PROFESSOR: So today, our goal is to really go through this the paper that you read maybe last night by Dekel and Alon "Optimality and Evolutionary Tuning of the Expression Level of a Protein." It was published in *Nature* in 2005. I think that it's a very interesting paper, exploring some kind of big general ideas.

I think it's also, in some ways, rather misleading. And we'll try to understand or discuss the ways in which the connections between experiment, theory, prediction, and so forth, how they all play out in the context of this problem.

Before we get going too much on the science, I just want to remind everyone that Andrew will not be having office hours today. He is off interviewing for MD, PhD programs right now. But if you had questions about the problems, I hope that you asked [? Sarab ?] last night.

You might be able to grab him after the lecture today, but yes. Any other questions about anything before we get going? No.

So I think that this paper in general, I guess, the lecture today is really a combination of trying to start thinking about maybe laboratory evolution or kind of population level phenomena in general, as well as this question of optimization in terms of protein expression.

So can somebody just maybe summarize the big idea of this paper? Yes, please.

AUDIENCE: Protein expression levels evolve to optimal values for cost-benefit questions.

PROFESSOR: Right, so that's the argument at least. And they have a very nice first sentence here. "Different proteins have different expression levels." You know, it's hard to argue with that statement, nice, concise. But the question is, well, why?

And I'd say that there is a range, different philosophical opinions out in the world. I said that some group that is very much reflected in this study is trying to think about this in the context of optimization.

Well, maybe the reason that we see a given level of expression of some protein is because, at least over evolutionary time, in some ancestral environment that we don't know, but maybe it evolved to optimize some cost-benefit problem.

And then I'd say that there's another kind of general philosophical approach that tends to be a little bit more agnostic or just maybe more of a sense that certainly things could have evolved to optimize something. But we can never really know where they evolved in, so we shouldn't be going out on a limb on these things.

And given that this is philosophy, I will maybe not require that you agree with any particular standpoint. But I will say that it's at least worth thinking about the question and maybe you can do measurements to illuminate whether all these ideas might make sense.

And then we'll try to, over the next hour and a half, figure out to what degree this paper maybe should convince us of this optimization in the context of this particular protein.

Now, even if it's the case that somebody convinces you maybe that expression of the lac operon maybe does optimize some cost-benefit analysis. That does not prove that every protein optimize things. So don't get overwhelmed or underwhelmed or whatever it might be.

Let's just first make sure that we understand what we mean by costs and benefits in this case. Can somebody pick one of them? Now, what is a cost and benefit in the context of maybe this paper? Yes.

AUDIENCE: Producing protein requires some kind of resource.

PROFESSOR: Right, requires--

AUDIENCE: [INAUDIBLE] energy--

PROFESSOR: --requires resources of some sort or another to express these proteins. And this can manifest in many different ways. But certainly, if you were not making these proteins, you could have been making some other proteins. And so if these proteins are not helping you, then maybe something else would have.

But they're are many different ways of looking at this. But there is some finite number of things that the cell can do. And the benefits, of course, in the case-- and this is in particular in the case of a lac operon, what does this network allow us to do? Yeah.

AUDIENCE: You get to consume the energy of lactose.

PROFESSOR: Yes.

AUDIENCE: Lets you go faster.

PROFESSOR: That's right, you get to consume lactose in this case. Now, we've already spent some time thinking or discussing the lac operon.

What were the two key components in here in the lac operon? If you were a cell and you wanted to eat lactose, what would you need to do? I'm picking somebody to-- yes, please.

AUDIENCE: It's a gene that you should express, the lac gene?

PROFESSOR: OK, right. So the lac genes. But maybe in a little bit more detail, what do we mean when we say the lac genes? Well, I mean, it's not just lactose. But I mean, what are the things that have to happen if you want to eat anything, I guess? Your cell--

AUDIENCE: Import.

PROFESSOR: Right, so you first have to import it. Now, in some cases, this can be done maybe for some-- maybe nutrients, it could be done even passively, if it crosses the membrane easily. But for most of the things that you might think about, you actually have to do active import.

So this is done by what? Anybody remember? lacY. So lacY is a membrane protein that imports lactose. And then what do you need to do?

AUDIENCE: Break the two apart, and then you can metabolize.

PROFESSOR: Right, then you have to eat it somehow. Now, of course, metabolism is a very complicated thing. But the key thing that's different between lactose and maybe the simple sugars is that you first have to break down the lactose into its constituent parts.

A lactose is a disaccharide composed of two simple monosaccharides. So what you need is you need this lacZ, beta-galactosidase, in order to cleave that bond. And then you have the two simple monosaccharides that can be eaten.

Now, the lac operon also has this lacA. And it's not quite obvious what that thing does, so nobody ever talks about it. But there is a third protein there. But what we always talk about is lacY, that's require to import the lactose and then lacZ that is required to break the lactose down into its monosaccharides.

And then the idea-- and that's not sufficient. You don't take those monosaccharides and instantly make more cells out of it. But the idea is that the rest of the metabolic machinery is kind of there any ways to do other-- that's kind of some assumptions.

Can somebody explain how it is that they measured the cost of expressing these proteins? Yes.

AUDIENCE: So they [INAUDIBLE] expressed these proteins at different levels using different concentrations of IPTG.

PROFESSOR: Right.

AUDIENCE: There was no lactose around, so it was only the cost to no benefits. And then they measured [INAUDIBLE].

