[Transcript] FYI - For Your Innovation / Bettering Human Health Through Artificial Intelligence with Sean McClain and Joshua Meier of Absci

Welcome to FYI, the four-year innovation podcast.
This show offers an intellectual discussion on technologically-enabled disruption because investing in innovation starts with understanding it.
To learn more, visit arc-invest.com.
Arc Invest is a registered investment advisor focused on investing in disruptive innovation.
This podcast is for informational purposes only and should not be relied upon as a basis for investment decisions.
It does not constitute either explicitly or implicitly any provision of services or products by Arc.
All statements may regarding companies or securities are strictly beliefs and points of view held by Arc or podcast guests and are not endorsements or recommendations by Arc to buy, sell or hold any security.
Clients of Arc investment management may maintain positions in the securities discussed in this podcast.
My name is Simon, Arc's Director of Life Sciences Research, and today we'll be discussing Absi, a public company harnessing generative AI to create more effective medicines faster and less expensively.
I'm joined by Sean McClain, Absi's founder and CEO, as well as Joshua Meyer, Absi's Chief AI Officer.
Thanks for taking the time, guys.
Yeah, absolutely.
Thanks so much for having us here, Simon.
So, you know, our audience hasn't had the luxury, like me, of getting to know you guys beforehand, so before we dive into the company, Sean and Josh, would you mind briefly talking to us a little bit about yourselves and also why you're so passionate about the role of AI in drug development?
Yeah, absolutely.
So I founded the company 12 years ago, actually, in a basement lab.
And the original idea was not applying generative AI to biologic drug discovery.
It was actually to engineer E. Coli to produce antibodies.
And we're the first company to be able to produce an antibody in E. Coli.
And what that actually enabled you to do was actually do what's called a pooled approach.
You could basically take a billion-member antibody library, put it into a test tube of E. Coli, and now you have a billion antibodies produced.
And this gave us a huge data advantage, being able to essentially produce these and screen these at very, very high throughput.
And it ended up becoming a realization to me about three years ago that, wow, this is what is needed for generative AI to really unlock biologic drug discovery or what we like to call drug creation.
It's essentially going from this paradigm of drug discovery, where you're looking for a needle in the haystack, to drug creation, where you're actually creating the needle.
And in our case, it's the biologic.
Being able to get the biologic or the antibody with all the attributes you want, the first

go around.
Mm-hmm.
So I think a lot of people are probably familiar with the terms antibody and antigen and understand that antibodies are a really critical component of the body's immune system and help us to attack and defend against disease.
But for those who may not have a very intimate understanding of what the drug discovery and development processes look like, specifically for antibodies, which is what you guys are focused on, would you mind briefly zooming in and discussing the importance of the things you talked about, like with yeast and generating these libraries?
I think that would help people understand a little bit more context.
Yeah, no, absolutely.
So unlike a small molecule or a pill in the bottle where you have a chemist making it, you have to have a living organism make an antibody.
And that's for the production.
But then to discover an antibody, you then have to use usually a mouse or what's called phage display or yeast display.
I'll focus in on the mouse where Regeneron was really the pioneer in creating a humanized mouse where you essentially could take, let's say, a cancer target or antigen of interest.
You inject it into the mouse.
The mouse uses its immune system to then create antibodies towards that target.
And then you extract out the blood and you're able to then find antibodies to a given target.
But the issue with that is that you can't tell a mouse to generate an antibody that hits the specific location of the target that you want.
That has the affinity or how tightly it binds to a target or the developability or manufacturability.
You have no control over how a mouse generates an antibody.
And that's what really leads to these long times lead times to get into the clinic as well as low success rates in the clinic of less than 4%.
And again, what we're doing is completely changing that paradigm and using AI to design the attributes you want the first go around and actually have control over the biology for the first time.
So before I kick it over to Josh, now that it seems like we've fixed the Wi-Fi, I do want to double underline something you said there, Sean, because I think it's going to come up time and time again as we talk about this is when you're trying to develop an antibody against a particular antigen or disease target, regardless of the disease, the idea is that you're exploiting the specific binding reaction and you use the word affinity between those two things, almost like a lock and a key.
And I like that you walked through the first vestiges of this approach with using animal models and immunization, which has its pros and cons.
But moving further downstream, and we'll talk about the in vitro, like the yeast display and E. coli and the bacterial display technology.
And then the ultimate golden goose of how much of this can we just do on a computer?
And so we'll get into that.
But before we do, Josh, I'm going to flip it over to you just for a brief intro on yourself,

how you got involved with AbSci and why you're so passionate about the role of AI in antibody discovery and development.
Yeah, absolutely.
Thanks for having us, Simon.
So a bit about my background.
I've been working at this intersection of AI and biology really before this thing became cool.
It sounds like everyone these days has kind of established that this is going to take over the industry.
But when I first started working on this, everyone thought it was crazy.
Like, why are you vetting your career on AI for biology?
Just go work on at least traditional AI.
Even that wasn't as big then.
I started training my first neural networks back in 2013.
This is when we first got deep learning, really working on GPUs and the space started to explode.
But I come from a family of doctors.
I was always excited to kind of deploy this technology to just better human health.
So I was always trying to find an intersection, but just the technology wasn't there yet.
AI wasn't good enough.
The data wasn't good enough.
And the problem space wasn't really worked out yet.
But fast forward a couple of years later, I was working at OpenAI.
And this was around the time of GPT-1.
So we were first seeing the first signs of life that language models could learn some really interesting things.
And of course, within OpenAI, we were kind of like the believers of this stuff.
We felt that everything we're seeing today was going to happen eventually.
But the application I was really excited about then was like, well, if we can get this thing to translate between languages or generate new text, could you use this to generate new biological text?
Can you just output DNA sequences and protein sequences?
And felt that'd be even bigger than just the NLP stuff.
Because if you just think about NLP, you and I can just write stuff on a computer, but we can't sit down and start clinking out DNA sequences.
You're not going to get anything interesting.
So left OpenAI shortly after that GPT-1 project to go join Meta and help start up an AI for science initiative there.
And that's where we did this first work on language models on protein sequences, published some of the first papers in this area demonstrating that these models could learn some really interesting things.
And a couple of years into that, really the thing that was clear it was missing is that when you're working in a Facebook or a Meta, you have really strong conviction in AI, which

