EASY MATHS
FOR
GRADUATE PROTEIN BIOCHEMISTRY

Nicholas Ostan

Easy Maths for Graduate Protein Biochemistry

Fundamentals for the Mathematically Uninclined; Myself Included

FIRST EDITION

Nicholas Ostan

This resource was written as a handbook and learning resource for the University of Toronto Advanced Biochemistry Lab (BCH478H).

A sincere thank you to the contributors who took the time to review the accuracy of this material. They are listed below.

- **Dr. Trevor F. Moraes**
- **Dr. Anastassia Pogoutse**

Table of Contents

Module Schedule

WEEK 1	Tues, Oct. 15 3pm – 5pm	--	**Module 1 Term Test**
	Fri, Oct 18 10am – 5pm	MSB2384	**Introduction** **SDS-PAGE/Quantification**
WEEK 2	Tues, Oct. 22 3pm – 5pm	MSB2384	**Kinetics Lecture** **Preparations for BLI**
	Fri, Oct. 25 10am – 5pm	MSB2384	**Biolayer Interferometry Experiment**
WEEK 3	Tues, Oct. 29 3pm – 5pm	MSB2377	**BLI Data Processing & Analysis**
	Fri, Nov 1 10am – 5pm	MSB2377	**Sparse Matrix Crystallization** **Lysozyme Crystallization**
WEEK 4	Tues, Nov. 12 3pm – 5pm	MSB2377	**Crystallography Lecture 1** **Crystal Observations**
	Fri, Nov. 15 10am – 5pm	MSB2377	**Crystallography Lecture 2** **Crystal Optimizations**
WEEK 5	Tues, Nov. 19 3pm – 5pm	MSB2377	**Crystallography Lecture 3**
	Fri, Nov. 22 10am – 5pm	MSB2377	**Solve Crystal Structure** **Build Models**
WEEK 6	Tues, Nov 26 3pm – 5pm	MSB2377	**Crystallography Lecture 4** **Electron Microscopy Tour (?)**
	Fri, Nov 29 10am – 5pm	MSB5231	**Discussion Session** **Presentations**
EXAM	**Tues Dec. 3** **3pm – 5pm**	**CCBR Black Room**	**Term Test**

Preface

"*To those who do not know mathematics it is difficult to get across a real feeling as to the beauty, the deepest beauty, of nature. If you want to learn about nature, to appreciate nature, it is necessary to understand the language that she speaks in. She offers her information only in one form; we are not so unhumble as to demand that she change before we pay any attention*."

Richard Feynman, 1967

Conquering Your Fear of Math

As somebody who used to be terrified of mathematical notation, trigonometric functions, and the like, I know what it feels like when somebody puts up an equation on the board that looks intimidating:

$$\rho(x,y,z) = \frac{1}{V}\sum_{hkl}|F_{hkl}|e^{-2\pi i(hx+ky+lz)+i\alpha(hkl)}$$

In the same way I think it is much more effective to learn something by starting from—and truly understanding—the fundamentals, it is likewise not effective to throw entire equations at you without first building up what mathematical expressions are really saying. When we break down mathematical expressions, it is actually quite comical how elementary humans are in their thought processes. If you read through this guide, you will be required to understand mathematical expressions; especially the one above, but we want to give you the tools to do so. We promise that by the end of this text, you will understand the above expression in its entirety.

We are going to assume you have a very basic mathematical background (early high school level), and try to present you with the foundations of calculus through a more explanatory and intuitive approach. You will not be required to carry out algebraic processes like differentiation and integration, but rather I want you to understand what the crystallographic equations really mean by understanding conceptually how these processes work. Though, there is no better way to get a good grasp of the material by practicing calculus and linear algebra problems.

If you want to truly understand crystallography, sorry – you're going to have to learn the math. Having the desire to learn a subject but not being willing to learn the language is ***lazy!*** It is akin to wanting to study ancient Greek tablets and historical pieces; but refusing to learn the ancient Greek Indo European language. Sure, you can get your friend to translate things for you, but at this point you are just memorizing re-stated facts (and worse, taking their word for it!). I will try my best to do a lot of the 'translating', but it is up to you to connect it with the math I present.

Expressions in one language never translate perfectly to another language because there seems to always be a beauty present in the context. The same is true for math. If you try to 'skim' the material for a pure intuitive understanding of crystallography, then when your committee inevitably confronts you with a crystallography question, you will be forced to scrounge and dig around in the 'box of facts' that you have memorized. "Hopefully it is in here somewhere!" you'll think. If rather you take the time to understand things mathematically, it has great reward. This approach provides a glimpse into the workings of nature, and will allow you to answer questions by deriving from first principles.

Hopefully I haven't scared you off already by calling you lazy. Give my teaching method a shot! I came up with a method to refresh your calculus for each of the scary-looking concepts like summation, integration, differentiation, and so on. We are going to pretend we have been hired by a blind genius who is asking us to transcribe his ideas that he says aloud to us, and we need to write them on paper for him. We will call our blind employer ***E. Hall***. Don't ask me why I named him E. Hall. In our conversations with E. Hall, we will learn that mathematics are not scary at all; instead, the expressions in mathematics are some of the most beautiful things to behold in the universe! I did not want to continue our conversations with E. Hall throughout the entire text because I felt it may be distracting. Thus, our conversations will be limited to the preface. Return at any time to review this section.

My favourite part about mathematics is that at its core, it uses very simple ideas. When these simple ideas are built on top of each other, we tend to get scary-looking equations, but they can always be distilled into more accessible 'chunks'.

Richard Feynman on mathematics and the character of physical law:

"Nature uses only the longest threads to weave her patterns, so each small piece of her fabric reveals the organization of the entire tapestry."

Summation

E. Hall was just at a carnival and he met a five year-old kid named Billy. Billy told him that every year on his birthday, his parents get him a cake that has twice as many candles in it as years he has lived. So, when he was 3, he had 6 candles, and so forth. E. Hall is wondering how many total candles this amounts to in Billy's lifetime, and he wonders how much money his parents must have spent on candles. We overheard this conversation between E. Hall and Billy, so we can ask some clarification questions.

Intuitively, I am sure you could calculate this in your head, but mathematicians like to use scary 'general expressions' that intimidate us simple folk. Let's explore this idea.

E. Hall: "I want to add a whole bunch of things together!"

$$\sum$$

You: "Okay how many things do you have that you want to add together?"

E. Hall: "Five things."

$$\sum^{5}$$

You: "Do you want to start adding from Billy's first birthday, or a later birthday?"

E. Hall: "His first birthday."

$$\sum_{1}^{5}$$

You: "What is the actual quantity you want to add for each of these years, from 1 to 5?"

E. Hall: "His age times two. This is how many candles he told me were in his cake."

$$\sum_{a=1}^{5} 2 \times a$$

You: "Alright, done. Anything else?"

E. Hall: "Multiply the number of candles by the price of a candle."

$$\sum_{a=1}^{5} 2 \times a \times p$$

You: "Done; what do you want to call the quantity?"

E. Hall: "I didn't like Billy. Call it money wasted."

$$M_{wasted} = \sum_{a=1}^{5} 2 \times a \times p$$

E. Hall: "Now ***evaluate*** the expression for me, assuming Billy is 5, and each candle costs $5.00."

$$M_{wasted} = (2 \times 1 \times 5) + (2 \times 2 \times 5) + (2 \times 3 \times 5) + (2 \times 4 \times 5) + (2 \times 5 \times 5)$$

$$M_{wasted} = 10 + 20 + 30 + 40 + 50 = \$150.00$$

When you find yourself confronted with a complicated mathematical equation, try to break it down into chunks. Deconstruct it, reverse engineer it, and understand what it is trying to *tell you*. The reason it is useful to have a general expression like the one above, is because now when Billy turns 21, we can change some of the values, and calculate how much money his

parents wasted again. When he moves to Dubai where the price of candles is $17.00 each, we can easily substitute this value in for p.

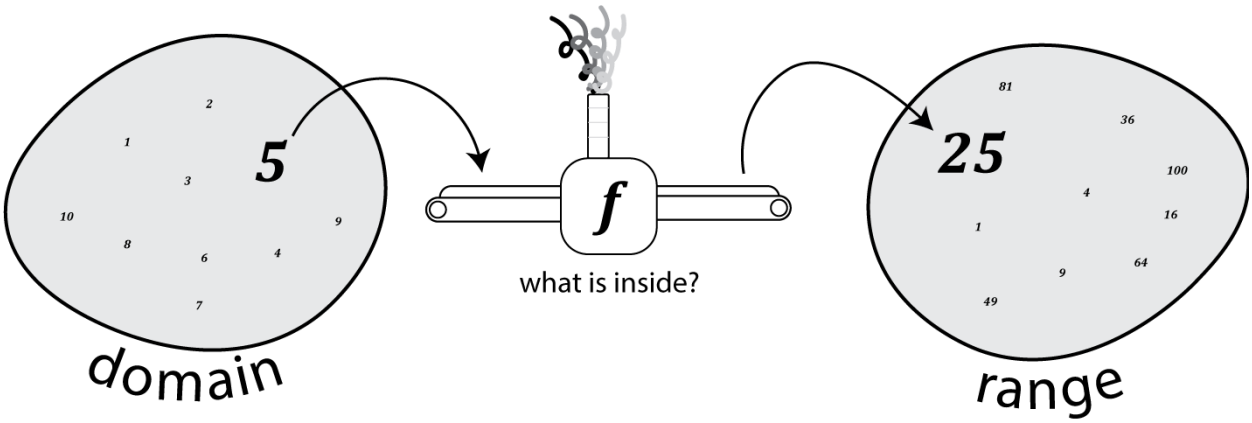
Functions

The idea of a ‘function’ can be daunting. I remember back to my first introduction to this nomenclature:

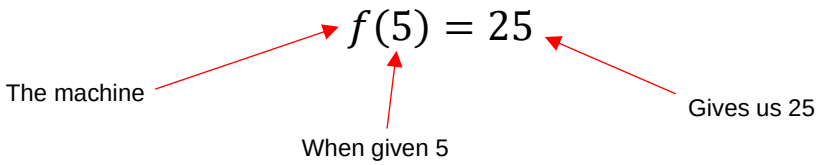
$$f(x) = \cdots$$

… I never really understood what it was trying to say.

A function can be thought of a machine that ‘eats something’ and then ‘spits something out’. In the image below, we have a function depicted with unmatched artistry as a machine with two conveyor belts and a smoke stack. The ‘domain’, depicted as the grey blob on the left represents all of the things our function is willing to eat. Let’s take the number 5, and put it on the conveyor belt to see what the machine spits out. We put it on, and it spits out 25. (The fact that the number 5 is in larger typeface than the rest of the numbers does not mean anything, I have just done this for clarity).



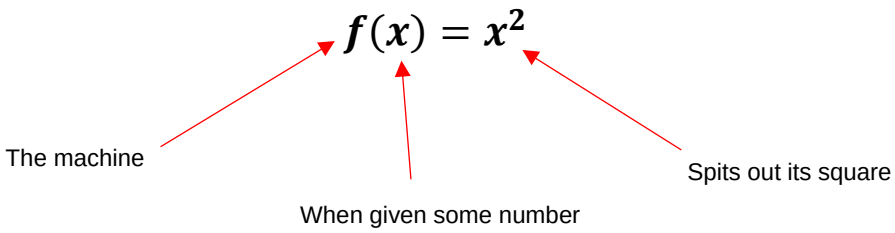
In this case, we could say the machine, when given 5, spits out 25. Or:



What did the function ***do*** to the number five in order to get 25? Well, maybe it added 20, since:

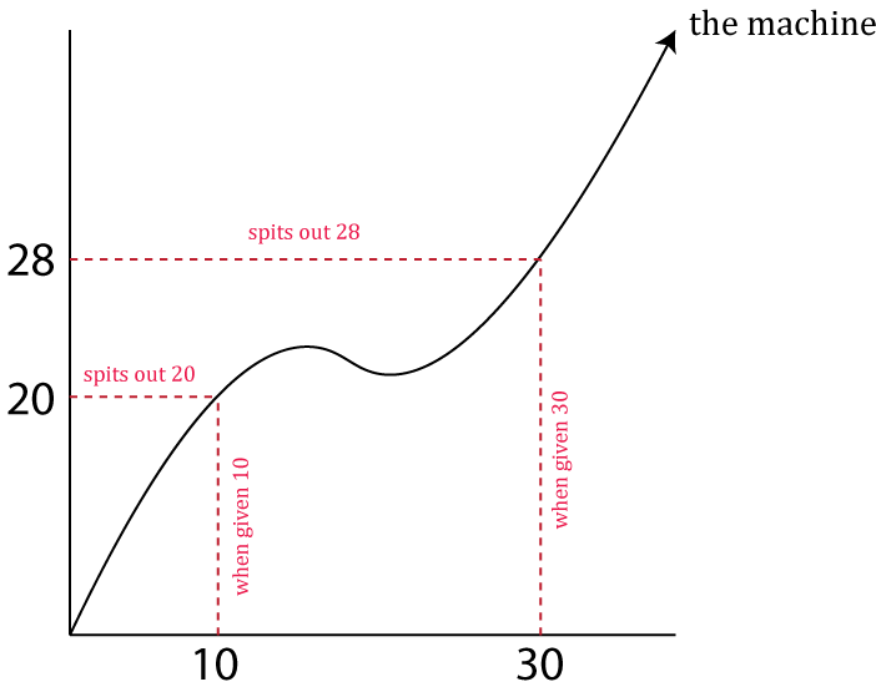
$$5 + 20 = 25$$

But if we look closely at all the numbers the function is capable of spitting out, also called the "range", again outlined with a grey blob to denote "the total space of all output values", we see that many of the numbers are perfect squares. Perhaps the function is squaring whatever we give it, and outputting the square. So, a general expression for this function would be:



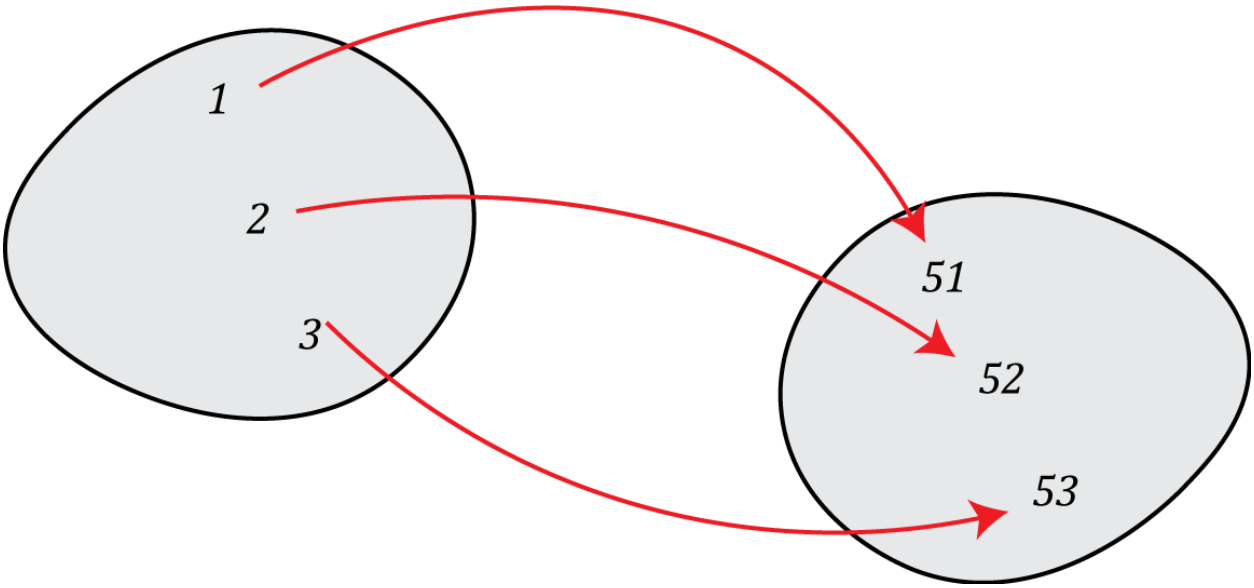
This is a very simple function; but functions can take on many forms. Whenever we introduce a new function in this text that looks a little bit different, we will try to break it down.

It is often very visually helpful to plot a function as a graph. Why is this the case? A graph lets us see all of the things a function will spit out for any value we feed it:



You may already know what a function is and want to skip this section, but I highly encourage you to continue reading and broaden the way you think about functions conceptually. On the graph, we only labelled inputs 10 and 30, but you could imagine drawing lines for any number of points in between, or outside.

A more general depiction of a function is something that simply maps information or values between two spaces, like this:



Which might be described as:

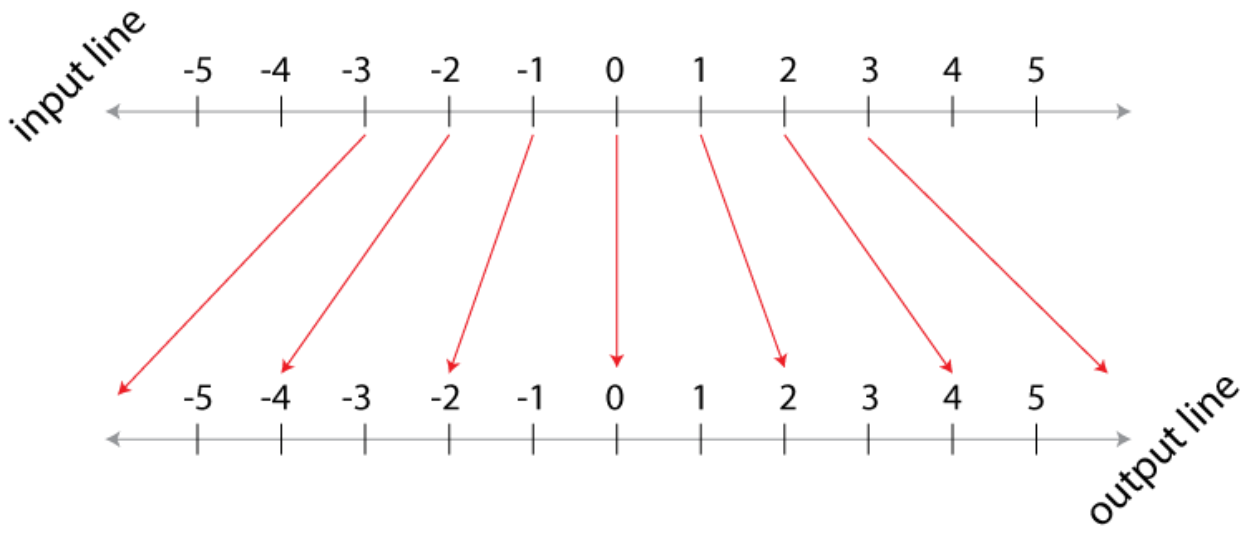
$$f(x) = x + 50$$

The machine

When given some number

Adds 50 to it

For values that take a single input and provide a single output, we can think of it as mapping between two number lines, like below. This is a very important idea and we will return to this idea in Chapter 4.



We might describe the function above as:

$$f(x) = 2x$$

The function

When given some number

Maps the input to a number twice as big on the output line.

OR

Stretches/scales the output number line by 2, keeping the origin position constant

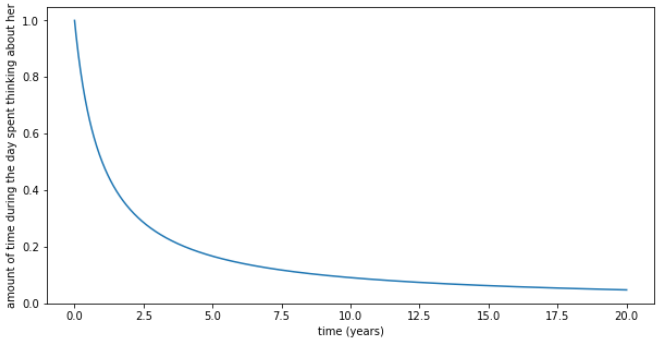
As functions get more complex, they start to take multiple inputs, and provide multiple outputs. In other words, the ***dimensionality*** becomes higher. Do not worry about this for now, we will come across these functions later with more context.

Limits

E. Hall has given us a function which he says represents how his love has changed for his wife over time. That function is:

$$f(x) = -\left(\frac{x}{x+1}\right) + 1$$

The quantity $f(x)$ is a measure of his love. The quantity x is time, measured in years. We can assume $x = 0$ was the day they got married. Let's plot this function.



E. Hall: "How much did I love my wife on our wedding day?"

You: "You spent your whole day thinking about her!"

E. Hall: "What about after 5 years?"

You: "At that point, you only thought about her for about 20% of the day."

E. Hall: "We have been married for 20 years, how much do I love her now?"

You: "It seems like you barely think about her at all anymore."

E. Hall: "Calculate for me how much I will love her on our 50th anniversary"

$$f(50) = -\left(\frac{50}{50+1}\right) + 1 = 0.019$$

You: "Seems like you will think about her for 0.019 of a day… which is about 27 minutes."

E. Hall: "I guess I'm a bit of a romantic. I want to write a love letter to my wife to tell her how much I'll love her after 1000 years of marriage."

$$f(1000) = -\left(\frac{1000}{1000+1}\right) + 1 = 9.99 \times 10^{-4}$$

You: "That comes out to about 1.43 minutes of the day, sir."

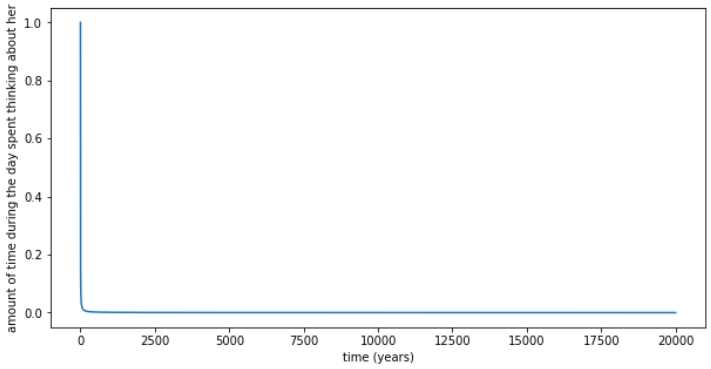
E. Hall: "Can you tell me if I will ever stop loving my wife!?"

You: "I don't know… it seems like as time goes on, the numbers get smaller and smaller, but never actually amount to zero."

E. Hall: "Well then take the limit as time approaches infinity!"

$$\lim_{x\to\infty} f(x) = -\left(\frac{x}{x+1}\right) + 1 = 0$$

We can't plug infinity into our calculator, so let's plot what the function looks like after a huge amount of time. Let's say after 20,000 years.



It appears that as time extends to infinity, we approach closer and closer to zero. The limit of our function, as x approaches infinity, is zero.

E. Hall begins to write…

Dear Susan,

On the day we were married, I thought about you all day long. Five years later, I still thought about you for 20% of my day, even when you weren't around. Today, on our 20th anniversary, I still think about you for half an hour per day! In a thousand years, I will still think about you for at least 1.43 minutes. The only way I will ever stop loving you is if time extends to infinity.

Your romantic husband,

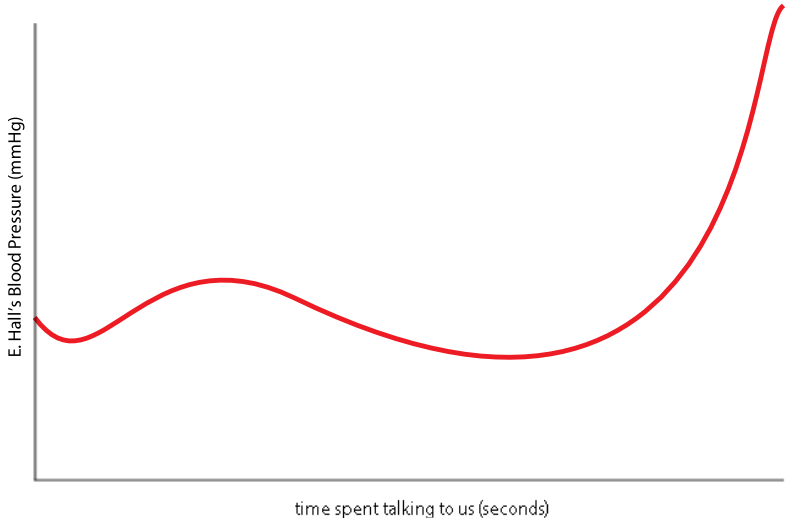
-E

Limits are central to the processes of calculus. We often sum together a huge number of 'infinitesimally' smaller, more manageable quantities to calculate some larger quantity. This is the idea behind integration, which we will explore soon. We will introduce the practical use of a limit with regards to differentiation, as it is probably easier to digest.

Differentiation

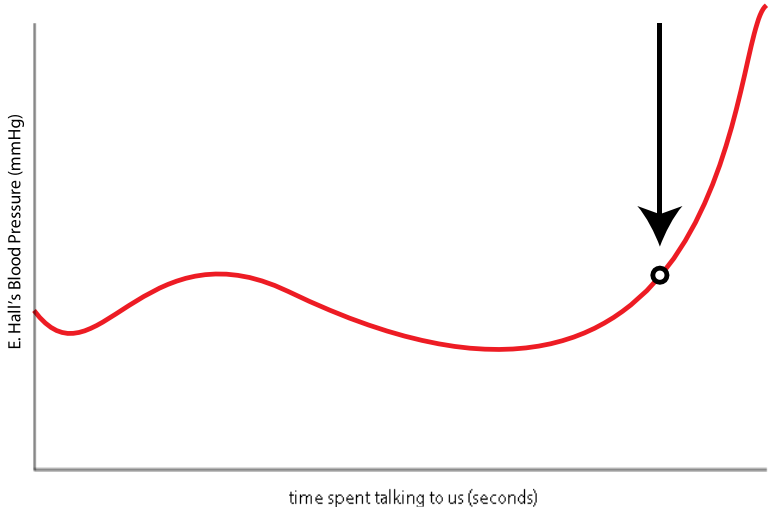
E. Hall is a very irritable person, and he is quite vocal about it. He lets us know constantly that he doesn't like talking to us. To prove his point, he took measurements of his blood pressure as he carried out a conversation with us. At one point, we began talking about our hobbies and family, which E. Hall didn't particularly care about. To prove a point, he wants us to compute his blood pressure at the moment we turned the conversation to ourselves.

E. Hall: "Plot my blood pressure over time for me."



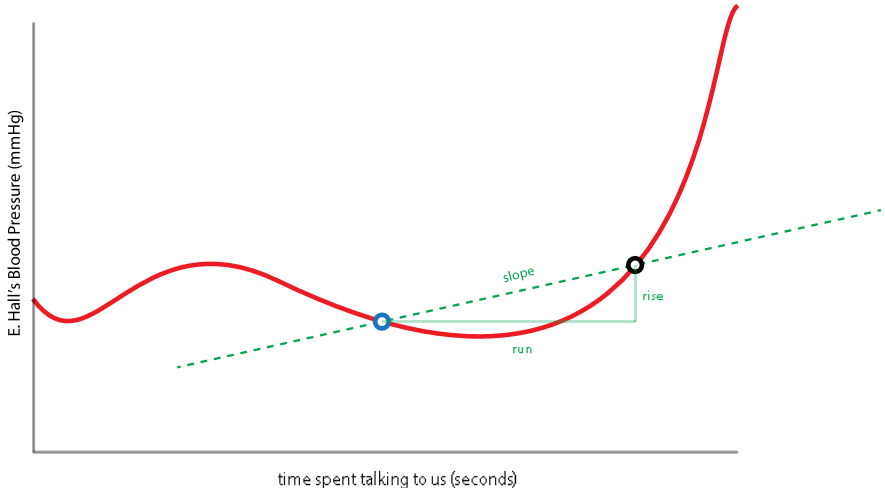
You: "Okay, now what?"

E. Hall: "Mark the time where you selfishly started talking about yourself."



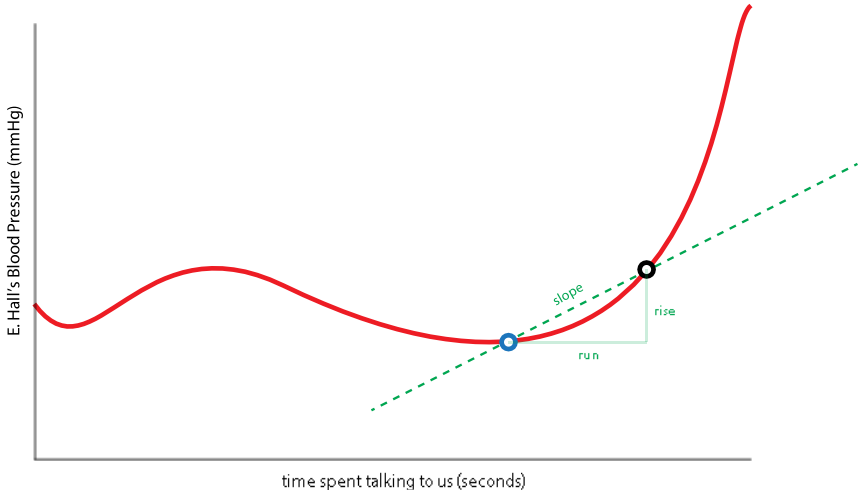
You: "Done. What next?"

E. Hall: "I want to know how much my blood pressure was increasing near this time."



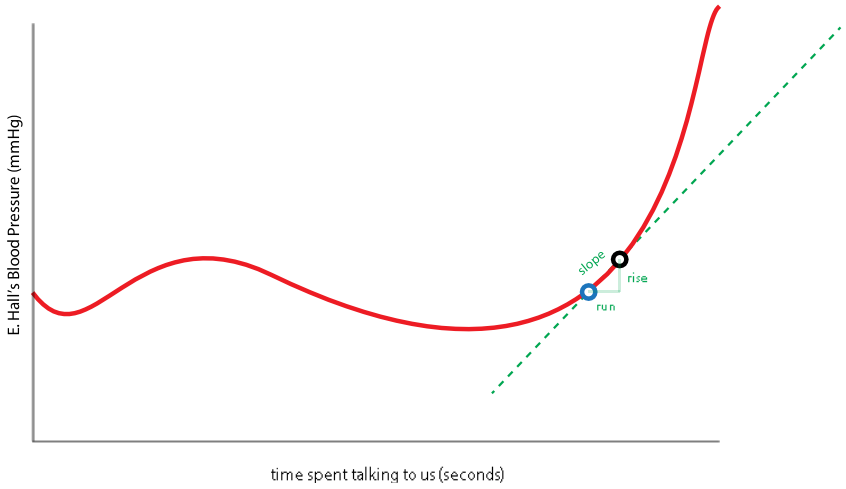
You: "Okay, I took a point that was 20 seconds before (blue point) and found the slope to when I mentioned myself."

E. Hall: "You idiot! 20 seconds is way too far back. Go closer!"



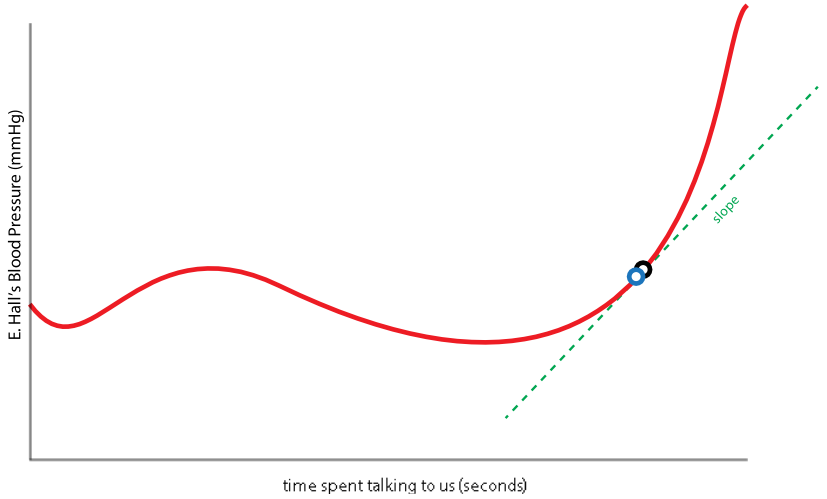
You: "Okay – I did it again but this time 10 seconds sooner."

E. Hall: "Idiot! Still too soon! Go closer to the point!"



You: "How about 2 seconds?"

E. Hall: "Idiot! Still too soon! Go closer to the point!"



Finally, by taking an adjacent point close enough, we are able to compute the slope ***at*** the point E. Hall wanted. We again notice this theme of allowing some quantity to approach closer and closer to another quantity, but not quite reach it. We have come across a limit again. The slope here represents the instantaneous rate of change (which is arguably a flawed statement, because change does not occur in an instant – it occurs over some duration of time). We have just made our duration of time so small that it is 'perceivably' instant to the human time reference frame).

E. Hall's blood pressure the MOMENT we started to speak

E. Hall's blood pressure JUST before we started speaking

$$the\ slope\ at\ point\ \boldsymbol{a} = f'(a) = \lim_{h\to 0} \frac{f(a)-f(a-h)}{a-(a-h)}$$

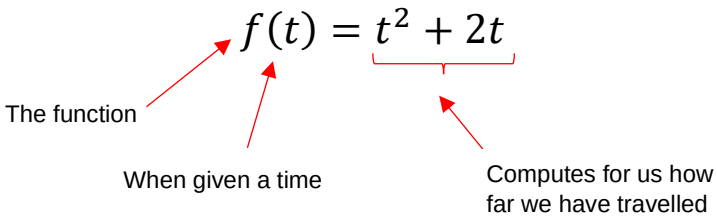
The time at which we started talking about our self

The moment just before we started talking about ourselves

The Power Rule

The only reason I want to introduce a method for differentiating a function (aka taking a function's derivative) is so that you can realize that when a function is differentiated, you obtain another function that describes the rate of change of the original function. The simplest way to illustrate this is with the classic position vs time graph, and its related velocity vs time graph.

Let's say we have a function describing our position in a car (in meters) along some road:



So if we take our position at $t = 0$;

$$f(0) = 0^2 + 2(0)$$

$$f(0) = 0m$$

Or if we take our position at $t = 10$;

$$f(10) = (10)^2 + 2(10)$$

$$f(10) = 100m + 20m$$

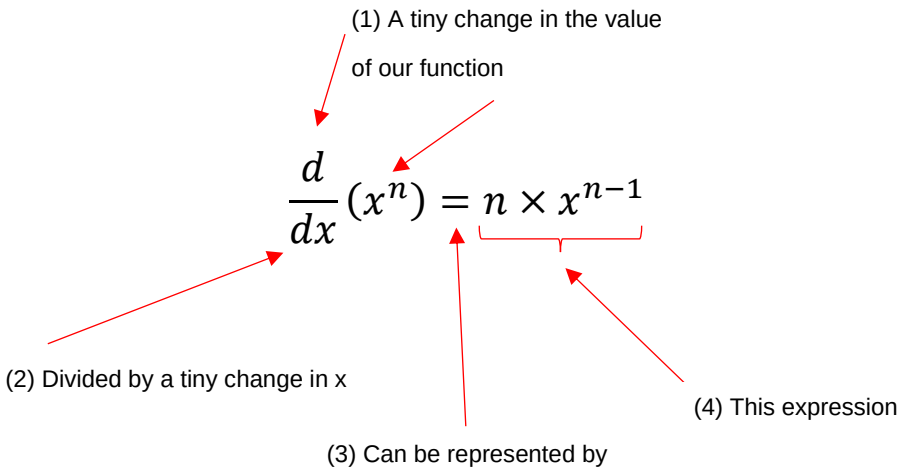
$$= 120m$$

We found in our last example, when talking about differentiation, that by taking a closeby point and finding the slope between the two points, we can find the 'instantaneous' rate of

change. Instead of having to compute the slope manually every time, we can have an expression, called the derivative, do it for us. The power rule states:

$$\frac{d}{dx}(x^n) = n \times x^{n-1}$$

This is notation we have not seen before. Let's write it out again so we can deconstruct it.



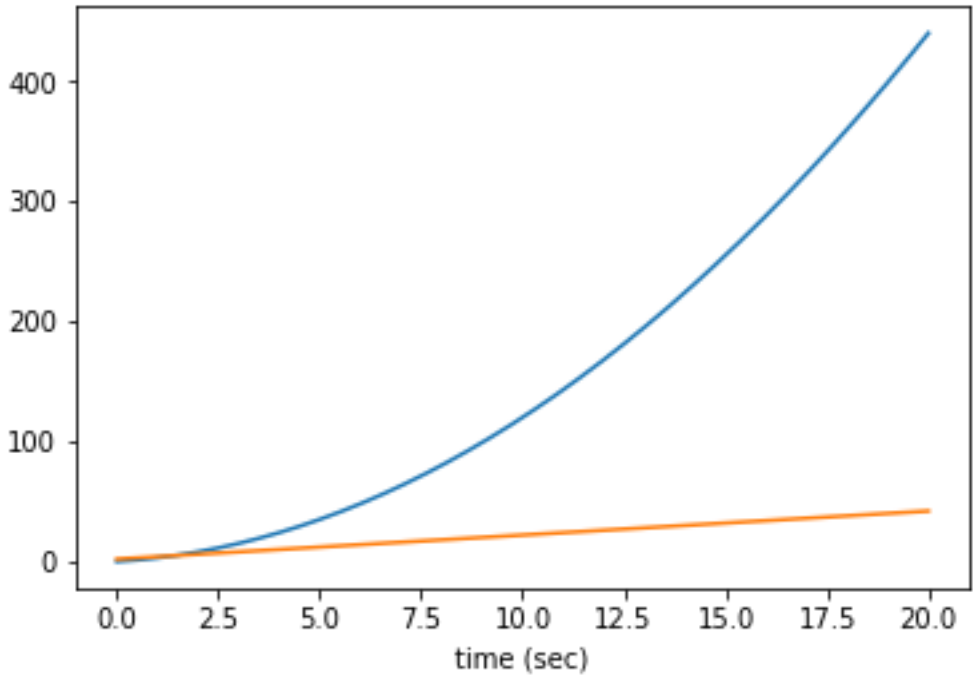
The d in this case really just corresponds to 'an infinitesimally small change'.

Let's carry out the power rule on our position vs. time function;

$$\frac{d}{dx}f(x) = 2t + 2$$

This really just says "the derivative of our function is $2t + 2$". What is interesting about the derivative of our function is that it represents our velocity at any point t. Just by looking at the expression, we can tell that we are accelerating, because as we substitute in larger and larger values for t, we will always obtain a greater velocity. In other words, our speed is always increasing, implying acceleration. We could take the derivative once again to obtain the acceleration function, but this is not necessary. The main concept here is that integration

(introduced next) is the inverse operation of differentiation. If we are given the velocity function, for example, we can obtain the position function by integration.

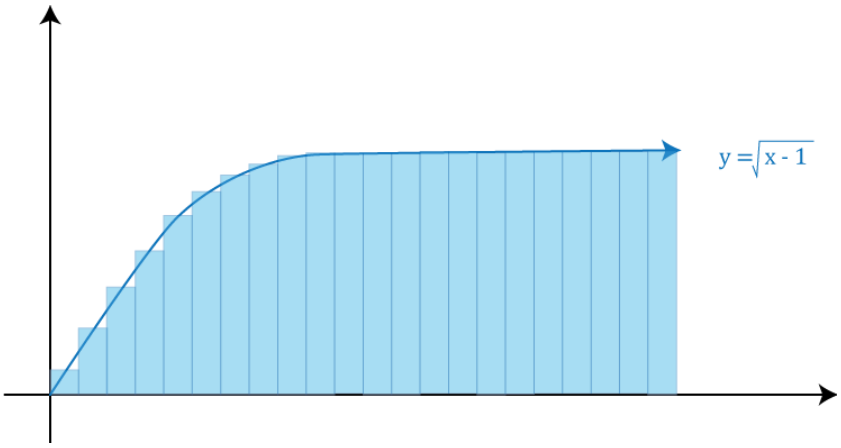


Here, the position vs. time graph is in blue, and the velocity vs. time graph in orange. You can know the velocity of the vehicle at any time by consulting the y-value on the orange line at that same time.

Integration

The ***fundamental theorem of calculus*** states that differentiation and integration are inverse processes. In the previous section we found the derivative of a simple function using the power rule – arguably one of the simplest differentiation techniques in calculus. If we wanted to undo our differentiation process, and re-obtain the original function, we can use integration. The interesting thing about integrals (specifically definite integrals) is that they allow us to find the area under the curve. That is to say, if we can generate the derivative by measuring the slope at extremely tiny intervals on some function, we can generate the anti-derivative by measuring the area under graph at the same extremely tiny intervals. In physics and crystallography, integration helps us understand some quantity as a whole, as it is built up of its smaller components.

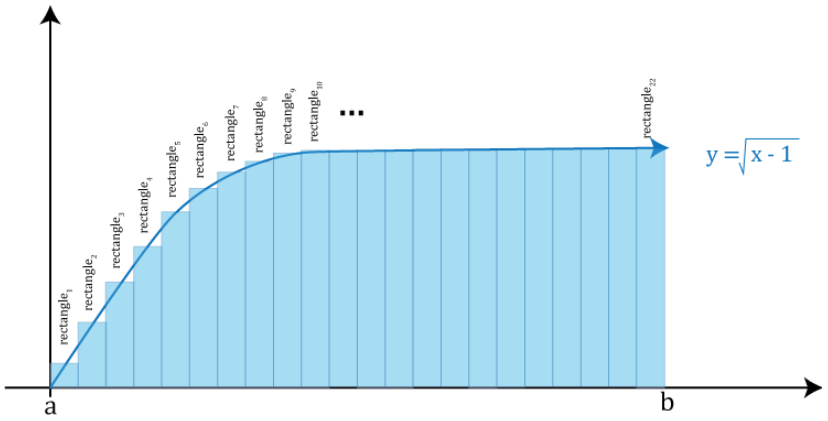
Let's say we have some arbitrary function:



If we were tasked with finding the area under the curve, we could break the area down into small rectangles. We of course know how to compute the area of a rectangle – that is easy:

$$area_{rectangle} = base \times height$$

But how can we get the values for the base and height of each triangle? Well, the base is simply the base width that we choose. I took a certain interval between **a** and **b**, and divided it into 22 triangles. Like below:



So, we know that the width of each rectangle must be:

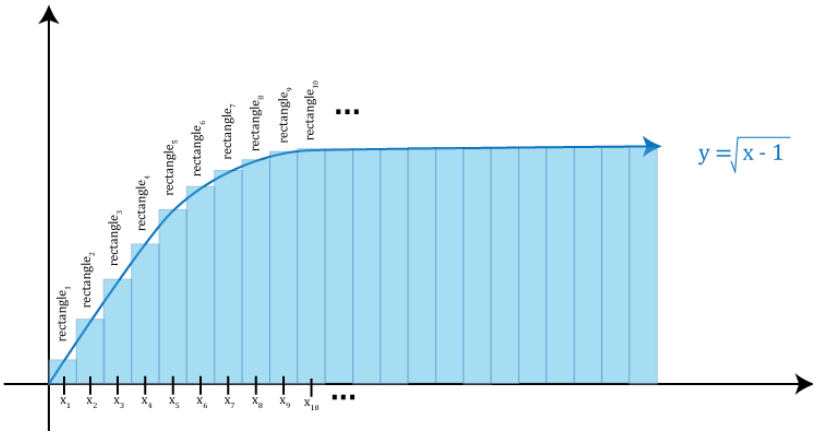
$$\Delta x = \frac{(b-a)}{22}$$

We can see that a wider base will likely give a more inaccurate approximation of the area. As the rectangles get *narrower and narrower (sound familiar? Think limits…)*, they will more closely approximate the curve. Since we are of course allowed to choose any number of rectangles, we can make a general expression:

$$\Delta x = \frac{(b-a)}{n}$$

Where n is the number of subdivisions (rectangles) we want to use.

The height of the rectangle is easy to express; it is simply the value of the function at x. We can of course see that in general, every rectangle will have a different area, so we need to treat each rectangle separately. Let's call our first rectangle "rectangle 1", our second rectangle "rectangle 2", etc. The rectangle that we are talking about is defined by which x-coordinate we are talking about. To illustrate this:



In doing so, we can come up with a general expression for any rectangle; lets call it $rectangle_i$.

$$area_{rectangle_i} = \underbrace{\Delta x_i}_{\text{base}} \times \underbrace{f(x_i)}_{\text{height}}$$

I'm going to switch around these two terms:

$$area_{rectangle_i} = f(x_i) \times \Delta x_i$$

Okay now lets sum up all of the rectangles:

$$area\ under\ graph = \sum_{i=1}^{n} f(x_i) \times \Delta x_i$$

Now, remember back to the idea of a limit. If we let n get bigger and bigger, this means we will have more and more rectangles approximating our area. As the rectangles get narrower, there are smaller gaps between the function and the rectangle (white space):

$$area\ under\ graph = \lim_{n \to \infty} \sum_{i=1}^{n} f(x_i) \times \Delta x_i$$

$$\int_a^b f(x)dx = \lim_{n\to\infty} \sum_{i=1}^{n} f(x_i)\Delta x$$

The expression on the left is called a definite integral. All it means is:

"Create a bunch of rectangles by multiplying an infinitesimally small width (dx) by the height of the function above the x-axis ($f(x)$), and do this between the values a and b; then all 'em all together!"

It should now be obvious that differentiation and integration are inverse processes. In differentiation, we are dividing by a tiny change in x (how many meters did I move ***per*** second?). In integration, we are multiplying by a tiny change in x (if I was moving at 5 meters per second, how far did I go in **some number** of seconds?). This is of course represented by the area under the curve. You'll notice the integral sign looks like an ***S***. This is because the symbol was derived from the meaning of '***S***ummation' (to add a bunch of things).

Trigonometric Functions[1]

The trigonometric functions are incredibly interesting functions. When I start wondering about the meaning of life, I often find that so much of our universe is related to triangles and circles. Trigonometric functions are useful to us for a couple reasons; they allow us to break down vectors into their components, which makes math and geometry much easier. They also model electromagnetic radiation (and other harmonic phenomena), and since this book is about X-ray crystallography, you can imagine why this is useful. If you're reading this, you have undoubtedly had to use trigonometric functions in your math classes. At their most basic level, they relate the angle of a right-angled triangle to the ratio of the triangle's side lengths.

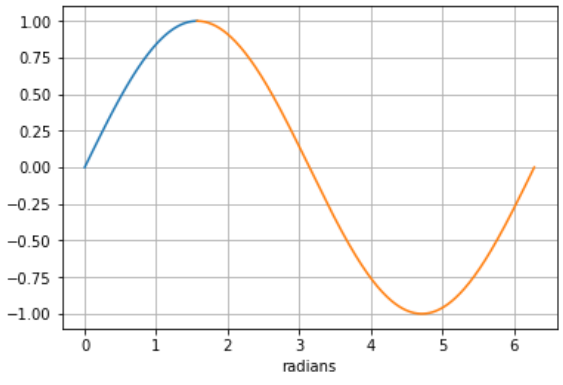
$$\sin(\theta) = \frac{opposite}{hypotenuse}$$

The sine machine

When given an angle of a right-angle triangle

Spits out the length ratio of the opposite side to the hypotenuse

If you have a calculator handy, you can try this for any angle you like. Let's think about a triangle like the one below. We will slowly squish it inward so that the angle increases, and then we will compute the sine of that angle.



But we have a problem. What does it mean to take the sine of an angle greater than 90 degrees? Right angle triangles never have such angles, so when we think of a traditional sine curve like the one above, how is it defined for angles between 90^o and 360^o? Each function is defined on the unit circle. As we move a point around the unit circle in a counter-clockwise fashion, the triangle drawn by the point and the two axes (keeping the origin at 90^o) is what we use to define the function. So, this gives us 4 triangles (one for each quadrant), amounting to 360^o.

Trigonometric functions are not quite like our other functions that we have looked at so far where we could see the internal mechanisms of the machine, and understand that they squared the number, or multiplied the number by two, or added fifty, etc. What I mean by this is: when you type $\sin(90)$ into your calculator and it is spits out 1, what did the calculator do to the number 90 to make it 1? How did it map the number 90 to the number 1? The calculator must be performing some mathematical operation on the number because it would be unreasonable to think that the calculator had the infinitely populated real number line populated with a map to another number line. The answer is the sine function which, as with other trigonometric functions, is approximated using something called a Taylor series. While this isn't important for our understanding of crystallography, I wanted to clarify that trigonometric functions aren't quite the same as other functions we have seen so far.

The other trigonometric functions that we care about are listed below. You will not be expected to do much (if any) algebraic manipulation with these functions; they are important to understand for the purpose of geometric break-downs.

$$\sin(\theta) = \frac{opposite}{hypotenuse} \qquad \sin^{-1}(x) = \arcsin(x)$$

$$\cos(\theta) = \frac{adjacent}{hypotenuse} \qquad \cos^{-1}(x) = \arccos(x)$$

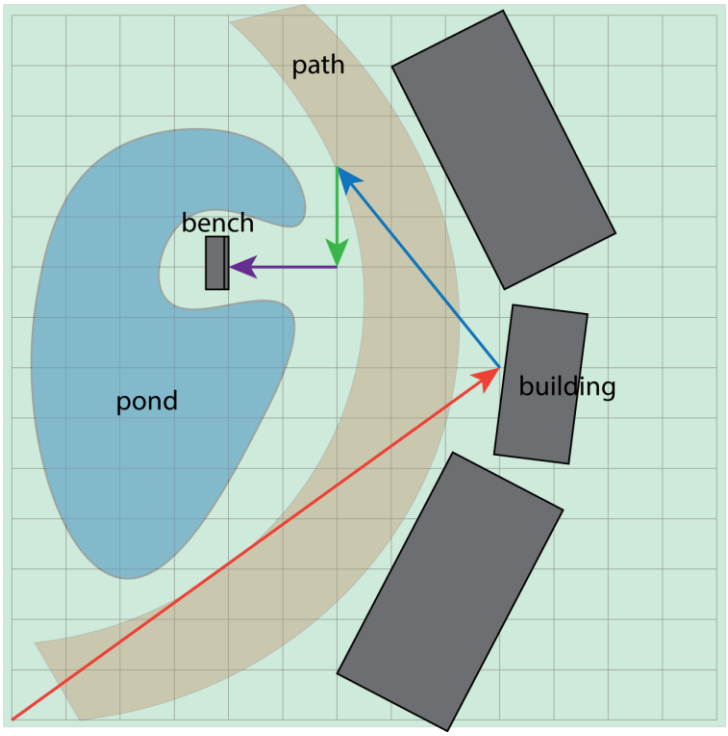
$$\tan(\theta) = \frac{opposite}{adjacent} \qquad \tan^{-1}(x) = \arctan(x)$$

Vectors

E. Hall likes to occasionally go for a stroll in the park. He doesn't go often though, so he doesn't have his route memorized. His favourite part of his walk is sitting on a bench located at a small nook near a lake. Today, on his way there, he ran into a building (remember, he is blind), then walked past the nook, and had to backtrack before finally arriving. At each part of his journey, he counted his paces, and he has given them to us so we can plot his displacement on a map of the park.

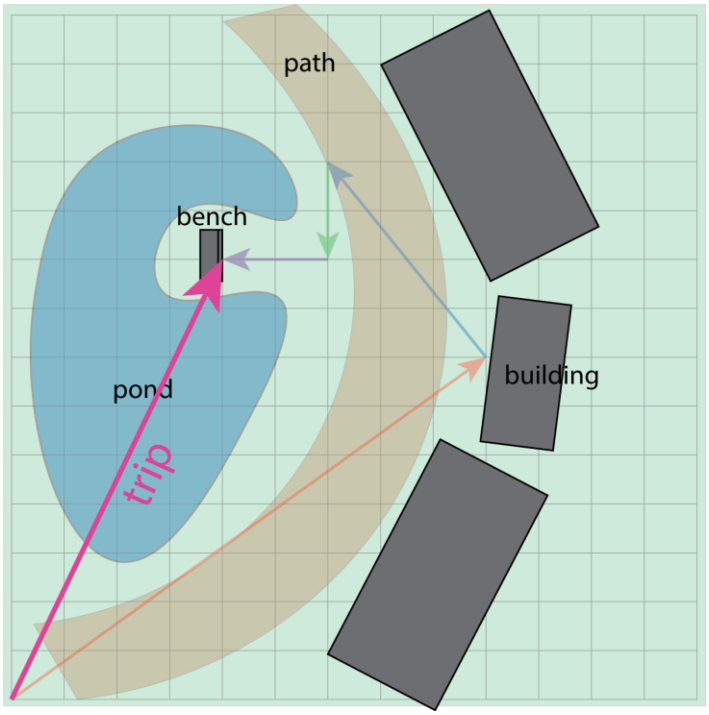
E. Hall: "First I walked 9 paces east, and 7 paces north. That's where I ran into the building. Then, 3 paces west, and 4 north. A passerby told me I had gone too far, so I walked two paces south. Then, two paces west brought me to the bench. Make these into vectors for me!"

$$trip = \begin{bmatrix} 9 \\ 7 \end{bmatrix} + \begin{bmatrix} -3 \\ 4 \end{bmatrix} + \begin{bmatrix} 0 \\ -2 \end{bmatrix} + \begin{bmatrix} -2 \\ 0 \end{bmatrix}$$



You: "Okay, I have your trip plotted on the map. What would you like to know?"

E. Hall: "I'm tired of getting lost all the time. What is my shortest path to the bench?"



You: "Well, the trip you made could be reached if you just walked 4 paces east, and 9 paces north.

E. Hall: "Wonderful. Thank you. Finally, you're worth something to me!"

You: "No problem!"

E. Hall asked for the shortest path. Not the driest path! This is a book on crystallography, not morals, so we are okay to do this. How can we understand this vector mathematically? The resultant trip vector can be obtained by summing the components of each of the vectors above.

$$trip_x = 9 + (-3) + 0 + (-2) = 4$$

$$trip_y = 7 + 4 + (-2) + 0 = 9$$

We will not aim to have a rigorous definition of vectors, just one that is good enough for you to understand scattering diagrams. You can think of a vector as an ordered list of numbers. For our purpose, we will only deal with real numbers. A vector looks like:

$$\begin{bmatrix} 3 \\ 1 \end{bmatrix}$$

This vector has two numbers, or components in it. The first component is 3. The second component is 1. The dimension of a vector corresponds to how many components it has. The vectors we have been dealing with so far are two-dimensional vectors. We can also write a 3-dimensional vector like this:

$$\begin{bmatrix} 2 \\ 4 \\ 5 \end{bmatrix}$$

While it is useful to understand and visualize three-dimensional vectors, most explanations in this book will be carried out in two dimensions.

When we add two vectors, we simply add the components. Geometrically, this corresponds to aligning the vectors “tip to tail” in the cartesian plane, drawing a vector from the origin to the resultant vector, and then defining its components. With vector addition, it doesn’t matter which order you add them; you will always get the same resultant vector (*we will learn later this is called commutativity*).

We can also subtract two vectors. We do this geometrically by placing the vector tails on the same point, and then drawing a line from the tip of vector one to the tip of vector two. The resultant line is our vector. **This is an important idea with scattering vectors, later in Chapter 4.** Mathematically, we just subtract the components. Vector subtraction is not commutative, as we will see later. This means the *order* in which you subtract one from another matters!

Unfortunately, we won’t be doing a lot of ‘easy’ vector math where we are adding and subtracting real numbers from the components of vectors. It will generally be more abstract than that.

Dot Product

In the previous section we added vectors together. It is also possible for vectors to participate in multiplication operations. This means that if I have a vector, and want to multiply it by vector of dimensionality 1 (also called a scalar, or as you probably think – a 'regular' number), the vector will be scaled by that amount. We simply multiply each component by the scalar.

$$3 \cdot \begin{bmatrix} 5 \\ 7 \end{bmatrix} = \begin{bmatrix} 15 \\ 21 \end{bmatrix}$$

You can imagine the consequences this has on the vector. The ratio between the components (5/7) and (15/21) is the same value before and after; meaning the vector points in the same direction. It has just been scaled up by 3 times.

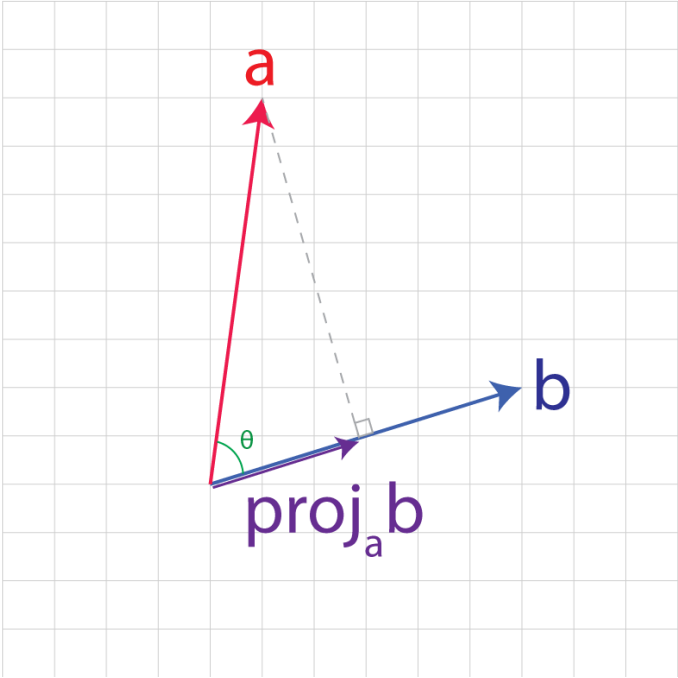
But what about multiplying a vector by another vector? Is this possible? Sure it is, and there are two major processes for doing so, namely the ***dot product*** and the ***cross product***. Both are critical for crystallography, but we won't be doing much lattice geometry in this book, so we can place the cross product on the back burner. Let's look at how to take a dot product, and then look at what the quantity it outputs actually means.

