DOUG LAUFFENBURGER: So we shall start. I haven't had the pleasure of meeting most of you. I'm Doug Lauffenburger. I'm gratefully invited for a guest presentation here. So I'll definitely enjoy it.

There should be plenty of time. I'm not racing through a lot of material, so feel free to interrupt me with questions. And of course I'll try to respond as best I can.

OK. Who has looked at the background materials that were posted on the web a long time ago, last night? Who already will admit to having looked at it? Good. All right. I guess that means I should do this because otherwise if you've read it already then there'd be no point, right? OK. OK.

Well, where we are in your semester-- you're learning a lot of things across the whole spectrum of computational systems biology. I hope I'll add something in here. It's actually a very specific topic.

We talk about modeling of cell signaling networks, and in particular, one approach is worth going through today and that's the logic modeling framework.

So I'll give you a little bit of a conceptual background for the first 10 or 15 minutes. Then we'll launch into the particular example that was in the main paper. And a little side light with an application of it to a particular cancer problem. And then that should take us pretty much to the end. OK.

OK. The biological topic here is cell signaling, primarily mammalian cells. Certainly applicable to microbial cells in a simpler sense. So just to place the context, in mammalian cell biology, I'm a bio-engineer and a cell biologist at the same time.

We're very interested in what controls the cell behavior, their phenotypic response.

We know that it's in fact controlled by what it sees in their environment, growth factors, hormones, extracellular matrix, mechanical forces, cell-cell contacts.

A variety of queues in the environment and the way these govern phenotype or control phenotype is that they influence, they regulate what I would call the execution processes. The crucial execution processes such as gene expression, transcription, and translation are governed by extracellular factors. Metabolism, synthesis of new molecules, cytoskeleton, motors, forced generation.

These things all carry out phenotype governed by the extracellular stimuli or cues. And it happens via these biochemical signaling pathways that are activated primarily by cell surface receptors in the plasmid membrane-- cascades of biochemical reactions, mostly enzymatic. Some protein-protein docking, mostly post-translational modifications. Kinase phosphate reactions adding and taking off phosphate groups that change protein activities at locations and so forth.

It could be other types of post-translation modifications. It could be second messengers, calcium, ATP aces and so forth. So, the extracellular-- now my battery's dead. That's not good. Oh, there we go. Extracellular stimuli, generate the signals. They regulate gene expression, metabolism, cytoskeleton. They carry out phenotype. OK.

So we want to learn about cell signaling network operations. There's actually multiple pathways involved. We really need to study many of them in concert to understand what the cells are doing.

And a big question is, what kind of information do we need to study this? And on the end, we'd be interested in how phenotypic behavior does arise from variations and mutations in the genomic content of cells.

But that genomic content, of course, is not modified, but its effects are influenced by what's the environment to these extracellular cues, log ins and so forth. So they influence what message is expressed.

From that message they influence what's actually translated into protein. From those proteins they influence the post-translational modifications and what the proteins are actually doing.

And so, in the end, the phenotype is carried out by these protein operations. And the question is, what information level that we might want to study.

And of course you would love to have the information content at all levels-- genomic information, transcriptional information, translational information, post-translational information.

So integrating all those different data levels can be extremely valuable. In terms of the models I'm going to talk about today, they've essentially been living at the level of protein activities in these signaling pathways. OK.

That will be the kind of data sets you'll see that will be analyzed with respect to the models. Obviously, they arise from these underlying mechanisms that, as influenced by the environmental context, altering the signaling protein activities.

OK. And what's very interesting and there's going to be more and more progress in the coming years is relating what's in the genomic information-- mutations and variations to what's happening at the protein level. And some of the other instructors in this class are really some of the world's experts in figuring out how to do this.

I'd like to just show this example as a motivation for this kind of approach. And that is if you do gene sequencing of many patient tumors-- in this case, I believe this was a paper on pancreatic tumors. This has been shown for pretty much every other type of tumor since then. In any given patient tumor, each one of these bars, there's dozens of mutations in each tumor.

And a variety of types-- deletions, amplifications mutations and by and large, they're all different. There's very few mutations themselves that really carry over to a substantial proportion of one patient's tumor to another. There's some special cases that are fairly pervasive, but the predominant of these dozens and dozens of mutations and variations are different from one patient to another, and even in the same patient.

So what's emerging as a productive way to think about this-- How do all these different types of mutations and specific mutations actually lead to classes of similar pathologies? And that is they tend to reside in what can be identified as pathways-- circuits, machines, things that are actually carrying out function at the protein level.

So for instance-- I'm losing this again. For these pancreatic cancers on this wheel are about a dozen different signaling pathways and self-cycle control pathways and apoptosis controlled pathways. And if you look at any individual patient tumors, like this green one or this red one-- two different patients.

If you actually look at the mutations at the genomic level, they're entirely different in the green patient tumor versus the red patient tumor. So if you're just trying to match gene mutation to pancreatic cancer, these two patients would look entirely different.

But, it turns out, that you can line up their mutations into the same pathways and say, OK, the red tumor and the green tumor both have mutations that affect the TGF beta pathway. They're different mutations, but they've just regulated that pathway.

And similarly, you can do that with pretty much all of the other mutations. That these tumors have been dysregulated in terms of particular pathways. But patient to patient to patient, it's happened by different genomic gene sequence mutations.

So that the ability to look at these protein level pathways is a way of making really good productive sense of the gene sequencing data. So there's lots of labs trying to go from gene sequence up to pathway modulation. In our case, we're not going to show you that here. We're going to say, this is a motivation for starting at the protein level.

And I'd like to show this picture too. Number one because it's such an anachronism. This is a circuit board from decades and decades and decades ago that none of you would recognize.

But, in the molecular biology world, this kind of a picture, and in its modern form is viewed as a very appealing metaphor for how to think about these signaling pathways and signaling networks that take the extracellular information and turn it into governance of transcription, metabolism, cytoskeleton, and phenotype.