allows you to go into these areas like AI for science and AI for biology.
But the thing that you're then missing is that wet lab component.
So how do you actually validate these designs?
How do you build differentiated training data?
And it became clear that this stuff was going to work for biology and who was actually going to read the value.
I felt that at some point it wasn't going to be as much meta as it was a fantastic place to do this kind of research, but sort of aligning with a very forward-looking biotech company that was also going through this generative AI transformation made a lot of sense.
So to that point, I kind of connected with Sean and these visions really started to collide.
And I joined AppSign.
It's been pretty amazing to see the kind of research and science that can happen in such a short time frame when you really bring these two differentiated edges together.
Yeah.
And it's actually a really funny story how Joshua and I met.
The team had put together a list of probably the top 50 AI researchers in the space and gave it to me.
And I went on LinkedIn and I wrote emails to all of them.
And Joshua responded.
And yeah, him and I totally clicked on the vision and what we could do together.
And I would say the rest is history.
Absolutely.
Well, I wanted to maybe segue back to the main conversation around this point that you're making Josh around the data and what you're able to do at a big tech company versus a company that's trying to combine multiple disciplines, which honestly feels like the field of life sciences is like inexorably headed is like a complete breakdown between the walls of AI and biology.
And if you look at some of the large language models that are being used with human language, you can train them on enormous amounts of information, ostensibly the whole internet and all the texts they're in.
But the issue with life sciences is like a lot of the data generation techniques have been artisanal, poor quality control, databases are fragmented and poorly annotated, and sure we're improving along all these dimensions, especially with sequencing and the breakout explosion and cost decline of DNA sequencing and other kind of ancillary technologies.
But I wanted to focus the conversation on data for a moment and talk about it in the context of Abside.
Sean, you mentioned billions of data points with the E. Coli expression.
And Josh, you know, you and I have had this previous conversation around data being the rate limiting step in training these models.
So I wanted to discuss both of those points together and really focus the conversation back on Abside and what you've done between in vitro and in silico work.
Yeah, Sean, why don't you kick it off as the one who really invented like the core microbial platform we're using here?

Absolutely.
Yeah, as I had previously talked about, we were the first to engineer a very simple organism E. Coli to produce an antibody.
And I guess stepping actually back for just one moment, like, how are antibodies currently produced?
They're produced in mammalian cells or choe cells, and you can really only scale that up to maybe producing thousands or tens of thousands of antibodies in a given week.
And that's just not enough throughput or data to actually start training models.
And so by being in a microbial organism and engineering E. Coli to produce antibodies, what you can do is what's called a pooled approach.
You can basically take a test tube of your engineering E. Coli, take a billion member DNA antibody library that encodes a billion different antibodies, throw that into your test tube, and now you have every single E. Coli making a different antibody.
So in that single test tube, I now have a billion antibodies that have been produced.
So I've gone from thousands or tens of thousands of antibodies to billions.
Now the second question you then have to ask or solve is the functionality.
Now that we've produced them, what is the functionality or potential efficacy of these?
And this is where we develop this really breakthrough assay that we call our ACE assay, where we're able to interrogate every single E. Coli and look at the binding affinity or how tightly it binds to a target of interest.
And so in that experiment, we can then be able to look at billions of protein-protein interaction data points.
And when you're developing an antibody, there's really two important aspects outside of the developability and manufacturability.
It's where does the antibody bind the location or what we refer to as the epitope?
And then how tightly is it binding?
Does it have high affinity or low affinity?
And these are the two attributes we're really able to hone in on very rapidly and train our AI models.
And I'll let Joshua talk about this.
But this data has allowed us to really build these extraordinary models or train these models that have allowed us to actually have a huge breakthrough in the industry, where we were the first to use generative AI to design an antibody from scratch on a computer.
And actually, I think this is a really great time to hand it over to Joshua on what that is and how we take this data, train our models to ultimately kind of see our big vision through of being able to design a biologic at a click of a button.
Sure.
Thanks, Sean.
So if we look about how we're using data at AppSign and even just taking a step back, first of all, and the importance of data in this space, I think the whole language modeling world is waking up to this today.
You look at something like chatGPT and GPT-4.
One of the big things that's advertised there is something called reinforcement learning

