit zones that generate walking. This suggests many sites along the spinal cord may be worth stimulating.

is expanded on in a later section. Much remains unknown about such neurons, what allows them to transmit beyond the injury site, and what recruits them to do so.

Rehabilitation proves to consistently aid circuit reformation, but lack of care causes poor reorganization that can worsen pathologies. Notably, when supraspinal[103] connections are severed, motor control relies on sensory signals only. One example being *V2a* neurons discussed in a previous section. Another similar set of neurons includes *dI3* neurons[104], another set of interneurons. While these aid in recovery, many changes are harmful to it. As a general trend, it appears that generally less activity, due to injury conditions like hypoxia, immune cells, etc. leads to greatly increased sensitivity of receptors and ion channels alike. Below the injury site often sees large increases in excitatory synapses. This combination leads to abnormal reflex responses, spasms, neuropathic pain, etc.

As a brief digression, circuit reorganization is of considerable interest in stroke patients as well. One interesting finding is that a stroke in one hemisphere is often worsened in a positive-feedback like manner during recovery. That is, the uninjured hemisphere continues to inhibit or suppress the injured one. One way to avoid this is by physically forcing a patient not to use the uninjured side of their body (i.e., to rely as much as possible on the injured side) to maintain and even promote further activity. Such treatments do not work if no residual function exists. However, it does provide an interesting potential route of treatment using BCIs.

Effects of Neuromodulation.

It would be good to mention something about the Hebbian principle, and this paper[105], for demonstration of time-dependence in plasticity.

Again, neuromodulation is covered in considerable depth in Part **V**. Both activity and artificial stimulation of the motor cortex promote regrowth of the corticospinal tract. This can be done non-invasively. Timing this with spinal cord stimulation enhances connections. There are a couple of paper's exploring this in depth in NHP[106], and in humans[107]. It would be good to look through these papers in the future. Fascinatingly, indirect neuromodulation, such as through the vagus nerve, too has been shown to enhance response to rehabilitation. Too, deep brain stimulation has been done on the MLR[108], and when positive yielded results comparable to stimulating the motor cortex[109]. Though, the MLR's region is less clearly defined, and results were more variable.

Stimulating the spinal cord is primarily done at the dorsal root entry zones or through EES. It is thought that low-frequency EES works through propriocetive afferents, which help circuit reorganization. Damage to the sympathetic nervous system is an essential problem to solve, as it can lead to life threatening drops in blood pressure. One such way to do so is in stimulating the low thoracic spinal cord's pre-ganglionic sympathetic neurons, causing release of norepinepherine and thereby constrict vessels. Similarly, EES can restore the sympathetic response via large afferents recruiting sympathetic

[103] Supra meaning "above" the spine, i.e., brain-to-spine neurons.
[104] lhttps://elifesciences.org/articles/21715
[105] https://pubmed.ncbi.nlm.nih.gov/24448593/
[106] https://www.cell.com/neuron/pdf/S0896-6273(13)00762-9.pdf
[107] https://www.pnas.org/doi/abs/10.1073/pnas.1505383112
[108] Another example of DBS promoting rather than inhibiting—kind of interesting.
[109] It would be nice to build a patient profile of those who may benefit the most from motor cortex vs. MLR vs. vagus nerve etc. stimulation.

neurons.

For stroke, successful trials have been done to show that, when BCIs implement feedback, circuit reformation can be achieved. In one example, visual feedback was given when patients entered a sensorimotor state consistent with flexing their hand, which helped reinforce their ability to contract this hand. Interestingly, such BCIs typically rely on kinematic control (velocity, acceleration, position) rather than kinetic (forces, torques)—in other words, without muscle involvement. It seems likely that better circuit reformation would be achieved with this integration. Such a trouble can be ameliorated with the addition of EMG signals in feedback (hence, the title of this book!).

Effects of Biological Manipulation.

Biological strategies seem most successful when attempting to enhance the projectome of surviving supraspinal neurons. One such way to do this is through removing damaged tissue and grafting in more malleable, healthy tissue. As part of this, physicians may attempt to ablate the fibrotic scar. Interestingly, partial ablation, and not complete ablation, has shown symptom improvement after SCI.

Pharmacological promotion of axon regeneration has inherent limitations. Such investigations into axon regeneration are covered in better depth in chapter **21**. For example, the canonical player PTEN (discussed in other sections) has shown robust regeneration, capable of penetrating the astrocytic border, but only in certain neuronal populations. Too, even in this case neurons abruptly halt when reaching the fibrotic scar. Somehow, one must induce global, or brilliantly targeted, growth to achieve true functional recovery.

Another application commented in earlier sections is the conversion of inhibitory neurons into excitatory, or the deletion of inhibitory neurons. One way to do so is via chondroitinase (ChABC) which dissolves glycosaminoglycan sid chains of chondroitin sulfate proteoglycans (CSPGs). CSPGs are known for their role in stem cell differentiation, and bookend the periods of neurogenesis. Another such attempt is in inhibiting nogo receptor reticulon 4 (RTN4).

23.1.1 Interneurons and KCC2

This is from the paper[110], but it seems KCC2 is an interesting topic of research in a couple different pathologies, such as in minimizing neuropathic pain, and downregulation increases spasticity after SCI. Some of the background for this paper is in better understanding how spinal cord stimulation leads to regain of function, with one hypothesis being that it helps reawaken "dormant" circuits.

To investigate, they used staggered, lateral hemisections performed at the T_7 and T_{10} sites. This injury meant all neurons would be severed, except those crossing the midline between T_7 and T_{10}. They verified near complete paralysis after injury, and used this to conclude that while some neurons were spared, they must be "dormant" and or unable to recover function. They performed a mini-screen of drugs known to modulate neural activity, and landed on CLP290 as the only one showing a phenotype. CLP290 is known to activate KCC2, among other things. They determined that CLP290's function required some preserved axons—as in their complete lesion tests, no functional recovery was found. They further verified this by determining no enhanced regrowth was found due to CLP290 in

[110] https://www.sciencedirect.com/science/article/pii/S009286741830730X

their hemisection model. To determine if CLP290's function was through KCC2, they tried a KCC2 overexpression line and verified similar functional recoveries.

They then used the Cre-Lox system to verify that KCC2's expression specifically in inhibitory interneurons (driven by Vgat-Cre) leads to recovery. They presumed these neurons interact with / enhance the ability of propriospinal neurons to transmit information, and to resolve the spatial component, they tried local injection of AAV-KCC2 in the lumbar segments and did not find recovery. They believe it is specifically interneurons within the relay site between the two lesions that are required for repair. Their explanation being that when administering systemically, the blood-brain-barrier is compromised only at the lesion site.

In development, intracellular Cl^- is high. During this period, Cl^- channel opening is excitatory, rather than inibitory in fully-developed cells. KCC2 upregulation is important in removing this Cl^-, converting Cl^- channel opening into being inhibitory. They believe activating KCC2 helps restore a more physiological state. They also did work to inhibit inhibitory interneurons, and excite excitatory interneurons, and determined that only the former was beneficial in achieving recovery after SCI.

23.2 Non-coding RNA in Rewiring and Plasticity

Background.

Non-coding RNAs (ncRNAs) have been studied quite a bit in recent years[111]. ncRNAs are classified as small (less than 100 nt) or long (above 500 nt), and among the small ncRNAs are microRNAs (miRNAs) of $\approx$ 20nt. They are typically produced through canonical means and play a role in post-transcriptional regulation within the cytoplasm.

As for miRNAs, gene silencing occurs by binding to a complementary untranslated region of an mRNA. This region is sometimes called a *seed region*. From here, a miRNA-induced silencing complex (miRISC) forms. Then, some of the standard RNA degredation mechanisms can occur like decapping. Some of the relevant long ncRNAs include long intergenic RNAS (lincRNAs), which are linear, circlar RNAs (circRNAs), and antisense (AS) RNAs. CircRNAs are produced but noncanonical means, and generally seem to function in tandem with miRNA in some manner, and or, they regulate miRNAs. AS RNAs are born of the opposite DNA strand, most often modulate this specific locus, and may be thousands of nt long. They are particularly relevant in early development. ncRNAs work in a symphonious way, usually involving multiple types coordinating together. This Review makes it a point to state that recent data has been overwhelming, due to advancements in technology. Therefore, it remains difficult to stay up to date or understand the complexities as new functions continuously emerge.

Regulation of Local Growth.

Local translation was first determined to be mediated through miRNAs in the case of miR-9's ability to decrease the expression of microtubule protein MAP1B via the brain-derived neurotrophic factor (BDNF) path within growing axons. Similarly, BDNF increased presence of miR-132, leading to increased axonal branching. In Xenopus, miR-182 was shown to be necessary for axon targeting via multiple modes of tublulin and actin regulations and their usage in the growth cone's guidance.

[111] `https://www.cell.com/neuron/fulltext/S0896-6273(23)00341-0`

In the dendrites, miR-9 has also been implicated. miR-9 is under a negative feedback loop, in which the repressor element 1 (RE-1) silencing transcription (REST) factor is activated by miR-9, which in turn decreases its expression. miR-9 downregulated the actin regulator Diap1, inhibiting further growth. In short, miR-9 inhibits growth of both dendrites and axons.

In mice, loss of miR-101 led to increased network excitability and impaired memory through increased expression of Na-K-Cl cotransporter 1 (NKCC1) and increased synaptogenesis, thought to be mediated through *Kif1a* and *Ank2*. Fascinatingly, miR-101 acute loss did not lead to such a result, but rather only when lost in development. The lincRNA ALAE was the first example of an lncRNA that locally regulates mRNA translation during growth by blocking the function of RNA-binding protein, and therere allowing GAP43 synthesis, causing axons to grow[112].

Circuit Remodeling.

In general, it appears that miRNA expression is bad for remodeling, and long-term potentiation (LTP) specifically. miR-134 expression was found to inhibit long-term potentiation through inhibition of CREB1. Under physiological conditions, SIRT1 inhibits miR-134 expression. miR-26a and miR-384 also inhibit LTP by different means. Many such other examples exist[113].

Interestingly, miR-153 was upregulated in the dentate gyrus during LTP. It seems to be required for exocytosis, suggesting it has a role in excitatory coupling. miR-218 and miR-137 were also named as promoting LTP. I think that's about all that's worth saying, for now.

23.3 Synaptic Pruning

23.3.1 Ube3a E3 Ligase Promotes Pruning in Flies

In general, it is rare for a purely fly paper to make it into the big journals like Science and Nature. Papers that include fly work typically use the later figures to validate their work in higher organisms, like mice—but this paper didn't, and is still in the current issue of Science[114]. The topic of synaptic and dendritic pruning is of considerable interest, as it has been implicated in a number of neurodevelopmental diseases. For example, in autism spectrum disorders, synaptic pruning occurs improperly. Flies have long been used as a model to study dendritic pruning, because as they go through metamorphosis (from larva to adult fly), a significant amount of dendrites are pruned and connections are remade. The study of synaptic pruning is more novel, and that's what this paper focuses on.

They investigate using fly class IV dendrite arborization (C4da) neurons, a group of snensory neurons. They identify the ubiquitin ligase Ube3a as being specific to C4da neurons and in presynaptic, but not dendritic, pruning. In UBe3a mutants, or RNAi knockdown, a significant amount of synapses are maintained through metamorphosis. They can visualize this with a marker called brunchpilot (Brp) which markes synapses. They found that Ube3a undergoes anterograde transport toward the synapses with the help of the kinesin motor. They found a missence mutation in Ube3a that was conserved

[112] I think part of the issue with this kind of paper is it is almost purposeless trying actually learn this. It is simply letters followed by numbers doing things redundantly.

[113] I hate to kind of "yada-yada" through this, but again, it's just miR-xxx does different, harmful things.

[114] `https://www.science.org/doi/10.1126/science.ade8978`

between flies and humans with Angelman disease—a disease where low Ube3a dosage disrupts pruning. This mutation disrupted Ube3a localization but not expression. Naturally, they wanted to test Ube3a's role in degrading proteins. Among its previously established targets is the the bone morphogenetic protein (BMP) receptor Thickvein (Tkv). Overexpression of Tkv induced similar defects as seen in Ube3a mutants. Therefore, their hypothesis is that Ube3a activation in metamorphosis downregulates signaling pathways like Tkv, which overall lead to synaptic pruning.

One of the ways this paper shines is in investigating a mutation conserved between humans and flies, which helps provide new insight into Ube3a. However, the paper would be significantly strengthened if they had mouse work or human cell cultures to support their findings.

Chapter 24

Stem Cell and Tissue Engineering

In this section we'll discuss a few labs in much depth, such as the Tuszynski lab at UCSD and the Cullen lab at UPenn—some of the pioneers in three dimensional tissue engineering. My understanding is that most researchers would agree stem cells alone are likely to be insufficient to attain functional recovery after SCI. Even the most staunch supporters would agree an integrated approach, combining neuromodulation and rehabilitation, is certain to be better. The ependyma is a source of stem cells, which contributes to glial scar formation after SCI. Connexin (gap junction) signaling between ependymal cells is a factor involved in their differentiation, notable because it opens up the possibility of manipulation through epidural electrical stimulation, discussed in chapter **27**.

Injections of both oligodendrocyte progenitor cells and spinal-cord derived neural stem cells have been grafted into patients without complications, demonstrating promise.

24.1 Cell Transplantation Therapy

In this section, we'll primarily cover[115]. The majority of implantation investigation focuses on neural stem and progenitor cells (NSPCs, or NPCs), Schwann cells, oligodendrocyte precursor cells (OPCs), olfactory ensheathing cells (OECs), and mesenchymal stem cells (MSCs). NSPCs are multipotent progenitor cells and are able to differentiate into neurons and glia alike.

Naturally, one of the primary benefits of grafting stem cells is that they may replaced damaged neurons and form new functional connections. Alternatively, they can form relay stations where endogenous axons synapse onto the newly grafted cells. The physical addition of cells itself plays the role of giving neurons something to synapse and grow onto. If, for example, a neuron was capable of growing, the unideal milieu, such as open cavities filled with interstitial fluid, do not provide anything to grow on. This is sometimes called a *bridge*, especially when it allows neurons to grow rostrally to caudally. Through production of ECM proteins like laminin and collagen, a new, more favorable environment can form. However, such bridges still have great limitations, causing heterogenous regrowth due both to the neurons inherent abilities and because the graft may be dissimilar from certain axon's needs.

It is also thought that many such implanted cells allow for neuroprotection after SCI. This will be commented on a bit later, but one issue with this is neuroprotection is strongest when the implantation is done in the acute phase. Implantation after the first day sparsely shows protection. Yet,

[115] `https://www.nature.com/articles/nn.4541`

implantation is normally done at 1-2 weeks after injury, because when done in the acute phase, the graft tissue is more prone to dying. This creates a slight Catch 22, where to maximize protection one must implant early, thereby sacrificing the graft and limiting the other potential benefits. The proposed mechanisms include protecting oligodendrocytes via healthy signaling, production of myelin, and the sprouting of new axons. All of these factors may contribute to the observed enlargement of the white matter near the injury site after a graft.

Other mechanisms of study include revascularization. Revascularization begins around 3d after injury and reaches its maximum around 14d after injury[116]. OECs, in particular, appear to increase vascularization and enhance its directionalization along the graft and or spinal cord. Immunomodulation is similarly of interest, as correlative studies demonstrate grafting decreases harmful inflammation.

A general concern is where implanted cells may migrate. Therefore, you'll likely see the mention of halting migration in further reading. One strategy to doing so is encasing the stem cells within a matrix, typically of fibrin.

24.1.1 Long Distance Growth of Stem Cells

Stated in the 2019 Giovanni Review: "We believe that the most promising preclinical strategy to promote long-distance regeneration of axons in the spinal cord has come from recent studies in the **Tuszynski laboratory**, which used human spinal cord-derived NSC grafts combined with a cocktail of growth factors." Indeed—let's begin with such a study[117]. The idea of this paper is to use stem cells cultured in a fibrin matrix, containing a cocktail of growth factors, and then graft this entity into rats after SCI. Growth was mTOR dependent, but not Nogo dependent[118].

Rats underwent complete T_3 transection. Rat embryonic stem cells were labeled with GFP and differentiated primarily into neurons after grafting. Cells did not migrate much beyond the graft / lesion site either. Grafted neurons were capable of growing long distances, with axons extending multiple vertebrae away both rostrally and caudally. They determined approximately 29,000 axons grown after grafting, and a similar density among all subjects tested. In fact, grafted neurons even became wrapped by host oligodendrocytes! Their justification for the growth not being Nogo dependent is that grafted cells still expressed Nogo, so therefore it is unlikely that Nogo inhibition promoted the growth. They tried inhibiting mTOR and found a significant decrease in axonal growth.

They then repeated these studies using human-derived stem cells. I didn't see this quesiton answered directly in the text (maybe I just missed it) but I am wondering how much the neurons actually matter, especially for functional recovery. I.e., the matrix contained growth factors—could the growth factors alone be promoting growth, independent of the stem cells?

24.1.2 3D Printed Spinal Scaffolds

The ability to create 3D structures of cells is becoming ever more popular. One such example is again from the Tuszynski lab[119]. The goal of this paper is to generate a scafford capable of promoting growth, more so than creating a CNS mimic. They essentially designed a cross section of the spinal

[116] Take this with a large grain of salt.
[117] `https://www.cell.com/fulltext/S0092-8674(12)01018-5`
[118] Dr. Strittmatter will be upset!
[119] `https://www.nature.com/articles/s41591-018-0296-z`

cord with pores / channels where axons can grow into, synapse onto neural progenitor cells, and those cells can grow in the opposite direction and synapse onto neurons below the injury site. It is kind of like a well structured relay station for signals. They call their method microscale continuos projection printing (μCPP), and remarkably it is able to print such scaffolds in a matter of seconds. They showed an interesting potential application where a 3D model of a patient's SCI lesion was taken and then a corresponding scaffold was printed to fill the gap.

The device itself is quite impressive. They are able to print on a 1μm level. In their proof of concept runs, they printed scaffolds that resemble different modes of SCI. Notably, this is meant as an almost complete replacement for the spinal cord—i.e., it requires a very invasive surgery to implant and the similar removal of leftover nervous tissue. The ones they used were 2mm in length and made of polyethylene glycol-gelatin methacrylate (PEG-GelMA). Fortunately, this paper had some of the most in-depth controls, including comparisons to stem cell only injections and scaffolds made of agarose. Another key control was that the implantation did not worsen inflammation. However, at the inflammation site there was still a reactive astrocyte layer. While it was less prominent in the PEG-GelMA scaffold than in agarose, it was still present.

Very interestingly, implantation of stem cells often requires waiting many weeks. This is because dissociated cells like this, which do not have support from blood vessels and are exposed to reactive oxygen and immune hurdles, do not survive well. The group wondered if by providing a supportive scaffold, if implantation can be done earlier after injury—and indeed it could. Acutely implanted scaffold retained their stem cells a month later. It seems like in general, the results were a bit inconsistent. In long term studies they found that some neural progenitor cells became myelinated and the tubes became vascularized. In fact, they even claim that astrocytes' feet lined the vessels, signaling a restored blood brain barrier. Interestingly, the walls of the scaffold thinned by $\approx 50\%$ at the 60 months post implantation mark. It is unclear if this has any meaningful impact, though, as once neural connections are made, perhaps the scaffold itself is not needed.

Onto the part that matters most. Animals with the scaffolds exhibited significant functional recover, reaching a Basso score of $\approx 6-7$, compared to the controls reaching $\approx 1-2$ at 5 months after injury[120]. They also used an electrophysiological method where they stimulated the motor cortex and read motor neuron outputs, and again found the scaffold to exhibit significantly better outcomes[121].

[120] This is a sidenote, but geez—it is kind of inhumane to keep a mouse alive with complete immobility for 5 months, solely to ensure it doesn't recover.

[121] As another sidenote, this whole thing is really... odd, in a sense. Of course, the FDA and medical boards cannot allow researchers to do whatever they want. But, these mice had a complete transection and were able to get significant recovery. Surely it can do some good in humans? This is some serious motivation for an enhanced Right to Try bill. If one is completely paralyzed, the option should be open.

Part V

Brain-Spine Interfaces

When we stand at the edge of the ocean, we can not understand its vastness. We know only that it is big. Ittō Ittosai is big. And what about Kojirō?
– Reworded from Vagabond by Takehiko Inoue

24.2 Perspective

Papers to Read:

I am so behind, it's actually insane.
– Neuroimaging guidelines: https://www.nature.com/articles/s41393-019-0309-x
– munc13 https://www.sciencedirect.com/science/article/pii/S0896627321007753
– courtine in NHP https://www.nature.com/articles/nature20118
– Myelin: https://www.nature.com/articles/s41577-023-00907-4
– Another scaffold: https://www.science.org/doi/10.1126/science.abh3602
– CD8 t cell paper https://www.science.org/doi/full/10.1126/science.abd5926
– Growth cone review 1 https://www.nature.com/articles/nrm2679
– Growth cone review 2 https://www.nature.com/articles/nrn3176
– a third, focusing on mechanics https://www.frontiersin.org/articles/10.3389/fncel.2015.00244/full
– literally focuses on glia-neuron growth cone lol https://www.frontiersin.org/articles/10.3389/fnins.2020.00203/full
– Attwell SCN model https://onlinelibrary.wiley.com/doi/full/10.1002/dneu.22601

Chapter 25

Interfacing with Nerves

25.1 Overview and Historical Dealings

A proverb in a writing called *Testament* by Daitō Kokushi reads:

> I beg you, try to find the fundamental source.
> ...
> Like our great predecessors,
> Do not merely pinch off the leaves
> Or concern yourselves only with the branches.

Naturally, a large focus of this book is neural repair. Therefore, we will consider brain-computer interface teachnology primarily through this lens. The fields of neural regeneration and neural stimulation feel novel, as investigating their complexities appear to require new age technology. Indeed, the term "brain-computer interface" itself is only a few decades old, originating in 1973 from Jacques Vidal who used EEG readings to control a computer. Thus, we're infected with the allusion that all key advancements have come only recently.

However, greatness spans time. In the 1930s, one of the most brilliant neuroscientists the world has ever known, Ramón y Cajal, suspected restoration of neural circuits after injury required the emulation of developmental conditions. Such predictions continue to hold weight today. Using electronics to study neural circuits, too, has a long and storied history. One can think of Hodgkin and Huxley, whose ubiquitous achievements evade a need for citation. Using an operational amplifier to clamp the voltage of giant squid axons, they revealed the fundamentals of action potentials in the 50s. Or, one can think of Fitzhugh, whose work inspired Nagumo to devise analog circuits to model excitable systems in the 60s. In the case of solving spinal cord circuits, so too have electronics long been used, with examples dating back to the 40s[122]. Similarly, stimulating the spinal cord originates from the theory of central pattern generation, first characterized in the 1980s[123]. Even integration of deep brain and spinal cord stimulation in to the clinic date back nearly a century.

I find the history of investigation to be quite interesting as a stand alone subject. A large part of what we must do as modern scientists is to not forget the questions posed by our predecessors. While they were limited by technology in their ability to investigate, the validity of their curiosities can never be

[122] https://journals.physiology.org/doi/epdf/10.1152/jn.1943.6.2.111
[123] https://nyaspubs.onlinelibrary.wiley.com/doi/10.1111/j.1749-6632.1998.tb09062.x

lost. Thus, I'll try to sprinkle in earlier works and thinking whenever appropriate.

Often, "brain-computer interface" refers to the use of brain-waves controlling some external device. However, this work will largely use it as a catch all term regarding any biology to electronics communication, whether it be from machine to neuron, vice versa, or both. A BCI can be either closed-loop (entirely self contained, with feedback incorporated within the device) or open-loop (requiring some external input). A BCI may be either *assistive*, meaning it intends to replace lost functions in some way (such as using a patient's brain waves to control a speech producing device) or *rehabilitative* (such as using some sensorimotor feedback to restore the patient's speaking ability).

This area needs major cleaning—it feels very scattered / disorganized.

25.2 Overview

25.2.1 Electrostimulation

This section is so bad... definitely needs re-writing.

Stimulating nerves is not straightforward. The amount of current necessary to stimulate a nerve is highly dependent on the location. For electrodes especially close to myelinated fibers, currents as small as 0.1 μA have been successful in depolarizing neurons[124]. As a general trend, the faster conduction velocity of a neuron, the lower amount of current is required to simulate it (presumably due to low capacitance of the membrane). However, these sorts of measurements are often thrown off by indirectly stimulating a neuron of interest by depolarizing nearby cells, causing one to mis-attribute the source stimulation required. For direct stimulation of myelinated axons, the distance from electrode to the nearest Node of Ranvier is the key metric.

The effect of a current supplied by an electrode generates a voltage spread resembling something of a double-sided Lenhert-Jones plot. That is, injecting a positive current may generate depolarize the region nearest to the electrode, while the surrounding regions are actually hyperpolarized. Therefore, for an anode, one would expect a core of cells to be depolarized, while those of the periphery to be hyperpolarized. The inverse would be true for a cathode. If one were to use two electrodes (i.e., an anode and a cathode nearby) then one has the option of longitudinal stimulation or transverse stimulation (i.e., either parallel or perpendicular to the axon).

Note: Perhaps it would be good to go through the different electrode options? I.e., where one places ground, whether current or voltage is applied, etc. Though, I have had trouble finding information on this.

Biomaterials Used.

Building electrodes or bio-compatible devices requires a combination of factors. A primary focus in finding suitable biomaterials are those that are conducting, flexible, and non-reactive[125]. Polypyrrole

[124] https://www.sciencedirect.com/science/article/pii/0006899375903649
[125] https://onlinelibrary.wiley.com/doi/epdf/10.1002/term.383

(PPy) is one such option. PPy alone is quite rigid, but is a good accompaniment with other materials. For example, you may coat some glass or fabric in PPy in order to make it more biocompatable and conducting. Too, PPy is a good candidate for boosting cell adhesion to one's biomaterials. Others have used PPy in a tubular fashion, applied some current, and used it to guide regenerating axons. Another similar product is polyaniline (often called PANI). Polyaniline can be oxidized and reduced, providing an additional layer of manipulation. PPy and PANI are two of the most studied conducting polymers, but many more exist.

Carbon nanotubules (CNTs) are another way one can manipulate the direction of growth. Interestingly CNTs appear to be cytotoxic when suspended in media, but strongly promote growth when integrated into a scaffold, keeping them immobilized. CNTs frequently have diameters similar to axons themselves.

The use of piezo-electric materials too provides interesting avenues of study. Piezo-electric materials being those that convert mechanical energy to electrical energy. An example is poly-vinylidenefluoride (PVDF). Generally, piezoelectric materials are considered inferior to standard conducting polymers because external control is limited and activation is often too localized.

Ramblings.

The axon-like size of CNTs seems very interesting to me, as they provide a neat route of usage, i.e., as artificial myelination. Perhaps one can develop a CNT with a membrane permeable enough for extracullar signaling to get through, but conductive enough to function like myelin. I am curious if you could produce a multi-layered scaffold, where certain tracts fit axons, certain tracts fit glia, etc, and it can act like repaired oligodendrocytes after SCI.

Another interesting thought is the deep implantation of Piezoelectric materials and then activating them using sonication. It would be neat to develop some kind of material that can be mechanically stimulated at specific sonic frequencies, preventing random stimulation from body movement.

25.2.2 Flexible Brain-Computer Interfaces

For the moment, I'm not sure where to put this, so I'm adding it here. The geometry of BCIs is quite important to their function, as was covered in a review[126]. The key idea of this work is that hard electronics present a geometric mismatch with the body's tissues. This review comments on several marvels of the brain, including that the dura mater, skull, and scalp function as low-pass filters! Which, now we know about in considerable depth from chapter **3**. This leads low frequency signals, like α and β waves, have biased detection by modes like EEG. Another interesting point made right away is that billions of neurons are distributed over volumes on the order of centimeters to meters, while neurons themselves are on the order of tens of microns to hundreds of microns, meaning neurons do not occupy the majority of brain space. Similarly, neurons fire on the order of milliseconds, yet brain wave changes and behavioral outcomes take months to manifest. The vast, vast majority of models simplify neurons and neural circuits as a network of wires. The 3-dimensional geometry of neurons is regularly neglected, and for good reason. However, these long term changes, and the vast non-neuronal volume of the brain, suggest our biological computer is more than just transistors.

[126] https://www.nature.com/articles/s41928-022-00913-9

Hard electronics are less flexible and adaptable. If BCIs are to improve, it'll mean that they must interface with more neurons and capture more data. Such a request is not feasible with hard electronics. Similarly, hard electronics offer potential harm to neural tissue via mechanical damage. Slight migration in hard electronics on the order of microns is inevitable. Therefore, the tracking of single neurons over long periods is impossible without re-calibrating and accounting for such migration. This work also harps that single unit recordings are not possible over long periods anyway, as gliosis and scarring will eventually worsen the signal-to-noise ratio so that individual neurons become invisible in the static.

The mismatch between hard electronics and neural tissue is often measured by Young's modulus, which designates the flexibility in response to applied force. Much thinner electronics are required to match the electronic and neural tissue moduli. Indeed, such attempts have been done, which require reducing transistors down to nanometer thicknesses. Tissue-like electronics that have been produced prove to be much better at recording single neuron activity over long periods. A considerable consequence of shrinking and shrinking electronics is that there may be accidental crosstalk between nearby transistors.

The brain expands and contracts in a non-negligible way. This is especially true in the developing brain, or in a highly degenerative brain. Thus, electronics would ideally be stretchable to accommodate volumetric changes. This is similarly true for segments of neural tissue that may require flexion, such as peripheral nerves or spinal nerves. Stretchable electronics have been achieved in the past when made as a mesh. Well designed meshes with channels can integrate optogenetics and chemogenetics as well, allowing for the possibility of cell-type specific neural recording and modulation.

Soft electronics like this come with considerable design and production challenges. For example, scaling up high densities of transistors embedded in meshes is extremely difficult. Similarly, these devices are very sensitive to ion or water influx. Further, parasitic capacitance (capacitance due to electronics proximity to one another) is largely unavoidable at small scale. When dielectrics are used as insulators, again possible mechanical damage due to flexion is a possible source of ion influx and thus worsened function.