PROFESSOR: All right, perfect. OK, so there are several key things in here. So first of all, normally, what we do is it's lactose inside the cell that causes this lac repressor to fall off and then you get expression of the lac operon.

But in order to kind of sidestep or circumvent that normal network, what we are doing in this case is adding IPTG. So IPTG allows one to get expression of-- and what IPTG is that it stops the inhibition of this lac promoter, where you get lacZ and lacY.

Now, the idea here is that you can control the level of expression of this operon, because what we really want is we want to measure a plot of something that you would call cost-- and we'll explore a little bit more what that means-- as a function of the lac operon expression.

And this is often done relative to the full induction of the wild type lac operon. And this is a relative growth rate reduction. So basically, this is a percentage, say, decrease in growth rate.

Now, there was a key thing that you brought up, which is that you want to measure the growth rate in the absence of lactose. Because otherwise as we increase the level of expression here-- so we're controlling this by IPTG, so there's some mapping from IPTG concentration to the level of expression here.

But we want to be able to measure the cost separate from the benefits. So it's important then to grow this in the absence of lactose. So say, no lactose.

But if I just take bacteria and I put them in a tube with say minimal media, salt, so forth, but no lactose, are they going to grow? They need to draw something. So what is it that the authors have done? Yes.

AUDIENCE: Glycerol.

PROFESSOR: That's right, they added some glycerol and in different parts. I think it's 1% glycerol. Does anybody happen to remember? I think, for most of it, it was 0.1%. I tell you what, we'll say a little bit of small concentrations of glycerol.

So the idea is that this is kind of a second rate carbon source. The bacteria are not super happy, but they're OK. And then given this, what they were able demonstrate is that, if they did add lactose, they would have grown faster.

So there's a sense that the lactose does help the cells. But you have to have some glycerol. Otherwise, you can't really measure these things. Yeah.

AUDIENCE: Why is it that-- you were saying if you put like a very good carbon source--

PROFESSOR: Well--

AUDIENCE: You're not going to see any [INAUDIBLE].

PROFESSOR: OK, so first of all what I was saying is that you have to have some carbon source.

AUDIENCE: Sure.

PROFESSOR: Right, so you have to do something. And it's just good conceptually to make sure you think about how you would actually do this experiment. Now, you have to add some carbon source. But the question is, well, what happens if you just added a bunch of glucose?

Now, in that case actually, for some of the other experiments, I think that would have caused problems in the sense that then there would not be any benefits associated with growing or with adding increasing lac operon expression.

For this experiment, in principle, one could have done that, although you really want to measure the costs and associated benefits in some environment, which you're to be doing in later experiments.

So I think it's really from a conceptual standpoint, in principle, you can measure this in glucose, but then you'd always worry, oh, well, maybe it's different. Yeah.

AUDIENCE: [INAUDIBLE].

PROFESSOR: Oh, yeah, right. So you could have broken down-- So the other issue is that, in principle-- and they don't talk about this here-- but yeah, if you add a bunch of glucose, then you would have to have another mutant in order to break the glucose repression, because if you have this preferred carbon source, glucose, then you'll naturally repress the CRP, all of the alternative modes of carbon metabolism just because glucose was kind of the best.

And what was the key conclusion from this first data plot?

AUDIENCE: It's nonlinear.

PROFESSOR: All right, nonlinear. The cost is a function of the lac expression. And it grows super linearly. I always forget what the difference is in concave and convex is. I don't know if other people have this particular brain problem.

But the second derivative is positive. In particular, that means that if you do draw some sort of like line, then they have data that looks something like-- so here is 0.5. We have something that kind of falls below here.

They had about a 0.25. And it was also a little bit below that crossed. They had a 0.75. And then they had a 1.

Why is it that they can't go above 1 here? Why do they not have more data out here?

AUDIENCE: Because you can't have more expression than full expression.

PROFESSOR: You can't have more expression than full expression with this promoter, because what they are doing is they're adding IPTG, so they titrate between 0 and maximal expression from this promoter. In principle, you could always get another one. And then you should be able to go out further, right?

And at maximal expression, they measure about a 4% growth deficit, 0.04, just to give you a sense of scale. So this is 4% deficit. Now, I want to ask a more general question.

So let's imagine that you are measuring some quantity. So we'll say this is some quality y as a function of x. And let's imagine that the true y as a function of x looks like something.

Now, you go and you measure at multiple values of x this curve, because we're very interested in what this curve looks like. Now, the question is, what fraction of the error bars will contain this curve and, of course, contain this is true curve?

So I'm assuming that this curve is the god-given actual thing that you're measuring. And so you measure this quantity with noise. So we measure this some number of times, some number of times. Do you understand the question?

So here, contained the curve. There, it didn't. So what fraction of error bars will contain that curve?

AUDIENCE: [INAUDIBLE].

PROFESSOR: Right. And indeed, what we want-- it's always good-- what were the error bars in the figure 2A in this? Right, well, OK, so they're experimental error, right.

Incidentally, how is it that they actually measure these things? Does anybody-- And so these are actually a result of growing on a nice [INAUDIBLE] well, like a microtiter plate, where they used a checkerboard pattern. And they take 48 different cultures. And they measure the growth rates for each one individually. And then they're plotting the standard error of the mean.

Do you understand what I'm trying to ask you?

