The Bitter Lesson is Coming for Proteins - Alex Rives, BioHub

70 MIN · YOUTUBE · HTTPS://WWW.YOUTUBE.COM/WATCH?V=XDEVS0GSUIQ
https://www.youtube.com/watch?v=XdevS0GSuiQ

SUMMARY

ESMC is pioneering a novel approach to programmable biology by leveraging predictive modeling to design proteins and antibodies. This method utilizes vast databases of protein sequences and evolutionary patterns to enhance the understanding of protein structure and function, ultimately aiming to accelerate scientific discovery in biology and medicine.

- *ESMC employs a world modeling perspective to design proteins, including antibodies, by predicting molecular structures based on evolutionary data.*
- *The model builds on previous iterations, integrating billions of protein sequences and structures to enhance predictive capabilities.*
- *A significant breakthrough came from incorporating metagenomic data, which expanded the training dataset and improved model performance without diminishing returns.*
- *ESMC demonstrates that proteins can be designed effectively without relying on multiple sequence alignments, which are commonly used in other models like AlphaFold.*
- *The initiative aims to create a comprehensive atlas of protein biology, linking evolutionary data with functional insights to facilitate new discoveries.*
- *The future vision includes integrating experimental data with computational models to create a "lab-in-the-loop" system, enhancing the understanding of cellular biology.*
- *Biohub's commitment to open science and collaboration aims to democratize access to advanced tools and data for the scientific community.*
- *The overarching goal is to develop a digital representation of biology that can predict outcomes and guide experimental interventions, ultimately leading to breakthroughs in disease understanding and treatment.*

So ESMC is is also approaching programmable biology, but I would say in a very different way. It's approaching it from this kind of world modeling perspective where the idea is basically you have a predictive model and you know you're going to search the world model to find protein molecules that satisfy kind of whatever design criteria that you have. So we've been able to use this to actually now go and design um many protein binders. But I think sort of most excitingly, we've been able to use this to actually design antibodies, SCFVS.

>> Hello, welcome to the latent space AI for science podcast. I'm R.J. Haneki, CTO of Muromix. >> Yeah. And, uh, I'm Brandon today. It's a pleasure to have Alex Reeves, uh, head of science at Biohub. Yeah. Would you like to introduce yourself real quick? >> Yeah. Yeah. Thank you for having me here. It's great to be here. Um, I'm head of science at Biohub. I'm a computer scientist uh and I work on AI for biology and a lot of my work has been on language models for biology.

>> By the time this podcast is released, you will have put out several new exciting interesting models. Going over them, I couldn't help but have the kind of thought that you might be the most bitter lesson person in protein biology right now. Can you give a little context about what that means for biology and you know why you're so committed and excited to this route? >> Well, I'll take that. Um, I believe in scaling laws. So, you know, I guess I've been working on this for, you know, since since the summer of 2018. Um, and so my team when we were at Metaphair trained uh really the first transformer language model for protein biology. And so I guess you know I I've always thought that there would be kind of emergence of biological information as you train a model to predict the next token that evolution creates. So our team has really explored that idea over a number of different years and we've really kind of I think seen the scaling curve and really seen as we have have increased models by an order of magnitude kind of in each generation that you know there's this emergence of new capabilities.

>> Yeah. So you've been you say emergence of capabilities scaling over generations. You've been working at this as you said for I guess it would be 8 years now or something like that. It didn't always work that way right like there was signs that scaling might work. You know we'll be getting to some new results where I think really you've kind of clearly demonstrated this hypothesis in a way that hasn't happened before. But you seem to have like a strong commitment to this in a way that I'm not necessarily sure I would have been so convicted that it would work in the same way. I mean proteins are not the protein language is not the same thing as natural language. There are similarities but if you start sampling a transformer at you know a normal language transformer at temperature you're going to get gibberish. you sample a protein language model at infinite temperature, you're going to get something which is a valid protein if not a not interesting protein despite the fact that is a different domain for a different reason.

I'm not necessarily sure that I would I primarily assume the natural language model insight would transfer over. So what is specifically about proteins that you thought was special or you you know that would make this also valid? >> Yeah, I mean it's a really interesting question. I think kind of a deep question across AI right now more broadly and you know I think you know what's what's so interesting is AI right now is is such an empirical science and so we don't have you know theory that can always guide us in these things but we have this really strong empirical evidence of scaling the thing that I was motivated by is you know if you think about evolution and you know you think about the data that we we have around proteins we have databases that have billions of protein sequences. And you know, those those sequences contain patterns and you know it had had been long been known so that you know this is going back you know decades kind of before you know we started working on this with language models but that there are patterns the sequences of protein families that come there because of the constraints that evolution is operating under. So you can think about, you know, like a um a protein sequence that folds into a three-dimensional structure in space. And you can, you know, imagine that there are two residues or amino acids that are in this sequence that might be in contact in that folded structure. And so evolution isn't free to choose those independently from each other. If it makes a choice at at one position, it kind of has to make another choice that's going to be compatible at the next position. So going back, you know, all the way to the beginning of gene sequencing when people first began to be able to to look at this and kind of look at different related, you know, the same protein and related organisms, you could start to see these kind of patterns that are reflecting the fundamental underlying biology. So the idea behind ESM, kind of the thinking behind ESM was, okay, what if you were to apply this principle of across all of evolution, across

kind of the vast diversity of proteins that have been generated across all of life and, you know, basically have a language model kind of predict the amino acids that evolution will choose to place in proteins across all of those biological contexts. So you can think that there's just this this kind of like incredible amount of information in that total picture about the underlying biology of proteins. And so that was really the idea that sparked this is is you know as as a model is having to predict the next token and actually we train these models with mass language modeling. So they're predicting kind of tokens that are masked out of various parts of the sequence that it would have to learn something about those kind of underlying constraints that are shaping which tokens evolution can choose.

>> Yeah. So maybe for a bit of history um so you know you have you you just released um evolutionary scale modeling Cambrian, right? Is that what it's called? Yeah. And this is like the maybe fourth or fifth in a series of models. I think maybe even more if you go back before they were called ESM. >> Well, they they were called ESM from the start. Yeah. We had sort of various branches of the different models. Yeah.