Chapter 26

Peripheral Nerve Stimulation

Peripheral nerve stimulation[127] has shown promise in treating nerve damage, but naturally is greatly limited by technology. One such limitation is the difficulty in targeting peripheral nerves that may be lodged deep within the body. Too, once reached, one must avoid off-target stimulation, movement of the electrodes over time, and swelling in surrounding regions. Removal of leads after the completion of neurorehabilitation can cause further damage as well, as tissue may become attached to it, resulting in a second injury.

26.0.1 Adaptive, Conductive, and Electrotherapeutic Scaffolds

Adaptive, conductive, and electrotherapeutic scaffolds (ACESs) is one implementation of electrical neurorehabilitation after injury[128]. The primary structure is electrodes embedded in alginate (an anionic polymer, which creates a maleable gel-like structure when combined with water, and is often used for creating molds) and polyacrylamide (a cationic polymer most known for its use in SDS page gels, and can be used to suspend solvents at various stiffnesses)—which can be dissolved, allowing for removal of leads with minimal damage to tissue. Because the electrodes are located within the gel, glial encapsulation is not to occur (or, if it does, it will be at the borders of the gel and not the electrodes themselves)—therefore, stimulation should be un-dampened. Gold particles were used to increase conductivity of the gel.

In rats, in order to test accessible nerves, sciatic nerve transection was performed, and electrodes were used for 6 weeks. Stimulation was delivered every other day for 30 minutes in anesthetized rats. It was delivered to each electrode at 20Hz (every 200 ms), 2 mA[129] amplitude, and a 100 μs pulse width. They used a stainless steal electrode by Plastics One[130]. Stimulation was powered using the Intan system[131]. As two electrodes were used, their pulses were offset by 100 ms. The two stimulating electrodes were proximal and distal to the injury site, and a third, ground electrode was located at the tibialis anterior muscle. EMG recordings and response to sensory stimulation were used to validate functional success of regeneration, as well as axon counts. To test percutaneous nerve stimulation, they performed a similar procedure on the vagus nerve of swine.

[127] Annoyingly, peripheral nerve stimulation is often abbreviated as PNS, so do not get confused with peripheral nervous system vs. nerve stimulation.

[128] `https://www.sciencedirect.com/science/article/pii/S2666634023001654?via%3Dihub`

[129] It is not clear to me why they choose to deliver current rather than voltage.

[130] Interestingly, I can't seem to find a product that resembles the one found in their supplemental images. Still it is probably comparable: `https://www.protechp1.com/search?q=MS303&type=products`

[131] `https://intantech.com/RHS_system.html`

26.0.2 Vagus Nerve Modulates Circuits via Acetylcholine

Note from future Jackson: I was a bit critical here lol. It would be good to ... rework ... this section.

Vagus nerve stimulation seems to be much more complex than one might expect on surface. One such example is[132]. Vagus nerve stimulation (VNS) has been shown to enhance motor recovery after nerve injury when paired with rehabilitation. Canonically, VNS is used to treat epilepsy and depression—but new treatment options are expanding, such as recent papers looking into treatment after a stroke. Notably, VNS leads to widespread brain activity, but seemingly has a narrow, targeted effect. This paper identifies M1 neurons as being partly responsible for enhanced motor control following VNS. **Importantly, this was tested only in healthy mice, and not any form of nerve injury.** Their work found that VNS specifically *after* successful completion of a task helped reinforce learning of this task—therefore, it may not work as a therapeutic in the case of total injury, but could help recover partial injuries back to normal levels.

Awake, active mice were stimulated via cuff which wrapped the Vagnus nerve. Machinery was mounted on the skull, from which wires would extend. They tested a few different methods, and found stimulation following success to be the only positive difference from the sham. Stimulation was done with 100μs pulses at a frequency of 30hz (or, every 33.3ms) a total of 15 times (entire duration summing to 500ms) at an amplitude of $\approx$ 0.5mA. Looking at their figure 1, there are some questions. Firstly, what they decide as early and late in the training days seems arbitrary (i.e., first 4 days is early, second 10 is late). I am curious if you thresholded this differently, if their significance levels would change. Secondly, the VNS stimulation results were significant, but the % increase in task completion from WT to stimulated seems only to be about 10-15%. Similarly, the difference in early vs. late trials seems to be about 10% as well—making me wonder how physiologically relevant this is[133].

To determine what is downstream of VNS, they used a tetrode implanted into the brain to read the response of neurons in the basal formation, as they wondered if it was here that encoded for reinforcement. They found that about 43% of neurons showed some response (either suppression or activation), and about 28% showed an increased activity due to VNS. Now, I'm no expert, but those don't seem like great odds to me. That seems like a coin toss. I didn't see much mention in this paper of trying different VNS parameters. I would postulate that they'd get a much better response in tweaking this. An experiment they could run right away is using the tetrode, and varying peak duration, amplitude, and frequency of stimulation to maximize this increased activity parameter. They could then repeat their behavioral studies and see if it increases success rate. Interestingly, they did something similar in the motor cortex, investigating temporal relationship of activation or suppression of neurons over a range of amplitudes. They found that neuronal activation and suppression occurred at disparate times succeeding VNS. Fascinatingly, to test if this oberved interaction was ACh dependent, they injected ACh antagonists and allowed them to act globally—effectively invalidating the entire experiment[134].

As a concluding remark, it seems from their data that VNS may contribute to motor learning reinforcement, but certainly is not the essential driving factor. Similarly, I did not see a "negative control" anywhere in their paper (i.e., a vagotomy, for example) demonstrating that motor learning was severely hindered when this pathway was stopped[135]. Many of the decisions made in this paper

[132] `https://www.cell.com/neuron/fulltext/S0896-6273(22)00555-4#%20`

[133] I.e., would a patient consider brain surgery to have a 10% increase in some motor activity? Absolutely not.

[134] I must be missing something, because this experiment makes no sense.

[135] Perhaps it is buried in the supplements, or done in a previous paper, though.

are quite confusing to me.

26.0.3 Hand Movement Recovery via Intrafascicular Stimulation

This group utilized intrafascicular, rather than intramuscular, stimulation as a means to recover fine motor skills in the hand of primates[136]. Courtine is a middle author on this paper. A drawback to intramuscular surgery is it requires electrodes placed on multiple muscles and is typically quite invasive. Very interestingly, intramuscular stimulation has been found to recruit large efferent nerves before small, contrary to normal function and leading to faster fatigue. In other words, faster fatiguing muscles are activated earlier than expected in this case. Stimulating an entire nerve fascicle requires high amplitude and is prone to overstimulation or recruitment of unintended nerve fibers. Intrafascicular stimulation is an opportunity to resolve this.

Such a method requires first identifying the ideal implantation site and building an electrode design accordingly. They call their device a transverse intrafascicular multichannel electrode (TIME). They developed a model of the monkey's fascicles in order to model recruitment[137]. They settled on a 16-electrode design, withheld in a polyimide shaft.

The bulk of the paper is stimulating individual or a combination fibers to initiate a variety of different grips. The end is when they take it to the next level, so-to-speak, in giving the monkeys some voluntary control. They pharmacologically block the monkey's spinal cord circuitry. They used intracortical recordings to modulate hand stimulating, using what they describe as a simple program. Their design was too simplified to activate complex grips, but did prove to work in simple grasping tasks.

[136] `https://www.science.org/doi/10.1126/scitranslmed.abg6463`
[137] I don't know how they did this, but it would be good to learn.

Chapter 27

Spinal Cord Stimulation

Note: These next two paragraphs deserve their own sections and likely should be in a different section. These labs are instrumental.
Restoration of the signals eliminated after spinal cord injury through therapeutics has been equally explored. For example, drugs such as clonidine, α_2 receptor agonist, can promote walking in cats with spinal cord injury[138]. This effect can be blocked completely by yohimbine, an α_2 receptor antagonist[139].

The ability to generate walking-like movements from electrical stimulation of the lower spine, the T10-L1 region, has been long known[140,141]. In these instances, "tonic" stimulation was able to generate step-like movements. Although the term "tonic" is dubious because other papers by the same lab identified frequency dependence in their stimulation[142]. While one can generate step-like movements at once frequency, continued extension can be generated with another.

A large part of spinal cord stimulation, particularly epidural electrical stimulation (EES), is in fact not neural stimulation directly, but rather CSF stimulation[143]. EES requires the implantation of electrodes onto the dorsal surface of the spinal cord. As was identified through this computational model, depolarizing signals spreading through the CSF are capable of triggering large, afferent fibers. Such fibers are most easily accessible at their dorsal root entry zones. Their activation leads to motor neuron activation, or the conversion of an inactive circuit to an excitatory one. The bulk of this investigation is done at the lumbar level, as its focus is on restoring walking after SCI. However, other studies have replicated the idea in hand dexterity when stimulating the cervical dorsal root entry zones.

An unsung hero which operates by similar means is transcutaneous stimulation—done through the skin, making it non-invasive and easily performed. Of course, it lacks the precision of EES, but has been used to promote motor recovery after SCI.

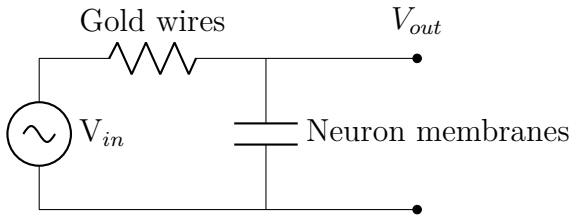
[138] https://journals.physiology.org/doi/full/10.1152/jn.1998.79.6.2941
[139] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2278596
[140] https://www.sciencedirect.com/science/article/pii/B9780444521378000188
[141] https://journals.sagepub.com/doi/pdf/10.1177/1073858417699790
[142] https://link.springer.com/article/10.1007/s00422-004-0511-5
[143] https://www.jneurosci.org/content/33/49/19326.short

27.0.1 Electronic Dura Mater

A couple years ago now, Courtine and Lacour groups developed a framework for electronic dura mater[144]. One of the primary problems addressed here is that dura mater is hard and protective while neural tissue is soft—so if one were to remove the dura and replace it with an electronic, it would need to meet both of these criteria. Their "invention" is called e-dura. It is meant to go beneath the dura mater, and is composed of silicone with gold wiring[145]. The silicone has a similar kPa to the dura itself. Most implant are done above the dura, so the soft nature allows for implantation beneath it.

One of their comparisons is in implantation of soft vs. hard electronics in locomotion (this is without spinal cord injury). Notably, the soft electronics performed the same as the sham—while the hard electronics were much worse. They similarly quantified damage to the spinal cord, and expectedly found that soft electronics did less damage, and similarly that there was more inflammation in hard electronics. They also made a model of the linak cord using a hydrogel to show similar effects—perhaps useful to keep in mind.

I've noticed this same thing in a few papers. But, they make the claim that impedance stays constant over the course of a few weeks. However, the error bars show an impedance of $\approx 50\text{k}\Omega \pm 30\text{k}\Omega$. Firstly, this is quite high impedance—but that may be a protective feature. Secondly, and more crucially, the spread is humongous. My wonder is if the Reviewers did not notice or understand this, and saw that the average stayed around the same. Impedance does seem to be quite variable / strange in every paper. So it also may be that controlling this is quite hard. There is also quite a large phase angle, reaching $\approx -30°$. Notably, *in vitro* there is no phase shift nor is the impedance near 50kΩ. The $\approx -40°$ shift is suggestive of this being a low-pass filter at their tested 1kHz. Very fascinatingly, the circuit may look something like this[146]:



In vitro, you would see this phase shift because there are no capacitors to ground. Let us take this a step further, even. If:

$$\begin{aligned} A_{out} &= A_{in}\sin(\omega t + \phi) \\ \frac{V_{out}}{V_{in}} &= \frac{1}{\sqrt{1+(2\pi f RC)^2}} \\ \sin(\omega t + \phi) &= \frac{1}{\sqrt{1+(2\pi f RC)^2}} \\ \sin(-40°) &= \frac{1}{\sqrt{1+(2\pi \times 1\text{kHz} \times 6\text{k}\Omega \times C)^2}} \end{aligned} \tag{27.1}$$

[144] `https://www.science.org/doi/abs/10.1126/science.1260318`

[145] I didn't realize this until just now, but the increased impedance over time is due to micro-cracks in the gold. Interesting to keep in mind.

[146] This is why you should read the electronics section.

I estimated R to be 6kΩ by eyeballing their graphs. I couldn't find any raw data in the paper. Solving this gives us a C value of 3.16×10^{-8}F, or 31.6nF. This does seem a bit low, as I believe the standard estimate for a cell membrane is around 100μF. However, there are likely many factors at play.

I digress. The functional tests they did were using a 3×3 array of electrodes, and the motor cortex of mice were stimulated using channelrhodopsin-2. The e-dura demonstrated the ability to record signals sent from the cortex. They then tried delivering stimulation after spinal cord injury, induced via severe contusion. They delivered serotonergic replacement therapy through microfluidic chambers in the e-dura and delivered continous stimulation at 40hz, 0.2ms, and 50 to 150μA. Naturally, stimulation improved locomotion.

27.0.2 Electronics with Shape Actuation

The goal of this paper is to address the gap in stimulation devices[147]. That is, one can choose either percutaneous electrodes (i.e., puncturing through the skin and covering a small area) or a large paddle requiring invasive surgery. This paper developed a flexible paddle, which can be rolled up to a diameter of less than 2 mm, and injected similar to an electrode but expands upon entrance. Too, the flexible shape is fitted with microchannels. Applying air pressure to such channels causes expansion. It is composed of materials like gold, parylene-C, silicone, polyethylene, and polyimide.

The device is made in two stages, but is monolithic. The first silicone layer is laid, and before it is fully cured, secondary materials are added. Then, the final layer of silicone is laid, and the entire device cures together. To validate it electrically, they used "electrical impedance spectroscopy techniques[148]" Interestingly, the impedance was calculated to be around 1000 Ω! And it was quite variable; their box plots show an IQR of about 600 to 1200 Ω. This is, indeed, not ideal. One concern is how much power will be dissipated by current running through the device, thereby generating heat (and possibly causing device expansion). They also tracked increase in impedance due to mechanical stress, and it seems to increase from about 500 to 1000 Ω—representative of an approximately new device vs. approximately 6 years old within the body[149].

The device has a diode-like IV curve[150]. They look exactly like a diode would in the positive half (i.e., current exponentially increases at about 0.8V). But the negative voltage range looks quite strange. To apply voltage, they used the Intan device, like in the ACES paper.

The claim is that they are able to implant the entire device using a syringe (presumably under the lamina) into cadavers and then expand the device over the dura mater. Somehow, the how the device is able to spread over the spinal cord while, despite all of the connective tissue and fat filling the lamina-dura space[151]. To find the desired site of implantation required fluoroscopy, meaning the device itself had to be visible under fluoroscopy. Another confusing aspect of this paper is the biphasic nature of SCS to treat neuropathy. As one always undergoes a trial period, the question is will this device complement the trial phase properly?

[147] `https://www.science.org/doi/10.1126/sciadv.abg7833`

[148] It is not clear to me what they are referring to, beyond simply driving different currents. There is a paper discussing this here, though.

[149] How they could approximate a 6 year old device is questionable to me.

[150] For lack of a better word, the IV curves look insane.

[151] All of this is very odd to me. I am having trouble understanding how this is possible mechanically.

The big takeaways, to me, are that expansive devices seem possible, and fitting it with microfluidic channels is a fascinating way to apply therapeutics in addition to electrical stimulation. The next question is, of course, can we do the same thing with ECoG—thereby not requiring a highly invasive procedure to record brain waves?

27.0.3 SCI Specific Cells

Using a simplified device similar to what will be discussed in the upcoming Lorach *et al.* 2023 section, the Courtine group used electrical therapy to re-establish walking in a number of SCI patients[152]. Their paper goes into depth on the cell-type specific response to this electrical modulation. They called this model epidural electrical stimulation (EES) + rehab, or $\text{EES}^{\text{REHAB}}$. They first approached by measuring metabolic consumption in the spinal cord before and after $\text{EES}^{\text{REHAB}}$ using PET scanners, and found that metabolism actually **decreases** in the face of $\text{EES}^{\text{REHAB}}$, despite regaining the ability to walk. Therefore, they hypothesized that $\text{EES}^{\text{REHAB}}$ "cleans up" the circuits a bit, causes activity to become targeted for walking.

To investigate which specific neurons undergo changes, they used a mouse model and performed RNAseq. They found excitatory interneurons as being enriched upon $\text{EES}^{\text{REHAB}}$. These neurons were identified by the markers $Vsx2$ and $Hoxa10$, and thought to be *V2a* neurons. These neurons projected exclusively to the ventral spinal cord and formed synapses with varying neuron types (glutamatergnic, GABAnergic, and cholinergic in no prefered proportion). The synapses were found in dense appositions (i.e., many synapses side-by-side) and SCI caused a reduction in such appositions. Ablating/inactivating these neurons alone in non-SCI mice did not halt their ability to walk. However, inactivating them did halt walking under $\text{EES}^{\text{REHAB}}$. Therefore, it is hypothesized that these neurons are specific to the injury condition and required for recovery after SCI.

The main takeaway from this paper, in my opinion, is simply the recapitulation that excitatory interneurons play a key role in recovery from SCI. This has been established now through many routes, and is certainly of considerable interest in the future. There are a few open questions, with the biggest being: how exactly does electric modulation drive circuit reformation through these interneurons?

27.1 Clinical Applications

Reduction of Pain.

This description comes from Dr. Iahn Cajigas, and this video[153]. In reducing neuropathic pain, the mechanism by which it works is postulated to be Gate-Control Theory[154]—in other words, stimulation promotes analgesia (the inability to feel pain). After this publication, spinal cord stimulators became available in 1965. The hypothesis was painful signals can be overridden by other pathways. Namely, fast, large Aβ nerves can override the smaller, pain carrying Aδ and C fibers. It is thought to occur at the level of the dorsal horn through inhibitory interneurons, but this is not known with certainty.

Failed back surgery is the most common need for spinal stimulation of this kind. Stimulation is a good treatment option if the patient's pain is localized, and no identifiable structural defects are found.

[152] https://www.nature.com/articles/s41586-022-05385-7
[153] https://www.youtube.com/watch?v=MuLbgftczf8&ab_channel=CongressofNeurologicalSurgeons
[154] https://pubmed.ncbi.nlm.nih.gov/5320816/

That is, if one's spinal cord is still improperly aligned, stimulation should not be considered. It is essential to attain extensive pre-operative imaging, as structural issues can also lead to hardware malfunction. For example, a slipped disc may cause unforeseen warping of hardware. Interestingly, psychiatric symptoms is a common indicator of poor patient outcome. Repeated operations usually indicate less effects as well.

Implanting an electrodes is done in two stages:

1. Trial: A lead is inserted in order to find the site of pain. That is, exactly where to place the electrode permanently is found first in a trial. In the trial phase, patients are usually kept awake.

2. Permanent: After such place is found, a permanent surgery can be scheduled for long-term treatment.

A variety of stimulators exist, ranging from linear leads to paddles. Paddles will require laminectomy by a neurosurgeon, but less invasive leads do not necessarily need this.

The stimulation itself has changed greatly. That is, the frequency, amplitude, etc. has been widely investigated. For example, bursts of stimulation, as opposed to continuous stimulation, has been recently tried. Notably, burst stimulation is usually imperceptible to patients, making it preferable in many cases. It is hypothesized that burst stimulation may inhibit both the physical and emotional aspects of pain.

Regaining Motor Control.

This is from this video[155]. EES is performed on SCI patients following a long period of physical therapy. The physical therapy has two purposes. First, it is to determine whether or not someone can recover at all without surgery. Two, it is to "prime" the cells for rehabilitative EES.

The operation requires complete anesthesia, placement of the electrode on the outside of the dura (i.e., epidural stimulation). It requires a laminectomy. The device they used is the same one as is used in treatment of chronic neuropathies. Importantly, the electrode is not placed nearby the injury site. In this patient, the injury was in the upper thoracic area, but the electrode was placed in the T_{12} area. They also used an EMG with electrodes in the muscles in order to trial and error the placement of the electrode. The entire surgery took $\approx$ 6h. The battery pack is also added within the patient a few days later.

Absolutely remarkably, the patient was able to regain some motor control the first time the device was turned on.

27.1.1 Paresthesia and SCS

Paresthesia is defined broadly as abnormal sensation in the skin, including tingling or numbness. In both the case of deep brain or spinal cord stimulation, surgeons use the presence of paresthesia to find an upper-limit on stimulation levels. For example, one would increase the stimulation amplitude until some tingling emerges, helping to establish a maximum amplitude. Interestingly, in the case of SCS for chronic neuropathies, sometimes paresthia is actually desired by the patients. Some slight buzzing

[155] `https://www.youtube.com/watch?v=OubpMu-2EcE&ab_channel=MayoClinic`

sensation gives some patients the feeling that treatment is occurring. For some, there is comfort associated with the replacement of pain with a non-adverse buzzing, rather than the absence of any feeling.

This work[156] describes how paresthesias are absent at a higher frequency, while they would be present at 40-60Hz when the pulse-width increases enough. On face, this is a surprising finding, as a higher frequency suggests more total charge is delivered. Gate control theory suggests touch afferents are capable of inhibiting pain afferents, providing the basis for electrostimulatory relief. Notably, the low-threshold afferents that convey touch ascend the spinal cord ipsilaterally, but those high-threshold neurons that convey pain cross over and ascend contralaterally—signifying a physical separation in such signals. One, theoretically, can simulate one and not the other.

This work uses the acronym "kf-SCS" to refer to kilo-hertz frequency SCS[157]. It remained ambiguous how high frequency stimulation was capable of blocking paresthesia, as evidence seemed to imply dorsal column (DC) axons are not activated, yet dorsal horn neurons are depressed by inhibitory neurons—leaving questions of how such inhibitory neurons were becoming active. Paresthesia is lost at the same time evoked compound action potentials (ECAPs) are lost when ramping the frequency. The hypothesis proposed by this work is that DC neurons become de-sychronized when stimulated at a rate faster than their refractory period. Notably, they used the very high amplitude of 9.6mA, which they established by finding the onset of paresthesia for low frequencies (150Hz). This was the first work to observe an increase solely in frequency is sufficient to relieve paresthesia, in part because clinicians typically vary all parameters in order to find the best treatment option. Similarly, previous works rarely extended into the kHz range. The blocking of ECAPs was verified in pigs.

Via single-unit recordings in the DRG, both kf-SCS and conventional SCS were found to activate DC neurons—which contrasts with previous findings. To further verify this, they performed both calcium imaging and immunostaining for pERK, which is said to be a marker of neuron activity. A de-synchronization due to high high frequencies was measured via the interspike interval, i.e. the time between spikes. This interval exactly matched the pulse times in conventional SCS, but were sporadic (typically around 3-6ms) for kf-SCS. They explored this computationally using a time-constant based analysis and an additional Hodgkin-Huxley-esque model. Both models supported their findings, and were then added to a spinal cord model.

Importantly, because they find asynchronous firing to block paresthesia, but not pain relief, it suggests synchronous firing is not necessary for relief. They suggest this is because inhibitory effects by such neurons integrate.

An Aside.

As it turns out, Chloride regulation is key to neuropathic pain. A great deal of work has focused on Cl- dysregulation, and how transporters like KCC2 can be manipulated to limit pain.

27.1.2 Restoring Feedback in Phantom Limb

Broadly, restoring somatosensory feedback is key to BCI advancement. As mentioned in other sections, restoration of walking in BCIs can feel "unnatural" to patients when no sensory feedback is

[156] `https://www.cell.com/neuron/pdf/S0896-6273(23)00802-4.pdf`

[157] Do we, as scientists, need acronyms for everything?

felt. Phantom limb is one of those neuroscience marvels. One who has an amputation can perceive sensation from the now missing limb. On occasion, this can result in the feeling of pain the missing limb, called phantom limb pain (PLP). BCIs employed in these patients may reduce pain, and have added benefits like improving balance and preventing falls. Indeed, that is what this work achieved[158].

The article first comments that $\approx 150,000$ people each year undergo lower limb amputation. Previous works have attempted complex surgeries to implant devices for peripheral nerve stimulation. While perhaps successful, this work sought out a spinal cord stimulation technique, which would greatly reduce the surgical variability. Interestingly, these authors previously showed that cervical SCS can be used to restore somatosensation in upper-limb amputee patients, and this work focuses on the lumbosacral spinal cord for 3 patients with below-knee amputations.

The work began by applying 1s of stimulation in the spinal cord and having each patient describe the locations they perceived sensation. Sensations perceived in the missing foot were variable over the first 2 weeks, but stabilized after. In general, higher amplitude of stimulation was required to get sensation in the missing limb. Frequency, surprisingly, did not have a noticeable effect in the missing limb—though they do not describe the domains they probed. The amplitude range of detectability was $\approx 0.6 - 4.0$mA. Interestingly, the group calculated the just noticable difference in stimulation amplitude, which turned out to be in the range of $\approx 0.05 - 0.3$mA—seemingly quite small. They also found an approximately linear relationship between stimulation amplitude and sensation. The group compared using monopoles vs. multipoles, and found comparable results for such metrics.

To provide sensation when walking, they chose a single electrode which consistently provided perceived sensation in the phantom foot. For whatever reason, they only did this for two of the three paritic-pants in the study. Upon stepping with the prosthetic foot, and under varying pressure, stimulation would be delivered. To test improvements in gait and balance, they used a Sensory Organization Test (SOT). The SOT test seemed to be designed to throw the participants off balance by moving the field of vision and or swaying a platform. Both participants tested showed improvements on this metric. Similarly, they used the Functional Gait Assessment (FGA), which assesses one's ability to walk in various conditions, such as with one's eyes closed. Here, only one of the participants showed improvement.

Finally, their reduction in pain tests were variable. Depending on the metric, different participants reported different degrees of relief, some of which were "sub-clinical", and or, not clinically meaningful.

Notably, there was considerable heterogeneity among the patients. Different patients used different devices and had amputations due to different reasons (either traumatic or from diabetic neuropathy). This work was rudimentary, in that their stimulation surgery and techniques were all built upon those used for neuropathic pain. Plausibly, more specialized systems could be helpful. Interestingly, they note that the bilateral sensation felt by these patients during stimulation was not preserved for their other work focusing on upper-limb amputees. They postulate this may be because nerves entering the lower levels of the spinal cord do so in a denser manner, are more tightly packed, and thus are more susceptible to crossover in the face of the large stimulation device. The fact that it took multiple weeks to get sensation in the phantom limb is also an important point.

To conclude, I will share some wisdom from Dr. Iahn Cajigas that I received (which I might comment on elsewhere as well—this document is getting long!). Typically, things like deep brain stimulation

[158] `https://www.nature.com/articles/s41551-023-01153-8`

or spinal cord stimulation that focus on one's perception (such as psychological disorders or pain disorders) are not the ideal routes of perfecting technology. This is because quantification relies on subjective information. Conversely, if one perfects the technology using, for example, reduction of tremors in PD or recovery of ambulation in SCI, it can be faithfully quantified. Point being: a barrier to advancing this sort of treatment will always be the lack of perfect quantification.

Chapter 28

Brain Stimulation

28.1 Deep Brain Stimulation

Deep brain stimulation is an ever increasing field of study. First popularized for treatment of tremors, it is not becoming generalized for many more disorders, including OCD. In the case of nerve injury, noninvasive methods like transcranial direct current stimulation and transcranial magnetic stimulation have been done with some success, especially when combined with rehabilitation, particularly in recovery of hand dexterity after injury. It is thought that recovery occurs through growth of descending corticospinal tract neurons. Needless to say, brain stimulation is an expanding field with many routes of exploration.

28.1.1 Technology Overview

Primarily drawn from[159]. Deep brain stimulation was first performed at Columbia by Dr. Lawrence Pool for psychiatric patients. It's use in treating Parkinson's disease was coined by Dr. Carl Wilhelm Sem-Jacobsen. Early usages were based on cardiac pacemakers. The first closed-loop works were done in 1980.

Electrode Design.

The linear electrodes we know and love are composed of platinum-iridium wires and nickel alloy connectors.
Questions for Dr. Cajigas:

1. "...integration with the extension cable surgically challenging"—what?
2. Why current over voltage? Surely you're depolarizing the extracellular space, so a clamped voltage makes more sense. I'd think you'd want a unipolar device driving voltage. Note from future Jackson: It's because current is considered to be more consistent in delivering charge. It is also easier to channel current, and or, ensure it is applied over a region. That is, you could supply a constant voltage, but said voltage drop may occur over an ambiguous region. Another comment, which is probably worth putting elsewhere, is that the "impedance" measured by technicians is very ambiguous. Presumably they perform a frequency sweep, but it is difficult to nail them down on what they are exactly doing. It's also not clear that impedance will rise over time.

[159] https://www.nature.com/articles/s41582-020-00426-z

If we take the cell membrane to be a capacitor in series with a resistor, driving a voltage to one side seems to give much more reproducible results than does current—especially when you consider different membranes to have different capacitance.

28.1.2 In the Clinic—Parkinson's and Essential Tremor

For more information on pathology and degeneration, see section **33.5**.

For this section, I'll discuss broadly how deep brain simulation works clinically, including an outline of the procedure. I'll discuss primarily from my experiences with neurosurgeons Dr. Casey Halpern and Dr. Iahn Cajigas.

Before Surgery Begins.

Candidacy (which drugs to try), HiFUS vs. DBS, etc.

28.1.3 The Surgery Itself

Deep brain stimulation is done in two steps. The first step is the primary brain surgery itself, which constitutes implantation and initial testing of the electrodes. The second is implanting the battery, somewhere in the chest, and connecting it to the electrodes. The steps are split because the duration of surgery is highly correlated with infection rate, so minimizing total length is helpful in suppressing infection whenever possible.