AUDIENCE: So in that case, I mean, the size of the error bars, you just want a scaling or something, [? if that's right, ?] because the size of the error bars--

PROFESSOR: Right, well--

AUDIENCE: I just--

PROFESSOR: Yeah, OK, so this is a good question. We'll find out. So it depends on n, where n is the number of samples that we took at each location. Question, yeah.

AUDIENCE: Yeah, the standard error is just the [INAUDIBLE]?

PROFESSOR: Right, so standard error of the mean, this is an important question. What you do is you calculate the standard deviation, divide by the square root of n-- OK, now, I always forget whether it's n or n minus 1, now.

We already did one n minus 1, right? So it's you measure the standard deviation of the data, the standard deviation in y divided by root n, where n is the number of measurements you took at that point.

But of course, when you measure the standard deviation, there was already an n minus 1, right? Have I lost a minus 1? Do you guys-- OK. Yeah.

AUDIENCE: Isn't the standard-- I thought the standard error of the mean and not the actual standard deviation [INAUDIBLE]?

PROFESSOR: Yes. And we're going to spend a lot of time talking about what the difference is between a standard deviation and a standard error of the mean. And it depends on what you're trying to ask. Do you guys understand what I'm trying to ask here?

All right, well, let's just see where we are, and then we'll discuss. OK, ready? 3, 2, 1. All right, so we got many A's, B's, C's. Nobody likes D. OK, but it's very common to see that.

Let's go ahead and-- it's worthwhile, I think there's enough variation to decide. And in particular, between your neighbor, try to agree on why or why not it might depend on n and so forth. We'll just have a minute to think about this.

AUDIENCE: [INTERPOSING VOICES].

PROFESSOR: So what do you guys think?

AUDIENCE: We're still [INAUDIBLE].

PROFESSOR: OK, no, that's fine.

AUDIENCE: [INTERPOSING VOICES].

PROFESSOR: Why don't we go ahead and reconvene, so we can kind of try to figure out what is going on here. I just want to see if anybody has changed their opinion as a result of discussing with their neighbor. All right, let's see it. 3, 2, 1.

Some people are not even willing to-- all right, OK, so it's interesting. So now, actually, it seems like there is some convergence to this. Should I feel like that you guys in general have more accurate votes than past years somehow. I don't know.

So let's try to figure out why it might be that and what this thing standard deviation is. Let's try to figure out what all these things are.

So the idea is that we're going to measure some quantity, but it's a measurement with error. And for now, we'll just assume that the measurement error is Gaussian distributed, because otherwise, we get confused and everything.

So let's say-- so what we're going to do is we're going to measure some quantity with error. OK, so it's-- Now, what we're interested in is not really the width of the resulting distribution, because that's a result of how accurate, how good we are as experimentalists.

What we're really interested in is this true quantity, so the mean of our distribution. We want to know mean. Now, if you read the supplemental section of this paper, what you'll see is that there's a significant standard deviation to their measurements, where the standard deviation, they don't actually quote exactly what it is.

But they have plots of the histograms, where like, for example, this is a histogram of the different growth rate measurements across those 48 samples, and actually, in this case, even more than that. But what you see is that the standard deviation might be 3%, 4%.

So the standard deviation is actually something that's big. Now the question is, what we really want to now is, how the mean of these distributions are shifting, because we want to know something about this true underlying growth rate deficit, because each individual measurement is a rather noisy measurement.

And indeed, in this case, the noise is larger than the signal. But if we believe that we don't have a shifting systematic error, then we can average that out just by making many measurements.

So the question is, so the standard error of the mean, what it's telling us about is that if you measure this quantity n times, you get some mean. So let's say that this is a-- ooh, it's a little bit of a broad somehow Gaussian.

So this is a histogram of our measurements of this thing. And what we want to know is the mean of this distribution. So this is similar to our discussion of super resolution microscopy.

And the question is, how will the mean be distributed if you have these n measurements? It's a Gaussian distribution. And it's certainly a Gaussian distribution, because of course, if we-- what we're doing is we're measuring a bunch of Gaussians.

And we're going to add them all together. And then we're going to calculate the mean. So we definitely get a Gaussian. And indeed, because of the central limit theorem, this is also saying that even if your errors were not distributed super Gaussian, even if they were a little bit funny shaped, the resulting distributions of the means will look more like a Gaussian.

Now, what we often plot is the standard error of the mean, which is kind of the plus or minus 1 sigma of the distribution of the mean. So if we go and we sample from this distribution n times, we'll get some value. If we sample from it again, we'll get some other value, so forth.

Now, the distribution of the means we're going to calculate is not going to be a representation of the full standard deviation. But rather, it's going to be suppressed by this root n, where n is the number that we're sampling.

So if you look at the histogram of the means, you're going to get a Gaussian in here-- OK, that's not a very nice Gaussian, but-- with a width that is the standard deviation divided by root n.

Now, if we assume that we don't have any systematic error, then this distribution of means that you would have calculated-- it's Gaussian, it's centered on the right value, but about a third of the time, it'll be beyond the plus or minus 1 sigma.

And what that means is that about a third of time, if you plot this standard error of the mean, it should fall off of the kind of true curve. And this basically does not depend on n. And can somebody say why it doesn't? Yeah.

AUDIENCE: Yeah, I think I was sort of confusing myself, but this makes sense. So yeah, I mean, you know that these error bars will shrink, if you take more measurements. But on the other hand, the actual measurements will be closer--

PROFESSOR: That's right.