So, so this one I would say is is kind of a a fourth generation model. Um it's actually a model that we trained a little over a year ago. Now that we're at Biohub, we're um we're we're open sourcing this this model fully under MIT license for the first time. So, we're really excited to do that. But kind of the the big thing that is new here is that we've really kind of built a world model of protein biology. So the foundation of that is ESMC. But you know using the representations of EFSMC, we've kind of now built a a structure prediction model. Um and this is the next generation ESM fold model. And then we've also used the techniques of of of mechanistic interpretability and sparse coding to really start to look deeply into the representation space of the language model and kind of be able to pull out the underlying features that the model actually uses to represent protein biology. So bringing all of this together, we're able to, you know, really make predictions for protein structure. um predictions about kind of the underlying features that that proteins are made out of that allows us to build linkages across evolution.

We're able to take this model and invert it to design proteins. And we've we've we've used this to kind of create a comprehensive picture of protein biology. So we we put together kind of all the world's largest protein sequence databases. And so that kind of amounts to 6.8 billion non-redundant proteins. And then we've we've resolved predicted structures for 1.1 billion of those. And and we've also computed features across all of those so that we can make these linkages basically all across um evolution and protein biology.

>> 6.8 billion of which you've resolved structure for 1.2 is that >> 1.1 >> 1.1. So what about the others? Well, so so basically what we did is we took that database and we clustered it at 70% sequence identity. So it's it's really resolving structures for everything in the sense that for each cluster we kind of have a cluster center. We're predicting the structure there and then we can expect that the other proteins are going to have a similar template structure. There be be small variations but they have the same fold.

>> 1.2 billion or so clusters >> that are that are kind of covering the 6.8 billion. Yeah. >> Okay. Interesting. And yeah, maybe since we're talking about scaling, how do you know that um this is the right number, right? Like uh how do you know that focusing on these 1.1 billion and that's the right resolution for this model? >> Well, we've

chosen them so that they really cover that entire space. So, I think what I can say about this database is it's really the most comprehensive picture of protein structure and function that's been created. It's adding, you know, hundreds of millions of structures to our knowledge of of kind of protein the diversity of protein structure and it's also creating this uh feature space that allows us to find these linkages between proteins across evolution. So we can see kind of really interesting themes emerging across evolution. you know linking for example um gene editing systems which are very far apart in sequence but you know they share some kind of underlying functional um patterns structural homology that the model's able to bring together and and find those connections now we're talking about the mechanistic interpretability part so you have if I understand correctly you use sparse autoenccoders and other techniques maybe to understand okay what are the when I activate the network using a protein Then what are the patterns of outputs that I'm seeing and how do they relate to each other if I understand correctly is that you have these sequences that are unrelated or only partly related based on the actual sequence but in terms of behavior they have similar behavior and therefore they are activating similar networks. Is that kind of the summary of what you just said?

>> Yeah. So I mean basically what we've done is we've trained sparse auto encoders across all the different layers of the ESMC model family. So there's actually three models in that family. There's a 300 million parameter model, a 600 million parameter model, and a 6 billion parameter model. And then we've looked really we've done kind of a very deep analysis of the feature space of that 6 billion parameter model, which is really the state-of-the-art protein language model. And so what we find what's really interesting is there's there's kind of this um you know this hierarchy of features that emerges.

What's really interesting about it is it's it really kind of corresponds to the reductive picture of biology that has been developed over you know many decades a century of of um biological experiments. But but what's so so cool is this is emerging you know without any prior knowledge. It's been learned by the language model. So the the interesting thing about SAEs, right, is they're really just revealing the intrinsic structure of the representation space. So this model's been trained on protein sequences. It's been trained just to predict the amino acids that evolution will choose. And then you know somehow this is leading to the emergence of this like kind of very ordered feature space that has this hierarchical structure where you can really see everything from the basic biochemical properties and kind of the basic structural building blocks of proteins to these very large kind of functional themes, these kind of abstract concepts that you know connect to to how kind of the human picture of of protein function. And do you have a hypothesis or feel for why if there are relationships between the sequences themselves even if they're like shifted and and cut up and recombined in different ways like I can imagine that might work because you have these you know proteins are kind of hierarchical in their nature as well. So maybe the hierarchy moves around but the same sequence but the functional units I guess those have related structures.

What is the hypothesis here? >> I mean it's it's a really interesting question, right? I think I think I can speculate about it. I don't think we we kind of completely understand this, right? But let me give a concrete example. So, you know, the nucleophilic elbow is this kind of like core kind of functional motif that people have thought that you know maybe this actually has emerged independently in evolution, you know, different times in different protein families. But it has this you know very very clear um structural motif that you can kind of see in a in a crystal structure. You know what we found basically is that the model has a kind of a single feature for this

nucleophilic elbow and it's activating across these like very evolutionarily diverse families. You know really completely different structural topologies proteins that probably evolve like entirely independently from each other. But you know the model is kind of using this one feature to represent that. So why does it do that? I mean I I think it's a really interesting question. I mean, I think one answer is sort of the idea of of compression and the idea that, you know, the model needs to have some kind of underlying latent variables that it develops to help solve this this kind of sequence prediction task because the nucleophilic elbow, you know, it's going to be a function of what's so what's so interesting, right, is the choice of any amino acid is kind of like completely entangled with with the choice of all the other amino acids in the sequence.

So this is a very complex task to try to predict what amino acids should be where in a protein. But to really do this well, you know, the model would start to have to have these kind of hidden variables that are representing the biology that allow it to, you know, look look at a protein and say, okay, what amino acids should be there in all these different contexts? So, you know, that that's sort of I think the intuition. I mean, I would draw the parallel to language modeling, right? And so I guess I was I was like very influenced by a paper by um uh Zelic Harris. It's called distributional structure from from 1954.

And I think that a paper that influenced a lot of people in the language modeling field as well, you know, but I I think it has so it focuses on on language and and it really articulates this idea that um the set of contexts in which a word appears are determined by the meaning of that word. And so what Zeli Harris kind of imagined is that you know it would be as you looked at the statistical patterns of you know what words appear in what context sets you would be able to derive the meaning of language you know you would you would have this kind of statistical structure that would mirror the underlying meaning of language. For me at least, that's one of the most convincing explanations for why, you know, a language model that's trained on the text of the internet is going to learn something about meaning.

It's going to learn something deeper and more fundamental. And so so I think you know you can think about the same thing in in biology where the contexts in which an amino acid can occur are really determined by you know the structure the function of the protein its biological roles you know these I mean very complex phenomenon u both the intrinsic biology of the protein and its relation to all of the other proteins and the function and evolution and so but but those are what determine the context sets and so you would imagine that then those statistical patterns in the use of amino acids, they directly reflect those underlying hidden variables and so the model is going to learn something about those hidden variables.