In its most basic definition, the dot product between two vectors is taken by multiplying each respective component, and then adding the result. This means the output is a single number, or scalar. So we can take any two vectors, take their dot product, and obtain a 1-dimensional value:

$$\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \cdot \begin{bmatrix} 3 \\ 4 \\ 5 \end{bmatrix} = (1 \cdot 3) + (2 \cdot 4) + (3 \cdot 5) = 3 + 8 + 15 = \mathbf{28}$$

It seems hard to understand how this quantity could be a useful value? What does the number 28 have to do with the relationship between these two vectors? For our purposes, we only care about the geometric interpretation of the dot product. The answer is: the dot product is a way of viewing a projection of one vector along another vector's direction. So, taking a dot product is

equivalent to finding the component of a vector along another one, and then scaling it by the vector you projected onto. Let me draw this out:



The proof of this has to do with the law of cosines, which I will not write out. We only care that we can find the component of one vector along another vector by:

$$|b|\cos(\theta) = \frac{a \cdot b}{|a|} = proj_a b$$

While it may be tempting to skip comprehension of this idea; I assure you it is critical for understanding scattering diagram, and thus how crystals diffract. To reiterate; the thing we're interested in here is the purple line. It's magnitude is the component of a along b. When we get to scattering diagrams, I will use a dot product to specify a distance. Come back and review this section.

How to Approach Learning Crystallography

Protein crystallography is an immensely vast topic that draws from many fields of science. In reality, the processes important for understanding crystallography all fall under the umbrella of physics, but humans have categorized our understandings of seemingly distantly-related phenomena into different research fields as a way of piecing together this giant puzzle of hierarchies we come across in life. We have research fields such as statistical mechanics, thermodynamics, quantum chemistry, biochemistry, kinetic theory of gases, etc., but in the words of Ernest Rutherford, "*All science is physics or stamp collecting.*". A mathematician would argue that all physics is just, at its core, mathematics. I prefer to understand this framework from the words and perspective of Richard Feynman:

> *"We have a way of discussing the world, when we talk of it at various hierarchies, or levels. Now I do not mean to be very precise, dividing the world into definite levels, but I will indicate, by describing a set of ideas, what I mean by hierarchies of ideas.*
>
> *For example, at one end we have the fundamental laws of physics. Then we invent other terms for concepts which are approximate, which have, we believe, their ultimate explanation in terms of the fundamental laws. For instance, "heat". Heat is supposed to be jiggling, and the word for a hot thing is just the word for a mass of atoms which are jiggling. But for a while, if we are talking about heat, we sometimes forget about the atoms jiggling- just as when we talk about the glacier we do not always think of the hexagonal ice and the snowflakes which originally fell. Another example of the same thing is a salt crystal. Looked at fundamentally it is a lot of protons, neutrons, and electrons; but we have this concept of "salt crystal", which carries a whole pattern already of fundamental interactions. An idea like pressure is the same.*
>
> *Now if we go higher up from this, in another level we have properties of substances- like "refractive index", how light is bent when it goes through something; or "surface tension", the fact that water tends to pull itself together, both of which are described by numbers. I remind you that we have to go through several laws down to find out that it is the pull of the atoms, and so on. But we still say "surface tension", and do not always worry, when discussing surface tension, about the inner workings.*
>
> *On, up in the hierarchy. With the water we have waves, and we have a thing like a storm, the word "storm" which represents an enormous mass of phenomena, or a "sun spot", or "star",*

which is an accumulation of things. And it is not worthwhile always to think of it way back. In fact we cannot, because the higher up we go the more steps we have in between, each one of which is a little weak. We have not thought them all through yet.

As we go up in this hierarchy of complexity, we get to things like muscle twitch, or nerve impulse, which is an enormously complicated thing in the physical world, involving an organization of matter in a very elaborate complexity. Then come things like "frog".

And then we go on, and we come to words and concepts like "man", and "history", or "political expediency", and so forth, a series of concepts which we use to understand things at an ever higher level.

And going on, we come to things like evil, and beauty, and hope...

Which end is nearer to God, if I may use a religious metaphor, beauty and hope, or the fundamental laws? I think that the right way, of course, is to say that what we have to look at is the whole structural interconnection of the thing; and that all the sciences, and not just the sciences but all the efforts of intellectual kinds, are an endeavor to see the connections of the hierarchies, to connect beauty to history, to connect history to man's psychology, man's psychology to the working of the brain, the brain to the neural impulse, the neural impulse to the chemistry, and so forth, up and down, both ways. And today we cannot, and it is no use making believe that we can, draw carefully a line all the way from one end of this thing to the other, because we have only just begun to see that there is this relative hierarchy.

And I do not think either end is nearer to God. To stand at either end, and to walk off that end of the pier only, hoping that out in that direction is the complete understanding, is a mistake. And to stand with evil and beauty and hope, or to stand with the fundamental laws, hoping that way to get a deep understanding of the whole world, with that aspect alone, is a mistake. It is not sensible for the ones who specialize at the other end, to have such disregard for each other. (They don't actually, but people say they do.) The great mass of workers in between, connecting one step to another, are improving all the time our understanding of the world, both from working at the ends and working in the middle, and in that way we are gradually understanding this tremendous world of interconnecting hierarchies."

In an effort to try to develop this comprehensive teaching resource for fundamental concepts in crystallography, I have tried to draw from many of these fields as to give an as-

accurate-as-possible account of crystallography, explaining as we go, the conceptual images of the fundamentals. It is heckin' hard to do this! Despite being advised against by Mr. Feynman, I have ***tried*** to draw, carefully, a line all the way from one end to the other (as it relates to crystallography), between the topics that relate to crystallography. As you will see, I have failed—as it is a jumbled mess, and I should have listened to Mr. Feynman. The desire to do so came from my frustration with how I was taught crystallography, and how many concepts within were labelled as a 'black box' because there was no 'easy way' to explain things like photon scattering, or the Fourier transform for those who are not experienced in mathematics. Consider this a 'crash course' in everything you need to understand in order to 'get' what is really going on in biomolecular crystallography.

We are all eternal students; the resource itself was developed as on opportunity for me to learn the concepts better as well! Do not feel overwhelmed when studying this resource; it is highly complex and not something that can be immediately grasped on the first read-through. When reading this resource, do not view it as 'knowledge required for an exam' but rather read it from a philosophical point of view; trying to connect the hierarchies in your mind. You are not alone in thinking it is difficult. Seek help from experienced crystallographers when necessary and make sure to ask questions, being relentless until you understand. When you do finally approach comprehension of the subject matter, it is more beautiful than any visible mountain range or sunset on earth!

This page unintentionally left blank.

Chapter 1

Protein Quantification

SDS-PAGE

Sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) is a very common technique typically used to separate mixtures of proteins by their molecular weight (MW) for analytical purposes. The mechanism of SDS-PAGE relies on there being a) two separate (but physically connected; see *Figure 1.2*) polyacrylamide gels each housing a different pH, b) the presence of SDS in the protein solution and electrophoresis buffer, and c) the use of an electrical current.

The purpose of SDS in SDS-PAGE is to unfold/denature proteins into a mostly linear state and simultaneously coat them with the negatively charged head group of the SDS detergent. Since many proteins have a hydrophobic core, we can use the hydrophobic tail of SDS to drive misfolding of the protein while retaining water-solubility by having an ionic head-group. The sulphate group at the head of the detergent acts to unanimously coat the protein and negate any inherent charge – such that all proteins from the mixture have roughly the same charge.

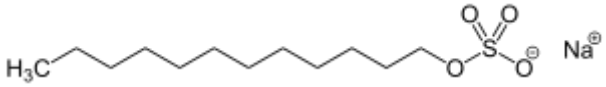


Figure 1.1 Sodium dodecyl sulphate.

The SDS-PAGE gel consists of two gels – an upper 'stacking' gel, and a lower 'resolving' gel. The purpose of the stacking gel is to act as a medium by which all proteins can migrate through to hit the resolving gel at the same time. The stacking gel is polymerized at a lower percentage of acrylamide to allow for large pores in the polymer, allowing for large proteins to migrate at approximately the same rate as smaller proteins. When the gel polymerizes, its pores retain the buffer in which it was situated during polymerization until either a current is run through the gel, or diffusion causes the solutes to escape the gel over long periods of time. In this way, we are able to polymerize the stacking gel and resolving gel at two different pH's (6.8 and 8.8, respectively). The SDS-PAGE buffer consists of Tris HCl (buffering agent which contains chloride ions), SDS (denaturant), and glycine (zwitterionic amino acid). Once the protein is loaded on top of the gel and an electrical current is applied, glycine from the buffer enters the gel at pH 6.8 and changes to the zwitterionic charge state, causing it to migrate very slowly. The chloride ions which

enter the gel migrate very quickly towards the anode. The proteins that are loaded have an intermediate speed of migration and are thus ‘sandwiched’ between these two fronts – a slow-moving glycine front in the back, and a fast-moving chloride front in the front. When these fronts reach the pH = 8.8 resolving gel, the glycine becomes negatively charged and migrates quickly away. This means all of the proteins – regardless of their size – hit the gel simultaneously. Once inside the resolving gel, larger proteins a

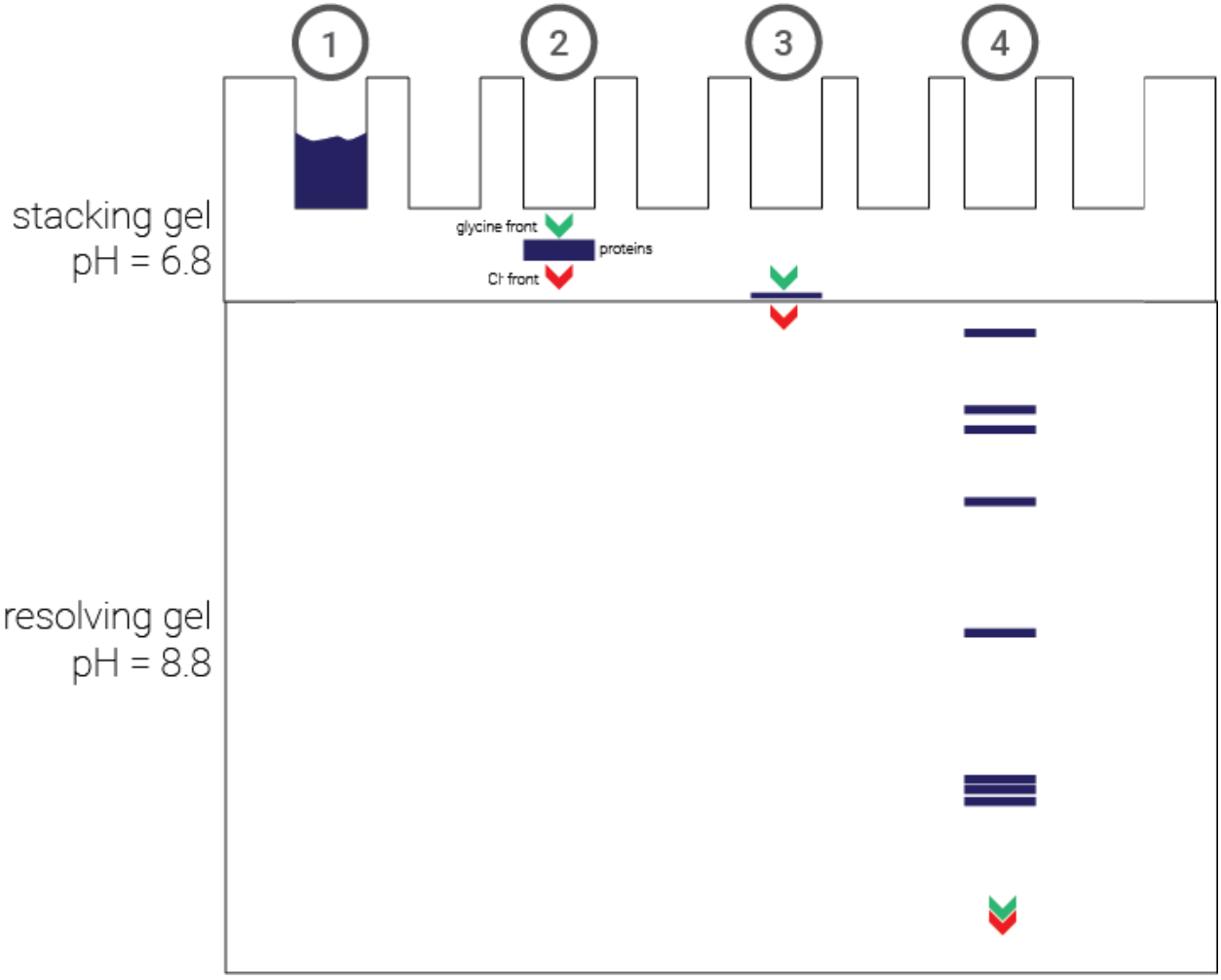


Figure 1.2. Different stages of SDS-PAGE. *(1) Protein sample containing glycerol, bromophenol blue, SDS, and a reducing agent is loaded into the well. (2) After application of an electric current, the protein mixture is compacted and pushed quickly through the stacking gel. (3) The proteins*

all hit the resolving gel at roughly the same time. (4) After the fronts dissipate, proteins are free to migrate at a speed based on their mass alone and resolve from one another. The gel is then stained for visualization.

Making the SDS-PAGE Gel

1. You will need two glass plates – one 'spacer' plate and one 'short' plate. Place the short plate on top of the spacer plate such that there is a gap between the two plates.
2. Insert the pair of plates into the green casting cassette. Ensure the plates are correctly lined up and that their edges are flush and parallel.
3. Place a sponge down on the casting tray and clamp the casting cassette into the casting tray. Now, you can begin preparing the reagents to polymerize the gel.

4. We will be making a 15% resolving gel and a 6% stacking gel. Since the resolving gel sits beneath the stacking, it will need to be poured and polymerized first. Combine the following reagents in two separate 50mL conical tubes – but **do not add the TEMED until you are ready to pour the gel as it will cause the acrylamide to polymerize**:

2.3 mL H_2O

2.5 mL Tris pH 8.8

5.0 mL 37% acrylamide

0.1 mL 10% SDS

100 µL APS

---

8 µL TEMED

2.1 mL H_2O

0.38 mL Tris pH 6.8

0.5 mL 37% acrylamide

50uL 10% SDS

50 µL APS

---

4 µL TEMED

5. Once all the components except for TEMED have been added, mix the tube thoroughly but not vigorously (to avoid foaming from the SDS). Add the TEMED quickly to the resolving gel, and mix thoroughly once again for about 10 seconds. Use a serological pipette to pipette the mixture in between the glass plates up to about 1.5 cm from the top of the plates. Quickly distribute about 1 mL of isopropanol on top of the mixture to create a flat interface during polymerization. Allow 25 min for the gel to form.
6. Once polymerized, dump the isopropanol from the gel and allow it to dry for 2-3 minutes. Now add the TEMED to your stacking gel, and quickly pipette it into the remaining space between the plates (up to the top). Carefully fit a gel comb in between the plates and allow another 25 min for polymerization.

Protein Quantification

There are several different ways to quantify how much protein is present in a given sample, however the most common (and likely quickest) way is by measuring the absorbance of light of your sample at a fixed wavelength of 280 nm. The aromatic portion of hydrophobic amino acids within a protein such as tryptophan, phenylalanine and tyrosine absorb light at 280 nm. In this way, we can utilize Beer's law to calculate the concentration of a protein in a sample, assuming we have the extinction coefficient for that protein. A good approximation of the extinction coefficient for a protein can be determined from primary sequence alone, using tools such as **ProtParam from ExPASy**.

We will be using a NanoDrop spectrophotometer to measure the concentration of our proteins. This instrument is advantageous in that it has a very small path length and thus only a small volume (1-2 microliters) is required to accurately measure the concentration. It is important to have a good idea of how much protein is in your sample before running on a gel so that you don't end up over-loading your gel.

Using the NanoDrop

1. Prepare a blanking solution (buffer only) and your protein.
2. Using a slightly wet KimWipe, wipe the NanoDrop pedestals clean (upper and lower).

3. Dry each pedestal using a dry KimWipe.
4. Pipette 2 µL of your buffer onto the lower pedestal, and blank the instrument by selecting ‘Blank’ from the upper-left menu. Allow 10 seconds for the instrument to blank.
5. Wipe clean the pedestals of your blanking solution, and pipette on 2 µL of your protein. Select ‘Measure’ from the upper-left menu, and allow a few seconds for the instrument to take a reading.
6. Record the concentration and the A280 measurements on the right-hand side of the screen.

Creating Standard Curves

We will be creating a standard curve so that we can estimate the concentration of a protein based on the absorbance of a protein solution which we know the concentration of. We have prepared 6 solutions for a BSA standard curve with Bradford reagent. BSA is dissolved in 1X phosphate-buffered saline at concentrations of 0.0 mg/mL (‘blank’), 0.2 mg/mL, 0.4 mg/mL, 0.6 mg/mL, 0.8 mg/mL, and 1.0 mg/mL. You will need to add Bradford reagent to each of these solutions, and measure their absorbance of light at 595 nm (the wavelength at which the Bradford reagent absorbs light). You will then plot these absorbance values as a function of the protein concentration to obtain a standard curve.

Once you have obtained a standard curve, you will measure a dilution series of your anti-CRISPR protein and estimate its concentration by interpolating its position on the curve.

Loading and Running the Gel

Protein loading dye is a general term for a mixture of SDS, glycerol, reducing agents, and a dye (bromophenol blue) and is used to load our protein into the gel wells. The SDS denatures the protein, glycerol allows the mixture to sit in the bottom of the well without freely diffusing when introduced to the running buffer, the reducing agents break disulfide bonds to linearize the protein, and dye is used for visualization purposes.

1. Remove the gel comb from your gel once it has polymerized and place it in a running cassette such that the open faces of the glass plates face inwards. Place a buffer dam or another group's gel in the other position in the cassette and clamp it shut.
2. Place the cassette in the buffer container (line up the anode and cathode ports), and fill the container with SDS running buffer.
3. Mix 15 µL of 2x SDS loading dye with 15uL your protein at an appropriate concentration (1-10ug total protein) in a 1.5 mL Eppendorf tube.
4. Boil your sample(s) at 95^{o}C for 3-5 minutes. Some proteins are extremely stable even in the presence of SDS. This step ensures that they will denature and become coated with SDS.
5. Using a gel loading tip, take 20 µL of your sample and carefully load it into a well. Ensure you load a protein ladder on the gel for visualization later.
6. Connect the leads from the container lid to a power-pack and run the gel at 180 V for 60 minutes.
7. Shut off the power and remove the lid.
8. Remove the running cassette from the container and dump the excess buffer back into the container or down the sink.
9. Carefully separate the glass plates using a green spatula and rinse your gel with excess ddH_2O.
10. Place the gel in a Tupperware container and add ~20 mL of Coomassie Blue stain. Let sit for 10-15 minutes.
11. Pour the stain back into the stain container and rinse excess stain from the Tupperware container with ddH_2O.
12. Pour ~20mL of destain into the container and fold a piece of paper tower neatly inside. Allow the gel to destain for several minutes before viewing over a lightbox. To fully destain the gel typically takes a few hours.

Weekly Questions

A) Research and explain the mechanism behind Coomassie blue. Can you think of any limitations of this stain?

B) What are the components of the staining and destaining solution? Explain why each of these are included.
C) SDS PAGE is very good for separating proteins by their molecular weight. Comment on why it is not used frequently as a purification method.
D) Were there any bands on your gel you did not know the identity of? If there (realistically or hypothetically) were, what technique could you employ to identify the protein within the band?
E) Can you think of a situation when you would not be able to quantify protein using absorption of light or with the Bradford reagent? How else might you be able to measure the concentration of a protein in solution?

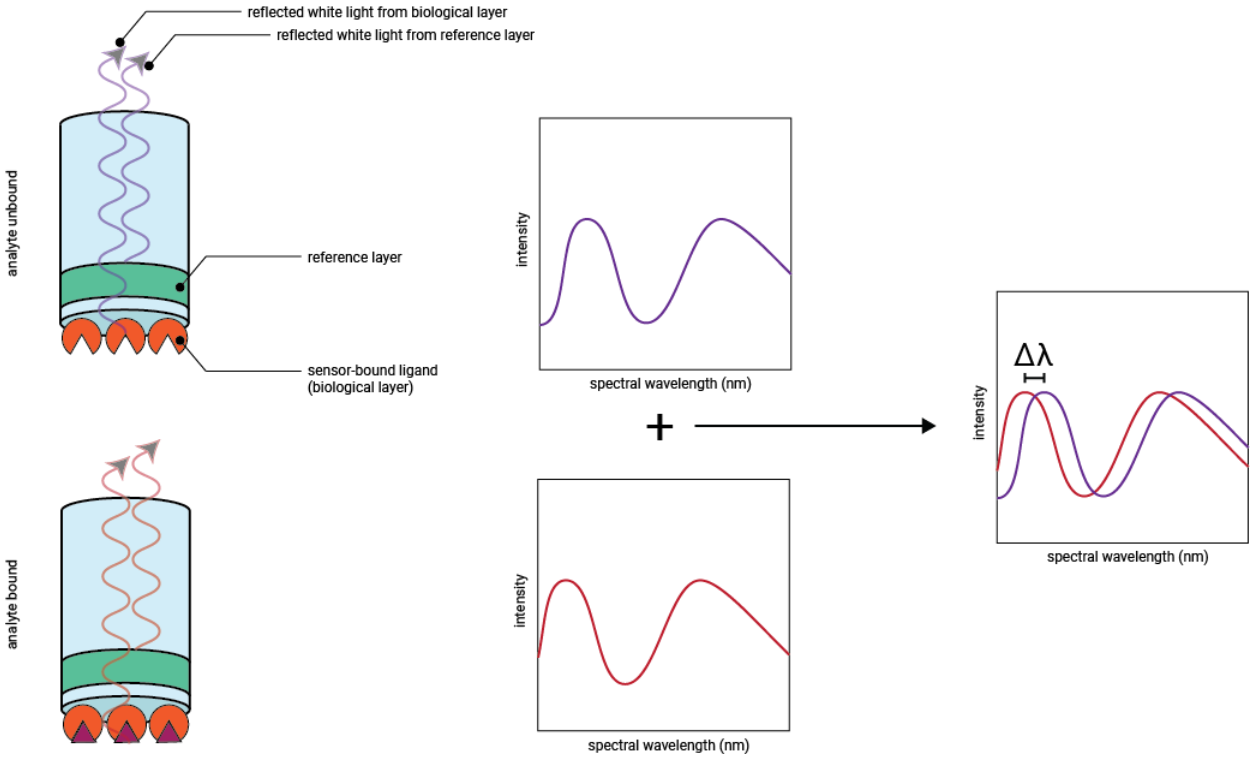
Chapter 2

Protein-Protein Interactions

Biomolecular Interactions (BLI, SPR, ITC or MST)

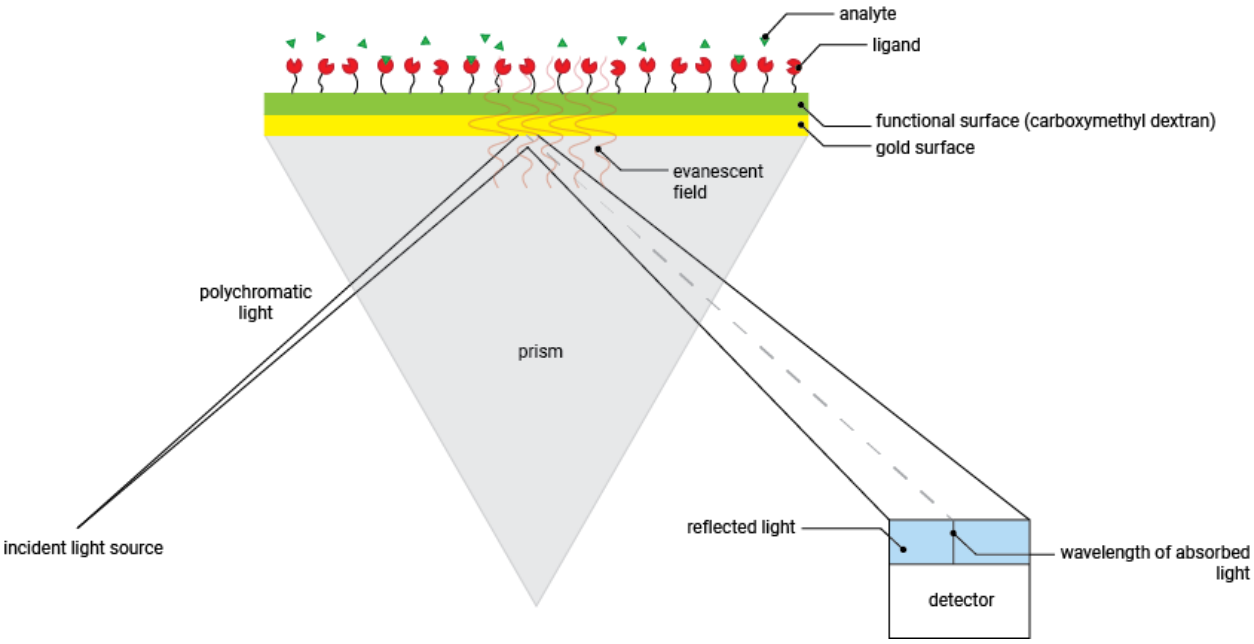
We will be using a biophysical technique to characterize the interaction between your *Cas* proteins and *anti-CRISPR* proteins.

Bio-layer interferometry (BLI) is an approach that measures the binding kinetics and affinity between two proteins (or molecules) using interference of electromagnetic radiation (light). White light is shone downwards and reflects off of two surfaces – 1) a reference layer and 2) a biological layer. The interference pattern between these two signals gives an indication of the thickness of the biological layer. Using this idea, we can attach a protein of interest to the sensor (ligand) and incubate it with a molecule we suspect it interacts with (analyte). Monitoring the signal in real-time and performing steady-state equilibrium analysis allows us to calculate on/off rates as well as equilibrium dissociation constants for biomolecular interactions.



Note – while it is commonly taught that a ‘ligand’ binds to a ‘receptor’, the terminology used in kinetics usually refers to the sensor-bound biomolecule as the ‘ligand’ and the free-flowing biomolecule as the ‘analyte’.

Surface plasmon resonance (SPR) provides similar information to that of BLI, however it uses different physical principles to monitor ligand/analyte binding. It utilizes light passing through a prism that is interfaced with a gold surface. Once a photon of polarized light hits the gold surface, it excites photons within the surface layer birthing ‘plasmons’. These plasmons propagate within the gold parallel to the surface and generate an electric field that extends perpendicular to the surface (upwards toward the sample and downwards into the prism) to a depth of 300 nm. This wave is known as the evanescent field or evanescent wave. The evanescent wave amplitude decreases exponentially as a function of distance from the gold surface, meaning it is more pronounced closer to the surface. Instantiation of this evanescent wave is dependent upon light interacting with the metal surface at some critical angle. This angle will change depending on the composition and mass of the sensor layer, and so in this way, we can gather information about biomolecules which bind or dissociate from the surface. So, as we alter the composition of the sensor surface, we are altering the refractive index. Small changes in the refractive index of the sensor (when a biomolecule attaches or detaches) then disallow plasmon formation (assuming you are measuring with monochromatic light). Altering the incident light angle until resonance can occur again provides us with our measurement – response units (RU). One RU is equivalent to changing the critical angle by 1/10,000th of a degree (10^{-4}).

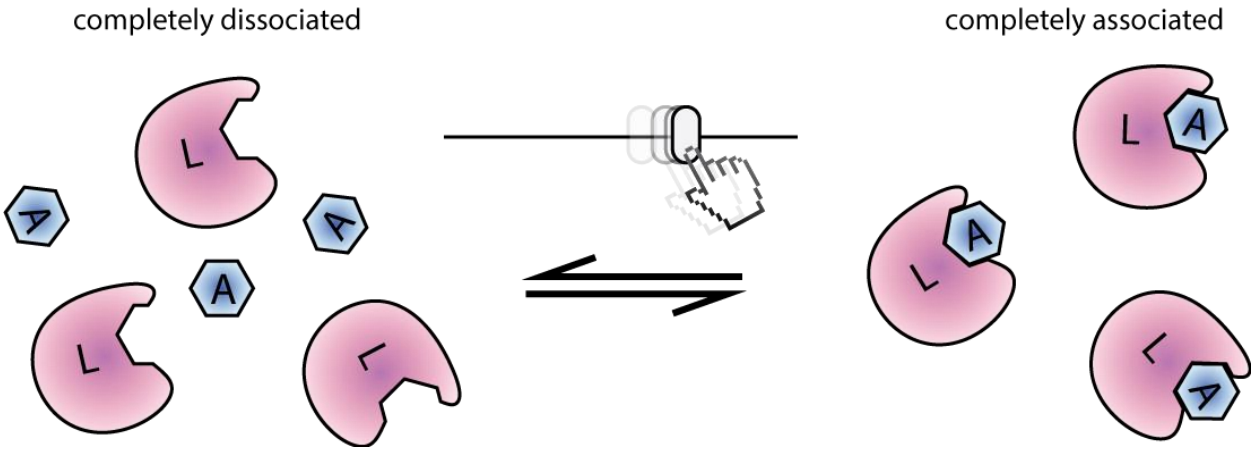


Put simply, any approach used to monitor biomolecular interactions measures some property of binding saturation (how it does so is specific to the instrument) by the parameter of analyte concentration. For example, isothermal titration calorimetry (ITC) measures heat released as a function of concentration. BLI and SPR monitor response units as a function of concentration. Microscale Thermophoresis (MST) measures thermophoresis as a function of concentration. The techniques are similar in principle – they just achieve measurements by different means.

Kinetics – Equilibrium Binding

We can describe the binding relationship between two proteins (for example, your Cas9 and ACR) using a chemical equilibrium between a **L**igand and an **A**nalyte:

$$\mathrm{L} + \mathrm{A} \leftrightharpoons LA$$



The mathematical representation is an analytical way of expressing the two extremes, or states that the system can be in. In reality, for a large number of ligand/analyte pairs, the situation is somewhere in between —a statistical average. You can think of there as being an adjustable slider that alters the equilibrium. As we move the slider to the left, more and more proteins become unbound from analyte. As we move it to the right, more and more become bound. In the lab, we can actually 'adjust' the slider, and alter the affinity of interaction between these two molecules by changing buffer conditions, making mutations, etc. It is our job however, to measure ***where the slider is naturally positioned at physiological conditions***; as we are interested in how these two operate in a natural setting. We call this the ***binding constant***. It is important to also note that binding constants under non-physiological conditions are just as mathematically valid. Though, for this course we will focus only on conditions that closely mimic physiological conditions.

Since all chemical equilibria can be characterized by an equilibrium constant (products over reactants at equilibrium), we can define an 'association' constant which is a measure of the ratio of complexed to uncomplexed species:

$$K_a = \frac{[LA]}{[L][A]}$$

This fraction is an expression of where the slider is positioned. A high number for $[LA]$ and a low number for $[L]$ and $[A]$ means a high concentration of bound protein, and low concentration of unbound protein. What happens when we divide a big number by a small number? The number stays big. In other words, we have a large K_a. K_a has units of M^{-1} because we have two concentration terms in the denominator, and only one in the top.

We can rearrange the above equation and see that the K_a is directly proportional to the concentration of free analyte:

$$K_a[A] = \frac{[\boldsymbol{LA}]}{[\boldsymbol{L}]}$$

Think about what this means! This expression holds a critical idea. As we increase $[A]$ (the amount of unbound analyte), the ratio ($\frac{[LA]}{[L]}$) is going to get bigger. If this ratio gets bigger, it implies it has a larger numerator and/or smaller denominator, which in turn implies there will be more complexes present in solution.

But we have one problem. When we add analyte into our system to measure binding saturation, the free analyte in solution binds to the ligand. A particle that was once contributing to the value of $[A]$ is now contributing to $[LA]$! *We cannot reliably measure binding saturation if our analyte concentration is changing*. One way around this is to use a very small amount of ligand, and an excess amount of analyte. In this way, when free analyte is bound up by sensor-

attached ligand, the concentration of the analyte solution does not change appreciably because there is so much present. We do not especially like having too many variables in our equations, so we have effectively 'gotten rid' of one with a trick! We have designed our experiment in such a way that $[A]$ will not change – by making its concentration so high relative to the ligand that it will never appreciably change.

The instruments we use to measure these interactions are highly expensive and technically impressive – but they are, at the end of the day, still quite limited. We of course cannot simultaneously measure the state of every protein molecule in solution. The instrument can sense only what is going on 'as a whole'. Can you blame it? Small things are hard to see! It looks at its sensor, or chip, and tries to estimate "how much of my chip is covered in analyte?". As such, we talk about 'fractional binding'. The best instruments are only so impressive as to be able to measure generally what percentage of some 'sensor' is covered with analyte (note that not all instruments are chip/sensor-based, this is just an example). So, with this idea in mind, we need to use math to cater to our incompetent instruments.

The instrument measures a parameter, θ, which represents the fraction of analyte-binding sites on the ligand that are occupied:

$$\theta = what\ percentage\ of\ my\ chip\ is\ covered\ in\ analyte?$$

Which should be restated as:

$$\theta = \frac{binding\ sites\ occupied}{total\ binding\ sites}$$

How can we write these phrases mathematically?

$$\theta = \frac{[LA]}{[LA] + [L]}$$

Alright great. We made it this far, but it seems like we are no further ahead. We need to somehow get K_a into our expression. It is, after all, the quantity we are interested in (slider position)! Not only that, but I just told you our apparatus cannot look at its sensor and individually count which ligands are bound and which ones are not bound, so how does this expression help us? We need to do some algebraic manipulation. Let's think back to our equilibrium expression and see if we can substitute out $[LA]$, because we know we cannot measure it.

$$K_a = \frac{[LA]}{[L][A]}$$

We can re-arrange this like so:

$$K_a[L][A] = [LA]$$

Now, we can finally substitute:

$$\theta = \frac{[LA]}{[LA] + [L]} = \frac{K_a[A][L]}{K_a[A][L] + [L]}$$

We see the ligand concentration is in every term! Let's factor it out, and divide it by itself.

$$\theta = \frac{\cancel{[L]}(K_a[A])}{\cancel{[L]}(K_a[A] + 1)}$$

Things are looking much simpler now.

$$\theta = \frac{K_a[A]}{K_a[A] + 1}$$

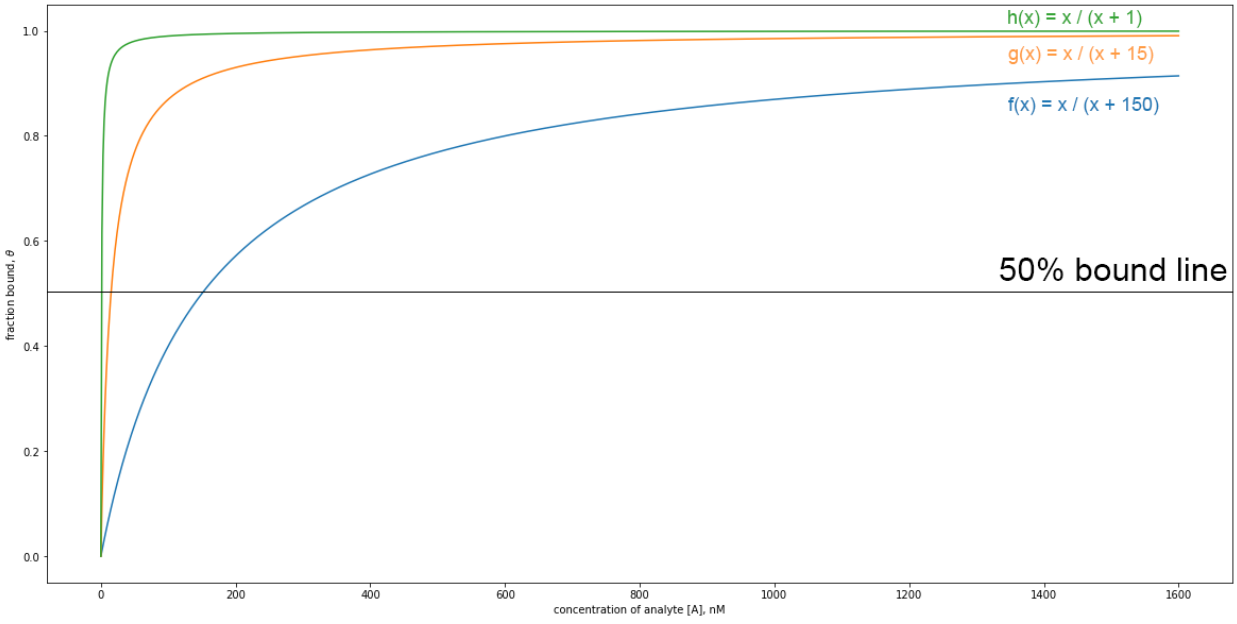
Let's now factor out K_a, and again divide through:

$$\boldsymbol{\theta = \frac{[A]}{[A] + \frac{1}{K_a}}}$$

This is the function we have been looking for! We were able to mathematically describe our association constant in the only two things we can measure in practice: the fraction of our sensor bound, and the concentration of our analyte that we are putting in excess. Keep in mind $\frac{1}{K_a}$ is a constant. In general, for a given protein-protein interaction in some environmental condition (pH, ionic strength, etc.) it does not change. In fact, the inverse of the association constant is actually called the dissociation constant, or K_D. So this function takes on the form:

$$f(x) = \frac{x}{x+c}$$

Watch what happens when I plot this function in Python for a couple different values of c.



Note that when the constant is big, the hyperbola has a more gradual curve, and requires a higher concentration of analyte to reach 50% bound. Smaller constants require a lower

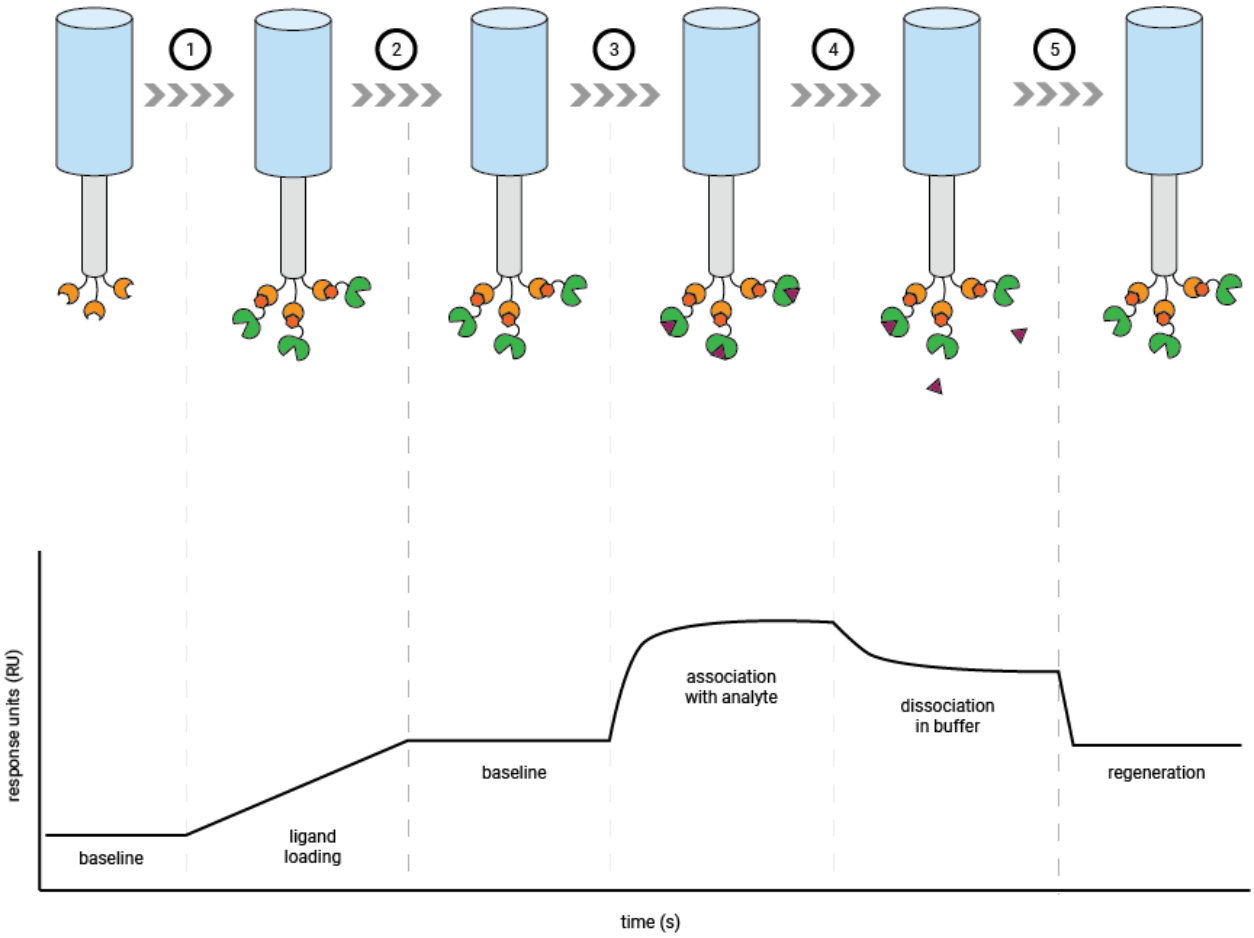
concentration to reach 50% bound. Remember that slider at the beginning of this section? That slider is simply the dissociation constant. This is the property of the dissociation constant. The dissociation constant is a concentration that measures the concentration of analyte required to half-saturate the ligand. Draw a vertical line down from the 50% bound line at the point each function intersects the line, and you will see that it equals the constant.

Finally, and majestically, we have arrived at our ultimate function!

$$\theta = \frac{[A]}{[A]+K_D}$$

Setting up for BLI

We will be using the **Octet RED96** for our experiments. A typical kinetics experiment on this instrument involves loading your protein of interest onto streptavidin-coated sensors, stabilizing the signal, and then carrying out association/dissociation reactions at varying concentrations, allowing the association to proceed until it reaches equilibrium. By plotting the binding signal at equilibrium against concentration, we can construct an equilibrium binding curve and perform something referred to as a ‘steady-state analysis’, because it represents relative saturation of our ligand at different concentrations during equilibrium. Often times, the analyte will not dissociate entirely from the ligand and as such we need to place the proteins in a harsher buffer condition to allow total regeneration of the sensor surface so that all (or almost all) of the original binding sites are available again for binding at the next concentration.



A typical kinetics run on the Octet RED96. *The ligand is loaded onto SA-coated sensors and a baseline is established. The sensor is then dipped into analyte and kinetic parameters are measured. The sensor is then regenerated and the process (starting from after ligand loading) is repeated several times at different concentrations*[2]*.*

Conducting a steady-state analysis requires that we measure association and dissociation at several different concentrations (typically five). Concentrations are chosen based around what the estimated K_D of interaction may be. For example, if the K_D is believed to be roughly 50 nM, we choose concentrations that flank this value by 10X in each direction (concentrations between 5 nM and 500 nM). Because we know that binding curves are not linear in nature, we try to select points that accurately construct the inflection point of the equilibrium binding curve as to have a more accurate model of binding. One way to do this is to select concentrations using a logarithmic scale.

$$\log(500) = 2.69$$

$$\log(?) = 2.19$$

$$\log(50) = 1.69$$

$$\log(?) = 1.19$$

$$\log(5) = 0.69$$

We need to select two more appropriate concentrations on this scale to satisfy the 5 points required for curve construction.

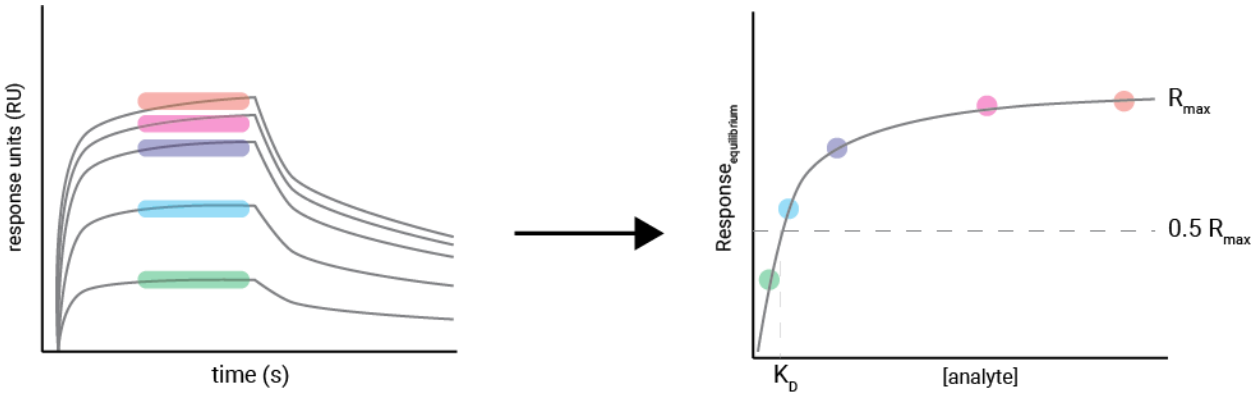


Figure 2.4. Construction of steady state analysis from real-time data. Response values at equilibrium are calculated as an average of data points within a range during equilibrium (coloured bars). These points are plotted as a function of their concentration to obtain an equilibrium binding curve or a steady state binding curve. The K_D of the interaction is calculated by looking at the concentration of analyte required to saturate half of the total binding sites.

The following reagents will need to be prepared before access to the instrument:

- Kinetics buffer
    - 1X PBS, 0.002% Tween-20, 0.1 mg/mL BSA

- Regeneration buffer
    - 100 mM sodium citrate, pH = 4.5, 50 mM EDTA, 100 mM NaCl
- Biotin-labelled ligand **(provided by TA)**
- Analyte at appropriate concentrations (5 tubes) (200 µL minimum/each)

The sensor in BLI is ‘dipped’ from well to well in a 96-well plate housing the different solutions. You will need to set up the plate with the solutions in the correct locations.

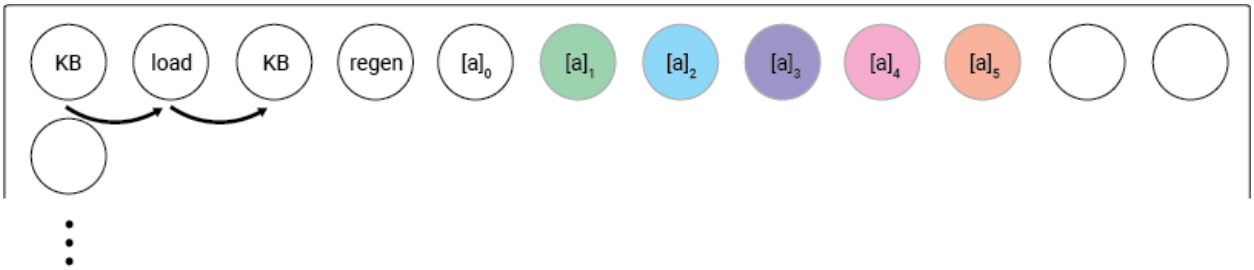


Figure 2.5. A single row from a 96-well plate for a BLI experiment. Kinetics buffer (KB) is loaded in several wells to be available for baseline steps. Biotinylated protein is loaded into ‘load’ well. Regeneration buffer in ‘regen’ well. Analyte concentrations 0 (representing just kinetics buffer) through 5 (highest concentration) are loaded in sequence. Wells can be re-used. Try to draw arrows representing how you would instruct the computer to dip your sensor in order to obtain data for a steady-state analysis curve.

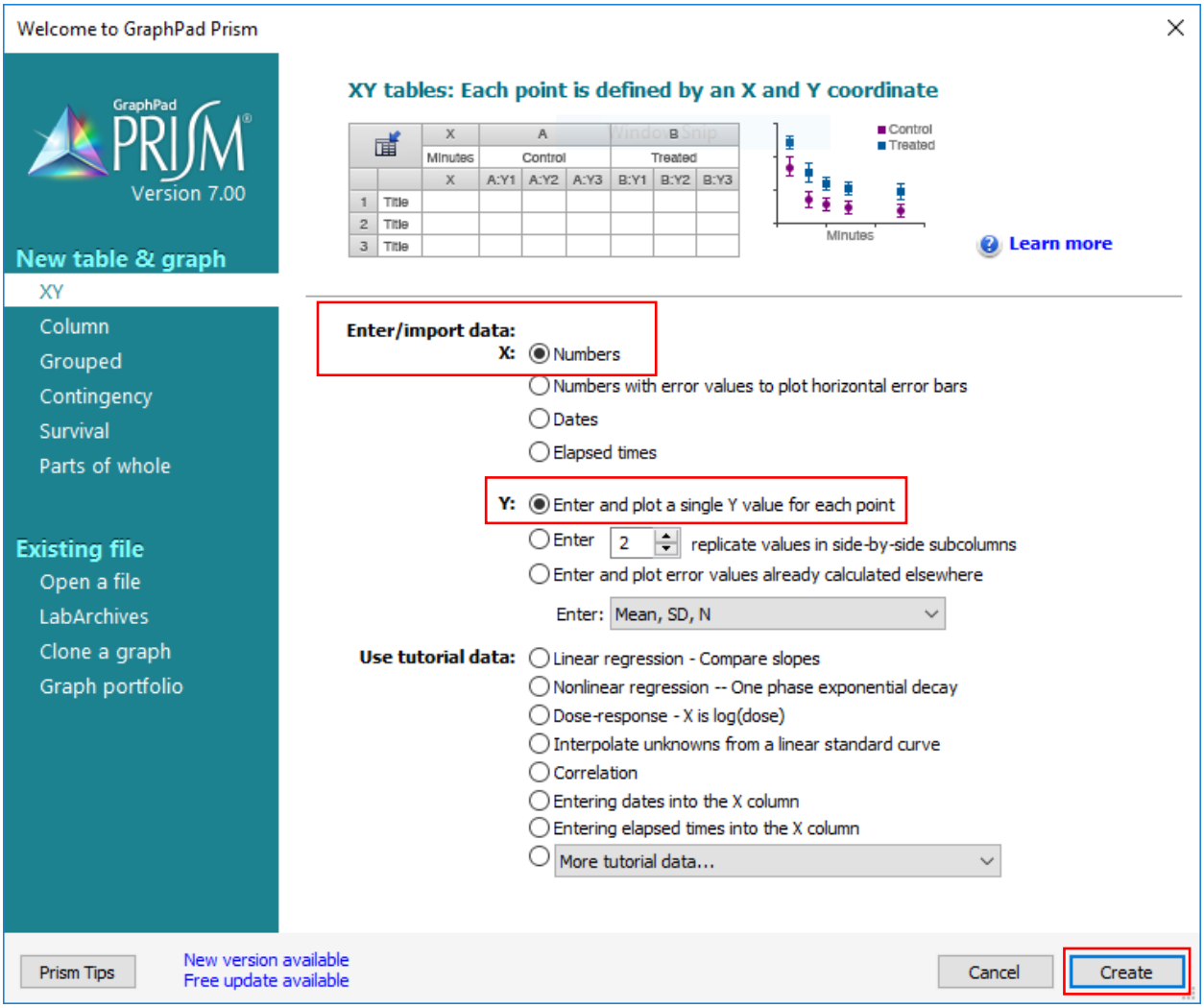
Once your plate is set up, consult with your TA for running the instrument.

Curve Fitting in Prism

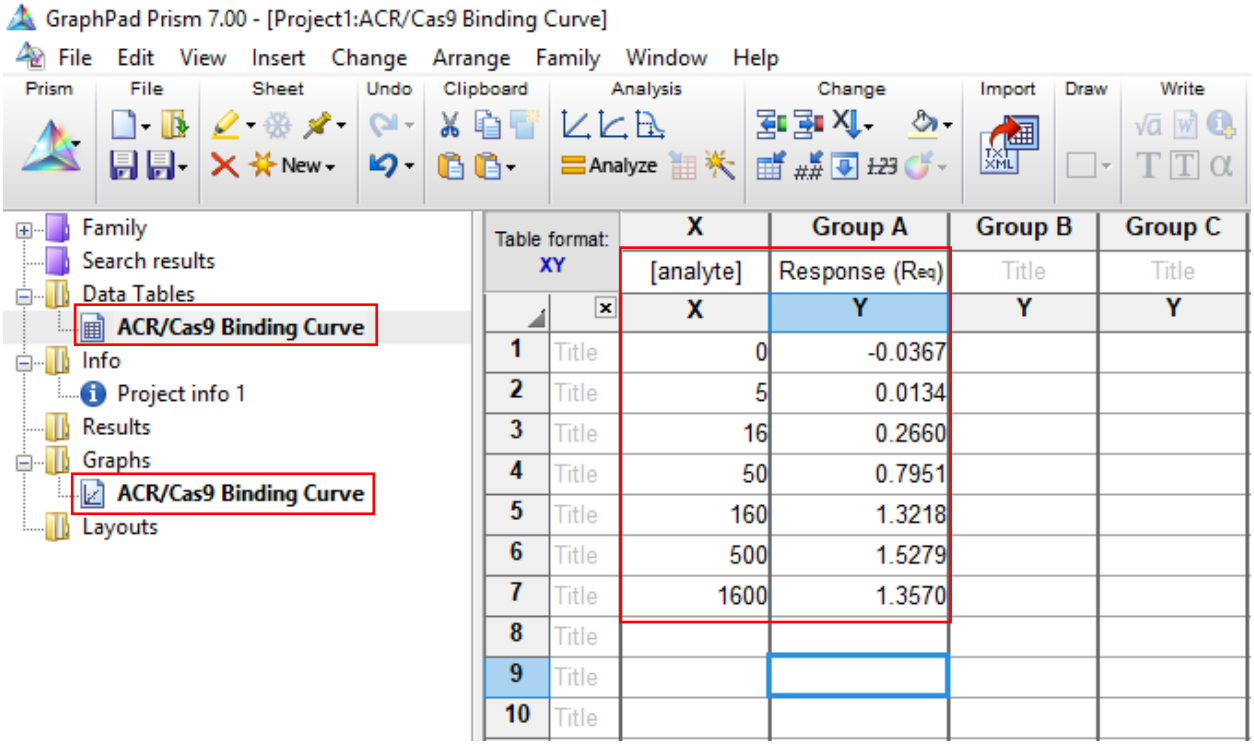
To fit curves to our data, we will need a copy of GraphPad Prism. A 30-day free trial for Windows and Mac is available from:

https://www.graphpad.com/scientific-software/prism/

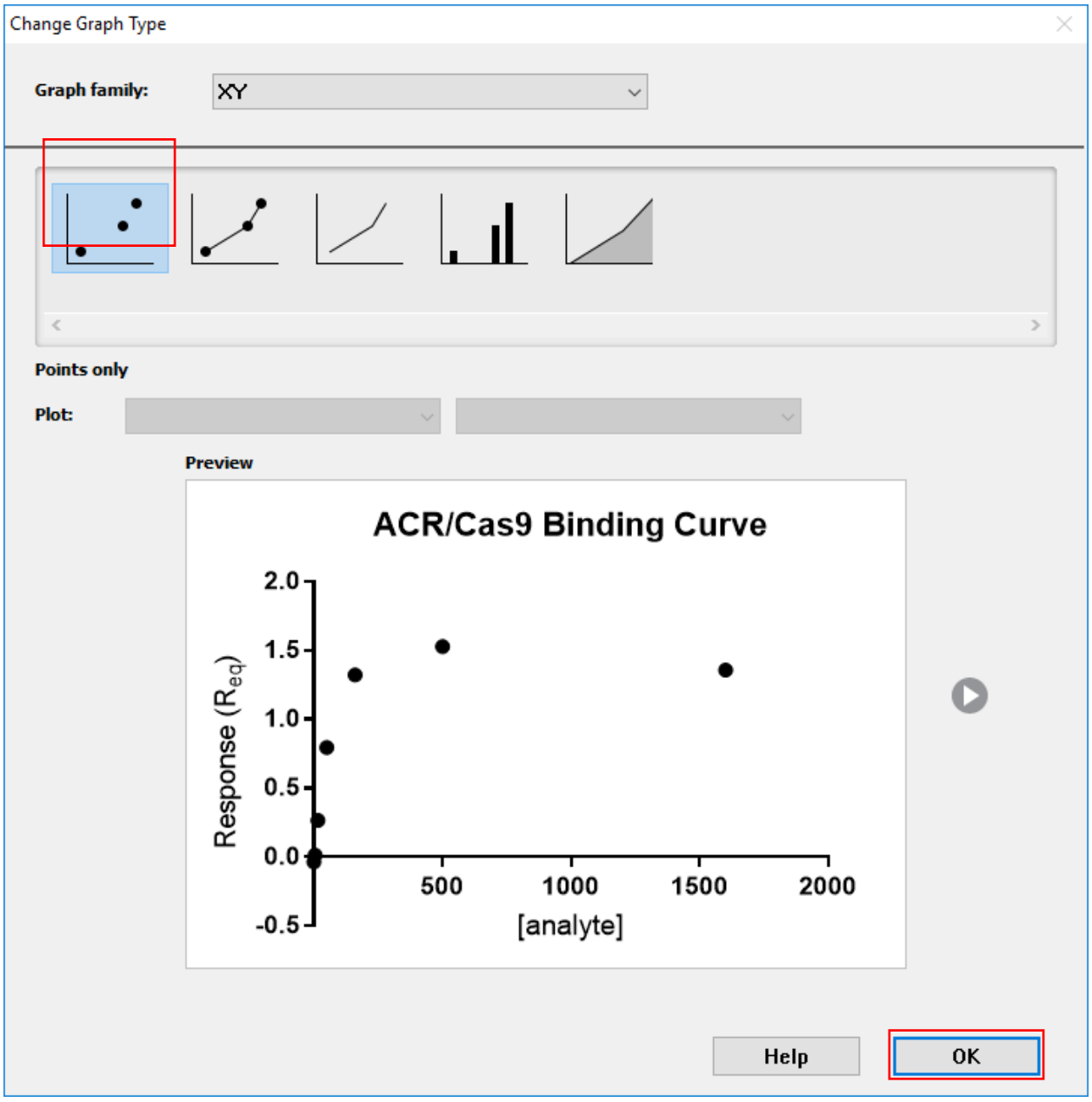
1. Install this software and open it. Under the Enter/Import Data header, select X – numbers, and Y – enter and plot a single Y value for each point.



