SHIVA MANDALA: And so just as an overview, today we're going to be talking about different techniques that are used to determine protein structure. We're talking a little bit about the protein data bank, or the PDB. And then the latter half of the recitation will be on-- we'll be doing a worksheet to look at the structure of ubiquitin and di-ubiqiotin using PyMOL. So that's just to get you familiar with using PyMOL.

And so for the first question I'd like to pose to you is why should we determine protein structure, what we can learn from determining protein structure? And so I look to you for answers, a lot of answers. But does anybody have any ideas? Yeah?

AUDIENCE: [INAUDIBLE]

SHIVA MANDALA: Yeah, absolutely. Knowing putting structures does help you determine enzyme mechanisms. Anything else?

AUDIENCE: Structure can indicate function.

SHIVA MANDALA: Structure what?

AUDIENCE: Can indicate function.

SHIVA MANDALA: Yes, absolutely structure does indicate function. Can you be a bit more specific with respect to that?

AUDIENCE: [INAUDIBLE]

SHIVA MANDALA: Yeah, absolutely. Yeah you can determine active site of enzymes. Any other ideas? Still a lot more to go. What about can you learn interactions with other macro molecules? OK.

Well, let's just go through some of them. So yes, structure does, indeed, determine function. And the idea is if you know the structure of a protein, you can learn a lot about its function in vivo. And so some of the things that you can study is you can study enzyme mechanisms. It's

relatively hard to do in x-ray crystallography, because every time you solve a crystal structure, it's only one snapshot of the enzyme. But you can definitely do it. You can design drugs or substrate that bind to the protein if you know what the active site looks like. You can design a high affinity inhibitor, and this is used in the pharmaceutical industry a lot. You can study translation and transcription, which is what we'll be learning about in class this week, and we'll be learning next week, as well.

You can also make co-crystals of proteins with other proteins and nucleic acids and study interactions between macro molecules. And this is something that has been emerging. It makes it harder to solve a crystal structure, but x-ray crystallography is a very powerful tool for this. You can also study immune system functions. So this is more on the biological side of things. You can study host-pathogen receptors and their interactions. And many, many more. Really there's no reason why you shouldn't have a structure for whatever protein you're studying.

And so the key idea is that if you know the protein structure, that allows you to carry out biochemical studies. But then also if you determine the structure, that can help rationalize results that you get from biochemical studies. So it goes both ways. And the final point that is not perhaps not emphasized that much is that structure is a result of sequence. And ideally, what we would like to know is if we know the primary amino acid sequence of a polypeptide, we'd like to be able to predict its complete three dimensional fold. And that's sort of the idea behind protein folding. But we're not quite at that stage yet computationally. And so that's why we need experimental data. But that's something we're moving forward, moving towards as a science.

And so just to go over, I'm going to cover the three main techniques in protein structure determination. And so the most common one is X-ray diffraction. And the I'll go over some of the details of X-ray later. I'll be focusing on X-ray diffraction in this talk. Some of the selling points of X-ray diffraction are that you can study a protein of any size-- small proteins, large proteins. You can study complexes of proteins. There is a need to crystallize your sample, which makes it challenging. It's very hard to crystallize proteins. In your body, no protein is crystallized. So you're trying to make proteins do something that they don't really like doing.

You can obtain a high resolution structure. So a resolution of two Angstroms or even less is pretty common for good protein X-ray structures. But then also it's difficult to observe dynamics. So an X-ray structure is just a snapshot of the protein at a specific time. And so you

need often a series of X-ray structures to really learn something about the mechanism or the dynamics of these proteins.

The second most popular technique is nuclear magnetic resonance, or NMR. NMR is typically used to study small proteins. And the reason for this is that if you look at-- so this is a 2D NMR spectra. If you look at the 2D NMR spectra, there's a lot of peaks, right? And the more residues you have, the more amino acids you have, the more peaks you have and it gets very crowded. And so that's the limiting factor with using NMR to study large proteins is that you can't resolve all the chemical shifts.

The plus point is that there is no need to crystallize your protein. You can study it in solution state or solid state. Solid state NMR is typically used for memory in proteins. You can use solution NMR to study any soluble proteins. You do need isotopically labeled samples. So these are 13c or 15n enriched samples, which is very hard and expensive to do. So that's one of the drawbacks with NMR is that it's a relatively expensive technique while X-ray is more accessible.