>> I definitely buy that seems plausible. Um maybe just I mean I in fact I I want to clear I actually do really believe in this direction but there are a lot of like ways I think about this where maybe I could say maybe I would imagine maybe it wouldn't work and one of them is like data availability right like what type of data do we normally uh have what type of sequence data do we normally get I think that ESMC in particular has some new data sources compared to like previous models which might be helpful but often times the type of sequences we have available have like a very strong bias towards certain specific needs for medicine or you know human biology or disease biology right so it's not necessarily that if you take just a naive data set you're going to necessarily get an interesting scaling law so I'm curious about like what in particular was the sort of

breakthrough in ESMC so I mean maybe we can go back a bit and talk about some of the other ESM you know predecessors which got here before ESMC and how like you know they were, you know, their strengths but also maybe some of the limitations that ESMC overcame and like what the developments there were.

>> Yeah. Well, I I'll admit that I am I am better less in film. I am scaling film and so >> I do think that I mean you know just just kind of increasing the data and increasing the parameters and having that compression is is going to just lead to more powerful models. But you know it is also true and I think you're you're absolutely right the structure the underlying kind of structure and distribution of the data is really critical and so you know some data sets will be far more valuable for kind of learning these these general principles than than others but I think it goes against a lot of biological intuitions about collecting data I guess is what I'd say because normally when you think about what data do you want you're trying to answer a very specific scientific hypothesis you want you know a very wellontrolled experiment you know, you really want multiple replicas, you know, it's it's it's something that's very focused, you know, is the way that I would put it. So, I think the change in the way of thinking is to think, okay, what you really want if you want to learn a general representation of proteins is you want to see amino acids in as many evolutionary contexts as possible.

That's really what you want. That's really kind of how I think about data. And I think if you what changed between ESM2, which was kind of the previous generation model, and ESMC, which is this new generation model, because they're both at the approximately the same scale. And you know, >> the same scale of compute. >> Same scale parameters. Yeah, SM2 got a lot of compute, but ESMC got even more compute. >> But it's not just the compute. The data was was really the critical thing here, actually. So when we trained ESM2 we observed two things. The first was that as we increased the number of parameters and compute you know we saw improvements. So we had a model kind of at the billion parameter scale. We had a model at the 10 billion parameter scale and you know the larger scale model is better than the smaller scale model. But if you kind of look at a plot that of of uh parameter scale, you know, sort of a log plot of parameter scale versus capability. And so for capability, you know, we're looking at kind of the representational fidelity. How well does it capture protein structure? You could see that there's there's kind of diminishing returns in ESM2. ESM2 is trained on unref.

And for ESMC, we added metagenomics. So we added billions more sequences to the training data. >> Could you explain what univer and uh metagenomics mean? >> Yeah. Yeah. So UNIRF is um I'd say sort of the the gold standard data set of sequence biology is kind of you know taking sequences from across a wide variety of different sequencing resources. It's clustering them you know to kind of remove some of this redundancy that that you were mentioning. and and so it kind of creates a definitive coverage of of of protein biology. What has happened in parallel to the classical gene sequencing is is this idea of metagenomic sequencing where people go out into, you know, all kinds of different biomes and environments and collect samples from the world and just kind of sequence the the natural diversity that's present there. So, you know, proteins from a hydro hydrothermal vent or proteins from a from a a frigid environment near the South Pole, you know, or or the deep ocean or, you know, soil or the human gut, you know, all all kinds of different environments. So, this is a very different way of collecting data. Instead of you are trying to understand a specific genome of a specific organism or trying to understand a specific protein, you just collect a bunch of stuff, mix it up in a pot, get the sequences out. You have no idea what organisms these are from. You don't necessarily even know if a given

sequence is a protein, but you can guess based upon certain contexts and you say, "Okay, we throw these together. These are likely protein sequences we find.

We're not assigning them to an organism. We're not assigning them to like a larger context. We're just saying this is probably a protein. Let's train on it." >> That is right. Yeah. And you you don't even get the full genomes. You just get these kind of contags that often are broken and have even partial proteins. So, the data is really noisy. One more like little nerdy question that I have here is that if I understand correctly, you're not actually looking at using a device that sequences proteins. You're sequencing the DNA that would manufacture those proteins. So, you're finding DNA and then looking for markers that indicate the beginning and end of a protein sequence. Is that kind of >> Yeah, that that's exactly right. Yeah.

Basically sequencing, you know, genetic sequences and then you translate the the proteins from those sequences. So you're digging up like sewers and not you but >> me personally there are sewers like probably many York City subway all kinds of things. >> Yeah. So the natural question to me is so you built this model and you think that you've kind of dduplicated it so that you have a good representational set without a lot of redundancy in it.

How much more is there? Like if we had an order of magnitude more resources, do you think that there is an order of magnitude more proteins to discover? >> I think so. I I I'm not entirely sure, but there are a lot of proteins and um I I think we've we've barely scratched the surface of measuring Earth's biodiversity. So there are core proteins that are conserved across all of life. So we we I think know that, right? But but as you go into these different environments, there's just, you know, new new genes and no new proteins constantly being created by evolution.

And this is a lot of my understanding is a lot of this is viruses and bacteria and other >> microorganisms or so those and these guys are in this basically longunning conflict with each other that causes them to >> recombine their DNA in ways that help them to survive in these extreme or whatever environment. And so that that's what's causing this incredible diversity of proteins.

>> That's right. Yeah. And just 4 billion years of of life running experiments in parallel all across the earth in all kinds of different ecological niches. And we just we see the outcome of all of that. And so the combinatorial that's why you believe that yeah there's there's going although that maybe from a macroscopic perspective when we look at it there's maybe not that not even nearly as much diversity as there will be at the microscopic scale because you have this incredible combinatorial effect.

>> Yeah. I mean there's just I think tremendous tremendous diversity there. So so kind of going back then there. >> Yeah. I know it's it's great, right? And I think it's really I mean we could also talk about data and building models of the cell and kind of really going from the molecular level to you know to to to higher levels of biological complexity but but to to to complete the the the description of of ESMC. So so that that's really the big change was kind of adding these metagenomic sequences and then you know what we saw basically is is is there ar are no longer diminishing returns to scale. So that's really saying that ESM2 was kind of data limited rather than compute limited for ESMC. There's a there's a really beautiful scaling law that we can plot where we

can look at um we can train models, you know, to make the larger models. We basically train models at the at the smaller scale and we can we can really look at the best representational fidelity that they can achieve for a given compute budget and just draw a line of extrapolation out that that kind of beautifully predicts what the larger scale models will will be able to achieve in their representational fidelity. So there's there's this really beautiful scaling and you know the really the only I mean there's some changes to to to SMC just to make it a more efficient model for for training but you know I I think the data is really you know really the big thing there that's that's driving that >> so it still is basically just a standard vanilla transformer a few tricks everyone has a few tricks at this point language model and just a lot of data >> so I mean this is very much in contrast to something like alpha fold right where you have a lot of inductive bias in built into the model in order to be able to >> predict protein structure.