So, just this metaphor of circuitry, where in white, the extracellular ligands, growth factors are somehow wired to the blue. The cell surface receptors, or B for instance-- they're wired too. Kinases and other signaling proteins-- they're wired to transcription factors, self-cycle control regulators, apoptosis regulators.

So these very famous folks in cancer biology say, what you've got to understand is, these signalling networks as circuitry. And if the circuitry is dysregulated somehow, the wiring is different, then that's what's underlying malignant behavior.

So, this is really beautiful but it's pretty much useless, right. Because there's no prediction or calculation or even hypothesis generation one can do from a picture like this. Yes, it's circuitry, but what do I do with it?

So, what I want to show you today are efforts to turn them into what I would call an actionable model, a computable model. Yes, it looks kind of like circuitry, but in fact you would know how to do a calculation that would fit it to data and predict new data. And then you have, in fact, a model rather than a metaphor. That's the idea.

So, one question is, if you want to turn that into a formal mathematical framework for circuitry that you can calculate-- what kind of mathematics might you use? And in this class you're learning a whole spectrum of things.

And one can think about it on one hand, if we knew all of those components and how they interacted, and could estimate rate constance and so forth, we could write differential equations for maybe the dozens and dozens of components and interactions and predict how they would play out dynamically with time.

For most systems with the complexity that's really controlling cell biology, at this point in time, this is almost impossible. There's only rare cases where enough is known about signaling biochemistry to really write down differential equations for what's going on.

At the other extreme, of course, is the type of mathematics one gets out of very, very large data sets, sequencing data sets, transcriptional, and so forth. More informatics type of analysis, where it has to do with multivariate regression and clustering, mutual information.

And what we've been working on is someplace up in the middle where you don't have enough mechanistic prior knowledge to write this formal of physics, and yet takes you someplace beyond statistical associations. And this is one of the areas that might be worth your learning in this class.

OK, this is really the same set of computational methods, just like it's cast in a little bit different form that delineates competition modeling, really into two kinds of classes.

What's traditionally appreciated in most fields of engineering and physics are differential equations that are very theory driven. You have a theory. You have prior knowledge for what's happening. You're writing down the components involved, you're writing down how they interact.

And typically, algebraic equations for those differential equations describe your theory, describe your prior knowledge. And now it's formalized and you estimate rate constants and so forth.

Another whole class of information is data driven, in which, you really don't have a good theory about what components matter and how they interact. And so you start with data sets and from it you do classification or typologies or associations with different types of mathematics that at least try to make sense and get hypotheses out of these large data sets, where you don't have any theory.

One reason that logic modeling appeals to me, is that it actually can be applied in either the theory driven or the data driven mode. You can say, I know nothing about my system. I just generate large data sets of signaling network activities induced by different stimuli, but I'm going to try to fit a logic model to it that says how the different components influencing each other in a logic way.

Or, you could say, well, I know something. I have some prior knowledge. I may have interact ohms the say what molecular components are present in signaling networks.

And so in principle, I kind of know who's involved and who might be influencing whom. And I could write a logic model based on that prior knowledge. And then run calculations and see if it actually makes predictions about experimental data.

So that's one nice thing. It's a mathematical formalism that can either be run in data driven mode or in theory driven mode and go back and forth. So that's one reason-- given one lecture to offer, I've decided to offer it on this topic.

All right, with me so far? Any questions? Philosophy? OK.

So, what we're going to do today is almost take a hybrid of these two. We're going to say, what prior knowledge do we have, and then recognize that it's really not enough. And so how do we now integrate that with empirical data to now come up with logic modeling that, in fact, is actionable and computable?

OK. So what kind of prior knowledge do we have? Let's say we wanted to have a logic model for what's in these signaling networks down stream of growth factor receptors, or hormone receptors, or things like that, that then govern gene expression, metabolism and so forth.

What prior knowledge do we have? And you folks probably have already seen some of this in the class. There's all kinds of databases of stuff. What's in those databases that might be relevant here?

AUDIENCE: [INAUDIBLE] that the protein-protein interactions-- if you switch proteins, they interact with each other, but maybe not necessarily what pathways they're in.

DOUG LAUFFENBURGER: OK, good. And have you seen databases like that?

AUDIENCE: Several of them have come up.

DOUG LAUFFENBURGER: OK. Are you the only one who's seen them? Or is there anybody else that kind of noticed them in passing too? OK, good. Second, third, fourth, all right. That's a critical mass if I ever saw one.

OK. So, I'm just going to allude to those. So there are pathway databases. And this is actually an old slide of a few years ago, so I'm sure the numbers are all different. And, in fact, there's new ones. I just haven't taken to updating the slide.

But we'll, based on literature, take certain numbers of gene products, a few hundred of them, and organize them into pathways based on biological knowledge.

There's other databases that are more interactomes, usually based on other kinds of experimental data-- yeast II hybrid, mass spectrometry, literature curation, that also tries to say who's physically interacting.

So these node-- these pathway databases don't necessarily say, somebody's physically interacting, they say somebody might be upstream and downstream and so forth.

And then they interactome databases say, component a and component b, there's some evidence that they have a physical association someplace along the way.

So these are two complementary types of databases that, in fact, can be put together.

OK. So an interesting thing about these-- there's a number of these databases. And so in principle you could say, well if I want then to start-- if I want to generate a logic model for signaling networks, all I have to do is take what's in the database and say what pathways are there and what's known with their interactions, and now I've got a starting point. You know, I can actually draw a graph with lots of molecular nodes and lots of molecular interactions.

So, you can do that. And so you can choose one of these databases and say I'm going to draw a graph that has what's believed to be true about nodes and pathways and interactions and signaling networks. But then you choose a different database and another database. And you'll actually get different information.

OK. We actually did a study on this-- I probably should have given you the citation of that-- that said if you look at six or seven of these databases, they are not coincident. They have a very small intersections. Most of their information is non-redundant.