from human feedback.
And the thing really to think about there is the human feedback part, where you can go and scrape basically infinite amounts of data from the internet, although some would say it's not really infinite.
We're almost running out of tokens now to feed these models as we're continuing to scale.
But if you just think about that human feedback, it's really critical to give these models this very chat-like capability.
When you're training it on, people actually interfacing with the model and teaching it in a very direct manner.
And the data is even more impactful than just finding random data on the internet because the model is involved in that data collection process.
So in that view, this is something that we've had at AppSign for a couple of years now.
And we've really built up the experimental platform and AI integration accordingly.
So specifically what that means is that almost all the data we're collecting in the lab is actually AI-designed.
So if we go in the lab and we're going to go collect 100,000 or a million data points, a million unique sequences, these aren't just random sequences that you find on the internet or find in a mouse immunization campaign.
These are sequences that the AI model is designing.
Or rather, the AI scientist is designing.
Sometimes there's different benchmarks or baselines that you want to throw in.
There are different controls.
But at a high level, it's the machine learning model is actually creating that data, telling us what data it wants us to collect.
And it's very similar then to this reinforcement learning from human feedback idea.
I mean, on a technical level, it's not exactly the same, but at a high level, it's the same sort of finding that if you allow the model to help you generate the data, you actually end up with a really nice flywheel there, where then you get that data back, the model becomes smarter, and then you can do this again.
So that's really allowed us, I think, to scale the model really quickly.
It's also allowed us to just run massive number of experiments and see what works.
Machine learning is a very empirical field.
A lot of people used to refer to deep learning as black magic, where you would just have AI scientists who just have some intuition about how the models work.
And that's still, in a sense, how you come up with the next generation of these architectures, like the transformer that people are scaling these days in NLP.
Like, how did people come up with that?
I mean, they can kind of give you reasons about it.
But at the end of the day, there's just really strong intuition that goes into this that's built up through just a lot of time spent training these models and evaluating them.
And that's where this experimental feedback loop is also critical.
So the AI scientists can be doing dozens of experiments in a month, take different kinds of models that they're training, test all of them in the lab, and really start to develop

insight and intuition about what's working and what isn't.
And I also would credit as one of the fuels to many of our recent successes.
If I distill it down to a couple of key points here, it's like, you know, you have to solve for a few things to make this type of project work.
The first is, you know, you're generating a ton of data.
You've created a in vitro technique to generate a ton, you know, billions of data points.
The second half of it is you have to actually create, you know, features and labels and get that functional data out, you know, for every part of that library.
So you've done that with the ACE assay.
And then you're describing, and I actually wanted to dig into a nuanced point here, because I'm not sure if it comes across every time, which is that you're getting data on every different, you know, combination or perturbation that you're having in these libraries, not just the top decile, you know, high affinity binders, like in other types of display technologies, you know, you're basically you're physically like washing away all the things that don't stick because they don't stick.
And you're, you know, maybe this is the wrong way to think about it, but you're kind of biasing that data set towards only the things that work.
And if you're talking about reinforcement learning and like, you know, penalties and loss, I want to get into that as well to try to understand the importance of actually collecting the whole thing, not just, you know, the fraction that works.
Yeah, that's that's absolutely right.
So we've actually developed a number of variations to our core, what we call ACE assay, based on that microbial system that that Sean introduced before.
And we can run the assay in a number of ways.
So one way we can run it is similar to past techniques in a binary way, where we're just looking at whether a sequence is binding or not binding.
And what's really nice about that is there are some cases where you just really want to profile hit rates in a very accurate way.
We find that when we run the assay that way, that precision recall of that assay compared to like lower throughput, but kind of gold centered assays is over 95%.
So that means that the information we're getting out there is highly reliable for us to compare the hit rates of different models to each other.
But then we've also developed an alternative way of screening, which actually gives us quantitative information.
So this is something that's very difficult to get with a traditional phage display or a yeast display, at least to do it in a batched way.
What this means is we can go screen hundreds of thousands of sequences, and then we can get a quantitative label for each of those sequences, a score.
And that score correlates very well with gold standard measurements of affinity.
You're seeing piercing and spearmint correlations above 0.8 in most cases.
So taken together, these are really the two tools that you want for any body engineering.
First, you want to identify a binder and you want to be able to profile like what fraction of your sequences are actually binders.

And then among those binders, you want to be able to profile the affinity in a highly quantitative way in a also very robust and accurate manner as well.
There's usually a tradeoff in biology between throughput and accuracy.
And I think we're at a very sweet spot with the assays that we've tuned over time to be able to get very meaningful information for the models.
Yeah.
And I think there's a really important point that Joshua hit on.
It's not just using this wet lab technology to get data to train the models, but it's the validation as well.
There's a lot of manuscripts that come out that don't show wet lab validation, but every single model that we design and we train on, we then go into the wet lab and validate it.
And we can validate roughly three million AI-generated designs in a given week.
And so training on billions and then validating on millions and then being able to have a cycle time, which you can go through all that in a six-week time period, just allows you to make very, very rapid progress in a way in biology that really hasn't been done before.
Before I get into some of the manuscripts, which I'm eager to talk about, I did want to ask just this general question.
I'm going to lean a little bit on a blog that a friend of mine, Pablo Lubroth, wrote, I think last year about KPIs in AI-enabled drug discovery.
You know, and it pulled from a lot of analogies from the SaaS industry, right?
Like all these investors that are looking at SaaS companies, there's like a lingua franca of, okay, we'll use terms like ARR and we have like very rigid definitions of like what everyone's managing to, but on both the investment side and the entrepreneur and the founder side.
And as the space matures, you know, as an investor, and I'm sure a lot of people are in the same position, like you get fatigued hearing about every possible mashup of AI and drug discovery because there are a lot of them, right?
And so I appreciate the point you're making, Sean, about the importance of wet lab, you know, gold standard validation is like a key part of that.
But I wanted to ask, like, what are some red flags and some green flags to you when you're thinking about AI and drug discovery come together?
What are the things where you're like, oh, this is, you know, legitimately differentiated or there's some value here.
And also to the extent that you're able to talk about it, I'd love to know what are the KPIs that matter to you as you're tracking your own progress against, you know, this ability to generate fully in silico antibodies without clear, you know, commercialization events or contracts, like what are those metrics?
Yeah, absolutely.
I would say that there's like three key metrics that we look at.
Well, you know, first is, you know, being able to specifically hit an epitope that you want.
So again, an area of the target that's of interest and being able to hit that and not having, you know, any sort of poly specificity towards anything else.
The second is then being able to have the accuracy of the model be good enough to predict the exact binding affinity that you want, you know, let's say you wanted in this particular