AUDIENCE: --to the true value.

PROFESSOR: That's right. So what happens is that as you sample from this distribution a larger number n times, then your error bars shrink, but your measurements get closer to the curve. And those two effects cancel. So you should end up roughly with 2/3 of the errors bars containing this curve, or 1/3 falling off.

And I think that this is a little bit surprising, because there's always a sense that we feel that there's something wrong with our measurements or something wrong with our model or whatnot, if any error bar does not contain the line.

I mean, I feel like I often see there's this effort that people have to try to make it so that these error bars always overlap with some underlying curve that is supposed to represent reality. But that's not, in principle, supposed to be true.

Are there any questions about where we are right now? OK. Now, what I want to do is something slightly different, which is ask-- let's say that this is a curve that is not the underlying reality but is instead a fit to the data. How does this change anything that we've said? Or does it?

All right, well, OK, let's-- OK, so we're going to say do fit. The question is does this change the thing here? Do you understand? Change, A is Yes, B is No. Yes.

AUDIENCE: Do you have the same modeling for the fit as we did for the original--

PROFESSOR: Yeah, well, let's say that this was a curve predicted by some fancy theory but that you have to specify the mass of something and the-- so I don't know, there are two things that are specified. So what you do is you fit.

And the question is, does it change what fraction of the error bars you expect to contain the true curve? Ready? Is it not clear what I'm asking?

AUDIENCE: But the true curve is determined by the god.

[LAUGHTER]

PROFESSOR: Right, so the truth curves-- we don't need to get too much into this. But I mean, the reason we're doing science is to try to look into the mind of God, right? So we were doing a fit to try to--

AUDIENCE: But you can fit anything to anything. You know, what does that mean? Do you see what I mean? Like, I could--

AUDIENCE: It depends on whether--

AUDIENCE: I could get curve that passes through each and every of these points points, if you give me enough time with it. So I guess I don't understand the question.

[INTERPOSING VOICES].

PROFESSOR: All right, yeah. OK, but I think you're arguing for something already maybe. But let's just say that this was a-- I mean, let's just for concreteness let's say that I measured at 15 values of x. I have some error bars and some error.

But then I needed three parameters to characterize this curve. And so those I used to fit. Are you happier with three fitting parameters and 15 measurements? All right, let's just see where we are. OK, ready, 3, 2, 1.

OK, so we have a majority of A but a significant minority of B. So just to be a lot more concrete, can somebody say why they're saying yes? Yeah.

AUDIENCE: I guess intuitively, [INAUDIBLE] we try to optimize the number of error bars that go through.

PROFESSOR: Yeah, so the fit is somehow trying to get the curve to go near the error bars. And typically when we do a fit, we're typically trying to minimize this mean squared error or deviation from our curve to the data point.

How much you expect this to make a difference? So for concreteness again, let's say that I had 15 values of x that I was measuring things at. Now, we expect say five of them-- five will miss true curve, we decided roughly.

Now the question is, what happens if we, instead of having this true curve, if we do a fit using these three parameters? How much of a difference should it make to this very, very roughly?

We'll see-- Now, I'm asking roughly how many of these error bars do you expect to then miss the fitted curve? And this is we used three fitting parameters, say. That was parameters over there.

Do understand the question? So instead of plotting this god-given curve, instead we're plotting a curve that I'm giving you, where I use three fitting parameters to fit to the data.

And I'm just trying to get it roughly. I think that this is not a rigorous statement I'm about to make, but just so that we're all roughly on the same page. All right, ready, 3, 2, 1.

Right, so it'll be somewhere in here. And I think this is not quite true. But the idea is that, in particular, if you make n measurements and then you use n fitting parameters, in general you will get a perfect fit, i.e. the curve will go through every single data point amazingly perfectly.

So if I give you 15 measurements across here and then I give you a 15-degree polynomial-- I guess, we only need a 14-degree polynomial with 15 free parameters-- then that polynomial will go through everyone of your data points spot on, not even a question of whether it goes through the error bars. I'm saying literally-- and that's just because you're just solving an equation at that stage.

Now, this is a stupid statement, except that once you're kind of like in the heat of the moment, eagerly trying to do some fitting for your advisor or whatever, it's easy to fall into this trap, where you just kind of like add extra parameters.

I mean, I definitely remember in graduate school, I was surprised. I was like, oh, this thing, it works wonderfully. It's like it seems to magically goes through all my data. And then I felt very stupid like 30 seconds later. But this is just a very easy thing to screw up and forget about.

So what this is saying is that, if you see a curve-- if in the course of your work or if you're reading a paper and you see some curve and you want to know something about how much information is it or whether things look reasonable given the data, it's useful to kind of orient yourself relative to these statements, that depending on how many free parameters you're kind of using, you expect a larger or smaller number of these data points to kind of go through the curve that you see.

But I would just want to stress that you don't want to be anywhere close to the point where you have a number of parameters equal to the number of kind of measurements that you're making.

And for any sort of reasonable curve describing what you hope is a reality, you expect some of those data points with their error bars to kind of miss the curve. And that doesn't mean that they're sloppy experimentalists. It doesn't mean whatever.

OK, now coming back to the task at hand, do you understand why they're plotting the standard error of the mean rather than the standard deviation?