And so you could try to put it all together. And we did this, again, in this paper that I'm not giving you a citation for. And so here's a number of nodes and signaling pathways downstream of receptors.

And all the colored nodes are those in which they appear in only one of these-- one, two three, four, five, six databases. So if something's colored green, it's only in GeneGo and it's not in any of the others. If something's colored purple, it's in PANTHER and none of the others. OK.

If they're gray-- some of these gray ones, they're in at least two. But out of these six, there's an exceedingly small number of nodes interactions that are in all six databases. Which was a real surprise to us when we did this.

So what this means is, if you want to start with some prior knowledge graph that you're now going to fit a logic model to by mapping it against data, you first even have the choice, well, what am I going to start with?

What is my prior knowledge? There;s not really consensus prior knowledge. So you can start with six different interaction graphs. Or you could try to put them all together and get a consensus graph.

So you have all these choices. And right now, it's not as if is there's detailed analysis of what the best choice would be for your starting point.

But I want to stress that, with respect to our approach, this is a starting point because one of the issues with the database information is that it's typically very diverse with respect to contact.

What cell type did this information come from? What treatment conditions did it come from? If there's different cell types, different species, different mutations.

So if I see interactions or if I don't see interactions, are they in conflict? Or they're just-- this one was in a lymphocyte, this one was in a hypatocye, this one was in a cardiac myocyte, and they're actually different.

OK, so if I had a cell type specific database, or pulled that information out, that would be good. It would be a smaller number of things. But then under what treatment conditions?

Because remember I said starting with the genomic content, what you actually see in terms of molecular interactions will be very strongly affected by what matrix were the cells growing on? Or was this in vivo? Was this in a multicellular culture situation?

So, that's why this is a starting point and can't really be used to describe any particular experimental situation with much confidence.

The other thing-- and this is what I've been trying to emphasize from the start-- is that there's no calculation you can do on this. There's a group of folks in this field who propose some ideas that I think are very intriguing, but which, at least to me personally, there's not that much evidence for.

And that is, that there's topological characteristics of these graphs, that then tell you what's important. So if I have a node that's somehow connected to more other nodes, that is going to be a more important node, and might be associated with the disease, versus a node that's connected to fewer.

OK. Some of these are very, very appealing ideas conceptually. If you actually look for the experimental evidence that they're valid notions, it's very thin.

But, that's where some folks would claim, oh, you can do predictions on the hypotheses based on these graphs because there are these graph theory characteristics that somehow might be biologically meaningful. OK. But I'd say, jury's out on whether, in fact, any of that is true.

So, our view is-- OK, this is a good starting point, but in fact, needs to be mapped to empirical data in order to gain confidence about calculations you can do.

So that's the goal of this kind of approach, is to say, let's stipulate that we start with some prior knowledge scaffold. This particular one is from the Ingenuity database. You could get one from any other database.

You could get a consensus one from three or four if you want. And so it has, up here, extracellular stimuli, growth factors, cytokines. They're connected in their interactome II receptors. They're connected to scaffolding proteins and signaling proteins and kinases and so forth. They're connected to transcription factors, metabolic enzymes. So you can draw this graph. Say this might be what's going on in my cell.

And then what we'd like to do is to turn this into a formal logic framework that's capable of then fitting experimental data, predicting new experimental data, and giving you a chance at biological hypothesis and testing. All right, so conceptually you get it? Two aspects-- some kind of starting prior knowledge, that's kind of a scaffold, a graph, for your network. And now you're going to turn it into a computable logic model by mapping it against empirical data.

So, merely what it takes is the kind of conceptual diagram you see in any cell biology paper, any signaling paper, that says, well, a and b both influence e positively, and b influences f negatively, and c influences f positively. Then there's a feedback from g to a. That's inhibitory. You can draw those. But now, how do you turn it into a computable algorithm?

So, what I'm going to spend most of the day on is, just conversion of this to a Boolean logic model that any one of these interactions is and-- a and b being active makes e active. c being active, but b not being active, allows f to be active, and so forth. You turn these into formal logic statements that you can compute on.

At the very end, if we have time, I'll show how to relax this from a Boolean framework that's just on off, to something that can be more quantitative.

All right. So that's the notion. Now what I'm going to do for the rest of the time is go through the specific example paper that says, OK, how do we in fact do this? What is a way to accomplish this?

So now let's go back to a biological problem where there's going to be empirical, experimental data that we're now going to map against one of these prior knowledge interactome graphs.

This particular study-- this was done with Peter Sorger, who's now at Harvard Medical School-- had to do with liver cells. Liver cancer-- you'll see some application of that at the end-- that says we have liver cell hepatocytes.

And we want to know how they respond to different growth factors, in cytokines in their environment. How that'll change their proliferation or death? Or the inflammatory cytokines that they produce. And we'd like to take-- this is just a pictorial diagram that could be in any cell biology paper, and make this calculable.

So we could say what's different from a primary normal hepatocyte liver cell that's not cancerous? It might have a signaling logic. But if then we compare the signaling logic to a liver tumor cell type, or four different liver tumor cell types, what's different?

If we can find some logic that's different for the tumor cell lines versus the normal primary lines-- some logic from here to there or to there-- that now tells you biologically, where the differences might be that have arisen from the genetic mutations.

And where good drug targets might be, or predictions if I intervene here, if there's no difference in that logic, between normal and tumor, well then that won't have any effect. I want to look for the places where there is a difference in the signaling logic. And that would be a better drug target.

OK, so the measurements are made in across 17 of these different signaling molecules here, pretty much all by measurement of a phosphorylation state. So if you've done cell biology or biochemistry-- in these signaling pathways, many of the activities in these kinds of pathways that regulate this kind of cell behavior are kinases that end up affecting transcription factor activities and so forth.

And it's the phosphorylation state of any these proteins that matters. If a phosphate is on some particular amino acid, the enzyme might be active. If it's not there it might be inactive and so forth. So, just measurement of phosphorylation states of 17 different proteins in these pathways distributed across multiple pathways.