instance to get the biology wanted, you wanted a medium binder, you can have the model generate that for you.
The third aspect is, you know, I totally lost my train of thought on the third aspect, but I will hand it over to Joshua.
If you agree on kind of those, at least those two kind of major areas of, you know, epitope specificity, as well as being able to, you know, just predict the affinity.
And actually the third thing that I was actually going to mention was the accuracy of the model.
So can we hit the epitope 25% of the time, you know, 25% of what is coming out of the model is hitting that epitope, you know, and, you know, ultimately increasing that over time.
So being able to get up to greater than a 90% accuracy is really where we want to be.
Yeah.
So those are exactly some of the problems that we're focused on at Absight right now.
I think one of the exciting things about this space is that it's moving very quickly and those KPIs will definitely continue to update over time as well.
Right.
So it's not like you were talking in a business setting, right?
There's like some AIR, very clear metric that you want to go after it.
What is your revenue for SaaS business?
I think in AI drug discovery, you have to be more creative than that.
You have to think through what are sort of the unmet needs that your application can bring in and just be laser focused on solving for those.
And then once you solve those, then you kind of move on to the next problems afterwards.
So like Sean mentioned, some of the things that are just not, you can't really do with existing assets or things like epitope specificity.
It's something that makes a lot of sense for an AI model because you can think of it as like prompting.
You can prompt the model.
I think it should chat to you sort of sunny prompt the model with a specific epitope that you want to hit or potentially other properties that you want your molecules to sort of fit.
Another thing is about accuracy as well.
So you want a model that's very calibrated.
One of the things that we see with our models is that it's a phenomenon that we're calling hit rate decay.
As you screen more and more sequences from the models, the accuracy started to go down.
You know, we're screening hundreds of thousands of sequences here.
And we're like, wait, you know, at some point, doesn't the model run out of binders?
You know, where, where are all the binders?
You know, is it in the first couple of the model is pulling out, or is it all the way through the end?
Do you really need to screen 100,000 in order to discover a binder?
And turns out the answer today is no, like we don't have to do that anymore because the model is actually giving us sequences from the first, let's say, 10 or first 100 that you're pulling out of the model.

So that's, for example, another KPI that we look for.
And we've put a lot of thought into the kinds of metrics and statistics that we've developed to sort of evaluate our models in that way.
So this is another way that we're able to really benchmark our models against each other.
And therefore, you know, abilities are focused on making progress here.
And the other thing I'll mention too is that we're not just developing, you know, AI within biologic drug discovery for, for the sake of it, like we're really wanting to utilize this to be able to discover, you know, new biologies.
You know, one of the things that we're really excited about is actually utilizing this sort of technology to generate antibodies towards GPCRs, which are notoriously hard to drug with standard immunization or phage display approaches and being able to specifically target, you know, the epitobon of GPCR.
This really starts to unlock new biology and new targets, which is ultimately going to be, you know, best for, for patients.
And, you know, it, you know, becomes highly differentiated, which we're, we're really excited about.
And I think that's what you're going to start to see more and more is, is generative AI unlocking new, new biology in a faster way than we've ever seen before.
And so earlier, Sean, too, you mentioned this concept of humanization.
When you were talking about, I think a transgenic mouse model, I wanted to blow up this point about humanization as well.
Because we've talked a lot about affinity.
And of course, there are multiple things that can go wrong, you know, even if you have an excellent binder.
And I wanted to talk about this a little bit because it's the subject of one of your papers, too, on this concept of naturalness, which is like, forgive me if I mischaracterize it, but it's sort of like an ensemble of a lot of different, you know, key aspects of what it means to have an antibody that is developable, meaning it can be manufactured.
It's hopefully rid of downstream liabilities, you know, a human body is able to take it in without any sort of, you know, unwanted immunogenic or, you know, side reaction.
So I wanted to learn about this like multi layer Swiss cheese model of optimization, you know, past just affinity.
And maybe a little bit more of a technical add on to that question is, are these optimizations happening in parallel, or is it more sequential?
Right? Does that make sense?
Like, you know, you're starting and kind of going downstream.
So I'd love to know more about one.
Yeah, no, absolutely.
So if you just look at an antibody sequence, like there's more sequence variants or drugs you could design than there are atoms in the universe.
And so the search space is ginormous.
And, you know, you look at like evolution, like a mouse or humans evolved to have a particular immune system, and they're going to have a particular immune repertoire where