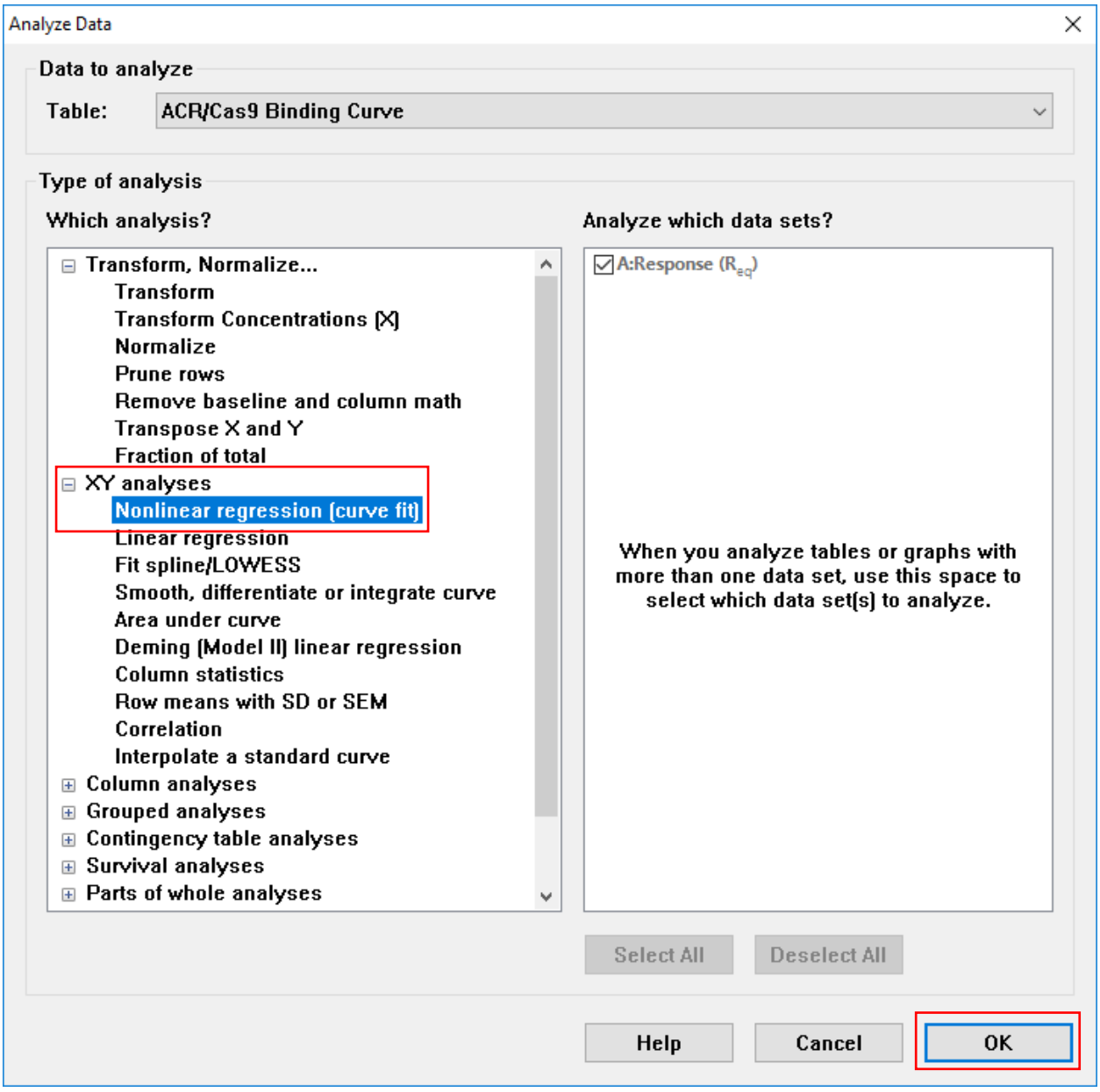
2. Press create, and enter in the data you obtained from the Octet Red, like the sample data below. You will likely have a sheet named ‘New Data’. Right click this and rename it to something that describes your binding interaction. We have chosen **ACR/Cas9 Binding Curve**. Label the axes with **[ACR],** and **Response (R_{eq}).** You will notice that there is a folder below your data tables folder named ‘Graphs’. Here, GraphPad will plot changes to your data in real time.



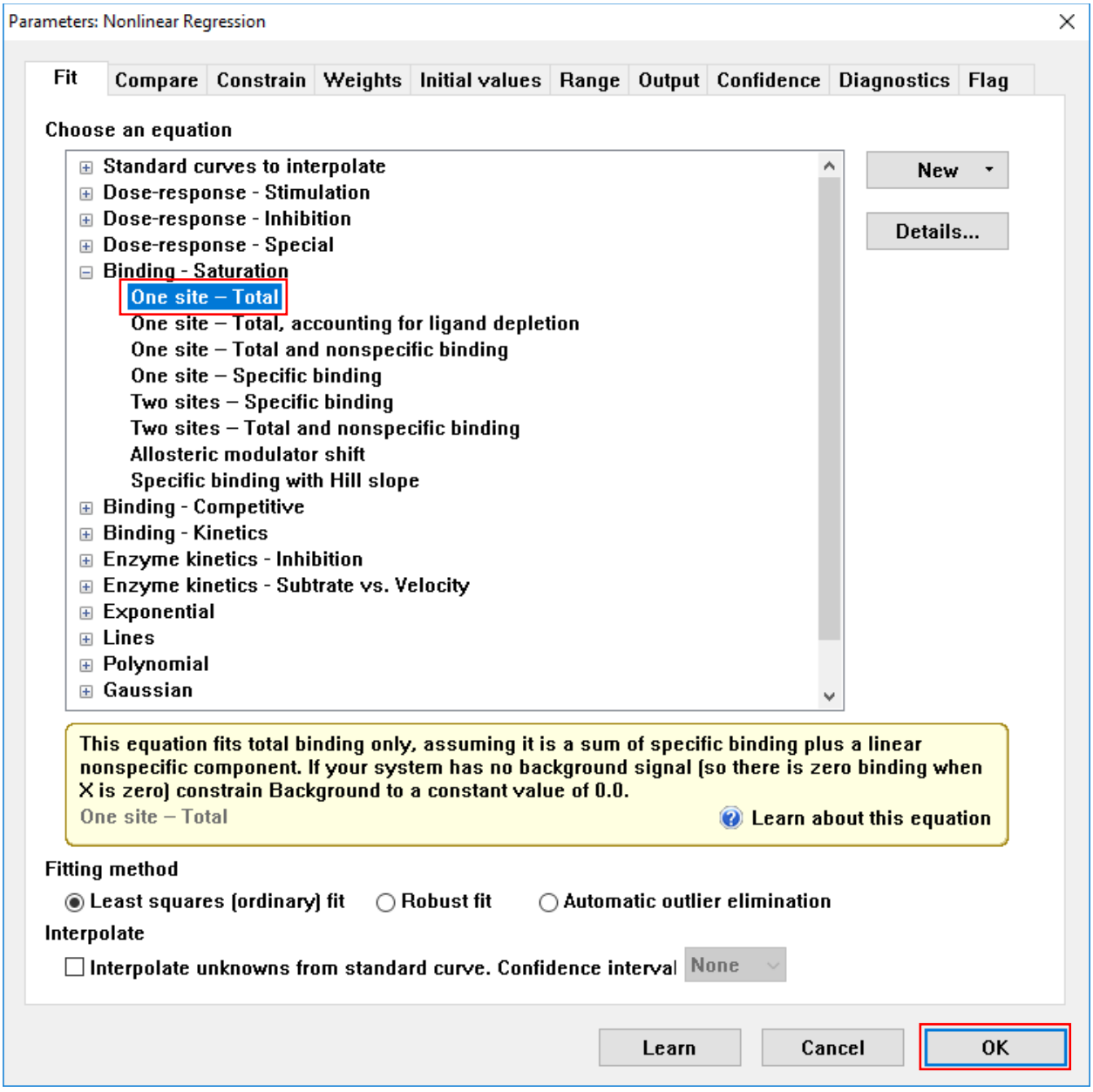
3. Clicking on the graph sheet should prompt you to choose which type of graph we want. Of course we don't want to connect our points with lines, but rather fit a mathematical curve to them that describes them best. So, choose **'Points Only'**, and press **'OK'**.



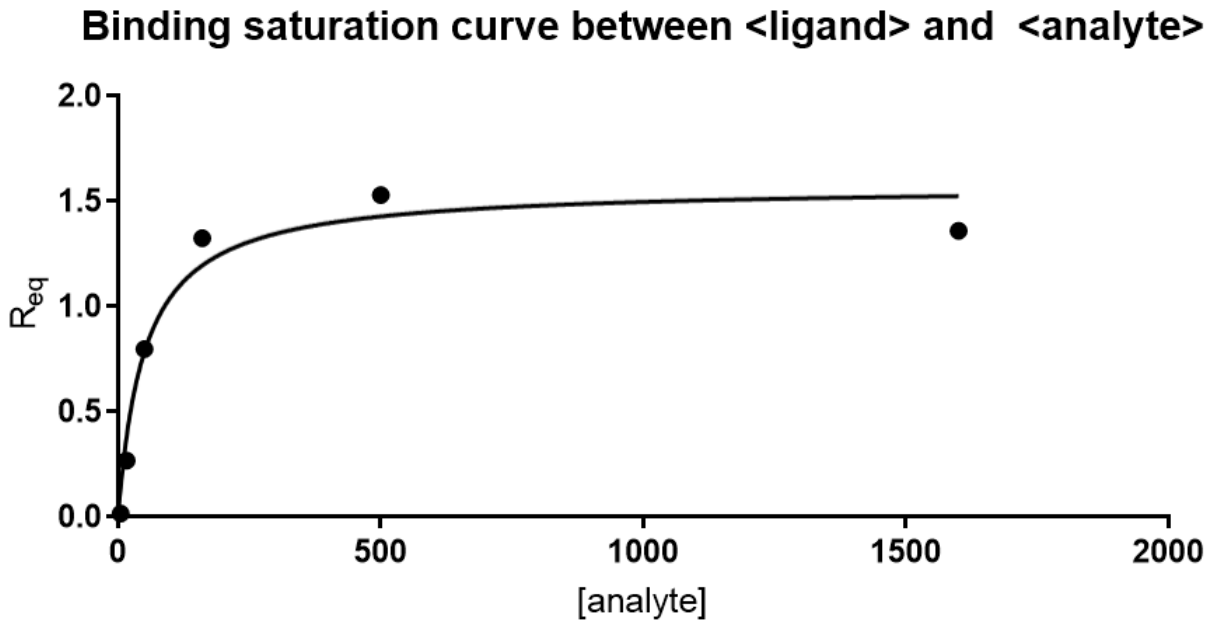
4. Select ‘Analyze’, and choose ‘Non-linear regression – Curve Fit’, since we are fitting a curve to nonlinear data.



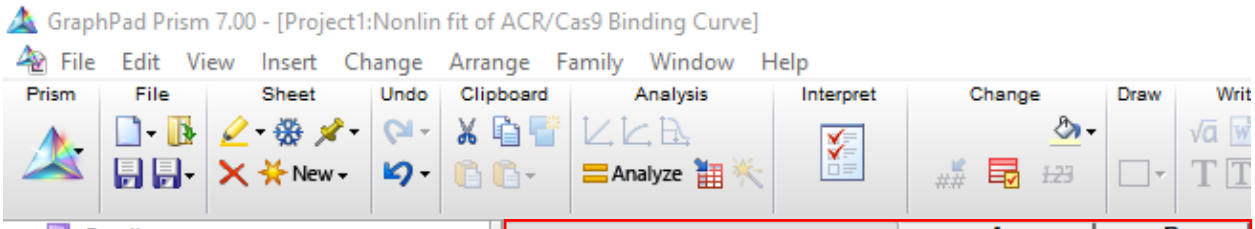
5. We want to construct a binding curve, and since we did not subtract any curves off of the data beforehand, we will select One site – total binding.

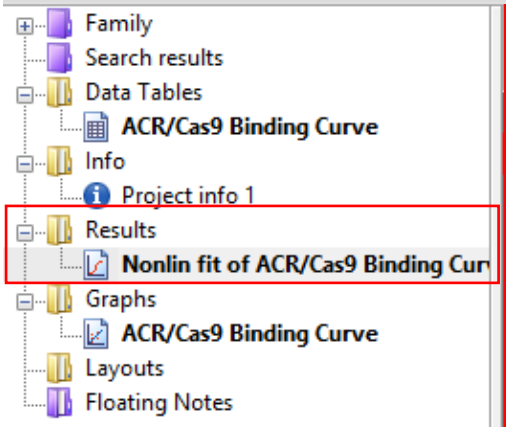


6. Your graph should now be fit with a curve like the one below.



7. In the left-hand pane, you can now select the sheet in the 'Results' folder. This table has information about the goodness of fit, the properties of your binding interaction (i.e. R_{max}, K_D, etc.) and their error values.



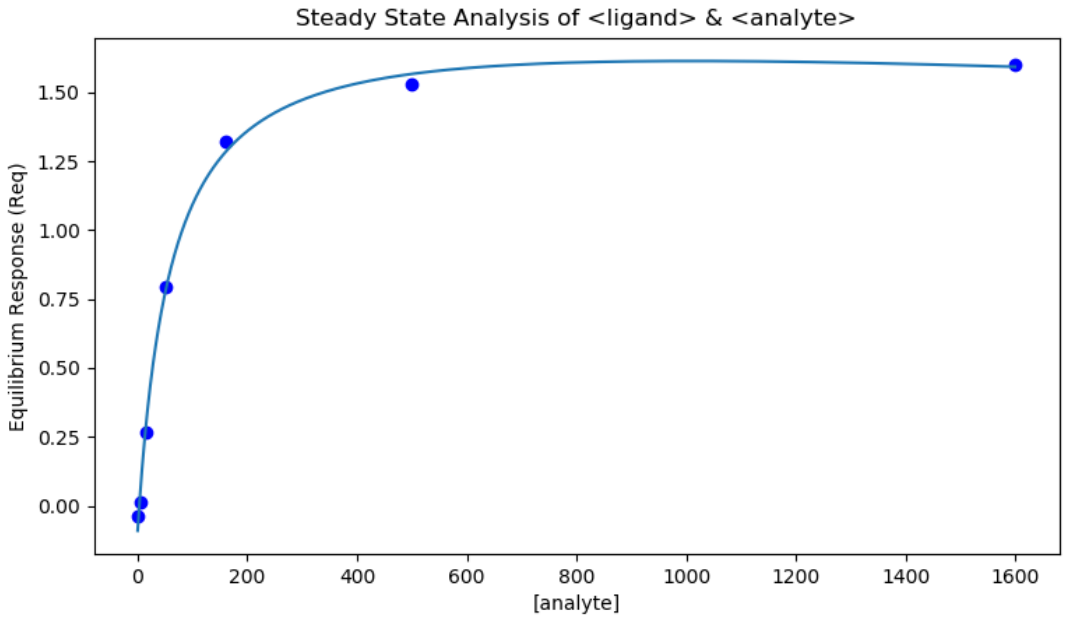


Nonlin fit	A	B
	Response (Req)	Title
	Y	Y
1 One site -- Specific binding		
2 Best-fit values		
3 Bmax	1.57	
4 Kd	50.57	
5 Std. Error		
6 Bmax	0.1086	
7 Kd	15.03	
8 95% CI (profile likelihood)		
9 Bmax	1.317 to 1.851	
10 Kd	25.14 to 98.71	
11 Goodness of Fit		
12 Degrees of Freedom	5	
13 R square	0.9687	
14 Absolute Sum of Squares	0.08452	
15 Sy.x	0.13	
16		
17 Number of points		
18 # of X values	7	
19 # Y values analyzed	7	
20		

Curve Fitting in Python - The Code

```
#Lets import and alias our libraries
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
from scipy.optimize import curve_fit

#Load the excel file
file = 'Data/sample_kinetics_data.xlsx'
df = pd.read_excel(file, sheet_name='Sheet1')

#Use pandas to obtain our data points
concs = df['concentrations'].tolist()
reqs = df['req'].tolist()

#Define the function for fitting
def EquilibriumBindingModel(conc, Bmax, Kd, NS, background):
    return ( (Bmax*conc) / (Kd+conc) ) + (NS*conc) + background

#Lets guide the curve fitting algorithm by supplying it with initial
guess parameters
p0 = [1.5, 50, 1.0, 1.0]

#Use scipy's curve fit function to fit minimize the square difference
between our data points and the model
c, cov = curve_fit(EquilibriumBindingModel, concs, reqs, p0)
print(c)
print(cov)

#Print to console the value we are most interested in
print("The KD of the interaction between ligand and analyte is " +
str(c[1]))

#Plot our data points and the fitted curve
plt.plot(concs, reqs, 'bo')
plt.plot(np.arange(0,1600), EquilibriumBindingModel(np.arange(0,1600),
c[0], c[1], c[2], c[3]))
plt.xlabel('[analyte]')
plt.ylabel('Equilibrium Response (Req)')
plt.title('Steady State Analysis of <ligand> & <analyte>')
plt.show()
```

Matplotlib should have generated a figure similar to the one below:



Weekly Questions

A) List the components of BLI kinetics buffer and explain what the role is of each of the components.

B) Which of the techniques you learned about this week would be ideal for monitoring the binding of a small molecule to a protein?

C) What was the K_D of the ACR/Cas9 interaction? What does this number actually represent?

D) Why might it be useful to subtract off binding from a protein (using ACR as the analyte, and BSA as a sensor-bound ligand) from a binding experiment?

Chapter 3

Crystallization Thermodynamics

Protein Crystallography

Determining the structure of a biomolecule often greatly informs us on what function it may carry out. Understanding the structure also allows us to carry out bioengineering applications on proteins and enzymes, for example to improve their efficiency. There are three dominant methods for determining protein structure – X-ray crystallography, cryo-electron microscopy (cryo-EM), and nuclear magnetic resonance (NMR). There are also some emerging mass-spectrometry based techniques, but in this course we will only touch on crystallography.

Protein crystallography and the methods for solving structures are quite complex and mathematically involved. As a result, we will only touch on the practical and surface-level aspects of the technique that you should understand when entering the field.

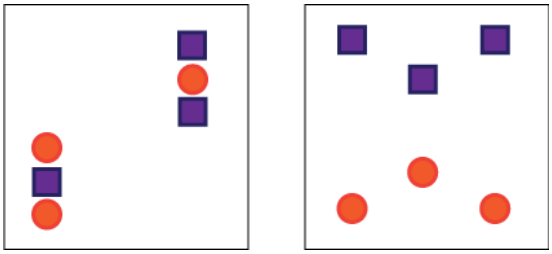
In short, x-ray crystallography involves thermodynamically- and kinetically-driven packing of many copies of a single protein into a crystal lattice to ultimately form small proteinaceous crystals. When exposed to x-rays, the repeating array of electron planes within the crystal act as a miniature amplifier to alter the direction and phase of the X-rays. The manner in which the X-rays scatter tells us positional information about the electron density within crystal, and thus the protein. Getting started with this technique requires the ability to purify large amounts of protein to a very high degree of purity (milligram quantities at 95% purity). Purity is important because we are forming crystals – we do not want heterogeneity within the crystal lattice, otherwise the signal will be disturbed during amplification, or the crystal may be unable to even form in the first place. Even if one is successful in obtaining protein crystals and exposing them to X-rays, only the amplitude of the resultant diffracted rays can be directly measured with X-ray detectors. The other component of the wave – the 'phase' – cannot be measured directly and must be back-calculated using a variety of experimental phasing techniques.

Entropy is *NOT* Disorder![3]

My goal in this section is to convince you that physics is even more interesting than biochemistry (in fact, all fields of science are fundamentally governed by physics), and then remind you that you're in fourth year, so it's too late to switch anyways. Having a thorough

understanding of thermodynamic parameters (entropy, enthalpy, temperature, etc.) is essential for having a solid understanding of biochemical phenomena, crystallography included. I will aim to ensure you have a rigorous conceptual and supplementary mathematical understanding of each of these parameters. I want to ensure that every time we put up an equation for you to look at, you know ***EXACTLY*** what each term means, inside and out. Otherwise, it just looks intimidating.

Unfortunately, the concept of entropy is often misrepresented when students are first learning the concept (I fell victim to this) because terminology such as 'disorder' and 'mixed-up-ness' were used by founders of this concept as a way of conceptualizing this phenomena. These terms however are misleading and pave a path that makes understanding higher-level concepts more challenging. Disorder, as a concept, is subjective. What one person sees as more disordered may be more ordered to another person. In reality, any system at any moment in time has a value for its entropy, regardless of how one person interprets its level of disorder. For example, a glass of crushed ice is 'less entropic' than a glass of liquid water. But how do we know this? Well, we need to first establish a concrete definition of entropy that allows us to make quantitative, objective measurements, instead of describing how 'random' a system is.



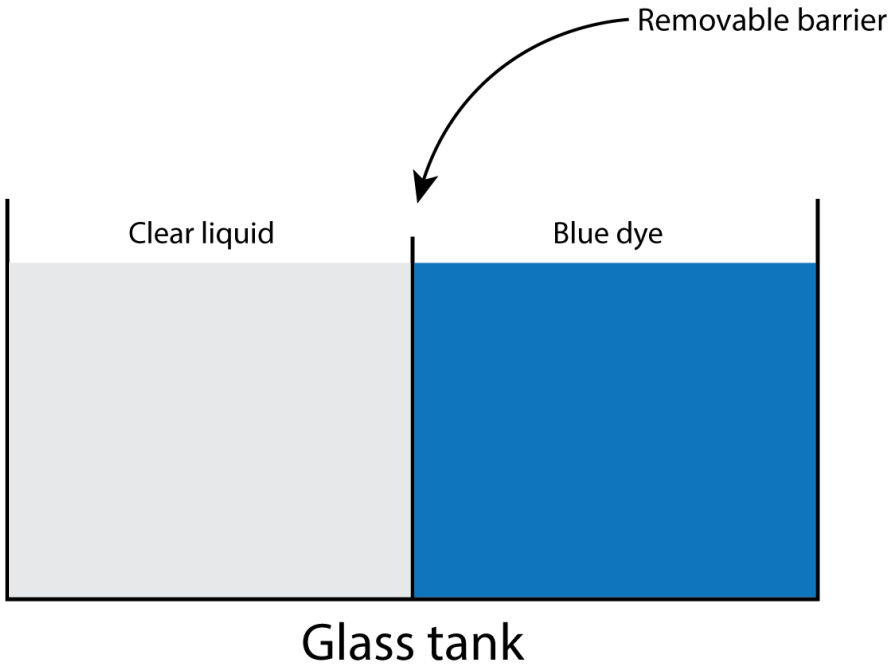
Which arrangement of these pseudo-molecules in a box is more entropic based on the definition you learned in high school or first-year? Does your friend agree with your decision? Diagrams like these are highly confusing and do not properly illustrate the concept of entropy. These diagrams are frequently used in course-required textbooks, though the authors are completely wrong to draw things this way. So, ignore them. Never fully trust what you read – even this lab manual! Only if an idea is experimentally verifiable should you entertain its merit. We hope to show how you might experimentally verify the concept of entropy in the following paragraphs so that you can decide for yourself if you believe us or not.

Entropy can be described in two ways. It has a thermodynamic definition, and a statistical mechanics definition. Both are certainly connected, and are just different ways of thinking about the same thing. We will give an example of each in the following sections, starting with the statistical mechanics definition.

What distinguishes the past from the present from the future? If you were shown a video of somebody doing a cannonball into a swimming pool in reverse (without being told it was in reverse), with a person flying out of a pool, and all of the water droplets re-organizing themselves from the air into a neat flat surface on the pool, you would have an intuitive sense that it is ***highly unlikely*** that this event would occur spontaneously in a 'forward' manner. The video ***must*** be in reverse! In order for this reverse-cannonball event to occur in a forward manner, kinetic energy in the atoms composing the ground would have to spontaneously 'concentrate' and enact a force on the water sitting on top of it, fly into the air, condense into droplets, and fall back neatly into the pool. We never observe things like this happening because the probability of it is so phenomenally small (though it *technically* is possible). On the macro-scale (for the time being, think of this as things that we can *see*), this is a product of the fact that anything we can see is made up of huge numbers of smaller particles. The more particles something is made up of, the lower the probability there is of observing these 'atypical fluctuation events'. Probability is at the root of entropy, and one concept to keep in mind in the following sections is that entropy is not driven by a 'magical mysterious force', but rather what we observe is simply the most likely scenario to occur.

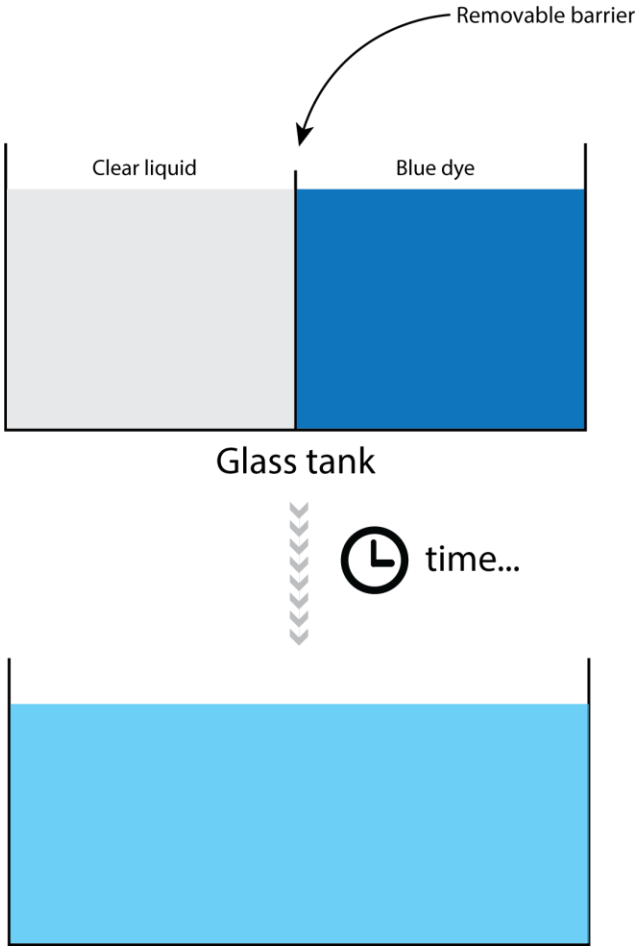
Imagine you are living back in the 1800s. You have no idea of any of the scientific progressions or achievements that have been made since then. At this point, there is serious doubt even about the existence of atoms. Let's make up a completely random name for ourselves… how about… Rudolf Clausius. We are trying to understand the properties of the things around us. What makes me, me? Why is this wall hard when I hit my head against it, and why can I swipe my hand through the air freely? Let's set up an experiment and observe these 'things' around us in their natural habitat. What will they do when left up to their own devices? Are they on some sort of mission (**no**, but we don't know this yet)? We will build a rectangular glass tank with a removable barrier in the middle of the tank whose purpose is to prevent two liquids from

mixing. In the left side, we will pour in water. In the right side, we will pour blue dye. The liquids cannot mix until we remove the barrier, as in the figure below.



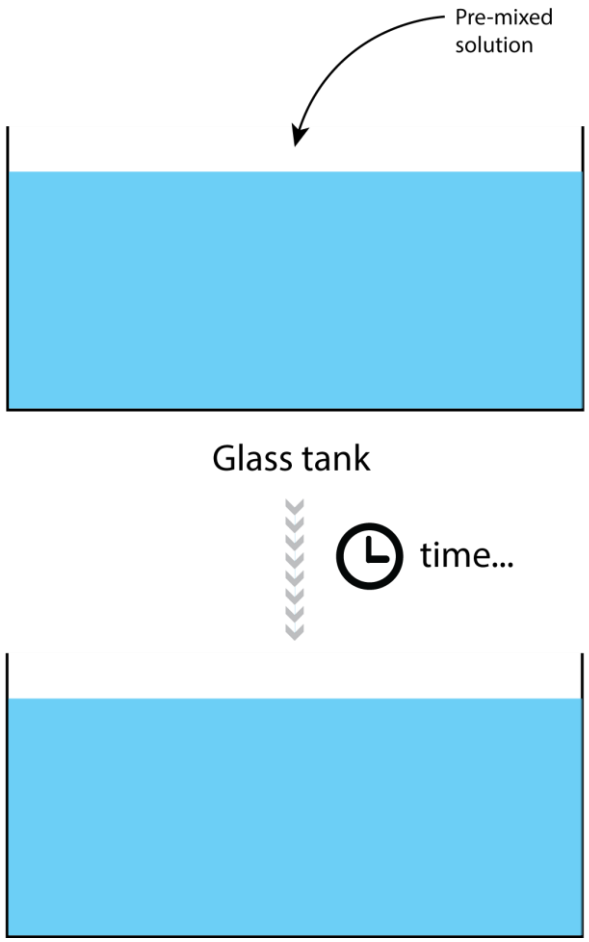
Now, lets see what these things (water and dye) ***do*** when we remove the barrier. We will not instruct them to do anything, after we remove the barrier, we will hold our breath, stop thinking thoughts, and they will do whatever they do of their own volition.

We remove the barrier and … wow … the liquids begin to mix slowly at the middle interface. We never told them to do that! It occurred ***spontaneously***. The clear water starts to turn slightly blue, and the blue dye begins to fade in color slightly. Eventually, the motion of these tiny 'things' causes the entire tank to become a slightly more pale blue than the original dye color, until fully mixed, as below:

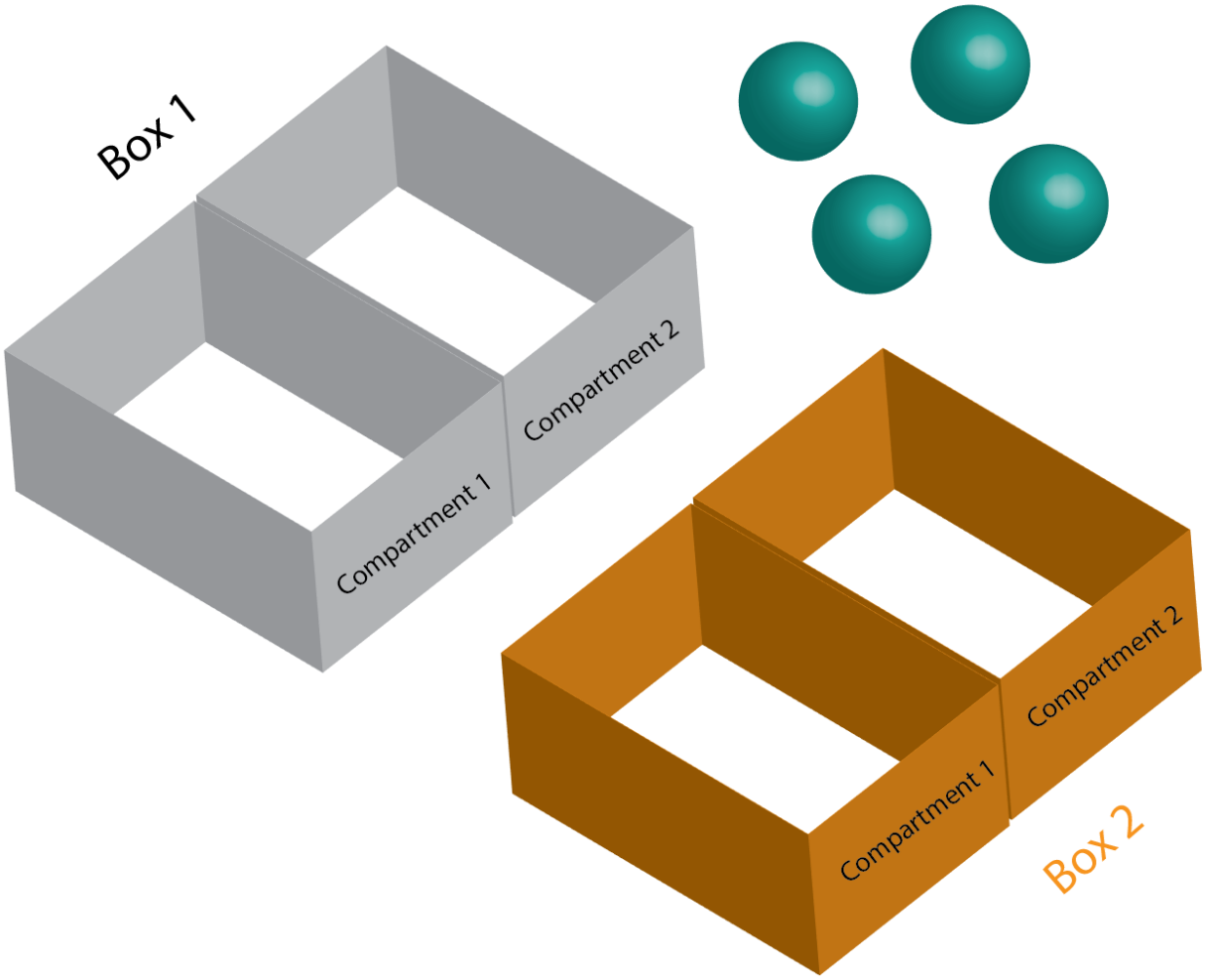


Okay! Awesome, but who says we can't do things in reverse? Who says the dispersion of these tiny things that make up these two liquids can't start from inter-mixed and unmix themselves? Let's try it! We pre-mix our dye and water with a big stirring stick, and we sit and watch it. Surely it will form back to half-and-half on either side of the tank right? We watch for hours and hours, and hours, and hours, and hours, and hours, and… it never happens. Incredible! We have just discovered the second law of thermodynamics. It turns out that in order for us to observe all of the dye molecules self-associating and all of the water self-associating on either side of the tank, it is possible (due to 'quantum fluctuations'), but we would need to wait several times longer than the age of the universe. The molecules in this tank arranged themselves in a way that allows them to ***maximally disperse their energy***. Entropy of a system is always

increasing. ‘Okay – but you never told me what entropy is…’ (that is me impersonating you). At least now you have a sense that there is some directionality to entropy, and perhaps how it is linked with time. But now let’s derive a formal definition.

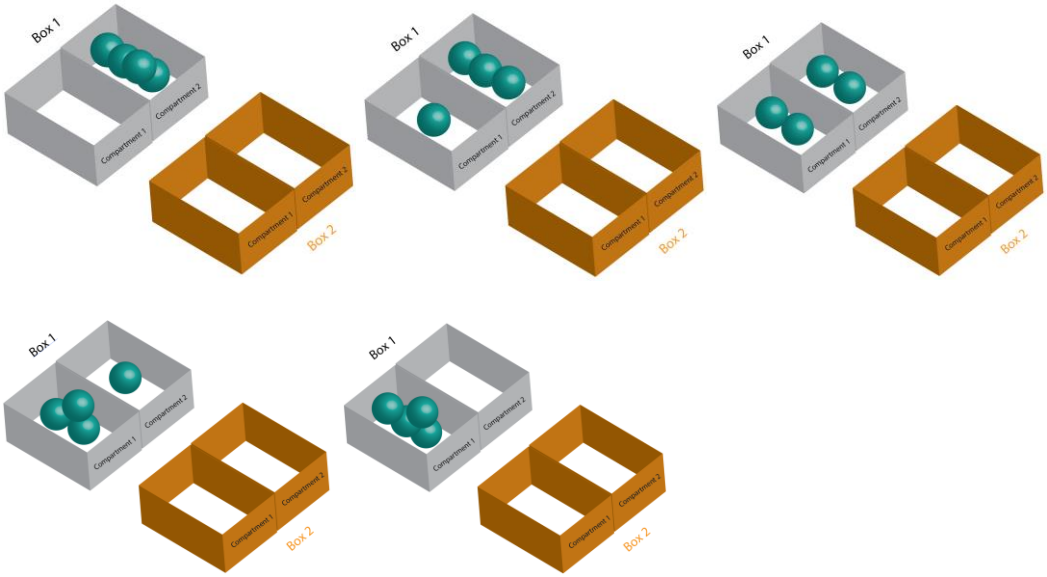


Let’s look at some examples with probability. Imagine two containers that can hold plastic balls in two positions. The containers are side by side, and we are free to move the balls around however we please. We can place as many balls into any compartment we like. I chose four balls for simplicity’s sake, but really we could have any number of balls.



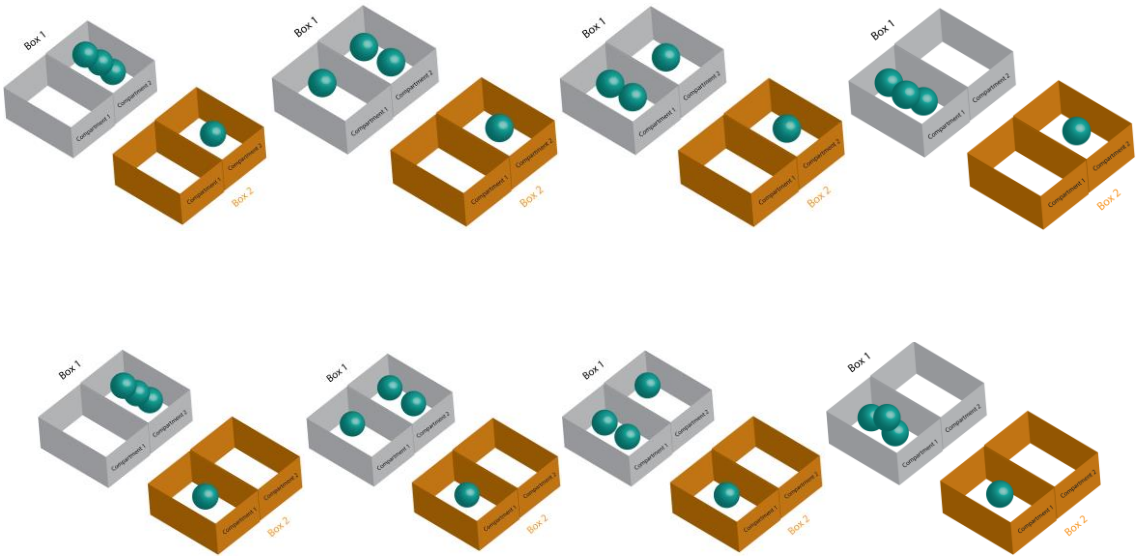
The question now is, how many ways can we stack the balls between these two boxes? Let's list out all the ways we can do it starting by only putting balls in box 1.

- Position 1: 4 balls
- Position 1: 3 balls / Position 2: 1 block
- Position 1: 2 balls / Position 2: 2 balls
- Position 1: 1 block / Position 2: 3 balls
- Position 2: 4 balls



Thus, there are **5 possible ways** of arranging the balls to all be in a single box. Now, lets keep three balls in the first box, and allow one block to move over to the second box.

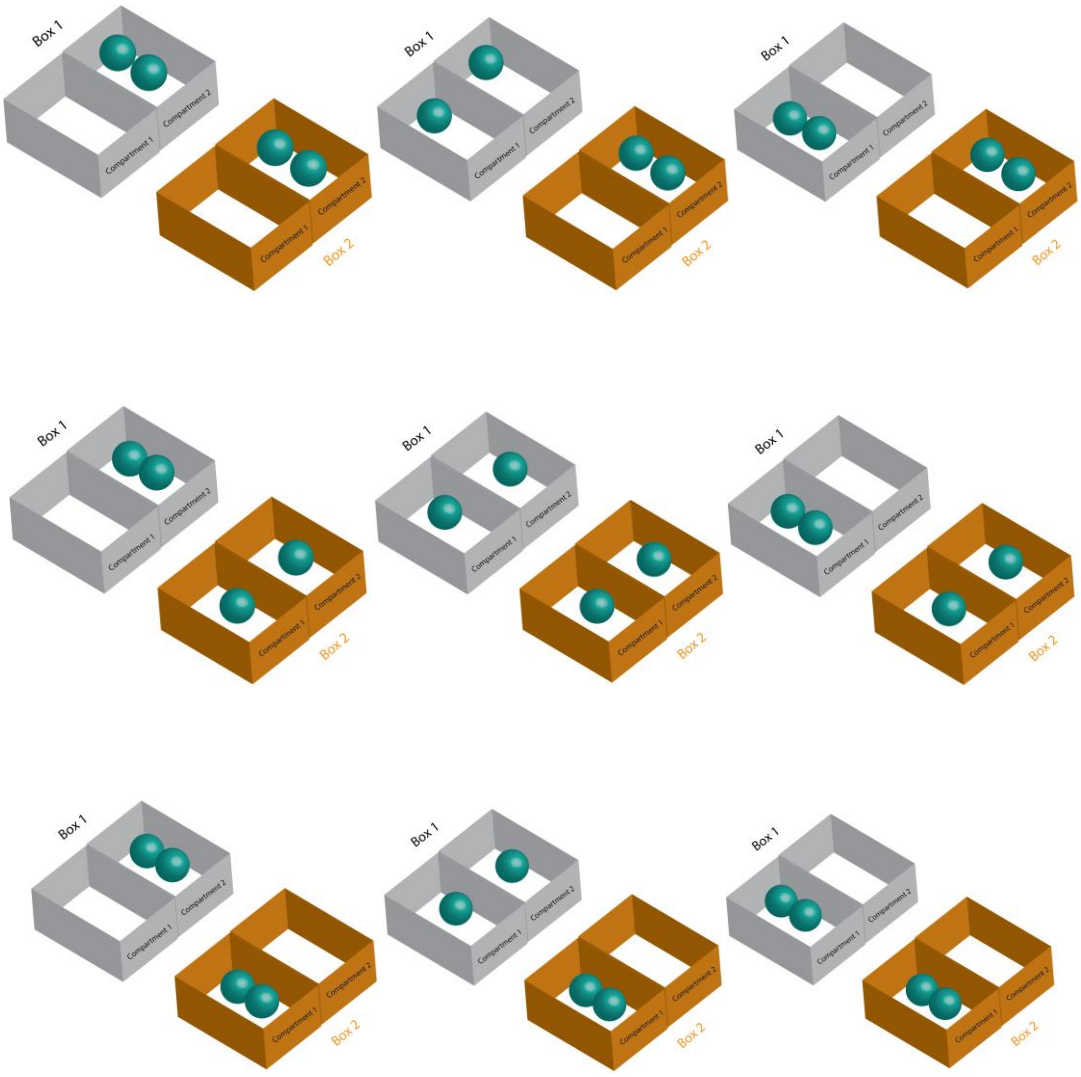
BOX 1	BOX 2
• Position 1: 3 balls • Position 1: 2 balls / Position 2: 1 block • Position 1: 1 block / Position 2: 2 balls • Position 2: 3 balls	• Position 1: 1 block • Position 2: 1 block



So, we see there are four ways of orienting three balls in the first box, and two in the second. Since we can have any combination of these, we have $\mathbf{4 \times 2 = 8}$ **possibilities**.

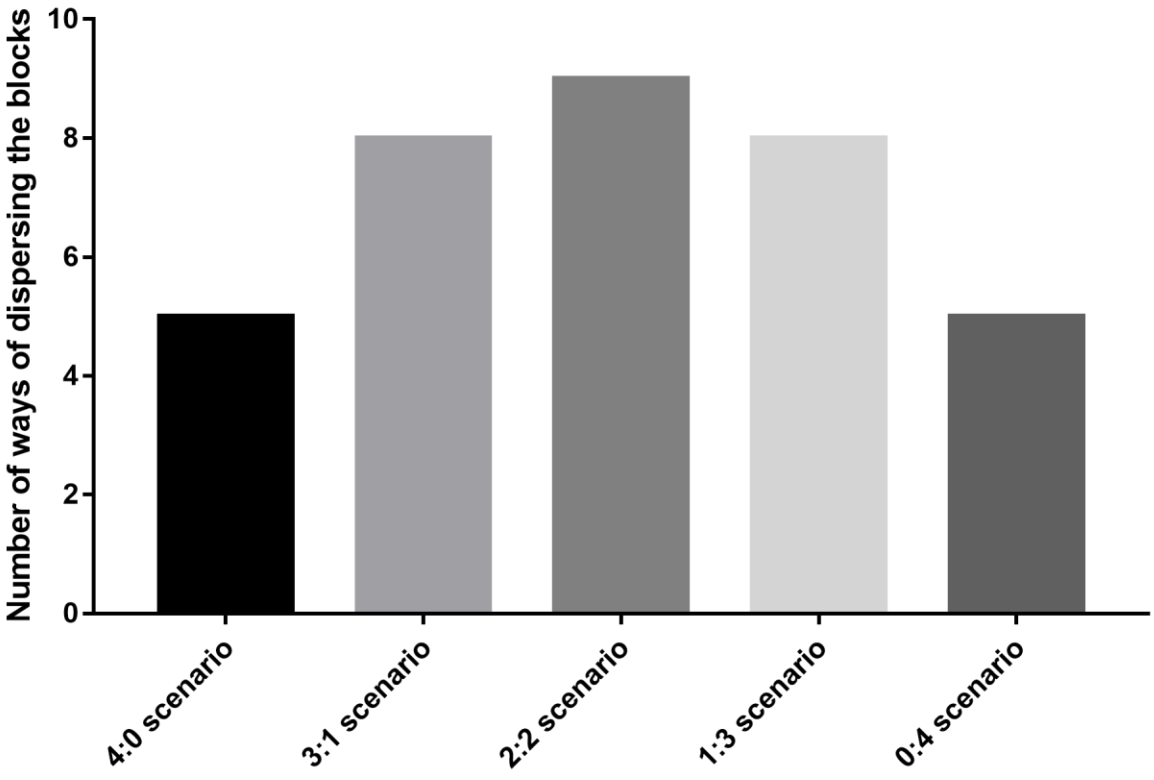
We are almost done counting balls. Let's try a final scenario where each box is allowed two balls.

BOX 1	BOX 2
• Position 1: 2 balls • Position 1: 1 block / Position 2: 1 block • Position 2: 2 balls	• Position 1: 2 balls • Position 1: 1 block / Position 2: 1 block • Position 2: 2 balls



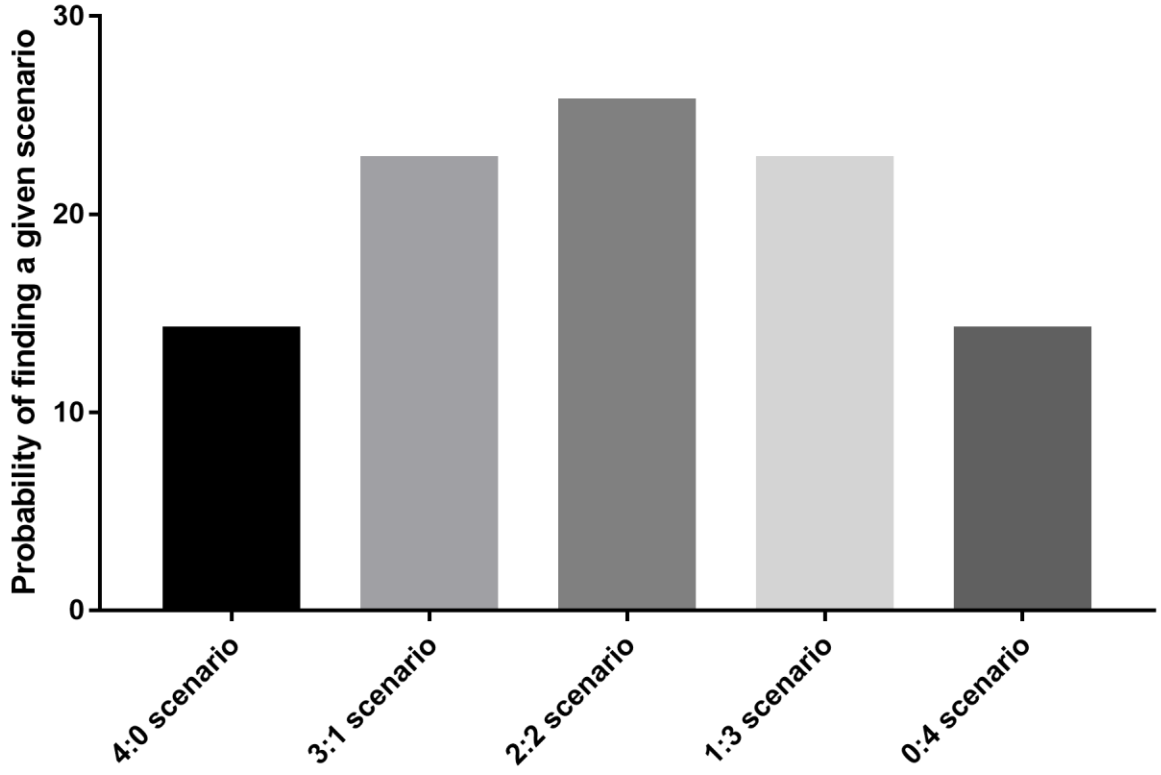
This time, it seems we have three possible ways of orienting the balls with two in each box. That means there are **3 * 3 = 9 possibilities**. Keep in mind that while we started by giving box 1 all four balls, we very well could have done this for box 2. This also applies in the 3/1 shared scenario, and the results are the same (5 possible ways, and 8 possible ways). We almost fully understand entropy! Just like… two more paragraphs, I promise.

Let's plot our results…



Woah! It looks like a distribution! Ok now what if we imagine we put a toddler in a room with these balls and just let them arrange them in one of these configuration as we walk out of the room. Can we predict the probability that when we re-enter the room after the toddler has

finished playing, that we will find a given configuration? We have to make one assumption; that **all of these arrangements are equally likely** (the toddler has no preference for one arrangement or another). With this assumption made, we just simply divide the number of ways by the total number of ways (5 + 8 + 9 + 8 + 9 = **35**) to obtain the probability, and change our y-axis to a percentage:



What we find is that the **HIGHEST PROBABILITY**, when we re-enter the room, is that the balls will be **as dispersed as possible** between the two boxes. Now, I just need to explain Planck's constant to you, and everything will make sense.

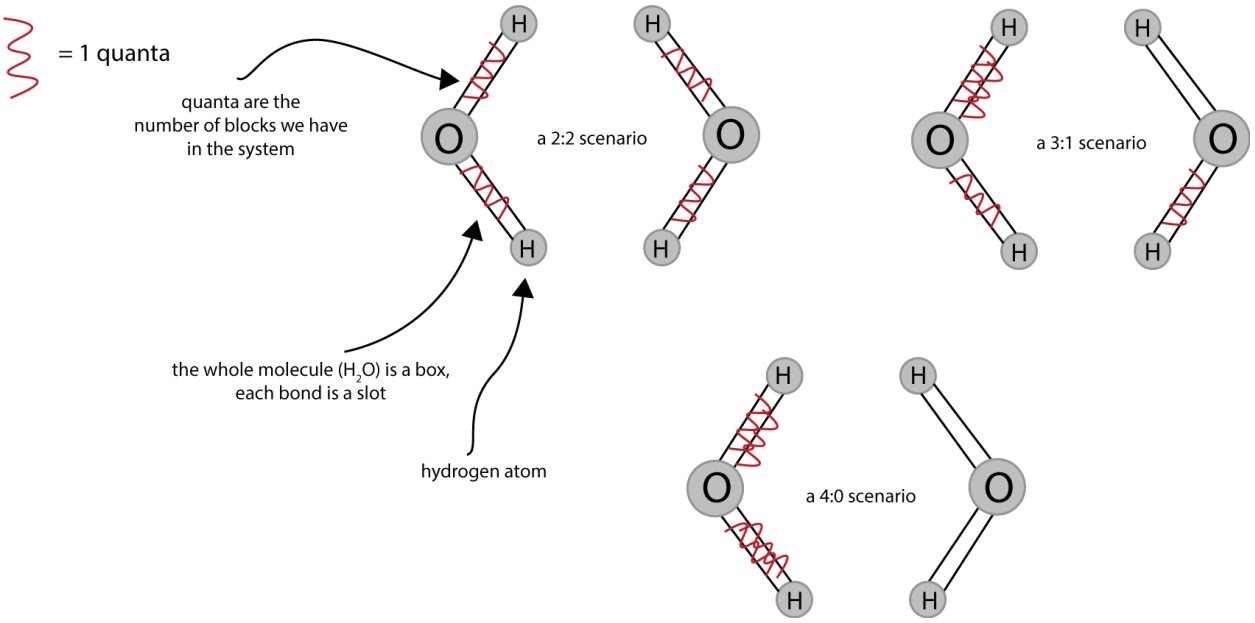
What is Planck's Constant Anyways?

If you're like me, you spent your undergraduate degree plugging the number **6.62607004 × 10^{-34} m^2 kg / s** into equations that you memorized, but never really understood what the number

meant, or what the quantity was that you were arriving on. Max Planck was studying why metal glows red when it is hot due to an industrial battle between Germany and England to become the producer of higher-efficiency lightbulbs (competition breeds success!). He developed a law called the 'Blackbody Radiation Law', which predicts the frequencies of light that are emitted by an object at different temperatures. For example, at room temperature, a piece of metal sitting on a desk emits radiation mostly in the form of infra-red electromagnetic radiation, which is invisible to our eyes, but we can detect it with an infra-red camera. When you heat it up, we can see it glow red – the frequencies of the light that the metal emits have changed (from infra-red/invisible to red/visible)! Planck's major breakthrough came when he 'realized' (as more of a mathematical trick) that the body ***could only emit energy in discrete 'packets' of energy***. These packets are called **quanta**. These packets exist in multiples of Planck's constant.

So, it turns out that Planck discovered our world is truly discontinuous! Now, the next time some hipster tells you that vinyl records sound better than CDs 'because we are analog creatures, dude… we don't live in a digital world, dude!', you can say 'no! the energy inside of all things is quantized, or discrete, fool!' (though I still think vinyl records sound better than CDs).

Now for the best part. From our discussion about entropy before, replace the word 'box' with molecule, and the word 'balls' with quanta. Each 'slot' in the box is a chemical bond on which quanta can reside. This is precisely how energy is shared between molecules. If a molecule has more quanta (discrete energy 'balls'), it is hotter. If it has fewer, it is colder. When we place two objects with varying energy levels in close proximity, we find that they always proceed to a state where the quanta are **as dispersed as possible**. This is why energy flows from hot things to cold things. Energy does not (principally on large scales) flow/concentrate from a colder object to a warmer one.



Speaking in terms of statistical mechanics, we can look at the equation engraved on Ludwig Boltzmann's grave. Statistical mechanics is just a way of describing the behaviour of macro-systems by accounting for the energy dispersion of the micro-parts (what is the **pressure** (macrostate) of an air canister (macrosystem), given its **microscopic components** (air molecules) have some given energy distribution (microstates)?). In our case, each 'WAY' (note the spelling of the word WAY, and the fact that it begins with the letter ***W***) of arranging the quanta is a microstate.

At the top of Ludwig Boltzmann's grave, we can see the equation engraved;

$$S = k \log W$$

What are each of these terms? S = entropy, *k* = Boltzmann's constant, and W = the number of **W**AYS (microstates) in which the system (and thus arrangement of the molecules) can be different microscopically. Even if you're not a fan of math, this equation is very elegant and simple. If we want to increase entropy (***S***), what do we have to do? Well, ***k*** is a constant, so ***W*** must increase.

The number of microstates must increase. The universe is always moving towards the scenario with the highest number of ways of dispersing quanta.

If we were to increase the number of boxes, positions the boxes could hold balls, and increase the numbers of balls that we had, we would find that it would become increasingly unlikely that the balls 'concentrate to one side'. While in our example we have a 14% chance that all balls are found in one box, as the number of balls/quanta increases, and the number of positions to hold increases, this number becomes infinitesimally small relative to the number of ways you could disperse the balls roughly equally. This is why large objects (composed of trillions and trillions of atoms, with tons of energy) do not spontaneously heat up, or perform strange feats like knock water off the ground into a perfect cannon-ball reverse splash. The probability of this happening becomes a number with a very large negative exponent (such as $10^{-250000}$; a very

long decimal number); a number that would require us to wait longer than the age of the universe before we could observe.