I've made these measurements on five different cell types, four tumor cell types, and the primaries in order to try to see what's different between primary and tumor. And then what might be different, patient to patient.

In response to seven different extracellular stimuli, some of them growth factors, some of them cytokines, some of them actually bacterial metabolic products. We all know about the effects of microbiome these days.

And, to further populate a database that might be capable of helping validate a model, a number of seven, in fact-- intercellular inhibitors. A small molecule, these things in black. One that might inhibit this kinase. One might inhibit that kinase. One might inhibit that kinase.

So now if you add all those inhibitors too, then you start to change the network activities and the downstream behavior. So that's how extensive the data is. And this is actually for a few different time points.

So the data looks something like this. Let's focus on the one on the left. This is just the primary, normal, human cells. It came from a liver donor. OK.

Each row is one of the 17 different signals, essentially measurement of the phosphorylation state of Akt or CREB or P52 of staph 3. OK?

So measurements of its phosphorylation state that has something to do with its signaling activity. Each of the big columns are the seven different treatments-- the different growth factors and cytokines and so forth. And the control. No stimulation.

And within each one of these treatments, in each one of these stimuli, then there's seven different inhibitors that were used for the different pathways. So seven stimuli by seven inhibitors plus controls. And then three different time points. Sort of zero, 30 minutes, and three hours.

So the data looks something like this. If there's really no change, due to the stimulation or the inhibitor, you'll see something in gray. So in these gray bars, there was already phosphorylation of this transcription factor [INAUDIBLE] and it didn't really change under most treatments.

If it was yellow, what it meant was, whatever the treatment was, you got a quick activation of that signal and then it went away. If it's late-- purple, then it didn't happen in the first half hour, but it started to show up a few hours later. And if it's green it showed up in the first half hour and it stayed sustained. So that's what the color means. But this is the real experimental data.

And over here on the right is one of the tumor cell lines. And you can just see by inspection, it's different, right. The colors here are different from the colors there. All the same treatments, stimuli inhibitors. The colors are very different. You know, therefore that the signaling activities are very different. OK. Just by visual inspection.

OK. So what we're going to try to do is build a logic model for this. A logic model for this. Compare them and say, oh where are the key differences in how the signaling pathways are getting activated? Downstream of the same stimuli.

So, we start with our prior knowledge. This is from the Ingenuity database, which actually happened to be missing, even basic information about insulin signaling. So we just added our own information about what the insulin receptor does. It's kind of hard to believe. This is a database that cost a lot of money and they didn't have really much information about insulin receptor signaling. Very strange.

So, downstream of our seven stimuli, down to the transcription factors of interest, there are about 82 molecular nodes and a hundred some edges that you'd pull out of the Ingenuity database. So here's our starting guess at what this looks like. There's no logic in here, but this is just, potentially, the things that the logic might operate on, downstream of stimuli, and when inhibited, and so forth.

All right. So here's the process. This was the actual algorithmic process that I'll walk you through. On the left-hand side is the computer part.

It said, OK, from the Ingenuity database, we had this prior knowledge about who was upstream, downstream, who affected whom. We strip this down some, because in terms of the measurements on the perturbations, there are some of the nodes that you just would not be able to see any measurable difference.

OK, there was no stimulus upstream, or no perturbation. And it was not measured so you really wouldn't be able to tell if it changed or not. So you just take those out.

Of everything remaining, now you don't know the logic. You know the potential. And so you say, well, of all the nodes and interactions remaining, I could have AND gates, I could have OR gates, you could have NOTS.

And you say, OK, in principle, I could have, then, many, many, many, many, many different logic models that could work. So how do I know which one? Well now you skip over to the other side and say, well, but we have all this experimental data. We have the data from all the different stimuli and all the different inhibitors for any given cell type.

And so, we have that data under all these different conditions. And what we're going to do is just run hundreds or thousands of these potentially appropriate models. Compare them to the data of whether any given node is activated or not, activated under treatment conditions, stimuli inhibitors. And we'll calculate the air. How good was any one of those models at actually matching those data? Simple as that.

And then it's a matter of finding what are the best fit ones from the best fit ones. Could you improve them and make them fit even better? And in the end, how did you go from an initial prior knowledge scaffold to something that, in fact, fit the data really well, from which you could make new predictions. OK. So you get the approach here? All right, good.

Now, in terms of figuring out how well any given model matches the data and how to go through model selection, there's a myriad of different approaches to this. And I'm not claiming that what we did was the absolute best approach. There's alternatives to it that one could consider and then perhaps could work even better. If you read the paper, you'll read the reasons for these choices. OK. So I'll let you do that.

The way the model quality was calculated was to have an objective function that said we want to minimize some number, theta. And how do we calculate theta? Well, first of all, for whatever that model is, we're going to fit-- whether the model says some nodes should be on or off, one or zero. And we're going to compare it to the experimental data.

Now the experimental data, I need to emphasize, isn't one or zero, it's normalized to go between one and zero. But the actual measurement might be 0.7 or 0.25. OK, so you're going to have error against the Boolean model even if all the edges are absolutely correct you'll still going to get some quantitative error.

So you calculate that. The Boolean model says zero or one. The experimental data says 0.250, 0.7. And you say, OK, I'll calculate that. But then you might think, all right, well, somehow I've got to penalize bigger models with more nodes and more edges because surely the more nodes and edges I put in, I could capture more of the data. And I don't want to make the model infinitely large just to get the best fit. So I need to penalize that. Turns out it's not true, but nonetheless it's worth doing.

So, you take a parameter that's the size of the model. It's basically just the number of nodes. The more nodes in it, the more you would be suspicious of the model for just fitting because it has too many components. And you multiply that size by a penalty parameter, alpha.

So you have a bad objective function if there's a lot of error with the data, or if your model's too big. A better model would be, better fit to the data and smaller. That's the calculation. OK.