they design certain, you know, antibodies that, you know, don't have a immunogenic response.
And essentially, you know, that's what humanization is, is making sure that the antibody you design isn't going to be targeted by the immune system and have what's called an anti drug antibody.
And when you look at, again, the overall possibilities and our ability now to search, to start to search that space, you start to become concerned with, you know, immunogenicity, like is what the model is designing these kind of sequences that have the same functionality, but are very different from, you know, what looks like a normal antibody.
So that's when we started to build out this naturalness model, which I'll have Joshua kind of talk about that.
But what we showed with this model is that it's inversely correlated to immunogenicity and has pretty good correlation to developability and manufacturability.
And so you're ensuring not only are you getting, you know, the functionality out of the antibody you want, but you're also ensuring that you have low immunogenicity or it's as human like as possible when going into the clinic, because there are clinical trials that do fail due to high immunogenicity.
And so being able to control for both of these is really important, you know, aspect.
And I'll let Joshua kind of dive into, you know, how we went about, you know, doing this and then, and then how are we, you know, doing this in terms of are we doing this in parallel at the same time, or is it a sequential?
Yeah, so when we think about the naturalness model, the key insight that we were trying to build towards is how can we take sort of this universe of antibodies that we know about and then use that information to distill the key factors that make an antibody so-called natural.
And that's really what the naturalness score is.
So the way that we train the model is we took hundreds of millions of antibody sequences that you find
within humans, that you find within animals, really naturally occurring.
And then we asked the model to give us some score.
So given a new sequence, what is essentially the likelihood that you would see that sequence within one
of these natural immune repertoires?
And it turns out, and, you know, it's kind of intuitive that this is the case.
But if a sequence of the model thinks it's more likely to have been found in immune repertoires, then it's more likely to have all of these properties that Sean was talking about before, like the developability properties having low immunogenicity.
And the reason why we built the model this way is it turns out that there is significantly more immune repertoire data that you can get access to than there is even some of the developability and immunogenicity data that's out there.
So on a fundamental level, you can think of the model as like bootstrapping from all the information that's available here.
And it's a very similar insight to what you observe with, you know, something like a GPT3 or a chat GPT, right, where you pre-train the model and lots of information.
And then the model is sort of learning, like, what are the real semantic rules that go into

language? We're doing the same thing here now for what are the rules that go into an antibody. OK. And I imagine, like, if you're if you're working along some, you know, like multi parametric optimization, you want to constrain this space as much as you can with the early steps. And so is what you're doing basically, like, you know, before you even get down to like what could be a good binder or not, let's throw out all the things that we know are going to be like highly toxic, or maybe they have like a premature stop codon or something, you know, that would truncate it like, well, you know, where along the process are these different steps working, I guess is what I'm curious about.

So for each one of these things we're building, there's actually multiple ways we can kind of combine them together, depending on the problem that we're trying to solve.

It goes back to your points earlier about a KPI, for example, you need to be very thoughtful about what you're applying this technology to.

One of the issues I think that the the field has broadly is you have a lot of smart AI people who are building hammers and looking for nails.

And, you know, as we all know, you know, if you start with the nail, you know, as the saying goes, it's going to be easier to find that solution.

So when we think about combining these together, it really depends on the problem that we're trying to solve. So usually in the case of like a drug discovery campaign, you're going to bring a naturalness as a way to select the molecules that are most interesting to you.

So we're at a point now where we can take our AI models, and then we can come up with hundreds, even thousands of potential sequences that could be brought forward for your preclinical testing. And the question is, how do you prioritize between these hundreds of different molecules? And that's where we want to bring in this information about naturalness, for example. That's one way that we use it. Another way that we can use it is actually just bring all these properties together from the start.

So maybe you have some lead, and what you'd like to do is optimize that lead for various properties. Maybe you don't have any sequence that has the affinity or naturalness profile that you'd like, and you can just co-optimize for all these properties together. So that's another one of the ways that you might use this information.

I imagine, you know, once you get past affinity and specificity, we're just trying to like systematically remove as many of the downstream like gotchas and surprises that keep drugs from ultimately making it to, you know, commercialization. And so we've talked a lot about some of these, I guess you could call them known unknowns, like we know what they are, but we just don't know where they rank or how dangerous they are. And I would imagine like over the years, people have always kind of had this like, you know, maybe it's hubris, but maybe it's also just strategy or thinking about the problem. But you try to get rid of all these downstream question marks. And like, is there a set of just, you know, unknown unknowns, things that we're not even really able to query or look for that could still be a surprise? Like I think about this with, and I'll use a specific example to maybe guide how I'm thinking about this, but like post-translational modifications that are not like directly being measured by DNA sequencing, or even in some cases like, you know, in like mass spec or protein sequencing.

If those show up in a, in a CDR or one of those active regions, like you could still, I imagine, change the binding properties of an antibody. And I know there are some tools

to predict those liabilities and things like that, but I'm generally just curious about like the unknown unknowns with antibody development and like places where you think no one's really paying attention to like, you know, how to ensure that these things actually do make it all the way to the end and to the finish line.
Yeah, absolutely. I mean, I think like one thing that comes to mind is a lot of developability, you know, attributes. Let's say that, you know, you end up finding out that you can't, you know, you can't get good enough viscosity in order to do sub-sub-q dosing. And so you're having to do like, you know, infusion instead of being able to do sub-q injections. And so, and it all comes kind of back to data. Like, you know, that's actually one area that I think like pharma actually has a lot of data on is the developability, you know, attributes. You know, like us, like, we don't have technology to scale up, you know, being able to screen for viscosity in a kind of a high throughput manner. And so for us, that could potentially be, you know, a blind spot for us is, you know, we ultimately get down the road. And it's like, we really want to do sub-q dosing, but we don't have the data to, you know, train our models to predict for that. And I think that's where, you know, some really interesting ideas that we've always had of like, how could, you know, some of the data that large pharma has that's not necessarily for the drug itself or like the efficacy and the functionality, but it's kind of on the developability. How could you actually kind of develop like a consortium where you could take all of this data, it's for a greater good, and everyone could have access to the models that are trained on this data to really help with kind of these developabilities where, you know, it is an unknown unknown for us, you know, when we, you know, get to get to the end, but others have, you know, the data. I think these are some interesting kind of ideas that we've had on how, you know, the industry can, you know, can collaborate and form, you know, potential interesting, you know, consortiums. Yeah. And I wanted to maybe just take a little bit of a side step and touch on this topic, you know, earlier, Joshua, and you were talking about large language models and the analogies to, you know, to biology. I mean, the analogies are really beautiful. Like, if you think about, you know, proteins having essentially like 20 canonical amino acids and, you know, the English alphabet has 26 characters and they're structured into words and those words form paragraphs. Like there are a lot of, I think, similarities between those things. So I wanted to ask a question along the same sort of axis, which is, you know, we've been talking for the last 30, 40 minutes about antibodies, which are a subset of biologics, right, proteins. The models that you're building are certainly, you know, not limited to one particular application or one particular modality. And I would like to understand a little bit more about, like, what's common, what's different when we're in our local neighborhood of proteins? And then as we really zoom out, like, what modalities do you think are most amenable or maybe the most kind of, you know, recalcitrant to in silico design? Yes. Are you thinking outside of biology or outside of antibodies more generally? Yeah. Just starting with antibodies kind of zooming out a little bit to biologics and then maybe taking a big step back and thinking about everything from small molecules to whatever else. Sure. So first starting with antibodies, you know, antibodies are really interesting because most of the binding is conferred by the CDR regions, componentarity determining regions in these antibodies. And that's something that turns out that the model can learn extremely quickly, right, because it's something that is very clear from the data. And that also means you have a very focused design task to work on designing those