The first surgery takes approximately 4 hours total, from bed, to operating room, and back to bed (4 hours bed-to-bed, as I call it). It begins with a pre-operative CT scan, which is required to guide the electrodes into place. The beginning portion of the surgery is done while the patient is under anesthesia, but the later stages allow the patient to be awake. A frame is placed on the patients head, and a rotating coordinate system allows the team to precisely mark where the electrodes will enter. A surgeon scrapes away the outer layers of the skull, dura, and arachnoid space to reveal the brain. Using the frame, an electrode is inserted and slowly transverses the layers of the brain until it reaches the desired target. Such transversing is accomplished with a nearby physiologist, who records brain activity from the electrode as it is lowered into the deeper segments. Naturally, the way one does this depends on which target you are aiming for. A fascinating metric one can use, if the GPi is the target, is that just below the GPi is the optic tract. Thus, the team may turn all the lights off in the roon and flash a light on the patient's face. In doing so, they'll stimulate the optic nerves, and a faint signal can be seen when reaching the bottom of the GPi.

Once the electrode is in place, and its location is confirmed with a second scan, the patient can be awoken. The purpose of this is to do some initial testing and verify that the stimulation is not impacting nearby sites. The current is slowly ramped, and you ensure that the patient is cognitively intact and doesn't experience any tingly. If the patient has strong tremors, there is sometimes an immediate benefit. Typically, though, initial benefits are to be taken with a grain of salt. This is because the entrance of the electrode itself can cause inflammation and damage to the site, which may in turn halt symptoms. Benefits in stimulating the STN may arise quickly, but stimulating the GPi usually requires more time.

Targets.

Device Options.

Boston Scientific has most up-to-date rechargable battery, which lasts approximately 15 years.

28.2 Literature

28.2.1 Recruitment of Neurons During ES

An interesting piece of wisdom I got from Dr. Benjamin Arenkiel is that the study of endogenous neural circuits itself, while interesting, may not be as clinically relevant as the study of recruited neural circuits. Having a strong foundation of endogenous circuits, of course, is necessary for understanding what changes occur in the face of deep brain or other electrical stimulation.

Here we discuss[160]. They primarily use 2-photon imaging with global GCaMP6s labeling, and cell-specific Tdtomato labeling. The high-frequency stimulation used was 250Hz on head-fixed mice, and recording was done in the mouse visual cortex. This paper claims to be the first to record inhibitory neuron responses during electrical stimulation. While the authors do not go further into the cell-specific responses, it's worth noting their statement that: inhibitory interneurons make up $\approx 10 - 20\%$ of neurons in the brain and can be broken down into three major classes, parvalbumin (PV)-expressing, somatostatin (SST)-expressing, and vasoactive intestinal peptide (VIP)-expressing. Theoretical extensions of this work could delve in to the type-specific responses after ES, which would provide a stronger understanding of how to manipulate certain neural circuits.
They comment that, in general, low frequency stimulation is typically considered to be excitatory, while high frequency stimulation is inhibitory.

They observe a higher degree of neuronal suppression at low stimulation amplitudes. They suspect that this is due to greater inhibitory neuron activation. Because inhibitory neurons have a higher degree of axonal arborization in these cortical layers, they postulate they are more readily activated at such stimulation levels. They show that high frequency stimulation recruited inhibitory and excitatory neurons at the same rate in the mouse cortex. Interestingly, higher amplitude increased density of recruited excitatory neurons, and volume of inhibitory neurons. However, the data for this is not immensely convincing. The sample size is small, and there appears to be a trend in increasing volume of excitatory neurons, that may have been statistically significant with a higher N.

Prolonged activity inhibited excitatory neuron responses. They found that pre-stimulus activity tended to suppress excitatory neuron activity, but resulted in variable responses from inhibitory neurons. Such pre-stimulus activity-dependence was not seen when the stimulus amplitude was high enough.

28.2.2 Living Electrodes

This is from the work by Dr. K. Cullen, and featuring Dr. I. Chen[161]. Part of the problem they are trying to solve is the rigidity of standard, mechanical electrodes. To solve it, they developed im-

[160] https://www.cell.com/neuron/pdf/S0896-6273(23)00927-3.pdf
[161] https://www.science.org/doi/full/10.1126/sciadv.aay5347

plantable cortical neurons embedded within a hydrogel tract. This paper mentioned another drawback of standard, inorganic designs causing overheating of surrounding tissues—this is the first time I've seen this mentioned, and I have frequently wondered about it.

The platform they use is called microtissue engineered neural networks (μTENN). The gist is they make a hydrogel tube, culture neurons in kind of a cluster (or, aggregate), place this aggregate at one end of the tube, fill the tube with ECM-like compounds to promote axonal growth across it, and then implant the tubes where the synapses will be formed at the end opposite of the cell bodies. The columns are $\approx 400\mu$m wide, with only the inner $\approx 180\mu$m being made up of cells. They are made between 2 and 9mm in length. Action potentials can propagate bi-directionally in some models, or uni-directionally, depending on how it is made. Their experimental setup in this paper, for example, is to use a uni-directional setup with optically activated neurons, and record the induced activity through a second, bi-directional column (that is not optically activated). Channelrhodopsin-2 (ChR2) was used as their optic input, while RCaMP was used as their output reader.

In vitro, interestingly, they show that the max $\Delta F/F$ increases with increased optic power. However, they do not show it levelling off at any point—which would have good to know (i.e., where the point of max stimulation is). They also didn't show the baseline intensity over time (for more than 20 seconds, that is), which might have been interesting to see. The fluoresence plots over time seem quite variable. That is, RCaMP $\Delta F/F$ is not uniformly affected at the point of each stimulation. *In vivo*, the non-uniformity is even worse. I would assume there is some accumulation of RCaMP fluoresence over time, as repeat stimulations slowly build up more calcium. Parallelly, one might expect the baseline intensity to go down as RCaMP becomes bleached.

There are many limitations to consider. For example, RCaMP's k_{on} or k_{off}, reset time for ChR2, refractory period of neurons, the time it takes to grow neurons (which seems to be closer to a week or more), necrosis within the aggregate, etc. Too, you lack any real understanding of the neuron if using living electrodes as sensors. That is, you don't get much localization or intensity information (if any). Confusingly, from watching their supplemental videos, the potential seem to be wave-like, flowing horizontally across the columns rather than one instantaneous pulse. Perhaps this is due to the k_{on} of the calcium indicator used, but it seems to indicate some strange temporal differences in activation—i.e., the depolarizing of one neuron within the aggregate may cause the depolarization of another nearby that is independent of its intent as a sensor. This would be comparable to a spreading depression one might expect in a migraine, but instead of being within a brain, it is within this cultured cluster.

Ramblings.

What is fascinating about reading papers like this is I go into them every time thinking "Wow, this method sounds remarkable! Perhaps this will be the secret to next-gen BCIs." Yet, I always finish the paper with more questions / critiques than I began with.

Chapter 29

Muscle Stimulation

We have neglected the *muscle* part of brain-spine-muscle interfaces... until now!

29.0.1 Injectable Prosthesis for Muscle Rehab

Here we discuss[162]. This work uses hydrogels as their bioelectronic basis, but moves away from patch-designed electronics toward a flexible, injectable device. Patch-type electronics are limited in their ability to reach complex or tight surfaces in a tissue. Hydrogel-based, injectable electronics typically rely on reversible bonds, which make them more susceptible to degradation. The addition of more irreversible bonds comes at the cost of injectability. Therefore, this paper serves to develop an injectable tissue-interfacing, conductive (IT-IC) hydrogel. It was developed using phenylborate (PB)-mediated crosslinking[163]. The PB-bonds allow for rearrangement and self recovery after injection, which wouldn't be achievable with stiffer, covalent bonds. The introduction of gold nanoparticles (AuNP[164]) allowed for increased conductivity and strong coordinate bonds. They found the ratio NaOH / Au^{3+} equalling 4 to confer the best mechanical properties, so the remainder of the paper uses "IT-IC@4."

To test the biocompatability, they treated neuronal cell lines with "hydrogel releasates" and found 99% survival rate. Their mouse muscle cell injury model was done via cross sectional cutting, resulting in complete severing. Hydrogel implantation (50μL) was done in to the sciatic nerve, and stimulation immediately evoked muscle cell responses, recorded by EMG. Injection of the hydrogel (350 μL) directly into muscle cells elicited EMG recordings comparable to un-damaged muscle (around 30mV). They then tested the ability of IT-IC@4 to regenerate muscles. Using a different mode of injury where a large section of muscle was removed, they injected the hydrogel to fill the gap. It is thought that the electrical properties alone, in the absence of growth factors, would promote myocyte proliferation. Indeed, after 4 weeks the muscle cell had regenerated. The immune response, and triggered fibrosis around the hydrogel, was considered to be small. Dissociated gold particles showed accumulation in the liver, kidney, and spleen, but supposedly did not confer toxicity.

Again on the sciatic nerve, hydrogel was implanted in a spherical manner so that a conventional wire could be attached directly to it. Different mechanical movements were tested to see if mechoreceptor relay information could be read from the hydrogels. Indeed, it was. Conversely, stimulation was applied and resulted in corresponding muscle contraction. Both EMG recordings and stimulation was

[162] `https://www.nature.com/articles/s41586-023-06628-x`

[163] This is why you should have paid attention in Organic Chemistry.

[164] Again, why does everything need an acronym?

provided for both a notched sciatic nerve and a severed muscle, demonstrating success of the IT-IC@4 in both conditions.

Finally, they tested its use in closed-loop rehabilitation for muscle injury. They used a closed-loop robot-assisted rehabilitation (C-RAR) with IT-IC around the sciatic nerve and a corresponding wire meant to deliver electrical stimulation. A similar setup was used on the muscle. The robot helps the mouse walk on the basis of the EMG's recording—when the EMG reading reaches some threshold, the robot helps the leg lift. Without added nerve stimulation, this threshold could not be met, leading to loss of ambulation. Point being: it was a success.

Chapter 30

Reading Thoughts

In the case of spinal cord injury, plausibly the limiting factor in bridging from the stimulation applied by Tator and Minassian to a clinically relavent, closed-loop brain-computer interface is a reliable way to read ones objective through their brain waves. Similarly, using one's mind to control robotics in patients with neurodegeneration or other is strongly blocked by better reading techniques. Therefore, let us consider the options, a bit of history, and advancements in this field.

30.1 An Overview of Reading

30.1.1 Reading Devices and Methods.

As is inherent to device implantation, there is a tradeoff between the resolution and the damage you will induce on the brain. This is described well by Schwartz in 2006, but notably technology has advanced since then[165]. The four classes of "reading strategies" or "depths of resolution" are:

1. Electroencephalography (EEG)
2. Electrocorticography (ECoG)
3. Local field potentials (LFPs)
4. Single Unit AP

EEGs, are non-invasive and have a reading range on the order of a few centimeters. The tissue that these signals passes through largely obscures any neuron-level information. A comparable option is that of magnetroencephalograms (MEGs), which use superconducting quantum interference devices (SQUIDs) to measure magnetic fields induced by neural activity. MEGs are less susceptible to the capactive effects of the meninges and skull. ECoG are sub-dural implants with reading potential on the tenths of centimeters level. ECoG typically employ platinum-iridium or steel electrodes, but can be improved with flexible and denser electrode designs. As they come in direct contact with the cortex, safety is a much larger concern. LPS are read on a millimeters level. Single unit AP is as the name describes, and reads individual action potentials (sometimes this is written as "single unit activity" or "SUA"). Both LPS and single unit require the electrodes be buried within the cortex itself, thus causing damage to the brain tissue. These leads are typically metal, glass, or silicon. Therefore, EEG is typically preferable, as it does not require surgery. It is highlighted by Schwartz that many negatives can occur long term from invasive procedures, such as degeneration, volume displacement, or glial encapsulation. Therefore, it is likely that reading thoughts is the limiting factor in progressing

[165] https://www.sciencedirect.com/science/article/pii/S0896627306007264

the field.

Importantly, all such devices read *scalar* quantities (electrodes read voltage with respect to a ground). The term "potential" in "local field potential" is a misnomer, as these devices do not read electric fields (expanded on in a later section, **30.1.2**). LFP is also sometimes used as a general term to describe now neurons affect the voltage in a given area. I digress.

We will generally avoid discussing individual company's devices, as such devices progress rapidly, and would render this section out of date quickly. However, when instrumental to the work, we will highlight it, such as the WMAGINE ECoG discussed later. Similarly, we will discuss individual lab's work.

30.1.2 What One is Actually Reading

Before going on, one may want to reread section **2.1**, because there is discussion of "ionic flow" in the extracellular space, fields, etc. In actuality, it is ambiguous to me what this truly means in a biological context.

To begin, devices like EEG or ECG are not directly measuring the transmembrane potential of neurons or the heart—they are measuring the polarity of the extracellular space. This contrasts to something like patch-clamp recordings, which directly measure the voltage or current across the cell membrane. The heart is the largest contributor to the body's extracellular polarity, therefore ECG leads can be places around the body and measured with respect to each other. The quantity being measured is the superposition of many contributing cells, which sum together to read a voltage with respect to a ground. There is a strange thing you'll notice about ECGs if you pay close attention, and recall that an ECG requires 12 leads: the three humps (the P, QRS, and T parts of the wave) correspond to atrial depolarization, ventricular depolarization, and ventricular repolarization respectively. Yet, all three humps have the same "sign"—despite that some show depolarization and some repolarization. This is because, again, we are not directly reading the transmembrane voltage in the heart. Rather, we are taking a linear projection of the 3-dimensional, extracellular space at each of the leads, onto a single 2D plane. Point being: the measure is slightly indirect.

Let us discuss[166]. For this section, voltage is referred to by V_e, and is a scalar quantity. The difference in V_e measured at two points is an electric field, and or a vector quantity in volts / distance, and or the negative gradient of the voltage ($-\nabla \cdot V_e$). Accompanying an electric field is a magnetic field, which can be read by a magnetoencephalogram (MEG).

While the read voltage is the composite of all ionic fluxes into all cells, synaptic activity seems to have the strongest effect. This contrasts to something like fast action potentials, induced by sodium currents, which occur on the order of $1-2$ms and are unlikely to be synchronized with other neurons (therefore, when summing them with many other signals, their contribution is minimal). Because synaptic transmission requires a coordinated swing in voltage on a relatively slow timescale, it can be read via electrodes.

Positive ion influx through AMPA or NMDA channels at the post-synapse leaves an extracellular *sink* (cations flowing in to the cell). To restore electroneutrality (see section **13.1**), a *source* (cations

[166] `https://www.nature.com/articles/nrn3241`

flowing out of the cell) is required, which generates a *return current* from the intracellular to the extracellular space. This then generates a temporary pole, depending on the geometry of the space. A monopole contributes to V_e with factor $1/r$, and a dipole contributes to V_e with factor $1/r^2$ (where r is the distance from the electrode). A neurons morphology contributes to how strong of a pole it can produce. For example, pyramidal cells dominate in the cortex and, due to their long, asymmetric dendritic projections, generate a large spatial separation between source and sink. Therefore, this creates a large electric field in the extracellular space. This separation generates an *open field*. This contrasts to a *closed field*, which might be generated when multiple dendrities are stimulated in, for example, thalamocortical neurons that have symmetric, spherical projections. Notably, pyramidal neurons exist in parallel, allowing their dipoles to be compounded. Across different species the number of pyramidal neurons and the projections of neurons running in the opposite direction of them vary—leading to varying degrees of dipole cancellation. Further, the larger scale geometry of the brain, including the folding of gyri, reorient such dipoles. However, geometry alone is not sufficient to predict signals. For example, the cerebellum is highly regular but acts primarily locally, thus does not produce global synchrony.

Much like how V_e is a reading of many cells superimposed, an individual cell's transmembrane current is the summation of all of its ion channels. The magnitude of such ion channel's contribution to the overall voltage depends on where the membrane voltage is in reference to each ion's Nernst potential, and the cell's ion channel composition.

Calcium spikes may be triggered by synaptic activity and contribute strongly to V_e, particularly because Ca^{2+} spikes occur over longer timescales and confer large voltage swings. Interestingly, it was commented in this paper that little is known about Ca^{2+} spikes *in vivo*. The resonant properties of a neuron, largely contributed to by I_h (inward hyperpolarization, think HCN channels) and I_T (low-threshold, think T-type VGCCs) currents, can similarly be found when particularly synchronized. That is, many neurons may fluxtuate at particular *resonant* frequencies, and when networks of neurons are synchronized, this may be visible on an EEG. Inhibitory interneurons, for example, may synchronize to gamma (γ) frequencies in the $30 - 90$Hz range. After a period of calcium bursting, or in response to certain stimuli, higher calcium levels can lead to the opening of Ca^{2+} activated potassium channels (KCa channels). Such channels lead to afterhyperpolarizations (AHPs) that may be synchronized and are thought to produce long lasting voltage shifts (on the order of $0.5 - 2$s. Such AHPs mark and modulate state transitions of neurons, such as in sleep.

Interestingly, the magnitude of local field potential power (the square root of the Fourier amplitude, and or, the energy of the sine frequency) is inversely proportional to the frequency with some scaling. For example, it may scale with $1/f^n$, where $1 < n < 2$. It is thought that this is due to dendrites acting as a low-pass filter (which you should know all about if you've been reading properly, reread chapter **3** if not). Remarkably, high frequency signals (around 100Hz) sent to dendrites are greatly attenuated once they reach the soma, which is observed to a lesser extent with lower frequency signals (around 1Hz). Such an effect depends on the dendritic morphology and its membrane time constants, and the aforementioned pyramidal neurons act as very strong low-pass filters because of this. Variance in ion channel expression can alter conductance and input resistance of the membrane, thereby allowing a neuron's filtering abilities to change.

30.2 The Electroencephalogram (EEG)

Note that this section will receive considerable upgrades through the coming semester.

This needs clean-up already.

Introduction and Naming Conventions.

We'll begin with the canonical device: an electroencephalogram (EEG).

EEGs are effectively reading spikes from vertical, pyramidal cells of the cortex (as discussed in section **30.1.2**). Such readings are on the order of 30-80μV typically, and are a composite of frequencies (the named frequency ranges are β, α, θ, δ), with the most characteristic being called the "dominant frequency."

β waves are between 13 and 25Hz and are dominant in normal consciousness. α waves are between 8 and 13Hz, and characterize wakeful relaxation. They are particularly dominant through the occipital lobe and are in sync with the pacemaker cells of the thalamus. Upon opening one's eyes, α waves are suppressed. θ waves are between 3.5 and 8Hz, and are said to occur when neural networks are being recruited. This can be from a variety of sources. For example, priming of the hippocampus to interpret incoming signal or the combination of sensory and motor information in learning. The slowest are δ rhythms, in the range of 0.7 to 3.5Hz, observed during non-REM sleep and under anesthesia.

Lesser discussed brain waves include μ waves, within 8 and 13Hz like α waves, and observed over the motor cortex but are suppressed when motor movements are performed or imagined. There are also γ waves, which occur between 25 and 100Hz, and are associated with problem solving and may coordinate different regions of the brain.

In a waveform, *spikes* are the fast, $<\approx$ 100ms peaks that stand out above the background noise. *Polyspikes* are successive spikes. *Spike & wave* refers to a fast spike followed by a distinctive, slow wave form which is quite prominent. Spike and waves can be overlaid and or compounded, which one would refer to as *Spike & wave (theta)* if polyspikes were seen over theta waves. A *sharp wave* complex is one which occurs $\approx 100 - 200$ms, which may again occur in conjunction with slow waves.

30.2.1 Therapeutic Applications

Types of EEG Reading.

The EEG was first developed by Hans Berger in 1929, allowing for the non-invasive recording of the brain[167]. In 1968, Wyricka and Sterman showed that one could control the EEG signals read from a cat brain via reinforced learning. That is, α waves were reinforced via feeding mild to a starved cat, causing such waves to appear more frequently and correlated to behavioral changes in the cats. Similar conditioning based responses in monkeys and humans were found in the coming years. This demonstrated a correlation between brain waves and behavior. However, scientists then were heavily limited by technology, and more importantly, data interpretation. When investigations began in patients, case study findings were difficult to generalize to broadly.

[167] `https://www.nature.com/articles/nrneurol.2016.113`

EEGs can read *slow cortical potentials*, particularly from electrodes places in the fronto-central region of the skull. These potentials are marked by larger shifts in voltage occurring on the order of $1-10$s. One can also read *sensorimotor rhythms* from electrodes above the somatosensory and motor cortex, which occur in the α frequency range. Both of which mark the activity of neurons in the area, and paralyzed patients with EEGs can learn to control basic devices, including moving a computer cursor, via these types of recordings.

Another is called the *P300 event-related potential*, which relies on the strength of positive depolarization after a novel stimulus is presented. Depending on how much a patient focuses on such a stimulus, one can discriminate choices. For example, this is applied to spelling-related activities, where different letters can be presented, and patients can focus on the desired letter to spell words. Similarly is *steady-state visual evoked potentials*, where a readings are done from above the occipital lobe and can be used to select a visually presented signal. Conversely are *error-related negative evoked potentials*, which present $200-250$ms after an incorrectly picked choice and can be used to delineate between correct and incorrectly picked letters.

30.2.2 Electro-convulsive Therapy (ECT).

This section contains information learned from Semple's book titled "Pragmatic guidance for EEG interpretation."[168] This book focuses on EEG reading in the context of electroconvulsive therapy—which may not be precisely what we are after, but surely will have considerable overlap. Let us dive in.

Electro-convulsive therapy (ECT) aims to induce seizures in patients with psychological disorders, and is done under anesthesia. EEGs are used in ECT to ensure the induced seizure was appropriate for treatment. ECT-induced seizures occur in 7 phases. **Phase 1** is seen as a large increase in voltage followed by a period of low-voltage, fast activity (i.e., it looks comparable to a hill followed by a bumpy plain). This "hill" occurs on the order of 1 second. **Phase 2** is characterized by low-voltage, high frequency spikes in a recurring rhythm, which converts slowly to polyspike, which occurs on the order of 5-10 seconds. **Phase 3** is marked by polyspikes, where the amplitude rises. This marks the "early clonic phase" of a seizure. The phase is expected to last less than 30 seconds. Notably, both phases 2 and 3 may be absent, as it depends on the strength and propagation of the electrical simulation. **Phase 4** is likely to be seen and constitutes the majority of the seizure phase. It is characterized by polyspike waveforms of about 3-7Hz and lasting on the order of 10-20 seconds. This phase ends in four main waves, concluding in suppression of all activity, signifying the termination of the seizure, and or, **Phase 5**. This may occur via an immediate stop (called *abruptly ending*), slow conlusion marked by decreasing amplitude and lessened spike frequency (called *gradually breaking up*, a very similar mode of conclusion which I cannot differentiate from the previous (called *diffuse slowing*), or increased disorganization which collapses into low voltages (called *becoming disorganized*). The final stage is then **Phase 6**, which neural activity is less than the original baseline observed prior to ECT, called *postictal suppression*. This is generally around 20% suppression and is quantified by a so-called *postical suppression index*. Okay, and, the final-final stage would be that of returning to baseline, a **Phase 7**. This can take a couple of minutes to occur. This phase may not be observed, as patients undergoing ECT may be moved to a recovery unit before this can be see on an EEG.

[168] `https://journals.lww.com/ectjournal/fulltext/2019/12000/pragmatic_guidance_for_eeg_interpretation__david.20.aspx`

If sufficient seizure is not induced in such patients, a second round of stimulation will be done, with a 30s - 1min delay, with an $\approx 50\%$ increase in dose.

A Word on Setup.

Lose leads, or other technical anomalies are persistent and unavoidable. One can ameliorate these by performing two baseline tests: the first is done while the patient is awake, to ensure readings are normal, and no technical issues are present. The second baseline is performed after anesthesia is provided, which helps account for differences in baseline due to the anesthetic.

A loose lead, or a moving patient (particularly movement in facial muscles), may be visualized as rapid, high voltage spikes accompanying shifts in the baseline—giving the appearance of high noise. In the case of ECT, this may be suppressed using muscle relaxants in the facial muscles.

Seizure End on an EEG.

Determining the end of a seizure on an EEG can either be obvious, with an immediate drop off in voltage leading to clear suppression, or non-obvious with gradual petering with no clear stopping point. Having a well-characterized baseline is key to mapping the seizure start and stop. Erring on the side of a longer recording time is also preferred. The minimum amount is considered to be 10 seconds of non-seizure activity on the EEG. Sometimes, artifacts on one lead may lead to the assumption that a seizure is still occurring. Therefore, another strategy is to identify the conclusion of the seizure at the end of recording and work backwards.

30.2.3 Artifacts

Of the Body; Eye Movements.

A common artifact is muscle twitching, which is seen as ultra high frequency with a enveloping structure at baseline. During a seizure, this may be seen as a superposition of high frequency jittering on a slow-wave background. Importantly, tremors of PD patients can be seen a slow wave contractions, particularly when the tremor is in the face (such as jaw tremors).

Movements in the eye are obvious in this mode of EEG, as the two leads are positioned nearby the orbitals. The eyeball is a dipole, with positive at the cornea and negative at the retina—thus, the direction of eye movement can be detected by it's sign on an EEG. An eye blink can be seen as a depolarizing slow wave, which can be distinguished between voluntary and involuntary by duration. Upward movement of the eye shows a positive spike, while downward shows a negative spike, due to repositioning of the dipole. Blinking also causes the eye to be re-positioned (called Bell's phenomena or oculogyric reflex), where closing causes an upward movement (positive) and opening causes a downward movement (negative). This reflex can be suppressed by anesthesia. Interestingly, closing one's eye while furrowing the brows introduces more noise on top of the oculogyric reflex.

The tongue is also a dipole, where the tip is negative and the base is positive. At baseline, when the patient swallows, this dipole can be see (as the tongue moves) and generally has two peaks. Different treatments may require bite blocks, or other modes of stabilizing the tongue, making this less of a concern.

ECG interference sometimes occurs as well, and can be spotted when the signal from the left electrode is stronger and displays peaks around 1Hz. Similarly, if an electrode is placed directly above a pulsing vessel, it may show broader waves around 1Hz in frequency. Finally, slow rythmic activity associated with breathing may also be observed, demonstrated by a longer sinusoidal wave at baseline. Similar effects can be seen if sweating, or other skin phenomena, occur, causing slow baseline changes.

Of Technology.

Artifacts due to technological limitations are also common. The most common is when a single lead becomes disconnected (sometimes referred to as popping). This is seen as fast spikes on a single electrode. Most transients observed on a single electrode are considered artifacts, unless there is good reason to believe otherwise.

Large spikes can occur when electrodes are accidentally touched by other medical professionals, such as anesthsiologists bumping the oxygen mask against mastoid electrodes, generating a quick spike. Detached electrodes can lead to highly variable signal, or loss of signal. Similarly, a detached ground electrode can cause malfunctions in all other electrodes, particularly when high impedance is found in other electrodes, forcing the grounded electrode to start reading a non-zero signal. Though, this type of error is rare, as the ground is typically sufficiently protected internally.

30.3 The Flagship in Reading Thoughts

Background

Patients with ALS are of particular interest in assistive BCIs (you may review ALS in section **33.4** as needed), but extend to all patients with limited motor control. One such class is augmentative and alternative communication (AAC) devices. Such devices are complicated for locked-in or completely locked-in patients. An early, remarkable example of this was done in 1998 by Kennedy and Bakay, who implanted a device including neurotrophic factors to promote growth. The activity of neurons which grew into the device's electrodes were used to control a computer cursor.

Work from the Donoghue group came in 2012, allowing for control of a robotic hand in 3-dimensions for ALS patients. In 2015, the Henderson lab showed that an implanted neural prosthesis could be used to generate speech in an ALS patient.

30.3.1 The Works

Here we'll discuss both Stanford[169] and USCF's[170] amazing work. The senior author on the first work is Stanford neurosurgeon Dr. Jaimie Henderson, and UCSF neurosurgery chair, Dr. Edward Chang, on the second.

Henderson.

To begin, the problem: those with ALS lose much more than just their ability to walk. In many instances, they lose their ability to speak and interact with the world. The patient in this study

[169] https://www.nature.com/articles/s41586-023-06377-x
[170] https://www.nature.com/articles/s41586-023-06443-4

was unable to vocalize intelligibly, but retained some ability to move their face and produce noise[171]. Microelectrode arrays, $3.2 \times 3.2\text{mm}^2$, were placed into their area 6v (ventral premotor cortex) and their area 44 (a section of Broca's area). Neural activity in area 6v corresponded with different facial movements, which they characterized using naive Bayes (see section **12.1.1**). They used approx. 30 orofacial movements (92% accuracy), and 40 phonemes (62% accuracy), and 50 words (94% accuracy). Area 44 was less helpful in discriminating words, and therefore was not used in further study[172]. The two arrays placed into area 6v gave different, but both helpful information regarding movement and speech production.

Their Neural Net outputs a probability once every 80ms[173]. Notably, it must be retrained regularly, as neural changes occur even on an intra-day level. After a training period of approximately a week, the model could predict, in a 50-word vocabulary, with a 9.1% error rate. On a 125,000 word vocabulary, it was 23.8% error rate—thus, there remains room for improvement. Such error rates were slightly worse when the patient was silent (vs. when attempting to vocalize) and at a rate of approximately 60 words per minute. Offline testing showed significant improvement in decoding, with errors reaching closer to 10%—thus, thus suggests room for optimization. They also make the claim that doubling the number of electrodes will half the error rate.

In analyzing phonemes, they vectorized electromagnetic articulography data. This is essentially the mapping of facial movements together with phonemes—which is particularly difficult, as it was impossible to discern the exact attempted phonemes of this patient as they could not speak ineligibly. Thus, they based their data primarily off of able bodied patients at first. They compared this data with neural representations and with their patient, and determined that the so-called "articulatory code" for phonemes is preserved.

Chang.

Here, an ECoG on the articulatory vocal-tract is used on a patient with severe paralysis induced by a brain stem stroke. They achieve a median word rate of 78 per minute and an error of 25%, on par with the previous paper. The ECoG had 253 electrodes and was placed over the sensorimotor cortex and superior temporal gyrus—regions involved in orofacial movements.