And in the end-- and I'm going to show you how we did this. And I think the field is now really believing this. That what you're not after is a single best fit model. That one single model that gives you the very smallest data. Because honestly, within the uncertainty of the experimental data-- OK, there's a substantial number of models that could fit the data within that noise.

So if you demanded the single best one, you say, well, but these other 50 actually fit it almost as good and within the uncertainty of the data. How can you really reject them? And you can't.

So in the end, what's being striven for in most of the field is a family of models. And then you see what the consensus is and the differences within that family. The particular algorithm for generating and running through different potential models-- because you just can't exhaustively sample all of them.

OK, these networks are so large, that you can't exhaustively test all possibilities of all their logic and so forth. It's really prohibitive.

So there's many different ways you can go about it. This particular method maybe you've already learned this in class for other applications as a genetic algorithm.

So you start with some population. You start with your Ingenuity scaffold and then you randomly remove or take edges and things like that. So that if you've got a whole family, that's slightly different.

For each one of them you evaluate the objective function against the data. And you get some of those that then are the most attractive. They seem to be the best fit.

But, by no means would you imagine they are yet optimal.

So, now you create a next generation from this population by the analog of genetics. Some of the very best-- you say, OK, they're going to survive so I'm just going to take them as is.

Some I'm going to mutate, I'm going to have a probability of mutating an edge here or there. You can have crossover, actually mating between one model and another model, so that the daughter model gets some of the arcs from the mother model and some of the arcs from the father model.

So you just generate an ex-population, do it again. And once you've reached a set of models that fit your data within the criteria that you want, then you say, this is now my population.

And these are now my best-fit models. So it's not exhaustive. You can definitely find local minimum here. There's no question about that. Yeah?

AUDIENCE: Do you always take the best model into the next round? Or do you--

DOUG LAUFFENBURGER: Yeah, that's the elite survival. If you don't incorporate that, you might lose the best ones in any given round. But this ensures you take the best subset. Let them go for it.

AUDIENCE: Is there a worry that you might get stuck in [INAUDIBLE]?

DOUG LAUFFENBURGER: Yes. Yes, absolutely. So now you run this with a number of different starting populations. And you see if you get to similar consensus models. Yeah, because absolutely, this does not guarantee any kind of a global minimum. You will always get local. So you have to condition it on a different set of initial populations.

OK. Once you do this-- I'm going to show you some results first and then dig into some other ways to think about it.

So it's plotted here. This is one of the tumor cell lines. What's plotted here, is again, all the rows or all the signals that were measured. All the big columns or all the different stimuli, and all the little columns are the different inhibitors. And I should point out, this is only for the 30 minute data. OK. This isn't for the three hour or both, this is just the 30 minute data.

And basically where there's green, the model and data fit was considered OK. Where it's red, it's not OK. Where it's pink it's less bad. So by the shaded. And the yellow actually, the model really couldn't make a prediction.

Now, why that's the case is what's showing up here is just the initial Ingenuity scaffold. The very best one that didn't add or remove any arcs or nodes from the Ingenuity prior knowledge.

It's that all we're going to do is just run the best fit Boolean logic model we can on that. And it wasn't very good. It was about 45% error. Almost half of the nodes it got wrong.

So what that tells you if you just take a scaffold from one is interactive databases and without adulterating it, just fit the best logic model to some data-- OK, at least in this case, and we've done a number of others, it actually doesn't fit very well.

And the reasons being, you're trying to fit this now to a very specific biological context. Hepatocyte tumor cells under these grow factor and cytokine treatments. That network is likely very different from whatever aggregate you got from literature curation and so forth in a database.

There's going to be a lot of stuff in the database that's not applicable, because it came from a different cell type, a different condition, or there just wasn't enough experiments in the literature for hypatocytes.

Maybe it was never measured under treatment with interferon gamma. So there's data here that the database never had access to literature that it had explored. So lots of reasons.

Now when you go through the processes we just talked about, and in the end, the best fit models give you something like less than 10% error. So less than 10% of these squares are red or pink.

OK, so that's the kind of improvement that you can take by generating an improved model. By adding and subtracting arcs and nodes.

So this is what the model looks like in the end for this tumor cell line. And this is a consensus model from the 20 or so best fit.

And so the thickness of a line is how strong the consensus was. The strongest would be all 20 had it. And the point here being, you see some purple.

And I wish my pen wasn't fading in and out. If anybody has a pointer I'll be happy to have it. Where you see purple, those were arcs that weren't in the Ingenuity database and had to be put in to get the data to fit this well.

And it turns out, if you actually go back to the literature, you find that those purple arcs were already described in the literature. It's just that they weren't captured in that database.

Well that's green and purple. Then you see some blue and they were in some of the other tumor cell types but now in this particular hep G2 But you can generate a model that works very well.

And see that it's consistent with much of literature. It's a more stripped down than what's in the databases. And there's some new things in it, that in fact, if you go back to the literature you can find, because they just were captured in the database.

All right. A few insights about the analysis. So I want to show you, here is the objective function. How well the model fit and that's in red. OK. And in blue is the actual fit to the experimental data. And again, the hirer it is the worse it is. And the green gives you essentially the size. And this is plotted against the size penalty.

And what's very interesting, is even for very small size penalties, almost negligible, that the size of the model that turns out to be best fit is substantially smaller than what was in the database.

OK, so you actually generate a small model immediately. A smaller model immediately, even without any size penalty. So your intuition that a bigger model was going to be better actually turns out to be incorrect. That even without a size penalty, the model strips down.

And why is that? Why is that? Let me make that question number two. Just to see who's still awake. Why, in fitting this hepatocyte data, would a model that leaves out a lot of stuff in the Ingenuity database that's presumably going on actually fit the data better? A smaller model fits better? Why is that? Yeah.

AUDIENCE: This is a [INAUDIBLE]. Maybe the strength of the attractions aren't really taken into account here? And so the moving things out, essentially means that you're not sealing everything in the [INAUDIBLE]. You're just taking one.