>> That's right. And the idea here is you know really can we just learn the right structure you know don't give any priors just allow you know allow machine learning to figure out what that structure is. >> So you also had your own detour into priors with ESM3 right like or maybe not priors but uh using more intuition or more human design. Do you think ESM3 was a detour? Do you think there was I mean did you just end up saying like okay let's make C bigger and then suddenly it worked and now you learned that actually we don't need priors anymore. Is that like kind of a key insight or do you still think there's room for prior? I think we need both. I mean I think there's just you know there there's a place for both of them. So, you know, the the goal for ESM3 was to really make biology programmable. And so, we're trying to think, okay, like what is the programming language, right? How are you going to be able to allow biologists to to prompt a model and design structure and design function and all these things? And so, we really thought it needed the right tracks. But but I would say that ESM3 was like very consistent with the philosophy of ESM because you know what we did is we basically predicted structures for this vast array of evolutionarily diverse proteins and and we're using that as the training data. So the model's now just it's learning se from sequence patterns learning from structural patterns learning from functional patterns. But I think that same kind of synthesis that the model is learning on sequences, you could imagine that you know bringing in more multi-dimensional information would would build an even better representation space. If you are a coder or if you you're building uh language models and then building coding agents, you start with pre-training on everything and then you go to doing the programming part by some sort of post-raining probably RL. I mean, have you thought about post- training uh ESMC to try give you the same abilities for programmability? Do you think you could get programmability without doing all of the inductive biases which involve like atlas of structures and you know which mostly dist some sort of interesting distillation but uh I guess maybe that isn't some kind of prostraining um of a different model. Yeah.

>> Yeah. I mean I think it's a really interesting question kind of to what degree can you interconvert these models. I don't think kind of that's fully understood yet, but I I think that's it's a very kind of promising direction to think about um doing that and what are the right ways to do that. >> So, ESMC is is also approaching programmable biology, but I would say in a very different way. It's approaching it from this kind of world modeling perspective where the idea is basically you have a predictive model and you know you're going to search the world model to find protein molecules that uh that satisfy kind of whatever design criteria that you have. So, we've been able to use this to actually now go and design um mini protein binders, but I think sort of most excitingly, we've been able to use this to actually design antibodies, SCFVS, and we're seeing really, I think, exciting uh success rates in a small number of trials now.

>> Yeah. So, can can you explain what those SCFC's or are? >> Yeah. Yeah. So, an SCF is basically um it's it's a single chain antibbody. So it's it's um a kind of therapeutic modality that basically has so an antibbody has a heavy chain and a light chain and then it basically has a pair of of you know one heavy chain one light chain another heavy chain and one light chain that that come together to recognize a target. So there's different variations of of these kind of modalities that are used therapeutically. And so what's interesting about the SCF is it has one heavy chain and one light chain. So it is able to kind of form these very complex binding interfaces where you know you can kind of have two different subunits coming together to engage a target. These are kind of important therapeutic modality um something like I think a quarter of of of new drugs are are antibodies. So it's really I think you know one one of the one of the critical um modalities for for medicine.

And basically what we're able to see is that you know you can search ESMC and you can actually find um antibodies that are reaching the level of affinity that I should say or really at the level of affinity that is needed for therapeutic function and activity. >> The protein design space has kind of exploded in the last 5 years. You know everyone is doing protein design pretty you know many people are excited about protein design. uh my kind of high level naive understanding of the field is that things like mini binders um are are quite doable. people have done that you know quite routinely successfully you know in smaller by the time you get to like nanobodies and SCFs they're a little bit harder to design um and then antibodies are still actually quite out of reach often times one of the common reasons you know for this is if you're in the alpha fold paradigm you don't have MSAs right the evolutionary pressure for antibodies is actually the opposite in many ways of what the evolutionary pressure is for everything else they go for diversity rather and trying to be go evolve along a very like constrained path. So I'm curious, did you try larger structures and is that something that you've seen success on or is this something that you still think for some reason it might be hard to do?

>> We can actually take the um SCFVS and reformat them as as antibodies. So I think that would be kind of the quickest approach to do that. Um we've not tried full IGGs. I I don't see any reason why that wouldn't work. Actually, it's something we haven't yet. You know, we've decided we're basically kind of releasing this now because we feel like it's it's kind of reached a point where, you know, we're seeing, I think, a really significant step above kind of what's been possible in the past. And so, we just we wanted to get it out there, but you know, I think there's a lot more progress that's possible. So, we're, you know, we have a lot of collaborations to kind of look at some of the other applications here. You know, the thing about it, right, is it's a general model. So I I think to me that's the most exciting thing about it is just you know a general model for protein sequence structure and function.

You can search it and you know therapeutic design basically emerges from that search. >> Yeah. I mean to me the the you mentioned that you're not using MSA's multi-seas alignments which was one of the or maybe the critical insight that allowed Alpha Fold to work really well and the fact that you didn't need that in order to make it work basically as well as Alpha Fold 3 is really exciting to me because that means that your thesis of let's let's cover the space of possible proteins and as well as we can and see what the emergent behaviors are so that if this is an emergent behavior that we're kind of able to replicate what happens with multisequence align when we have use multi-sequence alignment what are the other things that maybe we don't have data for but that we are able to also do in an emergent way >> I would say actually you know we're we're doing significantly better on on antibodies so I think I think that's one of the things that's really cool that's one of the thesis that that we

had is you know antibodies are not going to benefit from um evolutionary information probably in the same way that kind of predicting the structural topology of of a molecule will so you know I think I think you kind of see that now where the the representation space is containing something that's really interesting about antibodies here >> I want to talk about cuz you mentioned something very interesting to me which was talking about virtual cell and how this maybe interfaces or did this work here I'm really interested to know were you able to find other things in your mechanistic interpretability. What were some interesting things that weren't just validating biology, but there's a pattern that was unexpected. Did you find anything like that?