Training was performed by presenting the patient with a phrase with which they would attempt to say, but were unable to due to paralysis. High frequency signals (called high gamma activity, and with a frequency of 70 to 150Hz) and low frequency signals (around 0.3 and 17Hz) were extracted (see chapter **3** to review). A considerable trouble was that, like the previous patient, no ground truth exists for their speech. Therefore, proper timing information for phoneme production is unknown. To get around this, they used a connectionist temporal classification (CTC) loss function and attempted to map the timing.

Decoding speech from brain waves was done on the basis of predicting the probability of 39 phones and silence in a recurrent neural network. Interestingly, they showed that the greatest source of error was the RNN, rather than the CTC. They also demonstrated very minimal increased in error rate

[171] Somehow this patient is involved in BrainGate2, but it isn't clear to me how.

[172] I include this only to say that this is very interesting! It is surprising that they could go so far as to implant the electrode array, and it end up not being very helpful.

[173] I still think this will be a problem in the future. In this case, it is likely fine, but for the BSIs, it seems insufficient as it is bordering on slower than the average human reaction time.

after 61 days between tests, which showed that the training is robust and neural differences in these months was minimal.

The final thing they did was synthesize speech audibly using a speech representation learning model called HuBERT. They similarly rendered a 3D avatar to emit the speech from in order to convey additional emotional complexity. The avatar's movements were driven by articulatory-movement attempts via recordings from the precentral and postcentral gyri—and recordings from the postcentral gyrus were key to decoding.

Takeaways.

Importantly, no aforementioned patients had complete locked-in syndrome. A brainstem stroke, too, is capable of causing locked-in syndrome, but not in this example. Prior attempts to train patients on implanted or superficial devices in CLIS patients have been unsuccessful. Some speculate that these sorts of devices will never be possible for CLIS patients, as it's plausible that their neural architecture is sufficiently altered by late stage disease that reading it is untenable. However, some works have shown error-related potentials are still intact in such patients, suggesting there is hope. Some speculate that in spite of this, an almost complete disconnect between stimulus, response, and intention will make it impossible for training with devices to occur. Further, because ALS patients may have degeneration in the basal ganglia and other crucial areas for reinforcement learning, it seems plausible that this skill is lost in far along patients. Theoretically, one could circumvent this using classical conditioning, rather than instrumental learning, as this requires no cognitive appraisal. Indeed, this was successful in some trials, where a year of training under classical conditioning enabled CLIS patients to answer simple yes-no questions.

Regardless of what the skeptics say, I, a self-diagnosed chronic cynic, still think it's silly to suggest this is a hopeless adventure. What we know of the brain, and our development of BCIs, improves so remarkably each day. It's hard to say with certainty that anything is truly unachievable.

30.4 Literature

30.4.1 Reading LFPs Without Surgery

As discussed elsewhere, there is a trade off between the invasiveness of a surgery and the level of information one can gather. However, vascular passages allow for the insertion of small microelectrode arrays that may be able to read deep neural signals without requiring tissue to be dug into. A 2023 article did just that[174]. Their array was injected through blood vessels, and requires vessels of $\approx 100\mu$m. For reference, the smallest arteries are around $\approx 100\mu$m, and arterioles below this threshold. Therefore, this is usable in a wide range of vasculatures. Such a device is considered a micro-endovascular (MEV) probe. Other attempts at this have been done in the past, but a primary limiting feature is that metal electrodes are inflexible and thus cannot be navigated through the array of vessels in the brain. One of this paper's primary advancements, then, was making it ultra-flexible. Like other groups, they use an Intan device to read and process data.

The device is injected through a microcatheter, which itself has a much wider diameter. Therefore, the location of implantation is limited in that it must be accessible by a wider vessel nearby. In their

[174] `https://www.science.org/doi/10.1126/science.adh3916`

example, a large vessel branched off into two directions. To select between the two branches, they had to change the mechanical properties of the device itself[175]. To do this, they stressed different parts of the devices mesh in order to bias its turning direction.

Their testing was done in anesthetized mice in both wildtype and disease models (such as induced seizures). In such circumstances, they were able to record local LFPs. In some instances, though, they were able to get single AP recordings on individual electrodes. Interestingly, this paper included both the filtered and unfiltered readings, providing some nice insight into the benefit of bandpass filtering!

Some confusions: I'm not sure exactly if this is meant to be a permanent addition or not, or if it is removable atraumatically. Similarly, I'm not understanding how they are actually reading from the device. Their setup in the supplements seems to require anesthetizing the mice and doing surgery to attach the electrodes to the Intan device—which I assume cannot be true, lest the whole purpose is defeated. So my wonder is: are the mice walking around with wires hanging from their skulls, or are the wires inaccessible without surgery? I should probably read more closely to understand...

NOTE: A K. Cullen paper cited in the above article[176]. Read later.

30.4.2 Neurograin

Upon re-rearding, I'm noticing more and more how half of my older writings was just me being confused. Comical.
Neurograin (a type of ECoG) is an interesting development that includes both the ability to read and stimulate the brain from small, rice-grain sized chips[177]. Neurograins are individual pieces, providing the immensely powerful ability to scale your device. They verified their method using 48 electrodes, but suggested the ability to scale to up to 770—providing quite impressive resolution.

Some confusions:
The core achievement of this paper is in developing microchips of less than 1mm in size, that are able to read and transmit information. However, I am slightly confused by the intended use. In their Figure 1, they show the neurograin array beneath the dura, and a wireless power/download port wirelessly connect from above the skull. It isn't clear to me how they intend to use this—as they'll need to remove the dura in order to implant the device. A possibility application would be to make a network that can be "woven" into the spinal cord, allowing you get to readings with minimal (if any) laminectomy.

One large question I have is how they are processing this data. In the case of a PET scanner, which has a similar 1GHz+ sampling, instead of overloading the computer with billions of data points a second, it has a 2-step sampling approach, where data is discarded unless it is "non-zero." That is, data is transmitted only when light hits a reciever. In this case, I'm not sure if that is possible, since brain waves will not be binary.

The bulk of the paper is very electronics heavy, so it will be good to reference in future builds.

[175] This is likely a considerable limitation. It is unlikely that this sort of device will ever become generalized or patient specific. However, its use in research might be considerable.
[176] https://www.frontiersin.org/articles/10.3389/fnins.2019.00269/full
[177] https://www.nature.com/articles/s41928-021-00631-8

30.4.3 Robotic Control with Eye.

Here we discuss[178]. Usually, when we think of neural prosthetics or brain-computer interfaces, we think of their use in treating spinal cord or traumatic brain injury. However, there are many of really fascinating uses of brain computer interfaces beyond this. One such example is in ALS. Interestingly, muscles of the eye are unaffected by ALS. Therefore, while the rest of the body loses motor control, theoretically eye movements can be used to interact with the world. This paper used EEG signals and resolved eye movements—which are normally considered noise in an EEG.

EEG, being a non-invasive reading technique, is ideal for patients. Notably, unintentional, automatic blinks produce a smaller spike on an EEG than do intentional, voluntary blinks. These researchers used 5 subjects and attempted to monitor voluntary blinks as well as left and right looks in order to control a robot. They did so through some basic thresholding algorithms, followed by a training algorithm. They used this approach for real-time processing of EEG data. As a side comment, they wrote this in Python, which is surprising to me, because EEG produces a ton of data, so I'm surprised Python is fast enough to process it all "real time." They did mention, though, that their real-time algorithm is much less accurate than their "offline" algorithm, which may be due to differences in how they process data to speed it up. Their accuracy was in the 75-80 % range for the three movements they tested.

Long term investigation and optimization of eye tracking will be helpful for a variety of neurodegenerative diseases, so this is a helpful step in this path.

30.4.4 Computing via Organoids.

Did I finish this section? I can't remember...
Here we discuss[179]. These authors begin by commenting that driving artificial neural networks (ANNs) is extremely energy intensive. Data and data processing is physically separated, known as the von Neumann bottleneck. There seems to be a strong need for advancements in computing on a hardware level. Some attempts, such as through resistors containing memory (memristors, bad name) to achieve this. This work went in the opposite direction, and instead pushed for more biologically-based networks—they dubbed this Brainoware.

Via culturing a variety of early stage neurons of different identifies on a multi-electrode array (MEA) they were able to build a "computational reservoir" capable of unsupervised learning. They found success in tasks such as voice recognition, which improved through multiple epochs. That is, electrical stimulation delivered to the organoid was capable of discriminating between different voices. Notably, the training process takes multiple days. Similarly, producing organoids takes months, and maintaining them is difficult. Naturally, "harnessing their computation" requires keeping them alive, and an organoid's typical necrotic core would present problems in this regard.

[178] https://www.nature.com/articles/s41598-023-44645-y
[179] https://www.nature.com/articles/s41928-023-01069-w

Chapter 31

Synthesis

The purpose of synthesizing the aforementioned parts together to create a brain-spine-muscle interface is so that you may create a closed-loop, feedback system allowing for rehabilitation.

31.1 The Flagship Brain-Spine Interface

This section needs to be improved greatly. I wrote this about [10?] months ago, back when I was much dumber.

[Currently Frankenstein-ing this section together—I was such a fool before!!]

Background.

Here we discuss 2023's quintessential work[180]. But, before we do so, let's talk about the history of BCIs in SCI. Naturally, the earliest integrations of BCIs in treating SCI were EEG-bsaed. Implanted reading devices allowed for significant improvement of reading brain waves, and a landmark example came in 2006 when the Donoghue group allowed for prosthetic device control upon implanting a 96-electrode device in a patient with upper-spinal cord injury resulting in tetraplegia, allowing for cursor control[181]. In 2013, the Boninger group used a 32-electrode ECoG in a tetraplegia allowing for computer cursor control[182].

Back to Courtine. As an aside, this group's works primarily uses the term "BSI" for brain-spine interface. I'll continue to use BCI as our catch all term. In 2012, they demonstrated the potential of using SCS-based BCIs in restoring movement in a mouse model[183]. The large issue, as briefly commented elsewhere, would be "closing the loop" in human patients. That is, how does one match stimulation and intention to induce plasticity? This 2016 precursor by the Courtine lab focused on an NHP model[184]. In rhesus monkeys, they implanted a 96-microelectrode array into the area of the motor cortex expected to control leg movement. Via mapping motor activation onto the spinal cord, they identified hotspots which helped select their desired stimulator location, which was centered on the corresponding dorsal roots. It is thought that EES activated motorneurons by recruiting large

[180] https://www.nature.com/articles/s41586-023-06094-5
[181] https://www.nature.com/articles/nature04970
[182] https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0055344
[183] https://www.science.org/doi/10.1126/science.1217416
[184] https://www.nature.com/articles/nature20118

proprioceptive fibers in dorsal roots.

To capture data, they recorded from their 96 μ-electrode device and transmitted data wirelessly to a computer. On this data, they performed bandpass filtration (with corner frequencies 0.5 and 7.5 kHz). Spikes were designated as deviations 3 times larger than the standard deviation. The data stream was captured in 150ms overlapping windows updating every 20ms, computed into component vectors, and characterized via linear discriminant analysis (LDA). The only three states they had were "foot strike," "foot off," and "neither." Pulses were sent to the spinal cord stimulator after a probability of state transition above 0.8 was reached—and the time it took from signal detection to stimulation was over 100ms (quite slow!).

Their spinal cord injury model included hemisection slicing and led to paralysis and partial recovery. They paper only used 2 monkeys, and their control group was simply these same monkeys when the device was turned off. When turned on, and without prior training, the device was capable of immediately restoring locomotion. Such principles were observed in mice as well in prior works. Perfecting the frequency of stimulation improving stepping, and similar stimulation not under brain control (i.e., delivered continuously or at random) did not generate successful stepping.

Now, it's time to translate to the final model organism.

Overview.

The goal of this was to restore ambulation in a patient with incomplete cervical (C5, C6) spinal cord injury. The patient had lived with the injury for the past 10 years, unable to walk. Years of physical therapy did not make much progress.

In order to map their intended motor movements, the group used surgical implantation into the skull, over the motor cortex, to record patient movements and wirelessly transmit this to the spinal cord. The technology used to capture these electrocorticographic signals was the WIMAGINE ECoG. The design features two antennas, the second of which transmits motor signals that are to be decoded and sent to a pulse generator. The pulse generator was the ACTIVA RC, the same that is used in deep brain stimulation, spinal cord stimulation, and pacemakers in the heart. Pulsed feedback was then delivered to the spine through a typical spinal cord stimulator, as would be used in treatment of neuropathic pain or other. Notably, this group did not directly develop any new technology—everything was already available and commonly used. Therefore, the key advancement of this work was not invention, per se, but rather *synthesis*.

31.1.1 WIMAGINE ECoG

I'll briefly discuss the ECoG used[185]. The device features 64 electrodes, and a patient would need to have two implanted (128 electrodes total) in order to read both sides of their cortex. It is implanted over the dura mater (i.e., it does not make direct contact with the motor cortex).

The device was tested longest in sheep, which was a 10 month trial. Surprisingly, at the end of the 10 month trial, through GFAP staining, they still found a great deal of glial migration and or build-up around the site. They did not quantify this—which is curious. It isn't clear if this residual gliosis

[185] `https://www.frontiersin.org/articles/10.3389/fnins.2019.00847/full`

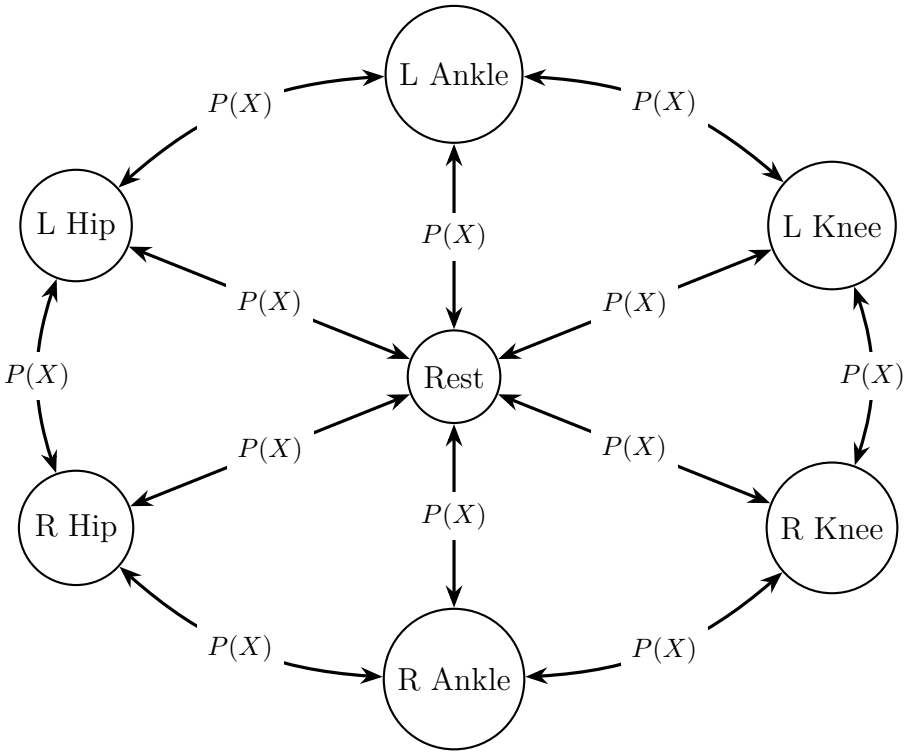
is clinically relevant. Secondly, calcification of the skull over the device had begun by thend of the 10 month grial. This is not inherently surprising, and perhaps is a good sign that the bone was not irreversibly damaged. However, as the device is wirelessly charged, and wirelessly transmits brain information, one must wonder how years of calcium buildup may impact the ability to send or receive information. Too, if the technology needs a dust-up, or removal, the surgeons will have to re-destroy these calcified layers. There is quite a lot of fuss made about the fact that the device is wireless, and how it was a decision made to best serve patient comfort. Notably, though, patients are having two large, 50mm arrays placed into their skull, so making the device wireless does not save a great deal of invasive-ness per se. It is interesting to ponder whether an 8×8 array is sufficient to extract all of the signals. Of course, this is the trade-off one must make in choosing between EEG, ECoG, and LPS, and for the purposes of their study, it seems more than sufficient. The WIMAGINE system digitizes information with 12 bits of resolution. Data is processed through pwelch spectral analysis (an implementation of the FFT).

31.1.2 The Work

In short, they created a device can be calibrated and programmed quite quickly. Device feedback, in combination with rehabilitation, over the course of months eventually resulted in restoration of ambulation.

Seven States.

Data from the ECoG was processed and used to edit the probability distribution of a Markov chain, which progresses the electrode paddle through seven defined states defined below. Such states differentially stimulate the spine, providing closed-loop feedback. A graphic of these 7 states are provided below, where only a single body part can be stimulated at a time:



The three shortcomings explicitly mentioned in the work's Introduction are that (1) one using this BCI must have motion sensors on in order to compensate properly, as the BCI has no direct feedback

information from the muscles, (2) that the patient's movement was not perceived as natural, and (3) that there was still considerable hurdles in traversing variable terrain. Many more shortcomings exist than this, and we will go through them below.

Importantly, the patient was able to walk without the BCI turned on (using crutches) after some time. This signals many things. First, the barrier to recovery is not as large as we previously thought. This patient went 10 years without success in rehabilitation, so one would assume their CNS was completely implastic. Further, that treatment 10 years post injury can be successful is quite a noteworthy point. Second, some residual function can be converted to natural recovery. The patient was able to evoke some small responses in EMGs, but not enough to regain movement. Therefore, these pathways seem capable of being enhanced by the BCI. We can ponder how necessary feedback from the brain is, but it is likely that potentiation from neural signals is important for inducing circuit reformation and plasticity.

Another consideration is that their signals are "un-sophisticated," as presented in this paper. That is, they use 16 electrodes in the spinal cord stimulator. Only seven states seems to limit the mobility of a patient. It is likely, then, that this stimulation hijacks a more reflex-based approach to walking—hence why one can not traverse variable terrain that goes beyond reflexive movement. Perhaps future devices with more complex stimulators and more states will be able to control other features of walking.

31.2 The Flagship in Synthesis

Founding Works.

In large part, tremors induced by PD have already been well solved through deep brain stimulation and high intensity focused ultrasound. However, this paper focuses on some of the non-tremor symptoms of PD, and the idea that one can use spinal cord stimulation to fix degenerative diseases is generally interesting, so let us learn from[186].The premise of this experiment is somewhat surprising, as stimulation is often thought to interrupt signals (as is the case in deep brain stimulation for PD, and SCS for neuropathy). In this case, though, they believe SCS **enhances locomotor signals**. That is, they are not treating PD tremors, but rather are treating PD induced akinesia. Experiments were performed in mice using dopamine-transporter knockout (DAT-KO) as their PD model. They also injected tyrosine hydrolase inhibitor (called AMPT)[187] to further deplete dopamine. Dorsal root stimulation was performed using electrodes into the upper thoracic levels of the spinal cord. The greatest locomotor improvement in locomotion was found at 300 Hz stimulation.

Notably, some hospitals are currently attempting clinical trials for spinal cord stimulation's use in treating Parkinson's. Some efforts include blocking pain and treating akinesia. Note from future Jackson: When I first read this paper a few months ago, I thought it was a bit silly. It was difficult for me to see why one would attempt this, when DBS is already so powerful. I figured it would go no where. Yet, read on:

[186] `https://www.science.org/doi/10.1126/science.1164901`

[187] I didn't bother Googling what this is, but intuition would suggest is aids in the conversion of tyrosine to catecholamines.

31.2.1 SCS-DBS Intersection

Here we discuss[188]. Dr. Ashwin Ramayya, an MD-PhD neurosurgeon I had the pleasure of talking to, commented that after seeing this work his first thought was "you don't have to be the best at everything." Indeed, this work is so awe inspiring that it makes one question their place in the field.

DBS treats a very small subset of symptoms associated with PD. Because of the success of EES in treating SCI, these groups wondered if a similar feat can be accomplished in PD patients. The first bit of this work was done in an NHP model (rhesus macaque). Because of differences in monkey and human mobility, they developed a scoring technique to analyze PD models in NHPs and compare them to human PD symptoms. The preferred PD model for them was administration of MPTP (see section **33.5**!), a drug able to mimic the death of dopamine producing neurons. They compared spinal circuits before and after MPTP and identified "hot spots" in the spinal cord. In implanting these devices, they stimulated different electrodes in the array in order to test their ability to cause muscle contraction.

After these tests, they then aimed to use a brain-spine-muscle interface as they had in their other foundational works. Microelectrode arrays were inserted into the cortexes of RHPs and were able to regularly decode motor signals. They then furthered this by inserting EMG electrodes into the legs of NHPs[189]. The neuroprosthesis together improved natural movement and posture. Finally, they added DBS to their circuit and found further improvement of movement.

In humans, the ability to decode motor movement may be difficult in people with PD. Therefore, their first step was to recruit participants willing to get electrodes implanted into their cortex for reading their intended movements. They considered this a success and moved forward. They enrolled a patient in their STIMO-PARK clinical trial. The patient already had DBS, but still had some severe symptoms. They used a SCS paddle used in treating pain, connected to an Activa RC IPG, guided in location by CT and MRI. Rather than implanting an ECoG to read the patient's motor cortex, they instead used motion sensors and attempted to read motor intentions and send updated pulses to the stimulator that way. Fascinatingly, DBS and EES seemed to treat different symptoms at times, and act synergistically. That is, DBS^{ON}-EES^{OFF} showed no improvement in "fraction of time frozen," while DBS^{OFF}-EES^{ON} showed total recovery. Their benefits were generally additive, though.

[188] `https://www.nature.com/articles/s41591-023-02584-1`

[189] I am really wondering if some nice code can use this kind of data to iterate and self correct. Need feedback, of course.

Part VI

Some Ideas

Man takes up the sword in order to shield the wound in his heart sustained in a far-off time beyond remembrance. Man wields the sword so that he may die smiling in some far-off time beyond perception.
– Berserk by Kentaro Miura

Chapter 32

Ramblings

This is all garbage—DO NOT READ.

This section will largely be messy—likely for the coming years. It is for me to jot down ideas I have through reading.

32.1 Idealized World

Regarding ECoGs.

Better ECoGs are really the path to incredible options. Of course, we have the mechanics already. Theoretically, one could simply build a mech-suit to accomplish any task—the trouble is getting it to respond. Somehow, it seems certain groups are able to resolve speech patterns and handwriting from brain data—so certainly it is within the realm of possibility. In this same way, ECoGs must become less invasive. The route to accomplishing this is likely through extremely intelligently made soft electronics. The resolution must be greatly improved by the inclusion of many, many more electrodes. In this way, powerful machine learning will be needed to process this data[190].

There are a few things to consider. First, glial encapsulation will probably destroy you if you try to use too small of electrodes. Second, hard electronics seem to have many drawbacks. If you are using ultra-micro electronics, they are still limited by their ability to read within the gyri, given that the electronics will be hard. Similarly, you simply cannot afford to do massive cranioplasties on a regular basis. It'll need to be improved. Accompanying this issue is the ability to test on humans. One must first test on mice, which parallelly means that the device must be scalable.

Some Experiments.

1. Inhibit L-type VGCCs—likely need to use some kind of hydrogel with antagonists built in.
2. Optimize stiffness of paddles—or hydrogel, if we use the ACES system.
3. Optimize ECoG—that is, take up less space, maintain resolution, etc. to minimize invasiveness of surgery.

[190] I should really start teaching myself ML.

Inhibit L-type VGCCs.

It is well known by now that electrical stimulation combined with rehabilitation improves patient's ability to walk after SCI.

ACES Method.

I like the ACES method as a starting point. It seems easy enough to implement, and can be built upon to arrive at some SCI method. Of course, I like that it is already integrated into a hydrogel setup. Some mild concerns are the 2 mA applied. It seems like quite a bit.

Biomaterials.

The biomaterials possibilities are quite cool too.

CSF-cNs.

A great, easy experiment is CSF-cNs ablation with EES, compare vs. WT, and see if recovery is similar.

Waveforms for Neuropathic Pain.

Perhaps one can use a feedback mechanism, where the frequency of stimulation can be altered in order to dampen neural signals. You'd need to both read and write, essentially. Surely you can record brain waves responsible for feeling pain, and have a circuit edit itself to minimize them? $--> $ actually, dampening neural signals is a great example of a time to use the peak detection method, because you're only interested in magnitude.

32.2 Device Idea

Another idea: frequency encoding of stimuli. Start by using EMG signal, digitizing it, using FSM to select a frequency output. In the distant future, this could be used to restore sensory perception after SCI or other injury.

32.2.1 Motivation

Much is made of the frequency and amplitude of neuropathic pain stimulators. Dr. Ben-Haim even called the search for such as being the "waveform revolution." Perhaps one can avoid this investigation altogether by simply connecting the entire pain pathway to the feedback loop of an opamp. That is, an opamps v_+ is grounded, the V_{out} is tied to a electrode in the spine, and the v_- is connected to some form of ECoG in the brain. Theoretically, the opamp's V_{out} will do whatever it can to cause the signal read at the brain to go to 0. Of course, you cannot assume the opamp's feedback will be predictive, nor can you assume it will fix sharp pains. But, it may be able to fix chronic pain. Let's think of an overall build.

32.2.2 Framework

Processing Brain Waves.

To process the brain waves we will likely:

1. Filter out background from individual electrodes.
2. Subtract and amplify from some reference.
3. Do some summing amplification from each of our electrodes.
4. Rectify our signal.
5. Convert via peak detection.

Feedback.

To implement, we will then:

1. Feed processed signal to a PID controller of some sort.

The only thing I am not sure about is if we will need a separate PID controller for each electrode in the spinal cord. I am not sure if we can just take all of the brain signals and combine them together, but I would assume that is fine. Another question is how much current we will need. If too low, we can add a push-pull to boost, but I doubt we will need a high current.

THANK YOU.

A FINAL NOTE:

The world is always full of the sound of waves.

The little fishes, abandoning themselves to the waves, dance and sing and play, but who knows the heart of the sea, a hundred feet down? Who knows its depth?

– Musashi by Eiji Yoshikawa

Part VII

Addendum

Overview

Hello there! You've reached not only the end of the book, but actually the parts beyond its conclusion! At least for the moment, there are topics I'd like to discuss, but they're only tangentially related (if at all) to the idea of brain-machine interfaces. Therefore, consider this the bonus parts that are not worth reading, unless you are immensely inclined.

One such topic is that of neurodegeneration. Naturally, there are not many direct parallels to be drawn from diseases like Alzheimer's and spinal cord injury. Yet, I do believe strongly in Musashi's epithet, that once you know the way broadly, you can see it in all things. For this reason, I think there is unknowable value is learning about degeneration. As it turns out, there are considerable parallels between the pathologies of traumatic SCI and neurodegenerative diseases, like degenerative cervical myelopathy (DCM) as covered in other sections like **20.3**. Thus, it is certainly worth understanding in-depth. Oh, and because I took a Molecular Genetics of Neurological Disease course and need somewhere to save my notes ;)

Chapter 33

Neurodegeneration

Much of this comes from Dr. Nancy Bonini's Molecular Genetics of Neurological Disease course.

This is more of a historical dive into the research, because so much of what we know about neurodegenerative diseases is likely wrong, overstated, or misinterpreted. There are many fascinating theories and little agreement. You will hear some, like Dr. Rothstein at JHU, argue that there is a viral component. You will hear some, like the Wallace lab at Penn, claim that Alzheimer's is due to mitochondrial dysfunction and not amyloid aggregation. So for now, let's call this a historical review of the research up to this point.

33.1 Alzheimer's

Overview

Alzheimer's disease (AD) is a neurodegenerative disorder found in many elderly patients, and the disease confers memory loss and other forms of cognitive decline. It can be familial or sporadic, and while onset typically begins much later in life, some early onset forms exist. Sporadic Alzheimer's is so ubiquitous that some hypothesize that it borders on the normal aging process. It is this ubiquity that makes it the *holy grail* and or the *white whale* of disease—solving it would help endless people. In the United States, it affects around 6 million people, making is more common than all cancers combined[191,192]. It was first characterized by Alois Alzheimer in 1906, who used dyes to stain tissues and reveal their histology. He found strange plaques present outside of neurons, and stringy tangles within them. Alzheimer's disease must be diagnosed after death because of considerable overlap with other disease symptoms—that is, nailing down memory loss to specifically AD is impossible without analyzing brain histology. Such a diagnosis is done on the basis of extracellular amyloid plaques and intracellular tau tangles, discussed later.

While, generally speaking, we are helpless in tracking and managing AD progression in patients, some tools exist. For example, a drug called Aricept blocks acetylcholine degradation, which can slow symptoms for a short time. Measuring amyloid buildup, or other markers, in the CSF is a tool to determine disease progression and or drug efficacy. Congo red is used to stain any amyloid buildup—that is, it

[191] https://www.alz.org/alzheimers-dementia/facts-figures

[192] https://www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures/cancer-facts-figures-2022.html

stains any β-pleated sheet and is not specific to the β-amyloid of AD, but still can be of help. Another method is using thioflavin T (ThT), which also interacts with β-pleated sheets. This has been used to make a radioactive compound called Pittsburgh compound B (PiB), which can be imaged on a PET scanner, and used to quantify the amount of plaques in a living patient. PiB is presently the best available compound and helps identify AD years before dementia begins. Other attempts have been generate compounds like PiB, but none have reached its standard. Still, though, PiB's predictive accuracy is still quite low.

All of our modern understanding of AD comes in just the last few decades. For the purpose of these Neurodegeneration sections, put yourself in the mind of someone in mid-late 1980s: genetic testing was a blooming but still a young field. Sequencing genes was immensely difficult, but worse was finding the correct region to sequence. Linkage studies helped give indications of location, but one still was leagues away from nailing down a couple kb region to sequence. It was completely unknown what genes, if any, cause AD. In fact, only a couple genes and proteins were known at all. We were totally blind to the human genome. But, we did have something: the protein.