So how are the number of microstates increasing in a tank of mixing blue dye and water? When we remove the barrier, the molecules are able to move freely in a larger volume, allowing them to disperse their energy more readily.

The first-ever written definition of the second law of thermodynamics was '*heat does not spontaneously flow from a colder body to a hotter one*', set forth by a man, coincidentally named Rudolf Clausius. We all know that when we put a pot of water on a hot stove, the heat in the pot (thermal energy, in the form of atoms and molecules of the pot jiggling, and the rotational, translational, and vibrational kinetic energy of the water molecules) does not make the burner observably hotter. This is a 'well, duh' moment and an 'I can't believe it took us thousands of years to prove that hot things make cold things hotter, but cold things don't make hot things even hotter'. The best part about these phenomena is how simple and intuitive they are when described, but how much information they give us about the inner workings of physics and Nature.

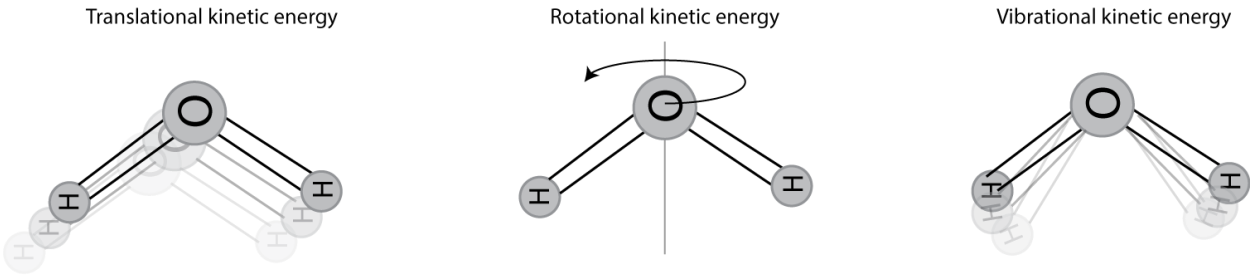
The take away message of this is that particles in any system in our universe will always disperse their energy as opposed to 'concentrate' it – not due to a mysterious force, but rather it is the most probable scenario. Molecules diffuse to be as dispersed as possible, providing them more ways to disperse their energy. Energy in the burner will disperse itself into its surroundings (the pot, water, food, and air molecules).

It is highly recommended that you read **"Disorder – A Cracked Crutch for Supporting Entropy Discussions"** written by Frank Lambert. It is freely available (just search the title on google). It is a short article that provides eight great examples of how entropy truly operates.

Enthalpy and Temperature Made Cool

I will not talk much about heat, other than to say it is in fact different from temperature. **Temperature is a *measure* of how hot something is *on average***. Temperature measures the average kinetic energy of a system, or body. Feynman often refers to how hot something is by how fast the atoms in the 'something' are 'jiggling'. Kinetic energy can take on the form of

rotational kinetic energy (how fast is the body spinning?), **translational kinetic energy** (how fast are the atoms moving up, down, left, right, frontwards and backwards?), and **vibrational kinetic energy** (how fast are the bonds in the atoms vibrating back and forth?). Temperature is thus a measure of the average of these three quantities.



Now for enthalpy. Enthalpy is an easier concept to grasp; it is closely related to the **internal energy** of the system. The internal energy is the total of all possible kinds of energy present in a substance. Probably the easiest way to think about enthalpy in the case of a solvated protein is as a measure of the **latent heat content ('hidden energy')** in the inter-molecular bonds around the protein.

$$H = U + PV$$

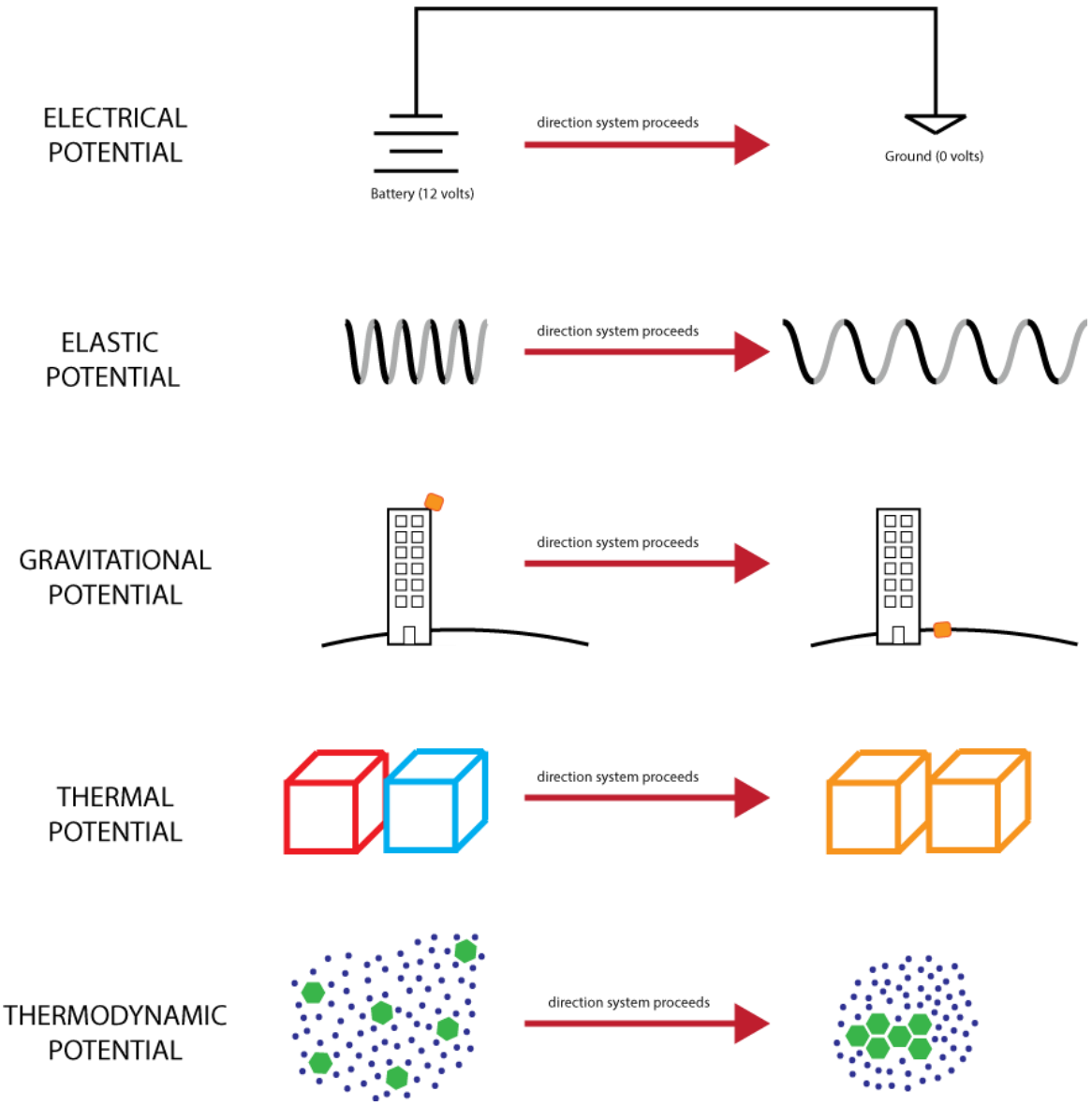
Since crystallization of our protein is an equilibrium between solid and liquid phases, there is a negligible change in pressure and volume, so we can assume. As such;

$$\Delta H \cong \Delta U$$

Intuitively, this enthalpy value is represented by the inter-molecular interactions between protein and water in the solvated phase, and between proteins in the crystalline (solid) phase.

Gibbs Free Energy

An important phenomena to understand is the concept of *potentials*. A potential indicates which direction a system will proceed on its own. As below,



If you have an electrical potential (also called voltage), we will get current flowing from high voltage to low voltage.

If you have a gravitational potential, we can tell which direction our system will proceed; our box will fall off the roof (think high altitude to low altitude).

If you have a thermal potential, we will get heat flowing from a hot body to a cold body (high temperature to low temperature).

If we have an elastic potential (compressed spring), it will proceed towards an uncompressed state (high tension to low tension).

A thermodynamic potential (high energy to low energy) is the difference in available energy between 'reactants' and 'products' of our chemical reaction. Potentials are the closest thing science has to a 'fortune teller'. It tells us whether something will happen by itself (spontaneously) or not. What are our reactants and products in this case? Well, our reactants are solvated protein, and our products are crystals. If we can calculate our **Gibbs free energy of the reactants,** and the **Gibbs free energy of the products**, taking their difference tells us whether our protein will crystallize or not! Let's look at what makes up the Gibbs free energy:

$$G = H - TS$$

You are an expert on enthalpy, entropy, and temperature now, so this should be a breeze for you. This equation says "*the available energy in your system is the difference of the enthalpy, and the entropy times how hot your atoms are, on average*". Or, more simply but less informative; "*energy equals energy minus energy*" Like I said above, if we take the difference of two systems, we can generally tell whether our reaction will proceed or not. This is because every system in the universe tends to want to proceed to a lower energy state. If what we are hoping to happen (crystals) have a lower energy state than what we currently observe (solvated protein), then it will proceed in that direction:

$$\Delta G = \Delta H - T\Delta S$$

Remember that the ***Δ*** symbol is called 'delta' and it means 'the difference'. Writing ***ΔS*** is the same as saying "entropy of the final system *minus* the entropy of the initial system". If the value for ***ΔG*** is negative, the reaction will proceed spontaneously. If ***ΔG*** is positive, it will proceed in reverse. That is,

$$if\ \Delta G < 0; spontaneous$$

$$if\ \Delta G > 0; not\ spontaneous$$

Sadly, in protein crystallography, with our current knowledge of how protein crystallize, **we can not know for certain** if the protein we are interested in will crystallize or not. It would be nice if we could calculate in advance whether the Gibbs free energy will be negative. If it is, then we are likely to get crystals. If not, then we won't bother; and we can try another technique to solve the structure. Only through hard work, lots of experimentation, and trial and error will we find out if crystals will form. Though, if you have an idea as to how to calculate this number before crystallization trials, you will become very rich, as the whole world will want to use your solution.

Crystallization Thermodynamics and Modifying Solubility

At the beginning of any crystallization experiment, proteins are solvated in solution. To form crystals, they must come out of solution – that means we have to create an environment for the proteins that causes them to preferentially form interactions with each other as opposed to interactions with solvent molecules (i.e. water). The solubility of the protein can be altered in different ways. One can make chemical modifications to the protein (mutating residues, modifying functional groups), and/or alter the environment of the protein (removing water from the system, changing pH, changing temperature, or adding precipitants). Do not view solubility as an 'entire new concept' which you must now understand too. Isn't solubility just the idea of changing phases from, say, aqueous protein… to solid? This change in states could easily be thought of again as governed by a thermodynamic potential. Solubility is no more than the difference in Gibbs free energy between phases (aqueous, and crystalline, or aqueous and aggregate). So, by changing the **solubility** of something, we are changing the **available energy** of the system.

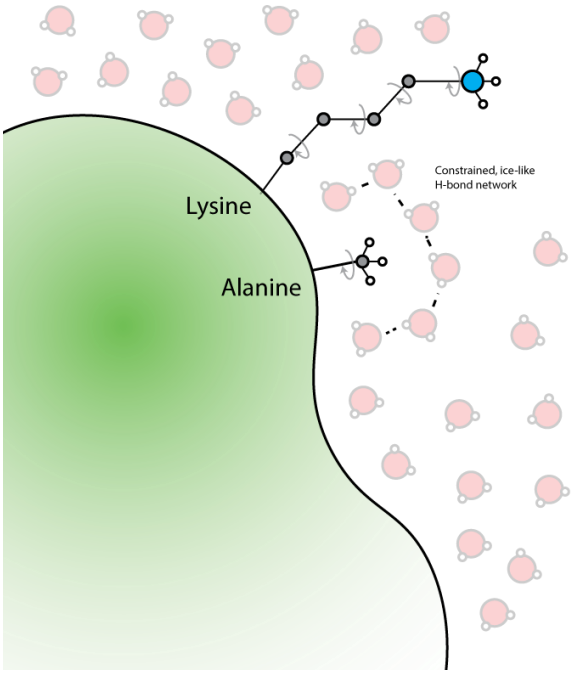
Assembly of protein molecules into crystals is a thermodynamic process that requires a net decrease in Gibbs free energy (***ΔG***). Two terms contribute to the change of ***ΔG*** – an enthalpic term (the systems internal energy, ***ΔH***), and an entropic term (energy dispersion, ***ΔS***):

$$\Delta G_c = \Delta H_c - T\Delta S_c$$

The enthalpy change associated with crystal formation is typically weakly negative and only contributes marginally to ***ΔG***. This is an indication that only weak intermolecular bonds form between protein molecules in the crystal lattice. The entropy change associated with protein crystal formation is actually a *decrease in entropy* – protein molecules suffer ***restrictions*** to their translational and rotational freedom, as well as conformational rigidity in loop secondary structures once assembled into a rigid crystalline environment, and this causes a large positive *destabilizing* effect to ***ΔG*** (subtracting a large negative value from enthalpy results in a positive value for ***ΔG***). Of course, you should now see, that suffering a restriction to translational and rotational freedom means the residues in the protein can no longer disperse their energy to the same degree. Before, they could adopt 5 or 6 rotamers, but now, in a crystal, perhaps only one rotamer is allowed. The number of microstates have decreased, meaning our entropy is lower. How do crystals then form if it seems thermodynamically impossible for them to do so? After all, didn't we say that systems have a tendency to proceed to a more dispersed energy state? The second law of thermodynamics states that entropy is always increasing! It is important to remember that crystallization occurs within a system that is surrounded by water; and that the total entropy of the system and the surroundings is what must obey the second law of thermodynamics. The entropic effect described above is actually counter-acted by an increase of entropy related to the dissociation of water molecules forming the *clathrate cage* (solvation shell) around the proteins. In simpler terms, the energy in the protein is becoming **less dispersed** through intermolecular interactions, and energy in the solvent is becoming **more dispersed** via dissociation from the protein. Thus, instead of the simplified equation above, we can describe the crystallization process by providing individual terms for protein entropy and solvent entropy.

$$\boldsymbol{\Delta G_c = \Delta H_c - T\left(\Delta S_{solvent} + \Delta S_{protein}\right)}$$

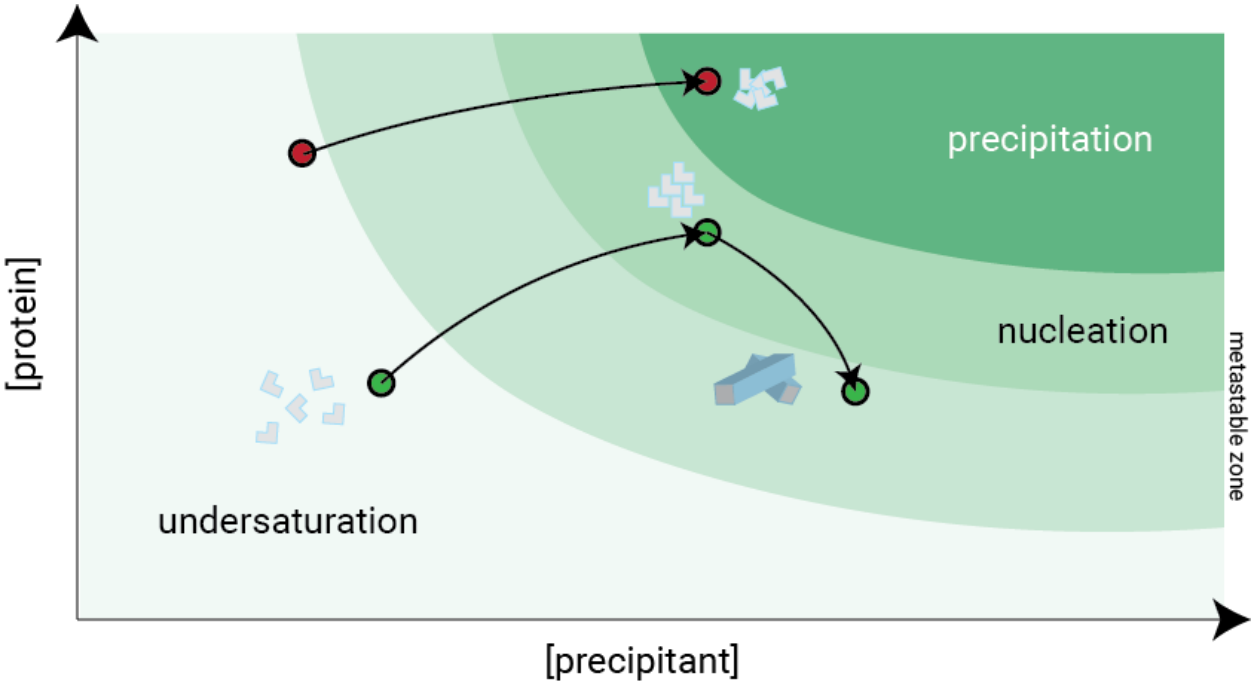
One can now imagine (since we have stated that entropy will always increase for a system and its surroundings) that the solvent entropy and protein entropy values are 'competing' in any given crystallization experiment. If the two values are of similar magnitude, small changes to either can drastically effect chances of crystallization. To give a practical example of this, in a technique referred to as 'surface entropy reduction', scientists mutate residues with high conformational entropy[4] (residues with a large degree of conformational freedom, which allows them to disperse their energy many different ways, i.e. adopt many rotamers) on the protein surface such as Gln, Lys, Glu, Asp, etc. to low entropy residues (typically Ala, Thr, Val). Such a change would reduce the entropic loss from residue motion and thus $\boldsymbol{\Delta S_{protein}}$ would become smaller. In this way, we try to minimize the entropic losses that are caused by conformational restriction of labile regions (side chains, loops) as a way of achieving a negative value for $\boldsymbol{\Delta G_c}$.



Which residue above has a higher number of **w**ays of orienting itself? What does this say about its entropy? Which residue causes more entropic restrictions to the water molecules surrounding the protein? Lysine has a higher conformational entropy. By mutating a lysine to an alanine, we are effectively making it less unfavourable (double negative is intentional here) for crystals to form; as should this protein crystallize, it would suffer overall less conformational restriction if

there was an alanine in place of the lysine. Additionally, we are raising the entropy of the water (upon crystallization) by giving the individual molecules the ability to disperse their energy more freely throughout the entire volume of the drop – as they are no longer constrained within ice-like cages around the hydrophobes. We are effectively trying to bias our potential to be in a certain direction (towards crystallization). We are asking; "*How energetically unfavourable can we make the solvation state whilst still retaining the protein fold, and not causing aggregation*?".

Unfortunately, achieving a negative ***ΔG*** is not the only requirement for crystallization. Two other processes – nucleation and growth kinetics – are required as well. To better understand how crystals might form, we frequently describe a 'crystallization phase diagram', shown below. The phase diagram represents how a protein's solubility changes with respect to its concentration and the concentration of the active precipitant. A **precipitant** is a molecule that reduces the protein's solubility, typically by competing with it for solvent molecules. Frequently used precipitants in crystallography experiments include salts (i.e. NaCl), organic solvents (ethanol, MPD), and polymers (polyethylene glycol/PEG).



Crystallization diagram. *Note; this 'phase' diagram is in violation of a true thermodynamic phase diagram because it includes kinetic information (nucleation zone), though we still include it

because it aids in the interpretability of the experiment. The red course represents a crystal drop that is set up whereby the concentration of protein and precipitant is too high, leading to instability and precipitation of the protein in unordered aggregates. The green course represents an ideal scenario where (after vapor diffusion) the system proceeds into a metastable state, nucleation occurs, and crystals begin to grow.

Reducing the solubility of the protein to a sufficient degree places the system in the metastable zone. From here, the system will require an 'activation event' known as **nucleation** in order to overcome the activation energy and begin to form a crystal. Collisions between protein molecules is a common event in such highly concentrated protein solutions. Should a collision event occur that forms favourable contacts between two protein molecules, the enthalpy associated with the resultant intermolecular bonds may overcome the entropic 'loss' of translational and rotational freedom – lowering the overall energy of the system, forming a nucleation site. The nucleus is susceptible to dissociation however, be it from spontaneous dissociation or via collisions with other particles. Though with enough favourable collision events, a stable nucleus will form that allows for periodic, long range growth of a crystal.

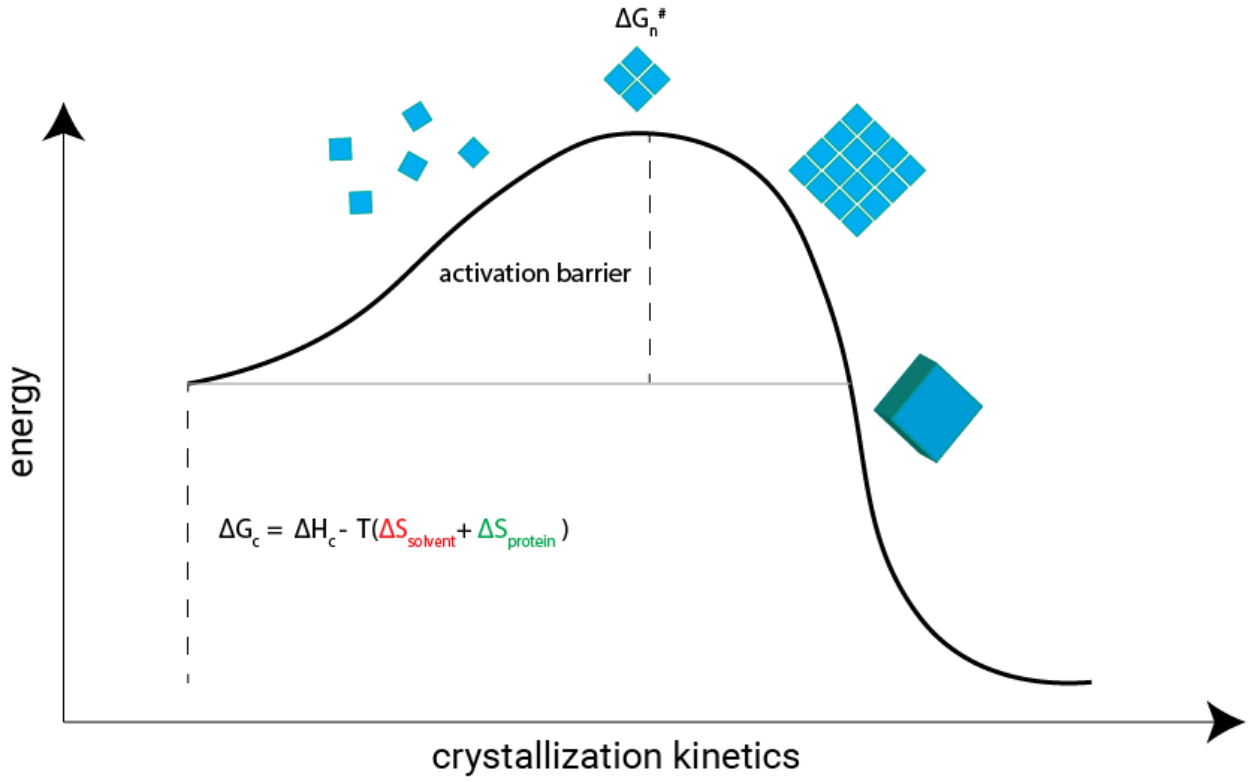


Figure 3.2. Nucleation energy. The energy difference between the reactants (free-floating protein) and products (protein crystal) is described by the equation noted on the graph. The activation barrier is overcome by a spontaneous nucleation event whereby protein particles collide in orientations that form intermolecular bonds and cause a net decrease to the energy of the system.

Setting up Crystal Screens Manually

We will be setting up crystal screens on two proteins – your anti-CRISPR as well as egg-white lysozyme. In practice, crystal screens are set up in high-throughput by a robot – typically in 96-well trays where each **well** houses a different crystallization 'condition', consisting of different buffers, additives, and precipitants. We will be using the Gryphon to set up a **sparse matrix** screen for your anti-CRISPR. The term 'sparse matrix' indicates that the screen covers a broad range of conditions and includes conditions that have allowed many proteins to crystallize previously – though they have not been selected through any rational experimental design.

While automation allows for many trays to be set up in a short period of time, it is also useful (and important) to be able to manually set up crystal trays for purposes of optimizing existing crystal hits. 'Optimization' refers to improving the crystal quality (packing, size, etc.) of your crystal hits by slightly modifying components of the crystallization condition.

We will begin by setting up a sparse matrix lysozyme screen manually, as it yields very large organized crystals within only several hours.

Each tray supplied is a hanging drop tray with 24 wells in a 6 x 4 format. The sparse matrix screen consists of 96 conditions which we have split up into four sets of 24. Each pair of students will set up a single tray using the 24 conditions that have been assigned to them.

To set up a hanging drop tray:

1. Fill the reservoir with the appropriate screen (500 µL)
2. Onto a circular cover slip, place 3 µL of the condition (from your reservoir).
3. Place 3 µL of protein into the drop on the cover slip, to make a 6 µL drop that is a 50-50 solution of the condition and protein.
4. Carefully grease the outside of the tray reservoir, and gently place the glass cover slip upside-down (so the drop faces inwards toward the reservoir) over the grease.
5. Gently push the coverslip down to create an air-tight seal using the grease.
6. Repeat this procedure until all 24 conditions are done.

Setting up Optimization Trays for Lysozyme

Lysozyme is a protein that crystallizes rapidly into very large, beautiful crystals. For this reason, it is great for crystallography demonstration purposes as you will be able to set up crystal trays and see crystals within only a few hours. Half of the class will be responsible for setting up crystal trays using Procedure A (next page), and the other half will use Procedure B (next next page). Consult with your TA if you are unsure how to set up the crystal trays.

Lysozyme Optimization – Procedure A: [protein] vs pH

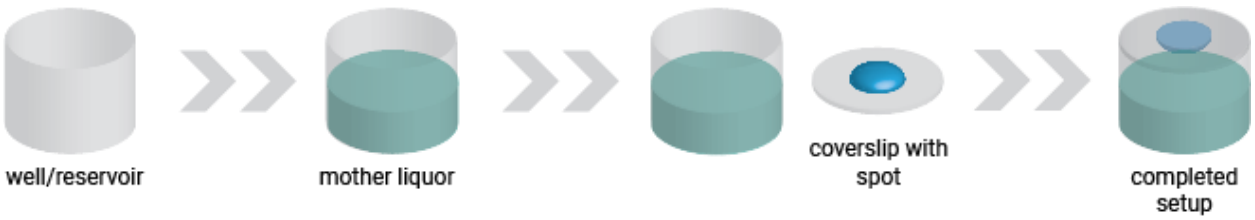
Set up your plate according to the layout below. The reagents coloured in orange constitute what is known as the 'reservoir solution' or 'mother liquor'. The bolded reagents are to be added to the cover in the 'drop' or 'spot', prior to sealing the well with grease.

1M NaCl 50mM NaAc pH = 3.5 **1uL protein** **9uL liquor**	1M NaCl 50mM NaAc pH = 4.0 **1uL protein** **9uL liquor**	1M NaCl 50mM NaAc pH = 4.5 **1uL protein** **9uL liquor**	1M NaCl 50mM NaAc pH = 5.0 **1uL protein** **9uL liquor**	1M NaCl 50mM NaP pH = 6.0 **1uL protein** **9uL liquor**	1M NaCl 50mM NaP pH = 7.0 **1uL protein** **9uL liquor**
1M NaCl 50mM NaAc pH = 3.5 **3uL protein** **7uL liquor**	1M NaCl 50mM NaAc pH = 4.0 **3uL protein** **7uL liquor**	1M NaCl 50mM NaAc pH = 4.5 **3uL protein** **7uL liquor**	1M NaCl 50mM NaAc pH = 5.0 **3uL protein** **7uL liquor**	1M NaCl 50mM NaP pH = 6.0 **3uL protein** **7uL liquor**	1M NaCl 50mM NaP pH = 7.0 **3uL protein** **7uL liquor**
1M NaCl 50mM NaAc pH = 3.5 **5uL protein** **5uL liquor**	1M NaCl 50mM NaAc pH = 4.0 **5uL protein** **5uL liquor**	1M NaCl 50mM NaAc pH = 4.5 **5uL protein** **5uL liquor**	1M NaCl 50mM NaAc pH = 5.0 **5uL protein** **5uL liquor**	1M NaCl 50mM NaP pH = 6.0 **5uL protein** **5uL liquor**	1M NaCl 50mM NaP pH = 7.0 **5uL protein** **5uL liquor**
1M NaCl 50mM NaAc pH = 3.5 **7uL protein** **3uL liquor**	1M NaCl 50mM NaAc pH = 4.0 **7uL protein** **3uL liquor**	1M NaCl 50mM NaAc pH = 4.5 **7uL protein** **3uL liquor**	1M NaCl 50mM NaAc pH = 5.0 **7uL protein** **3uL liquor**	1M NaCl 50mM NaP pH = 6.0 **7uL protein** **3uL liquor**	1M NaCl 50mM NaP pH = 7.0 **7uL protein** **3uL liquor**

Lysozyme Optimization – Procedure B: [protein] vs [precipitant]

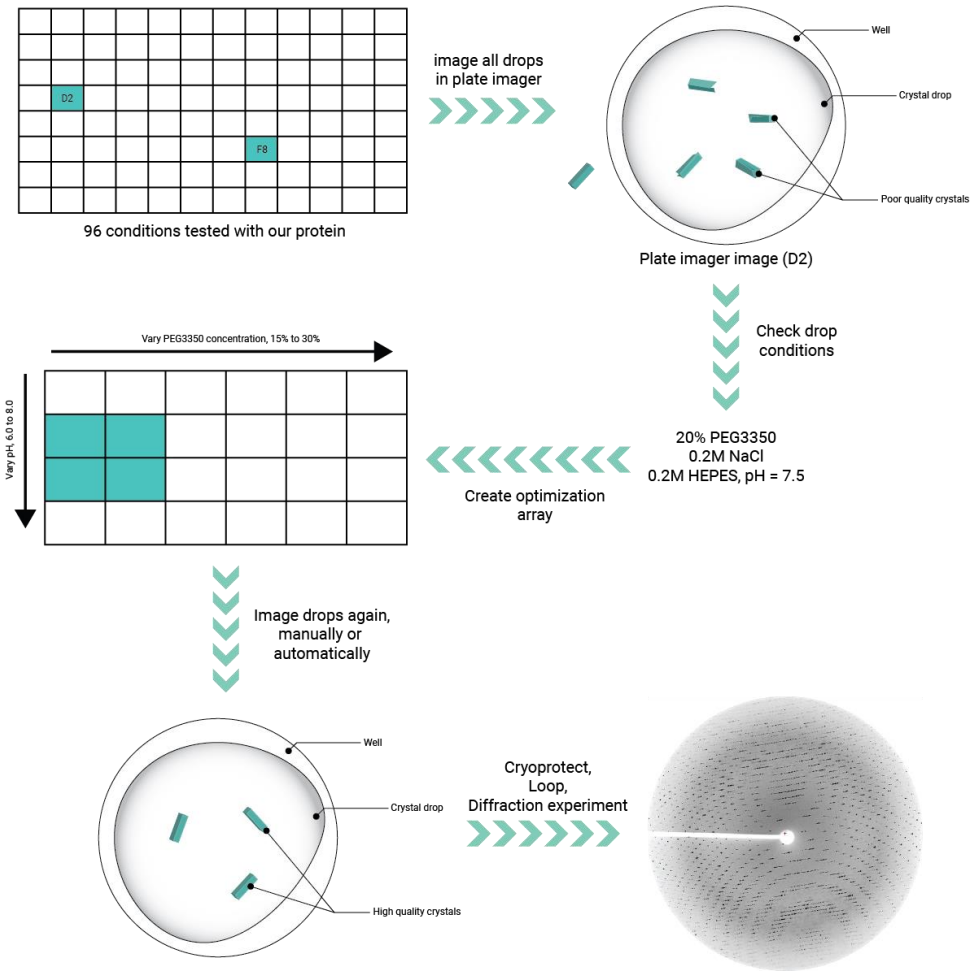
Set up your plate according to the layout below. The reagents coloured in orange constitute what is known as the ‘reservoir solution’ or ‘mother liquor’. The bolded reagents are to be added to the cover in the ‘drop’ or ‘spot’, prior to sealing the well with grease.

0.4M NaCl 50mM NaAc pH = 4.5 **1uL protein 9uL liquor**	0.6M NaCl 50mM NaAc pH = 4.5 **1uL protein 9uL liquor**	0.8M NaCl 50mM NaAc pH = 4.5 **1uL protein 9uL liquor**	1M NaCl 50mM NaAc pH = 4.5 **1uL protein 9uL liquor**	1.2M NaCl 50mM NaAc pH = 4.5 **1uL protein 9uL liquor**	1.4M NaCl 50mM NaAc pH = 4.5 **1uL protein 9uL liquor**
0.4M NaCl 50mM NaAc pH = 4.5 **3uL protein 7uL liquor**	0.6M NaCl 50mM NaAc pH = 4.5 **3uL protein 7uL liquor**	0.8M NaCl 50mM NaAc pH = 4.5 **3uL protein 7uL liquor**	1M NaCl 50mM NaAc pH = 4.5 **3uL protein 7uL liquor**	1.2M NaCl 50mM NaAc pH = 4.5 **3uL protein 7uL liquor**	1.4M NaCl 50mM NaAc pH = 4.5 **3uL protein 7uL liquor**
0.4M NaCl 50mM NaAc pH = 4.5 **5uL protein 5uL liquor**	0.6M NaCl 50mM NaAc pH = 4.5 **5uL protein 5uL liquor**	0.8M NaCl 50mM NaAc pH = 4.5 **5uL protein 5uL liquor**	1M NaCl 50mM NaAc pH = 4.5 **5uL protein 5uL liquor**	1.2M NaCl 50mM NaAc pH = 4.5 **5uL protein 5uL liquor**	1.4M NaCl 50mM NaAc pH = 4.5 **5uL protein 5uL liquor**
0.4M NaCl 50mM NaAc pH = 4.5 **7uL protein 3uL liquor**	0.6M NaCl 50mM NaAc pH = 4.5 **7uL protein 3uL liquor**	0.8M NaCl 50mM NaAc pH = 4.5 **7uL protein 3uL liquor**	1M NaCl 50mM NaAc pH = 4.5 **7uL protein 3uL liquor**	1.2M NaCl 50mM NaAc pH = 4.5 **7uL protein 3uL liquor**	1.4M NaCl 50mM NaAc pH = 4.5 **7uL protein 3uL liquor**



Setting Up a Sparse Matrix Screen Using the Gryphon

In practice, we have no way of telling what conditions may allow protein crystals to form. For this reason, we set up many conditions at once in high throughput in the hopes that one of them may facilitate crystallization. Using your purified protein (anti-CRISPR), we will be setting up one or several **sparse matrix screens** and placing them in a plate imager that records images of the crystal drops over a few weeks. In linear algebra, a sparse matrix is one in which most of its entries are zeros. This is funny, because it is reflective of the success of most crystallization trials. By consulting the images, we can easily check if crystals grew and try to reproduce the crystals by making the buffers corresponding to that crystal drop, and setting up the drop again. Some crystals may be of poor quality, but by altering drop conditions and seeding, we can **optimize** our crystals diffraction quality. If you are successful in obtaining crystals, we will also set up optimization trays.



Your TA will be operating the Gryphon, so you are only responsible for providing him or her with your ACR protein at an appropriate concentration.

Analyzing Crystal Drops

We need to look at each of our lysozyme drops to see if any crystals formed. Even if we do not see crystals, other species that form within the drop (i.e. precipitates, phase droplets, whether or not the drop is still clear, etc.) can tell us a lot of information about how we should proceed with our crystallization trials. Hampton Research has excellent documentation on the types of species you may find in a crystal drop, and what it may mean for your crystallization experiment. The figure below is adopted from the Hampton Research website:

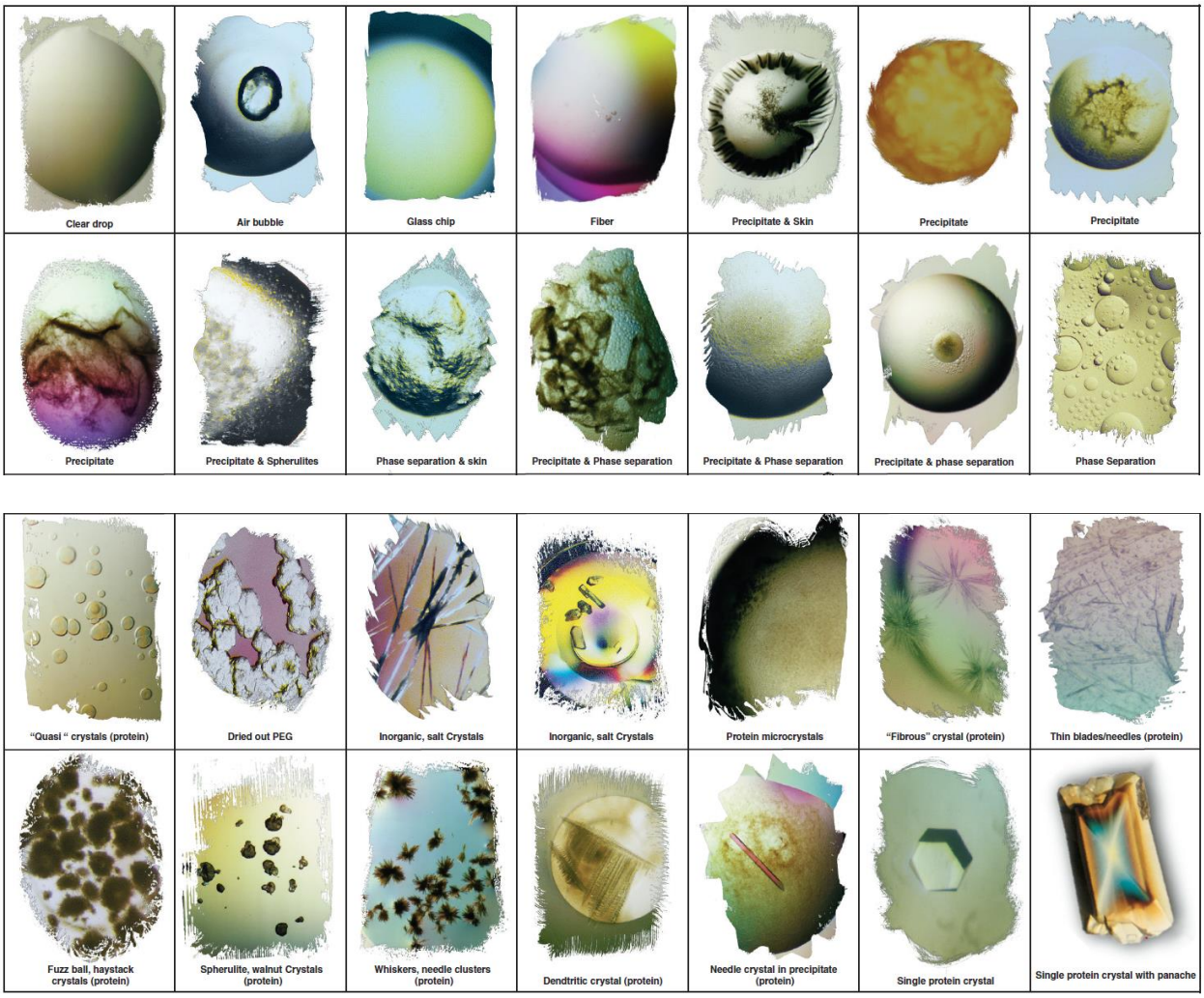


Figure 3.3. Species within crystal drops. Figure adopted from Hampton Research (https://www.hamptonresearch.com/documents/growth_101/9.pdf).

CLEAR DROPS: These appear exactly as the name sounds. The protein was not concentrated enough to reach a state that would allow for crystallization, nor precipitation. Nothing is visible within the drop.

AMORPHOUS MATERIAL: Fibers from clothing, dust, or any other generic debris can be seen in the drop. Sometimes, strange objects like this can act as nucleation sites for protein crystals.

These can always be seen at time zero (t = 0, when the drop is first set up) as sometimes these debris can look similar to crystals, though protein crystals cannot form instantaneously.

PRECIPITATE: Yellow or brown 'dirt-like' appearance within the drop indicates a clump of protein which is no longer in solution and cannot crystallize. Other precipitates that are less dense (light precipitates) still have a possibility to crystallize.

PHASE SEPARATION: These look like bubbles, but really they are 'oil droplets' consisting of either a component of the condition (i.e. organic solvent) or proteins themselves. Crystals can form from phase droplets and can sometimes be an indication of a high level of saturation.

CRYSTALS: Crystals are usually classified as 1D (needle-like, growing in a single dimension), 2D (plate-like, growing in two-dimensions), or 3D (growing in all three dimensions). They have edges and are usually very obvious. Once you obtain a crystal, the next step is to make sure it is not a salt crystal from a salt within your condition.

Create a chart in your lab notebook, and go through each of your wells of your lysozyme drops and label what you see, using the descriptions above.

Chapter 4

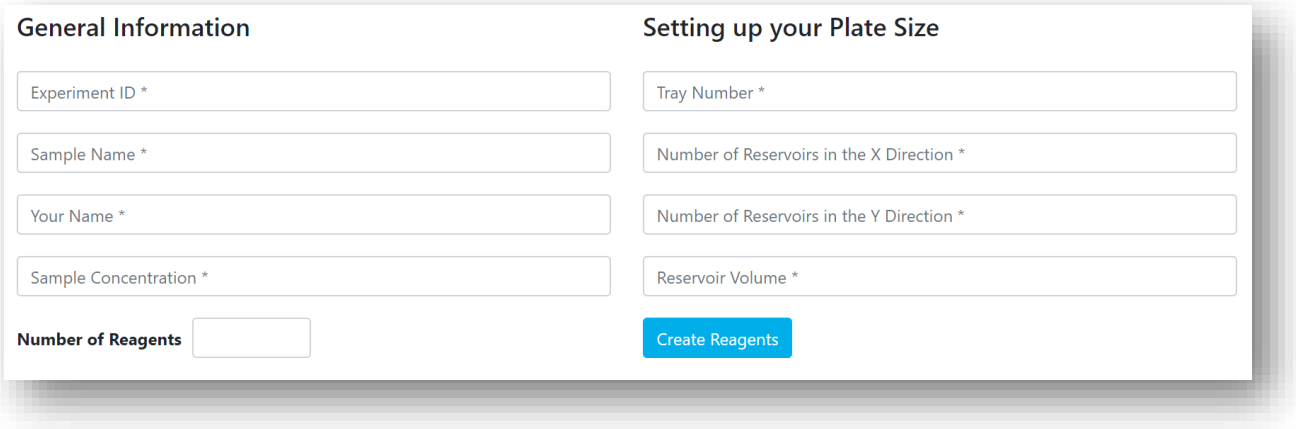
Group Theory, Photon Scattering, and Crystal Geometry

Setting Up Optimization Trays

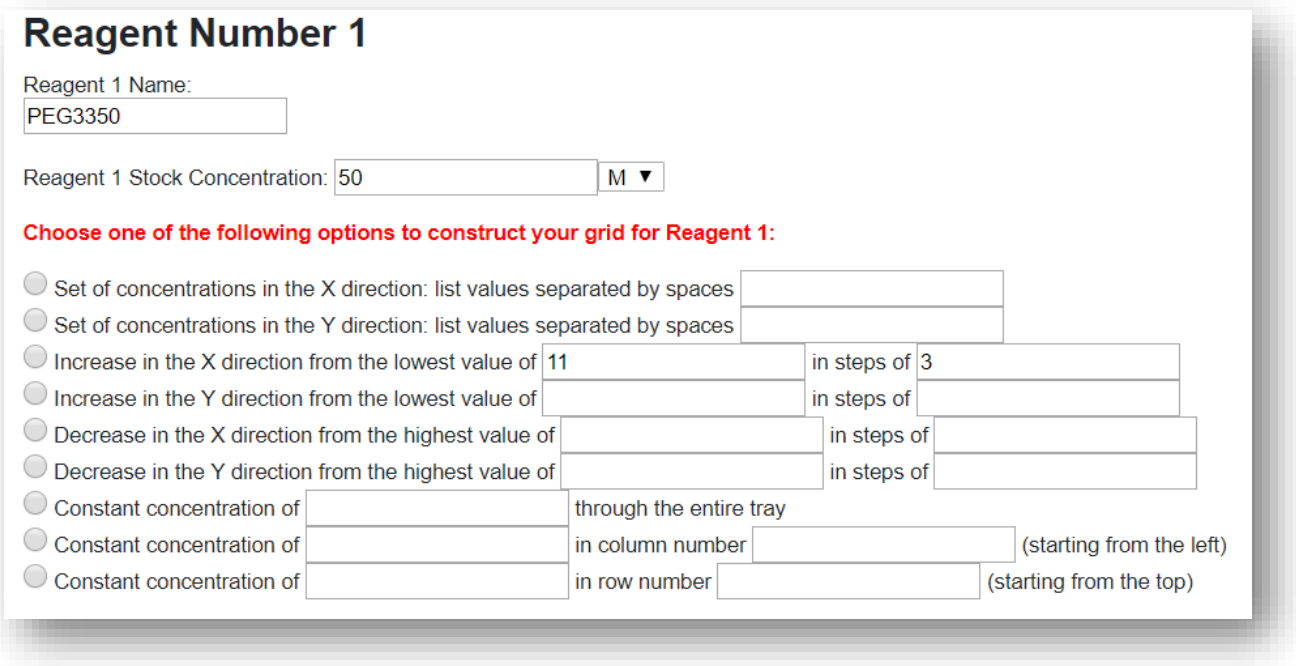
Once you have identified a 'hit' – a condition which has provided you with protein crystals, it is important to initially verify whether or not they truly ***are*** protein crystals and not just an artifact or salt crystals. Once the hit is verified, we want set up optimizations to ensure the crystals are **reproducible** and simultaneously (hopefully) create better quality crystals. There is a fantastic tool for creating optimization trays available on the Hampton Research website that will automatically calculate for you all dilutions required for your tray. It is available at:

https://www.hamptonresearch.com/make_tray.aspx

When you arrive at the website, you will be prompted to enter in some information about your experiment:

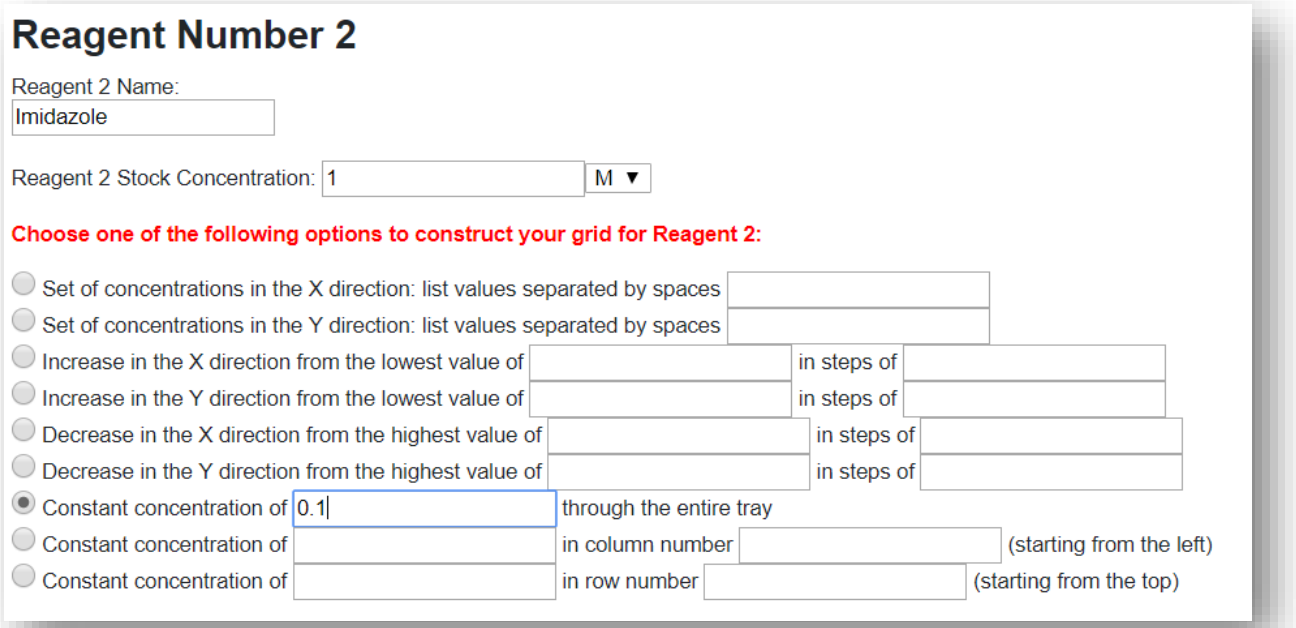


Our optimization trays are 6x4 arrays which are designed to have ~500uL in the reservoir. The number of reagents may vary if your hit consists of more or less than 3 chemical species. Once you have entered in the proper number of reagents, press Create Reagents. This will generate additional fields below that allow you to specify how you would like to vary your reagents. Using PEG3350 (in my example in the infographic from before), we will vary the concentration from 15 to 30% (v/v) by increasing in steps of 3%:



These parameters would give us drop A at 15%, B at 18%, C at 21%, D at 24%, E at 27%, and F at 30%. You will need to decide (and consult with your TA) on the best way to create your optimization tray. Unique/specialty buffers may need to be made by your TA.

As another example, we can also set a constant concentration, as below:



Reagent Number 2

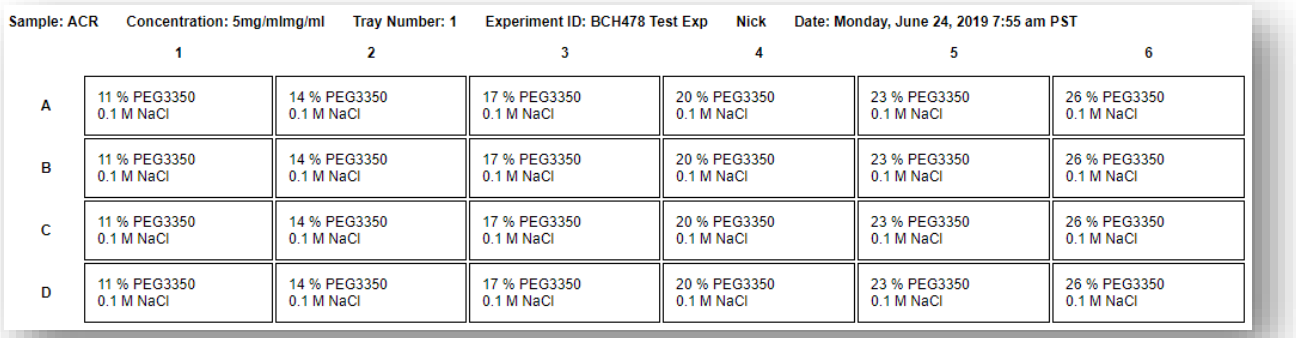
Reagent 2 Name:
Imidazole

Reagent 2 Stock Concentration: 1 M ▼

Choose one of the following options to construct your grid for Reagent 2:

- ○ Set of concentrations in the X direction: list values separated by spaces
- ○ Set of concentrations in the Y direction: list values separated by spaces
- ○ Increase in the X direction from the lowest value of ____ in steps of ____
- ○ Increase in the Y direction from the lowest value of ____ in steps of ____
- ○ Decrease in the X direction from the highest value of ____ in steps of ____
- ○ Decrease in the Y direction from the highest value of ____ in steps of ____
- ◉ Constant concentration of 0.1 through the entire tray
- ○ Constant concentration of ____ in column number ____ (starting from the left)
- ○ Constant concentration of ____ in row number ____ (starting from the top)

Now that our two reagents are set up, we select build tray to have Hampton generate our tray for us:



Sample: ACR **Concentration: 5mg/mlmg/ml** **Tray Number: 1** **Experiment ID: BCH478 Test Exp** **Nick** **Date: Monday, June 24, 2019 7:55 am PST**

	1	2	3	4	5	6
A	11 % PEG3350 0.1 M NaCl	14 % PEG3350 0.1 M NaCl	17 % PEG3350 0.1 M NaCl	20 % PEG3350 0.1 M NaCl	23 % PEG3350 0.1 M NaCl	26 % PEG3350 0.1 M NaCl
B	11 % PEG3350 0.1 M NaCl	14 % PEG3350 0.1 M NaCl	17 % PEG3350 0.1 M NaCl	20 % PEG3350 0.1 M NaCl	23 % PEG3350 0.1 M NaCl	26 % PEG3350 0.1 M NaCl
C	11 % PEG3350 0.1 M NaCl	14 % PEG3350 0.1 M NaCl	17 % PEG3350 0.1 M NaCl	20 % PEG3350 0.1 M NaCl	23 % PEG3350 0.1 M NaCl	26 % PEG3350 0.1 M NaCl
D	11 % PEG3350 0.1 M NaCl	14 % PEG3350 0.1 M NaCl	17 % PEG3350 0.1 M NaCl	20 % PEG3350 0.1 M NaCl	23 % PEG3350 0.1 M NaCl	26 % PEG3350 0.1 M NaCl

Sample: ACR Concentration: 5mg/mlmg/ml Tray Number: 1 Experiment ID: BCH478 Test Exp Nick Date: Monday, June 24, 2019 7:55 am PST

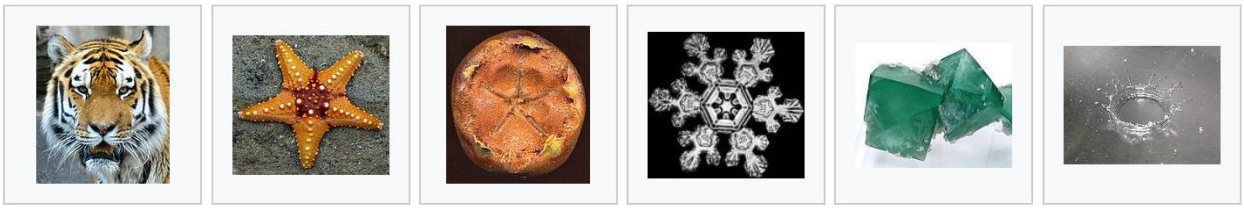
	1	2	3	4	5	6
A	220.00 uL 50 % PEG3350 20.00 uL 5 M NaCl 760.00 ul H2O	280.00 uL 50 % PEG3350 20.00 uL 5 M NaCl 700.00 ul H2O	340.00 uL 50 % PEG3350 20.00 uL 5 M NaCl 640.00 ul H2O	400.00 uL 50 % PEG3350 20.00 uL 5 M NaCl 580.00 ul H2O	460.00 uL 50 % PEG3350 20.00 uL 5 M NaCl 520.00 ul H2O	520.00 uL 50 % PEG3350 20.00 uL 5 M NaCl 460.00 ul H2O
B	220.00 uL 50 % PEG3350 20.00 uL 5 M NaCl 760.00 ul H2O	280.00 uL 50 % PEG3350 20.00 uL 5 M NaCl 700.00 ul H2O	340.00 uL 50 % PEG3350 20.00 uL 5 M NaCl 640.00 ul H2O	400.00 uL 50 % PEG3350 20.00 uL 5 M NaCl 580.00 ul H2O	460.00 uL 50 % PEG3350 20.00 uL 5 M NaCl 520.00 ul H2O	520.00 uL 50 % PEG3350 20.00 uL 5 M NaCl 460.00 ul H2O
C	220.00 uL 50 % PEG3350 20.00 uL 5 M NaCl 760.00 ul H2O	280.00 uL 50 % PEG3350 20.00 uL 5 M NaCl 700.00 ul H2O	340.00 uL 50 % PEG3350 20.00 uL 5 M NaCl 640.00 ul H2O	400.00 uL 50 % PEG3350 20.00 uL 5 M NaCl 580.00 ul H2O	460.00 uL 50 % PEG3350 20.00 uL 5 M NaCl 520.00 ul H2O	520.00 uL 50 % PEG3350 20.00 uL 5 M NaCl 460.00 ul H2O
D	220.00 uL 50 % PEG3350 20.00 uL 5 M NaCl 760.00 ul H2O	280.00 uL 50 % PEG3350 20.00 uL 5 M NaCl 700.00 ul H2O	340.00 uL 50 % PEG3350 20.00 uL 5 M NaCl 640.00 ul H2O	400.00 uL 50 % PEG3350 20.00 uL 5 M NaCl 580.00 ul H2O	460.00 uL 50 % PEG3350 20.00 uL 5 M NaCl 520.00 ul H2O	520.00 uL 50 % PEG3350 20.00 uL 5 M NaCl 460.00 ul H2O

Once you have generated your tray, save a picture of this set up or print it out. Assuming we have the necessary buffers, we can now begin setting up our optimization trays.

Your TA will instruct you on how to how to set up the optimization trays.

A Primer on Symmetry

You probably have an idea of what symmetry is in your head. Unless you have studied symmetry mathematically, it is likely that you associate symmetry with the ancient Greek definition of symmetry which described objects in terms of ‘perfection’ and ‘balance’. Surely you have observed things around you such as the crystalline patterns of snowflakes, helical symmetry in seashells, the projections from starfish, beehive honeycombs, the bilateral symmetry in animals (the left and right halves), and radial symmetry in flowers. Wikipedia has its own page for these patterns in nature. An example screenshot is shown below.