You can obtain a high resolution picture with NMR, as well, but it often requires more work than X-ray crystallography in that you need to do about five NMR experiments. That can sometimes take months to determine high resolution structure. X-ray is more accessible. But really, the big upshot of NMR is that you can observe dynamics within proteins. So you can really see-- proteins are living, breathing machines. And you can see that with NMR better than you can with any other technique.

Another quick point which I don't have written up here is that NMR is sensitive to protons. And you can study hydrogens with NMR. You cannot study hydrogens with X-ray diffraction, for reasons that I'll come back to later in the talk.

AUDIENCE: Can you explain how exactly NMR observes dynamics?

SHIVA MANDALA: Yeah, so there is a series of different-- I mean, I don't know how much detail you guys know about NMR. But basically, you can study the relaxation of nuclei. That's often one that's used. So you study T1 and T2 relaxation of nuclei. And more mobile residues and more mobile atoms relax faster. But really, the idea is you can use-- there are a whole bunch of different experiments in NMR. And you can access timescales from the nanosecond up to till millisecond. So 10 to the negative 9 to about 10 to the negative 3 seconds of motion. So it's quite-- and they use different experiments for different parts of that timescale. And we can talk

more about that in detail.

The third technique is electron microscopy. So this is restricted so far to large proteins. The reason for this is that resolution is not so good. Again, you don't need to crystallize your proteins. So that's an option. You don't need to use labeled samples, either. So the sample preparation is probably the easiest for electron microscopy. The picture that you get is sometimes lower resolution, but the technology is moving forward to the point where we can get a resolution as good as X-ray structures. And I know that there's a 2.2 Angstrom resolution structure out there definitely, and there are others that are 3.2 Angstroms. But maybe there's something better than that out there in the literature.

Again, it's difficult to observe dynamics. So similar to X-ray, it's just a snapshot of your enzyme.

AUDIENCE: Is that picture-- is the concept similar to a normal microscope?

SHIVA MANDALA: Yeah, absolutely it is very similar. And the only thing is you're looking at how electrons interact with with your sample as compared to light, right? Visible light, I guess. So each of these particles here is your protein, is a protein molecule. And then these are three dimensional reconstructions. So there's computer software that does this. So to go from this, you basically signal average over all of these different molecules. And then you signal average over all of your different orientations of the protein that are trapped in your static electron microscope image. And then using some image processing, you generate a three dimensional image of your protein that has a higher resolution than what you can see just with one single photo, I guess. So there's lot of computer processing that happens behind the scenes.

AUDIENCE: So is the electron-- the interactions with electrons, is that similar to fluorescence microscopy? Because that's where you're seeing where your proteins are located, right?

SHIVA MANDALA: So the difference with fluorescence-- I mean, here electrons can interact with any atoms, right? Any material.

AUDIENCE: Oh, so you can distinguish what different atoms the electrons are interacting with?

SHIVA MANDALA: Yes, because different atoms interact with-- different nuclei interact, and electron densities interact with electrons differently. But with fluorescence microscopy, you're usually looking at just a single molecule a fluorophore that's reporting on where your protein is. But electron

microscopy is a much higher resolution picture. It's actually an atomic level-- well, maybe a few atoms level-- picture. Fluorescence microscopy is usually just used to study where your protein is if proteins are interacting. So that's more macromolecular interactions. But you can get single molecule resolution with fluorescence microscopy if you use the correct techniques.

And so just as an introduction to the protein data bank, so the first graph tells you the number of structures in the PDB as a function of a year going from 1975 all the way to 2015. So you'll see today there are about 110,000 structures, of which 100,000 were determined using X-ray crystallography, and about 10,000 using NMR, and about 1,000 using electron microscopy. So really quite a nice ratio there.

And if you see the yearly increase in the number of PDB structures, you'll see that X-ray is, of course, really big. NMR has been fairly consistent over time. I think that has to do with the fact that it's expensive and it takes time to prepare your samples. But you also see a huge spike in electron microscopy of late. And so with the advent of cryo EM, a lot more people start using cry EM to determine protein structure.