DOUG LAUFFENBURGER: Yeah. That's essentially it. I think you've casted it an almost quantitative term, but I think it's true even in qualitative terms. And one way to think about it is-- let's say I have an extra arc or extra node. OK.

I might capture some more true positives. I might actually capture more of my data, but I could actually now, gain more complex with my data. Because now I've put in logic that, yes, it captures this measurement, but now maybe it messes up these other two or three measurements.

So you actually can make your model worse trying to capture some small piece, that in fact, adversely influences the effects on the other measurements.

So you get you get false positives, false negatives, along with anything and that's true. And it just so happens that in these kind of situations those can outweigh.

Then of course, as you increase the size penalty you can drive your model to be even smaller, fewer arcs, and now that of course does come at the expense of not fitting the data better. OK.

So where we decided that the size penalty best lived was where it was large enough to ensure stripping down of nonessential nodes and arcs, but not large enough to start compromising the actual experimental fit. OK. And so that lived someplace around there.

OK. An important thing-- and this goes back to the consensus model. If you think about, quote, model identification, can you uniquely specify one model a best fit model? You really can't. What's plotted here is for any of the arcs that would end up in a model.

Let's say we let's say we numbered them from one to I think it was 113 in the first place. One arc, another arc, another arc, another arc. And you say, how frequently did they end up in the best fit models?

Basically, only a small proportion of them were in all the best fit models. Some of them were in some models and some not.

Of course the higher the tolerance, the more air you allowed and now you started to get models that all fit to within whatever that criteria was in which most of their arcs weren't the same. You could have a lot of different network structures that give you that same fit. If you require a very, very tiny fit, compared to air, something like this, then more of the arcs in the models have to be in common. OK. So that makes some sense. But you can't really completely identify a unique model. That goes to what I said before.

OK. I was talking before about trade-offs between false positives and false negatives. You must know, I'm sure from previous things in this class, the receiver operating characteristic curves, where for every of your model parameter choices, you say, what are my results in terms of false positives versus true positives? And you're trying to find the optimal location along this type of path.

And so, what's shown here is that the best predictive model, in fact, is the one where we have the size penalty to be right on the edge of not making the experimental data fit worse, but still strips out the most arcs. So again, that demonstrates that the smaller model actually is in fact better, in terms of finding this type of--

And this shows if we actually put in some more arcs that tried to capture some more data, yes we decrease the false negatives, but, in fact, we increase the false positives. We actually shift ourselves on this curve. And so you decide whether that's desirable or not. Where you'd like to live. So you can analyze what you like about your best fit class of models in this kind of way.

OK, so now we have some confidence in this. What are you going to do with it? And one thing I'd like to do is just make a priori predictions. Say I now believe that on these hepatocytes or tumor cells stimulated with these kind of things, I can calculate what the experimental signaling activities should be.

All right. Let's see if we do that a priori. So let's now use new inhibitors that hadn't been used before. Combination of inhibitors, especially in cancer. People are always interested in combinatorial drugs. Experimentally it's prohibitive to run through all possible combinations. So this is one thing in the pharmaceutical field people believe these kind of models are really useful for. Let's try all possible drug combinations and see which ones are most promising.

And instead of just one ligand growth factor cytokine at a time, do different combinations. So this is all an entirely new data set. So different treatments that are different combinations, different inhibitors, different combinations of inhibitors.

And now you just run the model-- it's not trained on this. It was trained on the previous data. And now a priori predicts this data set. And now, again, you look for the model fit in the bottom. And again, you want the smallest number of red and pink boxes.

In effect it predicted to within about 11% error. About 11% of the boxes didn't fit well, but 89% percent did. And that's, in fact pretty close to the 9% that was on the original training model. So in terms of this, in this realm of studies, these a priori treatment conditions-- drug combinations, growth factors, cytokine combinations-- this is a pretty good validation that this model wasn't just kind of trained and fit.

That it, in fact, could predict then what was happening in these pathways.

And then of course, what it allows you to do, where all the red boxes are-- it say, OK, that's where we need more intensive study. Now maybe we go back to the literature and say, is there more known about those nodes that was captured in whatever our interactive database that we started with?

Maybe we need to supplement the scaffold with more information. That's out in the literature where more and more dedicated experiments are done. So it narrows down where the next set of investigations need to be, whether from the literature or from yourself.

OK. So this is just then some biological results. If you do this for the four different hepatocellular lines. Some of the signaling activities are the same, and some are different. I think I'll skip that.

All right, let me show this. So this says, where are the similarities and differences between the normal hepatocytes versus the tumor lines. Because this is where you would want to get the ideas for where the right drugs would be. Where is the logic different, between a normal liver cell and one of these transformed types.

So, this is the same kind of scaffold. It'll get us the consensus models and the thickness of the line is how strong-- what proportion of the models did that arc show up in? Along the best. If it's black, the arc was in the primary hepatocytes and all the cell lines. So black is just sort of consensus core. This is just invariably there.

The blue was in the models for the primary hepatocytes, but for some reason didn't exist in the tumor cell lines. So we're signaling logic that normal hepatocytes use, that the tumor cell lines have somehow lost.

Red, are arcs that weren't in the primary cells, but showed up in the tumor cell lines. So was logic that the normal liver cells apparently didn't use, but now showed up in the tumor cell lines.

And why would there be these differences? Well this is where it goes back to then the genetic mutations and variations. Because going from a primary to some tumor cell line, there's enough of the genetic mutations, that in this case said, OK, I've got some genetic mutation that interrupts the link between map three kinase and Ikk.

There was some docking protein or something that's now missing, not expressed as highly. It's got a mutation of amino acids and no longer docks right. It has a lower enzymatic activity.

So now you can go back and trace. Can I find some genetic mutation that has to do with the loss of that arc? Or if I've got a red arc that shows up-- like I said because there was something in my genetic mutations that now adds an activity here that wasn't there.