>> It's complicated. So, because we have to we have to now actually go and validate some of these things, right? So, I think what we saw are like interesting connections, right? So um you know what we can see for example is that kind of distantly evolutionarily related gene editing systems cluster together in this space in ways that are consistent with and kind of reflect our knowledge of the origin of those gene editing systems. So that's really exciting. But but kind of the thing is right there's a number of proteins that are in that map that are kind of brought together in different ways where we just we don't know what they are right now. We don't know what they do. So one hypothesis there is well these are kind of novel gene editing systems. I think in this atlas you know there's there's going to be some some really interesting basis for scientific discovery there. And if you think about kind of how people go out and look for new gene editing systems, for example, they're typically mining the large genetic sequence databases and they're looking for kind of different sequence patterns or structural patterns that are linked to that. Actually, the first version of the ESM atlas was was used by Funang's group to find um a new gene editing system. So, I think there's just a lot of biology out there that we don't understand that's waiting to be discovered and kind of being able to connect the dots between proteins so that we can go from, you know, what it is that we we know today to to kind of make those inferences about the unknown.

So, that that's what I'm excited about. And I I think, you know, there proteins for for so many applications that nature has probably invented. You know, you think about the thestable polymerase which enables PCR which came from a bacteria living in a thermal hot pool. You know, you have there may be the solution to to climate change, you know, somewhere in in protein biology. There are probably all kinds of building blocks for for completely green chemistry infrastructure out there.

There's probably new medicines and therapies, you know, but the question is how do you find those? And so I think, you know, kind of being able to connect the dots is is is really one way to to start to be able to open up that space of protein biology to discovery. >> I'm curious, you've uh one of the advancements of ESMC is a improvement in multimer. So basically protein protein interactions like um that structure predict the ability to predict the way two proteins interact. I think you now claim to do better than anyone else, right? If I correct me wrong.

>> Yeah. I mean I think we're state of the state of the art for open models. >> Okay. One thing which I know some people would find very useful for virtual cell is just an entire mapping of every single pair of proteins inside the human transcriptto. Have you thought about doing this in terms of um like kind of a beginning to a virtual cell like create that map. >> So I think something like that would be really valuable. I think you know

fast.

So the other thing about ESM fold 2 is a really fast model because it doesn't require the multiple sequence alignment. So you know you can do inference kind of you know directly from the sequence. Um it takes seconds you know you can get an atomic resolution prediction. So yeah that's I think one really interesting application at at biohub. I mean the other thing that we're thinking about is can we actually experimentally resolve this and so one of the things that we are we are building is cryeleron tomography and and we're really building systems that can greatly increase the contrast when you're looking at you know at the cell at the atomic level and so I I think one thing that I I hope to see is actually structurally empirically resolved interact at some point in the future and I think there are some some pretty big technical hurdles and technologies that have to be developed to to overcome that but I think that's something that that's going to be possible so we we can use computational methods to start to get the proxy of that and I think you know that's going to be really powerful but I think a lot of the future of structure prediction is going to turn into structure determination actually you know really bringing together these kind of tools that we have for modeling proteins and bringing them together with experimental data so that we can start to you know develop this picture that's you know informed by empirical biology by by what we can observe.

>> So is that the vision here if I'm understanding correctly is that you have maybe lab in the loop kind of thing where you have an agent that is talking to your you know C7 and whatever and then it predicts a property that you're interested in. It sequences the the the genome or it creates the genome. It creates the protein from the genome. It then observes it with some version of this uh microscope. What did you call the microscope again?

>> As a cryo electron tomography. >> Okay. Okay. And then you do whatever experiments or you observe it and then you use this as a lab in the loop to oh okay this folds this way. Therefore, I want to check the next one that I want to check is actually a different one and use active learning system. Is that sort of the vision that you're articulating here? >> Well, I I think they're going to be they're going to be a few fundamental principles for the next era of of biology. I think I think it's yeah, it's such an interesting time right now because I think we're really we're at the beginning of a new scientific paradigm. It's really just the beginning of it. And so, what is you know what is defining in that paradigm, right? And so I think there are there are a few principal data generation. I think that's going to be really critical. The second is computational, you know, predictive digital representations of biology. And we can kind of talk about, you know, you can think of ESM as as being, you know, first generation, Alphaold as being a first generation of those kinds of approaches. And so you can kind of start to think about what does that look like as we can model more and more biological complexity in that way. And then you have the principle of feedback and you have the principle of you know we have intelligence now that's scalable and so can be applied to every unit of a biological problem. What would it mean for all of that to come together? So, I think we're gonna we're going to have increasingly capable and accurate digital representations of molecules, genomes, cells, ultimately physiology.

That's where you want to get. We're going to have to have to go up that that complexity scale, the levels of of biological complexity that requires traversing a a data barrier. There's, I think, data that that does not exist that needs to be generated to achieve that level of predictive fidelity. And then we're going to have reasoning. And I

think you know what that will mean is that we can reason over thousands, millions, hundreds of millions of scientific hypotheses in parallel digitally using predictive oracles which can you know actually predict the outcome of an experiment. So the scale that we can ask questions and the kinds of questions that we can ask will just fundamentally change through that feedback is going to be critical.

You know the models are going to need to there's going to be sort of a scaling dimension of this which is which is you know building the data to have those accurate representations and then a feedback dimension where the models can learn from biology can reason digitally can reduce that to a small number of experimental hypotheses examine the outcome of each of those experiments update uh their their their understanding and and build knowledge in that way. So I think that's what's what it's going to look like and and we kind of have to build each of those components. What Biohub is really trying to do is to kind of bring together the experimental and the technology layer that will actually allow us to have these AI models interact with the biology and do experiments. And I think it's you know it's law you know we see incredible incredible advance in areas where we can get feedback computationally. Um, so enclosed domains, but of course, you know, experimental biology is is completely open-ended. And so the feedback principle there is is going to be very different. But, you know, something there's going to be something like RLVR, you know, with with experiments where we can, you know, have models that are that are just really building knowledge and learning from that knowledge and being able to develop more and more accurate representations.

>> You're the head of science science at Biohub. Maybe fun fact for those who don't know the science section of Leaden Space was um basically launched after or in response to Mark Zuckerberg and Priscilla Chan on this podcast about 6 months ago. It's actually very exciting to have you here and kind of come full circle and uh Mark laid out quite an ambitious vision for what Biohub was wants to accomplish and I think you just laid up a very natural you know successor to that. I think you had just joined at like you were like there two weeks.