AUDIENCE: Doesn't [INAUDIBLE] produce [INAUDIBLE] or you just put it in [INAUDIBLE]

SHIVA MANDALA: Yeah, it can be. But the problem with that is when you put your sample in the [INAUDIBLE], you can get chemical shift information. But chemical shift doesn't tell you anything about protein structure. I mean, it tells you a little bit. It tells you about what the electron density is at the atoms. But what you really need to get from NMR experiments are distance of strains. And so this is through space experiments. So you can say that, oh, this one carbon nuclei is at a distance of 6 Angstroms away from this other carbon nuclei. And you typically want to accumulate about five strains per atom. And so to collect five times how many ever atoms you have in your sample, it can take time. It's hard to do. So chemical shift by itself doesn't tell you much about protein structure. Any other questions so far? All right.

So now we will focus the rest of the talk on-- well, another part of the talk on X-ray crystallography. And so crystallography is the science of determining the t dimensional position of atoms in a crystal. And so what crystal is, a crystal is a solid material whose constituents are arranged in an ordered pattern expanding and extending in all three spatial dimensions. And so the key idea is that this translational symmetry-- so if you go in any of the three directions for a certain amount of time, a certain amount of length, you'll come back to the same pattern that you start off with.

And so this is a crystal of your protein of interest. What you want to know is you want to know how the proteins are packed or arranged within this crystal structure. And also as a result, how the atoms are arranged within the crystal structure. And the way this works is by diffracting X-rays through your sample of interest. And with this slide, I just wanted to point out that it's not restricted to proteins. You can study salts, you can study your favorite small organic molecule. Whatever you want, really.

And so the general workflow is that you have a source of x-rays that can be a singleton-- or local source, singletons are much brighter than local sources-- that you shine in your crystal. And you obtain what is known as a diffraction pattern. And so this tells you how the X-rays are in track-- this tells you something about how the X-rays are interacting with the atoms in the crystal.

And so this used to be collected on a photographic plate. This particular image is on a photographic plate, but now people use CCD sensors. It's a lot easier. And knowing the-- sorry, one more thing. Each of these dots, light and dark, on the diffraction pattern is called the reflection. And that contains some information about the electron density and the crystal structure.

And from your diffraction pattern, you can then back calculate the electron density in your crystal structure that gave rise to this diffraction pattern. And the way you do that is by looking at the intensity of these reflections. And you also need-- there's also something called phase, so you need to determine phase. And sometimes you'll see in the literature you'll see heavy atoms being introduced, or mercury being introduced. And that's often to determine the phase, which is essential for calculating the electron density.

Once you determine the electron density, you know what protein you started off with. And so you know what your protein looks-- you know the sequence of your protein. And so then you just take your electron density and fit whatever polypeptide change you have to that. And then usually this is all automated nowadays. So you press a few buttons and it goes through, software does everything for you. But it was much more challenging early on. And even now, the computers will get you up to a certain point. And then in the last, last stages of refinement, you always want to-- usually people do that by hand. Any questions about X-ray crystallography?

AUDIENCE: [INAUDIBLE] how strongly or complex, but how do you get that electron density from

diffraction pattern. Like in organic chemistry, in basic [INAUDIBLE] I thought that you can-- from diffraction pattern, you can learn the distance between atoms in the lattice points. But here with proteins, every point is itself a protein, right? In the lattice? No?

SHIVA MANDALA: No, no, no, no. Because you're still looking at every point in the lattice is still an atom if you're doing proteins. It's just that there are a lot more atoms, and the lattice is a lot bigger, which is what makes protein crystallography hard compared to small molecule crystallography. So it's harder to solve a protein crystal structure than it is a small molecule crystal structure just because there's so many more atoms in your lattice. But the idea is exactly the same as a small molecule. It's just a lot harder to do.

And for more information, I actually have a resource at the end that goes in-depth into the math of the process. But just briefly, this diffraction pattern is collected into what's called reciprocal space. And to go from this to electron density, you need to do a Fourier transform into Hilbert space, which is what electron density is spaced in. But I will provide a reference for more information on that. And so the next part of the discussion part of this recitation will be thinking about some of the limitations of X-ray crystallography. So there's a lot of them, but I'll turn to all of you for your inputs.

AUDIENCE: Is it difficult to develop crystals of certain types of proteins?

SHIVA MANDALA: Yes, absolutely yes. First point, it's really hard to purify and crystallize proteins. It's really not a trivial task. It can take months or even years to do so. And nowadays you have robots that can set up reactions under hundreds of different crystallization conditions. It's sort of a black magic sort of hard. It's hard to predict what crystallization conditions are going to give you a high quality crystal. Anything else? Yes?