Believe it or not, the idea of symmetry is at the heart of the universe. As stated by physics Nobel laureate Steven Weinberg, “*symmetry provides a key to Nature’s secrets*”. This statement is not to be taken lightly. Symmetry allows us to understand quantum theory and is what allows us to calculate crystal structures. When you inevitably go off and research symmetry on your own and find that the mathematical symmetries of the hydrogen atom are the same as those for our solar system, it is hard to not appreciate the idea of symmetry being part of something much bigger which we must aim to understand.

We should work towards a rigorous mathematical definition of symmetry, instead of an intuitive visual one. From Pieter Thyssen’s book “*Shattered Symmetry*”:

A (mathematical) object is said to be *symmetric*, or to possess a *symmetry*, when there is a *transformation* that leaves the appearance of the object unchanged.

Introductory Group Theory

When I was in elementary school, around grade 4 or 5, I had known my multiplication tables up to $\mathbf{12} \times \mathbf{12}$ for a few years because my mom would make me practice flash cards every night (or at least that's how it feels in my memory). She would flip a card from the deck, maybe it read $\mathbf{4} \times \mathbf{6}$, and I would quickly shout out the number $\mathbf{24}$ as fast as I could. But it wasn't until a day in class much later that I realized, (not of my own realization, but rather because somebody pointed it out) that the number $\mathbf{24}$ can be obtained by taking $\mathbf{6}$ sets of $\mathbf{4}$ things, or by taking $\mathbf{4}$ sets of $\mathbf{6}$ things. When I did finally comprehend this idea, I tried it with all of the multiplication tables that I knew; adding $\mathbf{7}$ to itself $\mathbf{8}$ times and seeing if it was the same as adding $\mathbf{8}$ to itself $\mathbf{7}$ times. Sure enough, it was. No matter what I checked, I always arrived on the same result. I had been memorizing the tables, but never really understood what it meant to apply the number $\mathbf{4}$, $\mathbf{6}$ times. You probably had no idea that multiplying two numbers belonged to this concept in abstract algebra known as group theory.

Let's first talk about *multiplication*.

If you take the action '4' and **apply** it '6' times, you get 24.

If you take the action '6' and **apply** it '4' times, you get 24.

If you take the action '12' and **apply** it '2' times, you get 24.

If you take the action '2' and **apply** it '12' times, you get 24.

It seems like there are multiple routes to the same destination. Does this same idea apply to *adding* numbers?

If you take the action '3' and **apply** '2' to it, you get 5.

If you take the action '2' and **apply** '3' to it, you get 5.

If you take the action '7' and **apply** '4' to it, you get 11.

If you take the action '4' and **apply** '7', it is the same as if you took '7' and **applied** '4'.

Why do I keep typing **apply** in boldface? The word **apply** here is the indication that we are **performing an operation**. In the first case, we are performing the operation of multiplication on two numbers. In the second case, we are applying the operation of addition on two numbers. We call this a **binary operation** because we only apply it to two numbers at once. But what happens if we have more than two numbers?

If you take the action '3' and ***add*** '2' and then ***add*** '6', you get 11.

If you take the action '2' and **add** '6' and then **add** '3', you get 11.

In more clear, mathematical notation;

$$(3 + 2) + 6 = 11$$

and

$$3 + (2 + 6) = 11$$

so

$$3 + (2 + 6) = (3 + 2) + 6 = 11$$

This property of real numbers being added is called **associativity**. In other words; it doesn't matter how you group your elements. The concept also applies for multiplication:

$$(2 \times 4) \times 5 = 80$$

and

$$2 \times (4 \times 5) = 80$$

so

$$(2 \times 4) \times 5 = 2 \times (4 \times 5) = 80$$

As somebody who just wanted to get mathematics 'over with', I took these principles for granted and never really thought about them.

Perhaps you are starting to understand what I am getting at now, so I will give you the formal definition of a group. So far, the groups we have talked about have been the **Additive Group of Real Numbers**, and the **Multiplicative Group of Real Numbers**. To keep things simple, lets talk about the Additive Group of Real Numbers first.

A group is just a set of 'elements', in our case, the real numbers (numbers like 5, 21, 66.2, -450.8, etc.) that has an associated binary operation, in our case, addition, indicated by the "+" symbol.

We denote a set of elements, $\mathbb{G}$, like this:

$$\mathbb{G} = \{a, b, c, d, \cdots\}$$

And we denote the group itself like this:

$$(\mathbb{G}, \bigstar)$$

This is an abstract, general notation. What it means is, we take elements from the set $\mathbb{G}$, and we combine them using the operator, $\bigstar$.

In order for a something to be a group, it must satisfy at least four axioms;

1. **Closure**: when we add any two real numbers, we should get back a real number. For example, adding 10 to 15 gives us 25; which is a real number. We do not get imaginary numbers; only real numbers. This property is called 'closure', and people often say 'the elements of the group are closed under addition', meaning; with the tools provided in the group, you cannot escape the group.

2. **Associativity**: we have already discussed associativity. It does not matter when we are adding numbers whether we add 2 + 5 to 6, or 5 + 6 to 2, we will always get the same result. Note that when we do so, we also get back a real number; this is again

closure. In other words; the grouping of our operations does not matter.

3. **Identity**: The group must have an element, e, called the 'identity' element, such that when the operation is performed with any element of the set, the element remains unchanged. What is this element in the Additive Group of Real Numbers? The number 0. We can add 0 to anything, or add anything to 0, and we will get back the same thing we started with. Again, we obtain a real number, implying closure. 5 + 0 = 5.

4. **Invertibility**: The set must contain an element that can undo the effect of combination with another given element. By undoing an action, you retrieve the identity element (because it is the same as if you did nothing). In the Additive Group of Real Numbers, if you perform the action '5', and then perform the action '- 5', you obtain 0; the identity element. Mathematically, this is stated as;

$$a \bigstar b = b \bigstar a = e$$

Where e is the identity element.

Let's look at a *concrete*, or sometimes called *realized* example. The set of elements in our group will be the real numbers, and our binary operation will be addition:

$$(\mathbb{R}, +)$$

where

$$\mathbb{R} = \{\cdots -3, -2, -1, 0, 1, 2, 3, \cdots\}$$

The real numbers of course include all numbers in between those listed, and this is called a continuous, infinite group. I'm just mentioning this for mathematical preciseness and to not get chewed out by mathematicians.

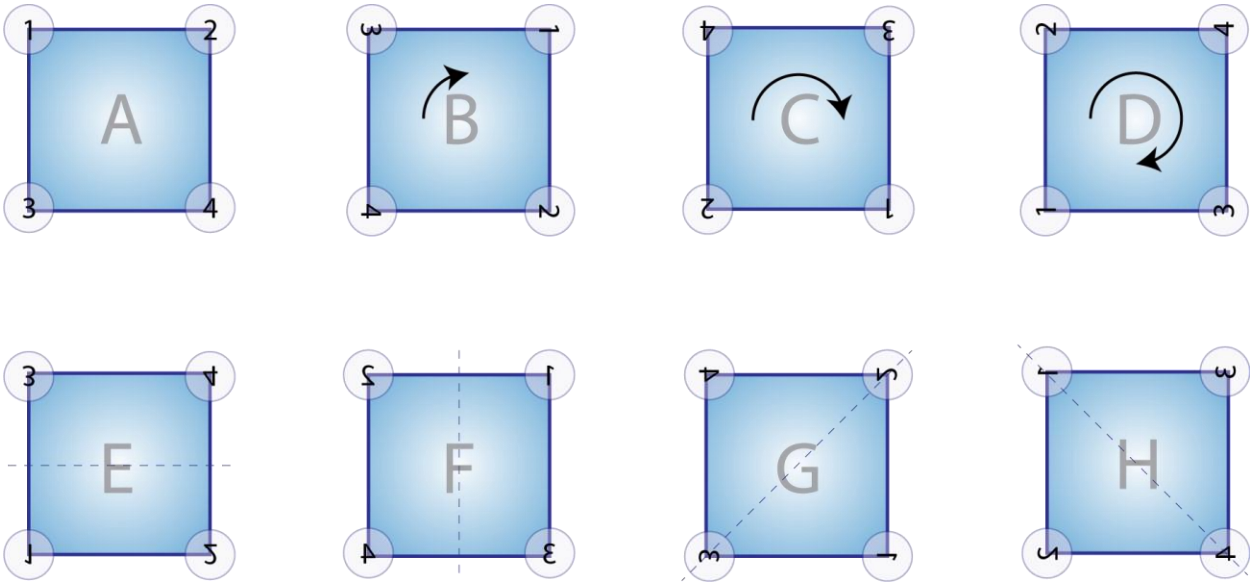
Let's simultaneously consider the Multiplicative Group of Real Numbers:

$$(\mathbb{R}, \times)$$

	Additive Group of Real Numbers	Multiplicative Group of Real Numbers
Notation	$(\mathbb{R}, +)$	$(\mathbb{R}, \times)$
Closure	$\mathbf{1 + 1 = 2}$ *2 is a real number.*	$\mathbf{5 \times 5 = 25}$ *25 is a real number.*
Associativity	$\mathbf{2 + (5 + 1) = 8}$ $\mathbf{(2 + 5) + 1 = 8}$	$\mathbf{10 \times (5 \times 2) = 100}$ $\mathbf{(10 \times 5) \times 2 = 100}$
Identity	$\mathbf{21 + 0 = 21}$ **0**	$\mathbf{61255 \times 1 = 61255}$ **1**
Invertibility	$\mathbf{10 + (-10) = 0}$ *Identity element returned.*	$\mathbf{50 \times \frac{1}{50} = 1}$ *Identity element returned.*

Let's now get away from numbers and look at a geometrical shape – the square. We call this group the dihedral group, and our operator (the star) can be thought of as 'performing some transformation' from the set of elements on the square. We could replace the star with a multiplication sign, but due to our preconception of how multiplication works with numbers, I do not want to confuse you with trying to understand what it means to 'multiply' actions on a square.

$$(\mathbb{D}_4, \bigstar)$$



e	Identity element; do nothing
a	Rotate 90 degrees clockwise
b	Rotate 180 degrees clockwise
c	Rotate 270 degrees clockwise
d	Flip about a horizontal axis in the middle
e	Flip about a vertical axis in the middle
f	Flip about a diagonal axis (45 degrees)
g	Flip about a diagonal axis (-45 degrees)

If you take e and then **apply** c, you get c.

If you take c and then **apply** e, you get c.

If you take a and then **apply** b, you get c.

If you take b and then **apply** a, you get c.

Let's make a table, as we did before, to see if we can prove that this is a group.

	Symmetry Group D4
Closure	Applying any action from the set always returns an element of the set.
Associativity	$(B * C) * B = \boldsymbol{A} = B * (C * B)$ or $(90° \curvearrowright * 180° \curvearrowright) * (90° \curvearrowright) = 90° \curvearrowright (90° \curvearrowright * 180° \curvearrowright) = 360° \curvearrowright = \boldsymbol{A}$ The grouping of our operations does not matter; we still get the same conformation back.
Identity	Applying 'A' does nothing; it is the identity element. $\boldsymbol{A}$
Invertibility	Any clock-wise element can be undone with a counter-clockwise rotation. Any inversion can be undone by re-applying the same (or inverse) inversion. Doing so returns the identity element.

We are starting to realize that adding number is just the same as performing sliding actions along number lines. Multiplying numbers is just the same as performing stretching and squishing (scaling) actions along number lines. Axial rotations seem to obey the same principles that numbers do! What we are aiming to display in this section is how multiplication of complex numbers can correspond with rotations around the origin of the Argand (complex) plane.

Throughout the rest of this chapter, we will continually allude to these groups until you begin to see the relevance, and their usefulness in mathematics.

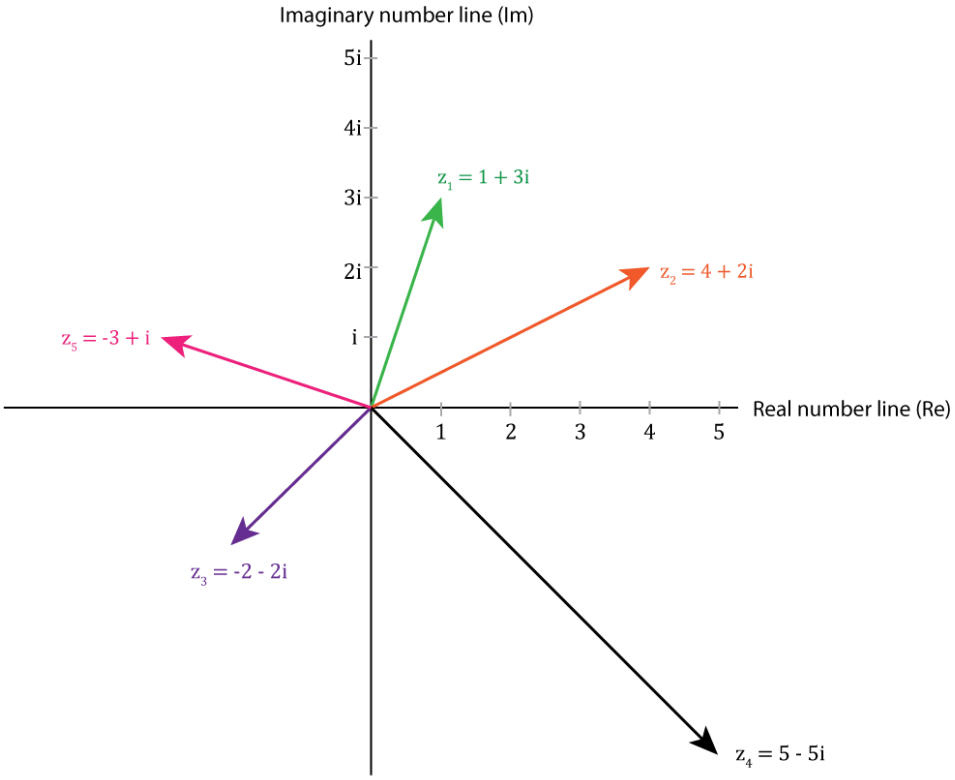
Complex Numbers

The reason we use complex numbers in physics and crystallography (and many other fields, especially electrical engineering) is because as mathematical objects, they are a great way to represent something with an amplitude and a phase. Electromagnetic radiation, has wave properties that are described by an amplitude, phase, and frequency. But what is a complex number?

A complex number Z takes on the form:

$$\boldsymbol{z} = x + iy$$

Where x is what we call the 'real' component, and i is an imaginary number multiplied by another real component y. What does it mean for these numbers to be 'real' and 'imaginary'? Well, from our discussion earlier, a real number is any point on the real number line. It encompasses all rational and irrational numbers (rational being numbers that can be expressed as the quotient of two integers, like $\frac{5}{2} = 2.5$; irrational being numbers that do not terminate, and do not repeat; like π). The imaginary component is composed of a real number, y, multiplied by the imaginary scaling factor, i , which is defined as being equal to $\sqrt{-1}$. You can visualize complex numbers by imagining them as a point, or a vector, on the 'complex number plane'. I have plotted 5 complex numbers on this plane below:



We have been discussing how groups have certain 'actions' associated with them. For example, when we add two real numbers, it is as if we are performing a 'sliding action' along a 1-dimensional number line; the distance that we travel is dependent on the elements of the set. So, for adding 2 plus 5, we slide the number line to the left 2 units, and then slide again 5 units. We end up at 7.

When we are multiplying two real numbers, we 'stretch' our number line by the scaling factor, as determined by the element of the set, keeping the origin fixed at zero. So if we want to multiply an element by 5, we scale our number line up 5 times, and find that the position 2 now sits where 10 sat on the original number line ($5 \times 2 = 10$).

When we are adding complex numbers, we add them much like vectors (though it is important to mention that complex numbers are ***not*** vectors) by adding their separate components. Geometrically, this can be visualized by adding tip to tail, as with vectors. For example, to add the following two complex numbers:

$$z_1 = 1 + 2i$$

$$z_2 = 2 + i$$

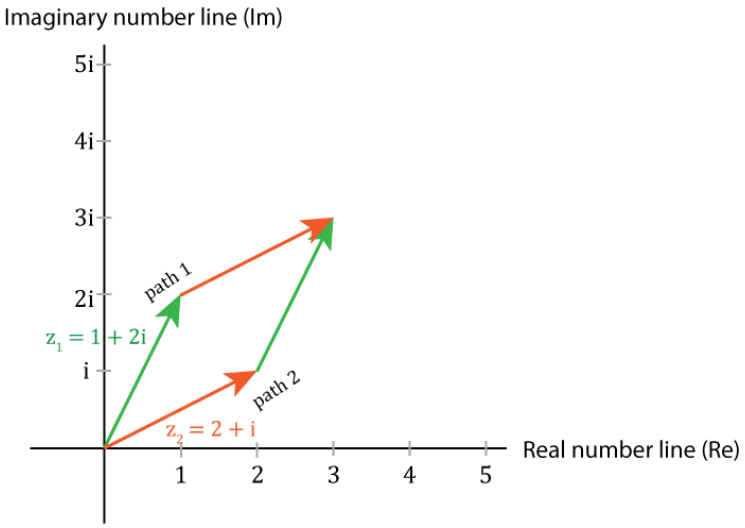
We can just add the real components $(1 + 2 = 3)$, and the imaginary components $(2i + i = 3i)$, to obtain:

$$z_{1+2} = 3 + 3i$$

As you can see, we added two complex numbers and it returned another complex number. This is yet another example of closure. It can be visualized again as a new point on the complex plane. Note that;

If you add z_1 to z_2, you get z_3.

If you add z_2 to z_1, you get z_3.



It doesn't matter which path you take, you will arrive on the same complex number (this property is called commutativity, and is a requirement for Abelian groups, though this is not important for our purposes). What is important to note here is that in the Additive Group of

Complex Numbers, we can break down how we get to the point defied by z_1 by breaking down our movement as a horizontal slide 1 unit to the right, followed by a vertical slide 2 units up.

Complex Number Multiplication

When we are multiplying complex numbers, the resultant vector has a magnitude equal to the product of the magnitudes of each vector, and an angle equal to the sum of the two complex numbers. In this way, multiplying stretches the complex vector, and rotates it about the origin. So instead of feeling bamboozled when you see multiplicative complex number algebra, all you need think about is arrows that are stretching up or down as they rotate around the complex plane origin.

Let's multiply two complex numbers to convince ourselves that this is truly the case.

$$z_1 = 4 + 2i$$

$$z_2 = 2 + 3i$$

To multiply these together, we treat them as if it was a binomial expansion (First, Outside, Inside, Last; FOIL)

$$z_1 \times z_2 = (4 + 2i)(2 + 3i)$$

$$z_1 \times z_2 = 8 + 12i + 4i + 6i^2$$

Now recall that i is equal to $\sqrt{-1}$, so taking its square would simply reduce it to -1.

$$\boldsymbol{z_3 = 2 + 16i}$$

We know that these complex numbers can be plotted on the complex plane just like vectors, so let's do that to convince ourselves that what we have obtained here makes sense.

First let's see if the magnitude of our new vector, z_3, is equal to the product of the magnitudes of z_1 and z_2. We know from the Pythagorean theorem that the magnitude of these vectors can be calculated with:

$$a^2 + b^2 = c^2$$

$$c = \sqrt{a^2 + b^2}$$

So,

$$\left||z_1|\right| = \sqrt{4^2 + 2^2} = \sqrt{16 + 4} = \sqrt{20} = 4.472$$

$$\left||z_2|\right| = \sqrt{2^2 + 3^2} = \sqrt{4 + 9} = \sqrt{13} = 3.605$$

$$\left||z_3|\right| = \sqrt{2^2 + 16^2} = \sqrt{4 + 256} = \sqrt{260} = 16.124$$

So now we check (including all decimals in our calculator):

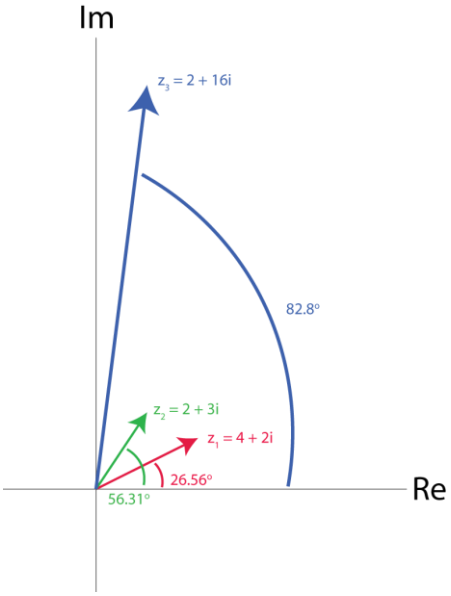
$$4.472 \times 3.605 = 16.124$$

The 'phases' or 'angles' of these complex vectors can be calculated with elementary trigonometry using **SOH CAH TOA**. When we are referring to the angle subtended by a complex number, we often write $\arg(z)$. So if I wanted to tell you the angle subtended by the complex number "z_n" was 223 degrees, I would write $\arg(z_n) = 223°$.

$$\arg(z_1) = \tan^{-1}\left(\frac{2}{4}\right) = 26.56°$$

$$\arg(z_2) = \tan^{-1}\left(\frac{3}{2}\right) = 56.31°$$

$$\arg(z_3) = \tan^{-1}\left(\frac{16}{2}\right) = 82.87°$$



From here it is clear that:

$$26.56 + 56.31 = 82.87$$

So far we have looked at both the Additive Group of Complex Numbers and the Multiplicative Group of Complex Numbers. Can we again generate tables to convince ourselves that these satisfy the group requirements of closure, associativity, identity, and invertibility?

	Additive Group of Complex Numbers	**Multiplicative Group of Complex Numbers**
Notation	$(\mathbb{C}, +)$	$(\mathbb{C}, \times)$
Closure	$(\mathbf{1} + \mathbf{2}\boldsymbol{i}) + (\mathbf{5} - \boldsymbol{i}) = \mathbf{6} + \boldsymbol{i}$ *6 + i is a complex number.*	$(\mathbf{2} + \mathbf{2}\boldsymbol{i})(\mathbf{3} + \mathbf{3}\boldsymbol{i}) = \mathbf{6} + \mathbf{12}\boldsymbol{i} + \mathbf{6}\boldsymbol{i}^{\mathbf{2}}$ $= \mathbf{0} + \mathbf{12}\boldsymbol{i}$ *0 + 12i is a complex number.*
Associativity	$(\mathbf{3} + \mathbf{2}\boldsymbol{i}) + (\mathbf{4} + \boldsymbol{i}) = \mathbf{7} + \mathbf{3}\boldsymbol{i}$ $(\mathbf{4} + \boldsymbol{i}) + (\mathbf{3} + \mathbf{2}\boldsymbol{i}) = \mathbf{7} + \mathbf{3}\boldsymbol{i}$	$(\mathbf{5} - \mathbf{3}\boldsymbol{i}) \times (\mathbf{5} - \mathbf{4}\boldsymbol{i}) = \mathbf{13} - \mathbf{35}\boldsymbol{i}$ $(\mathbf{5} - \mathbf{4}\boldsymbol{i}) \times (\mathbf{5} - \mathbf{3}\boldsymbol{i}) = \mathbf{13} - \mathbf{35}\boldsymbol{i}$
Identity	$\boldsymbol{z_i} = \mathbf{0} + \mathbf{0}\boldsymbol{i}$	$\mathbf{1}$ *All real numbers are also complex numbers.*
Invertibility	$(\mathbf{5} + \mathbf{5}\boldsymbol{i}) + (-\mathbf{5} - \mathbf{5}\boldsymbol{i}) = \mathbf{0} + \mathbf{0}\boldsymbol{i}$ *Identity element returned.*	$(x + iy)\left(\frac{x - iy}{x^2 + y^2}\right) = z \times \frac{1}{z} = 1 + 0i$

Group Isomorphisms

It seems quite convenient that multiplication of complex numbers gives rise to an additive rotation. How can we rationalize this? Consider the following abstract group:

$$(\mathbb{G},\times)$$

where

$$\mathbb{G} = \{e, a, b, c\}$$

We can map the elements of this abstract group to several different concrete group as below. Let $\mathbb{G}'$ represent the set $\{1, i, -1, -i\}$, and $\mathbb{G}''$ will be our square rotation group from before:

$\mathbb{G}$	$\rightarrow$	$\mathbb{G}'$
e	$\rightarrow$	1
a	$\rightarrow$	i
b	$\rightarrow$	-1
c	$\rightarrow$	$-i$

$\mathbb{G}$	$\rightarrow$	$\mathbb{G}''$
e	$\rightarrow$	$0°$
a	$\rightarrow$	$90°$
b	$\rightarrow$	$180°$
c	$\rightarrow$	$270°$

Look what happens when we construct multiplication tables for each of these groups, color-coding them by their abstract mapping:

	1	i	-1	$-i$
1	1	i	-1	$-i$
i	i	-1	$-i$	1
-1	-1	$-i$	1	i
$-i$	$-i$	1	i	-1

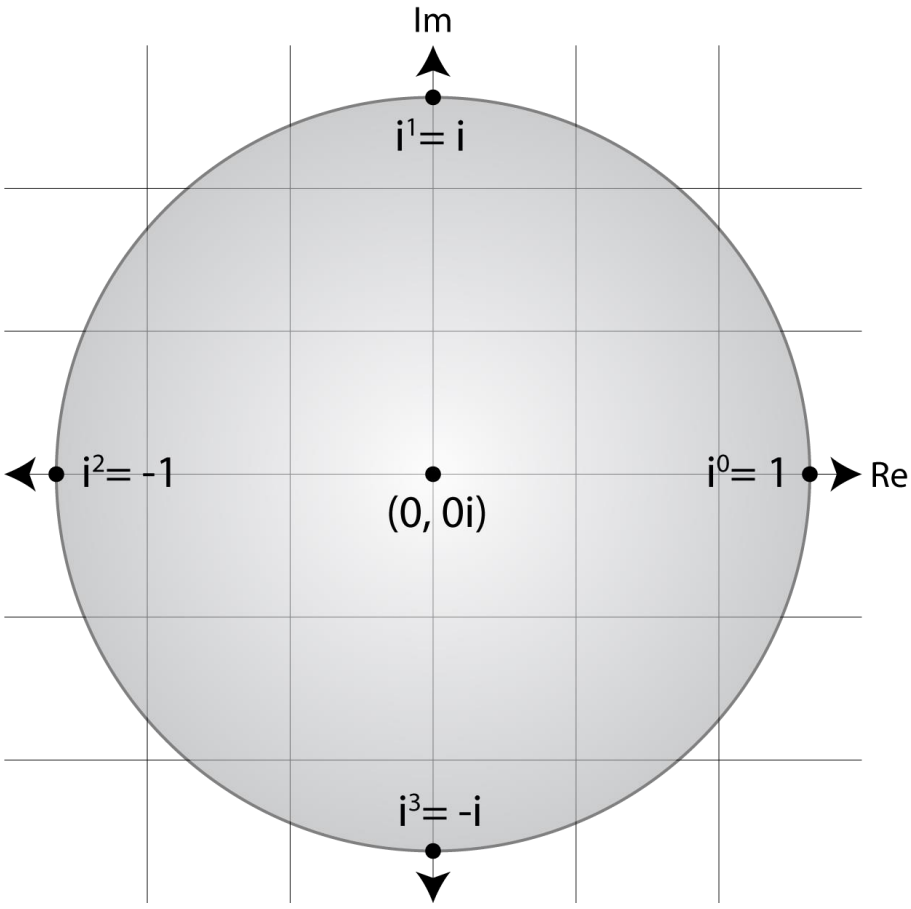
	$0°$	$90°$	$180°$	$270°$
$0°$	$0°$	$90°$	$180°$	$270°$
$90°$	$90°$	$180°$	$270°$	$0°$
$180°$	$180°$	$270°$	$0°$	$90°$
$270°$	$270°$	$0°$	$90°$	$180°$

Incredible! For the longest time, I had been trying to understand why, mathematically, it made sense that complex number multiplication could represent a counter-clockwise rotation in the Argand (complex) plane. These two groups are said to be ***isomorphic***. This idea will massively aid in understanding Fourier transforms. Group theory in action!

"A rotation of a square by $180°$ *followed by another one through* $90°$ *is the same as one rotation through* $270°$*."*

"The complex number i *multiplied by* -1 *yields the complex number* $-i$*."*

These are equivalent statements![5]



Polar Representation of Complex Numbers

It is important to understand that complex numbers can be represented in another way. Instead of separating them by their real (Re) and imaginary (Im) components on the complex plane (analogous to x- and y-components on the real plane), we can represent them with 'polar coordinates'. The polar coordinate representation of a complex number looks like:

$$z = |z|(cos\theta + isin\theta)$$

Do not be intimidated by the sudden introduction of sine and cosine functions into the complex number representation. Remember that the cosine of the inner angle of a right triangle represents the x-component of that triangle. The sine of the same angle represents the y-component of the triangle. As such, we basically have back our complex number nomenclature whereby we are adding a real component (x) to an imaginary component (y). So, $\cos(\theta)$ really just means $x_{component}$, and $\sin(\theta)$ really just means $y_{component}$. Of course, these fractional components must be multiplied by the magnitude of the vector, $|z|$, in order to have the appropriate lengths, otherwise, it would be of unit-vector length (1). The purpose of this section is to understand the "Eulerian" representation of a complex number. This representation is as follows:

$$z = |z|(cos\theta + isin\theta) = \boldsymbol{|z|e^{i\theta}}$$

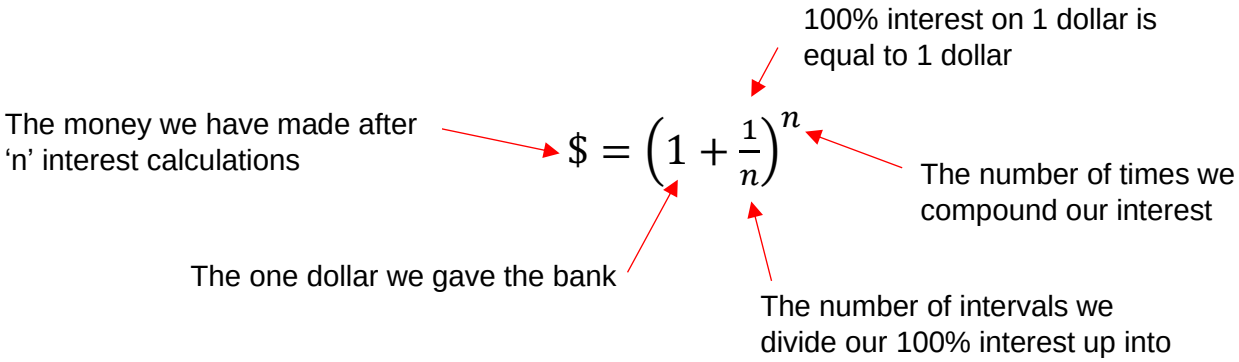
Do not worry if you don't yet understand how the expression in green represents a complex number in polar coordinates. As a primer; the $|z|$ component represents the 'length' of a vector drawn out on the Re-axis (x-axis). This vector is then rotated about the origin, counter-clockwise, using the operation $e^{i\theta}$.

So far, we have seen that additive groups typically perform symmetrical actions by performing sliding actions in one or two dimensions. Multiplicative groups seem to stretch, squish, and rotate. Let's review the rules of exponentiation

Why 'e^i' Representing Rotation Makes Sense

In order to understand how two abstract symbols, '*e*', and '*i*', when placed together in a mathematical expression could possibly act to create the phenomena rotation, we have to go back in time to origins of Euler's number. I was taught that Euler's number was an irrational number, equal to 2.718… and my highschool math teacher left it at that. I now realize that my math teacher had no idea what Euler's number really meant. '*e*' encompasses a whole idea of continual growth. It is amazing how a single number could encompass an entire *process*! When you think of the number '3.25', you think of how it represents, perhaps 3 discrete objects and a quarter of another one, and you leave it at that. Why would 2.718… be any different? Let's try to understand this idea with the classic explanation – compounded interest.

You have one dollar, and you decide to go to the bank. You tell the bank you will give them the one dollar *now* if they promise to pay you interest on this one dollar. They are willing to offer you 100% interest. They are going to let you choose how frequently you would like the interest to compound per year. Let's first create a mathematical expression that represents this scenario.



So, imagine if we chose that we only want the bank to compound our interest once per year. How much money will we make at the end of the year? Another way of saying this is "solve for $, assuming $n = 1$".

$$\$ = \left(1 + \frac{1}{1}\right)^1 = (2)^1 = 2$$

We have made $2.00. What if we make them calculate interest twice per year?

$$\$ = \left(1 + \frac{1}{2}\right)^2 = (1.5)^2 = 2.25$$

Nice! We made an extra quarter. Let's try three times per year…

$$\$ = \left(1 + \frac{1}{3}\right)^3 = (1.333)^3 = 2.37$$

Four times per year?

$$\$ = \left(1 + \frac{1}{4}\right)^4 = (1.25)^4 = 2.44$$

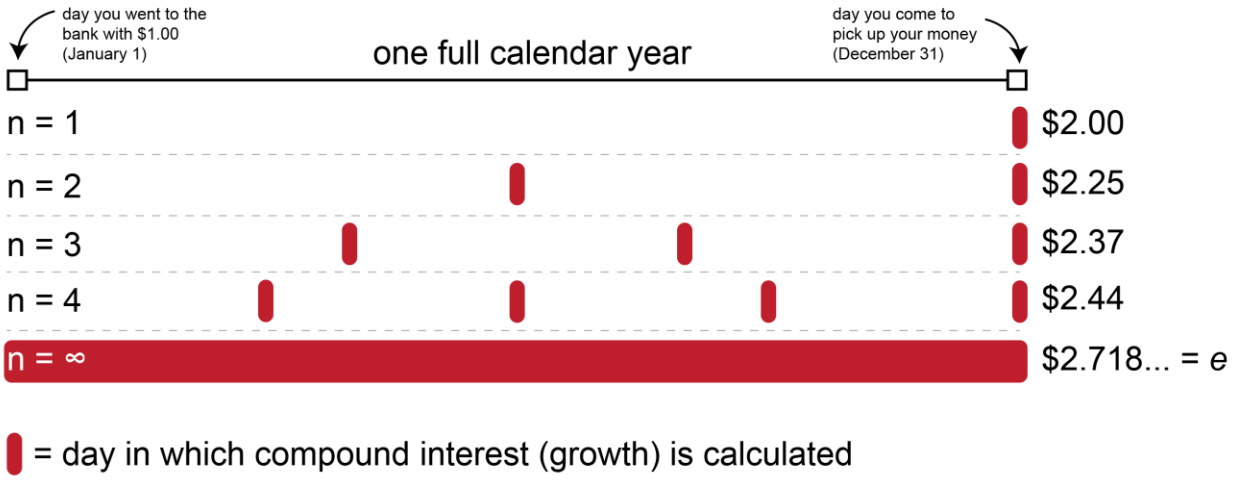
We have started to see a pattern whereby the more frequently we make the bank compound our interest, the more money we seem to be making. What is also interesting though is that the amount of interest perpetually decreases as we have more compounding intervals (the $\frac{1}{n}$ part). We then ask… what happens if we let n go off to infinity? This would represent continuous growth. As though every single instant in time from now to the end of the year, the bank will be adding interest (however small) onto our total and re-compounding it. One might think that we will become super rich. Before we try it with infinite compounding periods, lets try 1000 first.

$$\$ = \left(1 + \frac{1}{1000}\right)^{1000} = (1.001)^{1000} = 2.716923$$

The money we made relative to four compounding periods is not that much greater (it is not linearly proportional)… so something must be going on:

$$\$_{continuous\ growth} = \lim_{n\to\infty} \left(1 + \frac{1}{n}\right)^n = 2.718281284590452 \ldots = e$$

Euler's number doesn't just represent some magical irrational number. It represents the process of ***continuous growth*** … the way growth occurs in nature. It is for this reason you may remember in math class that its inverse is called the 'natural logarithm' or 'ln'. This strange number inherently encompasses this entire idea of continuous growth such that it represents how much growth occurs in one unit of time when you compound at infinitely small intervals. A diagram hopefully makes this idea more clear:



So, when you see 'e' in mathematical expressions – it is not always wise to simply think of it as its irrational self. Think rather of it being expressed as:

$$\lim_{n\to\infty}\left(1+\frac{1}{n}\right)^{n}=e$$

This process better encompasses why we use 'e' to describe natural phenomena. Things (on our macro scale at least) do not grow in discretized fashion. A child, when his height is measured between birthdays is noticed to grow 4 inches taller. This does not mean that he remained the same height for 365 days, and then suddenly shot up to be 4 inches taller on his birthday. Rather, he was **continually growing** for that duration, and for that reason, we need a **continuous growth model** to mathematically model this. Now imagine what we can do now that we have something

that represents growth! Imagine if we multiplied our growth rate by the imaginary number *i*! We can rotate things as they grow continuously, and change their growth directions to be circular! Why would we care to do such a thing? Because remember, properties of electromagnetic radiation are such that as they propagate, they vary their electric and magnetic fields like harmonic functions $sin(x)$ and $cos(x)$. In this way, they are cyclical, and trace out circles as they propagate and their phase changes.

Since we know that the inverse operation of exponentiation using base *e* can be reversed by taking the *ln* of that same number (and vice versa), we can represent the number '5' for example, as:

$$5 = e^{\ln(5)}$$

Thus, we can think of the number 5 as the number that our '1 dollar' grew to in one unit of time, using continual growth, with a rate constant of $ln(5)$. Remember, when we are putting *e* to some exponent, we are really putting that whole green limit above to the same exponent. This means the outer 'n' gets multiplied by that exponent, and it effectively increases the number of times we are compounding the interest. Thus, it speeds up the process, and that is why it is called a rate constant (some scalar number that increases the growth rate).

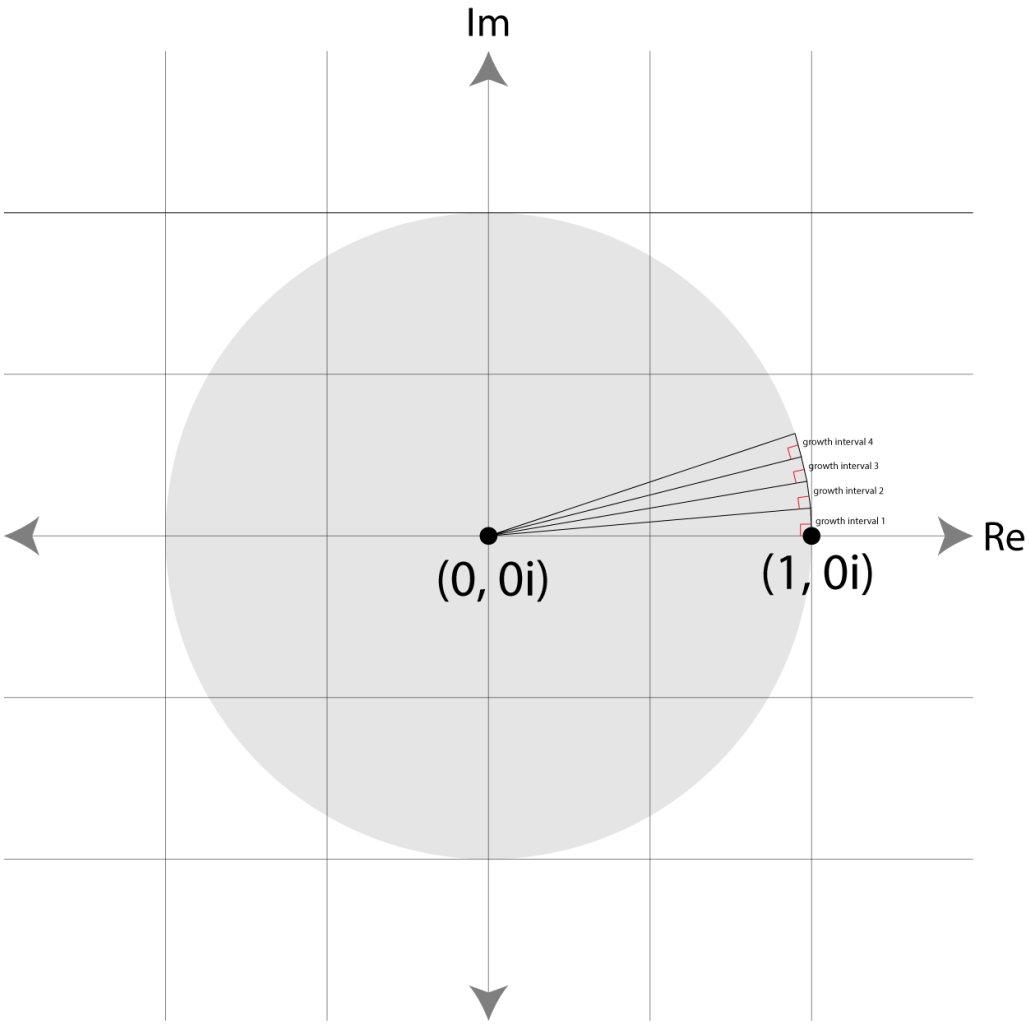
What happens if we raise this whole thing to the power of our imaginary number, i?

$$5^i = \left(e^{\ln(5)}\right)^i$$

From the power rule, we remember that we distribute the i into the brackets, multiplying it by any existing exponents:

$$5^i = e^{\ln(5)*i}$$

Remember from before that multiplication by *i* causes 90 degree rotation. Thus, as the number 1 (our 1 dollar from our limit) begins to grow, it will be continually pushed perpendicular to its current direction of growth. Visually, this creates a scenario like the following:

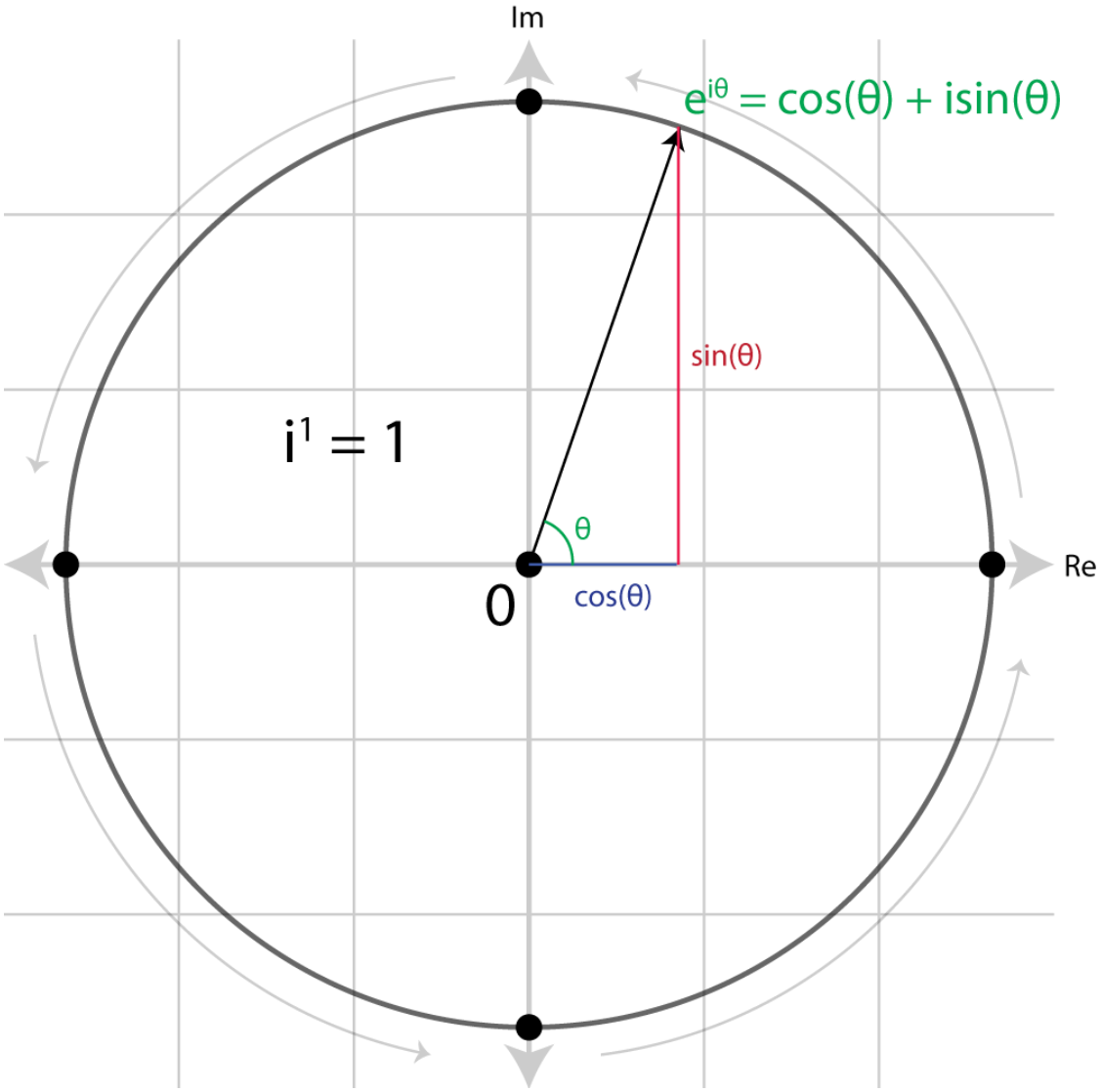


Note that the right angles in the diagram above are not perfect, since I cannot draw infinitesimally small triangles. The above approximation is good enough for visualization of how exponentiation to an imaginary number gives rotational growth.

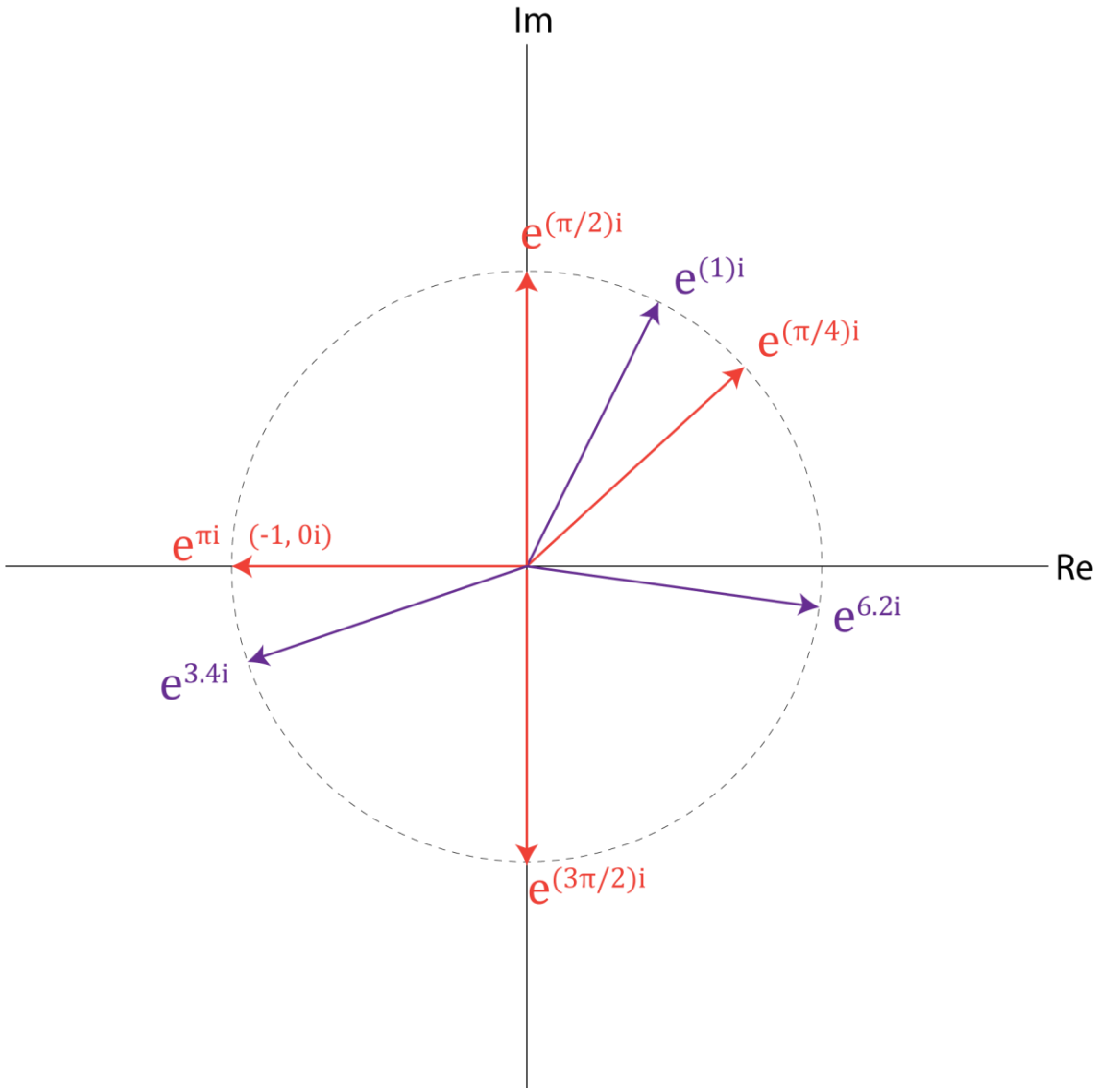
Euler's Formula

I want to now discuss Euler's formula. Euler's formula is arguably the most physically ubiquitous mathematical equation known to man as it finds so many uses in the natural sciences. This equation was a favourite of Richard Feynman's, who referred to it as "*the most remarkable formula in mathematics*" in his first volume on *Lectures on Physics*.

$$e^{i\theta} = cos(\theta) + isin(\theta)$$



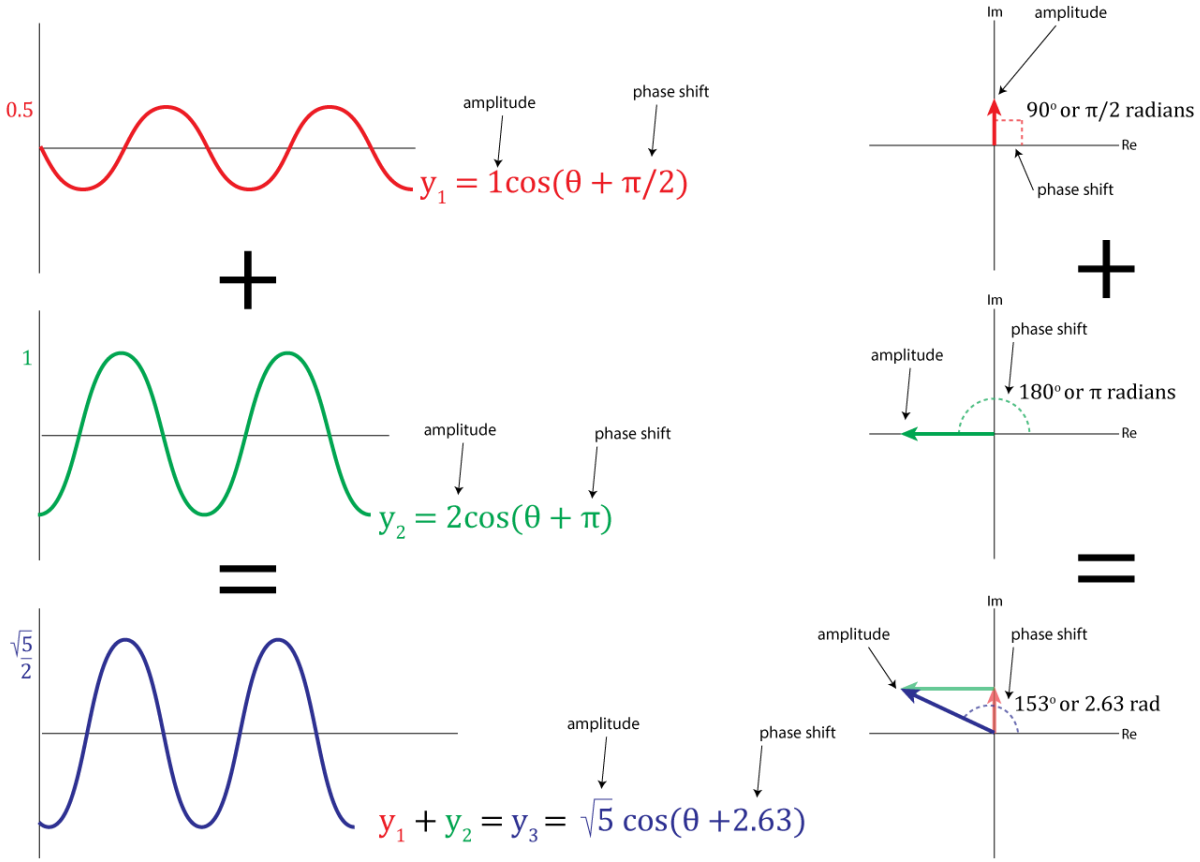
The derivation of Euler's formula is beyond the scope of this book. If you are interested, you can look up proofs of Euler's formula online. The main takeaway from this section is that when you see the base ***e*** raised to a complex exponential, you will know to think of it as meaning a rotation around the complex plane. For our purposes, the amount that it is rotated corresponds to the ***phase*** of a scattered wave. The ***magnitude*** (length) of the of the arrow corresponds to how many electrons contributed to that scattering event. Keep this in mind, we will soon be looking at the electron density equation.



X-Ray Scattering

It is not feasible to expect you to understand how a protein crystal creates a diffraction pattern when exposed to X-ray photons without first understanding more fundamental phenomena, such as the scattering of an X-ray photon from a single electron, or a single atom (unless you are of course a genius). We already discussed light and complex numbers in the previous sections, so let's tie these ideas together and talk about the **superposition of electric field vectors**. Do not be afraid of this term; you probably already understand much of the concepts we are about to introduce.

If you recall, complex numbers can be broken down into an object that is represented by a real component and an imaginary component, or more relevantly, a magnitude $|z|$, and a phase angle θ. Since a simple sinusoidal wave has two major components that we're interested in, amplitude and phase (wavelengths are typically kept constant for most diffraction experiments), complex numbers are fantastic tools for modelling waves. The manner in which complex numbers add together (if you remember, like vectors), is the same manner in which electromagnetic waves superimpose in space. This means that if we take two waves as below, and add them together, you can see the representation is equivalent to a sinusoid representation, or as a complex number representation:



We can see that adding waves can be done mathematically as the sum of sine functions, the sum of complex numbers, or geometrically. Using complex numbers/geometry is arguably easier to visualize, and it is certainly easier computationally.

Can we write a general expression that allows us to sum many different waves? If you remember back to the 'Eulerian' representation of a complex number, which is a vector of some length, rotated around the origin of the complex plane, it can be represented like so:

$$z = re^{i\theta}$$

So, if we want to add a whole bunch of these together, we can simply add a summation sign in front.

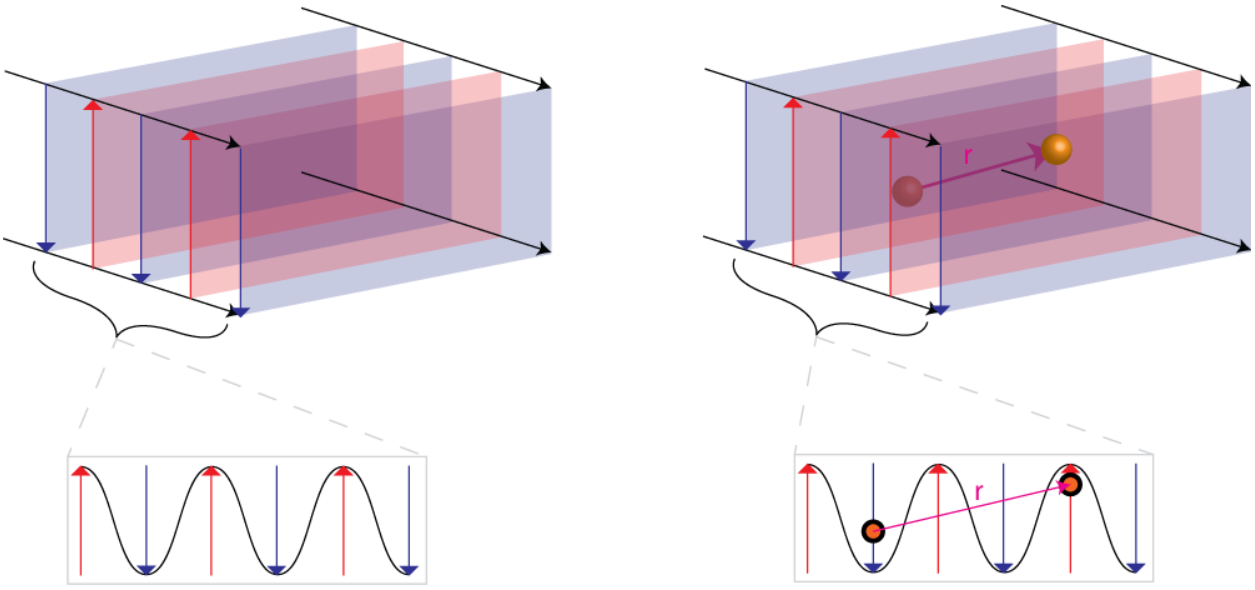
$$F = \sum_{j=1}^{n} r_j e^{i\theta_j}$$

This expression is simply what we did in the figure above, but for a whole bunch of waves. Specifically, n waves where n is some number. If $n = 10$, then you are adding 10 waves together which each (may or may not) have different amplitudes and phases. Since we remember that the Additive Group of Complex numbers is associative (and commutative), we can just place all of these vectors tip to tail, find the resultant vector, and that is our wave! Try to read this equation to understand ***what it is saying***, instead of being intimidated by the symbols. It says, "lets add together a whole bunch of waves that, as complex number representations, have a length r, each of which is rotated around the origin by some angle θ. In doing so, we will obtain our resultant wave, denoted **F**."