Because what your interest in, in principle, is not-- the question you're trying to answer is not how variable are their measurements but to what certainty can they claim to know the actual god-given, real cost associated with expressing these proteins as a function of the expression level. And for that, you really want to ask about the standard error of the mean.

Great. So now, we can come back and ask about, why did I just spend half an hour talking about standard error of the mean, standard deviations, fitting to data? Well, you guys are probably all asking yourself that question. But does anybody have an answer if I-- Yeah.

AUDIENCE: You can fit with different curves if you use different things.

PROFESSOR: You can fit with different curves if you-- yeah, I think that that's hard to argue that statement. But the statement is a little bit like "different proteins have different expression levels," but a little bit more concrete maybe. Yeah.

AUDIENCE: So in this case, I didn't check their calculations, but if you have a natural line, then you can't make this calculation of optimization.

PROFESSOR: Yeah, but I think that-- right, so--

AUDIENCE: In the sense that there won't be [INAUDIBLE].

PROFESSOR: Yeah, OK. So I think that this is a tricky thing. The data certainly do argue for a super linear cost. But I would say that they argued for it rather weakly, in that if you look at their data and you just fit a line, you would say, it's maybe OK.

And of course, once again, should we be surprised that the quadratic fits better? No. And this is a very dangerous thing, if you're comparing models. It'll always be the case, if you add another parameter, it will look better.

But the question then is how strong of a case should we make of this? And then how important is it for the conclusions of the study?

Now, in addition to the line and the quadratic, they had another curve in here, which looks like-- let me see if I can get it right for you guys. So this is fine, tricky thing. So it's the dashed line that looks very similar to the solid quadratic line.

Can somebody remind us what the difference was between those two non-linear curves that they had? Why do they have two curves that look so similar?

AUDIENCE: I think the dashed line responds to some model where there's only so much of this certain resource that--

PROFESSOR: Right, OK, so my dashed line is their red line, just to-- OK dashed red in the paper. So it's this line where there's a finite amount of resources or protein-making machinery that the cell has. And if you use them up, then you don't get any growth.

And of course, that statement kind of has to be true on some level. And the question is whether--

AUDIENCE: --that scale is--

PROFESSOR: --it's relevant here, right. Certainly, I would say that one question is whether you can reject the hypothesis that this cost function is a line. Another question is whether you can distinguish between the two quadratic or the two non-linear curves based on the data.

And I think the answer to the second question is certainly not. And they don't claim that they can. But it's important to just note that it's just impossible for them to assume-- I mean, those curves are so, so similar over the entire range where they have data that it's going to be possible to distinguish those two things.

But does it matter which of the two cost functions is the true cost function? Yeah.

AUDIENCE: Is it because the [INAUDIBLE] where the marginal benefits become zero is like inside the range where the cost functions are still exactly the same?

PROFESSOR: OK, right. So what you're saying is that the two cost functions they have they behave similarly over the range that is relevant maybe, so then therefore, it doesn't matter. Is that-- or am I-- OK.

So why do they have to cost functions there then, why two non-linear cost functions? Just to provide variety in our modeling? Yep.

AUDIENCE: They were doing another experiment later on, and they said something like something was saturated and that was modeled by the second cost function.

PROFESSOR: Right, yes. That's right. And what's the later experiment they're going to do, just so that we're--

AUDIENCE: You should ask somebody else to explain that, not me.

PROFESSOR: You regret opening your mouth. No, OK. So yeah, so what is the experiment that they're going to do?

AUDIENCE: Measuring the benefit?

PROFESSOR: So next, they're going to measure the benefit. But this question about the two cost functions is not somehow relevant yet for the benefit part. Yes.

AUDIENCE: So they're doing it in different concentrations of lactose and seeing if the protein expression could adapt [INAUDIBLE].

PROFESSOR: Right, after a long time. So they actually do laboratory evolution experiments, where they grow these bacterial populations in different lactose concentrations. And then they look to see what level of the lac operon expression does the population evolve to.

So what they're trying to do here is they're trying to say, OK, well, we can measure some cost as a function of expression. Maybe we can measure some benefits as a function of expression. And then from that, we'd like to be able to predict where the population will evolve to.

And they had these two non-linear cost functions, which based on the data they have, they can't distinguish. But they say, oh, well, they're both kind of reasonable cost functions.

And in some ways, maybe the problem here is that the two costs functions end up being wildly different in terms of predicting what happens for large lac concentrations, where you would want to express more of the protein.

Do you guys-- do you remember this or not? Sort of. And that's actually-- well, you might as well just look at that. So that's figure 4.

That's the normalized lacZ activity that the populations evolve to as a function of the lactose concentration they're evolving in. And what you see is that this red curve corresponding to the finite resources cost function, it explains the data. Whereas, the other ones very much do not.

And that's just because these other models then would predict that if you grow the cells in a lot of lactose they should express out to five times the lac expression, much, much, much more, which is not what they see experimentally. Yes.

AUDIENCE: Is there another way to put a bound on the expression, because of this expression we have? You mentioned for that promoter, it's not possible to--

PROFESSOR: OK, but the idea of evolution is that evolution can make it a stronger promoter. So you guys, one statement is, given this DNA sequence at that promoter, how much expression can you get? And the most you can get is this amount that's normalized to 1. But if you make mutations in that promoter, then you could go out further.

So the question now is, after we kind of tell you the results of these evolution experiments, how much should that favor this dashed red line, this super linear cost function with finite resources? And on one level, you'd say, oh, well, that's pretty compelling.