β-amyloid was purified in 1984 by Glenner and Wong[193], which allowed it to be sequenced early. The protein was small, but congregated together into plaques. They generated an antibody against it, allowing them to stain and identify regions of the brain affected by AD. After the β-amyloid protein was identified, other groups sequenced the surrounding nucleotides and found it to be part of a larger transmembrane protein—which was eventually called the amyloid precursor protein (APP). Thus, it was intuited that this little peptide was cleaved from the larger transmembrane protein. Fascinatingly, trisomy 21 patients almost ubiquitously get AD, and specifically early onset AD—suggesting there must be some association with Chr21, and perhaps some correlation to dosage. Other genetic linkage studies confirmed this, showing that in familial AD, there was some segregation with Chr21. So, we had the protein, and we have a general idea of location. Where do we go from here?

Finding APP

As mentioned, the Aβ protein was first sequenced in 1984. In 1987, four groups cloned the APP gene. However, it remained completely unknown if this protein was causative. It was in 1991 when Goate et al. confirmed the existence of causative mutations in the APP gene[194].

To begin investigation, groups had to focus on familial AD (fAD), otherwise finding a non-inherited mutation would be nearly impossible. In these days, only a small set of chromosomal markers were known. However, these markers allowed for the powerful scanning of different regions of the chromosome. That is, you could compare markers between different family members, indicating different degrees of recombination. Family members sharing the disease and similar chromosomal markers allowed you to narrow down to a region of interest, implying the causative mutation was somewhere nearby. This was quantified by logarithm of the odds (LOD) scores, where more positive numbers meant the association was more likely. Generally, LOD scores above 3 give a high level of confidence that the disease segregates with said marker. This required finding large enough pedigrees where this kind of analysis was possible. Of course, misdiagnoses would immediately ruin these sorts of analyses. Similarly, heterogeneity in a disease would greatly complicate things. Indeed, this is why APP, for a long time, was not thought to be the causative gene in AD—many families showed no segregation

[193] https://www.sciencedirect.com/science/article/pii/S0006291X84801904?via%3Dihub
[194] https://www.nature.com/articles/349704a0

with APP at all.

In their first early onset AD pedigree, Goate and colleagues found family members to exhibit two different alleles, which correlated exactly with AD. However, two unaffected family members exhibit recombinant alleles at different sites. The first individual shares the same allele at the telomeric end of the chromosome to their affected parent. This individual was 15 years over the mean age of diagnosis for AD, giving a high degree of confidence that they do not have the disease. This suggests the AD causing gene is not, in this family, telomeric to the APP gene. The second individual shared the same allele as the affected parent only centromeric to APP, and at the time of the study was unaffected. It's possible this individual would develop AD later in life. Therefore, based on their current pedigree, APP was the clearest possible disease causing gene, but they did not have absolute confidence.

From there, they used PCR to identify the site of mutation and developed primers specifically around the βAPP. This was a natural choice, as Aβ was already identified in AD plaques and, fascinatingly, dysfunction of the β unit was implicated in other diseases already. Via direct hand Sanger sequencing, they found a valine to isoleucine polymorphism in the βAPP encoding portion.

Digression: Restriction fragment length polymorphism (RFLPs)

Harnessing the power of restriction fragment length polymorphism (RFLPs) immeasurably advanced genetic testing. The benefits of RFLPs are that mutations cause a restriction enyzme cut site—allowing for very easy screening. Researchers would look for polymorphisms around a disease causing gene, which served as an easy technique for getting a broad sense of the genotypes in a pedigree. In other words, it helps you screen for markers segregating with the disease causing gene. Better yet, when a polymorphism that's thought to be disease causing is found, one can see if it alters a possible restriction enzyme cut site. For example, perhaps the polymorphism converts an *Eco*RI cut site into a non-cut site. Now, to screen possible patients, instead of sequencing every patient individually, they could use PCR and restriction enzymes and analyze the bands on a Southern blot (two would signify no mutation, one would signify mutation), significantly cutting down the time, effort, and cost of diagnosis. While sequencing may take closer to weeks, such enzyme gels can be done in an afternoon.

Indeed, a *Bcl*I restriction site was induced by the polymorphism, allowing them to detect the presence of the mutation on a gel. In this case, a non-affected person will have an uncut sequence of DNA of 319 bp in length, while those affected individuals will have a band at 199 bp and 120 bp. They were able to screen other families and find the same mutation. To verify these families were not related, they checked other polymorphic sites and found divergence, suggesting sufficient genetic dissimilarity.

Therefore, these authors identified a site on Chr21 that segregated, sometimes, with early onset Alzheimer's disease and in the gene encoding amyloid precursor protein (APP), specifically within β-amyloid. Importantly, this mutation was conserved between two fAD pedigrees, implying it was indeed causative. Because segregation of this gene is not found in all AD patients, the authors concluded there must be considerable heterogeneity among early AD patients. They also hypothesized that APP's lack of causative support may be due to previous studies using patients with misdiagnosed AD with other diseases that phenocopy it. Notably, they did not sequence the entire APP gene. Thus, it's possible that the true disease-causing mutation is not within the β encoding portion. Their Discussion included the suggestion that the mutation, being valine to isoleucine (becoming more nonpolar), makes the APP transmembrane domain anchored more tightly to the membrane—which we now know is false, but it's

still interesting to see where their heads were at in these all-important early days. They did correctly postulate that the deposition of β-amyloid is associated with a worsened disease, and based on the phenotype of trisomy 21, it is somehow dose-dependent.

Follow-Up Work.

Brain banks have been an instrumental, and highly controlled way to investigate neurodegenerative diseases. Access to a brain bank could make or break researchers. After this work, many of the labs around the world began sequencing Aβ from their brain banks, and a host of mutations were found. There seemed to be a correlation between the location of mutations and the site that a protease cleaves APP. A number of secretases cleave APP, in different orders. We now know that when α-secretase precedes γ-secretase cleavage, two fragments of Aβ are formed—this is not pathogenic. However, when β-secretase precedes γ-secretase, leading to slightly different fragmentation, problems arise. γ-secretase cleavage is also not hyper-specific to a given site on APP, and thus some variable cleavages too lead to worsened disease. γ-secretase cleaves many other proteins in addition to just APP as well.

Finding the other AD Loci

Other linkage studies had identified more chromosomal locations associated with AD. To this point, there were three loci known: βAPP, ApoE (discussed later), and an unnanmed one called AD3. This AD3 loci was associated with an aggressive form of AD, making it ripe for identification. It was Sherrington et al. in 1995 who finally found it[195].

The loci was loosely mapped to Chr14. Using a group of familial pedigrees and previously published data, they examined a set of genetic markers on Chr14. By analyzing the recombination events, they were able to conclusively narrow the causative region to between two markers. However, this region was still massive and likely contained many hundreds of genes. To clone the region, they started by generating contigs. They examined the partial haplotypes of families with the same ethnic origin, hoping to further narrow down a region of interest. Two clusters of genes were shared between families of the same ethnicity. This indicated to them the heritability of the gene, and allowed them to make more direct comparisons as the families were under similar "genetic backgrounds" (that is, comparing between very heterogenous groups can make it difficult to discern causative mutations and genetic noise). They analyzed the cDNA clones of this region and determined 151 of them to reflect spliced mRNA. Using RT-PCR from AD affected brains and non-affected brains allowed them to narrow their field of interest to 7 transcripts. Among these was a novel gene, represented by fragment S182—encoding for a newly identified, transmembrane protein. They identified mutations in this new protein that were conserved between families, indicating these defects are causative.

Using a Northern blot, they identified 2 major transcripts and expression in most regions of the AD brain. Notably, mouse tissue only identified one of the two, specifically the 3kb region—which the authors suspect may be the normal mRNA, and the other the other transcript may represent a more rare, alternatively spliced version.

There is considerable sequence homology between the mouse and human transcripts. They used computational algorithms to predict 7 transmembrane domains and an accompanying possible structure. Two other non-polar regions were found that appear unstructured and unlikely to form transmem-

[195] `https://pubmed.ncbi.nlm.nih.gov/7596406/`

brane helices. They identified potential glycosylation sites, which they suspect may help order these regions into structures along the membrane. It was not presented in their data, but Sherrington and colleagues commented on sequence homology with *C. elegans* proteins—it was hoped that similarity to proteins mapped in lower order organisms may give insight into what this protein was. Unfortunately, no great candidates stood out among those identified. A protein called SPE-4 seemed to be the most similar, which plays a role in spermatogenesis specifically through a golgi derived organelle formation—thus the authors throw out the possibility of S182 altering βAPP or Tau golgi trafficking. Of course, we know in hindsight that this is untrue. A similarly offered hypothesis is that this S182 is the mammalian equivalent of Ca-α1D[196]!

Within the ORF of this S182 gene was mutations in multiple AD affected individuals. Among families with such mutations, no unaffected family members were ever found to have these mutations. In short, this strange, S182 gene encodes for a transmembrane protein whose function is presently unknown. This gene would eventually be called PSEN1.

PSEN and the Amyloid Cascade Hypothesis

Background.

Soon after PSEN1 was identified, another gene was found on Chr1, which happened to be quite similar to PSEN1. We now know this to be PSEN2. Then, a *C. elegans* paper came out in which a worm homolog was found, identified through investigation of Notch signaling. In hindsight, we know that γ-secretase and PSEN are one and the same (or rather, that PSEN1 is the proteolytic subunit of the γ-secretase complex).

Now, for the formal introduction: Prenesilin (PSEN) is a transmembrane protein known to contribute to AD progression, and thought to play a role in amyloid processing. In *Drosophila*, PSEN modulates Notch activity, which thereby can influence transcription. Notch signaling is a key developmental path in many organisms. Notch is especially important in cell-to-cell communication during development, aiding in cross talk between cells and driving cell differentiation. A ligand released by a cell will bind to Notch's extracellular domain, inducing a conformational change and allowing its intracellular domain to be cleaved. Cleavage is done by γ-secretase, which cleaves the Notch intracellular domain (NICD). The NICD is trafficked to the nucleus, where it interacts with gene inhibitors, turns them off, and allowed genes to be transcribed.

Model organisms, like *Drosophila*, were instrumental to our understanding of PSEN, and much of this understanding came from Strul & Greenwald, and Ye et al., published in the same issue of *Nature* in 1999[197,198]. They found that PSEN is needed for Notch cleavage, allowing it to exit the membrane and enter the nucleus where it can exert its control.

The Key Works.

Let us begin with Strul & Greenwald. By 1991 work, it was known that β-secretase is required for APP's conversion to β-amyloid and such cleavage occurs in the extracellular space. Similarly, it was

[196]This is funny to no one but myself, as my first published paper was on Ca-α1D function in *Drosophila*, a cool 28 years after this paper came out.

[197]`https://pubmed.ncbi.nlm.nih.gov/10206646/`

[198]`https://pubmed.ncbi.nlm.nih.gov/10206647/`

known that γ-secretase cleaves APP in the transmembrane domain. It was shown that Notch cleavage occurs in a similar way. PSEN mutations cause AD by increasing the amount of Aβ. PSEN knockdown inhibits the activity of γ-secretase. It was not known at this point if PSEN was an activator or a component of γ-secretase. Notably, they specifically comment on PSEN in relation to an Aβ42 (42 amino acids in length) variant, which is mentioned in other literature. Some thought that variable cleavage in the APP gene leads to slightly different length Aβ products, which are more or less prone to aggregation.

The pair began by analyzing the phenotypes of PSEN (PS in flies) knockdown. They generated two truncated alleles, which both showed identical phenotypes. PS^- embryos exhibited clustered neuroblasts, which should typically be spread out. Midline cells, thought to be controlled by Notch, did not differentiate as they do in WT flies. Homozygous *PS* loss was pupal lethal. Clones of PS^- cells causes abnormal wing development, thought to be through Notch's role in dorsal-ventral cell communication. Though, embryos of PS^- flies show normal accumulation and membrane localization of Notch, which indicates that it is not upstream of Notch expression—suggesting a role in Notch signaling.

They then generated a Gal4 knock-in on the intracellular side of the Notch protein. This tool was used to determine Notch localization. This was marked by a UAS (a promoter), which drives β-Gal expression (in the presence of the Gal4 protein) that they can stain for. I.e., when Notch is trafficked to the nucleus, β-Gal is expressed and can be stained for. They then used three 3 lines, which they call: N^+-GV3, N^{ECN}-GV3, and N^{intra}-GV3. N^+-GV3 is like their WT control, which functions normally. In the absence of either PSEN (PS^-) or Notch's ligand (Delta, Dl^-), no nuclear transport occurs, marked by no β-Gal staining. N^{ECN}-GV3 includes a large deletion of the extracellular domain, which essentially decouples Notch from it's ligand and makes it constitutely active. Thus, in this case they found Dl^- embryos to still have considerable trafficking to the nucleus, while PS^- did not. Finally, N^{intra}-GV3 is missing both the extracellular and transmembrane domains—therefore suggesting it will never be trafficked to the membrane at all, causing it to remain cytosolic. In this case, neither Dl^- nor PS^- prevented Notch trafficking to the nucleus, so all embryos could be stained by the β-Gal.

Onto Ye et al: to investigate the role of PSEN in Notch signaling, they developed 5 deficiency (*Df*) lines and observed similar dysfunction mentioned in the previous paper. *Df* lines are large chromosomal deletions, spanning many genes. Therefore, to verify the phenotypes of these lines were PSEN dependent, they used a variety of DNA rescues in order to see WT phenotypes. One of their rescues omitted PSEN DNA and did not rescue the phenotype of these flies. Therefore, they felt confident that these *Df* lines served as good models of PSEN mutation and or knockdown.

When analyzing their *Df* lines, they identified a pupal lethal phenotype they considered to be similar to an Su(H) mutant, an effector of the Notch signaling path. They used a variety of wing disc markers (β-Gal driven by expression of genes related to wing development), and saw β-Gal missing or heavily altered in their *Df* lines. Similarly, sensory organ precursor (SOP) cells were found to cluster together with limited lateral inhibition, determined by enlarged territories and increased *achaete* and *scabrous* expression (early SOP markers).

The *Df* lines showed obvious neurodvelopmental issues, including unclear and or non-existant formation of a ventral nerve cord (VNC), scattered neural localization, and over-proliferation. This experiment was done via ELAV staining, a pan-neuronal marker. Interestingly, they saw that paternal PSN partially rescues neurodevelopmental troubles—VNC formation was visible, but clearly had

uncouth morphology.

Notch protein localization and expression was similar across their PSEN mutants and wildtype larva. This was consistent between both the intracellular and extracellular components of Notch, and for the Notch ligand, Delta. In retrospect, the finding that intracellular PSEN is unchanged may be considered surprising, since PSEN is now known to cleave Notch. The authors attribute this to the small intracellular quantity that is actually trafficked to the nucleus, which therefore may remain undetectable using immunostaining. To investigate how PSEN processes notch, the authors used Western blots. The largest, ≈120K band of the Notch protein migrates more slowly in the PSEN mutants, suggesting the piece is slightly larger. Among the 120K fragments, the largest dominates in the PSEN knockout compared to WT. This phenotype is specific to PSEN knockdown, as in the Su(H) mutant background, this phenotype is not observed (i.e., Notch is processed as WT is), and similarly in mutant Notch background the large 120K band is absent. The same fragments are unaltered in other proteolytic knockdown (specific to Notch's processing in the golgi), demonstrating a specificity for PSEN and that the cleavage observed occurs after its membrane targeting.

Expression of Notch's intracellular domain alone, and accompanying blots, showed that it is unaffected by the PSEN knockdown, suggesting PSEN plays a role specifically in releasing Notch from the membrane. Finally, PSEN knockdown did not affect Delta levels, as quantified by Western blots. Inactive Notch confers proneural clustering in both WT and PSEN mutants. Using only the intracellular Notch domain, no SOP differentiation occurs in both WT and PSEN mutants. Most SOP differentiation was suppressed in a membrane tethered background. It is supposed by the authors that PSEN is not required for membrane trafficking nor nuclear trafficking of Notch. Rather, they supposed PSEN is needed in some activation mechanism that follows ligand binding.

The experiments by both groups led to the conclusion that PSEN was specifically involved in cleavage of Notch, as opposed to acting as a ligand or aiding in it's downstream signaling cascade. It was concluded that PSEN cleaves Notch somewhere in its transmembrane domain. Naturally, the big takeaway was that PSEN may play a similar role in APP processing, and thus produces Aβ in the same way.

Making an AD Mouse Model

First, let me state that mice do not naturally get Alzheimer's disease. This is a key point that I think often remains understated. Flies, similarly, do not get AD. In fact, it is difficult to get protein aggregation in both of these models. Many fly ALS models (discussed later) do not show aggregation at all, while the same proteins in humans do. Therefore, one must always wonder if the models use are using non-physiological conditions to achieve them (e.g., absurdly high doses of some protein, which may, on their own, have odd implications).

To this point, we have a decent idea of some of "key players." But, we have little room to investigate them. Despite some attempts, no successful AD model had been established 1995. Goate et al. were the first to generate a meaningful one[199]. The group used a modified, transgenic APP gene in mice which reproduced many of the AD phenotypes. Such a model drove increased expression in many mouse tissues, but especially within the brain. They consider the promoter they used, PDGF-β, and the familial AD mutation they used (V717F), to be the two prime contributors to success.

[199] `https://www.nature.com/articles/373523a0`

Their Northern blots showed 3 different alternatively spliced forms of APP mRNA, and Western blots showed the amount of protein is was increased in the transgenic mice vs. WT mice, and that that Aβ was indeed being produced too. Recall that these mice did have endogenous APP of their own, which was not mutated. Therefore, the disease focused on the gain of function associated with APP aggregation.

Aβ buildup occurred only at 6-9 months of age and worsened from there. It was primarily local to the hippocampus, corpus callosum, and cerebral cortex. Aβ buildup took a variety of forms, including irregularly shaped bundles and other more characteristic, dense plaques. Reactive astrocytes were often found surrounding such plaques via GFAP staining. Aβ plaques were associated with synaptophysin, suggesting their localization to sprouting axons[200]. Staining for MAP-2 in addition to synaptophysin showed plaques disrupting and de-densifying nearby regions. Plaques also seemed to compress the surrounding neurophil (the unmyelinated, cell dense regions of neural tissue).

Tangled up in Blue.

The appropriate question to ask is: "what about tau?" Tau was not found in this model. Tau is notoriously absent from rodent tissues. At the point this model was created, there was the belief that the ratio of Aβ42 to 40 indicates the progression of the disease, which existed under the umbrella of the amyloid cascade hypothesis. The vast majority of the work done so far focused on the plaques, but with no insight whatsoever into the tangles.

In the 60s, tangles were purified to the point of visual analysis, indicating they were helical in nature and likely two filaments coiled around one another. This gave them the name "paired helical filaments" (PHF). In 1992, this PHF was tested against tau antibodies and stained with Coomassie Brilliant Blue[201], showing different forms of tau at variable molecular weights. Normal, wildtype tau was largely unaffected by dephosphorylation by a generic phosphatase. But dephosphorylating PHF substantially lowered the observed molecular weight. In short, this suggested PHF is a heavily phosphorylated version of tau. This principle is generally conserved in diseases, where a protein becomes abnormally phosphorylated and aggregates. Notably, tau mutations are never associated with familial AD—but are associated with other degenerative diseases.

The endogenous function is not fully solved, but we know the gist. Tau binds to microubules and is thought to help stabilize them. When phosphorylated, it unbinds from the microtubules, thought to cause axonal degeneration. Notably, tangle burden progresses with the disease more so than does plaque burden. Tangles appear later on, but correlated more strongly with increased symptoms. This bit of conflict created a rift among scientists on what the root of the disease was—Aβ or tau. Some supposed tau was a side effect of the greater disease, induced by Aβ aggregation. Others supposed that Aβ aggregation was largely benin, and it was when tangles arose that the disease truly began.

What About Tau?

Initially it was thought that tau tangles were solely a symptom, and or, a result of already dying neurons. In 2000, Lewis et al. helped challenge this idea with a paper that overexpressed mutant

[200]It is a bit unclear what is meant by sprouting here—I am not sure how definitions of sprouting may differ between groups.

[201]The section name is a play on the Bob Dylan song title.

tau protein found in frontotemporal dementia and parkinsonism (FTDP)[202]. The mice exhibited a wide variety of degeneration linked to the tau protein buildup. This included symptoms even in the spinal cord and peripheral nerves. Mutant tau (P301L) was expressed by the mouse prion promoter (MoPrP). They found that hemizygous lines expressed tau at approximately the same level as WT, and mice homozygous for the mutant expressed it at levels $\approx 2\times$WT.

A composite of behavioral assays, including escape responses, motor defects, and the inability to right (flipping over from its back onto its feet again) demonstrated sensorimotor dysfunction. Hemizygous mice had symptom onset at about 6.5 months, while homozygous at around 4.5 months. Eventually, they were unable to walk.

The spinal cord showed defects, namely fibrillary gliosis, axonal degeneration, and axonal spheroids (bleb). GFAP immunoreaction was found throughout the brain, signifying gliosis. Importantly, degeneration was global, including skeletal muscle and peripheral nerves. Neurofibrillary tangles were found throughout the brain, including in the brainstem and diencephalon. Interestingly, so-called "pretangles" were found in other regions of the brain, including the hippocampus. The fibrils were irregular in shape and specifically phosphorylated.

In the mutant lines, a greater proportion of tau was insoluble and shifted toward a higher molecular weight. This is consistent with being heavily phosphorylated, which was verified by dephosphorylating the protein causing the elimination of such high-weight bands. The authors showed the similarity between AD bands and their model's bands.

Implications.

To follow-up, whether or not APP modulates the tau protein buildup became a focus. Indeed, crossing this tau mouse with an APP mouse led to increased NFT formation. Later, groups tried combining multiple familial AD mutations in APP with PSEN mutants, leading to very early onset and more powerful, clear behavioral defects. After these tau models were verified, tau too was added. It isn't clear what the connection between fAD and sporadic AD is. Is inducing these severe AD symptoms in mice via fAD mutations actually comparable to the very slow progressing, late onset forms of AD? To what degree are the fAD and sporadic AD mechanisms conserved?

In clinical trials, antibodies against Aβ showed the ability to decrease plaques. However, no obvious clinical benefit was found. Naturally, this could be because degeneration had already run its course, or it could be that tau is directing progression at this point. Some take this as evidence against the amyloid cascade hypothesis.

Mutations Protecting Against AD

In 2012, Jonsson et al. found something totally unexpected[203]. By 2012, gone were the days of hand reading base pairs from an autoradiogram. The intense study of amyloid precursor protein's (APP's) role in Alzheimer's disease (AD) largely began when Goate *et al.* hand sequenced the β-amyloid coding region of individuals with familial AD and found a causative mutation. Some 20 years later, the β-amyloid allele of nearly every individual in Iceland became known. Iceland keeps quite good

[202] `https://www.nature.com/articles/ng0800_402`
[203] `https://pubmed.ncbi.nlm.nih.gov/22801501/`

genealogy records, and because people are quite related due to population isolation, their genetics are ripe for this sort of testing.

Via direct sequencing of around 2,000 Icelanders, Jonsson et al. generated chips for APP alleles, allowing for the genotyping of 70,000 more. The accuracy of their genotyping was verified through direct sequencing, allowing computational approaches to provide predictions for nearly all Icelander's APP allele. Fortunately, an otherwise very rare allele is present at a high frequency in some North European countries, like Iceland. This allele was A673T, the first mutation found to be protective against AD, and its discovery indirectly answered a number of open questions.

The group demonstrated this A673T mutation is protective through comparisons of the allele frequency in AD and non-AD patients. Notably, this allele was 7.52 times more likely to be found in cognitively intact individuals above 85 than in AD patients, suggesting it aids in the prevention of late onset AD. Like many familial AD mutations, A673T is present near the β-secretase 1 (BACE1) cleavage site. Because the A673T mutation is protective against late onset AD, it indeed showed APP's sequence is relevant to sporadic AD, and that there is some conserved mechanism between early onset and late onset AD. Through Western blot analysis of 293T cells (kidney cells), they sought to gain mechanistic insight. Interestingly, the work showed evidence against the importance of Aβ cleavage product ratios (namely Aβ_{42} to Aβ_{40}). The A673T mutation showed an overall decrease in Aβ, rather than affecting this Aβ_{42} to Aβ_{40} ratio. The amount of sAPPα was slightly raised, as might be expected if un-cleaved APP is present in higher levels due to weakened BACE1 activity, making more available for cleavage by α-secretase. These results suggest AD can be slowed or stopped via reduction of Aβ, particularly through the weakening of BACE1 proteolytic activity. They make the bold claim that the A673T mutation is protective against general cognitive decline, suggesting a function of APP beyond AD. This measure was done by the Cognitive Performance Scale (CPS), but is purely correlative in nature.

The mutations compared by Jonsson and colleagues were A673T and A673V, with A673V showing a significant rise in Aβ. Notably, A673V converts from the mildly hydrophobic residue, alanine, to the markedly more hydrophobic residue valine. Conversely, A673T becomes more hydrophilic. One can postulate that the cleavage of APP, releasing Aβ, requires a hydrophobic interface between BACE1 and APP around this residue. Further credence is given to this by a different familial AD mutation, called the Swedish mutation (K670N/M671L), which adjacent to the BACE1 cut site becomes more hydrophobic with leucine.

Mudher and Lovestone stated that proving the amyloid cascade hypothesis requires showing a reduction in plaques results in tau tangle abolishment, and thus functional recovery[204]. A considerable drawback of Jonsson and colleagues' work is that there is no direct insight into the interplay of APP and tau. The study of AD still has long to go before reaching its completion. But, as new tools become available to researchers, progress is continually made. Faster sequencing, computational drug design, and ever-improving mouse models allow for new routes of study. The journey to treating AD is long, but we can be sure that whenever a new roadblock arises, technological advancement will provide the path to new insight.

[204] `https://pubmed.ncbi.nlm.nih.gov/11801334/`

Digression: Other Means of Protection.

Interestingly, education proves to be a strong protective factor against AD. If memory problems have already began, using acetylcholine-esterase inhibitors can be helpful to help potentiate this transmitter and improve function (or, delay its decline). However, these forms of treatment tend to be short-lived. Exercise also turns out to be quite helpful in promoting neurogenesis, which may delay progression.

There's Still More?: ApoE

As mentioned in an earlier section, the effect of ApoE on AD had long been known. Surprisingly, ApoE protein was found within the Aβ plaques, and only after this was a linkage found to Chr19, around the APOE gene. ApoE is a small protein, only 229 amino acids. It is a lipoprotein, which helps move cholesterol through the bloodstream—therefore, one possible mode of management is controlling cholesterol levels. Four alleles exist, but the three key APOE alleles we consider here are:

1. APOE-ϵ2 (cys112, cys158)
2. APOE-ϵ3 (cys112, arg158)
3. APOE-ϵ4 (arg112, arg158)

The dosage of the APOE allele which confers risk of disease, as shown by Corder et al. in 1993[205]. The APOE-ϵ3 allele is the most common. It is thought, now, that ApoE2 may be slightly protective over ApoE3. But, the key player is APOE-ϵ, whose presence results in a much earlier age of onset. The APOE-ϵ4 allele is one of the strongest determinants of AD onset. The risk factor for AD increased by approximately 3$\times$ for each additional APOE-ϵ4 allele. Corder and colleagues showed that $\approx 90\%$ of those with 4/4 genotype had the disease, while only 19% did for a 2/3 genotype. Similarly, the age of onset was much earlier for those with 4/4 than other genotypes. Autopsies of these patients show a much higher degree of amyloid staining, suggesting the APOE-ϵ4 allele increases the plaque burden.

From then to now, there has been considerable skepticism over this finding. As APOE is not considered disease causing, but rather a risk factor, it was unclear the role it might play. Still, some were unconcerned with the fact that ApoE shows up in plaques, as it is not clear that amyloid is disease causing. Importantly, too, most of these genetic studies were done in European populations—which turned out to be of great relevance. APOE-ϵ4 is not a (big) risk factor in Hispanic populations like it is in a European background, further muddling the potential role that it may play.

Degeneration & Genome Wide Association Studies (GWAS)

Sporadic AD remains much more common than familial—so how do you go after risk factors involved in sporadic AD? Of course, APOE-ϵ4 is the highest risk factor by a significant amount, but are there more[206]?

One way to investigate this is through a genome wide association study (GWAS). A Manhattan plot is one way to represent this data, and probes the entire genome for SNPs of individuals and co-segregates significantly with diseases vs. controls. Initially, massive amounts of money were poured into GWASs,

[205] `https://pubmed.ncbi.nlm.nih.gov/8346443/`

[206] It is unclear to me if asking this question is a good idea. Perhaps we should focus on understanding the factors we already have identified first. Or perhaps we will find the missing link that connects all the puzzle pieces.

but were largely unsuccessful. Now, with much greater sample sizes, they have been more successful. Notably, 90% of polymorphisms are shared ancestrally. Around 60 de novo point mutations are expected per individuals, with 3 indels, and (on average) less than 1 structural variation. Therefore, much of the original GWAS studies were muddled by the different genetic backgrounds of the patients. That is, someone with the disease from one continent that is compared with an individual of another continent would show a significant, surprising degree of genetic association for factors totally unrelated to the disease. Combining this with the fact that a large portion of the genome is non-coding—GWAS struggled greatly with poor controls in the early days.

Now, suppose you find a peak—what next? You have to understand what is occurring. If the mutation is in an enhancer, what gene is it affecting? When tons of genes cluster closely together, and linkage is found, which one is the affected gene? When many nearby SNPs are found, which ones are relevant? In short, GWAS seem like an all-powerful technique, but often the results are difficult to interpret or meaningless. Later, in section **33.2**, we will discuss a hugely successful example of this mode of study.

Digression: AD and Sleep.

This is from a video featuring Holtzman called Sleep & the Pahophysiology of Alzheimer's Disease[207]. To start, dementia is considered a decline in cognitive abilities sufficient to impair daily life. Approximately 50% of people over the age of 80 have some form of dementia.