X-Ray Scattering from a Single Atom

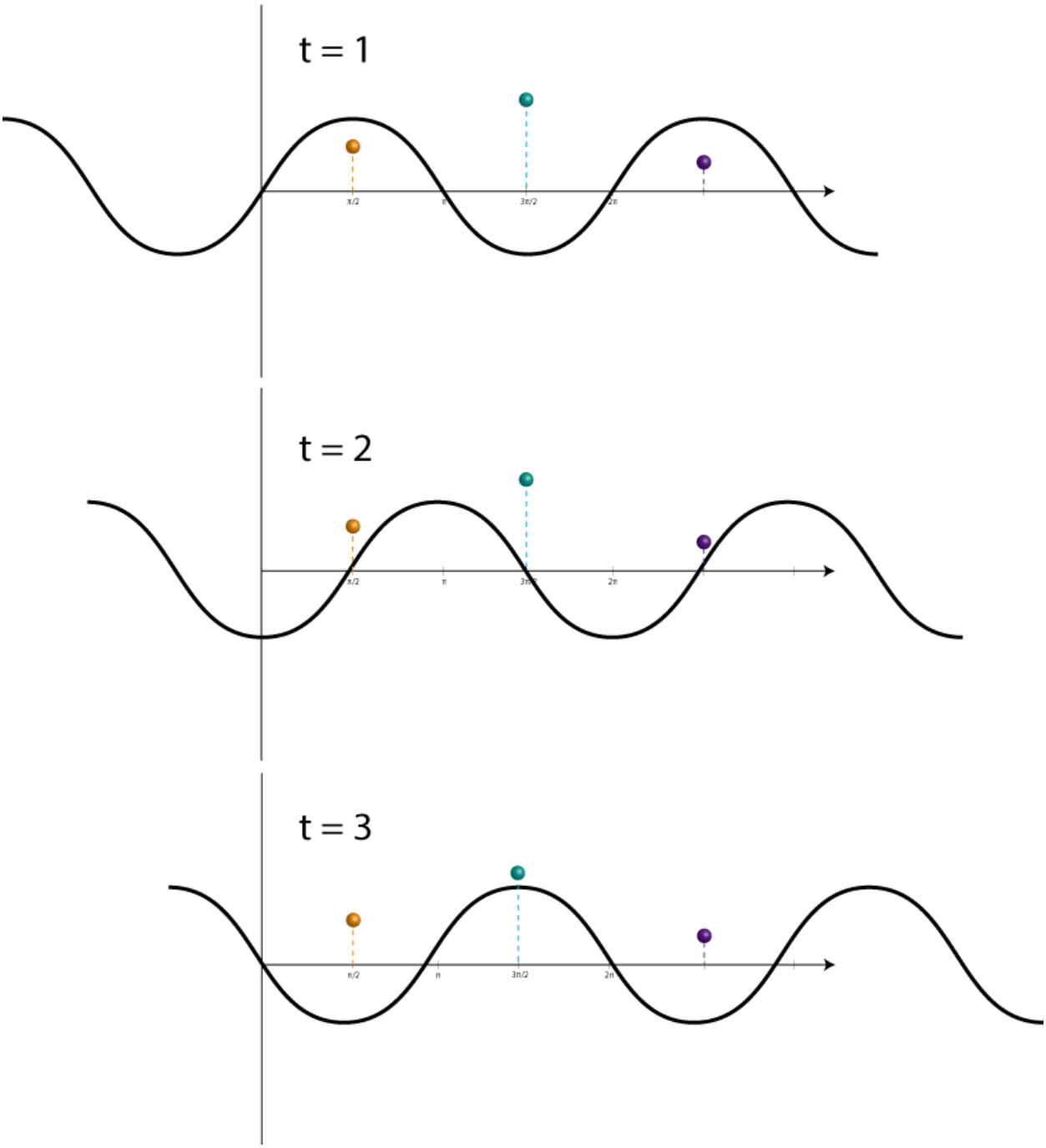
Atoms have, of course, *multiple electrons* arranged in probability distributions (orbitals) around their nuclei. Electrons do have ***discrete*** positions, but the positions they may occupy are limited to within these probability distributions. For our purposes, although not entirely accurate, we will consider the electron density surrounding our atom to be perfectly spherical (it is a good enough approximation to keep the math simple). Since any two electrons are not located at the same position in space and the wavelength of our X-rays is comparable to the distance separating any two electrons in an atom, we will have non-negligible phase differences arising from the scattering from each of these two photons. How can we account for this phase difference? Using a scattering diagram, we can compare the phases of scattered waves coming from electrons at different positions in an atom.

First, let me say, it is absolutely critical that you visualize the X-ray wave correctly in your head. We have been thinking of things in two dimensions, but this is not really the case in real life. When we talk about a 'plane wave', we are talking about a flat plane propagating in a certain direction. The plane carries with it an electric field vector. As the plane propagates, the same point in space will experience a wave maximum, and then a minimum, and so forth. That means that two points spaced some distance apart (***depending on their distance***) may 'feel' different electric field vectors at any given instant. Since it is these field vectors that accelerate the electrons in some direction, when one electron is experiencing a maxima (on the 'up part' of its oscillation), another electron a distance π radians away will be experiencing a minima (on the 'down part' of its oscillation) .



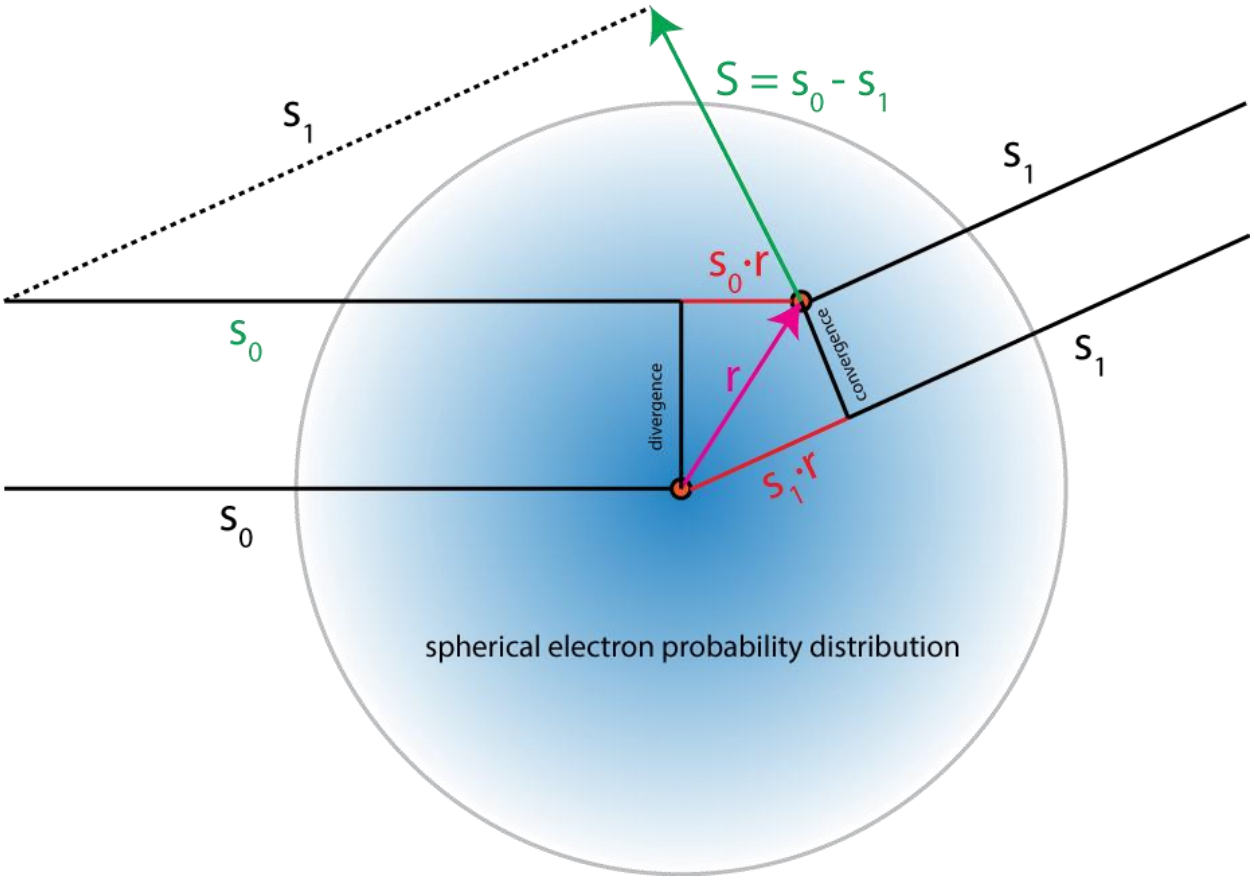
Depicted on the left is just the propagating ‘plane wave’, with a two-dimensional representation of a ‘side-view’. On the right, two separate electrons are introduced into the picture. These are a distance $\boldsymbol{r}$ apart. This diagram is of course a ‘snapshot’ of two electrons in an atom experiencing the electric field of an X-ray – frozen at some instant in time. The electron in the back experiences the ‘downward’ electric field vector, and the electron in the front experiences an ‘upward’ electric field vector. I can also illustrate this in a time-dependent manner, where the wavefronts move as the wave propagates. Let’s look at this idea, and see what kinds of oscillations are induced in electrons that are certain distances apart.

Below, I have labelled three arbitrary timepoints, t = 1, t = 2, and t = 3. The wave is moving on your page from left to right. Depicted are three different electrons located at arbitrary positions in space. The yellow and purple electrons experience the same maxima and minima at all times during wave propagation. As a result, they oscillate in unison. The green electron however, is always experiencing the opposite field vector as that of the purple and yellow electron. As a result, the scattered wave that it emits is exactly π radians out of phase with the other two.



This is only half of the story, though. It would be the full story if the scattering occurred solely in the forward direction. But, since the X-rays emitted from the oscillating electrons are emitted in some new direction, and since the electrons I drew do not have identical vertical

components (positions in space), we have to account for the distance that the scattered wave from the first (yellow) electron travels before its wavefronts 'meet up' with that of the purple electron some short time later. We can look at this effect with a scattering diagram:



It is tempting to look at the above and think that an incoming X-ray (s_0) hits the 'lower electron' and then another incoming X-ray hits the upper one. Do not think of it this way! As I mentioned before, there is a plane wave continually propagating through the atomic volume. The line labelled in 'divergence' is the point at which the resultant scattered waves being travelling different paths. The line 'convergence' represents the point at which there are no more path difference between the scattered waves (s_1). In other words, they have met up. So what is the distance before the plane waves meet up?

This distance is calculated by finding the difference of the two lines highlighted in red. Since they are sides of a triangle related by the vector **r̂**, we can calculate these lengths using the dot product (see preface section if this looks unfamiliar).

$$\Delta path = s_1 \cdot r - s_0 \cdot r$$

Now we factor out the r:

$$\Delta path = r \cdot (s_1 - s_0)$$

We can see from the diagram that our scattering vector, $\boldsymbol{S}$, is described as the difference of the incoming and scattered waves, so our term in brackets simplifies to S.

$$\Delta path = r \cdot S$$

Now we can convert this path difference into radians and express this as a phase difference (how many cycles of the wave elapse while the scattered wave travels a greater distance?)

$$\Delta\varphi = \frac{2\pi}{\lambda} \times r \cdot S$$

If we choose to define our wave vector as having a magnitude of $\frac{1}{\lambda}$ then our scattering vector will also have this unit, and this becomes reduced to the critical crystallography expression:

$$\boldsymbol{\Delta\varphi = 2\pi S \cdot r}$$

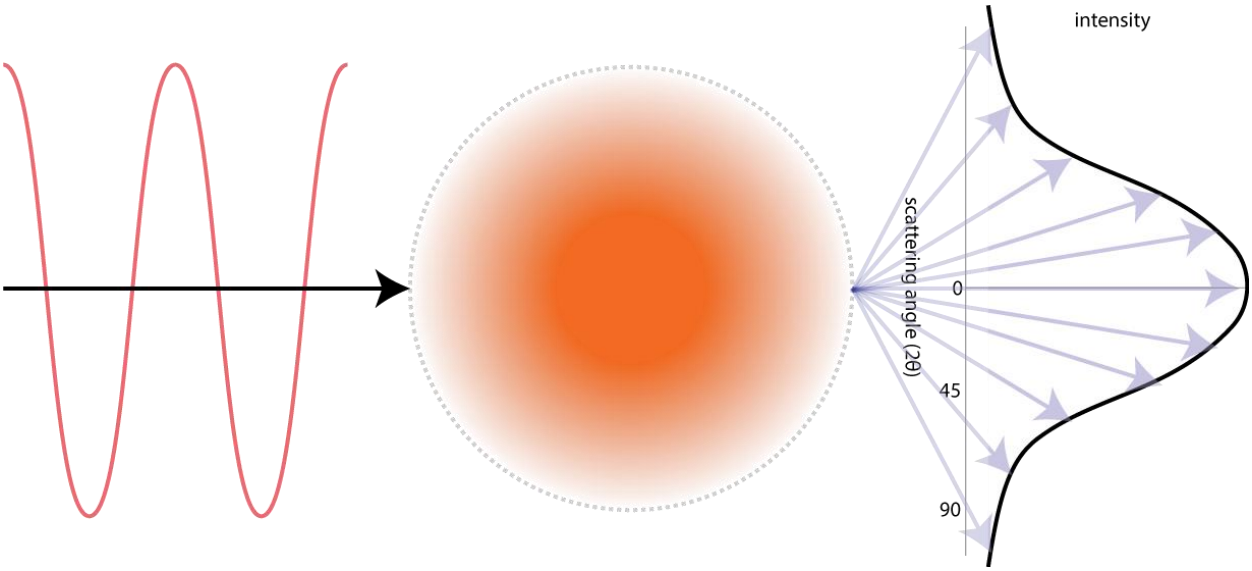
This says "the phase difference between waves scattered from two scattering objects can be calculated by the dot product of the scattering vector with the vector representing the distance between the two objects modulo 2π.

We have been representing our electrons as two points in space, but it's important to point out that electrons spatially occupy the surrounding atomic nucleus in a probability distribution. Thus, just as we arbitrarily selected two points to shoot electrons at, and for many photon

scattering events, we have to consider what the situation might look like for the entire spherical electron density. From our previous discussions, we remember that we can represent waves in Eulerian form. Mathematics allows us to consider every possible interaction between scattered waves from electrons at any point in the electron density cloud by infinitely summing together all of these possible wave interactions using integration:

$$f_s = \int_r^{V_{atom}} \rho(r) \times e^{2\pi i S r} \times dr$$

If we plot this function against scattering angle, we get a Gaussian distribution.



Bragg's Law

If you have done any supplementary reading on crystallography, you will have inevitably have heard of Bragg's law. Bragg's law is a beautiful way to simplify the 'net' result of the scattering and interference diagrams we have been discussing above. Many crystallography teaching resources start with Bragg's law as a description of what is going on, but I feel as though there are some elements of Bragg diagrams that can be very deceptive and misleading, as its depiction does not accurately represent what is happening physically. Some things to keep in mind before reading about Bragg's law:

1. Bragg diagrams are depicted as though there are two incoming, coherent (meaning they have the same phase) photons. This is not the case – it is one photon (remember the plane wave) exciting all electrons in its coherence length.
2. Bragg's law depicts reflections of photons from electrons along planes that act like mirrors; as if a photon comes in and bounces off the plane, then continues on its way. This is not the case – the electrons can scatter photons at any angle; the probability of each angle being given by the **Thomson equation**. The scattered wave is the sum of all of the virtual waves that result from interaction of the plane wave with all of the atom's electrons.

It is fine to ***interpret*** what is happening in Bragg's law as reflections from planes, so long as you understand there is more at play than the classical depiction of 'light bouncing off a mirror'. It is a nice simplification that helps to understand, on a larger scale, which objects will successfully contribute to diffraction.

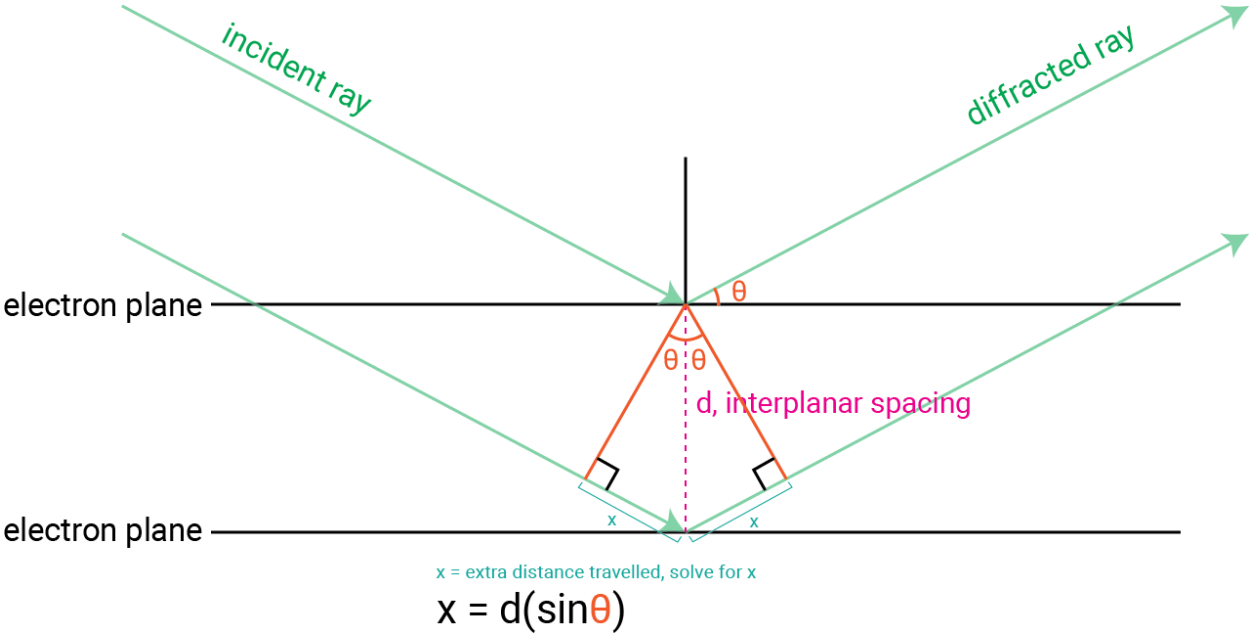
If we think back to the equation we just derived:

$$\Delta\varphi = 2\pi \boldsymbol{S} \cdot \boldsymbol{r}$$

We can interpret the difference in distance between scattering objects as the expression:

$$\boldsymbol{S \cdot r}$$

We should expect to get a maximum amount of constructive interference when our scattering objects are separated at a distance such that the wave can travel an integer (whole number) number of wavelengths. Bragg's idea was to turn the scattering vector on its side[6], such that the scattering vector points upwards:

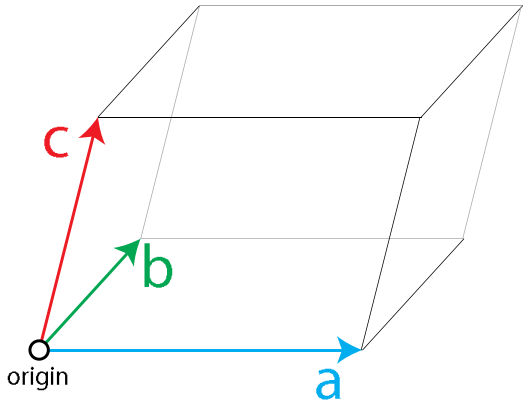


The 'extra distance' travelled by the lower X-ray is equivalent to $2x$. Thus, if we can determine whether or not the 'extra distance a ray must travel' for a given set of planes is an integer number, we can be certain that there will be constructive interference between photons.

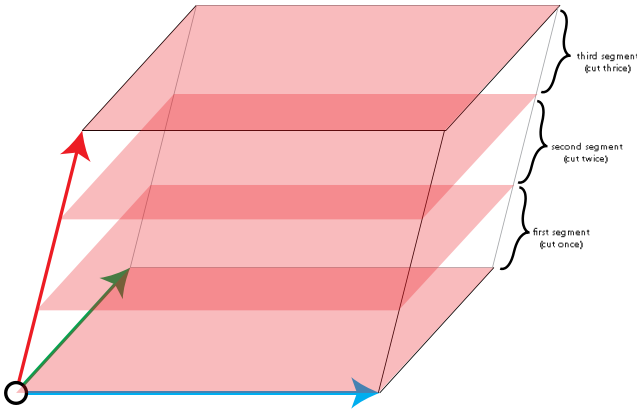
$$\boldsymbol{n\lambda = 2d(sin\theta)}$$

Miller Indices/(h k l) Planes

We have seen that maximal constructive interference occurs when our scattering objects are located a distance apart such that the second photon must to travel a distance equal to an integer multiple of the source X-ray wavelength (as to not offset the phase). Thus, we observe diffraction from ***discrete*** planes when we expose a crystal to X-rays. We therefore need a way to characterize and identify the planes that are providing us with reflections. We do this simply by subdividing our unit cell axes into integer numbers of segments. Let's take a look at a sample orthorhombic unit cell:



Now let's determine some discrete planes in this unit cell where we may observe constructive scattering from (provided we shoot the plane at the right angle!).



Here, we have cut the unit cell with 3 planes (the fourth one is technically part of the next unit cell). The three planes divide the cell into three segments. The axis being 'cut' is the **c**-axis. We

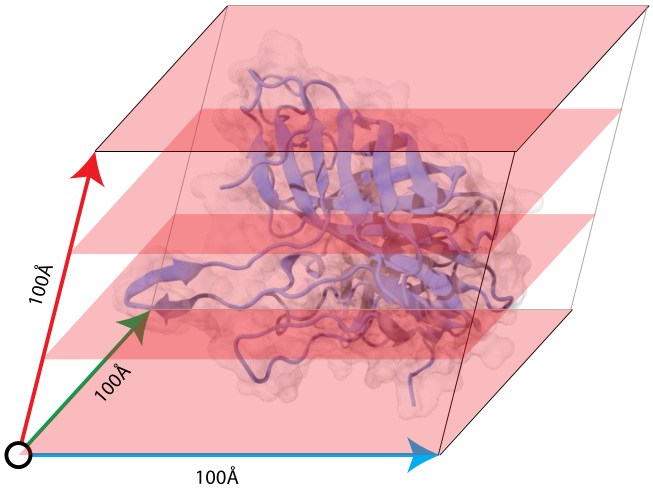
refer to this ***set of planes*** as the (0 0 3) set of planes. The general notation for a set of planes that divides a unit cell is:

$$(h\ k\ l)$$

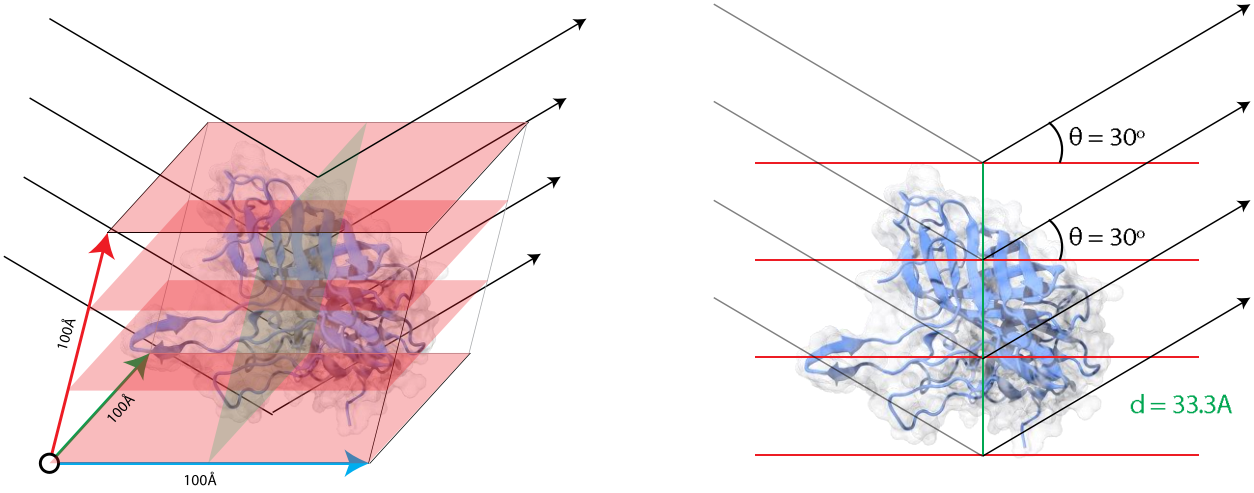
Where;

- h: the number of segments the 'a' axis is cut into by the set of planes.
- k: the number of segments the 'b' axis is cut into by the set of planes.
- l: the number of segments the 'c' axis is cut into by the set of planes.

Let's imagine now that we know the size of this unit cell. Let's say it is 100Å in each direction (in other words; **a**, **b**, and **c** all have magnitude = 100). Let us also place a protein into the unit cell so we have some electrons around that we can observe scattering from:



With this information, we can actually calculate whether or not we will observe scattering from these planes provided we know the angle at which we are shooting X-rays. In other words, let's draw in the pictorially-inaccurate but easiest-to-visualize 'incident X-rays' in our next drawing:



Above we have the 3D representation on the left, and the corresponding 2D representation on the right to make visualization of the calculations easier. Note that I have drawn a green plane in to allow us to focus on a single horizontal position as to make our calculations easier, but in reality, these photons may collide with any of the electrons lying along these planes. In fact, the incoming photons can collide with any electron they encounter – but what we are trying to see is whether or not we can 'observe' the collective scattering from the photons that have decided to collide with the (0 0 3) planes. The question right now is; will we see constructive interference if we shoot the unit cell at this angle? We need to know if our photon(s) travel an integer number of wavelengths between the planes before recombining. Only in this way is the phase conserved.

$$n = \frac{2dsin(\theta)}{\lambda}$$

The equation above is asking "how many multiples of the wavelength do the photons travel when colliding between these planes". We want to know if it is an integer value, that is;

$$1, 2, 3, 4, 5, 6 \ldots$$

So, let's plug in what we know about our scenario:

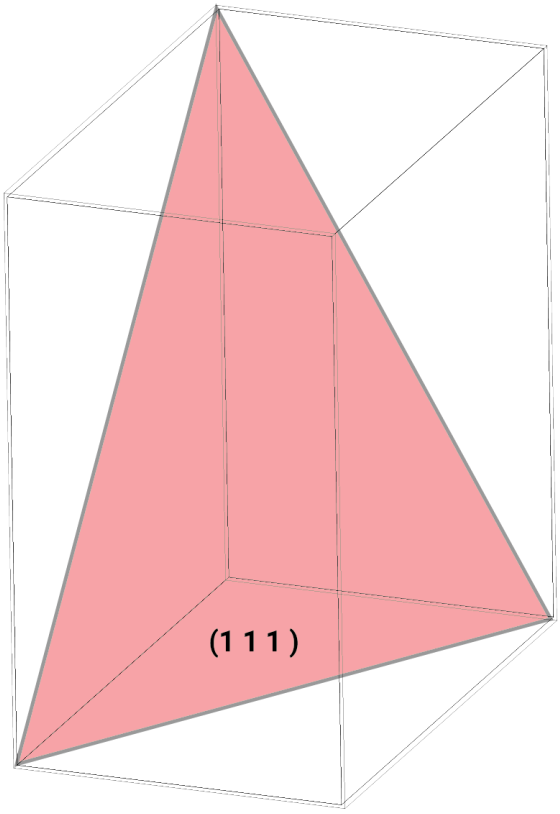
$$n = \frac{2\left(33.33333\text{Å}\right)\sin(30)}{1.54\text{Å}}$$

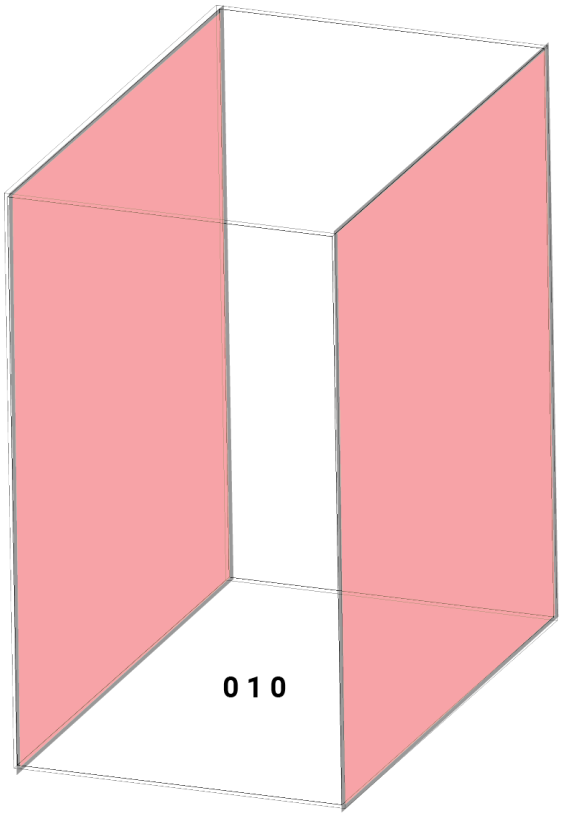
We are going to assume we are using a copper anode source for our X-rays which has a wavelength of 1.54Å. At synchrotron sources, you can actually tune your X-ray wavelength to almost anything you want (within a range).

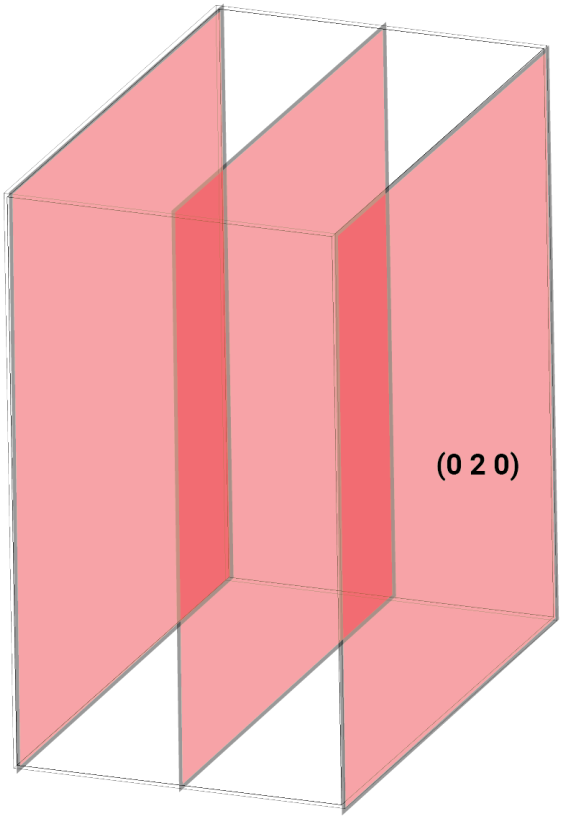
$$n = 21.645$$

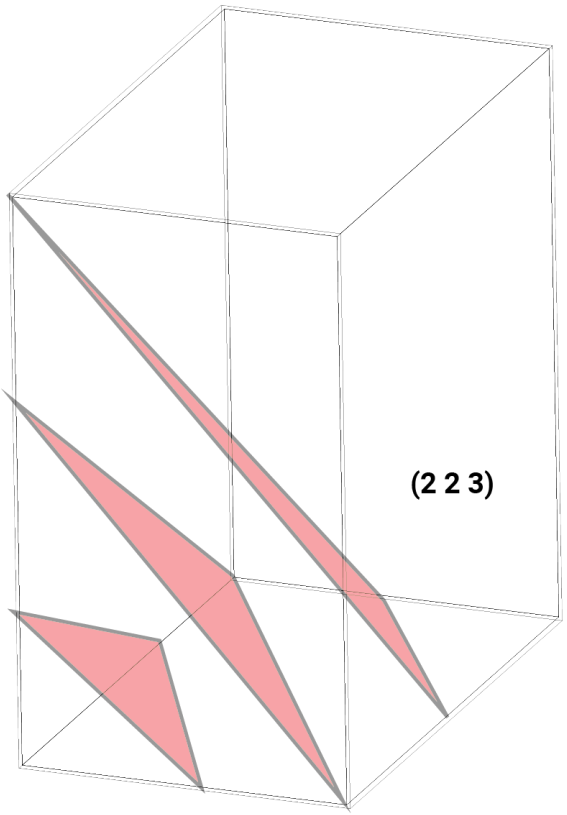
This is **not** an integer number of wavelengths! Therefore our waves will not constructively interfere. If we got the number 21, or 22, it would have worked. What if we shoot our planes at a much shallower angle… say… 1.32º? Try the calculation and see if you obtain an integer number (note you may need to round slightly; in other words… on a calculator, I would consider 0.997 to be an integer once slightly rounded 😊). Better yet, try to figure out several different angles you could should different planes of your choosing at in order to obtain constructive scattering (diffraction).

Here, you can see the utility of Bragg's law. I will give a few more examples of (h k l) planes so that you can familiarize yourself with the idea:


(1 1 1)


0 1 0


(0 2 0)


(2 2 3)

The Reciprocal Lattice

The concept of a reciprocal lattice is a difficult one to grasp. You will inevitably come across it in crystallographic studies – but what is it? Is the reciprocal lattice real? What does it mean for something to be in 'reciprocal space'? In our mind's eye we can picture real space with much ease, but how do we picture reciprocal space? The reciprocal lattice does not 'physically' or 'materially' exist. That is to say, we can physically ***see*** the real-space lattice under a powerful electron microscope. We cannot ***see*** the reciprocal lattice, but it still exists in theory, or, mathematically. In this way, you can 'imagine' a reciprocal lattice extending outwards from a unit cell origin. The purpose of talking about a lattice in reciprocal space is that it allows us to describe diffraction and the conditions required for diffraction very easily.

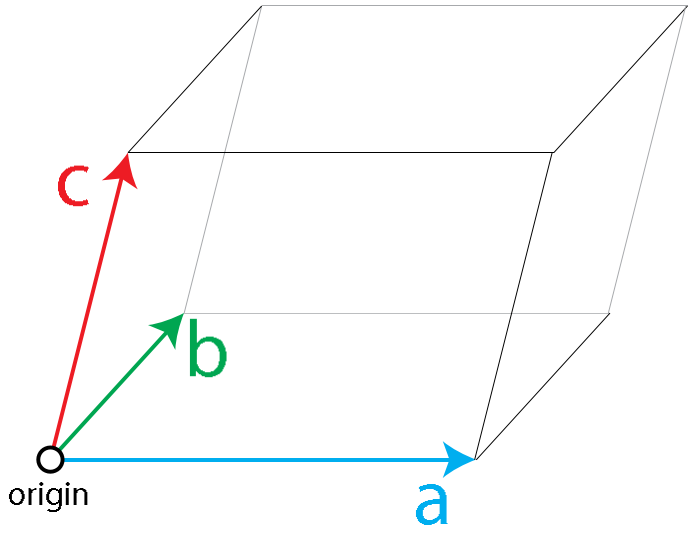
To fully understand the construction of the reciprocal lattice, it is necessary to have some understanding of linear algebra. Instead of diving into linear algebra, we can instead construct the reciprocal lattice of a unit cell using a set of rules that are *derived from* the linear algebra. In other words, let's learn how to construct the reciprocal lattice using plain-English math rules, instead of the math itself.

First let us understand some notation. We describe a real space lattice using the notation:

$$[0, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}]$$

Where;

- 0 represents the 'origin' of the unit cell
- **a** represents the first unit cell vector
- **b** represents the second unit cell vector
- **c** represents the third unit cell vector



When we talk about the reciprocal lattice, we indicate that we are talking about reciprocal space using an asterisk '*****':

$$[0, \boldsymbol{a}^*, \boldsymbol{b}^*, \boldsymbol{c}^*]$$

Where;

- 0 represents the 'origin' of the unit cell (it does not get a star – the reciprocal space lattice and the real space lattice share an origin)
- **a*** represents the first **reciprocal space** unit cell vector
- **b*** represents the second **reciprocal space** unit cell vector
- **c*** represents the third **reciprocal space** unit cell vector

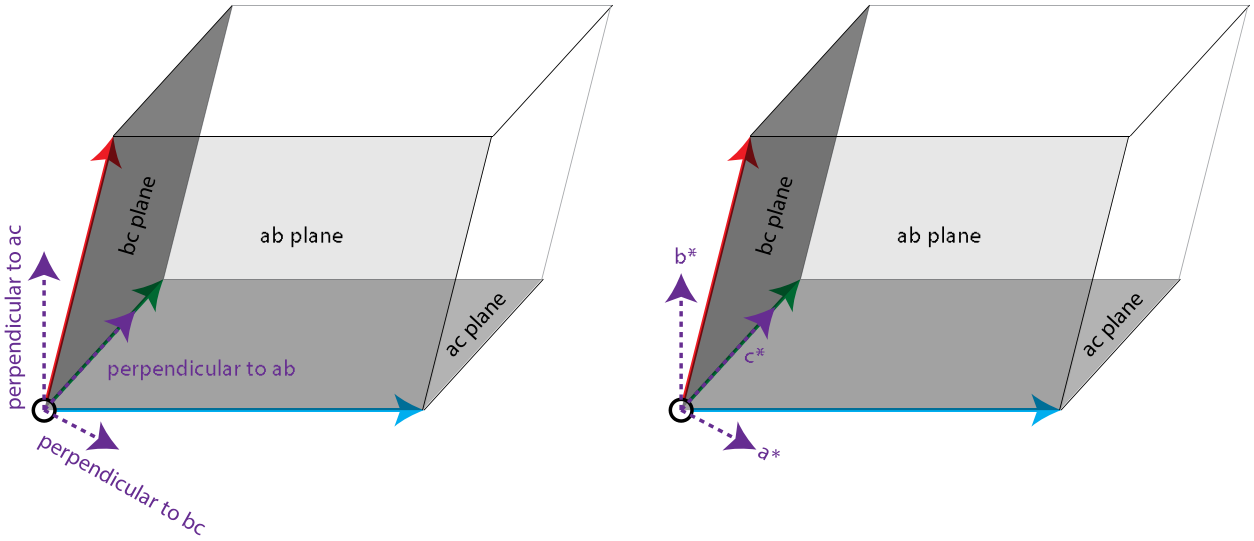
So how do we now figure out which direction our **a***, **b***, and **c*** point? How long are they? This is when we use the list of rules derived from the linear algebra relationship between the two lattices.

1. **a*** lies perpendicular to the plane created by **bc**.
   b* lies perpendicular to the plane created by **ac**.
   c* lies perpendicular to the plane created by **ab**.

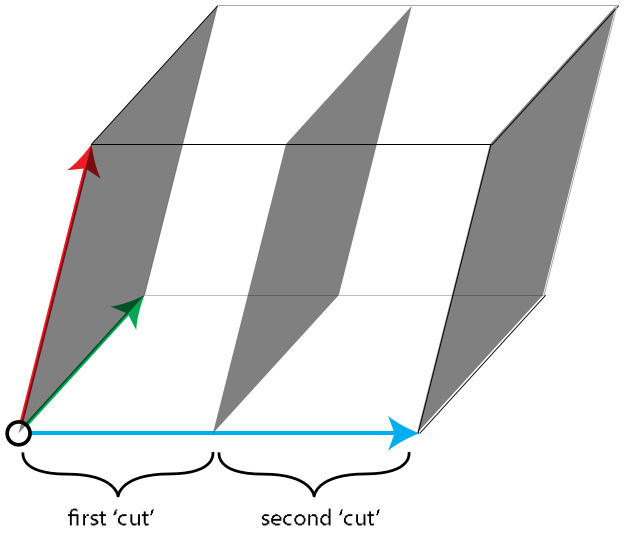
Remember; '**bc**', '**ac**', and '**ab**' refer to the real space lattice vectors.

2. The length of a reciprocal lattice vector is the reciprocal of its real space length. In other words, if in real space we have a vector of length 10, the reciprocal lattice length will be:

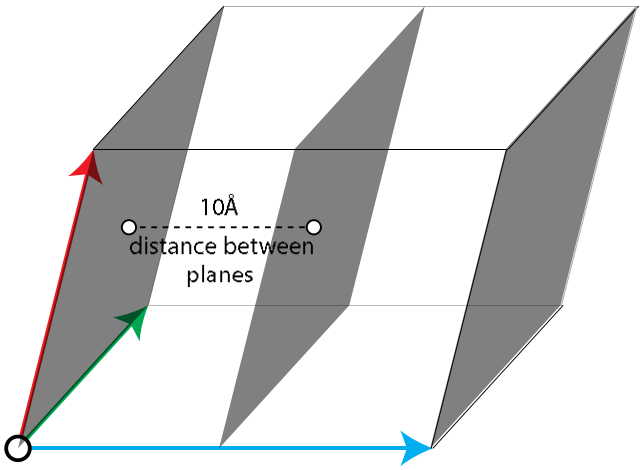
$$\frac{1}{10} = 10^{-1} = 0.1$$



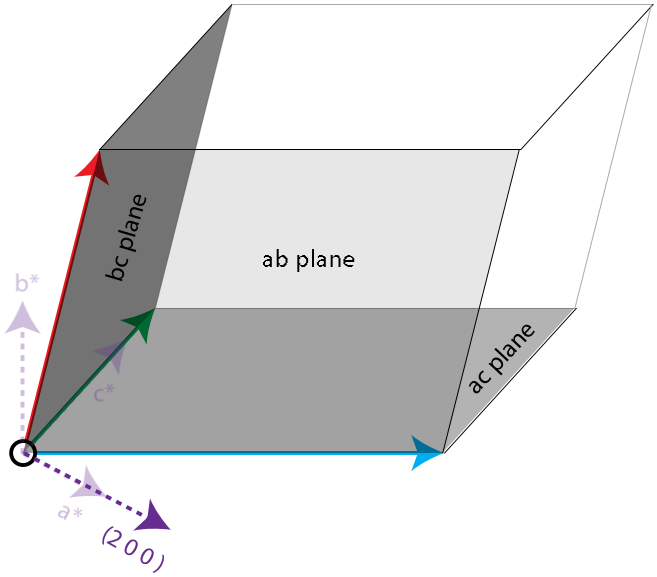
From how we have defined our reciprocal lattice so far, it is hopefully clear that for every (h k l) plane that we talked about in the previous section, we will have a reciprocal space vector that 'describes' that plane. To reiterate, I just told you that each plane in the real space lattice (ab, **bc**, and **ac**) is described in the reciprocal space lattice by the purple vectors. We can follow this exact principle for the hkl plane (1 0 0). In fact, this plane 'cuts' the **a** axis once, and runs along **b** and **c** – so, we have already described this set of planes using the reciprocal space vector **a***. Let's pick a different plane; (2 0 0). This 'cuts' the **a** axis twice, and runs along **b** and **c**. We can visualize it like so:



To fully describe this set of planes with a vector, we need to know the distance between the planes. Let's make up a number for this unit cell and say that the real-space **a** vector is 20Å long. That is to say, the unit cell in the **a** direction is 20Å in length. With this in mind, we can say:



I told you before that a real-space distance of 10 units has a corresponding reciprocal space distance of 0.1 units. So, now we know the direction and magnitude of the reciprocal space vector that describes this set of planes, and we can draw it like so:



Similarly, we could do this for the (3 0 0) plane, (4 0 0) plane, and so on. It would be pretty simple! They would all lie in the same direction as **a***, but the distances would **increase** as the spacing between the plane cuts became **smaller**. Let's prove this to you with some basic computation:

$$|(1\ 0\ 0)^*| = \frac{1}{\left(\frac{20}{1}\right)} = 0.05\text{Å}^{-1}$$

$$|(2\ 0\ 0)^*| = \frac{1}{\left(\frac{20}{2}\right)} = 0.1\text{Å}^{-1}$$

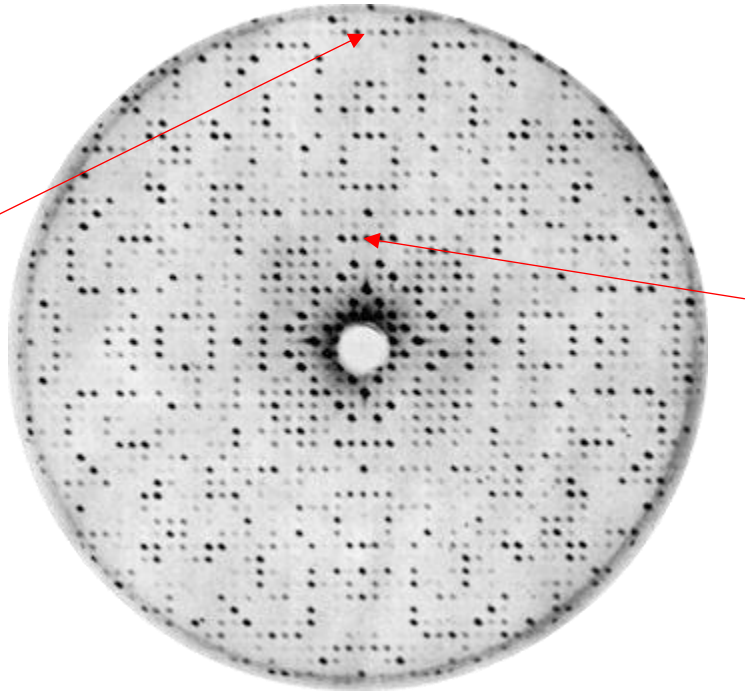
$$|(3\ 0\ 0)^*| = \frac{1}{\left(\frac{20}{3}\right)} = 0.150\text{Å}^{-1}$$

$$|(4\ 0\ 0)^*| = \frac{1}{\left(\frac{20}{4}\right)} = 0.2\text{Å}^{-1}$$

… and so on …

$$|(30\ 0\ 0)^*| = \frac{1}{\left(\frac{20}{30}\right)} = 1.5\text{Å}^{-1}$$

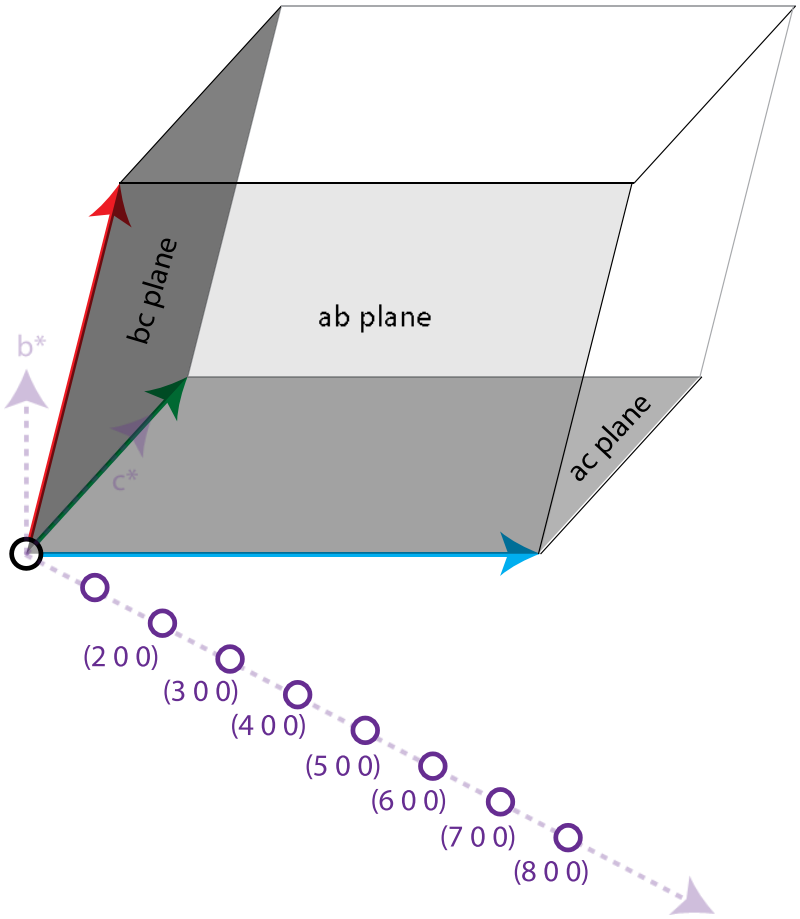
Notice how the reciprocal space distances grow larger as the size of (h k l) cut segment gets smaller and smaller (the **a** vector cut 30 times has a very small segment size compared to being cut 2 or 3 times for example). This is the concept of reciprocity and it is important to understand this when looking at the diffraction pattern (which is really just a visualization of the reciprocal lattice!). In other words, when we look at a diffraction pattern (precession photo), the reflections spaced closely to the origin correspond to planes with a large real-space spacing (low resolution information). The reflections spaces far from the origin correspond to planes with a small real-space spacing (high resolution information).



(23 0 0) plane represented in reciprocal space

(7 0 0) plane represented in reciprocal space

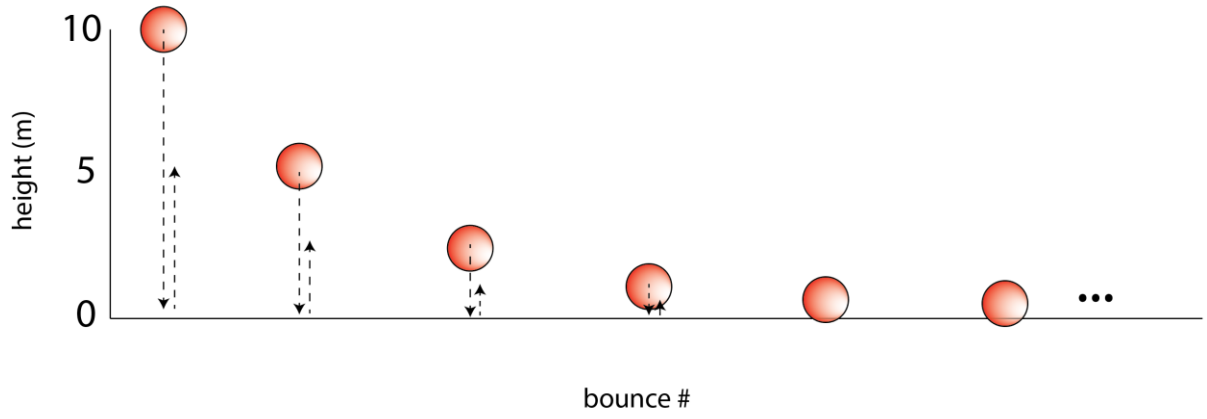
So our reciprocal lattice is starting to build up like so:



Each purple point here is a vector from the origin that describes a set of planes in real-space. To build up the rest of it, we simply would draw vectors perpendicular to all other planes – for example – (2 1 0), (2 2 0), (2 3 0) … and all conceivable combinations.

Fourier Transform – Mathematical Explanation

A **convergent series** is an infinite sum in which the limit is a finite value. For example, think of a ball on a ledge 10 m above the ground. Let's push the ball off the ledge. We know beforehand, based on this ball's physical properties that it will bounce exactly half as high as it did from the previous drop. So, it will fall 10 meters, bounce up 5 meters, fall 5 meters, bounce 2.5 meters, fall 2.5 meters, and so on. If we wanted to calculate the total distance that the ball falls, we would add together the terms 10, 5, 2.5, 1.25, all the way to the infiniteth bounce, until the sizes of the bounces got so small that each bounce contributed negligibly to the overall total distance travelled. In other words, the distance travelled by latter bounces is close to zero.



We can represent this mathematically with something called a ***convergent series***:

$$distance\ fallen = \sum_{n=0}^{\infty} 10\left(\frac{1}{2}\right)^{n}$$

Don't be afraid of the notation, the big sigma just means we are adding a whole bunch of similar terms together, from the range n = 1, to n = infinity. In expanded form, this would be:

$$distance\ fallen = 10\left(\frac{1}{2}\right)^0 + 10\left(\frac{1}{2}\right)^1 + 10\left(\frac{1}{2}\right)^2 + 10\left(\frac{1}{2}\right)^3 \ldots$$

$$distance\ fallen = 10 + 5 + 2.5 + 1.25 \ldots$$

$$distance\ fallen = 20m$$

These infinite series take on the general notation:

$$S_n = \lim_{n\to\infty} \sum_{n=0}^{k} a_k$$

What this equation means is, given an infinite sequence (ball fall distances: a_1, a_2, a_3,...), we are going to keep adding them together, all the way until n is infinity (a very, very large number). The 'complete picture' of how far the ball fell in total is given by the **solution** (S_n) to this series.

When talking about our situation with the bouncing ball above we say the ball bounces 20 meters after infinite bounces. This is called taking the limit of the series as n approaches infinity. In reality though, we may not be able to measure such small bounces with our instrumentation. Certainly a meter stick can measure the initial ~20 bounces, but as the bounces begin to reduce in their size, we need a very small measuring tape that can measure millimeters… and then micrometers… nanometers… etc. In this experimentally-ideal world, the ball continues to bounce forever, ignoring effects like our 'perfectly inelastic momentum transfer' during collisions, friction, air resistance, etc. Perhaps we are only able to measure the first 20 bounces. So, despite the ball

bouncing forever, if we were to calculate how far the ball bounced based off of our experimental data for the first 20 bounces, we would be taking a ***partial sum*** of our series (only the first 20 terms, instead of infinite). To compute the partial sum of a **geometric series** (the type we are working with right now), we can use this formula:

$$l = \sum_{n=1}^{k} a_n = a_n \left(\frac{1 - r^k}{1 - r} \right)$$

Where **l** is the value that the series converges to after you add together **k** terms, with common ratio **r** (which in our case is ½). It is not important that you know this formula, only that it is a quicker way of adding together many terms in one of these series (instead of plugging in to a calculator "10 + 5 + 2.5" for n = 3 and so on.

As a primer to understanding a critical idea about resolution in crystallography, lets imagine we are trying to figure out how far the ball fell with varying amounts of data. Perhaps we make three groups of students carry out this experiment with different measuring devices. The first group is only able to measure the height of the first three bounces accurately. The second can measure 5 bounces, and the third group can measure 10. Using our formula above, we can see how 'accurate' their assessment was of the total distance fell using the information available to them:

Group 1	17.5 meters
Group 2	19.375 meters
Group 3	19.9804 meters

With a greater expansion of this series being:

n = 1	10 meters
n = 2	15 meters
n = 3	17.5 meters

n = 4	18.75 meters
n = 5	19.375 meters
n = 6	19.6875 meters
n = 7	19.84375 m
n = 8	19.921875 m
n = 9	19.9609375 m
n = 10	19.98046875 m

We recall that when we took the limit of this series as the summation went to infinity, we calculated a value of 20m. We can see that as we get more information, our estimation of the situation becomes closer and closer to reality or 'the true value'. Group 1, who had less information, calculated a value further from 20 than group 2 or 3, both of whom had more measurements to include in their calculation. This idea relates to how closely you can calculate tiny position details in your crystal structure as well. How much information we can get about our crystal is a function of a few things including the wavelength of our X-rays (**$\lambda/2$** being our maximally-achievable resolution under ideal conditions), the sensitivity of our detector (can we measure very weak scattering from high scattering angles?), and displacive effects within our crystal (if the same position in two different unit cells have appreciably different electron density due to sidechain mobility, domain flexibility, etc., the scattering will undergo destructive interference, and we will lose this information for these positions). Another important analogy I should perhaps make is this: in order to get the value closest to 20 m for our bouncing ball, we cannot construct this solely with a high resolution information summation (for example, adding bounces 50 to infinity will still only give us a very small value). We need both the low resolution data (10 m, 5 m, 2.5 m) and high resolution data (0.001220703125 m on the 13^{th} bounce). In crystallography, we obtain scattering from 'low resolution' planes (i.e. 1 0 0 plane), and 'high resolution planes' (i.e. 22 28 16 plane). Summation of low resolution data provides overall morphology, and high resolution provides fine details. You can perhaps see why either of these

alone do not provide a comprehensive picture. Hopefully it is intuitive at this point why more diffraction data provides a 'closer to reality' model of our protein.

The title of this section is "*Fourier Transform – Mathematical Explanation*" and yet I have not talked about the Fourier transform at all. First, you needed to be introduced to series and their properties to understand how it is common in the physical disciplines to add together lots of little quantities to obtain a single larger quantity (this is the idea behind integral calculus as well). We have been looking at geometric series in our example, but the **Fourier series** uses **periodic functions** (such as $\sin(x)$ and $\cos(x)$), instead of our function $\mathbf{10}\left(\frac{\mathbf{1}}{\mathbf{2}}\right)^{\boldsymbol{n}}$ that we used before to represent the ball bouncing, to model some sort of phenomenon. For our protein structure case, the terms we are summing are no longer ball bounces but rather the structure factors (scattering from each ***hkl*** plane in the unit cell). Each ***hkl*** plane has its *own term* to be summed into the Fourier summation, and the more of these that we can measure (with accurate amplitudes/phases), the higher resolution and 'closer to reality' our model will be.