CDRs. And that, you know, you really want to bake that into the task that you're working on. You have these CDRs and then they're binding to a specific epitope. So we want to provide the model, for example, with some structural information. You know, this is a departure from, let's say, a traditional language model that's just in the sequence space. We're really thinking about how do we bring in meaningful information about that target protein structure so that we can really focus the model on that. I think this is something that's maybe more unique within the AI space, right? It's, this is not just to chat GPT off the shelf, train on biology sequences.
This requires really some deep thought about how do you best apply AI to these problems.
Now, when we zoom out a step, so we're focused on antibodies, they have a specific form.
When you move to a general protein, you don't really have that information anymore. So you need the model to be able to represent that information. I'll give you one of the challenges that comes up here. If you take like a general protein model and you apply to antibodies, if you look at a lot of the success of like protein language models within general protein design, essentially what the models are learning to do, or at least what we, it's hard to know for sure anything in deep learning. So it's really a little bit of like speculation, what these models are doing, but they're presumably taking a bunch of evolutionarily related sequences and using that to build some understanding of the protein world. So this is the way people used to do, let's say protein folding 10, 20 years ago, you would take a given protein sequence, you would go enumerate the most similar
proteins to that, and then you would start to do some statistics on top. Then you're like, okay, I see like this position is always the same. That probably means that position is like doing something really important. Or I see these positions covariate, it's always AB or CD, probably means that they're touching each other. Now, antibodies don't work that way. You know, antibodies are not the result of evolution. It's not random, you know, pieces of dirt that we're finding around the world that were involved in different ways. It's created by the immune system. So you need to figure out general models that can really pick up on all that information.
And I think that sort of segues into how do you apply this to an even broader picture where the modalities really start to change in a broader way, right? If you wanted to have a model that brings in proteins and small molecules, or maybe you wanted to bring in something like protein dynamics, or if you wanted to represent a whole cell or a whole physiology, again, you need to be thoughtful about what the domain, how the domain switches over time, and then how to sort of build the right biases into your models to pick up on all that information.
And you mentioned this point, you know, about off-the-shelf bottles and reapplying them.
And I wanted to just kind of give a reference frame, like if we're talking about experimental ground truth, you know, for those proteins that we can crystallize and use like cryo-EM or x-ray crystallography, we're getting on the order of what in terms of accuracy, like an angstrom, roughly? You're talking about for protein folding, predicting the structure?
Yeah, like what I would love to do is just for people who are thinking about how good these models are at predicting the general structure of a protein and how that relates to what experimental ground truth is, and like what's been that vector of improvement. And then I want to flip it back over and just ask this theoretical question about, you know, is there an asymptote? Like, is there a reason why they can't just blow right through that experimental, you know, plateau?
Yeah. Okay. So I think first of all, what you're citing here, let's say in a protein folding

setting, right? You have all these structures. Now you can use AI to predict those structures. For the average protein, you can get pretty close to experimental accuracy. Like we're talking about like an angstrom, as you just pointed out, right? That's sort of what you can get in the lab anyways. So these things are very highly performant. Now there's exceptions, you know, something like antibodies, it turns out are actually very hard to model. Those CDR loops that I mentioned before, the ones that actually confer a lot of the binding, they're one of the hardest things to actually model with protein folding tools. So one of the things we've done at Abside, for example, is develop state-of-the-art capabilities for antibody antigen folding. So we can really get a good idea of what our potential drugs are going to look like before we even go test them in the lab. But where things start to get really interesting is accuracy becomes a lot harder to measure when you're no longer in a prediction setting. You know, for protein folding, you have a sequence and there's some structure, right? There is some answer. That sequence folds up into some structure.

I mean, maybe it's dynamic. We can, you know, have a whole conversation about whether there is like one answer or not. But for the most part, it's a classification problem. When you move into a design setting, it's not like that anymore. I take some target antigen, I design antibodies against it. Like what is the correct solution? I mean, there is no correct solution. There are probably hundreds of correct solutions and it really depends on the constraints that you're putting into the model. So that's where evaluation becomes critical. The NLP people realized this, you know, a few years ago and started to put together all these benchmarks, which, you know, things like GPT-4 just put everything, you know, just blew everything out of the water.