The role of glia has been understated for a long time in the progression of AD. It seems that Tau and β-amyloid buildup activates a glial / autoimmune response. Like most neurodegenerative diseases, symptom onset marks the end stages of it. Pathology beginnings occur far before symptom onset.

APP is enriched at synapses in the brain. Interestingly, endocytosis occurring at the membrane leads to membrane bound APP being taken back up by vesicles into the synapse. It is here that, while in endosomes, APP is cleaved, leaving behind its β unit. These are then exocytosed. This can occur both pre- and postsynaptically.

The Aβ peptide quantity is somehow higher in the dark and lower in the light. The Aβ presence is also closely correlated with the amount of lactate, which sometimes serves as a proxy for synaptic activity. Aβ can be measured in the CSF. Chronic sleep deprivation leads to significantly increased Aβ plaque buildup, and the converse was true in sleep induced mice. This idea was repeated in humans and was conserved. These researchers also determined that it was due, in fact, to production and release as opposed to clearance (done using labeled Aβ). This is quite interesting because ApoE is thought to help clear Aβ. Also fascinatingly, Aβ seems to further worsen sleep.

While Aβ accumulation precedes symptoms by many years, tau buildup seems to correlate strongly with symptom onset. Acute neuronal activity increase and chronic increase both lead to buildup of tau. As wakefulness increases neuronal activity, so too does it increase tau.

[207] `https://www.youtube.com/watch?v=zCmngDk9VDU`

33.2 Huntington's

Overview

Huntington's disease (HD) is characterized by an autosomal dominant mutation, and symptom onset typically begins between age 30 and 50, and it is completely penetrant. Chorea (large muscle jerks, resulting in dance-like movement) and cognitive decline are two marked symptoms. HD also features genetic anticipation—the diseases get worse, with earlier onset, every generation. While Huntingon's was first recognized in the 1900's, it was largely ignored—researchers didn't know what to make of it, because the "gene" changing over time seemed confusing and inexplicable.

Repeat expansion—an increasingly long pattern of repeated nucleotides induced by instability in replication—was first identified in myotonic dystrophy 1 (DM1), a CTG expansion. This CTG expansion is in the untranslated region of the dystrophia myotonica protein kinase (DMPK) gene. It is marked by genetic instability, is autosomal dominant, and inherited. Presently, there are 30-40 known human diseases due to CAG repeats, HD being one of them. In HD, one cannot precisely predict the age of onset by the length of the CAG repeats, but the length of the repeat does correlate with the severity of the disease—just with considerably variability. The expansion is within the coding region of an exon, which encodes glutamine, becoming a polyGlu (polyQ). Many diseases share this mechanism with HD. These genes normally have endogenous CAG repeats, but in disease these repeats become elongated tens or hundreds of times more than in the wildtype gene.

HD is typically characterized as a movement disorder, and the most severe cases, marked by very long expansions, can cause juvenile HD. In patients, the caudate and putamen (which make up the striatum) are especially damaged, leading to enlarged central ventricles.

Interestingly, in diseases like spinocerebellar ataxia (SCA), another exon CAG repeat disease, many present clinically in the same way. However, their mutations are in different genes and the cell types affected are different. This causes degeneration in slightly different regions of the brain. Ataxic gait is classified by a lack of balance, which causes patients to widen their legs to maintain balance and jerk-like movement with stepping. It is especially clear upon asking patients to close their stance, in which case you'll see them lose balance. HD has less extreme, jerky movements and more stiffness, dis-coordination, and difficulty balancing[208].

Nancy Wexler was one of the first to "set out" to characterize HD. She traveled around the world and met many HD patients, looking for someone who was homozygous. She specifically went to Venezuela because they had a very large HD population. The median repeat length in this population was around 45 repeats. Most fall in the range of 40-50, but some rare outliers existed. Juvenile onset median was closer to 60 repeats.

Repeat expansion occurs more commonly in sperm production, so it is more common for an affected father to worsen the disease. In fact, a later discussed work (Mangiarini et al.) made the comment that 70% of juvenile cases are inherited from fathers. Interestingly, as mentioned, HD most severely affects the striatum, impacting movement. However, conversely to Parkinson's, dopamine levels actually increase, while GABA and acetylcholine decrease. Therefore, dopamine inhibitors are a way to treat the symptoms (not the disease progression) of HD.

[208] A rude comparison is to someone who is tweaking.

Tracking the Gene

Prior to this 1983, nothing was known about the disease pathology of Huntingon's disease. In fact, the closest that had been achieved before is that 20% of the chromosome was excluded from the disease-causing locus. The disease causing gene was solved by Gusella et al.[209]

The group began by generating stable cell lines from patient-derived lymphoblastoid cells. Picking high quality patients was essential, which included using experienced neurologists and well verified HD over the course of a couple of years. Finding large families with well documented Huntingon's disease and available tissue samples was quite hard. Similarly, recall that in 1983, very, very few genes were known. The library of DNA probes could nearly be counted on two hands. The group used They used 12 established DNA probes from the known genes, which hybridized to a single place in the genome. Shockingly, one gave a hit: the G8 marker showed linkage to the HD disease causing gene.

The prime roadblock in investigating HD was the lack of genetic linkage to any known markers. The G8 marker showed the first promising results. It was extremely difficult to extract meaning from blots, in these days.

The group characterized nearby restriction cut sites for *Hind*III (H) and *EcoR1* (R). The G8 probe is quite small, and the polymorphic sites are very close to one another. An imporant part of this was differentiating between affected and non-affected chromosomes in the heteozygous patients. Similarly, it gave indications of the patient's genotypes. The two pedigrees they analyzed had different haplotypes. Haplotype A was associated with HD in the American family, and haplotype C with the Venezuelan family. The haplotypes being different either suggests it's two different genes or that the mutations arose independently in the two families. There are instances in which individuals have haplotype C, but do not have HD. This could be due to different age of onset or that some recombination event separated this polymorphism from the HD gene. Recall that we still have absolutely no information about where G8, or the HD gene is, in the genome. They still may be many kb apart.

They showed that the G8 probe does not hybridize without Chr4 present. In fusing human-mouse cell hybrids, they essentially generated mouse cells with human chromosomes. Each cell may contain any combination of human chromosomes. They scored which chromosomes each cell had, by eye, and if it hybridized with the G8 marker. Because all cells with Chr4 have the G8 marker, they assumed G8 must be on Chr4. They tried testing recombination events of the HD gene with other Chr4 markers and found no association.

The G8 probe certainly doesn't hybridize directly to the HD gene, but it is likely close to it! The scientists had narrowed the disease causing gene down to Chr4 and had some indication of surrounding restriction enzyme cut sites, and the haplotyopes of affected and unaffected individuals. This had considerable implications, as you can now test haplotypes of offspring and predict if they'd get HD. Moving forward, more markers and more segregation analysis will be helpful in narrowing down the gene. Then, you have to start sequencing nearby genes.

[209] `https://pubmed.ncbi.nlm.nih.gov/6316146/`

Finding *HTT*

Led by Gusella, the Huntington's Disease Collaborative Research Group (HDCRG) was formed. For many years, they had to make markers and compare recombination of events around this G8 area, until they finally found the gene in 1993[210]. The goal was to narrow down the region more and more before beginning the dreaded hand-sequencing. Through continual investigation of recombination events in HD families, the group was able to rule out different regions of Chr4. They were able to narrow it down to between two markers in a 2.2Mb region. They showed corresponding restriction sites, which were helpful in mapping the haplotype of different HD kindreds, allowing them to narrow down their range of interest further to 500Kb. They look specifically for genes that were expressed by screening with a cDNA library.

They generated exon clones, and therefore cDNAs, IT15A and IT16A. Northern blotting showed hybridization with a similarly sized sequence, suggesting the sequences are derived from the same mRNA. By Northern blotting, they verified the existence of an 11kb RNA strand characteristic of the IT15 gene in both homozygous HD patients and normal samples. This gene, as it turns out, was quite big. They didn't have a single cDNA which covered it, so they pieced together multiple cDNAs. Using the cDNA they had (IT16A and IT15A) they did a walk, where they used overlapping cDNAs to get the entire IT15 sequence. It was, in total, predicted to be 3144 amino acids!

This gene contained the mysterious CAG repeats. After sequencing, they generated primers for the regions flanking the CAG repeats. Of the 173 chromosomes they sequenced, the vast majority had below 25 repeats, and chromosomes had 80% heterozygosity. They found that more than some number of repeats (42 in this case) was perfectly correlated with HD. They then tested as many HD families as they could. PCR and southern showed very variable repeat lengths.

They showed PCR products of different family members in the kindred, hoping to show that one allele was unaffected and one allele was affected in the HD group. Notably, though, some do not show a HD allele, which they write off as being due to the repeat lengths being too long to amplify. They verified this by southern blot, which showed that the repeat length was 100—and this patient had HD onset at age 2. Some people show more than 2 HD alleles. The hypothesis for why this occurs is that the cell lines (lymphocytes) are likely a bit more mosaic, since they actively divide much more—hence, there's a greater chance for oddities in replication, leading to further repeat expansion.

They were interested in testing sporadic cases of HD to see if this repeat expansion was still the driving cause. Indeed, they found parents with higher-end repeat expansion, and thus the offspring had repeats reaching the stereotypical 44+ range. They identify a similar haplotype in these chromosomes as in 1/3rd of all HD chromosomes, which suggests that this haplotype may predispose individuals for HD.

Takeaways.

The key findings were that there is this correlation between repeats and HD. This could be used as a diagnostic tool, or describe those who are pre-disposed to getting HD. At present, they do not know the mechanism by which HD acts, but they suspect it may be gain-of-function. They cannot say anything with certainty about this gene's corresponding protein, nor can they say that the repeats are translated at all. If it is translated, the polyQ would have considerable effects, and if it were untranslated, the effects are likely at the mRNA level. The next step that they propose is, indeed, to

[210] `https://pubmed.ncbi.nlm.nih.gov/8458085/`

understand the endogenous function of this protein which they call huntingtin (with gene *HTT*).

An Accidental Model

Prior to this point, questions regarding the nature of HD existed. It was unclear how the CAG repeat was affecting patients, but by then it was known that the glutamine residues corresponding to CAG were indeed translated. Generating a mouse model was difficult, as the HTT gene is incredibly long, so cloning was almost unachievable then. In 1996, it was Mangiarini et al. who improbably created the first successful mouse model—one which is still used today[211].

The story goes, the group was most interested in studying the instability in the HTT gene. Therefore, they generated a house line with only the beginning portion of *HTT* (the first exon), including $(CAG)_{115}$—$(CAG)_{150}$. Their transgene, using a fragment from the human genome, also included the endogenous HTT gene promoter[212].

After inserting the predicted region of human genomic DNA into mice, they probe the F1 progeny in a Southern blot using a piece of DNA that hybridizes to the C termus of the transgene. The purpose of this was to better understand where their insertion integrated into the genome. They make complicated predictions about the transgenes, regarding if there were multiple insertions, deletions, etc. This was primarily speculative in nature, but helped them get a sense of which lines were most successful. They determine the number of CAG repeats in each of their transgenic lines via PCR amplification and an ABI sequencer—in other words, amplified and then sequenced. The ABI sequencer plots show peaks corresponding to the length of the repeats—which were all in the range of about 110-150. The line that they end up choosing as their best is called the "R6" mouse—this is worth vaguely recalling, as future papers will use this very mouse.

Age of onset was usually around 9-11 weeks, and death around 10-13 weeks, meaning death followed symptoms quickly. The phenotype is described as complex, and that at birth no differences are observed (this is a common trait in good models—as with the human disease, no differentiating symptoms exist early in life). Symptoms included motor jerks and resting tremors, which worsened under stress. Significant weight loss is noticed by death, which is strangely confounded by the fact that their appetites are not reduced. Indeed, upon death food is found in their stomachs and stool seemed to be normally produced, but their bodies have atrophied. Strangely, too, they seem to urinate excessively.

They then analyze the location of transgene expression. In some lines, transgene expression was found in all tissues analyzed, suggesting to them that it was expressed constitutively under the promoter within the injected DNA fragment. Using an antibody called 1C2, they performed Westerns to analyze protein translation patterns—which appeared to be global. In one case, they identified a lack of translation, which they attribute to the protein not expressing the necessary epitope—which seems odd since it is meant to bind to polyQ. This generally brings up another point one may have noticed by now: many neurodegenerative disorders are due to proteins expressed globally—yet symptoms seem to be neurological in nature, and often restricted to just a subset of neurons. What confers neuronal vulnerability, then?

[211] `https://pubmed.ncbi.nlm.nih.gov/8898202/`

[212] A small comment that I have is: How do you maintain these mice? Won't the HTT gene always get purified, lest the lines will die due to their instability? I believe the answer is: mice experience a much lower degree of instability, therefore it is less of a concern. Still though, one must continually back-cross mice with others containing the transgene. Otherwise, it may be lost.

Finally, they characterized the neuropathologies of the mice. They identified universal shrinkage of the brain, not necessarily degeneration nor shrinkage localized to the striatum, on the order of about 20% smaller than normal brains—however, there was clear ventricular expansion. They observed no neural or glial loss, nor did they see reactive gliosis. In fact, even the basal ganglia and striatum had normal neural morphology, aside from shrinkge. This shrinkage was observed early in the disease progression, and did not worsen with age. This, obviously, contrasts to human HD. One hypothesis they suggest is that disease progression is so rapid in mice that there is not time for atrophy to occur. This, while perhaps fishy on face, is a common issue with mouse models. Many mouse models lack essential attributes of disease because progression is so rapid—in this case, what would occur over the course of a decade in humans occurs in a few weeks in mice. How this will alter the pathology is unclear.

Without the hindsight we have now, one might think Huntington's is a metabolic disorder.

Future Directions.

Afterwards, they further investigated the neuropathology associated with the phenotype. By EM, they found large clumps present in the nucleus, where HTT is not expected to be. They found these clumps stained for both HTT and ubiquitin, meaning it was tagged for degradation. It seemed that degradation was failing. Almost all neurological diseases are characterized by ubiquitinated proteins, including AD, and thus it is a strong marker for the buildup of aberrant proteins. Notably, they did not see staining for congo red, which suggested it was not amyloid.

They also found that brain weight decreased weeks before total body weight decreased—but up to this point, everything is normal. Further correlative data suggested neuropathology was induced by HTT accumulation in the nucleus, which was causal for brain weight decrease and therefore causal for whole body phenotypes. This suspected mechanism linked HD to AD.

In other works from different groups, protein context was found to be critical. It was shown that only when polyQ proteins are trafficked into the nucleus does a phenotype manifest. If a polyQ protein stays in the cytosol—it will not be pathogenic.

Huntingtin Aggregates

Scherzinger et al. 1997 focused on *in vitro* and *in vivo* model development and subsequent investigation of supposed amyloid plaque-like structures that form as a result of polyQ translation[213]. Prior works had shown the Huntingtin protein is associated with the cytosol, possibly around vesicles and or microtubules. Notably though, the HTT protein is expressed widely, which presents a strange problem when considering that degeneration is relatively specific to the striatum. Some hypothesized a gain of function as a result of the increased expansion, and some predicted that it was due to conformational changes in the HTT protein. Notably, polyQ are capable of forming β-pleated sheets. I'll let you connect the dots with that one.

Using the model presented in the previous section as their starting point, this work found morphological changes on a neuronal level. Small granular structures can be seen, and indentations in the nuclear membrane are found along with increased density of nuclear pores.

[213] `https://pubmed.ncbi.nlm.nih.gov/9267034/`

To build their model they used expression of exon 1 of the HD gene with expanded repeats fused to glutathione S-transferase (GST), which could be cleaved to produce free polyQ repeats. They included varying lengths of repeats, called GST-HDx (where x is the number of repeats). Such fusion proteins were made into plasmids, expressed in *E. coli*, and purified. Proteins were purified and stained using a Coomassie blue dye and an anti-Huntingtin antibody, both of which verified expected expression and variable sized bands predicted by their sequence and or number of repeats.

Via digesting the proteins with trypsin to cleave off the GST domain, lower molecular weight products are found in GST-HD20,30. However, a group of high molecular weight particles were present after GST-HD51 cleavage, suggesting there is some aggregation occurring specifically at this higher number of repeats. This also suggested that GST increased solubility of the protein when un-cleaved, as this higher weight form was not as present when not exposed to the protease. This difference was stronger when complete digestion of the GST tag occurred (i.e., due to longer time exposed to trypsin).

They developed a method of quantifying aggregation by collection of insoluble components on an acetate filter, which existed in their longer repeat line (GST-HD51). They find GST-HDx present in both the cleaved and uncleaved state on the nitrocellulose membrane. This contrasts to the acetate membrane, in which they only find protein in the cleaved GST-HD51 group. Because their other experiments showed GST-HD51, and not the other groups, formed aggregates, this suggested to them that only aggregated protein would be retained on the acetate membrane. Therefore, this was a useful tool for screening aggregation.

Via electron microscopy, they found uniform, oligomeric particles without aggregates in undigested GST-HD51. Upon digestion, many non-uniform, ribbon-like aggregates were found. Clot-like structures were found on the end of fibrils. It appeared the clots were due to incomplete digestion. This contrasts with GST-HD20,30, which did not form aggregates with or without cleavage.

Notably, this protein stained with Congo Red. The staining was similar to that which would be seen in prions or in amyloid. To determine if these products are found *in vivo*, they used transgenic mouse lines mentioned in Mangiarini et al. 1996. They indeed found, via SDS page gel and Western Blots using the anti-HD antibody, that there is some fraction of peptide present at a high molecular weight. This product was found in the nuclear fraction. They also used an anti-ubiquitin antibody, and found that it reacted with the aggregates, suggesting it is ubiquitinated.

Notably, the less aggressive transgenic lines they used showed characteristic fibrils, while the most aggressive transgenic lines did not. They think this may be because fibrils do not have time to form in such aggressive cases. One can postulate what this means for disease progression. This is a nice time to further ponder the relevance of protein aggregation in disease. If aggressive forms of the disease do not show aggregation, then is this merely a symptom? Why is so much study dedicated to aggregation, when evidence does not strongly suggest it causes disease?

HD is Reversible?

Yamamoto et al. 2000 is one of the flagship works in the study of HD[214]. The premise of this paper is that they've used a conditional expression system to drive mutant HTT. Upon blocking its expression,

[214] `https://pubmed.ncbi.nlm.nih.gov/10778856/`

Huntington's disease symptoms disappeared—leading to the hypothesis that continuous production of HTT is required for disease phenotypes.

The technology they use is called a tet-regulated system. A promoter drives the expression of tTA, which binds to the TetO elements, allowing for expression of the transgene that follows, which, in this case, was a HD$(CAG)_{94}$. A molecule called tetracycline (and the analog doxycycline) is able to bind to tTA, which blocks its binding to the TetO element, shutting off transcription. TetO is actually bidirectional, so the opposite side contained a β-galactosidase (lacZ) reporter, used to validate that expression is or is not occurring—a really elegant strategy. Initially, they observed a strong lethality phenotype which they attributed to strong expression of HTT in the developing fetus. To ameliorate this, they gave pregnant mothers doxycycline, which did the trick and stopped fetal lethality.

Strong expression was found in the striatum, septum, cortex, and hippocampus. They validated this both with β-gal staining, and with antibodies against HTT. They found diffuse nuclear staining as well, including aggregates of varying sizes. Though, they did find a significant amount of extra-nuclear aggregates as well, including those that were highly ubiquitinated. They saw expected phenotypes, including finding a significant size reduction in mice with HTT expression, and a narrowing of the striatum. Reactive astrcytosis was observed via GFAP staining, particularly in the striatum. Interestingly, they also observed D1 receptor loss[215]. Clasping behavior was observed here as well. Interestingly, the average lifespan of mice was still comparable to WT (approx. 2 years)[216].

Then, things get interesting. In comparing mice without dox treatment, and those that began treatment at 18 weeks, no HTT was found in the brains of mice experiencing treatment. However, some buildup was seen in other areas. Similarly, no brain weight reduction was seen after the dox was administered. After dox administration, the caudate and putamen no longer shrunk. GFAP staining did not increase, and in fact seemed to reduce. Finally, D1 receptor loss did not further decrease. Their clasping score actually improved quite significantly. Therefore, blocking aggregate progression indicates an overall blocking of disease progression, and or an improvement in symptoms.

Interestingly, they commented in their discussion that, while there is a significant volume loss in the R6 transgenic lines, there is not a clear decrease in the cell number. It is thought, then, that other factors may be decreasing, like the volume of the cells or their ECM.

SCA1 and Degeneration Treatment

Let us first begin by stepping back. Huntingon's disease is one of many diseases characterized by long and variable genomic repeats conferring neurological dysfunction. As mentioned above, HD was shown by Yamamoto et al. to be reversible through a carefully orchestrated mouse model that allowed for temporal control of the disease causing gene's transcription. To this point, degeneration, and or damage, to the implastic CNS had long been considered permanent. Thus, this concept was promising, as it suggested there was hope of rectification for patients already showing symptoms, but did not have a straightfoward mode of clinical application.

[215] It's unclear to me how what this could be attributed to. One might speculate it is related to cell loss, but that does not seem to be the answer, as commented on later

[216] This is a superb example of what it means to be a scientist, I suppose: 2 years for a single metric, which ended up having no difference from WT.

Spinocerebellar ataxia type 1 (SCA1), another disease characterized by variable repeats, manifests similarly to Huntingon's, with loss of motor control being one of the key symptoms. Repeats of SCA1 are within the Ataxin-1 gene, and the Davidson group postulated that knock-down of Ataxin-1 through RNA interference (RNAi) may accomplish similar results as the work of Yamamoto et al. We will discuss this flagship work by the Davidson group, Xia et al. 2004[217]. An important step of this paper is applying RNAi technology *in vivo*, as its efficacy was already established *in vitro*.

The group screened for a number of hairpins formed against ataxin-1 RNA and used HEK cells with shLacZ as their control. The most effective they identified were called F10 and F11, directly adjacent to the CAG repeats. This was verified through Western blots, which showed a decrease in protein expression, via probing for a flag tag. Such a decrease was dose-dependent. They tried with other promoters, and identified the modified CMV promoter (Pol II) as more effective in knockdown, as well as recapitulated the phenotype in cultured neurons. Knockdown efficiency was further improved via using an miRNA hairpin, rather than their original hairpin—they assumed this may be better because it more closely mimicks endogenous miRNA, and therefore might be processed easier (increasing expression). This is hypothesized to be due to enhanced nuclear transport. This held true for both mutant and non-mutant Ataxin.

After verifying their model works *in vitro*, they sought to replicate it *in vivo* via generating AAVs including it. The AAV included hrGFP, allowing for visualization of the transfected cells. Confocal imaging verified uptake, and expression of hairpins was shown via Northern blotting after 10 days.

Now, the paper truly begins. Using SCA1 transgenic mice, motor symptoms were assayed. Particularly, improper foot placement is noticed, which causes mice to struggle with rotarod tests. However, mice treated with the RNAi had significantly better performances on the rotarod test. They were not as strong as WT mice, but exhibited obvious improvements over saline treated mice. The natural next step is, of course, test the neuropathology. Calbindin staining, a molecule used as a proxy of Ca^{2+}, was used to survey the transduced and untransduced brains. By analyzing an overlay of GFP and Calbindin, and it is clear that transduced neurons were robust against SCA1 (i.e., calbindin and GFP staining were positively correlated). Comparative images show that on the left side of an individual lobule, which is not transduced, has thinning. The right side, which is transduced and marked by GFP, is robustly maintained. The width of this region of the cortex was quantified, and again was smaller in untransduced mice. They showed that nuclear inclusions of Ataxin-1 decreased, even 1 week after AAV injection. They attribute this to the rapid turnover of the Ataxin protein and because AAV expression is quick.

Implications.

Notably, their AAV expression was extremely limited. They injected only into a small portion of the brain. Their construct showed expression in $5 - 10\%$ of all cerebellar Purkinje fibers. It is surprising, then, that phenotypic improvements can be seen with such low levels of AAV coverage. One might wonder about using AAVs, which are not permanently expressed, verses using something more permanent like letiviruses. Other works showed that using lentiviruses can be bad—early clinical applications saw integration into the genome in places that caused lymphoma. AAVs, then, are preferable and can be made cell-specific.

[217] `https://pubmed.ncbi.nlm.nih.gov/15235598/`

The work serves as a great initial efficacy demonstration of RNAi-based therapeutics. However, it does not cover the potential off-target effects of the RNAi technology. From this work, it is unclear if there will be long-term side effects of this AAV injection. Similarly, as the work only surveyed phenotypes over the course of weeks and did not include later time-points measuring the RNAi's expression, it is unclear the duration of the AAV's effectiveness. Therefore, we are left to wonder if patients seeking treatments will require frequent AAV re-injections.

Despite some gaps in the work, RNAis seemed to have incredible promise. Yet, decades later, still no RNAi-based therapeutics exist to treat SCA1 or HD. Why hasn't this flagship work translated to the clinic? To start, delivery of RNAis to the brain is inherently difficult. As RNA does not readily diffuse throughout the brain, some neurosurgical operation will be required for targeted delivery, often with multiple sites of injection. This is particularly difficult when we consider the size and location of CNS degeneration in diseases like HD—which manifest themselves in central regions like the basal ganglia. Too, RNA is unstable, thereby requiring clever preservative technology, with today's research revolving around lipid nanoparticle delivery methods.

Another considerable drawback of the RNAi approach, as explained by the Davidson group's original paper, is that developing hairpins that efficiently knockdown expression is difficult. It took them many trials to get an effective RNAi for SCA1, and they commented in their discussion that they were unable to achieve similar results for HD. They mentioned in the discussion that the Huntingtin RNA, and particularly the CAG region, did not lend itself to RNAi (or was not accessible by the RNAi degrading mechanism). Though, the next year they found success in perturbing HD in mice[218], and years later they furthered this work and generalized beyond mice into non-human primates (NHP), namely in Rhesus Macaque[219]. Similarly, RNAi would target both the mutant and WT RNA equally—therefore, if the protein has essential endogenous function, that would be a major issue. For SCA1, this is not predicted to cause problems, as mice with Ataxin-1 knockout seem healthy (an arguable point that I discuss later). Finally, RNAi is not expected to completely ablate protein levels, so a partial phenotype may persist.

But again, their NHP work still extended only weeks, and did not track AAV expression over time. It is ambiguous if the diseases have a point of no return, too. Unfortunately, much of the strengthening data that they're lacking is likely due to time, money, and regulatory constraints in working with NHPs. Therefore, an inescapable roadblock exists: *where*, *when*, and *how* do you administer RNAis to make them clinically relevant?

In summary, RNAi technology provided an incredibly exciting avenue in the early 2000s, yet we still lack useful RNAi therapeutics. Some of the biggest barriers in using RNAi in the clinic include our lack of knowledge regarding their temporal and spatial effectiveness. Such questions will take time to answer, but are key to realizing the potential of RNA interference in treating neurodegenerative diseases.

CAG Repeat Length, Not PolyQ Length

Here we discuss a GWAS success in the study of HD. Recall that the age of onset is extremely variable, and does not seem to correlate with the the expanded CAG repeat exactly (that is, someone with

[218] https://www.pnas.org/doi/full/10.1073/pnas.0501507102
[219] https://www.sciencedirect.com/science/article/pii/S1525001616327009?via%3Dihub

repeat length x may have onset anywhere from 20 to 80). While there is a trend, the variance is still tremendously high.

Work furthering our understanding of this phenomena comes from the Genetic Modifiers of Huntington's Disease (GeM-HD) Consortium in 2019[220]. Because of this very high variability, this group decided to perform a GWAS strictly among those with HD to look for age of onset modifiers. After doing many GWAS, they pooled all of the data in a meta-analysis resulting in an $N \approx 10,000$. Large peaks associated with age of onset was found—one of which was, in fact, on or around the *HTT* gene itself. Again, note that this is not a study for the association of getting HD, but rather with symptoms of HD—making it surprising that there might be an association with the *HTT* gene itself.

They found an association between onset of HD and the repeats themselves. The normal HD allele includes a (CAA-CAG) sequence, which both encode glutamine. Individuals with delayed onset had $(\text{CAA-CAG})_2$, meaning it was a slightly "less pure" version of the CAG repeats. Regardless, it is still only glutamine that is encoded. Conversely, in fact, individuals lacking this (CAA) sequence altogether (resulting in less glutamine's translated) had earlier ages of onset—despite the fact that they would then have a shorter polyQ length. This suggests that, rather than the polyQ length conferring disease, it is instead the length of the CAG repeats specifically—the DNA, rather than the protein.

Follow-Up and Future Directions

Why would this matter at all? What is thought now is that the purity of the repeat impacts instability. Mosaic cells within the striatum were found to have widely varying numbers of repeats, which seemed to worsen when the CAG repeats were more pure. Similarly, those with earlier ages of onset were found to have much more unstable repeats.

The HD consortium recruited single cell RNAseq expert, Steve McCarroll, to help resolve the type of cells involved. Via identifying their cell-type, a greater loss of spiny projection neurons was found. The key question now is: what is the expansion length in each cell? Among glia and interneurons, there was very limited instability. However, in the spiny projection neurons, there was massive instability[221].

Fascinatingly, the other markers found to be associated with HD age of onset were all associated with DNA repair (many of which in the same complex). Interestingly, work in mice found that repeats are unstable in certain tissues, like the liver and striatum. Remarkably, in background with DNA-repair protein knockouts... this instability was eliminated. One of the linked genes is called MSH3. From other, very comprehensive genome studies, much data already exists on MSH3. MSH3 total knockdown is actually viable, meaning it may be a superb therapeutic target for destruction.

So, since *HTT* was found in 1993, much has been learned—but still, much remains to be.

[220] `https://pubmed.ncbi.nlm.nih.gov/31398342/`

[221] Evidently... at the time of me writing this, this work is not published! It is from a talk!