On another level, later people that have come and measured this find that it's basically a line. So are there any questions? So it seems to basically be not true within this range. It is the case that if you go out far enough, then the growth does go to 0. But that's much further out. Yes.

AUDIENCE: After they-- on the experiment, they had [INAUDIBLE] expression protein at the [INAUDIBLE] level. Why didn't they go back and do the experiment again, just to see

[INAUDIBLE]?

PROFESSOR: OK, so actually, one of these curves-- so the triangle, the sort of teal triangle, it is indeed higher up. And it's kind of here. So they do have a data point that is further beyond and is, again, above that curve. So that does provide somewhat further support for a non-linear model.

But again, there's a question of how strong that should be and so forth. And indeed, I'd say, for example, Terry Hwa has spent a lot of time characterizing growth rates as a function of many, many things.

And if you measure the relative growth rate as a function of a non-useful protein expression-- and what he finds is that this thing basically looks like a line in this axis.

And it saturates at around if you're at 30% maybe of total protein expression. So this is a lot. But this is kind of where the cell just can't handle that.

So Terry Hwa has recently been exploring a lot of these sort of phenomenological growth laws, where he imposes costs of various sorts and then looks at how the cell kind of responds.

And what he finds is just a remarkably large number of lines in various spaces that I find very surprising, but that he can understand using kind of some phenomenological modeling. But this is one of like a dozen lines that he sees of various axes doing things.

But the point here is that, as a function of the level of expression of these non-useful proteins, what he sees is that for a variety of different proteins-- including beta-gal but also beta-lactamase and other proteins that are not being used in that particular environment-- what he sees is that there's basically a linear cost growth, as you impose this non-useful protein expression.

So I'd say that this basic statement of it being not-- the statement of cost being super linear, I think, ends up not being true. Now, what does it mean for this paper?

AUDIENCE: I mean, they still presented with same hypothesis and had these data to back up some of it.

PROFESSOR: Yeah, right. So it's a very interesting hypothesis. They did nice evolution experiments, where they saw the population adapt to different levels.

But what does it mean about the predictions, in particular, in the sense that if you measure cost and benefits, then you want to predict where it's going to evolve. What happens? If it's the case that cost as a function of expression is actually linear, then what does that mean for their ability to predict what's going to happen?

AUDIENCE: Seems like if they use their same model for the benefit in this linear cost, that their predictions would be really off [INAUDIBLE].

PROFESSOR: Yeah, right. So the problem is that if you actually use a linear function here, then their model doesn't even predict that there should be an optimum, because their benefit function ends up also being essentially linear with the amount of this protein expressed.

So if you have two lines-- so overall growth is something like goes as benefits minus costs. And maybe this is a relative growth. So if you have a line here and a line here, no optimum. So that's kind of a bummer.

But it doesn't mean that that's-- in biology, eventually things are non-linear, so there should be some optimum. And actually, what I would say is that I think that the non-linearity is probably actually here.

That's the non-linearity that's relevant, maybe, is dominated on the benefit side rather than the cost side. My guess as to what's going on here is that rather than the costs growing super linearly with the expression level, rather the benefits will be sub-linear with the expression level. And why might that be?

AUDIENCE: We're just seeing them apart, splitting up more lactose that's useful, just so it can't metabolize more of it.

PROFESSOR: Right, you know, at some point, it's just that the cell doesn't need more sugar. And then it's not going to be as useful.

And even before you get to that regime, I think there are various ways in which cells may be able to use the sugar more or less efficiently, depending on how much they have it, which means that as-- and this is just like for us, the first slice of pizza is great. But then once you're at the fifth one, you start to feel a little bit full.

So in general benefits as a function of anything, should have some saturating behavior. And my sense is that this is basically why there's an optimum here.

Now, of course, I'd say that all these cost functions behave very similarly in here. So the predictions that they make in here are really not very sensitive to which of the cost functions they use. And those are all still then relevant and valid.

The question is just trying to predict what happens beyond the range that you have data is very hard, because it depends very much on what your curve does past that region.

So I guess I've made an argument that I think that maybe what's happening is that the benefit function here is non-linear. But what did they actually do to measure the benefits, because this is not I think totally obvious either?

So what should I be plotting? Well, this is still a relative growth rate. And here, this was actually lactose concentrations. So this is not lac expression, which is the most obvious thing that you would want to do, but that's harder.

And what they show is that their model is sort of consistent on this axis. This is external lactose. And the idea is that-- here is 0-- in the absence of any lactose, if you induce the lac operon, then you're at this minus 4 and 1/2% or whatnot.

So it kind of starts out down here. And then up here, it comes out up to above 0.1. So this is the first 4, maybe 4.5%. This is up here at 10%. And you end up with a curve that kind of goes from 4% or 5% deficit up to 10% or 11% advantage.

And this is at full induction of the lac operon. What this is saying is that if you're making the proteins to break down and consume lactose, then there's a cost. That's just how they plotted it.

But that the benefits do indeed outweigh the costs at some concentration of lactose. But then here, there's a saturation. And here, the saturation in their model-- they get a saturation just because of the dynamics of import.

So what they assume is that there's a Michaelis-Menten kinetics for import. So the import rate kind of goes as the concentration of the lactose divided by some k plus the concentration again of lactose, so Michaelis-Menten dynamics.