And now there's like, did you just memorize all the benchmarks? How do we even evaluate models anymore? And, you know, you're getting to that asymptote now. One of the big issues you're going to have now in NLP, the big issue will be evaluation. In biology, it's actually more convenient because when we have high throughput experimental capabilities, I don't need, if I have a hundred thousand proteins that I generate, I don't need like a hundred thousand human labelers to go look at each of those. I just have a hundred thousand bacteria do it. And it's actually a lot more scalable. So I think this is one of the funny things about biology. It feels very esoteric and it feels very hard to build the data. But going to that point about like asymptotic limits, for NLP, it's starting to become really hard. And this is something where biology will be able to pass, I think, because again, you can really evaluate all of this within the lab in a very scalable way. Interesting. So the problem is kind of flipped. It's like, you know, in one hand, it's dearth of data, but easy to do these valuations and then the mirror inverse of that for NLP. Yeah, I mean, that's pretty exciting. Like I would love if the, you know, the dominant sort of headlines around AI and just started to become mostly biological. And it seems like with AlphaFold, I think that it really became, I mean, I remember going home for Thanksgiving is the first time I've ever been asked about protein folding over Turkey. So that was, that was cool.

Yeah, no, thank you for that explanation. I think that makes a lot of sense. And, you know, I think the importance is like the devil's in the details with these things, like, you know, good enough can be fine for some downstream tasks. But to your point on antibodies, it's really got to be virtually perfect to keep going, you know, down towards the next thing. So maybe in the last couple of minutes, just kind of open ended questions around, you know, the pace of technological improvement, what this field looks like by, you know, let's say a decade from now. And

if you think about the enabling technologies on both the dry lab and the wet lab side that could really boost, you know, your capabilities of doing what you're doing, but also just for the field of drug discovery, you know, it's an exciting thing for us because we're looking at, you know, full stack like hardware, wet ware software. And I'm interested, you know, for you guys, if you have strong opinions about that. Absolutely. I can start off with this. So where this is all headed is personalized medicine. What we're going to start to see over the next five to 10 years is seeing that 4% success rate, you know, start to increase, you know, to 10% to 20% and so forth. And what that actually enables is you actually to go after smaller and smaller patient populations because you're no longer, you know, having to pay for as many drugs that that failed in the clinic. I mean, that's why drug development so expensive is because you pay for the 96% that ultimately fail. But as you increase that, you can go, you know, smaller and smaller patient population ultimately getting to the point where it's cheap enough to do personalized medicine and ultimately being able to take a patient sample, find the target that's relevant for that disease and then design an antibody that not only hits the epitope and has the affinity that you want, but the model actually starts to understand the biology. It knows that that target and it says, okay, this is the epitope and the affinity that you need in order to achieve the biology that you're looking for. And I think that's the next big step for us is being able to not only, you know, design the antibody hit the target we want and the affinity, but actually starting to understand the biology. And I think that's going to be a big next step for us is being able to scale data around the biology. So when I get a brand new target that comes in, I want the model to then again, give me the antibody that hits the epitope and the affinity that achieves my biology. And so I see synthetic biology playing a very important role in all of this. I mean, you see how important synthetic biology was for our success. I mean, we started out as a synthetic biology company. And that technology was what allowed us to scale biological data to train the AI models. And the synthetic biology technology on its own to get the data. Yes, we could get by with it, but it wasn't going to solve the 4% success rate. It wasn't going to solve the decreased time to clinic. You need to combine the two in order to solve these biological problems. And that's the exciting thing is because the next problem we have to solve is then in the wet lab. So how do you scale the biology data to train the models? And that's another CynBio problem. And so I feel like you're going to go, ultimately, CynBio is going to play a very, very important role for developing technologies that can scale data to answer the next question for AI. Got it. Got it. So it's basically two things, I think, that are shining through on that for me. The first is people may not understand the disincentives around creating drugs against rare disease. To your point, you underwrite this huge investment with a low, very low probability of success. And to recoup your investment as a drug company on the back end, you've got this finite window of time of patent protection to go out and make that money back before there's competition. And not only is it difficult logistically to actually source and find these patients when they're scattered about and there are not many of them, but there just aren't many. And if you have a treatment that in some cases, like we're seeing with some of these gene editing techniques that potentially could be cures, right, then you're deleting that person's status as a sick person, which is great, patient. And I think great for humanity. But to your point, the economic incentive has to be there. And so by increasing the probability of success, you make tractable a greater subset