>> I joined at the very end of October and launched at the beginning of November. >> Yeah. Yeah. One thing I'm curious about is in your eyes, you know, where is Biohub now? Like what what do you want to, you know, accomplish? What are your goals? Big picture goals for listeners who haven't watched um that the episode with, you know, Mark and Priscilla. Um we recommend, of course, that they go watch it. Link in the description. And then have you learned anything even in just the short time of six months you've been here? And like has the vision evolved and you know where where do you see this going? Um I think we can you know how does ESMC fit into this? Uh how does you know the virtual biology initiative that you recently announced fit into this? And then I think there's like several other things that you're working on that we haven't even touched on.

>> Yeah. Well, I'm learning things every single day. Yeah. But the way I think about it, we're building a scientific institution for this new paradigm. And you know, to do that, you know, it's it's it's an institution that's going to be powered by frontier experimental biology, >> frontier technology for for measurement, for observation, and it's going to be powered by Frontier artificial intelligence. >> And this is all open source, right? >> It's a philanthropy. So So our goal is to accelerate science. Our mission is to cure or prevent disease. And to so to do that, you know, our belief is that there's a fundamental gap in our understanding and we need to accelerate science to traverse that gap. And so we're really thinking about every layer of biological understanding that goes

from the most basic, you know, level like the, you know, the the atoms of a protein in a cell all the way to systems of cells in physiology and disease and how can we create models that can capture that complexity, can allow us to understand that complexity and I think, you know, if you think what is, you know, what does the cure to disease look like, right? It's it's not it's not a pill, right? It's not a medicine in the conventional sense. You know, it's it's going to have to be um a system that is capable of modeling and understanding, you know, the underlying physiology of disease in a way that's differentiated for every single human being, for every single different genome. And it's going to have to be able to link events all the way at the molecular scale to the manifestation of disease in in physiology. So, it's it's an incredibly complex, incredibly hard problem. And for us, you know, we're trying to ladder up those layers of complexity and we're trying to build kind of the foundational tools that scientists can use to, you know, answer the fundamental questions there. And so, we're creating atomic level imaging. We're creating light sheet microscopy that allows us to observe, you know, how how all the cells move and and develop in a in a developing organism. We're creating spatially and temporally resolved maps of of inflammation. We're creating, you know, cellular um programming and immune cell reprogramming to be able to actually design completely programmable therapies. And you know, we're creating these digital representations at each of these layers so that we can accelerate the science, simulate what's happening, you know, make biological matter and make proteins and cells and genomes programmable. And I think all of that has to has to come together. And I think if you have the focus and you build, you know, the biology and the the computational layers together so that they're tightly integrated, you know, that's how we're going to make the fastest progress. For the last 10 years, I think we've been one of the one of the big champions of open science. You know, we're we're an organization that does we both fund and we and and we build. And in our funding, we've we've always supported open science. And in our building, you know, we've always done open science. So that's that's something that's going to continue. It's it's just really fundamental. We're not a drug development company. We're not trying to generate therapies. We're trying to trying to build the technology that that that moves science forward.

>> So I think Mark had this concept about if you provide the right tools, then the entire scientific community can leverage them. Yeah. So obviously you believe strongly in protein language modeling as a tool. What is the next most important tool for advancing a general improvement in our ability to tackle human disease? >> Yeah. So I I think the next level of complexity that we have to address is is the complexity of the cell >> and I mean this is going to be tremendously hard billions of proteins.

>> So you say it's tremendously hard. Yeah. If you you come and say it's going to be easy peasy we would just >> Well, I think it's a it's a worthy challenge but but it Yeah. I mean it requires technology that doesn't exist today, requires new, you know, modeling approaches and and probably architectures and ideas in machine learning that probably don't don't yet exist. So there's, I think, deep and fundamental problems to solve. But again, I think, you know, you you you you take it step by step. And so, you know, we kind of start at the molecular layer and we know that that is really fundamental and we can begin to link that to, you know, observables in in cellular biology. I'm really curious because this has been the question that's been on my mind for a long time about we have virtual cell models, we have molecular scale models and there's been a few I've seen a few papers about trying to link them but what what are you guys doing because it sounds like this is becoming top of mind for you.

>> So let's maybe to make the analogy with protein biology. You know what what I think makes our digital representations of proteins powerful and useful is that they generalize. They're able to make predictions for proteins that are entirely unlike the proteins in their training data. You know, they're able to generalize so that you can design, you know, fundamentally new folds, new binding interfaces, new structures. So there's there's this degree of of of um yeah, what we call generalization or generality. um the in in short they can predict the outcome of an experiment that we haven't already made that they haven't already been trained on. So for for digital representations to be valuable you know they've got to be able to be used to answer a new question. So I think that's the critical thing. So I don't we're not there with cells. I think with with you know kind of the current generation of models that are that are being called virtual cells they are good representations of the underlying data but you know they have a very limited ability to predict what will happen when you make a novel intervention in a novel unobserved context but to be able to answer the fundamental scientific questions about cellular biology we need a model that can do that. So, so kind of you know our thinking about this starts starts with that idea is what's it going to take to to get there?

>> Going back to the protein protein interaction the human interact of the if you had that just you know predicting static structures static structures are in some sense not enough for a lot of understanding biology. Uh dynamics are probably for most people a much more useful tool to have. You can start with static. It can give you some insight but it's it's very rarely the full answer. So you have you know a model which is capable of predicting a lot of different proteins. We probably have almost we have a many of these resolved in the PDB. Uh some of them we don't. Given the dynamics and interactions are being are more important. How do you bridge that gap? is to me that seems like maybe one of the the key steps in going from like a really microscopic model of things to going to something which is closer to a virtual cell. You actually have to be able to model local interactions of local uh proteins or RNA or DNA or lipids or you know whatever else is floating in the cell. Is is that sort of like a a goal that you would try to bridge or maybe I'm misunderstanding. Is there like another way you would imagine bridging these two? I mean, one day it'll probably be possible to have a computer that can kind of simulate the cell from first principles, but we're very far from that, right? I I think I think that's far beyond reach of of current computational technology. I mean, kind of even simulating the the physics of the folding of of a single, you know, protein molecule. Basically, we could do it for a fast folding a few fasting, but that's really about it.

>> Yeah. So there's kind of this dual view of biology, this dual complimentary view of biology. One one view is is kind of that kind of first principles reduction, you know, where where kind of all of biology is explainable in, you know, more basic terms in in in basic physical, chemical, biochemical terms. And I think historically there's a long research line of research that's really sought to to understand biological phenomena and simulate biological phenomena in that way. And I mean I think historically you know the field had believed that the solution to the protein folding problem or the protein structure prediction problem would come from you know this kind of first principles simulation. And it really really kind of came came out of nowhere that you know this could be solved using essentially pattern recognition or you know this this type of machine learning approach. So I think I think historically it has been productive to understand biology through information theory through information and you know you can think about the cell as a computer as an information processing machine.