But of course, if you have more of the protein lacY, then you'll be able to import more. So just because you have saturation as a function of lactose does not mean that you'll have saturation in terms of the number of proteins that you're making. Do you understand why I'm saying that?

And indeed, I would say that many underlying models could have been consistent with this data as well. So I'd say that their data does not reject the hypothesis that the benefit function is sublinear. Yeah.

AUDIENCE: So that you just said if you have more lacY and import more and it would saturate to an [INAUDIBLE]. So you could imagine that by evolution something happened there. So why would you even expect the prediction of this cost-benefit analysis? You see what I mean?

PROFESSOR: OK, so you're saying that evolution might be able to change other things as well to kind of fiddle-- yeah. I think this is an important question. I think the basic answer is that there are some things that are easier for evolution to do than others. And also that somethings have maybe already been optimized.

Now, relevant to this point, so they did these laboratory evolution experiments, and there was one category of mutation that they did not see. Does anybody remember what that was?

What's the most straightforward way of kind of getting around all this cost-benefit discussion that we've just had?

So the one thing that they did not see was significant improvements in the enzyme. So they checked, and they found that they did not see any increase the lacZ activity normalized by the amount of the lacZ that was being made.

Now, that might make sense, because if this enzyme has already been gone through millions of years of optimization to break down lactose, then it's reasonable to say, oh, well, in the next five generations in the lab, it maybe won't improve.

Of course, you always have to be careful about this, because it could be that some sequence slash structure is best when you're thinking about-- when is it that E. coli might see lactose? Our gut.

So you imagine you have bacteria in the gut. That's a different environment than in the lab. So it could be very well that the enzyme, because of the pH and all these other things, the enzyme actually could adapt to the lab, even though it may have already been adapted to our gut.

So you have to got to be careful about this kind of argument always. But of course, once you see the result, then you say, oh, well, that's because of this.

So I just want to make sure that we know what these experiments look like. So they went for 500 generations. So it's useful to ask how long this experiment should have taken.

Is it closest to three days, three weeks? Anytime you read about an experiment, it's useful just to have some notion of what the authors went through in order to bring you the results you're reading about.

If you are not sure, you can just make a guess. OK, ready, 3, 2, 1. All right, so we have some number of A's, some number of B's, and a couple of C's.

Well, one thing you might say is, how fast can E. coli divide? OK, on one level, you may say oh, about 20 minutes. That should give us what? 75-ish generations a day. So we should be able to get here in a week or something, maybe.

But that's not what they did, for several reasons. First of all, this would be in rich media. In the environment that they are doing this in, it's a bit slower. But that would get you maybe to the two or three-week mark.

But that still is not what happened. They actually had to go for three months. And this is because experiments are not always keeping cells constantly dividing at their maximal rates.

The standard way that we do this is what's known as kind of daily batch culture. And does anybody know how much they diluted by each day? Yeah, so I think it was diluting by a factor of 100.

So it's daily batch culture with 100x dilution, which corresponds to about 6.6 generations per day. So this is very far from what you would think of as kind of the best they could possibly do. And what it means is that, yeah, it does take about three months for them to have done this experiment.

It also means that if you look at the number over the course of each day, this is n max. And they dilute-- this is n max over 100-- so they dilute by a factor of 100.

When you transfer cells from a saturated state into new environment, do they start dividing immediately, for those of you who have done this experiment? No. It's going to take an hour or two for them to get going.

But then they're going to start dividing. And this on a log scale maybe-- log N. And what you'll see is they kind of go-- they're dividing exponentially and then they saturate.

Indeed, they're going to saturate for about a fair amount of time. So this might be an hour or two. This might be say five hours.

But then you still have another roughly 20 hours to go before the next dilution. And then we repeat. So they actually saturated for a fair fraction of the day.

Now, in all these discussions of laboratory evolution-- and in many of the calculations we're going to be doing over the next couple of weeks-- we'll typically assume that what is being optimized is the growth rate, the rate of division.

But you can imagine there being other things that might possibly be optimized in the course of these sorts of experiments. Can somebody volunteer what are other things?

AUDIENCE: Maximum density?

PROFESSOR: Right, so you could imagine, if you could just eke out one more division out there, then you could get an advantage. And there's a whole set of interesting things, these growth advantage. It's stationary phase or the GASP mutants, where the focus is on trying to do well here.

And also you can imagine related maybe, if you do better out for this period, cells will start dying eventually. So if you have a lower rate of death at saturation, then you can also spread. Other-- yep.

AUDIENCE: Sorry, can I ask a quick question? What's a possible reason for the initial [INAUDIBLE] used [INAUDIBLE] at the beginning?

PROFESSOR: Yeah, right. So I think it's basically that when the cells are saturated, they generally enter a rather distinct physiological state, as compared to the dividing state. And I think the longer they sit in this saturated phase, the longer it's going to take them to get going in the next day, for example.

And it's also the case that cells in saturated culture tend to be more resistant to a variety of perturbations of various sorts-- so if you're talking about heat, salt, this, that, and the other thing. What's something else they could be optimized here? If you were imagining you're a cell, you want to spread, what would you do?

AUDIENCE: [INAUDIBLE].

PROFESSOR: OK, right. So we're saying that the media is specified by the experimentalist. So you're the cell in this Gedanken experiment.

AUDIENCE: You'd divide yourself.