33.3 Fragile X

Background

Fragile X is a developmental disorder characterized by immensely long repeat expansions on the X chromosome. It's associated with some degree of mental disability and clear visual markers, like a longer face and larger ears. Fragile X is considered to fall within the autism-spectrum disorders, and causes multiple unrelated phenotypes (pleiotropic). As autism is becoming increasingly common, Fragile X is of considerable interest.

As a brief digression: There has been considerable, considerable, considerable work over the years solving different neural circuits, on a single or clustered neuronal level. In general, it is unclear how this information may be useful clinically. For example, suppose one were able to know the exact neuron responsible for some behavior. How can one clinically access such a neuron without impairing others? Dr. Bonini commented that TMS treatment for someone with autism has helped alleviate some of their social symptoms—so indeed, the ability to resolve individual neural circuits is important if accessible via treatments like TMS.

Back to Fragile X. Fascinatingly, the Fragile X site itself can be visually seen, due to the nature and number of repeats—there are little nub-like structures on the ends of the chromosome. Using cytogenetic techniques, one can identify carriers before there is an affected son. However, this visual marker is not found in all cells—only in $\approx 40\%$. Though, knowing the gene is on the X chromosome provides a great basis to start research—much easier than spending years narrowing down the location in the genome, such as was done for *HTT*. Fragile X is characterized by $(\text{CGG})_n$ repeats in the FMR-1 gene—interestingly, it is specifically in the 5'-UTR, rather than an exon, meaning it is untranslated altogether (one would expect). A normal number of repeats includes the range from 6 to 54, and one can have no phenotypes for up to 200 repeats. Fragile X is the most common form of inherited mental disability, and the second most common total (behind Down's syndrome). It is considered X-linked dominant—yet, around 30% of female carriers show symptoms, and 20% of males with a fragile X show no phenotypes. This is referred to as the Sherman paradox—genetic anticipation in the Fragile X context. Notably, a reduction in FMR-1 mRNA is noted in fragile X patients. Specifically, a CpG island in the FMR-1 gene is associated with methylation, and such methylation is associated with decreased expression. It was thought, then, that the fragile X mutation alters methylation patterns in this region. Due to increased repeats, the chromosome is unstable and thus mosaic genotypes are observed.

Let us dive in to some of the key works.

Probing for Instability

In 1991, when Yu et al. published their work, little was known about the location of the gene, and the genetics of it were quite mystifying[222]. Through isolating a chunk of human DNA using a yeast artificial chromosome, they generated a probe they called XTY26. Via *in situ* hybridization, they found that this probe spanned the entire X chromosome region thought to be responsible for Fragile X. At one end of this probe, which they identified through generating contigs, a marker called Alu2, and at the other end one called VK16. They further scanned the region until they found an *EcoR1*

[222] `https://pubmed.ncbi.nlm.nih.gov/2031189/`

fragment that segregated with the fragile site.

Using these markers, they attempted Southern Blots against patient genes that were digested with *Pst1*. Via probing with segments of the previously mentioned *EcoR1* strand, they found that some hybridize to longer or shorter regions of DNA. Notably, because this region is very G-C rich, it was very difficult to PCR amplify. Therefore, the only insight they got was through Southern blots. Point being: due to the varying lengths of products, they concluded the region must be unstable.

Understanding the Sherman Paradox

A great deal of insight into the so-called Sherman Paradox came from Fu et al. in 1991[223]. To start, they used a shotgun approach (splicing the gene up into parts), they sequenced the region predicted to contain the Fragile X gene. Interestingly, they compared this sequence with one previously reported and found discrepancies, which they chalked up to differences in various cloning steps. They generated a PCR test to quickly screen for variability of genes in patients, which they validated via Southern blotting. Of those they sequenced, they found a range of 5 to 54 repeats encoding arginine in this area. Of the ≈ 500 chromosomes sequenced, 29 repeats was the most common.

A plot of the distribution of repeat lengths shows the high degree of variance. Notably, non-affected carriers of the Fragile X allele show mosaic genotypes ranging from 200 to 600 base pairs (they call this the pre-mutation), and affected individuals of 600 to 4,000 base pairs with greater mosacism (they call this the full mutation). Naturally, such long repeats make PCR difficult, and using a Southern blot limits one's ability to determine the size. While they include PCR products of such patients, they may have inaccuracies due to failed amplification. Regardless, they find repeat lengths centered around 70-80. They mention having sequenced some over 200, but chose not to include them due to the questionable PCR accuracy.

In examining a pedigree with their PCR assay, they found that affected males did not generate a product at all, unless they were mosaic for lesser alleles. This means the repeat lengths were so long, no rounds of PCR would be successful. Lanes 7 and 8 of their first test used PCR products for a CAG repeat region of an androgen receptor as a method of control. They show side-by-sides of pedigrees and the PCR products, which demonstrate the heritability and instability of this allele. For example, in their first pedigree, the carrying grandmother passed along a 73 repeat, which expanded to 83 in the mother. This mother's two sons then received a full mutation, which expanded to be not visible by PCR. Their second pedigree showed a reversal of the repeats. The mother showed an allele which could not be amplified contracted in the son to only 73 repeats (a premutation form)—which is quite curious, and suggests reversal may be possible. An individual in their third pedigree is thought to be mosaic for the full mutation (i.e., one who is affected), but shows a smear for the pre-mutation form—again, quite curious. Further examples of mosaic women are shown in later pedigrees. The takeaway from this is that the gene is highly variable between generations. While it may trend toward longer repeats in each successive generation, it seems possible that it may be shorter. Similarly, the line between pre-mutation and full-mutation seems blurred, and all of this data is confounded by the possibly unideal PCR.

They throw in a Southern blot on one family mentioned in a previous pedigree and include both *EcoR1* and *BssHII* restriction enzymes—intended to help differentiate between size and methylation patterns

[223] `https://pubmed.ncbi.nlm.nih.gov/1760838/`

of the gene. They probe with a pE5.1 probe utilized in Pieretti et al. 1991 (discussed later)—a cosmid. Carriers show two bands using the *EcoR1* cutting. In affected individuals, stereotyped bands are missing. But, as per usual, it's quite difficult to meaningfully interpret these results. There is a lot of background smudging that obscures what is or is not real.

A family in which a 54 repeat length allele was found with none expressing Fragile X was studied, showing that normal-length repeats show little instability. Offspring from the 54 repeat mother all had varying lengths, some of which expanded to 60, and other which shrunk to 52. Further comparison of parent and offspring alleles demonstrates that as the parental number of repeats increases, so too does the instability which causes altered size in the next generation. And they conclude with a graphical display of the Sherman paradox.

Much of this work is confusing and difficult to interpret, after all, a smear is but a smear. It goes to show you, science was much more difficult then.

Methylation in Fragile X

By this time, the gene had been given a name, FMR-1. We had a sense of these expanded repeats, but it was unclear what affect they had. Pieretti et al. 1991 focuses on methylation patterns of the FMR-1 gene in Fragile X patients[224]. A CpG fragment (which, strangely, stands for cytosine phosphate-backbone guanine) in the FMR-1 gene is preferentially methylated in Fragile X patients. They use an RT-PCR assay to determine the amount of mRNA, and identified very low levels in Fragile X patients. The primers they used were designed for RNA only, and do not overlap with the variable CGG region.

They show the RT-PCR products of affected males. In comparing affected, carriers, and unaffected, it is obvious that FMR-1 expression is lowest in the affected population—however, there is also a marked reduction in the carriers as well. Though, they don't call attention to this and rather say that they do demonstrate some product. They attempt to be more quantitative by limiting the PCR cycles and normalizing to their control gene expression, etc. The exact methods are probably unimportant as, essentially, they find no FMR-1 in affected individuals—but they do not compare this to heterozygous females and or carriers.

They then use a restriction enzyme that normally cuts at a GCGCGC site, but is incapable of doing so when these cytosines are methylated. The Southerns shown that there is significantly less *BssHII* digestion occurring in affected patients.

Fragile X Induced Degeneration

Interestingly, long ago it was observed that Fragile X patients have significantly more dendritic spines, and this was preserved in mouse models with FMR-1 knockouts. This was known before the FMR-1 gene was identified, but scientists did not know what to make of it. Complicating things, those with FMR-1 knockout experience no degeneration, while those with the expanded CGG repeats do. Similarly, it was noticed that patients with the pre-mutation, who may not experience developmental delays but show tremor or Parkinsonism later in life. This suggested that degeneration still occurs in these individuals, but at a much slower rate. Thus, it was hypothesized that the repeats confer

[224] `https://pubmed.ncbi.nlm.nih.gov/1878973/`

degeneration. Studies after death showed that Fragile X and pre-mutation patients had intranuclear inclusions that stained for ubiquitin.

Evidence for repeat-induced degeneration was found by in *Drosophila models*. For example, in 2002 Morales et al. found erroneous neural fiber extension[225]. They begin by comparing the sequence homology of different species FMR-1 gene, and the evolutionary origin of it as well (the family tree, that is). In *Drosophila*, this gene is called *dfxr*. To validate their mutation, they use a Western blot for the *dfxr* protein, and they find that their deletion lines show no expression, and one mutation is a hypomorph. They stained in both neurons and glia, and found expression in neurons but not glia. Further, they see that *dfxr* is expressed most in the neuronal cytoplasm, rather than nuclei.

One of the benefits of using a fly model is that the neural morphology is often very stereotyped, and that there are significantly less neurons to worry about. In this vein, they assay morphological dysfunction via a reduction, or mis-localizing, of neurons in some unimportant region bridging two sections of the *Drosophila* brain. The important takeaway, though, is that this region typically shows ≈ 10 neural fibers in it, but shows less in *dfxr* deletion. Similarly, such fibers extend erroneously, with poor guidance in *dfxr* knockdown. However, as mentioned above, degeneration does not occur in FMR-1 knockdown, so this result is difficult to interpret on face.

In 2003, Jin et al. used a *Drosophila* model in which they express varying length CGG repeats of the human FMR-1 gene[226]. The group used the UAS-Gal4 system (derived from yeast, where the Gal4 protein drives expression of the gene following the UAS promoter, allowing for cell-specific expression), and a premutation of either 60 or 90 CGG repeats was added to a UAS-EGFP sequence. Notably, this construct did not include the full FMR-1 gene—it was only the CGG repeats. By both RT-PCR and Western Blot, in which they used *GMR-Gal4*, expression of the human FMR-1 gene and EGFP was found in the retina.

They note the phenotypes for different lengths of repeats and the varying levels of expression in a few different tissues. They key assay used was examining the retinal degeneration of these flies, which was exaggerated with more repeats. Interestingly, via DAPI staining, they find nuclear heatshock protein 70 (Hsp70) inclusions upon expression of $(\text{CGG})_{90}$. Overexpression of Hsp70 had the ability to rescue the retinal degeneration phenotype. This rescue does not occur so when overexpressing a dominant negative Hsp70. Therefore, this work provided a basis of degeneration, and some potential targets for therapeutics.

An Aside: Rett Syndrome

Rett syndrome is an X-linked developmental disease, where symptom onset begins around 6-18 months of age and affects only girls. At the point of onset, there is a rapid deterioration, where children may lose their previously acquired skills—this is particularly terrifying to parents[227]. It is considered the severest form of the autism spectrum disorders. Communication and motor function delays are some of the earliest observible signs. Patients may also get seizures or other more severe symptoms. It is a lifelong disease with no treatments.

[225] `https://pubmed.ncbi.nlm.nih.gov/12086643/`

[226] `https://pubmed.ncbi.nlm.nih.gov/12948442/`

[227] As a brief aside, it is thought that this phenomena partially fuels distrust of vaccines, as sometimes parents may see their child regress soon after receiving vaccines and erroneously attribute it to such vaccines. However, of course, the timing is purely coincidental.

This is a rare disease, with an incidence of $\approx 1/10,000$. It was determined that Rett syndrome is indeed lethal in males, hence only females get it. 99% of cases are sporadic, however some very rare Mendelian cases allowed for insights into the gene of interest. The mutations were in a gene called MECP2, which aids in methylation of CpG islands. It was the first disease found to be caused by mutations in a *trans*-acting factor, i.e. a gene that modulates other gene expression at the level above DNA.

This was accomplished in 1999 by Amir et al[228]. Using PCR amplification, they scanned for mutations in the MECP2 gene. These mutations were shown to be *de novo* by comparing the mother's sequence to the daughters. While the mutations are sporadic, it is possible to have a mosaic germline, in which a mother can pass on the gene to multiple offspring (at a probability impossible if it were chance alone). The MECP2 protein has a methyl-binding domain and a nuclear localization domain, which is flanked by transcriptional repression domains, which recruit proteins to help shut off transcription. It is predicted that the mutations are loss-of-function. The hypothesis is that genes that would normally be repressed are now being expressed.

Fascinatingly, early mouse models with homozygous knockouts demonstrated phenotypes similar to human disease—i.e., were not lethal, but rather were born normal. Further interestingly, more sophisticated models including *Cre*-based knockouts in only mature neural tissue phenocopied these models.

Remarkably, it was shown that there is potential to recover, or improve, if one has Rett syndrome in 2007 by Guy et al.[229] They used a *lox-Stop* site in order to create a null-allele for *Mecp2*, but which can be excised (leading to normal expression of Mecp2) by tamoxifen (TM) treatment. The efficacy was shown via Western blotting and in situ immunostaining. As a control, they compared the survival rates of *Stop* and *Stop-Cre* construct groups to demonstrate that the *Stop-Cre* does not confer any extra survival when off (i.e., in the absence of TM).

Without going in to detail for all of the symptoms measured, the mice experience substantial improvement after injecting TM. This was one of the earliest examples showing recovery from a neurodegenerative disorder was possible.

33.4 Amyotrophic lateral sclerosis

Overview

Amyotrophic lateral sclerosis (ALS) was identified by Jean-Martin Charcot around 1825 in France. His style was innovative at the time, which was to interview, record, and imitate many symptoms in order to demonstrate them to training physicians. This is also around the time photography became possible, so in addition to his drawings, pictures allowed for further diagnostic insight. Amyotrophic refers to the degeneration of muscles, lateral refers to the region in the spinal cord, and sclerosis refers to the glial scarring in the nervous system. Charcot characterized other diseases as well, including multiple sclerosis.

ALS is a quickly progressing neurodegenerative disorder where motor neurons are lost in the motor cortex and throughout the spinal cord. Patients typically die of respiratory failure if not put on

[228] https://pubmed.ncbi.nlm.nih.gov/10508514/

[229] https://www.science.org/doi/10.1126/science.1138389

permanent ventilation. This progresses to the point of total immobility, barring some motor control in the places like the eyes. This is called locked-in syndrome. If all motor movements are lost, that is called complete locked-in syndrome (CLIS). The lack of degeneration in certain places is curious. This provides a pathway for developing therapeutics, as we can look for what is unique about these neurons. The typical onset is around 45 to 65. Initially, it was thought to be purely motor-related. However, it is now known that dementia can be a symptom as well—and such symptoms show a genetic correlation (i.e., depending on which ALS-causing mutation one has, they may experience slightly different symptoms). Specifically, some forms of ALS can segregate with fronto-temporal dementia, where one can present with both diseases, or only one. There was some initial, strong support for environmental factors causing ALS, but we now know there are 20 or so genes which can cause it.

Identification of SOD1

The genetic investigation of ALS had a fast and promising start. A key player was found by Rosen et al. in 1993, and remarkably is gene was already known and published in other works[230].

ALS can either be inherited as an autosomal dominant disease (about 10% of cases), or is sporadic. The two manifest similarly in the clinic, causing resarchers to become interested in using fALS for insight into sporadic ALS. Linkage studies showed a correlation between ALS and Chr21. Rosen et al. furthered this by identifying Cu/Zn-binding superoxide dismutase (SOD1) as causative in some fALS cases. This was a key, remarkable finding because SOD1 was already known at the time. Very few proteins and or genes were understood, but in fact this was one of them. Therefore, many thought ALS would be cured in due time. SOD1 is helpful in clearing reactive oxygen species, and thus the immediate hypothesis was its dysfunction leads to ROS buildup and thus cell death. Though, Rosen and colleagues commented that this may be counterintuitive because the allele is dominant, and it seems unlikely that a loss of function allele would be dominant.

Fortunately, in 1993 the sequence of SOD1 was already known and published. This allowed for the easy design of primers for SOD1 exons. A metric they used was single-strand conformation polymorphism—which uses the structure of single stranded DNA and looks for differences due to polymorphism. They identified conformational changes in a small group of fALS families, but none in non-affected individuals in the SOD1 gene. They followed this up with Sanger sequencing results and identifies mutation sites. Importantly, these were all heterozygous, as indicated by the multiple base pairs found. Thus, SOD1 was the first known ALS gene, but many more were soon to come.

A Mouse Model

Naturally, with SOD1 known, the immediate follow-up would be to develop a mouse model. That is exactly what Gurney et al. did in 1994[231]. To do so, they used human WT and mutant SOD1. Their model was verified first through Western and Southern blotting. Their strongest cell line, dubbed G1, became impaired between 60 and 150 days and die between 150 and 180 days. Their stride significantly worsens between 150 and 180 days. They quantified the total amount of SOD1 protein being produced, and found that it was comparable between WT and G1, suggesting that the G1 phenotype is due to the mutation rather than expression alone. Their neuropathological studies found decreased ChAT

[230] https://pubmed.ncbi.nlm.nih.gov/8446170/

[231] https://www.science.org/doi/10.1126/science.8209258

(choline acetyltransferase, a cholinergic neuron marker) staining in the G1 mice. They also stained for the presence of SOD1 in the spine and found a high degree in the horns. Neurofibrillary tangles were found in many, but not all, motor neurons.

A Link Between Diseases: TDP-43

Frontotemporal dementia (FTD) and ALS are two of the most common degenerative diseases for those under 65. They present with similar symptoms, suggesting possible mechanistic crossover. About 30% of FTDs are familial and many show linkage to chromosome 17. While the inclusions in FTD are hyper-ubiquitinated, FTD does not present with any tauopathies, leaving the protein unknown. Interestingly, FTD and ALS are frequently found together in families, suggesting there may be genetic conservation between the two.

Remkarkably, in 2006 Neumann et al. found that both frontotemporal lobar degeneration (FTLD) and ALS have inclusions due to TDP-43[232]. These inclusions are marked by hyperphosphorylated, ubiquitinated TDP-43.

The group, led by Drs. Virginia Lee and John Trojanowski, began by developing antibodies which specifically bind to patients with either FTLD type 1 or type 2, and these antibodies mark ubiquitinated inclusions—already an immensely impressive feat. To develop these they would have had to inject ubiquitinated inclusions into mice and isolate single cells within the spleen. To screen these possible antibodies they tested on slices from human brains. In other words, a robust and large brain bank played a key role in this. This work is all performed on patient tissues, demonstrating the importance of neuropathologists in the study of neurodegenerative diseases. On a Western blot, these antibodies show an $\approx$ 24kD band and an $\approx$ 26kD band for type 1 and type 2 FTLD respectively. Before running their blot, they used a urea-based purification technique, which was essential for removing any and all soluble proteins so that only the insoluble aggregates remained. These bands were not found in controls or AD brains, and similarly tau staining was not found in FTLD brains.

To determine what these antibodies might be binding to, they performed a 2D page gel and looked for a bands around 25kD. Because similar spots of $\approx$ 3.5pI were found on both type 1 and type 2, they predicted this was correct. They performed mass spectrometry and identified the protein as being TDP-43. Both fragments of TDP-43 are truncated at the C-termini. Notably, this gene is on chromosome 1. Using immunostaining, they further proved their TDP-43 hypothesis. In type 1 and type 2 FTLD, as many inclusions as were labeled by TDP-43 antibodies were labeled by ubiquitin. Interestingly, type 3 FTLD had inclusions not labeled by these antibodies. Additionally, other degenerative diseases did not show any TDP-43 staining, showing specificity for type 1 and 2 FTLD.

"Aggregated" TDP-43 was found only in the urea fraction after Western blotting. Non-aggregate TDP-43 was found in all fractions. Blots showed that the molecular profile of TDP-43 was different in FTLD than it was in other degenerative diseases. Specifically, extra bands around 45kD were found. We can postulate from previous papers that this may be due to hyper-phosphorylation. Indeed, dephosphorylation of proteins via a phosphatase collapsed them into a single band. Interestingly, the higher weight TDP-43 band is also ubiquitinated.

They extended this work to ALS by verifying TDP-43 also labels ALS inclusions. Interestingly, they

[232] `https://pubmed.ncbi.nlm.nih.gov/17023659/`

suggest TDP-43 is normally a nuclear protein, but in disease pathologies becomes cytosolic. A cleaved, C-terminal product, marked by a low molecular weight band, was found only in the frontal cortex. However, phosphorylated TDP-43, represented by all other bands, was found throughout other regions of the brain and spinal cord. Additionally, some of the FTLD families were mutant for a different gene called PGRN. The relation between PGRN and TDP-43 is unknown. Regardless, TDP-43 aggregates characterize nearly all diseased ALS brains.

Inclusions: What of it?

Soon after the publication of Neumann et al. detailing the potential role of TDP-43 aggregates, a 2007 piece from Rothstein was published criticizing it[233]. He comments largely on the over-statements frequently made regarding aggregates. He comments that some nuclear inclusions have been shown to be protective, rather than disease-causing, in some CAG repeat diseases. Rothstein points out that TDP-43 inclusions were found in non-motor cells and in regions clinically unaffected by ALS. Therefore, this calls in to question their relevance in disease progression. Similarly, because these inclusions are not found in fALS mutant for SOD1, there are further questions. The author proposes that, because sALS is a slower progressing disease than fALS, more time is given for aggregates to build up. Therefore, they are not necessary for disease progression, but rather another symptom.

An all-important point is that we only look at a brain once the patient is deceased. It gives no insight into how the disease was "generated." Therefore: perhaps the cells with inclusions are the survivors, rather than the sick ones. The work fits in to the grand wonders about the true role of aggregation in disease.

Follow-Up.

Follow-up work looked to sequence TDP-43, and many mutations were found in the glycine-rich region (near the C-terminus) of TDP-43. Other similarly structued proteins, which had RNA-recognition motifs (RRMs) include proteins like FUS—which was later found to aggregate in ALS as well. Interestingly, FUS does not typically have many mutations in the glycine-rich region, but rather it was in its C-terminus. Similarly, in FUS mutation, FUS is trafficked outside of the cell. The C-terminus area, then, was determined to be relevant in nuclear localization. However, TDP-43's mutations are not often in its nuclear localization sequence. Both FUS and TDP-43 have NLS and NES, meaning they can be trafficked back and forth between the cytoplasm and nucleus as needed—but both spend much more time in the nucleus.

Interestingly, both FUS and TDP-43 are localized to stress granules in these diseases. Important insight regarding this came from *C. elegans* embryos. Such granules begin scattered about the cell, but can be trafficked to a desired location. Fascinatingly, though, these granules do not get "moved," per se, but rather disappear and reappear. That is, they are not being dragged along a microtubule, and or, being pulled to the desired location like you might expect. Rather, they are phase-separated liquid droplets, which move out of phase and spontaneously back in to phase with a different localization. Stress granules are similar to the nucleolus—they are effectively membrane-less organelles. Both TDP-43 and FUS have "low complexity domains" which allow for the gel-like formations of these proteins. It's hypothesized that they are stored in liquid droplets and then transition somehow into being hard, solid aggregates. One of the most fascinating questions, then, is how non-mutant TDP-43 still exhibits these stress granule to aggregate conversions. Somehow, continued stress in the neurons

[233] `https://pubmed.ncbi.nlm.nih.gov/17469124/`

leads to this. What causes this?

I should say, before we continue, what is TDP-43 anyway? It is a nucleic acid binding protein which is capable of doing a lot in RNA maintenance and processing, including splicing. TDP-43 is considered a repressor, helping to prevent expression of certain exons. Therefore, both TDP-43 loss of function and gain of function are capable of producing different, new proteins with either extra or less than intended exons. TDP-43 especially represses "cryptic exons"—which are not found in mice (what does that mean for mouse models?). Therefore, we can only use RNAseq from humans to determine how TDP-43 dysfunction alters the transcriptome. Broadly, TDP-43 seems to make transcription more accurate and precise. Thus, it is wise to hypothesize that improper transcription of critical genes may lead to neuron death. One example of this is stathmin-2 (STMN2), which has a cryptic exon and is known to play a role in motor function and motor neuron outgrowth (it is a canonical regeneration marker). Again, this is not preserved in mice—so we cannot investigate it properly in a mouse model. As a final comment, TDP-43 is autoregulated, so it is not an ideal therapeutic target.

Most Common Allele Finally Found: C9ORF72

This is a very interesting story, which shows how our technological advancements were accompanied by some limitations. First, let's consider DeJesus-Hernandez et al. 2011[234].

It was shown years prior that ALS and FTD are linked to a gene on Chr9. Yet, finding this gene proved impossible. A large reason for this is the advent of next generation sequencing, which caused researchers to miss the key mutations. All of the exons in this region were sequenced, so the Rademaker group questioned if the causative mutations could be in the introns. Via attempts to PCR these introns, one intron failed to amplify. This work identifies this gene as *C9ORF72*, and that the causative mutation is a repeat expansion in a non-coding region.

The group used brain tissue from patients showing Chr9 linkage, and staining them clearly showed TDP-43 inclusions. They show a pedigree including patient haplotypes that implies a lack of transmission from affected parent to affected offspring. Via PCR amplification, they size different alleles in their suspected gene and find that affected individuals only show one size. This is curious, and suggests affected individuals are homozygous—which cannot be true. This instead implies that the gene cannot be amplified. They then turned to a Southern blot, and by probing this region, they found that diseased individuals had very long alleles. The expansions they found were on the order of thousands of base pairs composed of GGGGCC repeats—on the scale of previously mentioned diseases, this is quite long. After finding the mutation, they showed that it is more common than any other ALS or FTD mutation, even more so than SOD1, FUS, and TDP-43, etc. At many clinics around the world, it was on the order of 50% of cases included this repeat expansion.

Based on one SNP, the group can tell how the RNA is cut, and demonstrate the possibility that these expanded repeats cause alternative splicing. Patients with the expanded repeats have a much higher proportion of the V2 alternative splice, which perhaps alludes to a role in pathology. Finally, they show that the expanded repeat RNA forms nuclear inclusions, which they identify via a $(GGGGCC)_4$ probe.

A similar, parallel study was done by Renton et al. in 2011[235]. The Traynor lab used extensive

[234] https://pubmed.ncbi.nlm.nih.gov/21944778/
[235] https://pubmed.ncbi.nlm.nih.gov/21944779/

next generation sequencing to look for the changes in the critical region of Chr9. Next generation sequencing specifically creates a shotgun library of genomic DNA, sequences it, and aligns them together. Each fragment is on the order of 50 to 100 base pairs. This is not *de novo* sequencing, but rather comparative sequencing—one compares and or aligns with a reference sequence from the original human genome project. In affected individuals, they were reading with a depth around 170 (read depth describes how often an individual base is read, meaning each individual nucleotide was read around 170 times), meaning they had incredible depth. Part of the computation done is to align uniquely mapped reads, meaning segments of DNA which are conserved in different parts of the genome cannot be mapped—hence a struggle with repeat expansions. The Traynor group therefore was able to see a massive gap in the reads done by their sequencing (i.e., a large blank spot in the output), suggesting some error in computation, which they supposed was due to repeat expansion.

To sequence this unknown region, they employed a strategy called "repeat primed PCR," where a reverse primer and an anchor primer are used. Part of the reverse primer is homologous to the variable repeat, and the anchor is homolgous to a non-repeat expansion. The reverse primer will then be capable of binding anywhere along the repeat, generating various amplification products. Adding the anchor primer after allows you to amplify all of these fragments. You will then, in short, see a wide variety of products when there are repeats. Indeed, they found long strings of GGGGCC repeats.

Implications.

One possible implication would be that a long repeat leads to RNA accumulation, and soaks up tons and tons of RNA binding proteins, leading to a dual sort of gain-of-function-loss-of-function result. Notably, no point mutations have ever been found in this gene, suggesting loss-of-function alone is not disease causing. Opposed to something like FMR-1, it must be true that repeat expansions alone are causing the disease. Therefore, it may also be that the protein is translated strangely, including translation through introns, leading to aberrant protein production.

This gene is now part of the common disease haplotype that characterizes ALS. Healthy individuals have 2-23 units of the GGGGCC repeats, while affected individuals have nearly 700-1600. Yet, while this gene is extremely unstable, there is no observed genetic anticipation—a confusing finding. It thus remains unknown what the "minimum" length is required to cause disease, as there are no "middle" cases (i.e., a patient with mild ALS and 300 repeats does not exist). Interestingly, two non-ALS patient haplotypes exist: A and G. The A haplotype has more repeats seen than does the G haplotype, so some wonder if it be more prone to expansion. However, both haplotypes are still well within the normal range.

Another fascinating thing to note is this mutation can confer FTD or ALS, or both. Different family trees seem to be biased toward one disease or another, suggesting more genetic factors are involved, or environmental factors. Because of the similarity in this haplotype, and the localization to certain countries, it's thought that this repeat expansion may have originated in Europe from Viking times. I.e., it's possible this repeat expansion occurred only once, and has been carried about ever since.

The Repeats are Translated

Insight into the disease mechanism of *c9orf72* came in 2013 from Mori et al.[236] They showed that these repeats are translated via abnormal protein synthesis, resulting in protein aggregation. Hairpin formation can lead to the incorrect recruitment and initiation of translation. This is called RAN—repeat associated non-AUG translation. The G_4C_2 repeats found in ALS are capable of forming both hairpins and quadraplexes, implying RAN can occur. Interestingly, it's also possible for translation to occur on non-AUG, but on similar codons. This is typically seen under stress or in abnormal conditions.

Based on the sequence of *C9orf72*, it's unclear where the ribosome may be tacking on and initiating translation. Different translation start sites, however, will generate different proteins if they are in different reading frames (such possible proteins are called GA, GP, and GR, where repeats generate Poly-GA, GP, and GR). Interestingly, all predicted proteins would be very hydrophobic, suggesting a possible disease mechanism. The instrumental advancement was their uniquely generated antibodies—which they raised against the three possible proteins.