of diseases. So I think that's a really good point. And then on the synbioside, the genetic engineering, you're right. I mean, from this whole conversation we've talked about, there's a lack of biological data that you need to train these models using large animal systems that require a lot of time, immunization in the case of antibodies. It's just not a high enough throughput upscreen screening tool, right? We've got to figure out in vitro mechanisms for generating data that's not only abundant, but high quality, testable, functional. And to do that, we might have to do some genetic engineering of our own on these single cell organisms. So I think those are both great points. Joshua, did you have anything else that you wanted to add to the mix here? Yeah, I think maybe one thing on the technical side too that I find really interesting here is about where does colonization kind of appear across the value chain here? So one of the nice things I think about kind of working out this intersection of AI and data is that the inputs and outputs to our models, I think the costs seem to be coming down really quickly. So on the inputs, our model designs all these sequences, we need to go synthesize the DNA encoding those sequences. So you're starting to see the cost of DNA synthesis really come down over time. And then on the back end, we take all this DNA, we run it through our E. coli system, we have this flow cytometry based assay. And at the end of the day, you've got sequences. You've got DNA and you need to go sequencing. And sequencing, as we know, the costs are also going down dramatically. So I think one really nice thing about working in this field is that if we think about like the dollar cost of every, let's say, you know, data point we're creating, that number is going down over time, which will mean that, you know, for the same amount of data, we can say amount of money, we can just produce more data over time. So it's a really nice place to be. And it's actually reminiscent, I think, of like the early days in computers. So, you know, when Apple and Microsoft were building the first personal computers, it was just really expensive to get hold of that hardware. And what did Microsoft do? They kind of put out all the blueprints, how to make a computer. And people looked at that. And then all the OEMs started to make their own computers, they commoditized it, got really cheap. And it's like, well, you have all this hardware, but now you need the software and you go buy Microsoft OS. So I think it's a really interesting thing. It's a way that really massive companies start to get built when you have, you know, a real sort of advantage in the market because things around you are getting cheaper. But the problem you're working on is very hard and you have a differentiated angle on it. And I think at the end of the day, too, like the generative AI companies that are going to win at the end of the day, and this is a cross to like all industries, are those that own and control the data? At the end of the day, those are going to be the companies that ultimately win. And I think like, you know, if you're looking at kind of through an investor lens, especially in this space, it's the differentiation is the data. And where are you getting your data? How are you training this? Because that's ultimately like what's going to enable you to have a competitive mode. Because at the end of the day, when we go fully in silico, and we're able to design a drug out of click of a button, it's going to be very hard for people to catch up to us because we've, you know, spent so much time training the models on a ton of data to increase

the overall accuracy. And then we've done our own model designs as well. And so it comes down to data. Data is key to success in generative AI. Yeah. And I, you know, this is another one of those cool KPIs we were talking about earlier is like, what is your dollar per data point, you know, governed by the inputs and outputs is another one that I like. And maybe the last comment or question I'll open it up to you guys is, Joshua, I really wanted to zoom in on the statement you made about, you know, the cost of the data point coming down. It seems like if you look at the last 50 years of, you know, drug discovery, drug design, as that kind of improvement vector as as, you know, continued churning along and the cost per data point comes down, there's a point in which you almost like stop taking the traditional like hypothesis driven approach to your problem. And you kind of just start letting the data point speak for themselves and tell you what to do because it's cheap enough to do it that way. And of course there's experimental design, I'm not saying throw that all out the window. But like, how do you think about hypothesis driven science at an era where there's abundance now in, you know, the data that you're able to generate? Well, that's a great question. I kind of did this thought experiment myself recently. So I started just to, you know, I don't do enough coding in my day job anymore. So I said, let me just try to build something in a programming language I haven't worked with for for many years, right? Start building an iPhone app. And I use GPT-4 as kind of my co-pilot, right, to help me build it. And when I realized something really interesting happened, I never really done this before. I was literally just copying and pasting the code it was giving me and putting it straight into Xcode to kind of build my app, right? Usually I'm going to read it very carefully. I'm going to pull out the components I need. I'm going to rename the variables. I was kind of just going on autopilot at some point, right? And I imagine a similar thing might start to happen in drug discovery, right? Where the model just starts to get so good that you're just like, okay, the model says go test this antibody. You just do it. You don't even think twice because you just become second nature to trust this thing. So I think that might be where the field is heading. Of course no one knows, but just, you know, it's kind of cool to work in this space where you, you know, no, of course no one knows how things are going to play out in drug discovery, but you do see it playing out in an earlier field. So a lot of those like product experiences, you kind of get a sense into the future of what it's going to be like, you know, in this industry. So yeah, that's one thing that might happen, right? It could really change the game of how we do science here. So my question is what app are you building? It was just, just playing around with some, some cool AI capabilities. I mean, well, that was the cool thing about, you know, GPT-4 is able to build that app just in an evening, right? Because you can just, you know, maybe, you know, it's actually, Sean is making me push on this, right? How was I able to do it quickly? I had some playbook, right? I said, like, wrote down a piece of paper, this is what I want the thing to look like and started building it that way, right? So I think it's going to, it could lead to a world where we're just a lot more efficient, where scientists, instead of getting distracted about, like, fancy technologies, you just trust the model and you think about, you know, what is the, the real application that you want to go after, right? So you think about your disease indication, for example, you can be very passionate about that and allows us to be more strategic about biotech and stop spending as much money, kind of chasing a lot of fancy technologies. I think drug discovery is really hard, so it's going to take us some time to figure this out. And I think we're, you know, kind of being the trailblazers here on

that at AppSci, like, thinking about what that future looks like. So for folks who are excited about this, like, come join us and work with us on that journey. But yeah, I think it's a really exciting time to be working in science more generally, because AI is, I think, going to really revolutionize the way that we think and do our work. Well, I think that's a good place to end it. Guys, it's been a blast. For people that have been listening in and stuck with us through the end here, please go follow AppSci on Twitter. We'll be sure to link, you know, all the recent literature so people can, you know, stay engaged and learn about, you know, what you're doing and making sure that they're informed. But other than that, Sean, Joshua, thank you for spending an hour talking with us about this. It was a lot of fun and hope to see and talk to you guys again soon.
Yeah, thanks so much, Simon. Thanks, Simon. It's a lot of fun.
ARC believes that the information presented is accurate and was obtained from sources that ARC believes to be reliable. However, ARC does not guarantee the accuracy or completeness of any information. And such information may be subject to change without notice from ARC. Historical results are not indications of future results. Certain of the statements contained in this podcast may be statements of future expectations and other forward looking statements that are based on ARC's current views and assumptions and involve known unknown risks and uncertainties that could cause actual results, performance or events that differ materially from those expressed or implied in such statements.