PROFESSOR: Right, so you can eat the other cells, yeah. Well, and in particular, actually out here, this is part of how the GASP mutants spread, is that when other cells start to die, they lyse their contents. And then the cells that are surviving can actually eat the contents, yeah.

AUDIENCE: Is this a way to coordinate between different cells, so that they can sort of evenly distribute themselves in the media, so you don't have to many--

PROFESSOR: OK, right. So I'm actually assuming here it's well mixed, so that in principle would not be an issue. But yeah, so you can imagine spatial effects of various sorts being relevant.

I guess I just drew this up here to highlight that, in principle, you can also decrease the lag time. So if you start dividing more rapidly at the beginning of the day, then you'll get to spread before your neighbors and your genotypes will indeed spread. Yep.

AUDIENCE: So I just know so little about cells, but is it true that a lot of the cells could survive and be the same cell for that whole duration when they were in the stationary phase?

PROFESSOR: You're asking whether the--

AUDIENCE: Yeah, whether a cell that entered the beginning of the stationary phase, that same cell would have a pretty good chance of--

PROFESSOR: Yeah, over I think this sort of 12-hour type period, I think the answer is yes. But if you go for an extra day or two, then I think you can start getting extensive cell death.

AUDIENCE: Because then who knows? Maybe long enough, though, they would develop a little clock to let them know that it was about to split.

PROFESSOR: Yes, right. Yeah, so people have thought about-- and I'm not sure if this--

AUDIENCE: And it seems like it would.

PROFESSOR: --particular effect is-- yeah, right. But I just want to mention that this is something that you kind of maybe would expect, indeed, you see in these-- so they're a famous set of experiments done by Richard Lenski at Michigan State, where he's been dividing six or eight-- doing daily batch dilutions of equal E. coli cultures now for decades.

So he started, I don't know, late '80s or so. I don't know if you guys remember. So he's gone tens of thousands of generations and has seen a bunch of remarkable things. One of the things that he has seen, as you might have expected, is a decrease in the lag time of the vector area.

So what we have now is a situation where they add IPTG, so that all the cells are in principle start out expressing the lac operon. And then they grow the cells over time.

And what they see is that the lacZ activity, it starts out at being 1, normalized, for all the cultures, because there's IPTG, so it doesn't matter how much lactose there is. But what they see is that over time, they see things that look like this.

So the 0.5 millimolar lactose didn't change very much. But if you look at some of the others, like no lactose, there was significant decrease in expression. Whereas, up here at, for example, 2 millimolar lactose, they see an increase.

So what you see is that there really are evolutionary changes of these strains, because-- and it's very, very relevant that they had IPTG in the media. So if they did this experiment without IPTG, do you have any sense of what would kind of happen to the cells? I mean, how would that change the results? Yep.

AUDIENCE: The expression level would be determined by [INAUDIBLE].

PROFESSOR: Right, so the expression would be determined by the lactose. But let's say that after 500 generations, we put them all in a millimolar lactose. How different do you think they're going to be?

I mean, do think that the culture grown, for example, in the absence of lactose, do you think that it would still be able to eat lactose after 500 generations in that experiment? Hm?

AUDIENCE: Yes.

PROFESSOR: Yes, OK. And yeah, so what's the difference? I mean, why are you saying yes or what's the--

AUDIENCE: [INAUDIBLE]. I don't know how hard it is--

PROFESSOR: Well, yeah, right. So this is an experiment with IPTG. And now, I'm just trying to think about or imagine what would have happened if they had done the same experiment without IPTG just growing in that environment, in particular, if you grow minus IPTG and then minus lactose for 500 generations?

And then what I want to ask is, OK, let's say that you go over there and you just add lactose. Will the cells, do you think, be able to grow and the lactose? OK. And so why is it that here the answer seems to be no?

So here, we have evolved a population that not only it's not expressing the lacZ activity here. But indeed, if you put lactose in there, it doesn't express. So these cells can no longer grow on lactose. So what's the key difference here? Yep.

AUDIENCE: So I mean, there's no [INAUDIBLE] in this case.

PROFESSOR: Right. Now I think this is just really important. So in this case, there is approximately, we'll say, no cost to having the lac operon on there, because it's just not being expressed.

So then the only cost is associated with DNA replication. So the advantage associated with shutting off or removing the ability to grow on lactose is just really minimal. And indeed, in this culture, the authors did say, what happened.

AUDIENCE: Yeah, the entire gene is diluted, right?

PROFESSOR: Yeah, right. So it was almost a kB was just removed from the genome. And that kB included the promoter. And so that it just-- yeah. So it's not going to be able to grow on lactose anymore.

But the key thing is here, these cells were subject to this 5% cost associated with making the lac operon, which means that that mutant that appeared, it had a 5% advantage, and so it was able to spread throughout the population.

Whereas, what they could see is that the evolved lacZ activity indeed was different, depending on how much lactose they had in the culture. And this is in the presence by IPTG, so they removed that feedback loop.

And in these experiments, anyway you slice it, the normalized lacZ activity did not go above around 1.2 or 1.3. So there is some non-linearity that is somehow constraining those cells from going up to increased expression very much beyond the wild type.

We are out of time, but on Tuesday, we'll start talking about evolution, and in particular, in the context of neutral evolution, as kind of a null model to try to understand these dynamics.

And we will also talk more about why it takes as long as it does before you start seeing anything happening here. If you have any questions, please feel free to come on up.