AUDIENCE: Like the crystals you get may or may not be physiologically relevant?

SHIVA MANDALA: Yes, absolutely. Just on point with questions. But yeah, it's hard to tell whether the crystal structure that you get is depicting what's happening with the protein in vivo. And I mean, this is a problem that's inherent to X-ray crystallography, right? Every time you solve a crystal structure, you don't know whether it's relevant. But usually I think it turns out that it's pretty close, if not completely accurate in solution. But sometimes you do have to be careful about this. It's especially challenging for stuff like membrane proteins, where you don't really know. Any other ideas?

So when proteins are translated, do they usually-- are they usually used just like that, or does something else happened to the proteins in most cells?

AUDIENCE: [INAUDIBLE]

SHIVA MANDALA: Yes, absolutely. Post translation modifications. So a lot of proteins are post-translationally modified. And so when you're growing a crystal of your protein, you usually just use a purified version of your protein so you can't really calculate. And sometimes these PTMs are essential for the function of the protein. So you're missing some part of the picture. Anything else? What about movement? Can you tell what proteins are flexible, what parts of the-- sorry.

AUDIENCE: Well it's like if part of protein is mobile, then you won't have the density for it.

SHIVA MANDALA: Yes, that is true. You can discern anything about dynamics and flexibility. And the answer is you can tell something. You can tell something about the relative motion of different parts of the protein with respect to each other, but it's hard to tell something about the absolute motion of these proteins. So you can't see, say, larger scale motions, right? Most proteins are living, breathing machines, and it's hard to capture that in an X-ray structure.

And one more thing. So is there any element that you cannot detect in X-ray crystallography very well? This has to do with the way-- so in X-ray crystallography, you're setting interactions of X-rays with electrons, right? So does anybody know? Yeah, I heard somewhere.

AUDIENCE: Protons.

SHIVA MANDALA: Protons, yeah. So protons have one electron. And so their X-ray signal, so-called, is really weak. Really, really weak. And so you can't really see protons with X-ray crystallography. And so you can't really study hydrogens or hydrogen bonds. And if you look at the structure of proteins that has hydrogens in it, those hydrogens were put there as a result of an average bond length, the typical bond calculation. So it's not actually experimentally determined.

You can use neutron diffraction to get around this, but neutron diffraction is hard to do because you need to grow very large crystals to study. And I think there are about 80, I think, neutron-- around 100 neutron structures in the PDB so far. But for small molecules, neutron is much more accessible. Neutron diffraction? And so the idea is the same as X-ray crystallography, except for you're using neutrons.

And the final point is that one structure only tells you part of the story. Again, this is

emphasizing the fact that one structure is just a snapshot of the protein at a certain time. And you want to correctly interpret your data and learn something more about the protein, you often have to use complementary biochemical techniques. Are there any questions at this point?

So the last part of the talk is on how to assess the quality of structures in the PDB. They're large structures, and you want to be able to know whether the model that's presented to you is actually accurate, actually reflects the data that was collected. And so the first point is what is the resolution of the structure? And so the take home message is that a lower number means a greater resolution. And the resolution actually here is referring to the distances between the atoms in the plane. And so that's where that's coming from. So that's why if you have a lower resolution, that means you can resolve atoms. A one Angstrom resolution means that you can resolve atoms that are one Angstrom apart on parallel planes.

But the take home message at a one Angstrom resolution, you can see individual atoms and you can discern the identities of those atoms by looking at their electron density. But if you come bound to a four Angstrom structure, you'll see the benzene ring doesn't really have a clearly defined electron density, and there's no hole in the center. But if you look at the one Angstrom structure, you can see that there's even a hole in the benzene ring to confirm the electron density there.

And then if you look at-- so these are the data statistics. So this is just pulled from PDB off 2JF5. So this is the PDB ID for di-ubiquitin, and we'll be looking at the structure later in the worksheet. The resolution tells you, of course, about the resolution of the crystal structure. And so in this case, it's at 1.95 Angstroms or two Angstroms, and so that's pretty high resolution.