Informational terms, there are, you know, these these very basic principles that link the information coded in

the genome to the genes that are transcribed to the phenotypes of the cell that will result. And so, you know, if we could model and understand the cell at the level of its underlying programs, you know, that that sort of gives, I think, the right abstraction. What do I mean by the right abstraction? Well, I think I mean the abstraction that is possible today because we're in the era of information theory at scale. you know, we're we're in the CL Claude Shannon, you know, had this idea of of kind of the the ideal predictor of the next character and he he had this really beautiful paper where he tried to compute the entropy of the English language and imagine basically, you know, taking an infinite context and then, you know, what is the entropy of the next character and at that time I mean it was unimaginable to I I'd say it create it took a great leap of imagination to imagine that ideal predictor but today we're we we get closer and closer to being able to to build that and we can do that for text and so you know what would that predictor be for biology and that's kind of the idea of of ESM it would learn kind of the the the underlying structure of of all biological phenomena so if you think about that from the standpoint of the cell if we can collect enough outputs of cellular biology that we can observe to reveal the underlying programs patterns and structure you know then we could create kind of the information theoretic description of the cell and I think that would be sufficient to understanding disease.

>> This is reminds me of the a lot of the work that happens in signaling pathways right now right where you have protein in a cascade of different protein protein interactions that eventually cause a phenotypic change in the cell in some way. How do you translate that into something that can be sort of scaled into a or maybe it's something else but how do you for example that >> yeah going back to the bitter lesson >> going back let's just get back to the bitter lesson >> we we need data I mean I think you know why have these advances in protein biology been possible they've been possible because of you know decades I mean for for protein structure half a century of of of work to experimentally determine the structure of proteins and you know and the the you know the effort across you know the scientific world to you know sequence genomes and metagenomes and so that's created you know this this data set that you know you can really you know train a scale on and really learn these these deeper principles and so >> but those two different data sets are actually in many ways quite different like PDB a bunch of >> a bunch of very painstakingly constructed protein structures which many of which were an individual PhD thesis and then maybe the follow similar ones came later which might have been you know 10 of them for a PhD thesis then this people I estimate it's like 13 billion to create the GDB some like very large number the reason people created the PDB was because each individual protein was independently useful like people didn't create it with the sake of we're solving protein structure they saw that like oh this protein we believe is involved in this these these pathway let's understand this protein so we can target it and so on of course there's some caveats here, but at a high level, a lot of this genomic data was uh especially for humans or viruses or bacteria, you know, was sequenced for a very specific reason as well, right?

Well, it's great that these are useful after the fact, but I I wonder if now going forward, especially since, you know, with the virtual biology initiative, um Biohub's virtual biology initiative is like half a billion dollars, I think. Um, and I'm sure there will be more large initiatives coming from Biohub in the future. You know, you have the chance to be very specific, deliberate, and now collect data for the sake of solving a problem with ML rather than depending on a data set which was curated, created for some other purpose.

So given that new opportunity, how do you do things differently? How do you think about data collection um to enable science broadly when you have the option of doing basically anything from first principles? A little bit

of context. We announced a few weeks ago the virtual biology initiative. Um we basically said you know we're going to invest uh 400 million internally in data creation and development of technology to scale data generation to be able to increase the the number of modalities that we can measure simultaneously. We're we also um uh announced that we're going to commit 100 million to catalyzing efforts outside of biohub to generate data. And so, you know, we think that's, you know, a fraction of what's actually needed to to do this, right? But, you know, the hope basically is that by making this initial commitment kind of giving starting funds to some of the groups that are really thinking about this, you know, working to build different core areas of of the data that's going to be needed that, you know, that that's going to be a catalyst that's going to galvanize other other groups to come in and and contribute to this. So, that's what we we really hope to see. You know, the idea is that, you know, this is this is broad a broad-based effort. So, it's it's not just us. So, you know, I can I can say kind of what my perspective is on what data needs to be generated here or what what can be generated. But, you know, we also want to approach this really collaboratively with the scientific community. And so part of this is is also kind of hearing from from scientists what they want. So, so so from my view, you know, there are a few key principles here. The first is speed. Okay. So you know it took decades to build the data for proteins and we can't wait decades. You know this is we need to figure out how to do this in a couple of years and you look at the rate that that uh general AI is developing and it's just you know the limitation in biology we're going to be fundamentally limited by experimental science and data and so we really need to you know work to address that gap as quickly as possible. So I think that's one key thing is is looking at what are the technologies that we can scale up today to begin to you know give this picture of the information architecture of the cell. So there's speed and then there's also uh the idea of generalization. So kind of going back to what I was saying before you know we want models that can serve as oracles for the biology. They can predict an experiment that you haven't done. And so how are we going to be able to do that? we're going to need to look at a multitude of different interventions in a multitude of different contexts. And so it's it's kind of similar to the principle of training a language model on the internet or training a protein language model across all of evolutionary diversity. What does that look like for cellular biology? And so we have to scale interventional biology. So that looks like things like perturbation biology perturb seek measurements that where we can look at combined transcription imaging other other layers of the cellular information hierarchy and there you know number number of groups our our teams are working on this there are a number of groups across the scientific world that are that are working on problems like this that are ready I think to to scale the second is spatial biology I think that's going to be really important and so that's going to help us to really understand the cell in context and I think you know understanding the cell in isolation is is really not what we need it's not the goal right the cell is part of an incredibly complex system in the body and you know to be able to understand disease we have to understand how cells interact the systems that they form the circuits that they form so we need to see that so spatial biology I think is undergoing rapid progress and you know is an area that's really ready to scale up that's kind of what can scale Now I think and um Biohub has actually over the last 10 years really I think made kind of pioneering funding commitments in those those areas and so we've we've we've funded efforts like the human cell atlas and we've built uh tabulous sapiens which looked at uh built built large cell atlases and we've bu built cell by gene which is kind of a database of single cell transcrytoics and so we're you know we're we're really looking to kind of build on that and you know I don't know how many cells there are in kind of the the largest efforts we're probably around like a billion cells or something like that today. So that we've got to go you know multiple orders of magnitude from that. So you know that involves scaling the technologies that we have now but it also involves you know a new the next generation of technology. So we're also funding and supporting efforts in that

area. And there you know we really want to look more at cross modality. You know can you simultaneously see the phenotype observe the transcriptional layer understand what's happening proteomically link that to the genome.