I have placed the two series we have considered so far side-by-side below in their mathematical notation so you can getter a better idea of how they are different. In our case, we are trying to model the electron density. Think of the x-ray photons as our 'measuring device' and each hkl plane effectively serves as a ruler of electron density. We are shooting them through the crystal, and these wave packets (photons) are interacting with other wave packets (electrons). The manner in which they do so provides us with diffraction data (when scattering occurs from a periodically arranged lattice, we call it diffraction). More diffraction data at higher resolution means we can reconstruct a better and better model of our electron density in that crystal unit cell. It is in this same way that we use light from a light bulb in the room to detect photon scattering from the painted graduations of a ruler into our eyes (the photon detector) to measure the length and details of our ball bouncing, or the dimensions of an object, for example. Isn't it beautiful to see how when we observe the world around us, it is really not much different from X-ray crystallography? At the root of it all, dispersed matter of all kinds exists all around us. We analyze our surroundings and obtain information on this matter by reflecting photons off of it into our detectors (eyes) and having our brain generate an image of what we think is really around us. Which begs the question, do our eyes 'see', or does our brain 'see'?

Geometric Series	Fourier Series
$S_n = \lim_{n\to\infty} \sum_{n=1}^{k} a_k$	$f(x) = \frac{1}{2}a_o + \sum_{n=1}^{\infty} a_n \cos(nx) + \sum_{n=1}^{\infty} b_n \sin(nx)$

It is important to remember Euler's formula at this time:

$$e^{i\theta} = \cos(\theta) + i \cdot \sin(\theta)$$

Remember here that the cosine of some angle represents the x-component of a vector, and the sine of an angle represents the y-component of that vector. Euler's formula is a way of representing the same vector by rotating it around the complex plane by that same angle. Do you see how similar Euler's formula (if it was written as a summation) is to the Fourier series?

It is common to hear of the operation performed on the set of structure factors obtained from diffraction data and phasing to be an '**inverse Fourier transform**'. This is because typically, in fields such a signal processing, we are taking a signal in the time domain (think about an electrical signal, or a complex sound wave from a recording microphone) and decomposing it into its constituent frequencies. Many times, it is useful to see which frequencies make up the signal so that we can modify them for engineering purposes (maybe you want to eliminate a high-pitch buzz from an audio signal; so you take the Fourier transform, and delete only the frequencies above 18,000 Hz, and then back-transform to get back your cleaned-up audio signal). For this reason, the Fourier transform is often regarded as an operation that converts a function in the ***time domain*** to the ***frequency domain***. Just as frequency is 1/time, space and reciprocal space share the same relationship.

Is it not absolutely mesmerizing that when we fire X-ray photons at a crystal, we obtain the Fourier transform of the electron density on our X-ray detector? To again quote Richard

Feynman on mathematics and Mother Nature: *"she offers her information only in one form; we are not so unhumble as to demand that she change before we pay any attention."*

The most concise explanation I have found yet for the Fourier transform is from Stuart Riffle:

$$X_k = \frac{1}{N} \sum_{n=0}^{N-1} x_n e^{i 2\pi k \left(\frac{n}{N}\right)}$$

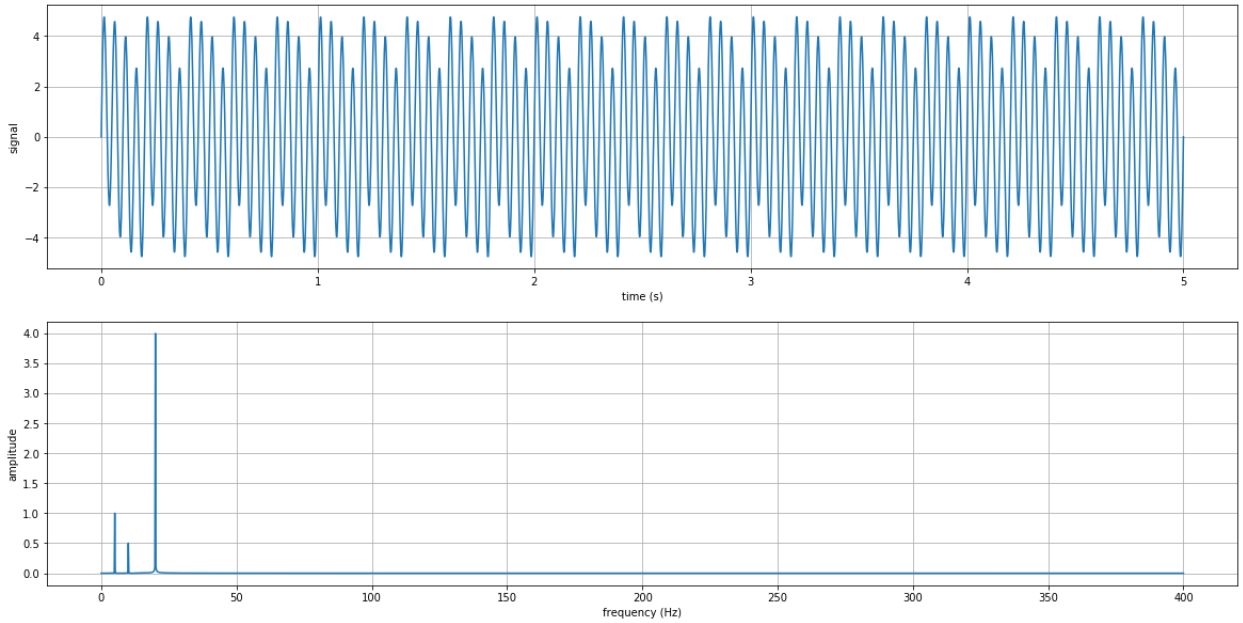
TO FIND THE ENERGY AT A PARTICULAR FREQUENCY, SPIN YOUR SIGNAL AROUND A CIRCLE AT THAT FREQUENCY, AND AVERAGE A BUNCH OF POINTS ALONG THAT PATH.

Example: 1D Fourier Transform

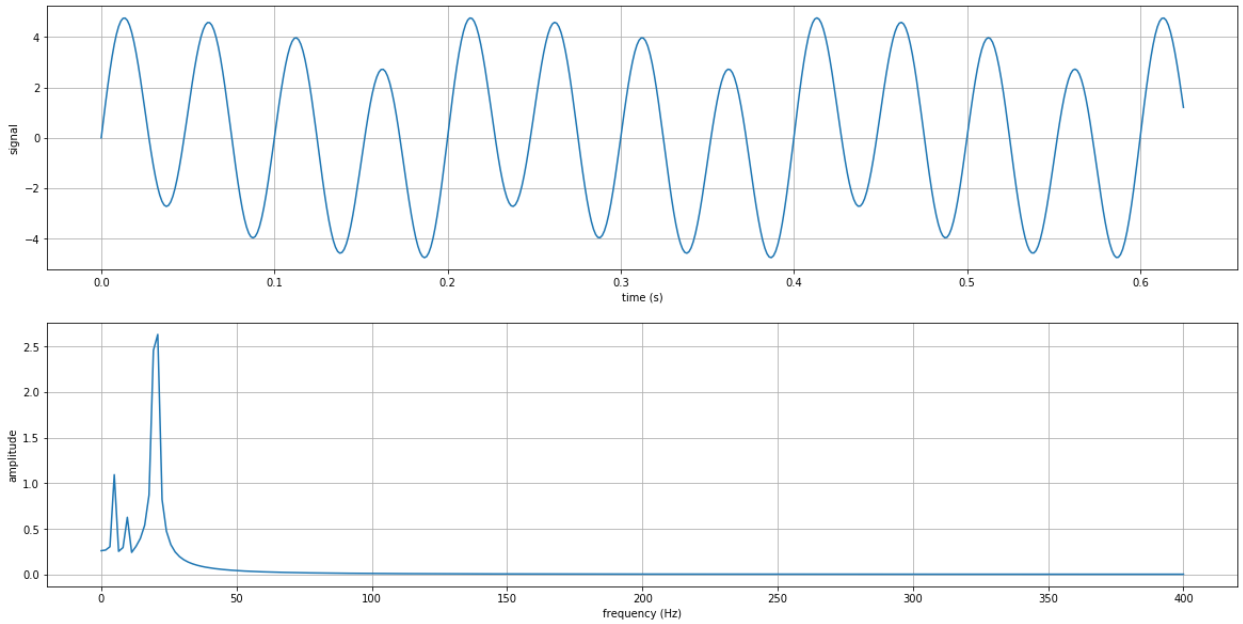
Using Python, I made a function that is simply the sum of three sine waves:

$$f(x) = \sin(5 \cdot 2\pi x + \pi) + 0.5\sin(10 \cdot 2\pi x) + 4\sin(20 \cdot 2\pi x)$$

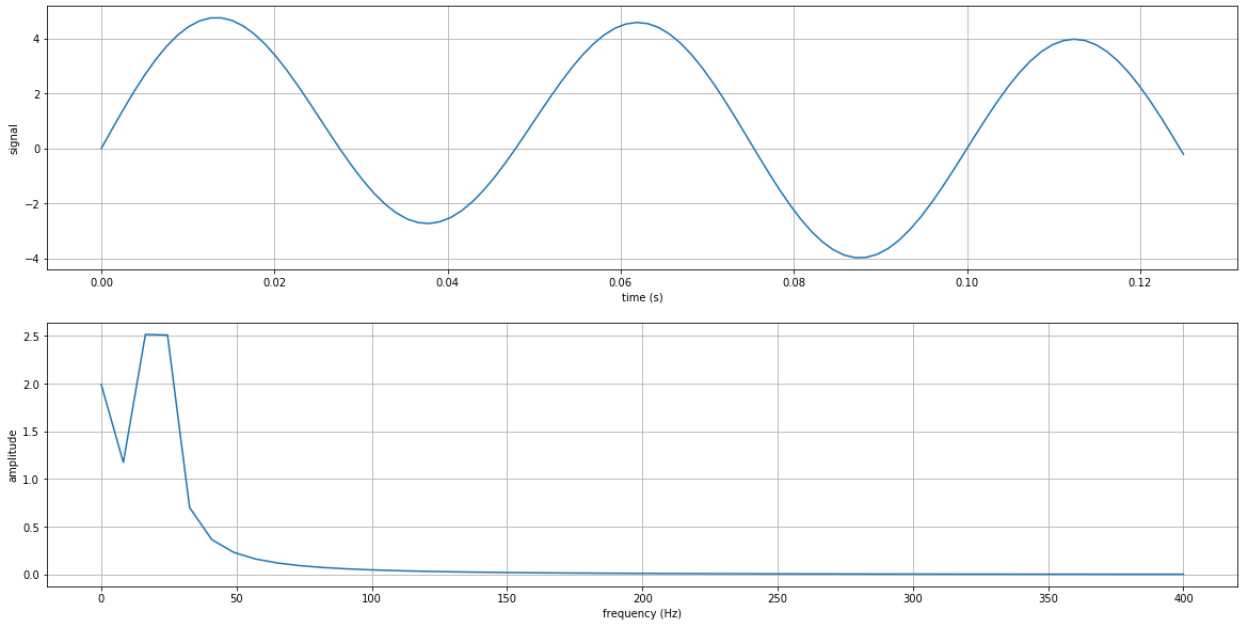
I then plotted this function over five seconds, with 4000 data points. The sum of these three waves can be seen in the upper subplot of the figure below. The lower subplot is the 1-dimensional Fourier transform of the upper signal, taking it from the time domain to the frequency domain. Three sharp peaks are observed at $x = 5$, $x = 10$, and $x = 20$. These values are the frequencies of the sine waves. The height of the peak represents the amplitude of the sine wave, or in other words, how much that specific wave 'contributes' to the signal.



Now, 4000 data points in this scenario implies very high resolution. That is why the peaks are so sharp in our frequency plot. What do you think will happen if I lower the number of points recorded in our signal? Note that our hypothetical instrument records one data point every 1/800th of a second. Let's lower it to 500, and see what happens.



With the number of points cut down by eight times, we only have signal for roughly half of a second. Our upper plot appears more ‘zoomed’ in, and it is perhaps slightly more difficult to see the periodicity of the function on that scale. Interestingly, the Fourier transform provides ‘broader’ peaks, because there is now less information available. If we cut down the signal record duration even further, to say, 100 data points:



Our frequency plot has become essentially useless. We can tell that the frequency is made up of some combination of low frequencies, but there is great uncertainty. We cannot even tell how many frequencies make up our signal anymore, let alone what the frequencies even are!

But what about the phases? If we were asked to reconstruct the signal from just the amplitudes, it would be very difficult. Sure, we know the amplitude of the three waves but they could be offset from one another in a variety of combinations. This is the phase problem. Fortunately, since this is just an electrical signal, the phase can be determined. With X-ray light, a wave with an electric field vector oscillating 3×10^{19} times per second, which is then travelling at the speed of light … you can see how this might be difficult to measure. Seriously… snap your fingers once every second. In the time it took you between snaps, an X-ray electric field vector shooting through someone's broken wrist over at Toronto General Hospital just oscillated 30,000,000,000,000,000,000 times. And then there are even higher energy gamma rays…

As a fun aside, it is this same idea that governs Heisenberg's Uncertainty Principle. If we treat particles as 'wave packets in space', the uncertainty principle states that we cannot precisely know the position and momentum at the same time. Measuring where a particle is located *spatially* can be intuitively thought of as a wave packet with a very sharp peak, but limited to a small region

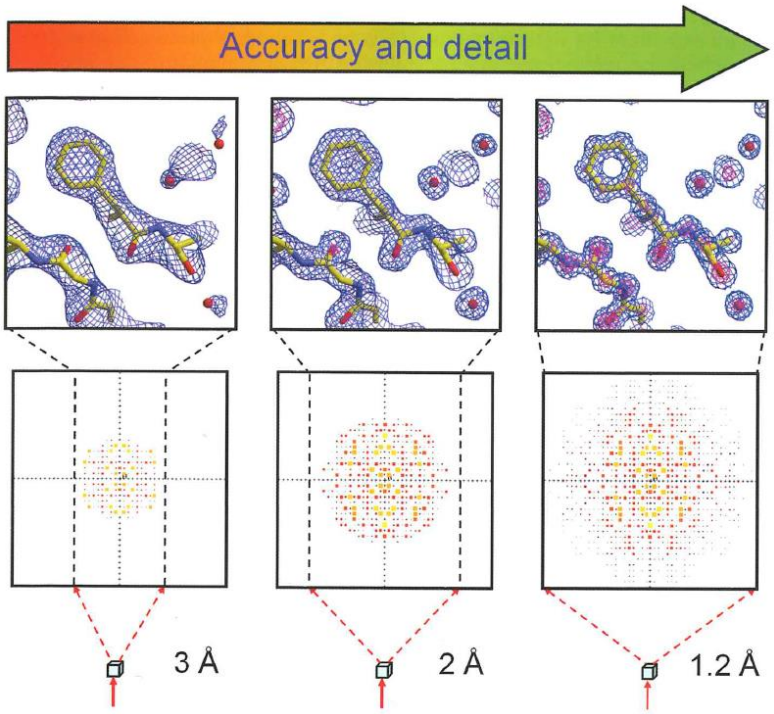
in space. A more diffuse wave allows for precise velocity measurement, but its position in space is more uncertain.

Example: 3D Fourier Transform

In our previous example, we have been looking at transforming some signal into some other representation:

$$time\ domain \rightarrow frequency\ domain$$

In crystallography, diffraction from the crystal gives us the signal in the reciprocal space domain. Reciprocal space is analogous to the 'frequency domain' construction of a time domain signal. It is our job to back-transform it into real-space. The figure below from Bernhard Rupp's *Biomolecular Crystallography*[6] shows how the resolution of our electron density map (and thus certainty in atomic positions in our final model) are affected by differing amounts of 'signal' from our crystal



Notice how the maximum scattering angle differs between the crystals. Can you think of how this relates to Bragg's law?

The Big Picture So Far[6,7]

Let's take a moment to recap what we actually need to solve a crystal structure, step-by-step.

1. Firstly, we need a crystal of our protein which is large enough in each dimension to provide useful diffraction data.
2. This crystal is mounted within a loop to a diffractometer or synchrotron source and exposed to photons with an energy in the X-ray region.
3. When the planes of electrons within the crystal are hit with X-rays, they oscillate, creating new X-rays of identical energy but of different *phase* and *direction*. **It is this process that defines the phenomena of diffraction**. At any given angle that we shoot our crystal at, only a subset of planes will be in *diffracting condition* (Bragg's law!).
4. These new waves recombine in both constructive and destructive manners, creating a diffraction pattern of many 'spots'. The spots that appear can be pre-determined by overlaying the reciprocal space lattice with the Ewald sphere.
5. These spots represent a **convolution** of the Fourier transform of electron density of the protein with the reciprocal space lattice.
6. Each 'spot' on our diffraction pattern is a structure factor ($\mathbf{F_h}$), composed of an **amplitude** (how dark the spot is) proportional to the square root of the number of electrons on that plane, and a **phase**.
7. We can measure the amplitude of the structure factor directly, but we cannot directly measure the phase. Unfortunately, the phase holds the important structural information – the amplitude, less so.
8. We use the *patterning* of the spots to determine the space group of our real space lattice. It is systematic absences of spots as well as symmetry within the diffraction pattern that allude to the symmetry within the unit cell.
9. We now want to calculate an electron density map using the equation below. Note that an electron density map is just an array of points with x, y and z coordinates, each of which has a density value. Coordinates which share similar density values are connected with contour lines and give rise to electron density maps.

$$\rho(x,y,z) = \frac{1}{V}\sum_{hkl} |F_{hkl}| e^{-2\pi i(hx+ky+lz)+i\alpha(hkl)}$$

10. The blue term represents the structure factor amplitude (measured directly), and the green term represents the structure factor phase. We now need to determine the phase component of each of our structure factors.

In practice, crystallographers employ three general methods to experimentally determine the phases of structure factors once they have obtained diffraction data. These three methods are **multiple isomorphous replacement**, **multi/single-wavelength anomalous dispersion**, and **molecular replacement**. For this course, it will be important to understand, in general, how each of these three methods work.

Phasing

By now, I have hopefully effectively conveyed the idea that in order to reconstruct a three-dimensional electron density model of our unit cell, we are adding together three-dimensional cosine waves in a variety of directions (related to their reflection planes). What the cosine wave represents is an approximation of the electron density perpendicular to the reflection planes. Now, think about what the amplitude and phase mean in this context. The amplitude will be an indication of how many electrons (or how much density) lies in a certain direction. The phase will contain positional information *about* this electron density. Amplitude = how many electrons? Phase = where are these electrons? How we offset the cosine wave approximations from each other is very important in correctly representing the electron density. If you have been doing some introductory reading on crystallography outside of this book, you have probably come across the idea that one can use completely random amplitudes and the experimentally-determined phases to reconstruct a model, or image of something with a Fourier transform – and it still looks roughly correct. Though if you use random phases and experimentally-determined amplitudes, you will obtain a completely invalid model.

So how do we determine phases? There are three major techniques used for determining the phases. In each case, the phases are calculated – not measured.

1. Multiple Isomorphous Replacement (MIR)
2. Anomalous Dispersion (SAD/MAD)
3. Molecular Replacement (MR)

Multiple isomorphous replacement relies on the incorporation of heavy atoms into pre-existing crystals. This can be done by soaking crystals in a solution containing the heavy atoms, which allows them to diffuse through the solvent channels and bind to the protein. During this process, it is imperative that the heavy atoms do not change the space group of the unit cell. This is because phases are calculated from comparing reflection intensities from the same Bragg planes before and after incorporation. If the diffraction pattern is suddenly different after incorporation, you can no long compare these reflections.

Single/multi-wavelength anomalous dispersion relies on the incorporation of heavy atoms directly into the protein. This is carried out during the over-expression step in whichever expression system the experimenter is using. An analog of methionine called ‘selenomethionine’ that has a selenium atom in place of the ordinary sulfur atom is added to the growth media. This causes proteins to have selenomethionine incorporated into them. Assuming crystals can still form with the selenomethionine incorporated, one can now shoot X-rays at a specific wavelength at these crystals and notice something interesting. The intensities of the reflections that are related through centrosymmetry of the reciprocal lattice will be consistently different. Using these differences, one can calculate the phases for each reflection.

Molecular replacement uses a pre-existing structure with high sequence identity to the protein you are trying to solve. The electron density for the structure is calculated and moved around (rotated and translated) within a virtual unit cell. The Fourier transform of the electron density of the virtual molecule is calculated, and compared to the experimentally-observed diffraction pattern for the protein you are trying to solve. Since the sequence identity is high, we can expect similar atoms to be in similar positions and thus provide similar scattering. Phases for the diffraction pattern are then inferred from the virtual protein in the virtual unit cell.

Calculating Electron Density

Once the intensities of the diffracted rays have been measured and their phases have been calculated in a **Fourier** synthesis, this information is used to calculate a three-dimensional map of the **electron density** of the repeating unit of the crystal (**unit cell**).

The proper equation is called the **electron-density equation**:

$$\rho(x,y,z) = \frac{1}{V}\sum_{hkl} |F_{hkl}| e^{-2\pi i(hx+ky+lz)+i\alpha(hkl)}$$

(x, y, z) = coordinates of a point in the electron density map;

V = the volume of the unit cell;

h, k, l = the components of the vector representing a diffracted ray;

$|\mathbf{F}_{hkl}|$ = square root of diffracted ray's amplitude;

α_{hkl} = the phase of diffracted ray

Weekly Questions

A) What are 3 methods could you use to determine whether a crystal you observed is composed of protein or salt?
B) If you were crystallizing a novel protein, how could you determine an appropriate concentration to set up screens with?
C) Why does altering the pH of the condition affect crystallization?
D) How do the concentrations of protein and precipitant change on a phase diagram after we set up a crystal drop?
E) Describe the three methods for calculating phases for reflections in a diffraction experiment.

Chapter 5

Practical Crystallography

Solving a Crystal Structure (in real life)

Solving crystal structures is largely automated now with the advent of high-powered computer processors and easily distributable software. Two recognized software suites for solving crystal structures include **Phenix** and **CCP4**. These are simply a collection of programs (algorithms) that do a lot of complex algebra and transforms for the user. For this course, we will be using Phenix because it has arguably the most intuitive graphical user interface. To visualize our electron density maps, we will use **Coot** – and it is invaluable to have **PyMOL** installed as well to view protein **coordinate files**.

Obviously, we have not collected diffraction data on our proteins – we are only in the screening stages. We can however carry out an exercise of solving the structure of another anti-CRISPR protein. Whenever a crystallographer aims to solve a structure of a protein, he or she must do so with a **strategy** in mind – how will they solve the phase problem? Does the protein have methionines (roughly 1 per 17kDa) that can be replaced with selenomethionine, or will they have to be engineered in? Is the protein similar to a known structure, or can it be complexed with one so we can solve using molecular replacement? These concepts are paramount in deciding the strategy to solve the structure, because heavy atom soaks are not very reliable as they often destroy the crystal, change its unit cell (making it anisomorphous), or do not incorporate at all. Probably the most relied-upon method to solve crystal structures is **single-wavelength anomalous dispersion (SAD)**. Funny enough, the name is technically incorrect considering no dispersive effects occur, since the dataset is collected at a single wavelength. SAD relies upon the incorporation of an anomalous scatterer, changing the intensities of the Friedel pairs within a single reflection set. This provides two **Harker circles**, leaving ambiguity in the structure solution. Thankfully, progress in **computational probabilistic methods** allow us to dedicate a bit of computational power towards figuring out what the phases are.

So, what do we need in order to solve a structure using SAD? We need to collect a series of diffraction images at an X-ray wavelength corresponding to the **absorption edge** of our anomalous scatterer (hereon chosen to be selenium). It is relatively straightforward to supply special amino acids to the organism used in our protein expression system, and so we replace methionine with selenomethionine (sulfur replaced by selenium). We then rely on the ability for

the protein to (hopefully) still form high quality diffracting crystals despite this incorporation. If successful, and the crystal provides anomalous signal during diffraction, SAD can be used to solve the structure.

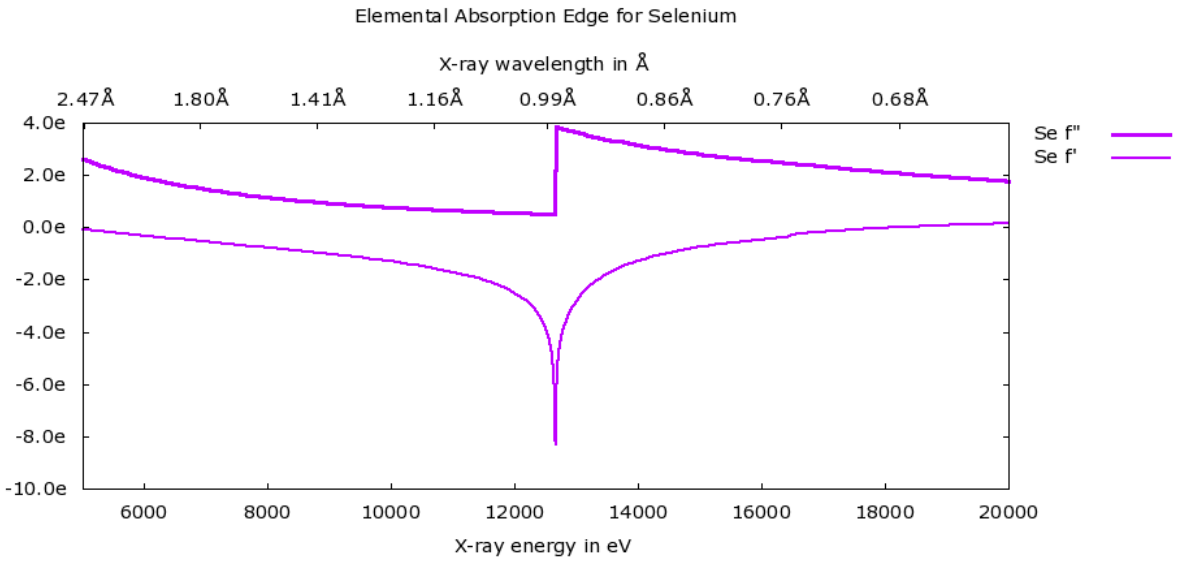
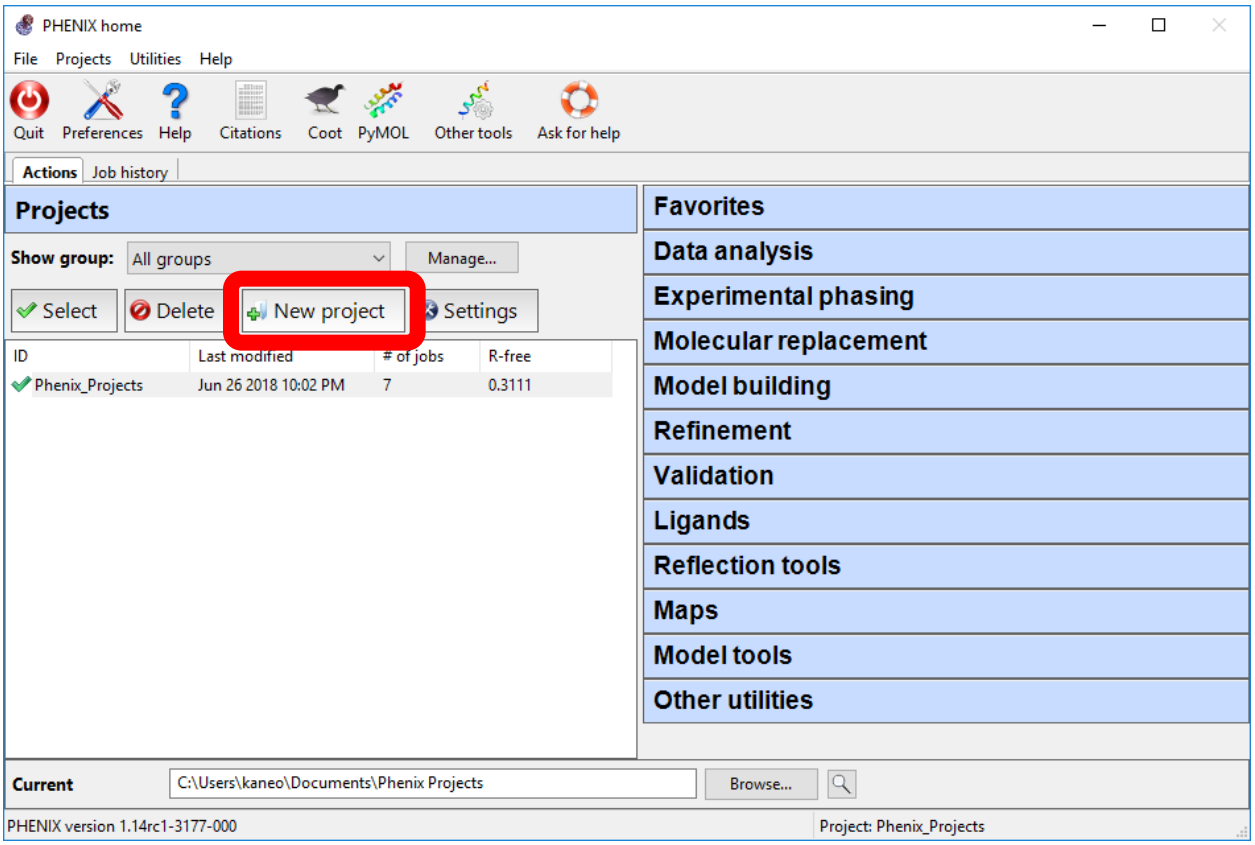
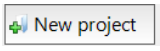


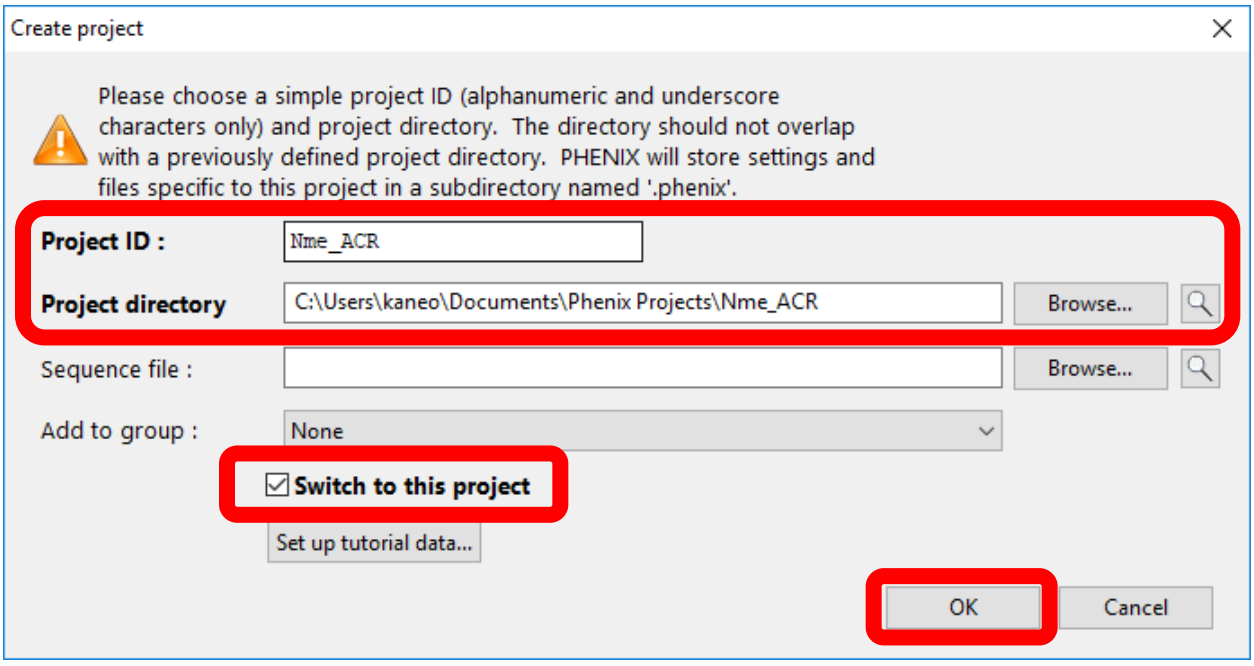
Figure 5.1. Elemental absorption edge for Selenium. Generated with Ethan Merritt's webtools (University of Washington).

Using Phenix

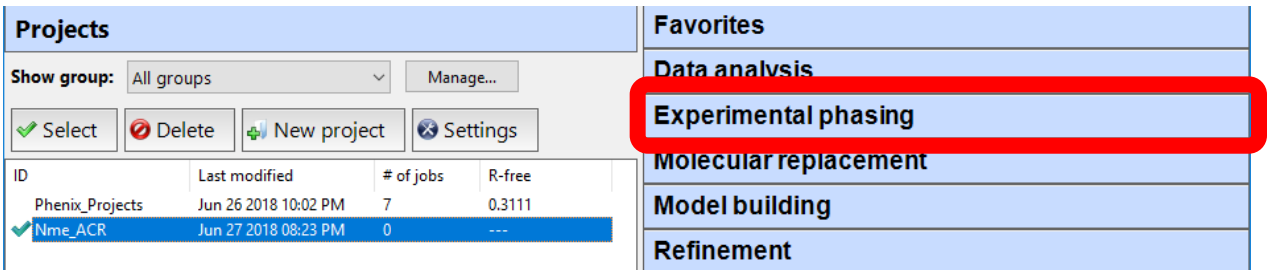
When you first start Phenix, you should arrive at a screen like so:



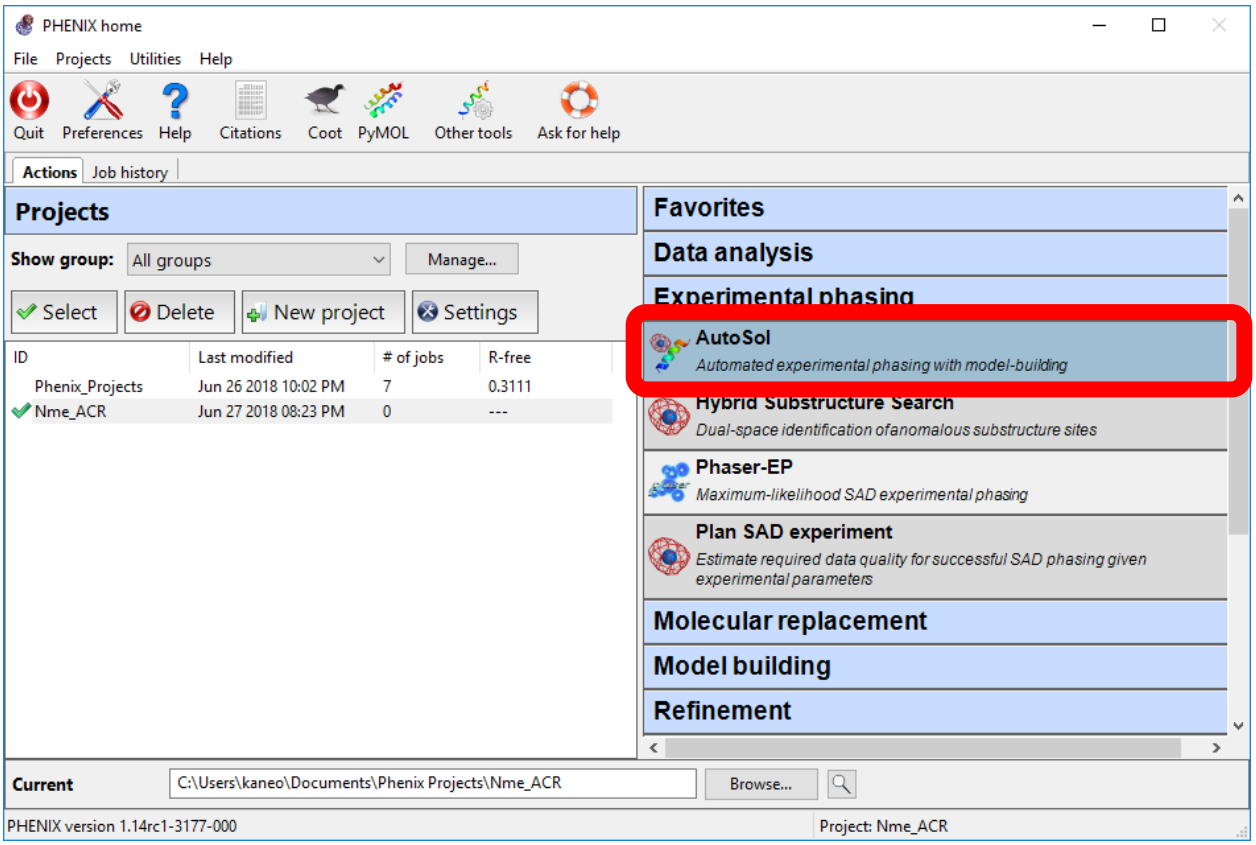
The left-hand window represents different projects (different protein structures you are trying to solve). The right-hand side represents the library of software that is available to you. We want to create a new project with our sample anti-CRISPR (ACR). Often, upon the first start-up of Phenix, it recognizes that no projects have been started and will prompt you to begin a new project. You can also create a new project by pressing the  button:



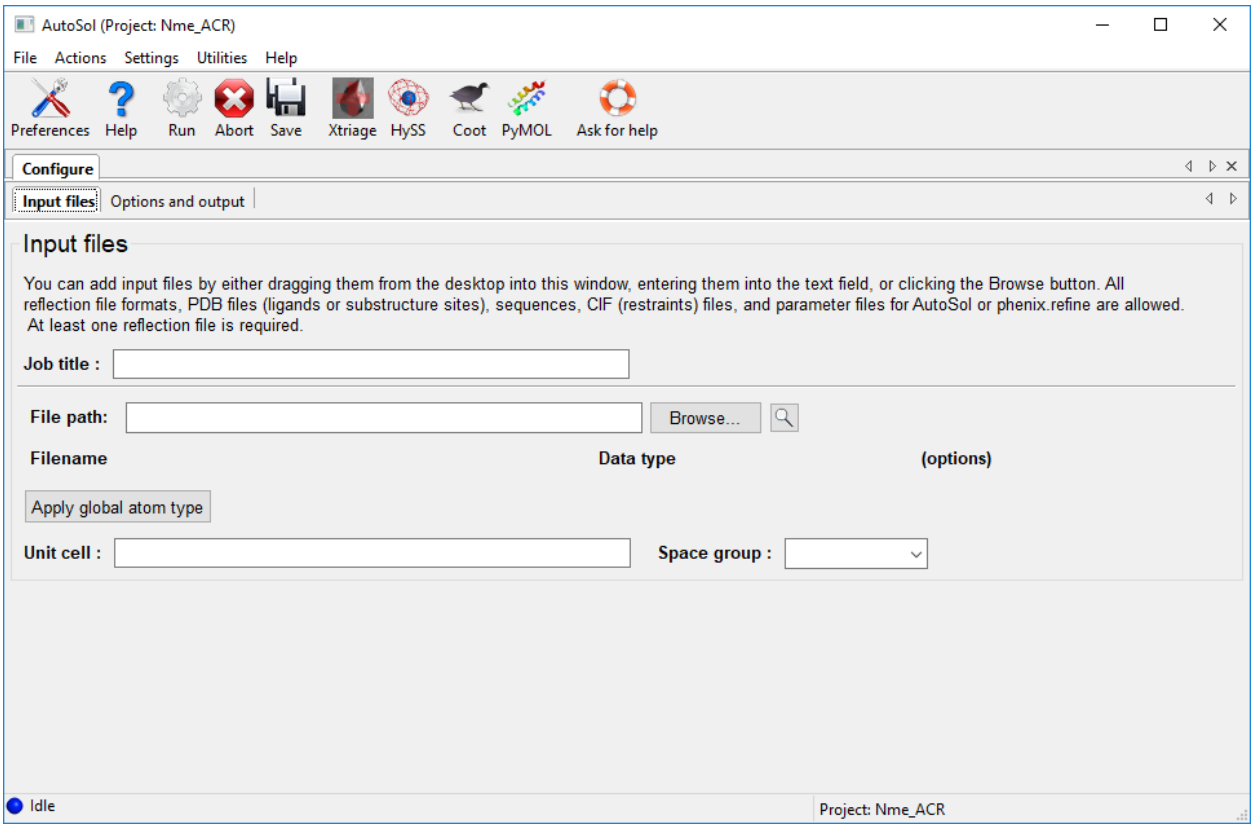
The ACR we will be solving is from a phage that infects an organism named *Neisseria meningitidis*. In short, we will call it **Nme_ACR**. Create a folder somewhere convenient on your computer (Desktop, Documents, etc.) named appropriately. For now, do not worry about supplying a sequence file. Press OK to finish. Now that you have created your first project, it should be automatically selected (green check mark beside the project). We are interested in performing **experimental phasing** as a way of solving our structure, so select it from the right.



You will be given several options to select from. Each of the selections are different pieces of software. Autosol is what we use to solve structures with SAD in Phenix.



Once Autosol is open, you should see a screen as below:

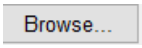


Realistically when solving structures, it takes many attempts with different datasets and strategies to do so. The "**Job title**" refers to one of many jobs that you will likely need to carry out in solving your structure. Luckily for us, we have a robust dataset which (with the correct parameters entered) should be able to solve on our first attempt. As stated before, we will need to supply the algorithm with our set of reflections (diffraction images turned into a language that the computer can understand). This file is referred to as a **.mtz file**. While I realize it may initially seem super uninteresting analyzing the innards of a file, one can realize that we are generating electron density, and thus a structure, from a list of numbers related to the crystal reflections. For this reason (and to have a comprehensive understanding of data collection), it is important to analyze this file manually. These files are binaries (you will not be able to open them with a text editor), but the contents can be dumped with a command line. For simplicity's sake, we have given a sample below of the header followed by some reflections.

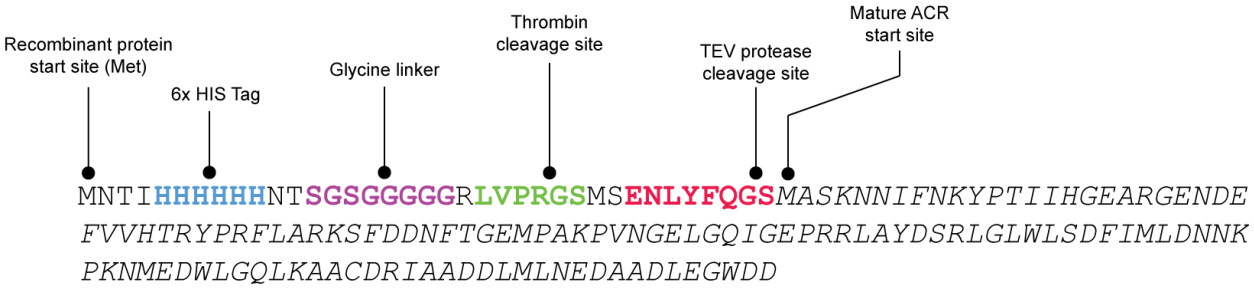
```
CELL    71.8800   71.8800  135.1400   90.0000   90.0000   90.0000
SYMINF   8  8 P    92            'P 41 21 2' PG422 X
```

```
0   0   2      626.00    112.00      3.00
0   0   4     9111.00    168.00     22.00
0   0   6      513.00    146.00     20.00
0   0   8     2610.00     52.00     10.00
0   0  10           ?         ?     11.00
0   1   1     1200.00     38.00     13.00
0   1   2     2244.00     55.00     21.00
0   1   3     2163.00     36.00      6.00
0   1   4     6057.00     82.00     13.00
0   1   5     3698.00     46.00     16.00
```

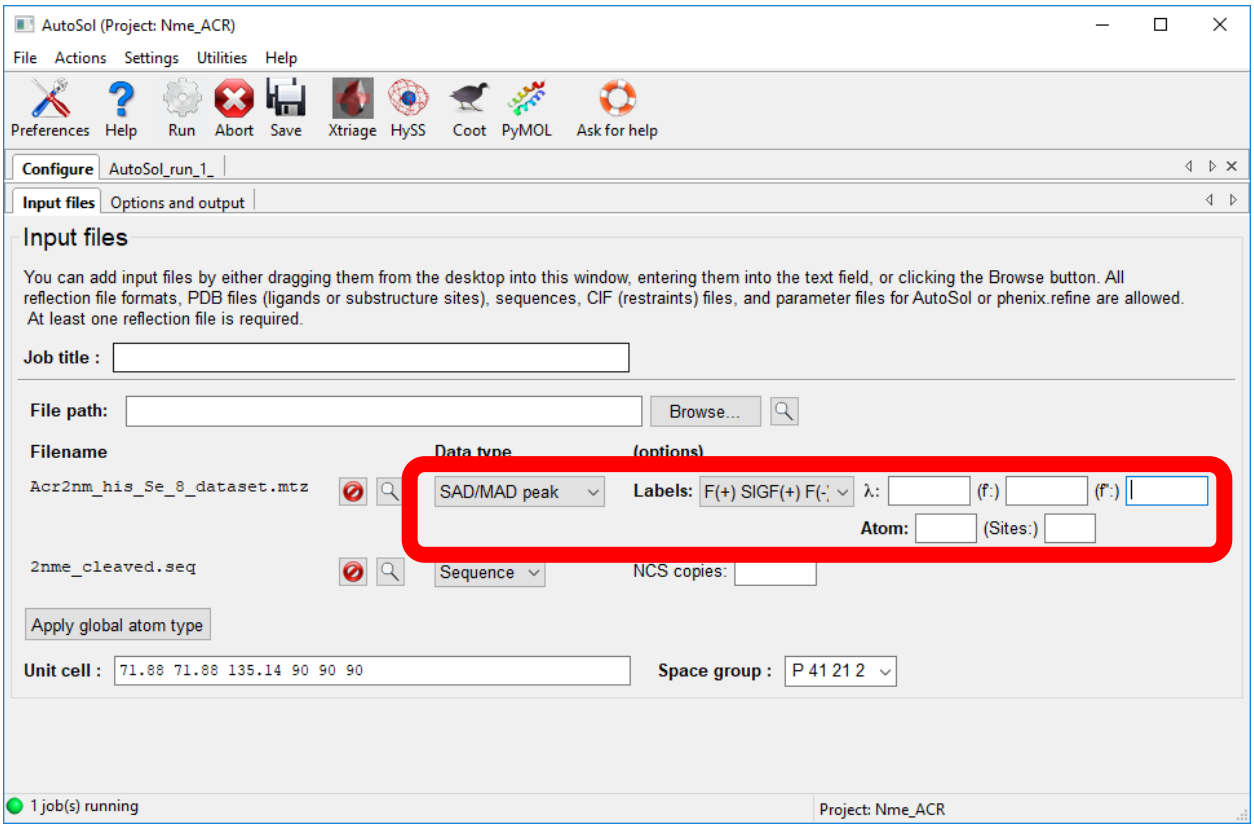
The first two lines (header) give unit cell dimensions, and the space group of the crystal. Remember, these are already known as soon as we get the diffraction pattern – it is the patterning of the spots (systematic absences and symmetry) in reciprocal space that tell us information about the symmetry within real space. The next ten reflections are structure factors related to each spot. The first line, for example, is data related to the **002 plane**. The next value **(626.00)** is the experimentally measured amplitude (related to number of electrons) of the structure factor, followed by the error in this measurement **(112.00)**. The mtz file contains many more data, but we will not discuss these values in depth. It is more important to notice how we can generate structures merely by looking at diffraction spots in numerical form.

Now let's stop nerding out and get back to solving the structure. We can now supply two files to Autosol – the reflections file (.mtz), and a sequence file. Do this by selecting the 

button beside file path, and start by selecting the .mtz file. Let's take a look at the sequence file before we supply it to Autosol:



This sequence represents the actual sequence of the protein that was cloned into an **expression vector** for bacterial expression. The protein, when purified, was purified with an affinity purification tag (His tag), and protease recognition motifs so that this tag could be cleaved off with Tobacco Etch Virus (TEV) protease or thrombin protease. These regions are positioned after a linker region to provide flexibility so that proteases can access the site easily. As discussed previously, unstructured regions (such as these tags) need to be eliminated if we want a high probability of our protein having conformational homogeneity (and thus success in crystallization). For this reason, the tag was cleaved off with TEV protease before crystal trays being set up, leaving us with a sequence of *SMASKNN*…. Edit the sequence accordingly and save the file. Now, supply Autosol with the sequence file. Your window should now appear as such:



The final critical information to include is highlighted in the red box above. "**Data type**" asks what kind of data the reflection file represents. Of course, we have one single dataset collected at a wavelength corresponding to the selenium absorption edge, so this is a **SAD peak** dataset. "**Labels**" simply refers to the columns in the .mtz file and these are automatically specified by the syntax within the .mtz file. The "**λ**" box is asking what X-ray wavelength the data was collected at. If we look back to our selenium absorption edge plot, we can see the peaks occur at an energy of **12657.8 eV**. This can easily be converted to a wavelength of **0.9795Å** using the $E = h \times v$ equation, where E is energy, h is Planck's constant, and v = frequency.

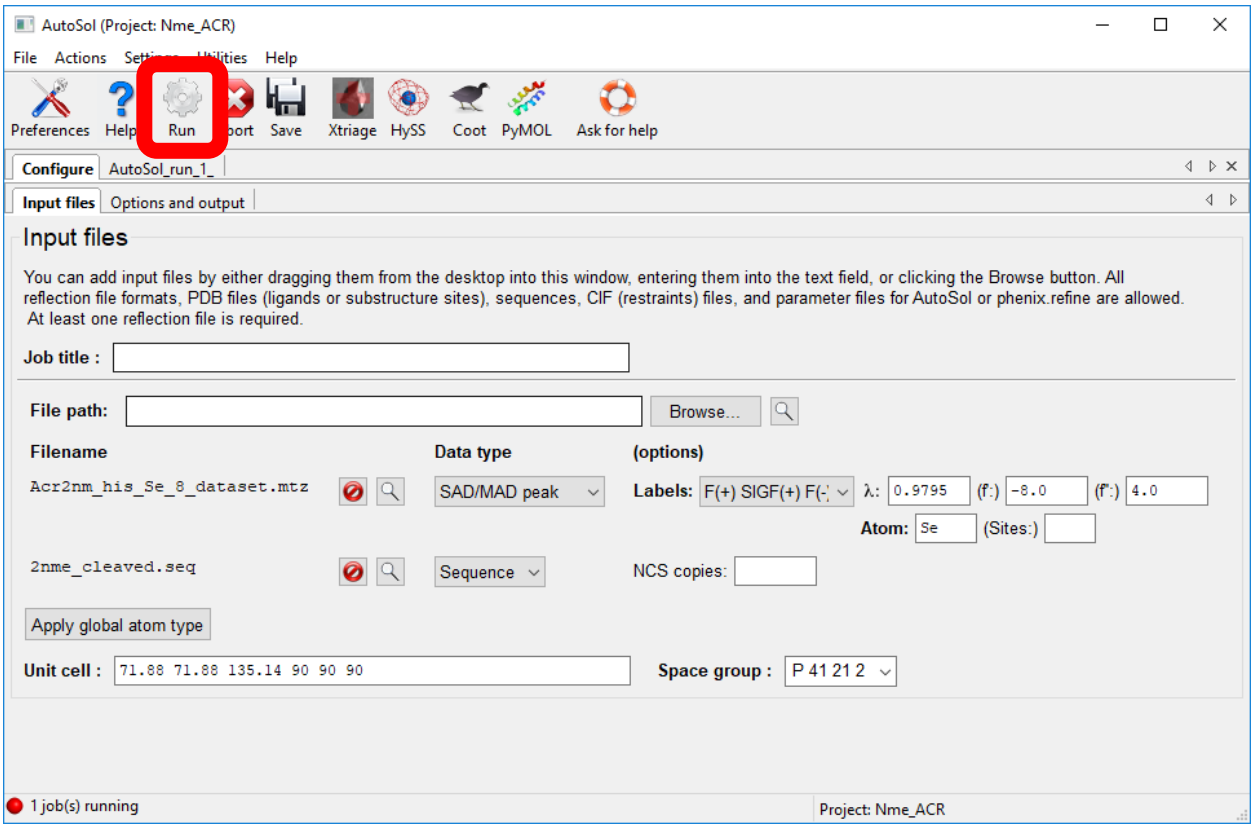
$$v = \frac{c}{\lambda}$$

$$E = h \cdot v$$

$$E = \frac{hc}{\lambda}$$

$$\lambda = \frac{hc}{E}$$

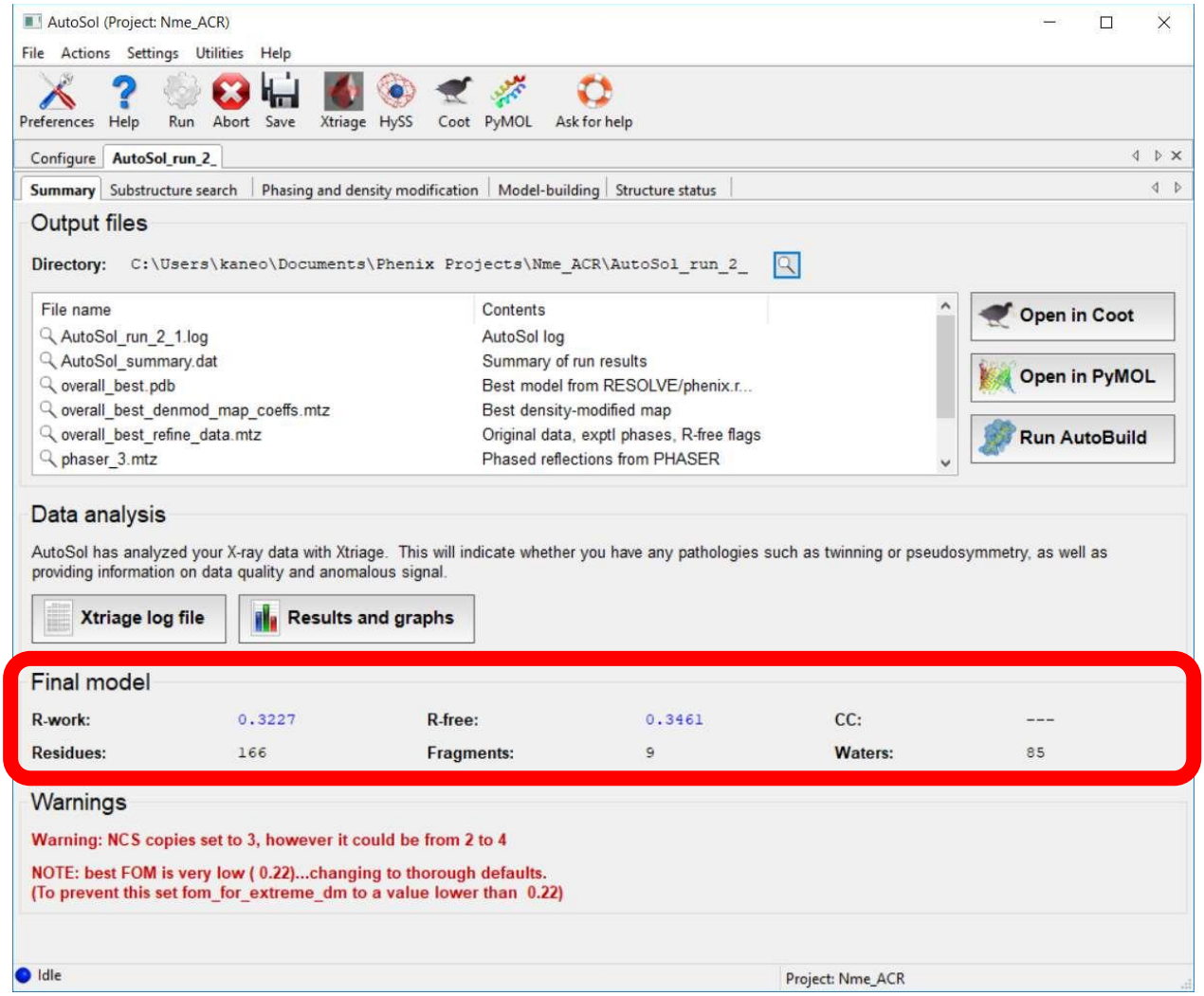
The **f’** and **f’’** are the real and imaginary scattering components from the absorption edge. The **f’’** contribution is not actually ‘imaginary’ but rather refers to an imaginary value on the complex number plane. Consulting our plot again, we find that these values are **-8.0,** and **4.5,** respectively. “**Atom**” is of course **Se**, selenium, and “sites” refers to the number of incorporations of this heavy atom within our unit cell. We can leave this field blank for the algorithm to figure out:



We can now select  and press "**Run Now**", and wait for Autosol to try to solve the phase problem for us. This may take a long time (30 min to several hours) depending on the power of your machine.