Reflections are each of those points in the diffraction pattern that you collect, and a unique reflection refers to the fact that you've only collected it once. So usually when you collect diffraction patterns, you put your protein a certain orientation with respect to the X-ray beam, and then you collect the diffraction pattern, and then you rotate your crystal a whole bunch of times, and you collect a whole bunch of different diffraction patterns. And then you superimpose all of those together to get the master diffraction pattern.

Redundancy refers to the fact of how often each reflection was observed. And so this is a signal averaging. The more redundancy you have, the greater number of times you observed

that particular reflection. Completedness refers to how many of the data points were actually measured. And so this is when you created your model, you can back calculate your diffraction pattern. And then you see how many of those reflections were experimentally observed in our data set.

And so for this usually you want as close to 100%, and anything above 95% is considered fairly good.

R merge is an indicator of how consistent measurements are. So this is a measure of what the difference between different measurements for the same reflection are. So if you look at the intensity for the same reflection a different number of times, you're seeing the standard deviation of that. So you want as low a number as possible. So usually you want it to be about 1/10 of the resolution of your crystal structure, which is 0.2 Angstroms in this case.

I forgot to mention this, but the values in parentheses are for a high resolution bin. So this is just a certain subset of this data set that's considered to be higher quality than the original data set. And that's actually coming from this value, which is signal intensity over sigma of signal intensity, and that's a measure of the signal to noise ratio. So you want a higher signal to noise ratio is better. And this is fairly good. And the higher the better. And cutoff is at 2 for the high resolution bin of your data points.

All, right so this is just the raw data. And then the refinement statistics tell you something about your refinement process that gave you the crystal structure that you calculated. So our crystallization, which is R cryst, which is also called R work, and also called the R factor. That tells you how well your model and your data match. And so this is where you calculate the difference in the diffraction patterns between what you experimentally observed and what you calculated using the model of electron density.

R free tells you how well your model and data match when corrected for overfitting. So the idea behind this is that when you collect your data set, you put aside about 5% of the reflections that you observe in the data set to prevent overfitting. And then when you've created your model, you go back and see how well those 5% of data points fit the model that you've come up with. And if they fit well, that means you've accurately predicted your crystal structure using the data that you have. If you don't fit well, that means you've just fit whatever data points you have to some model. You've used a bunch of data points and you fit it to some model, but that's not actually accurate to the protein structure that you determined.

And so for this you want it to be about 1/10 of a resolution. That's also true for R cryst. But then the other key point is that it should be very close to the R cryst, because it's telling you that it's random error, or it's the quality of your data set that's causing this and not sum overrefinement that you've done during the refinement process.

B factor tells you about how mobile atoms are in the crystal lattice. And this is something that's it's not particularly useful if you look at the bulk statistic. But if you need to, it's important to evaluate this by residue. And so if you look at amino acids in the loops of proteins, you'll find that they have a higher B factor usually, meaning that they're more mobile. And so you can often tell what residues are important for function by looking at the B factor. B factor is kind of-- there's inherent vibration motions in all atoms, right? So there will be a B factor at any temperature greater than 0. But it also does tell you a little bit about the disorder in protein structures. And the B factor of water is often included just so that people who are looking at the crystal structure afterwards can decide whether that water is really there, or really was part of the electron density, or whether it's just an artifact of ore refinement, or something like that.

And then the final statistic that you can look at is the RMSD from ideal geometry. So the geometries of these bonds and angles are usually well known. And so if you compare the results from your structure to the known stereochemistry, you'll find that this is actually just at the threshold of the cutoff for what is considered good. So 0.015 Angstroms, this is standard deviation of 1 lens of your model versus what is already known. And so 0.015 Angstroms is just about the cutoff that's considered good for X-ray structures. And same for bond angles. 1.5 degrees is considered the threshold of what is considered acceptable.

And you can also look at Ramachandran's statistics-- so this is looking at five sided back one angles-- to tell you if there are any static clashes, say, for side chains that really shouldn't be there. And you can if you do have static clash nowadays, you have to report it to PDB. The only exception is for glycine, which doesn't have a side chain. And so it can adopt a strange phi psi angle that's outside the Ramachandran plot.

Any questions on this? Anything else about X-ray crystallography? So this brings us to the end of the talk, and this is a resource that I found that has more information about X-ray crystallography, and the math behind X-ray crystallography, and more of the theory behind it. But again, I want to emphasize that today it's a very automated process is that you click a few buttons and anybody can do it.