You know we'd like to and the ep the the epigenetic state. You know we'd like to be able to see all of that. And so really pushing technology to be developed faster that can reveal more of those connections and more of that biology and do that in a more scalable way. It's interesting because when I hear most of those ideas, they're often times the things that people already think about in terms of scaling biology. What is the next technology that is going to allow for like enabling data collection technology? Um going to being back to the theme of bitter lesson for biology. You don't have just scaling laws on you know compute and parameters but now the scaling laws probably in data collection in some meaningful sense. where are the next big opportunities there like in TR. So you're talking about developing new technology is sort of like the um as part of the this initiative.

>> Yeah. So I mean I think it's basically the things that I'm saying is scaling what we have now kind of being able to expand the number of interventions that that we can look at expand the number of parameters that we can measure. So really kind of more and more multi-dimensional measurement um and you know drive down the cost and all of that. better gene sequencing kind of better ways of encapsulating cells and being able to measure what's happening not just the transcript dome but but other layers simultaneously. There's an interesting paro frontier there about uh if you have a fixed budget, how much time do you spend on improving your assay versus how much do you spend on actually scaling it? You know, where do you uh went out there?

>> We have to do both of those things, right? Cuz I think with current technology, you know, we can definitely kind of get get data 10x to 100x where it is today with like relatively reasonable investments, you know, but but then to get another 10x or more there, that's going to require >> require a lot more technology development. But the other the other really big principle is is going to be feedback. And so I think that's going to be really critical. And I think you can see that as a layer of of technology development that's that's going to need to occur. And I think there's a lot of great things happening right now kind of automation, flexible robotics that's going to accelerate um where where that can go >> and experimental design as well.

So we typically ask our guests what is a a bottleneck that you would remove that would you know sort of unlock things but we just spent a long time talking about that. >> Yeah I answer that question. So, but I want to ask it but I'm going to give it a spin which is like so maybe a little bit outside of your domain like so language modeling or supply chain something that is a bottleneck that is maybe nonobvious and not directly something that you are working on but that maybe has impact on on the work of biology or biohub in particular.

>> I mean it's a hard question to answer because there's just so many bottlenecks. I mean the one that you know that I always think about is compute but I think that's a pretty obvious one. It's the bottleneck for for all of AI in many ways right now. But, you know, especially because we're training these large scale models, you know, our we're always focused on compute. And I think we're you kind of limited both by the data and compute. I think we're in a you we're in a position where I think we we have, you know, incredible compute resources for a team working in biology. But I think like like all teams working in AI right now, really the limit is is just how much

compute power you So if you could 100x your compute, you think that ESMC would like be way better?

>> I mean, it would definitely be way better. We also need to scale data. So both of those things would have to happen in tandem. >> Have you basically exhausted what's available right now for >> I don't think so. No. No, I don't think so. >> Okay. The large data sets out there or you could >> well I mean more parameters, you know. So we trained ESM C up to six billion parameters. Oh, but I'm saying in terms of data available, like have you exhausted most of what's publicly available in terms of like >> No, no, not not yet. And you know, the atlas that we just built actually has more sequences and structures than ASMC was trained on.

>> So definitely have a little room to go. >> So like what's is that an order of magnitude jump or twice as much? Like how does that how does that work? >> Yeah, I mean I think the SMC is trained on you know say order a billion sequences. So there's there's definitely probably order 100 billion sequences. >> This is large a lot of them are largely redundant. >> 100 billion. >> Yeah. >> Oh, okay. To get that billion, you whittle down from six billion 6.8 billion. Right. So of those 100 billion, if you were to similarly cluster and find unique ones, what do you think?

Where do you think it would land? >> The sequences aren't actually redundant, right? It really depends on what you mean by redundancy because there's I think a tremendous amount that you can learn from small genetic variations, right? Cuz these are these are really revealing of you know kind of the the very basic determinants of protein structure and function at a very fine level. So I think that you know as we think about protein space you know having a vast diversity of sequences across a wide range of protein families is you know really critical for the emergence of this kind of structure prediction capability because I I think kind of large diversity is what trains the model to understand to develop a representation of structure but I actually think that to develop a representation of function it's these very small variations that are important and so I I do think that there is probably a lot more you know it's like the models haven't yet been trained at that level of kind of just like really deep understanding of these very small but critical patterns in and sequence I mean a single a single mutation is enough to destroy the the function of a protein >> so you could you could conceivably actually take all 6.8 8 billion of those retrain everything's the same but >> yeah you could train on more than that there even I mean that's kind of clustered down so >> yeah maybe the question is how far wouldn't you hit the law of diminishing returns here I mean it sounds like you have plans for an ESM4 or an EMD or whatever you want to developing yeah but I'm just wondering it's you know at some point is this actually something that you could exhaust you know people talk about exhausting the pre-training data in on Yeah. At some point. Yeah. At some point.

>> Yeah. But it's not actually something you could conceivably imagine imagine doing in the next few years or even if you don't exhaust it, you hit a lot of diminishing returns for you know the applications that you're trying to predict here where maybe your resources are better spent somewhere else. >> I mean it's it's basically it's an empirical question, right? It's truly an empirical question. >> And so we just we just don't know, you I mean with the SM2 we weren't sure because there were some diminishing >> returns with the SMC you know now now there aren't right so you can kind of look look at that extrapolate from from the scaling law there and you know there is enough data to train that next next model so >> and the other question that we usually ask is any call to action or what what do you want people to go take action on if the listeners want to get involved olve, get hired,

get build things. What would you ask people to do?

>> Well, we just announced or or I should say we are we are going at the time that this this podcast comes out, we will have announced ESMC and this this world model for protein biology. Um, it's going to be open source. It's it's going to be MIT licensed and we want people to use it. You know, we want we want this to be a tool that can unlock science. We're excited to collaborate. We have a team that that that works on that and we want to hear from people and understand, you know, what what we can build that can help to accelerate their science.

>> Yeah, we might have a um a demo slashpaper club of some sort on this channel. So stay tuned. >> Yeah, stay tuned for that. We'll we'll invite you and your team um whoever can make it. We'll feature this paper once it's in final preprint form and spend some time on it for an hour on the way in space paper club. >> Yeah. Uh, thanks for chatting with us. >> Awesome. Yeah, great to meet you guys.