Visualizing Electron Density Maps and Polypeptide Models

Once your job finishes, Autosol should have outputted a ‘final model’ or ‘solution’. **Make note of the R-work and R-free values for this model in your lab notebook, and indicate which these values represent.**



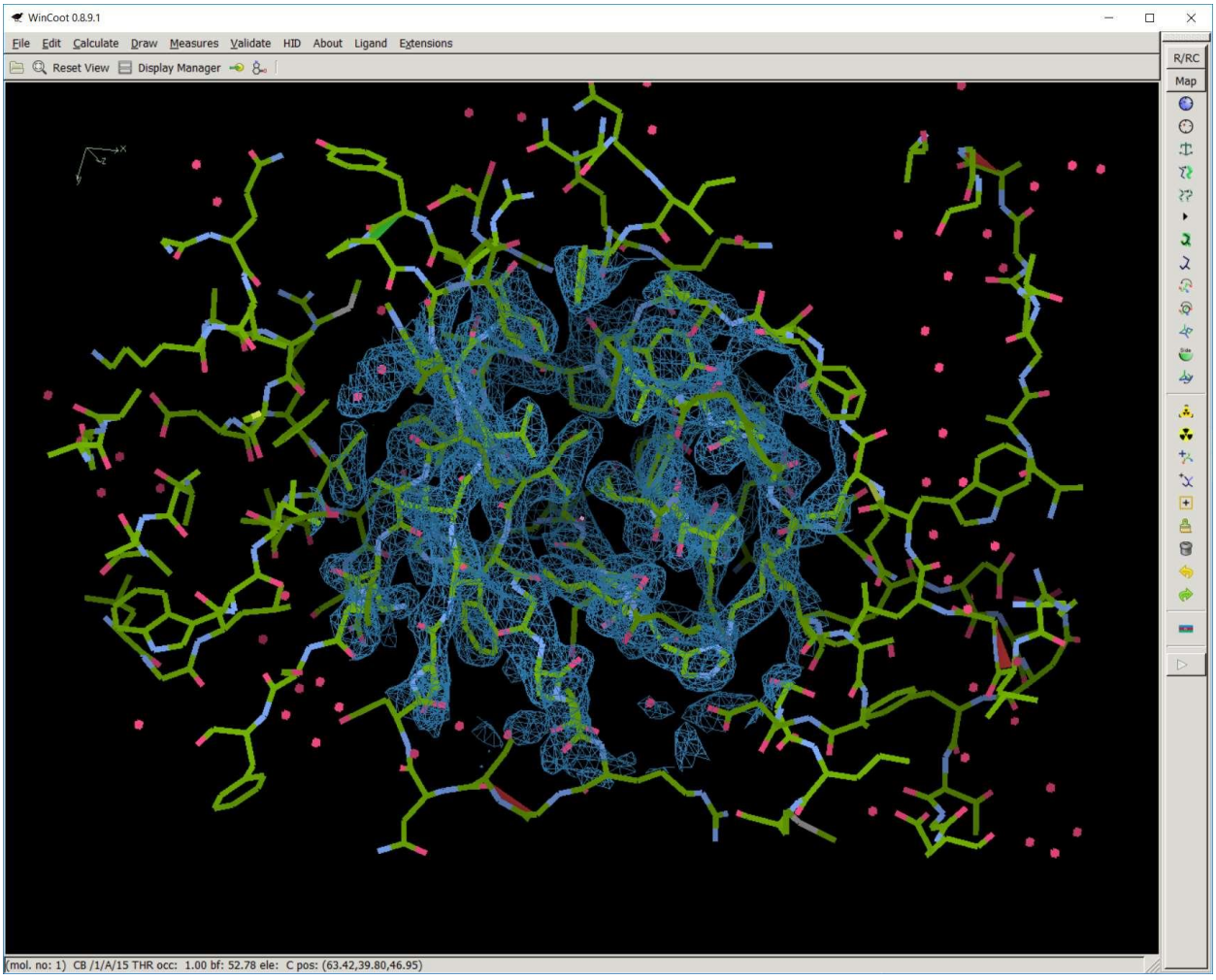
Go to the model building tab. Make note of how many residues have been ‘placed’. **Why might the algorithm not have been able to place all of the residues? If there are more residues placed than are present in the sequence file, what does this mean?**

Now we need to build residues into the electron density that the software cold not. We also need to refine the structure so that the bond angles agree with certain validation metrics and the electron density itself. Start by opening the solution in Coot by pressing “Open in Coot”. If Coot

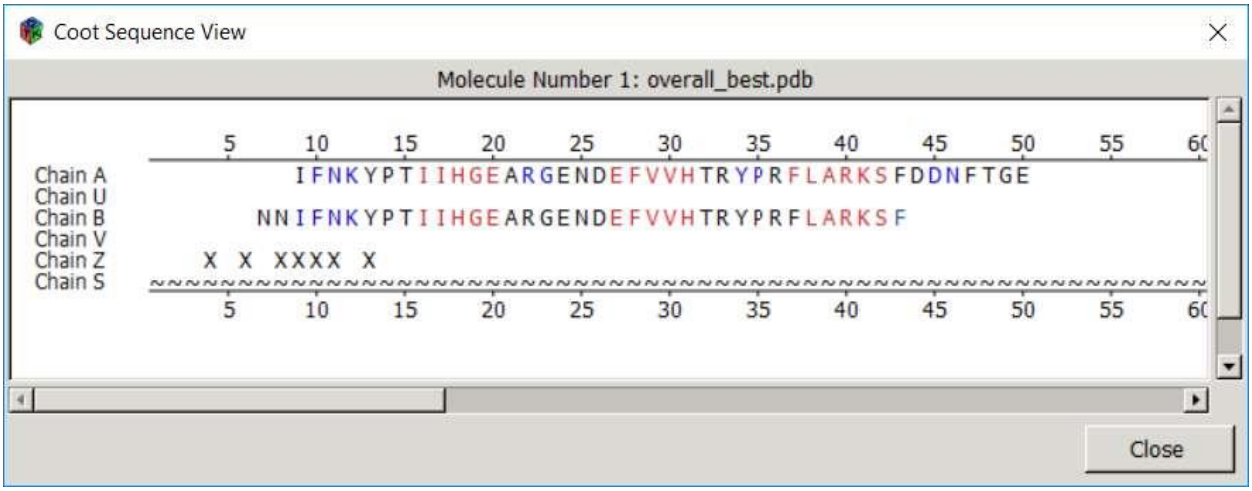
does not open, you may need to manually open Coot and go to **File → Auto Open MTZ**, then migrate to your project directory and your map will likely be named "overall_best_denmod_map_coeffs.mtz". If your protein coordinate map (pdb) does not also load with the density map, you will need to open it with **File → Open Coordinates**, and select the appropriate pdb file.

****Make sure throughout this process to frequently save your coordinate file using File→Save Coordinates. Coot will crash frequently and does not have an auto-save feature.**

Once your map and coordinate files are open, you should see a 'sphere' of electron density as per the image below.



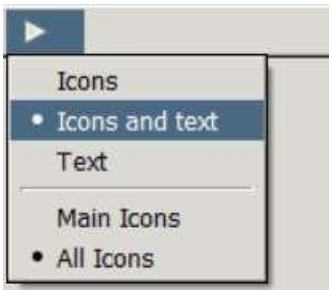
Now, go to **Draw → Sequence View → <filename>.** This should open a window like the one below. What is unusual about the sequence and chains in this window?



It is time to build and refine residues in the crystal structure. This is a slow and tedious but (arguably) enjoyable process. Start by going to **Draw → Go To Atom…** In Chain A, go to the first residue (likely an isoleucine; residue 9). You can also do this by double-clicking the residue, or clicking it once and pressing “Apply”.

Inspect the placement of the atoms in this residue. Does it agree with the electron density, or is it sitting outside of the density? If you think the residue is situated within the density appropriately, press your Spacebar to move to the next residue automatically.

Make sure on your right-hand sidebar, you select the small white triangle at the bottom of the tray and select **Icons and Text**.



The tools that pop up will be instrumental in moving residues around to fit the density.

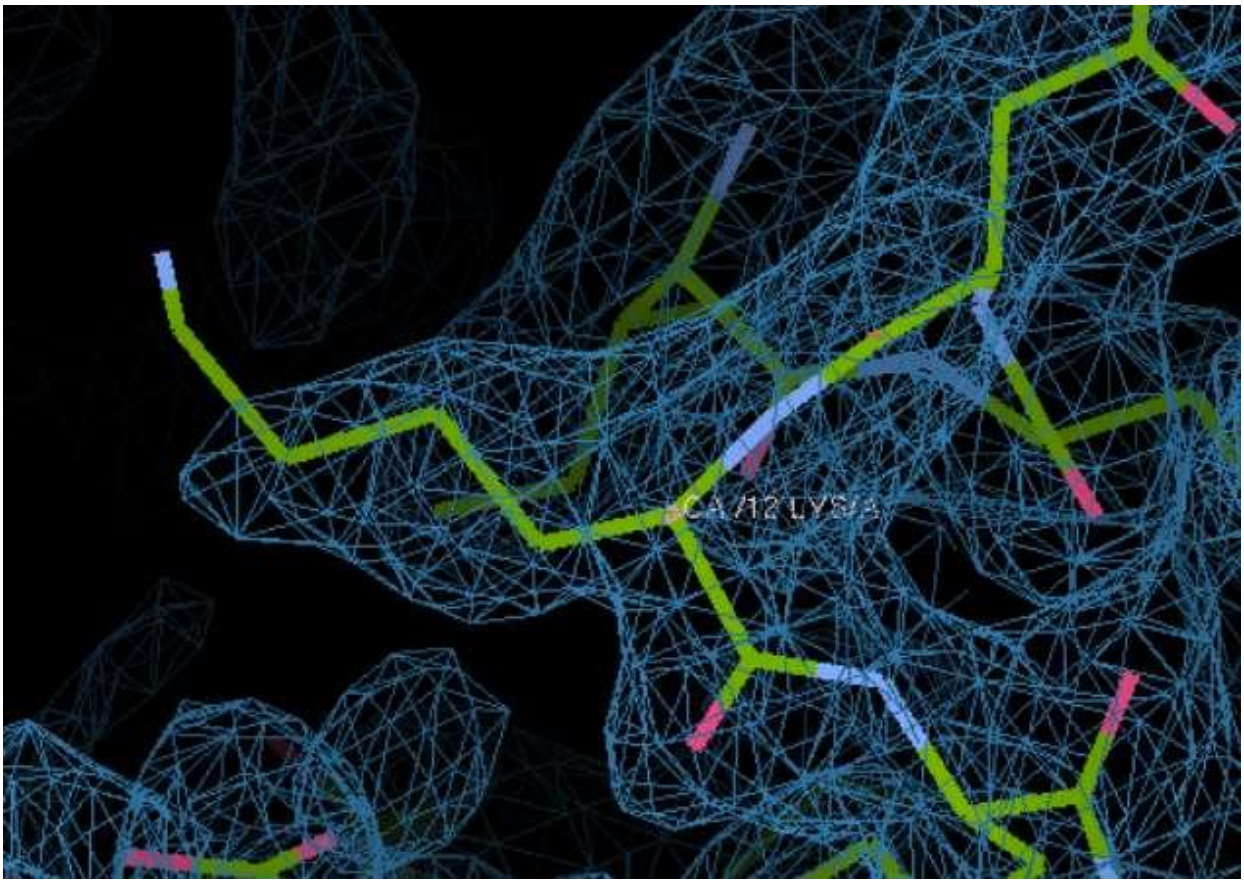
When you need to re-adjust an entire area of the peptide backbone, select **Regularize Zone** and select the region (by clicking two points) of the peptide backbone that needs adjustment.

When you need to move specific atoms in a residue select **Real Space Refine Zone**, and select a residue. It will change its color to white and you will be able to click and drag it into the density nearby.

If you come across a region where no residues are built in you will need to **Add Terminal Residue**. By default, it will place an alanine when you click on the C-terminus (or N-terminus). From here, you will need to select **Mutate & Auto Fit**.

Your objective is to get the best possible model, as measured by the $\mathbf{R_{free}}$ and $\mathbf{R_{work}}$ values.

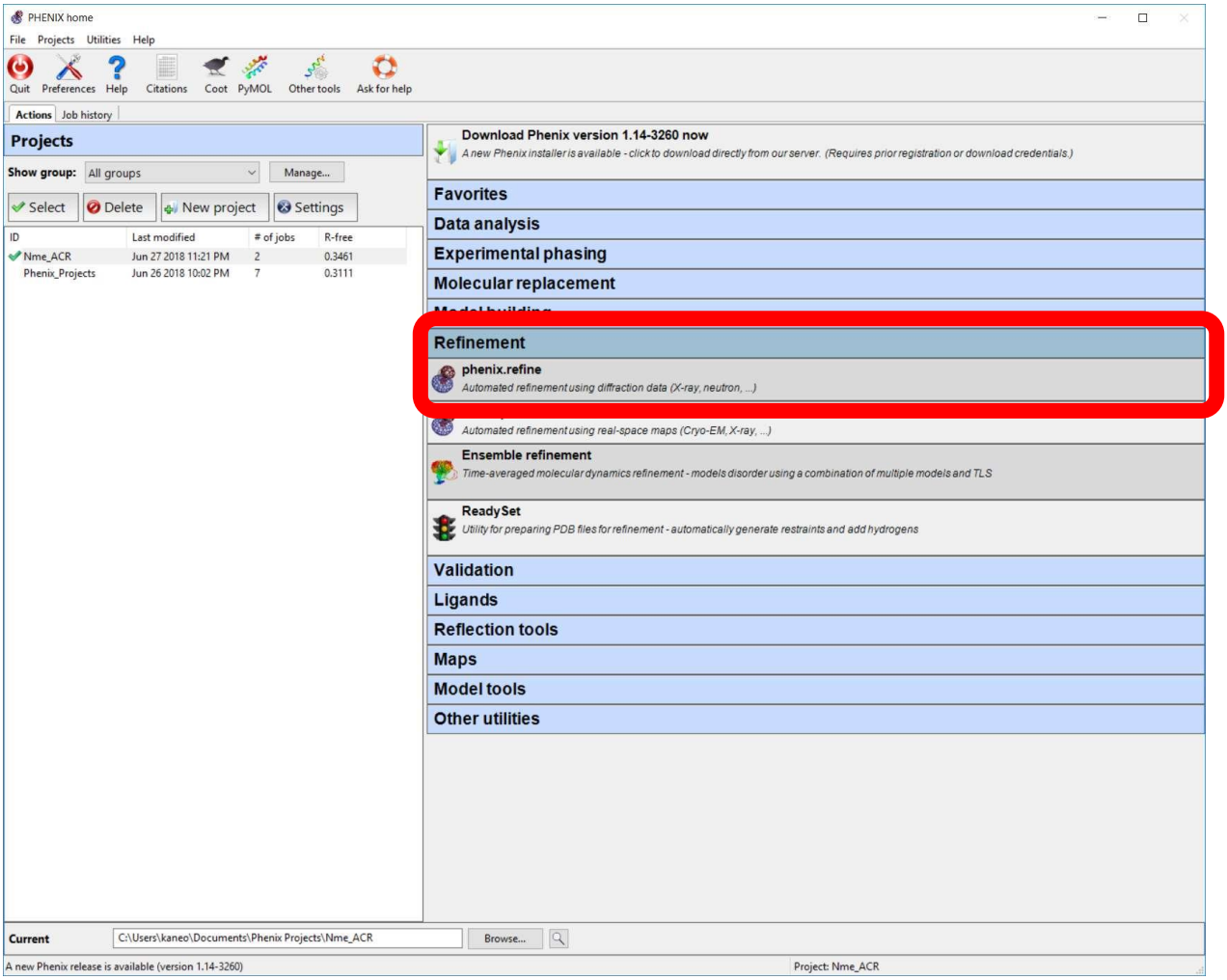
You may notice on lysine 12, the coordinates of the NZ atom and adjacent carbon lie outside of the density, as per the figure below. **Why do you think this is the case? Can you do anything to adjust for this?**



You may also come across stretches of alanines and glycines that are built into the density that are obviously not part of the actual protein. **Why is this the case?** You will need to mutate these to the correct residues.

Once you have finished building and refining your Chain A, save your coordinate file by pressing **File → Save Coordinates**. Save it in your project directory with a name that clearly distinguishes it from your other PDB files i.e. John_Real_Space_Refinement_1.pdb.

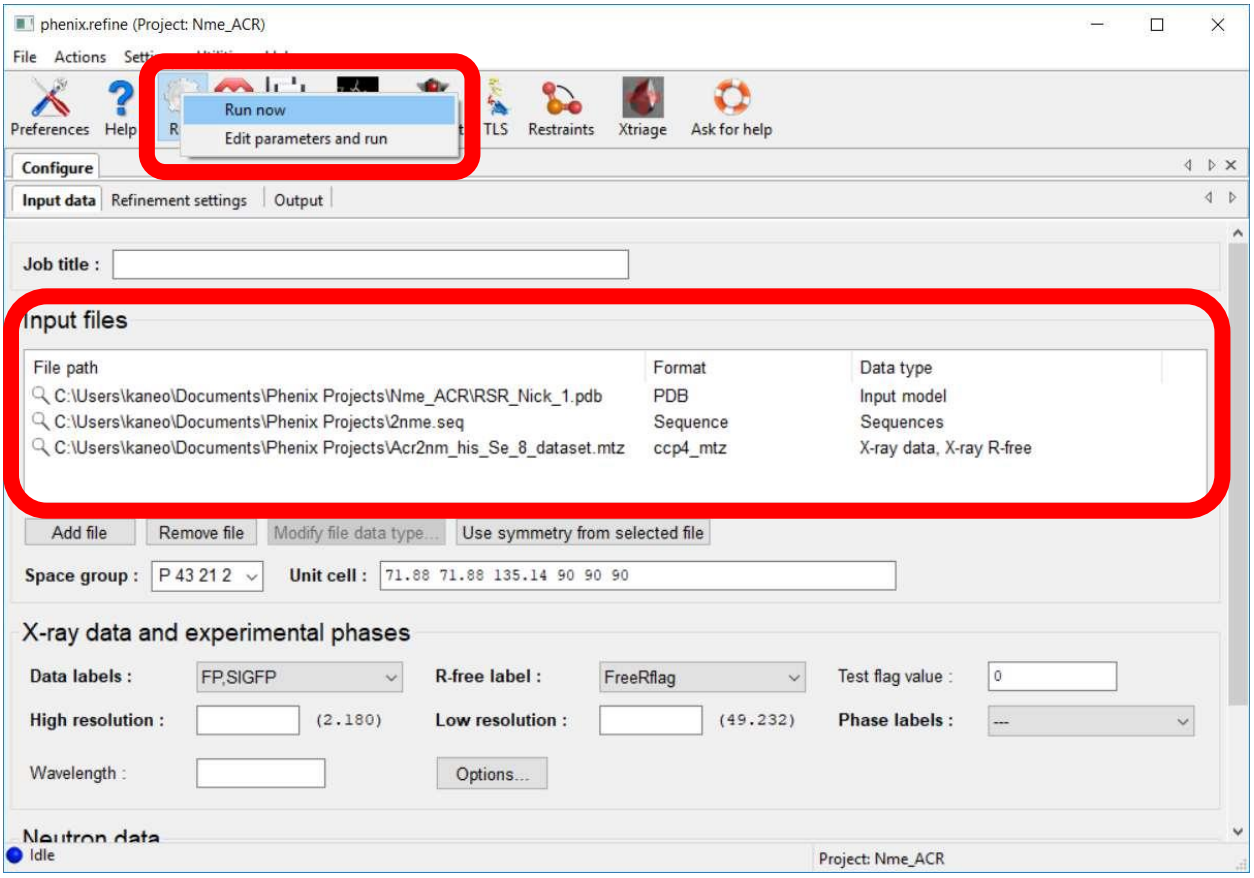
Once your PDB is saved, go back to Phenix and open the **Refinement** panel of tools. Select **phenix.refine**.



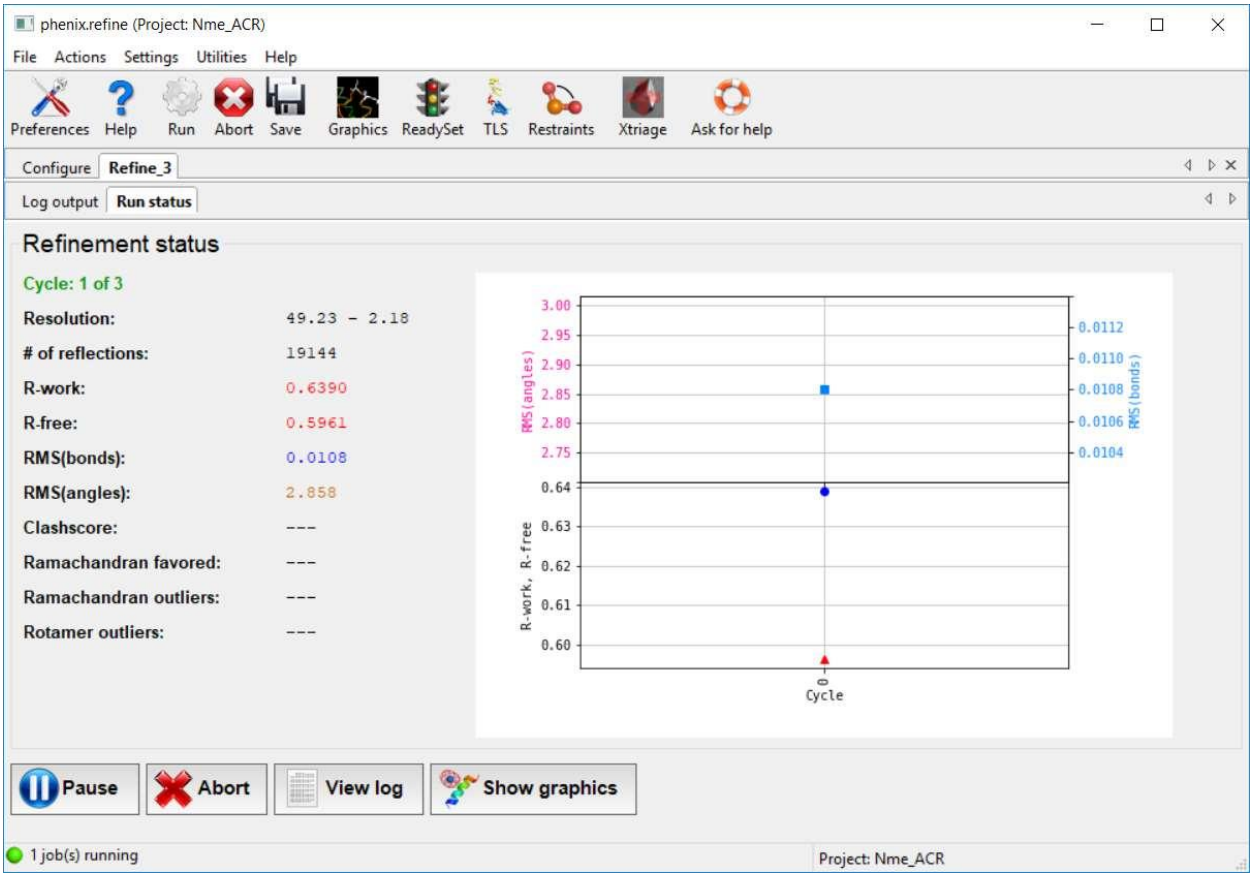
A window will open – in the input files panel, make sure you specify the path to the following three files:

1. Original MTZ file that was generated after running Autosol
2. Sequence file
3. Your saved PDB (John_Real_Space_Refinement_1.pdb)

After this, select “Run” and press “**Run Now**”, leaving all other fields with their default values.

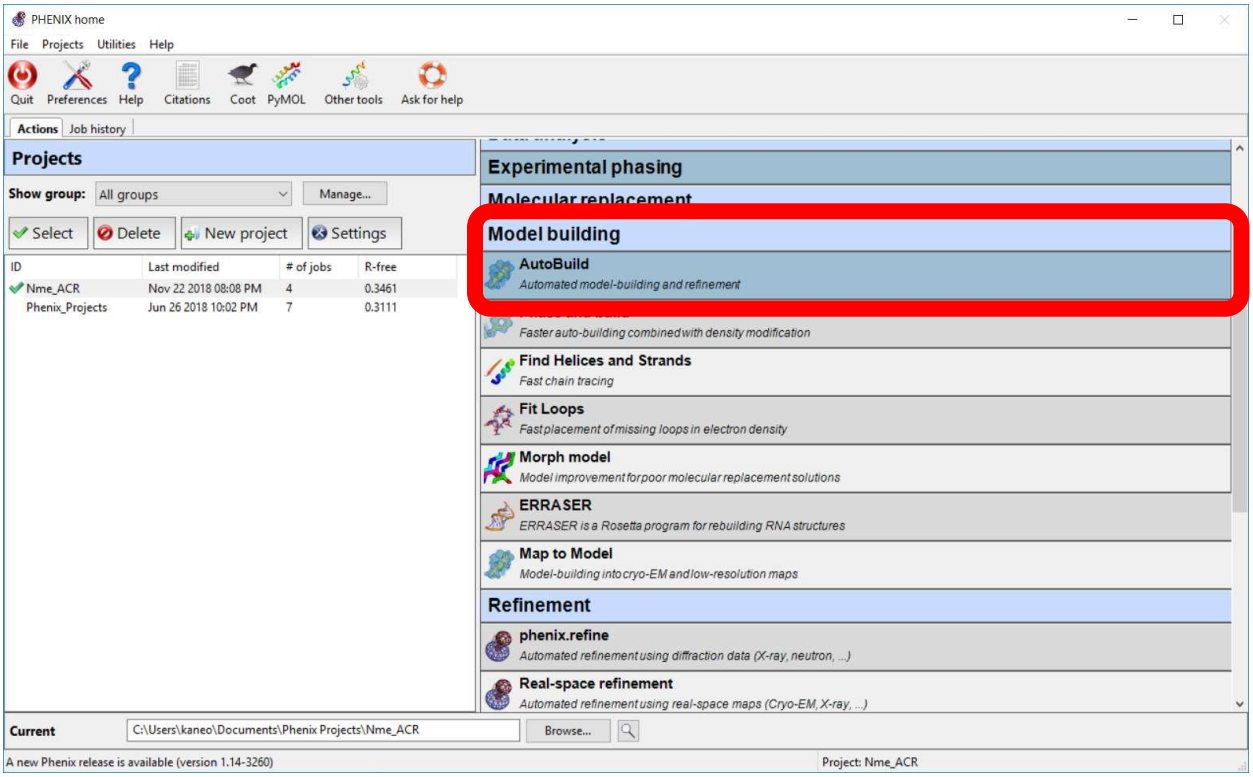


If your run began successfully, you should see a window like the following. If any errors occurred, consult with your TA for assistance. The job should take 3-5 minutes to run.

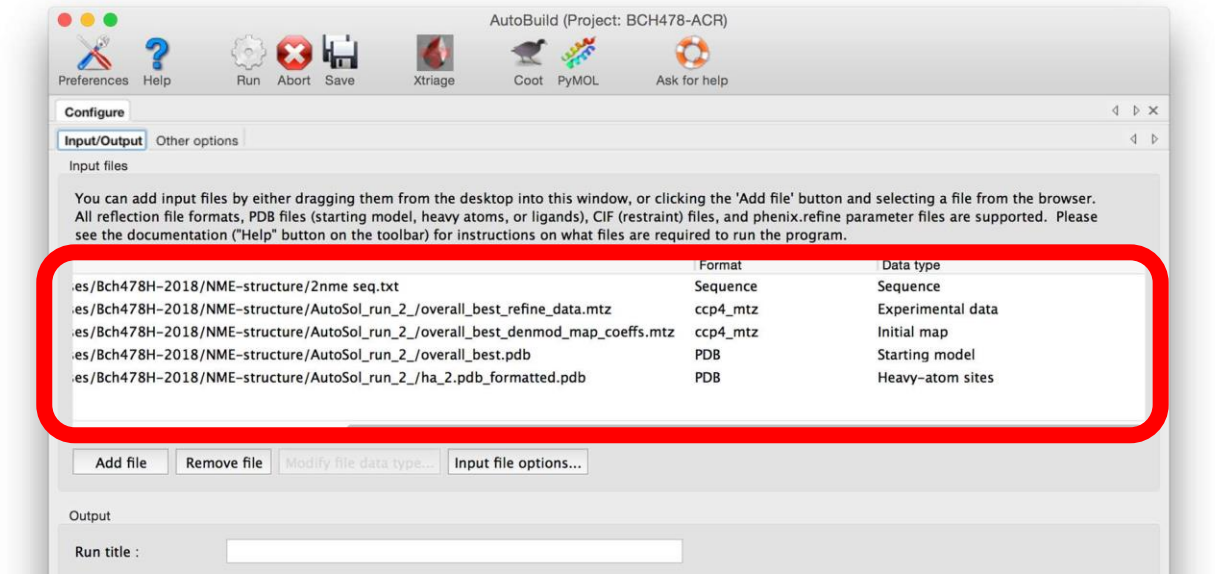


Record your R_{work} and R_{free} after this round of refinement in your lab notebook.

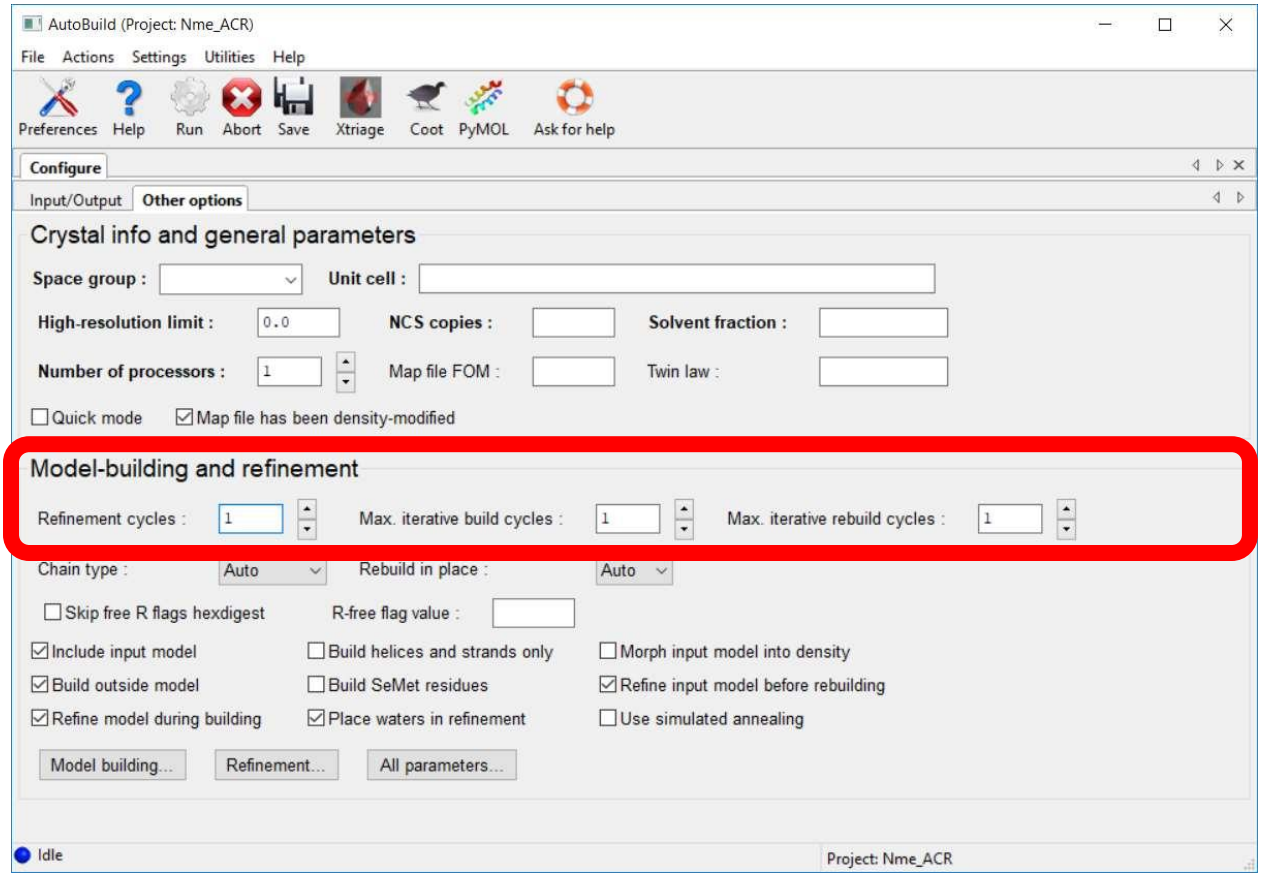
Go back to the Phenix “home” screen and select **Autobuild** under the **Model Building** tool panel.



Input the files as per the screenshot below:



Go to the '**Other Options**' tab, and under "**Model-building and refinement**", change the Refinement cycles, Max. iterative cycles, and Max. iterative rebuild cycles to **1**.

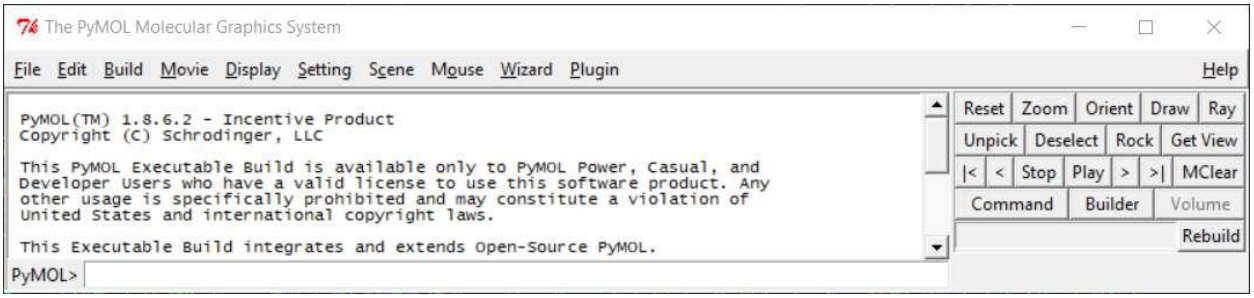


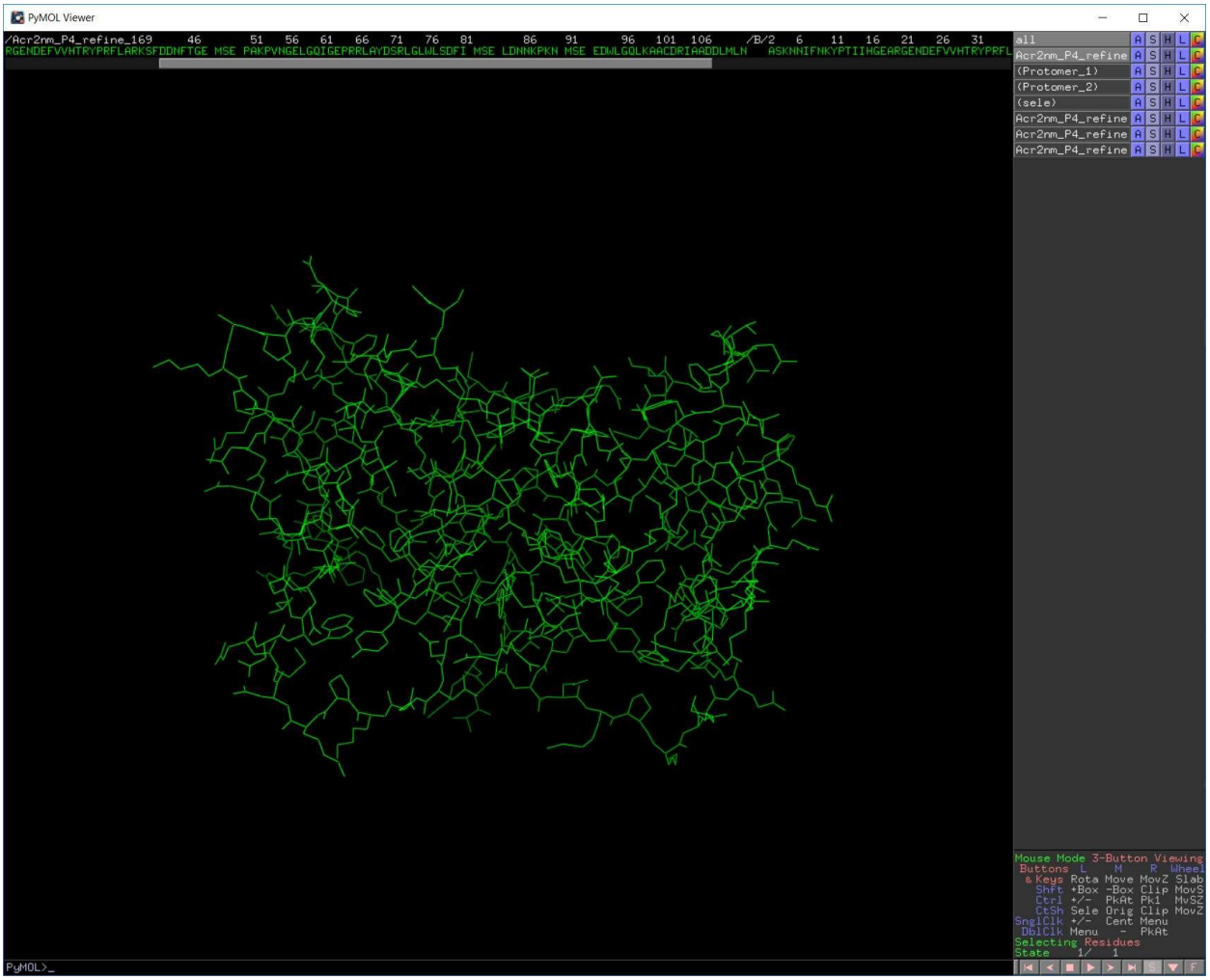
Press "Run", and select "**Run Now**". This will take about an hour to run. Lunch break!

Once the run is complete, open the manually refined PDB in Coot.

1. Select any two options under the validate tab to find and fix regions in the newly refined PDB.
2. Using the Ramachandran plot feature, find and list the outliers. Can you fix them?
3. **Why are these residues outliers?**
4. After fixing chain A and chain B, run another round of refinement.

Take your finally-refined PDB file, and open it in PyMOL. By default, PyMOL will open with two windows – a GUI +console for issuing commands and selecting commonly-utilized tools from a dropdown menu, and the protein viewer window itself:





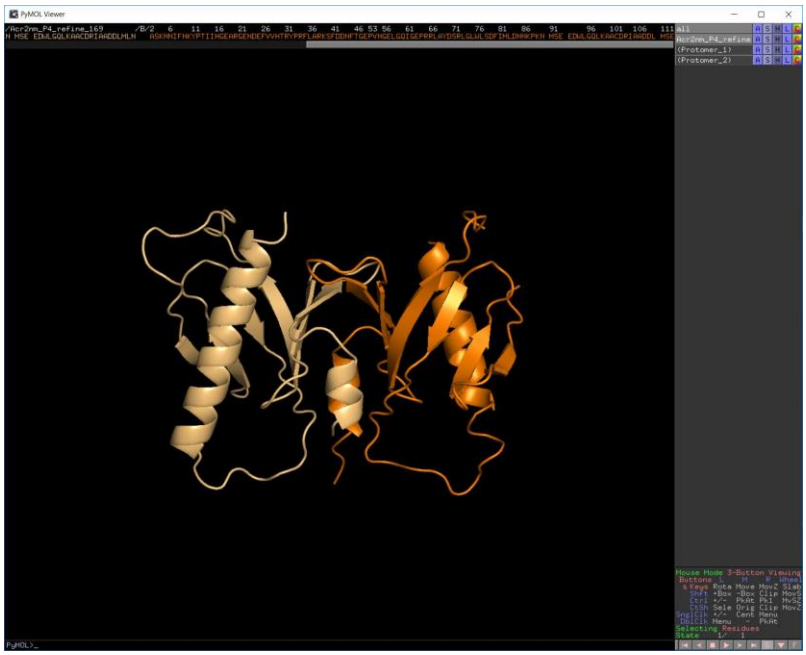
By default, PyMOL uses a line representation to show molecules. Change the representation to a cartoon by selecting the H button (which stands for “Hide”) at the top right of your screen and selecting ‘Lines’. You can also simply type hide lines in either of the consoles.

Now, at the bottom right of the screen, click S to show your protein sequence at the top of the viewport. You can click and drag on the amino acids to select regions of the protein. Drag along the sequence to select your first protomer in the dimer:

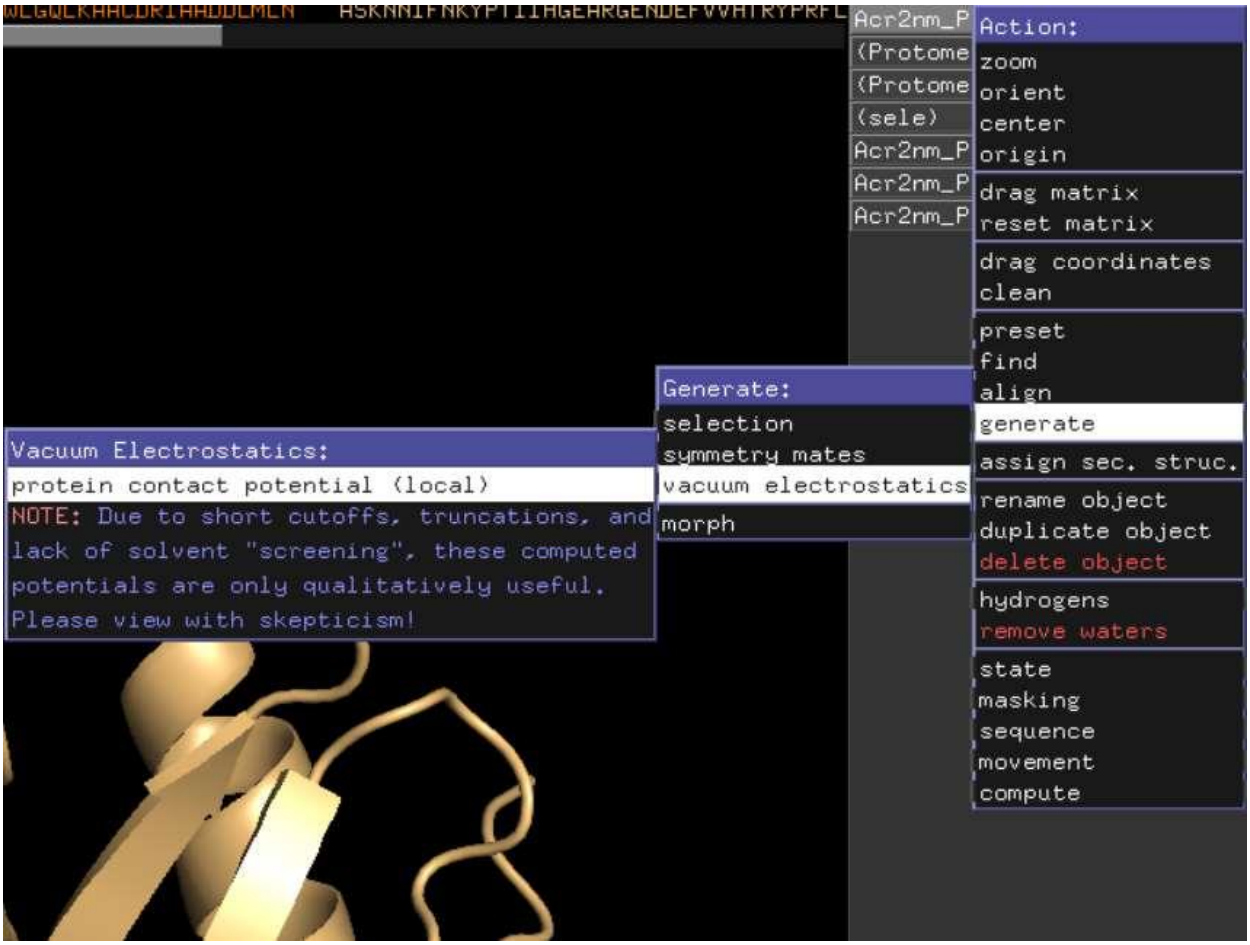
You can tell where one chain ends and the other begins by the **/A/** and **/B/** threshold markers. These are called chain labels, or chain identifiers. Once your sequence is highlighted, right click and go to **Actions → Rename Selection → backspace the ‘sele’ name, and call it “Protomer_1**”. This will now be its own ‘object’ in the right-hand panel that you can play with, independent of the rest of the atoms in the PDB file.

Repeat this process for the second protomer, and change the colors of each rotamer from the nasty default green to something fun and exciting, using the C button.

Now, select S for “Show” and pick cartoon. Your representation should have switched to a cartoon, as below. Do the same for a surface representation. **Save images of each of these representations for your lab notebook.**



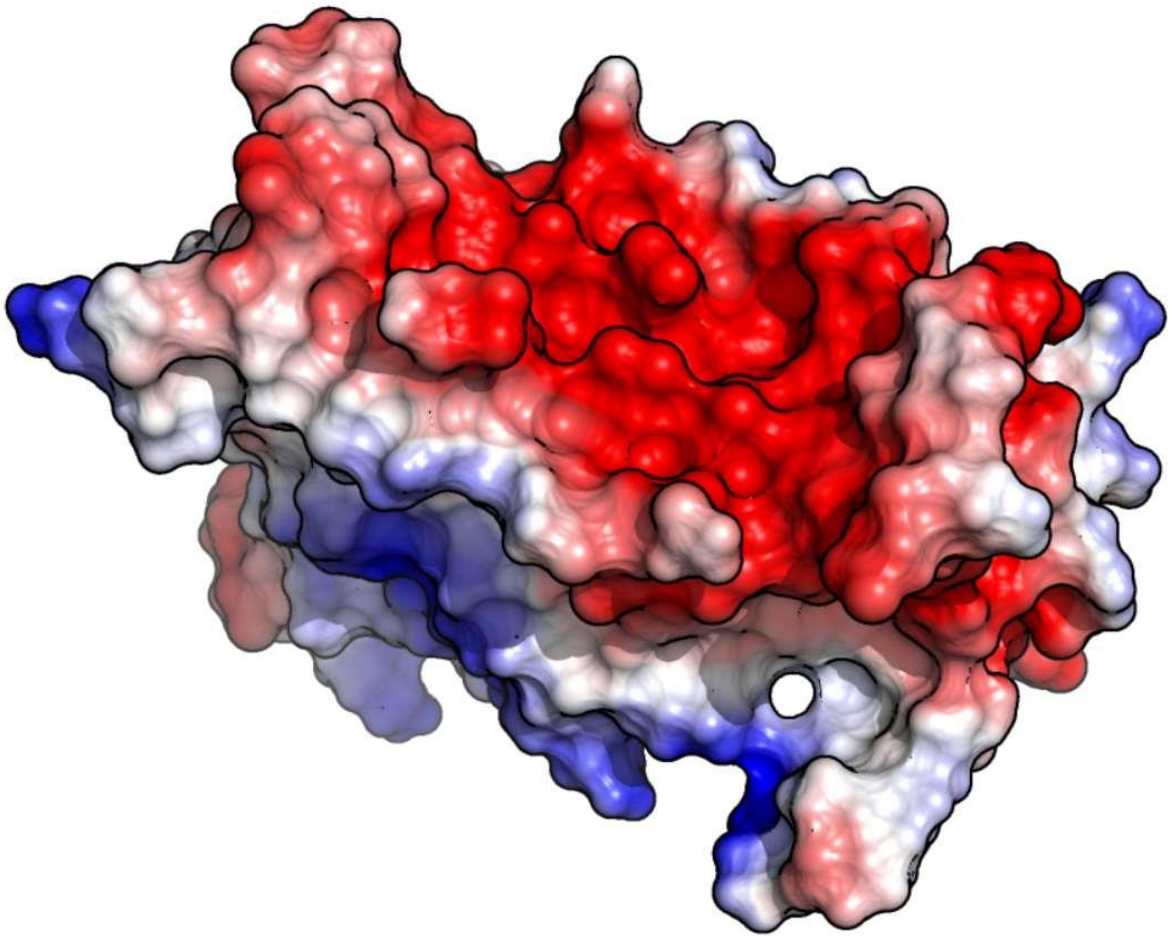
Once you have generate a few representations and feel comfortable in PyMOL, you will need to generate a vacuum electrostatic map. This can be done by going to **(Action) → Generate → Vacuum Electrostatics → Protein Contact Potential**. This can take several seconds to minutes. Your machine may temporarily freeze.



You should have an electrostatic map with electropositive regions in blue, and electronegative regions in red.

Create a vacuum electrostatic map for chain A and chain B.

If you want to save an image of your protein, type **bg_color white** into the console and go to **File → Save Image → PNG** which your protein is appropriately oriented in the viewport. Unfortunately, the educational version of PyMOL will not allow for raytracing. Fortunately for me, I have the licensed version, so I'll make a ray-traced image just to rub it in.



Can you speculate as to which surface might be crucial for the function of this anti-CRISPR? In your lab notebook, predict two mutations that would disrupt dimer formation.

Weekly Questions

1. What is model bias and how do you ensure that you aren't incorporating model bias in iterative rounds of refinement?
2. What is electrostatic surface potential? What information does it provide and why is it useful to calculate?
3. Name three crystallographic statistics that you could use as a read-out of data quality and explain what they are indicators of.

Chapter 6

It’s Finally Over!

Discussion Topics

1. Research the method of using isothermal titration calorimetry or microscale thermophoresis to measure binding constants. How do these technologies work and how is the information obtained from them inherently different when compared to 'real-time kinetics' experiments like BLI and SPR? Find an interesting paper that used ITC or MST to characterize an interaction and present the data.

2. Often times, crystallization experiments do not yield crystals, and we must resort to additional methods to 'rescue' the experiment by modifying some aspect of the crystallization setup. Discuss some methods one might use to improve crystallization chances of 'problem proteins'.

3. Advances in electron microscopy now allow for tomographic imaging of entire cells. Outline how the technique of electron tomography works and provide an interesting example of its application.

4. A bottleneck in cryo-electron microscopy studies is sample preparation. Research some of the problems scientists encounter when preparing for electron microscopy of proteins, and potential ways these problems can be overcome.

5. Integral membrane proteins make up a very small proportion of all solved crystal structures to date, yet they represent one of the most important classes of molecules for drug targeting, etc. Why are membrane proteins so difficult to crystallize, and what are some strategies crystallographers use to overcome these difficulties?

6. The invention of X-ray free electron lasers in serial femtosecond crystallography can now void the requirement for large, well-diffracting crystals. Outline the method behind XFELS/SFX and indicate when it is advantageous to use.

7. An emerging field in structural biology is that of structural mass spectrometry. This involves using different aspects of mass spectrometry (native MS, XL-MS, hydrogen deuterium exchange MS, labelling MS, etc.) to gather a variety of constraint-based information which can be used to solve structures of proteins. Select a few of these technologies/methods and explain how they are used in practice.

If you do not find any of these topics interesting, you may propose a different topic to Dr. Moraes, and upon his approval, present that topic. Note that the topic must be relevant to recent advances in structural biology.

Discussion/Presentations Grading Rubric

Marks will be given ***individually***.

		0	0.5	1.0
Student Name:	Slides were easy to read and interpret			
	Presentation followed a logical sequence			
	Slides consisted mostly of graphics. Minimal text.			
	Topic or paper was clearly explained			
	Pace was appropriate			
	Questions answered accurately and concisely			
	Speaker was confident, professional, made eye contact, etc.			
	Student made an effort to participate in discussions			

Student Name:	Slides were easy to read and interpret			
	Presentation followed a logical sequence			
	Slides consisted mostly of graphics. Minimal text.			
	Topic or paper was clearly explained			
	Pace was appropriate			
	Questions answered accurately and concisely			
	Speaker was confident, professional, made eye contact, etc.			
	Student made an effort to participate in discussions			

Final Lab Report

Maximum Page Limit: 8 pages, not including figures/tables.

Abstract (1/2 of a page)

In a few sentences, describe the project you carried out in the 5 weeks in the lab. Summarize the results.

Introduction (1 page)

Introduce the Cas9:ACR system, and discuss the rationale for the study.

Results (~2 pages)

Present and explain your results:

1. BSA Standard Curve
2. Nanodrop Readings
3. SDS-PAGE Gel
4. BLI Raw Data
5. BLI Steady State Binding Curves
    a. Cas9/ACR
    b. bLf/e4995
6. Lysozyme screen crystals
7. Lysozyme optimization crystals
8. ACR screen crystals
9. ACR optimization crystals
10. ACR structure

Discussion (~4 pages)

1. What useful information was obtained from running our proteins (ACR, bLf, BSA) on an SDS-PAGE gel?

2. Construct a standard curve from your A_{595} measurements of BSA with Bradford reagent.

3. How did you quantify your protein before loading an appropriate amount on the gel? Why is it not perfectly accurate to measure protein concentration via A_{280}? Why is it not perfectly accurate to compare to a standard curve of BSA? What method would you

suggest as a way of obtaining a concentration measurement that is as close to the true concentration as possible?

4. If you observed a protein on SDS-PAGE that was not migrating to a position that was appropriate based on its molecular weight, what are some reasons you might suspect the protein has an increased or decreased electrophoretic rate? Name three reasons.

5. If significant impurities were observed on the SDS-PAGE gel, how might this affect downstream kinetics experiments?

6. If you were tasked with characterizing the affinity of interaction between two novel proteins, and did not know if it was a high or low affinity interaction, how would you find the K_D and construct a binding curve?

7. What factors contribute to whether we treat one protein as a ligand or analyte? Could either protein serve as either the ligand or analyte?

8. Construct a steady-state binding curve for the ACR-Cas9 interaction and for the e4995:bLf interaction. Report B_{max}, K_D, and comment on the shape of your curve/fit model.

9. During the sensor loading step, which ligand (Cas9/e4995) loaded to a greater degree? Would you expect higher or lower binding signal for a ligand that loads to a greater degree? Is it okay to compare binding constants between two ligand-analyte interactions whereby one ligand loaded more than another? What are some limitations of having an over-loaded sensor and an under-loaded sensor?

10. Many interactions between proteins are not 1:1 in nature but rather have stoichiometries of higher degrees. Why does this complicate kinetic analysis, and how do we analyze these interactions?

11. Why is surface plasmon resonance a useful technique for measuring real-time interaction between proteins and small molecules? What property of the chip surface in SPR is one actually measuring during the course of the experiment?

12. How did we physically link (load) our ligand to the sensor in BLI? Research how proteins are commonly loaded onto SPR chips, and comment on limitations of this method.

13. What conditions in your sparse matrix screen facilitated crystallization of lysozyme? Comment on the role of each additive in the screen condition. Present an image of a clear drop, a drop with precipitate, and a drop with crystals, along with each of their drop conditions. Using a phase diagram to explain your answer, why do you think these drop conditions caused each outcome?

14. What aspects of the crystallizing condition for the ACR did you choose to vary for your optimization screen? Why did you feel that varying these properties would yield higher quality crystals?

15. Protein crystals are three-dimensional arrays of ‘tightly’ packed entities in a single conformation. What limitations does a crystal structure have in understanding the molecular function of a protein?

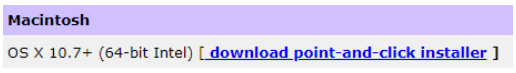
16. If part of a protein exhibited highly dynamic motion in solvent, how might you expect the electron density associated with this region to appear in the unit cell?

17. Why might a protein that has many surface-exposed hydrophobic residues be more likely to crystallize than one with many surface-exposed charged residues?

Appendix

Installing Coot, PHENIX, and PyMOL

*Software is already installed on the teaching lab computers. You are welcome to use either the teaching lab computers or your own computer. It may be more convenient to use the lab computers, especially if you are a Mac user, since all of my screenshots will be done on a Windows machine.

	Windows Users	Macintosh Users
PHENIX	Register first here: http://www.phenix-online.org/phenix_request/index.cgi Download here: https://www.phenix-online.org/download/phenix/release/  **Windows** (partially supported) 32-bit (Windows XP or newer) [**download point-and-click installer (recommended)**] 64-bit (Windows 7 or newer) [**download point-and-click installer (recommended)**] You will likely need the 64-bit version. Ask TA for help.	Register first here: http://www.phenix-online.org/phenix_request/index.cgi Download here: https://www.phenix-online.org/download/phenix/release/  **Macintosh** OS X 10.7+ (64-bit Intel) [**download point-and-click installer**]
Coot	Download WinCoot here: http://bernhardcl.github.io/coot/wincoot-download.html	Download here: http://scottlab.ucsc.edu/xtal/wiki/index.php/Installing_Coot_on_OS_X Ask TA for help.
PyMOL	Download here: https://pymol.org/2/	Download here: https://pymol.org/2/

Bibliography

1. Hill, T. *Essential Trigonometry*. (Questing Vole Press, 2017).
2. Pogoutse, A. K. *et al.* A method for measuring binding constants using unpurified in vivo biotinylated ligands. *Anal. Biochem.* (2016) doi:10.1016/j.ab.2016.02.001.
3. Lambert, F. L. Disorder - A Cracked Crutch for Supporting Entropy Discussions. *J. Chem. Educ.* (2009) doi:10.1021/ed079p187.
4. Cooper, D. R. *et al.* Protein crystallization by surface entropy reduction: Optimization of the SER strategy. *Acta Crystallogr. Sect. D Biol. Crystallogr.* (2007) doi:10.1107/S0907444907010931.
5. Thyssen, P. & Ceulemans, A. *Shattered Symmetry - Group Theory from the Eightfold Way to the Periodic Table*. (Oxford University Press, 2017).
6. Rupp, B. Biomolecular Crystallography: Principles, Practice, and Application to Structural Biology - Chapter 6. in *Garland Science* (2009).
7. Taylor, G. L. Introduction to phasing. *Acta Crystallogr. Sect. D Biol. Crystallogr.* (2010) doi:10.1107/S0907444910006694.