Maybe something is now constituently active. Maybe something is just expressed at a higher level. And all of a sudden that pathway comes into play. So that's the cool thing. You can trace what's actually in the genetic mutations if you have some methodology for that, to what's actually been altered in the network logic. Yeah?

AUDIENCE: Are the primary lines considered healthy lines? Or are they--

DOUG LAUFFENBURGER: Yes.

AUDIENCE: OK, so the--

DOUG LAUFFENBURGER: Yeah. So they're from donors but they're mainly like motorcycle accident donors that don't either liver anymore but the liver was fine. So, yeah, they're from healthy donors.

AUDIENCE: [INAUDIBLE].

DOUG LAUFFENBURGER: Yeah. Yeah. It was the lines at some point came from a tumor and have been propagated in a culture, yeah.

OK. What do I want to-- got a little bit more time. Let me do this. OK.

So here's another interesting thing that can happen. If you take these models seriously, it can tell you something about the biochemistry, perhaps of what's going on.

So see there's this dashed line here that I want to emphasize and we'll emphasize it again on another slide. That was one that had to be added. It just wasn't in the Ingenuity pathway, scaffold. Actually couldn't find it in any literature anywhere. But nonetheless you needed it to fit some data.

So we kind of kept our eye on that one. What the heck is going on here? This dashed line from I kappa kinase up to step three. No evidence for that signaling linkage in the literature anywhere. What could that tell you?

All right. Well, you go back to the data now and you say, well what of the data set, of the experimental measurements that we made, caused that arc to have to be there to fit the data well? OK. You can now ask that kind of question.

Well remember I said in the data set were inhibitors. Some small molecule inhibitors against this kinase or that kinase or that kinase that would perturb the network and then give us relationships at the logic model and had to account for.

Well, this one had to be there, mainly to account for data that came from an inhibitor of Ikk. That one of the kinases that we had a small molecule inhibitor for, inhabited this kinase. And somehow there turned out to be an effect on staph 3 phosphorylation. And so you needed that arc to be there.

So either the explanation that either there's, in fact, some real mechanism going on here. It might have been transcriptional that somehow the activity of this kinase affects the levels of expression and the responsiveness of staph 3.

Or you say, ah, maybe it's a problem with the drug? It's a problem with the inhibitor. That, in fact, what you thought was an inhibitor that just affected this kinase, has an off-target target effect on that kind of that kinase. And it's just an artifact. That's an alternative explanation.

Right, so that's the sort of thing you can test. And we did test it. And here's the data here. At the bottom is the kinase that you wanted the inhibition 2. And in the blue was the inhibitor that was actually used in the study, both in vivo and en vitro and it inhibited that kinase.

But then we looked at the potential off target effect on that other-- the JAK2 [? stat ?] 3 and it also did have activity on that. So it meant that that inhibitor had an effect, not just on the Ikk, but also on the JAK [? stat ?] 3.

And so that's why that arc had to be there, is because, in fact, that inhibitor, inhibited this kinase as well. So if we took that into account in terms of the algorithm, then we wouldn't have to have that arc because it was spurious and came from the arc, in fact, of that inhibitor.

But the interesting thing is that, by taking the model seriously, we can actually find that. Because it was not previously known that this inhibitor had an off-target effect on that kinase.

In effect, the interesting thing, pharmacologically, was that this small molecule that was aimed to be an inhibitor against this kinase was the best by far in treating lung airway inflammation, compared against a whole other set of other types of inhibitors for the same kinase.

So now the reason might be is, it's better because it's also hitting this other kinase. That this off-target effect actually is therapeutically efficacious and in fact a combination of drugs against this kinase and the other kinase is what's required for the therapeutic benefit. So that's something that could be explored. And that's the sort of thing this model leads to.

OK. Let me end by digging into this difference a little bit. Because I said, you see these differences between primary hepatocytes and the tumor cell lines. And the model said, just from examining the data sets, that the logic is different. OK. Is there any validation for that?

Well, so let's go back and look at those differences with respect to literature. So if you just blow up that part of the model, there's eight edges that are strongly disparate between the primary, normal cell types and the tumor cells and they're all enumerated here. One, two, three, four, five, six, seven, eight.

And they're essentially in three different pathways. So what the model is telling you is that there's three different pathways that are substantially different between a normal liver cell and a liver tumor cell. OK.

So is there any evidence that this is really true? So let's look at one. On to this pathway that I've got differences. And you see blue here and red here.

It says that this particular signaling node in normal cells is activated by this pathway. In the tumors, that regulation is lost and that actually comes through another pathway.

And it turns this is consistent with literature that, in fact, in the tumor cells, you get a higher activity of this downstream node. And now I've lost my light again. This HSP27.

Even though it's over expressed, you get less activation because this pathway is less strongly activated in red than the blue pathway is. So if you went by gene expression, you'd think in the tumor cells, this is a higher activated pathway.

Turns out the logic is different, and you actually get less activation of it because it's coming from a different pathway. So that turns out to be true in the liver tumor literature.

Another one-- I find this one really interesting. That in normal liver cells, to activate this Ikk pathway-- that's a very important kinase pathway, governing the transcription factor of NF Kappa b. In a primary cell, I need this combined logic between a pathway downstream of insulin receptor and a pathway downstream of a cytokine. Only if both of those pathways are on, do I now turn this on.

In the tumor cells, that check is lost. Only one pathway is required. OK. If this one is activated, I'm going to get this transcription factor activated. I don't have to wait for simultaneous activation of this pathway. Where as a normal says I have to. OK.

That turns out to be true that in the liver cells, the progression is associated with a looser regulation of this transcription factor.

And one more. I won't go into too much detail, but again, you see reds and blues here. In the tumor cell lines, you've now got activities downstream of insulin. That's normally just a survival factor, that's just not found in the primary cells.

And that, in fact, is shown in the literature too, that insulin signaling shifts from metabolism to proliferation. It's mainly metabolic, stimulus. In the normal cells it turns into a proliferative stimulus in the tumor cells.