They demonstrated, in human embryonic kidney (HEK) cells, that different dipeptide repeat proteins are aberrantly translated, depending on the reading frame and in correlation with the length of DNA. They found GA expressed over a range of lengths, GP expressed only at the longest lengths, and GR never expressed. However, upon using a filter trap to remove aggregated protein, they found GA, GR, and GP, still with GA being the most common. Therefore, the disease must somehow lead an odd intron to be removed from the nucleus and somehow transcribed depending on the repeat length. They further verified the presence of such proteins in *c9orf72* patient tissues. The proteins form inclusions that vary in both location and morphology, are distinct from TDP-43 inclusions, and are marked by ubiquitin and p62. It was long known that *C9orf72* patients are seen as having additional, non-TDP-43 inclusions, and their molecular identity was finally known. Due to their ubiquitin and p62 markings, it's plausible that they might induce autophagy.

Glia Are Key in ALS

The autonomous vs. non-autonomous nature of neurodegenerative diseases will become a large topic in the following few sections. A key example of the non-autonomous nature of ALS came from Clement et al. in 2003[237], where they investigated the role of glia in hastening or slowing disease progression. The key takeaway here is that both glia and neurons must express mutant SOD1.

Chimeric mouse lines were generated by combining the morula of mice with either WT or mutant non-neuronal cells. Such chimeras contained SOD1 mutants and those expressing low levels of human neurofilament subunits. Loss of motor neurons was found to be asymmetric—which was surprising, given that some of their lines included 100% mutant neurons. It was found that the neuronal sparing was higher on sides of the spinal cord where the proportion of wildtype, non-neuronal cells was greater. This signified some potential neuroprotection associated with non-neuronal cells. Similarly, it was hypothesized that mutant SOD1 expressing non-neuronal cells could be capable of killing motor neurons as well.

Another key finding was that, even in mice where nearly no WT motor neurons existed, there may still be no phenotype (i.e., all motor neurons having the mutation does not necessitate disease). All

[236] `https://pubmed.ncbi.nlm.nih.gov/23393093/`
[237] `https://pubmed.ncbi.nlm.nih.gov/14526083/`

of this data points to the hypothesis that non-neuronal cells, namely glia, require dysfunction before an ALS phenotype can emerge.

Another remarkable work in a similar vein is Nagai et al. 2007[238]. They showed that mutant astrocytes were directly able to harm neurons. In cell culture, transgenic astrocytes expressing mutant SOD1 caused more cell death than non-transgenic astrocytes, irrespective of if motor neurons harbored the same mutation. Fascinatingly, media conditioned by transgenic astrocytes resulted in greater neuronal death than did normal media or media conditions by non-transgenic astrocytes. This suggests mutant astrocytes release toxic elements, causing neurons to die.

This paper also found that non-astrocytes (that is microglia, myocytes, etc.) have no effect whatsoever on motor neurons—a very surprising finding. Similarly, non-spinal motor neurons are not affected at all by transgenic astrocytes. Even other motor neurons derived from other parts of the body were unaffected by these mutants.

Certainly, this work was a landmark in the field. It unquestionably shifted the investigation into ALS, and provided considerable promise. However, in the decades to come, little advancement was made. Many searched, and searched, for what was being released by the media. Endless screens were ran and came up unsuccessful. Today, what is left in this media remains unknown. In 2021, there was some promise when researchers found that the toxic product in the media may be lipids—specifically non-saturated lipids. A working hypothesis, then, is that non-saturated lipids are preventing membrane fluidity, somehow impacting neuronal membranes globally. This is all speculative in nature, though.

Taking Neurons out of Neurodegeneration

We will now digress to discussing X-linked spinal and bulbar muscular atrophy (SBMA) in order to make a point, specifically through Cortes et al. 2014[239]. SBMA is an X-linked neuromuscular disease caused by a CAG repeat expansion in the androgen receptor (AR) gene. Surface AR binds to glucocorticoids, leading to nuclear localization where they modulate transcription and play a key role in development and sexual differentiation. The disorder is inherited, but onset begins as an adult, manifesting first as muscle weakness. SBMA is part of a large class of poly-glutamine neurodegenerative diseases featuring protein aggregation, alongside HD and SCA. While past work has focused greatly on neurons, this work suggests the answer to treatment may be in non-neuronal cells, which may have broader implications for many more degenerative diseases. Indeed, Cortes et al. show that removal of the disease causing allele from muscle cells is sufficient to halt disease progression of SBMA, representing a possible paradigm shift away from the study of neurons in neurodegenerative diseases.

Cortes et al. were able to achieve this result from the well-tested model they build. They designed a transgenic mouse line encoding for the human androgen receptor with CAG genomic repeats. The insertion included around 50kb on both the 5' and 3' sides, which was helpful in ensuring the promoter and other regulatory domains were included with the gene. The transgene was "floxed" (loxP sites flanking both sides of the *AR* exon 1) and made from a BAC (bacterial artificial chromosome, which are able to carry very many kb). The key advancement, then, was allowing cell-specific excision of the gene via the Cre-Lox system—the transgene was open to excision by the Cre-recombinase. Therefore, these authors could selectively remove the disease causing gene from different tissues by expressing

[238] `https://pubmed.ncbi.nlm.nih.gov/17435755/`
[239] `https://pubmed.ncbi.nlm.nih.gov/24742458/`

the Cre-recombinase in them. Notably, the mouse androgen receptor remained intact and functional in all tissues, meaning their model focused only on the gain-of-function components of the disease.

To ensure their model worked as intended, they verified expression of the human AR gene containing the expanded repeats and that it was similar in strength to the endogenous AR. They validated that these mice experienced SBMA-like symptoms that were comparable to other, previously used SBMA models, including muscle weakness, sickly looking motor neurons, and early death. They then verified their ability to knock down the transgene's expression via a global Cre (CMV-Cre). Indeed, they found limited expression, no neuromuscular phenotypes, and no lethality in this negative control. Finally, they ensured that expression of the Cre-recombinase in muscles, via a muscle-specific promoter (HSA-Cre), would effectively knock down the transgene in only skeletal muscles. This experiment left the disease causing allele completely intact everywhere in the mouse body, including neurons, except muscle tissue. Remarkably, nearly all SBMA phenotypes were eliminated, suggesting that muscle cell pathology is required for disease progression.

As mentioned above, this work comes on the heels of others like it, which suggest that neurodegeneration is non-cell autonomous, such as Nagai et al. 2007. Other works found that expression of varying length polyQ repeats in the AR gene (ARxQ) show muscle degeneration first, or only myopathy, which proves to be sufficient for SBMA-like phenotypes in mice. This work was particularly promising, though, because it implies *treating* muscle cells alone will be sufficient to cure or suppress symptoms associated with SBMA. Skeletal muscle, being readily accessible, makes it a much more viable therapeutic target than neural tissue.

Implications.

However, there are some drawbacks. Their model itself shows clear neuromuscular symptoms, but the underlying pathology is noticeably different from the human disease. While motor neurons in the spine die in humans, they only become sickly and shrunken in their mouse model. This suggests their model may be a more mild version of the disease, or perhaps misses the mark on replicating the key components of the disease. Instead of seeing less neurons in the spine, they see a shift from larger spinal neurons to smaller ones. On face, this might not be such a concern. However, Cortes et al. showed that removal of the allele from muscle cells ameliorated symptoms and *some* aspects of the disease—but, protein aggregates were still found in the cortex of mice. This may mean that, in the human disease, these neurons would still die and confer SBMA symptoms. Another possible explanation is the endogenous mouse AR protein expressed in the background. While their model induced the gain-of-function phenotypes associated with AR expanded repeats, it lacked possible loss-of-function that may accompany it.

Because the gain-of-function associated with genomic repeats and protein aggregation is so striking, corresponding loss-of-function is often overlooked. Though, some works have focused on these aspects of both SBMA and other repeat expansion diseases, like SCA1. SCA1, mentioned previously, is caused by repeat expansions in the gene encoding for Ataxin-1. Complete loss of Ataxin-1 does not confer SCA1, which immediately draws attention solely to gain-of-function. However, some works have found that, while the gain-of-function mechanism seems dominant, some aspects of disease may be enhanced or sensitized due to loss-of-function as well[240]. While loss of Ataxin-1 may appear trivial, loss of the androgen receptor certainly is not. AR is responsible for broad male sexual development

[240] `https://www.sciencedirect.com/science/article/pii/S0021925820324558?via%3Dihub`

and transcriptional regulation; loss of AR confers other diseases, like androgen insensitivity syndrome (AIS). In an SBMA *Drosophila* model, the ability of AR to associate with some of its native binding partners was necessary for degeneration, suggesting the natural role of AR, and alterations to it, is relevant to disease[241]. In fact, there is crossover in AIS and SBMA symptoms, further suggesting a loss-of-function contribution to SBMA progression. Surprisingly, even wildtype AR overexpression alone in skeletal muscles is capable of conferring SBMA-like symptoms[242]. One must wonder, then, if the double dosage of mouse androgen receptor and human androgen receptor may have unintentionally contributed to the disease as well. In any case, leaving wildtype AR intact in these mice may not capture the full range of the disease, and may even lead to unintended side-effects.

Surprisingly, despite the capabilities of their model, they did not test knockdown in any tissues beyond skeletal muscle. To my knowledge, this still has never been tested. Readers are left to wonder, then, what the true role of neurons is in disease progression. Could it be that neurons are completely auxiliary? Or is it that disease progression is only possible when both neurons and muscles express this mutant allele? A great addition to their work would have been the same experiment, but with knockdown in neurons instead of muscles. Irrespective of the outcome, these results would have been very illuminating.

While they are still some open questions, Cortes et al. present a very profound finding: they show that SBMA, a neurodegenerative disease, can be impressively treated by only targeting muscle cells. The study of neurodegenerative diseases is constantly met with dead ends, or mazes which seem to lead nowhere. It is rare for a study to redefine how we view a disease, but this is surely one of them.

33.5 Prion Diseases, and the Prion-like Behavior of Others

Overview

Prions are transmissible diseases. They can be "caught" in that way. Prions confer what's called transmissible spongiform encephalopathy (TSE). Understanding of this disease was kick started with the scrapie epidemic, when farmers attempted to treat sheep with a vaccine. It was noted that the "scrapie virus," as it was known then, could not be treated via a vaccine. A second historical example was the Kuru epidemic, where women and children started experiencing degeneration, but no obvious genetic pattern or route of transmission was clear. Anthropologists postulated it was due to their consumption of human brain tissue, which only women and children did. Notably, the phenotype of the Kuru people and of sheep with Scrapie was similar in both behavior and neuropathology. Gajdusek then tried to inject the Kuru tissue directly into chimps to see if they'd develop the disease, and indeed they did—winning him the Nobel Prize.

In humans, Cruetzfeldt-Jakob disease, especially iCJD and vCJD, are the prime types of transmissible prion disease—the first from contaminated medical devices and the second from eating cows. Long ago, it was common to take dura mater from cadavers and transplant it onto individuals without properly formed dura mater, which could transfer prions. Similarly, growth hormone was regularly purified from cadavers, which also induced prion disease.

Initial rounds of testing showed that normal things, like boiling, UV, radiation, autoclaving, etc.,

[241] https://www.annualreviews.org/doi/10.1146/annurev-physiol-030212-183656
[242] https://www.pnas.org/doi/full/10.1073/pnas.0705501104

which are expected to kill viruses, were ineffective in killing prions. This is important, as sick animal tissue is never sold for human consumption. However, such tissue may be used, sterilized, and then sold for making feed for other animals. However, prions cannot be sterilized, meaning this enabled their transfer to other animals.

So, prions can be transferred in a variety of way. Ingestion, clearly, is one. Injection, clearly, is also one. And, finally, experimentally is one. Such transfer can occur between species. As touched on later: Prusiner suggested that prions may indeed not be virus at all. Perhaps they are protein only, which is now called the *protein only hypothesis* and led to an initially controversial Nobel Prize in 1997 (few accepted this hypothesis!). Prusiner determined that prions cannot be eliminated by proteases, and that their sequence is shared among all humans—therefore, it is infectious, and yet it is endogenous.

SNPs Relaying Symptoms

Fatal familial insomnia is a familial prion disease; familial prion diseases are mysterious. All mutations are found in the PrP gene, yet present very differently. Medori et al. in 1992 determined some of the disease causing alleles for fatal familial insomnia[243]. The prion protein is quite small—only on the order of 300 amino acids. It's predicted to be membrane linked. Interestingly, there are many polymorphisms that are not disease-causing, per se. Rather, once one has the disease, such polymorphisms can confer different phenotypes.

Protease K is a common assay to destroy protein, which was employed by Medori et al. Pion disease patients show a persistent band, even after exposure to PK, demonstrating it is not available for degradation. Among the two families analyzed, both presented with D178N mutation. However, whether or not they had a M129V mutation determined how their symptoms presented. This introduces the idea of *prion strains*. In this world, prion strains refer to the different conformations of the prion protein, and the disease phenotype they confer.

Protein, Not Virus

Telling et al., published in 1996, is considered the flagship paper because it helped to prove that different prion diseases are due to different conformations of the same gene, and further that prions are protein only[244]. The notation here is PrP^C refers to the WT protein, and PrP^{SC} is the "scrapie" form. In other words, this denotes the spontaneous change from an alpha helical protein (C) to a very stable, beta-pleated sheet protein (SC)—this event occurring spontaneously is incredibly rare in WT. The big open question was how there can be different diseases associated with the same, general pathogenic origin. The hypothesis was that there are many conformations of the same prion protein—prion strains.

The group used a mouse model in which they injected different prions. All mice had the same (verified through sequencing) transgene. Importantly, a "species barrier" exists between humans and mice, where differences in sequences make it difficult for human prion protein to induce errors in mouse prion protein. Therefore, transgenic mice are be preferred. Using a very small amount of brain tissue from fCJD patients and fatal familial insomnia patients and injects them into mice. Western blots were performed to compare human and mouse tissue. All protein was first digested with proteinase K

[243] `https://pubmed.ncbi.nlm.nih.gov/1346338/`
[244] `https://pubmed.ncbi.nlm.nih.gov/8953038/`

to eliminate non-prions. They were then also digested with PNGase to remove linked oligosaccharides. Comparisons are made between human and mouse brains (mice that were inoculated for with tissue from the human brains, that is).

As they all express the same control PrP gene, differences arising must be due to injected prions templating differently onto the endogenous proteins—thereby generating different strains. An essential component here is that these animals, when injected with the misfolded protein, require the endogenous PrP gene to template onto. That is, mice with PrP^{KO} do not get prion diseases.

Fall Out.

Pruisner would go on to win the Nobel Prize in just the next year. However, many were still extremely skeptical, as it was difficult to show with certainty that nothing else was injected besides proteins (for example, perhaps there could be oligonucleotides or lipids wrapped within the proteins). Contrarians called for the production of synthetic, amino acid only prions before this hypothesis could be confirmed. It took years for fully synthetic prions to be made by Pruisner, largely closed the door on the hypothesis. Pruisner also threw out the possibility that other neurodegenerative diseases, which show overlap in mutant proteins, may be due to similar differences in "strains" or conformations of such proteins—a bold claim.

Meanwhile, something quite curious was noted in another disease. As Parkinson's disease is caused by loss of dopaminergic neurons, implantation of dopaminergic neurons into the third ventricle of the brain became a treatment option. This treatment indeed helps alleviate some of the PD symptoms. Fascinatingly, though, after death this tissue was examined, and shocking there were aggregates found within the transplanted tissue. Therefore, it's thought that propagation of the disease causing protein, α-synuclein, was acting in a prion-like manner. But, it is not a prion disease, and cannot be transmmitted from person-to-person. How this can be reconciled will be... discussed later!

Parkinson's Disease Spreading

Background.

We are digressing now to Parkinson's disease. PD is the most common movement disorder, and the second most common neurodegenerative disease, behind only AD. Parkinson described the disease thoroughly in 1817. Parkinson's Disease is characterized by Lewy body accumulation, first characterized in 1923. Notably, though, other diseases are characterized by Lewy body formation, like Lewy body dementia. In PD, the substantia nigra, a darkly pigmented region (hence the name) around the floor of the third ventricle, is one of the most strongly affected regions of the brain. There are a wide range of phenotypes which progress over time. Many early symptoms are typical in non-disease contexts, like anxiety, depression, and trouble sleeping. It is not until there is significant neural loss, around 90%, that one would even pursue seeing a doctor, as it is only then that abnormal symptoms begin. Naturally, treatment would be much stronger if caught earlier, but this is largely an impossibility without some genetic insight.

Lewy bodies appear like large, circular modules within the cells and in 1997 were found to stain for α-synuclein, a protein which will form fibrils in vitro. $\alpha-$syn is a small, around 140 amino acid protein. This discovery led to the finding that mutations in $\alpha-$syn cause PD. Braak proposed a timecourse for PD where disease seemed to propagate out from the central regions of the brain. His model was

loosely classified into stages where PD pathology originates in the brainstem and leaks out into the cortex as the disease progressed. It was not clear how credible this description was, as it was initially based on a very small sample size.

Notably, it was originally thought that PD was strictly sporadic, and that there was no genetic component. In fact, many of today's world leaders in the study of PD will tell you that, while in medical school, they learned this fact explicitly. We now know, of course, that this is untrue—in fact, there are very many PD-related genes. Their loci are typically called PARK, and the gene encoding for $\alpha-$syn, *SNCA*, is PARK1. For a disease to be cleanly classified as PD, it requires both symptoms and the Lewy body pathology. A lack of Lewy bodies thus dubs the disease "Parkinsonism." Recessive genes often result in Parkinsonism, where disease progression may be slower and less severe.

Great insight into PD came from model organisms with homologs to PD loci. For example, in *Drosophila*, loss of function in mutations like Parkin and Pink1 result in mitochondrial dysfunction. It's thought that PD may impact mitophagy and or mitochondrial turnover (see section **22.1**). A clean assay is to look at mitochondria is in the wing discs of flies, which are typically packed with mitochondria and would be lost in a PD model. Additionally, this manifests behaviorally as a loss of flight.

Early evidence of dopamine's key role in PD came in the 1960s, when the use of L-dopa to treat comatose patients with encephalitis led to their awakening, covered in a movie called *Awakenings*. Further great insight came from drug addiction in 1982. Synthetic heroin was being made, which resulted in a side product being produced called MPTP. Remarkably, this compound was taken up specifically by dopaminergic neurons in the substantia nigra, and killed these neurons almost immediately. Therefore, these addicts immediately got PD—in other words, there was a sudden epidemic of PD all sprouting from one heroine lab. This has compound has become a very helpful way to generate PD models. Many such similar neurotoxins have been discovered in recent years, after great study, such as in pesticides which result in higher incidences of PD in rural areas.

Synuclein Spread.

Because treatments like L-dopa have been successful in treating PD, researchers began to wonder if grafting in dopaminergic neurons would be successful. Many rounds of this were ran in parallel, with one such key paper being Li et al in 2008[245]. These investigations took a considerable amount of time, and brain tissue could only be analyzed on the order of 10-15 years after grafting. Patients that passed earlier (i.e., 2-3 years after implantation) showed no noteworthy pathology, besides the impressive survival of grafted tissue. The graphed tissue was taken from fetuses, meaning the tissue was quite young. Therefore, as later term patients passed after a decade or so, it implies the grafted tissue couldn't be more than 20 years old, eliminating the chance of natural PD arising in such neurons. Still, shockingly, Lewy bodies were found in grafted neurons of patients who survived 10-15 years. Presumably, then, the sick tissue was infecting the healthy tissue. This was the first suggestion that Lewy bodies could, somehow, transfer from neuron to neuron. These bodies stained for both $\alpha-$syn and ubiquitin.

Similar papers came out at the same time as this one. Brundin, the senior author on this work, described in a review that grafted tissue was originally thought to be disease-resistant due to the

[245] `https://pubmed.ncbi.nlm.nih.gov/18391963/`

results of grafted neurons in early dying patients. However, the analysis on patient's that died after 2-3 years were done when $\alpha-$syn was not known. Therefore, it's plausible that pathologists then didn't know precisely what to look for, and perhaps Lewy bodies may have transfered in those early cases too. To my knowledge, this has not yet been looked at.

Fall Out.

Certainly, this provided some support for Braak's ideas of neurodegenerative diseases spreading from some central point. Braak's investigations, though, went beyond just $\alpha-$syn. He observed similar progression for tau. Tau is a hallmark of many different diseases, called tauopathies, and examples include CTE, Pick's diseases, and, canonically, AD. This suggests, that like with prions, tau exhibits different strains that can spread and confer different diseases.

Spreading and Strains of Tau

Given the previously mentioned $\alpha-$syn findings, researchers questioned if similar things can occur with tau. Clavaguera et al. 2009 tested this[246]. The core of their work was taking WT-tau expressing mice and injecting in mutant, filamentous tau generated in other mice. They observed the propagation of tangles throughout the brain, and thus were the first work to show evidence of the prion-like behavior of tau. Such tangle presence was marked by tau antibodies and Gallyas-Braak silver staining. It was found to spread and increase over time, on the order of 10 or so months, depending on the region of the brain. While tau tangles were initially induced by human tau, there was no staining for human tau in the mouse tissues, demonstrating that the filaments formed were specifically due to templating of the endogenous, mouse tau.

In 2014, Sanders et al. showed that tau has "strains" which result in different tangle morphology[247]. The key question of this work is if different strains of tau can be created in stable cell lines. They begin by injecting a number of different fibrils / amyloids into HEK cells—generated somewhat randomly in vitro. They find that Tau, Htt, and $\alpha-$syn all seem capable of inducing inclusion formation. Tau inclusions were halted when the cells' endogenous tau was mutated to be unable to aggregate. Cell lines showing tau inclusions at Day 3 after exposure could be selected and used to grow a cell line containing such inclusions. These lines exhibited "strains" conferring different morphologies, patterns, and molecular weights that were conserved through cell divisions and continual re-platings—i.e., presumably endogenous tau continued to re-template with each cell cycle and preserve the inclusions. Presumably, the cell lines did not die out because HEK cells are more robust than neurons.

They furthered this work by taking strains from different patients with different tauopathies and attempted to generate cell lines from them. Indeed they find very variable and seemingly stain-specific phenotypes in their generated lines. They characterized their phenotypes as no seeding, toxic, mosaic, ordered, disordered, and speckled. Interestingly, though, these phenotypes were relatively consistent within diseases. For example, AD exhibited typical speckled patterns, while corticalbasal degeneration (a different tauopathy) more often exhibited disordered morphologies. This suggested different strains of tau confer different sorts of tangles, which are consistent between diseases.

[246] `https://pubmed.ncbi.nlm.nih.gov/19503072/`
[247] `https://pubmed.ncbi.nlm.nih.gov/24857020/`

Meaning.

A very interesting point is that, because these fibrils were taken from deceased brain tissue, there is the question of selection bias. If we assume tau may seed differently, we can then wonder if the more deadly tangles are already lost by this point, only leaving behind the non-clinically relevant ones?

John Trojanowski spent a great portion of his career as a pathologist characterizing the different morphologies of tau, and other aggregating proteins, in disease contexts. Similarly, he isolated antibodies which were able to specifically label certain morphologies specific to different diseases. Today, techniques like Cryo-EM massively enhance our understanding of proposed "strains." As you'd likely predict by this point, different conformations of tau is found, depending on the disease. Thus, structural biology confirmed what pathologists had proposed decades prior.

Importantly, it does not seem that tauopathies are transmissible like prion diseases. Generating good PD models has been a consistent trouble in the field. Many reseachers preferred fly models over mouse models, because mouse models struggled to phenocopy patients. Some of the greatest PD models, which came from the Virginia Lee group, were from injecting human α-syn into mice. Importantly, prion researchers disagree that tau and $\alpha-$syn spreading makes these prion diseases, specifically because they are not transmissible. Thus, some call these "trans-synaptic spreading" diseases instead. Indeed, similar to the name's suggestion, one of the things Braak and others noticed is that spreading seemed to occur along neural connections, suggesting it occured trans-synaptically.

Comments on Advancements and Directions

Induced Pluripotent Stem Cells.

In 1957, Waddington proposed the "landscape model" of a cell, where a cell begins at the top of a hill (as an embryonic stem cell), so-to-speak, and roll down hill as it differentiates into some valley. In 1962, Gurdon combined the nucleus of a somatic cell and an enucleated egg could be combined in a way that resulted in a genetic copy of the first frog. This implied differentiated cells still contained the necessary instructions as the cells at the "top of the hill." Gurdon later shared the Nobel Prize for this work. In the 80s and 90s, many groups were working on this and found that different transcription factors could convert one type of stem cell to another. Yamanaka, in the early 2000s, wondered what the key factors were, and asked if one could return these cells to a state of pluripotency. This culminated in the Nobel Prize winning work Takahashi & Yamanaka, 2006[248].

Today, induced pluripotent stem cells (iPSCs) are about as ubiquitous as an experimental model can get. Pluripotent stem cells are those that are capable of differentiating into all three of the germ layers (ectoderm, mesoderm, and endoderm). Using embryonic stem (ES) cells as a method of therapeutic regeneration had already been considered, but of course there is the inevitable limit of not being patient specific. Therefore, a better technique would be to use pluripotent stem cells from the patient themselves. As mentioned, prior work showed that combining somatic cells with the contents of embryonic cells or oocytes resulted in their "reprogramming," suggesting some genetic contents might be capable of achieving the same for patients.

Takahashi & Yamanaka selected 24 gene candidates they hypothesized may induce pluripotency from

[248] https://pubmed.ncbi.nlm.nih.gov/16904174/

mouse embryonic fibroblasts (MEFs). MEFs still are not fully matured, but one has to start somewhere! Genes were inserted into MEFs via retroviral transfection. To determine if pluripotency had been achieved, they used a toxin and inserted a resistance conferring gene into the *Fbx15* gene. This gene is expressed specifically in embryonic stages, but is not necessary for development—therefore, it provides a good marker for induced pluripotency. Notably, no single factor from their list of 24 was sufficient to confer resistance (meaning Fbx15 was never expressed, and cells never reached pluripotency). All factors combined were able to, and generated a number of resistant lines. These lines showed embryonic cell-like proliferative properties and divided in a similar manner. Similarly, some promoters for genes expressed in the embryonic stages became unmethylated.

They narrowed down the lowest number of necessary factors called OSKM and or the Yamanaka factors: Oct3/4, Sox2, Klf4, and Myc. They performed expression analysis via DNA microarrays in order to demonstrate the transcriptional similarity between ES and iPSCs. While they were clearly not identical, iPSCs were markedly more similar to ES cells than to MEFs. Such factors are now key players in inducing pluripotency, which is used to derive patient-specific, or mutation-specific, cell lines of all sorts.

Antibodies.

Antibodies are often developed against things like Aβ with the hope of either conferring "immunity" or acting as a protein sink which will remove amyloid from the brain and pass it into the bloodstream. Adding in antibodies can either confer active or passive immunity. Simply implanting the antibodies will be passive, but other groups are working to immunize one against Aβ in order to kick-start an active immune response. Famous trials like Aducanumab were examples of attempts for passive immunity, but all trials have failed so far, either showing no promise or results with considerable side effects. Current trials for prompting active immunity are ongoing, but still none have succeeded or show real promise. Many open questions exist, such as if treating patients later in disease is even possible.

Antisense Oligonucleotides.

Antisense oligonucleotides (ASO) are 20 or so bp sequences of RNA which can hybridize with disease causing genes and target them for degradation. ASOs can also be used to alter splicing or to hybridize with miRNAs that block transcription, which can help boost transcription of your intended gene. ASOs cannot pass through the blood brain barrier, meaning targeting them to the CNS requires injections (and reinjections) into the CSF. Notably, once injected into the CSF they do a good job spreading throughout the entire CNS—an important feature not seen in all therapeutics.

Nusinersen (also called Spinraza), developed by Ionis with a trial ran by Biogen, is a very powerful, success story for treating spinal muscular atrophy (SMA). SMA occurs due to the loss of the *SMN-1* gene and its onset begins in childhood. The deletion of one exon from *SMN-1* is quite common, conferring disease. Notably, a paralog for SMN-1 exists, called SMN-2. In the wildtype case, SMN-2 is largely non-functional due to the exclusion of a an exon of its own. Therefore, in affected individuals, SMN-1 is nonfunctional, but there is the potential of hijacking this "dormant" SMN-2 gene via including this missing exon. That is, indeed, exactly the goal of this ASO. Therefore, even patients who are null for SMN-1 can recover via SMN-2 hijacking.

Another example by Ionis and Biogen is Tofersen, an ASO developed for knockdown against SOD1

in ALS. Indeed, SOD1 was reduced. However, ALS motor scores were not significantly different after the first trial. This is an often observed problem, where therapeutics have difficult to reach, "clinically substantial" goals, often on a very short time course. In some instances, like this one, continued experiments showed that there is indeed some improvement beyond the original studies boundaries. Indeed, after open-label extension, in trials which lasted around a year, a benefit was found for this ASO. Now, ASOs to SOD1 have been approved by the FDA.

An ASO to Huntingtin's was also developed called Tominersen, which got to a Phase 3 trials. However, these trials were halted because patients were getting worse. Recently, Gillian Bates, who was the original creator of the R6 mouse, determined that it is specifically exon 1 which confers disease. It turns out the polyQ expanded alleles generated buildup of toxic protein specifically from exon 1. The ASO was targeted elsewhere in the htt transcript. It's thought that the full length protein can potentially be protective, and therefore an ASO directed to only exon 1 could potentially be successful. It has not been tried yet, but potentially will be helpful in the future.

Concluding Remarks

This effectively concludes our discussion of neurodegenerative diseases. It seems we are leagues away from having all questions answered. Indeed, some of the most fundamental concepts, like the true role of neurons, and the importance of aggregates, are still not known. Let us hope that in the coming years, landmark studies help clear up these misconceptions so we can get to the heart of disease. Perhaps these landmark studies will be done by you. Best of luck.