OK. So, what this says is, just by mapping this logic scaffold, the scaffold against empirical data, developing a logic model, you in fact can find loci of differences between the normal cell signaling logic and tumor cells signalling logic for which there's evidence in the literature, none of which was in the original databases.

Finally, I'm going I'm just going to say that it turns out in another study, what you could show is those three pathways that the model predicts are the differences between the liver tumor cells and the normal cells.

That in order to kill these liver tumor cells, you need inhibitors against all three pathways simultaneously. You actually need combination drugs of three different pathway inhibitors to kill these cells. And it's exactly the three pathways that the model predicted of the differences between the normals and the tumor cells.

OK. All right, so I will end here and then see if there's any more questions. Something that comes up a lot is-- there's discomfort with Boolean logic because of zero, one. It's off, on, and of course we know biology, biochemistry doesn't work that way.

And so there can be so many artifacts, so many places that you can get things wrong, because you're trying to fit a model where the measurement is supposed to be either zero or one, and you're comparing it against a measurement that might be 0.6.

Well, 0.6, is that closer to 1, is it closer to 0? Is there some normalization that would shift it from one to the other. And instead of being a correct fit, it's now an incorrect fit. So you can see the room for artifacts by mapping quantitative data against a qualitative model.

So, one thing done more recently is to admit that and say, well, let's say just relax this a bit. And instead of having step functions from off to on, that they're more graded. It's like an analog transfer function.

So what you've essentially done is add one more parameter to every node, to every gate. Because of Boolean logic, there's essentially one hidden parameter. That's where you shift from off to on, right? There's some location of the level of the signal that you've decided is 0 or 1. So there's some parameter that you shift from, saying it's off to on.

Well here now in this formalism there's that, but there's also then the slope of shifting from off to on. Is it still fairly steep? Is it really mild? Is it someplace in between? OK? And this can go with AND and OR gates too. Now, instead of just one dimension, one component being off to on or on to off, now you got AND and OR gates that have these slopes as well.

So what this means is you require more data to fit this-- we call it a constrained fuzzy logic model because you've got-- if I've got 50 nodes in my system, I've got 50 more parameters I've got to fit. OK, so that requires more data.

What's the benefit of it, is that your predictions now, in fact, can be quantitative. So you can go into the model and say here's a transcription factor CREB. I'm going to predict its phosphorylation state and its transcriptional activity, perhaps, based on the activities of two upstream kinases.

And so if I had had an inhibitor for one of these kinases or another, how much would I shift the phosphorylation of this transcription factor? And what you actually see is these gradual curves, that if I start to inhibit [INAUDIBLE], OK, it gradually changes the phosphorylation of CREB.

Or if I inhibit the activity of P38, it even more gradually effects the activity of CREB. So you can turn these into quantitative predictions of strong effects, weak effects. And again, look at drug combinations.

So that's the advantage of going to this more analog transfer function logic model. You can deal with quantification much better, but at the cost of requiring more data.

OK, I think I'll leave it here. It's about 3:15 and so if there's more questions we can take them about any aspect of this. Most of you have stayed awake, I think that's a good thing. OK. More questions?

AUDIENCE: [INAUDIBLE]. When you have the model for the template of the Ikk story. And then it seems like it may not be as easy to back out the original data that led to that specific mode.

For example, you showed that one arc was from this one treatment. But because if you trained the model the same as that and it's not deterministic, then what-- could you just add--

DOUG LAUFFENBURGER: I think that's a great question. So let's say there's new arcs that you add, that weren't in the original scaffold. I mean that's what you got the biggest questions from. If you delete one, you say, ah, it's easy to believe why you would delete one.

Any arc that you add to get a best fit, I think you've got to ask questions about. So in all those cases where there are arcs that were added that led to a better fit model, the first thing we did was go to the literature.

Say, OK, is there literature on some affect of this node to that node? And it's just that that literature wasn't curated into that database or something like that. And most of the time we could find it there. OK.

So then there were the cases, and this was the most prominent one, where from some added arc, we just couldn't find it in the literature. In this particular case, it was very easy to trace it to this particular effect of this inhibitor.

I would say there's no reason to believe that that's always going to be the case. I don't have another example to show you where it was harder. Everything else we actually found in the literature.

But you could imagine, having some new arc that you really couldn't find in the literature and there's no artifactual explanation for it. And now, how you trace it back to what the data was that might give you a more nuanced hint.

It's a great question. I don't really know how we'll do that. I think, we and other practitioners who use this, I'm sure we'll run into it at some point. That's a great challenge to be thinking about. Yeah?

AUDIENCE: I might have missed this earlier, but I was wondering, is this model actually able to incorporate the heterogeneity of a tumor, for example? Or the population heterogeneity?

DOUG LAUFFENBURGER: That's also a really interesting question. Let me try to show something here. Yeah. So two things.

One is, what's shown here-- this is the four different tumor cells that we did. And what's shown in color is the arcs for each one of them. So yellow, orange, brown, red. So some places you see all four of those colors there.

In some places only two or one. It says if I had four different tumor types, there's some slight differences in logic among them. You could translate to that is, well I could imagine then having a tumor that's a mixture of sub types and how would I discern that?

One possible idea that's attractive to me-- although we didn't really explore this in any form of [INAUDIBLE]. We didn't really have the means to make experimental measurements on the tumor heterogeneity at the time. It's when you get to a set of consensus models. Right.

So let's say you get the 50 best fit models and you say, some arc is in 80% of them, but it's not in 20% of them, is it possible that that represents some of heterogeneity because you're getting an average of different subtypes? I don't know.

Sometimes that appeals to me as potentially valid. Sometimes I think there's a flaw in that reasoning. Just because you get an average that's not then as strong. Does that necessarily reflect a sub-population? I don't know.

So what we do know is, we can see differences when there are differences. How you actually see them then, if all you have is averaged data, maybe it's reflected in the heterogeneity of the consensus models. Maybe not. It would be an interesting to explore.

Anybody else? All right. All set. Thanks.