**CRAIG:** Hi, I’m Craig Smith and this is Eye on AI.

A couple episodes back, I spoke to DeepMind about AlphaFold, which harnesses deep learning to predict the shape of proteins. This week, I talk to Max Bileschi, a software engineer at Google Research, who is part of a team using deep learning to identify and label the characteristics of segments within the amino acid strings that make up proteins. The two approaches are somewhat complimentary and promise to enable new applications, particularly for healthcare.

But before we begin, I want to mention my sponsor, [ClearML](https://t.clear.ml/eye_on_ai), an open-source MLOps solution. Promotions can be distracting for regular listeners, but they are important in keeping podcasts like this alive.

[ClearML](https://t.clear.ml/eye_on_ai) is the best collaborative MLOps tool on the market, allowing you to easily Track, Orchestrate, and Automate ML Workflows at Scale.

Do me a favor and checkout ClearML – at [clear.ml](https://t.clear.ml/eye_on_ai). They are supporting this podcast, so if you’re a listener, let’s support them.

Meanwhile, I hope you find the conversation with Max as fascinating as I did.

**CRAIG:** So why don't you start by introducing yourself, telling us how you got involved in machine learning and. What your current work is?

**MAX:** Sure. So, my name is Max Bileschi and I'm a staff software engineer in Google Research.

**MAX:** I had a somewhat non-traditional path to research. I came from big data and software engineering and worked at Twitter before doing big data stuff. Then I came over to Google, did some more big data stuff, and then there are opportunities on machine learning. And the machine learning team is doing some biological research or bioinformatics research.

**MAX:** And that really interested me. I made a nice addition to the team, my skills in terms of being able to deal with data and quality software practices and whatnot.

**CRAIG:** So, you're not specifically a machine learning guy. You're a big data guy that's involved in a machine learning project.

**CRAIG:** Is that right?

**MAX:** Now I am a machine learning guy slowly putting on

**MAX:** that hat.

**CRAIG:** I'm always interested in people's journeys. How you got involved in big data and what your educational background is.

**MAX:** Yeah. So, I went to university at Buffalo state school in New York and I did math and computer science.

**MAX:** I did my master’s in computational complexity so more of a math background. And then instead of going to grad school, I had an internship at Google, and I said, wow, this is tremendous. This is way better than I thought it would be. My quality of life was tremendous. I then accepted an offer to go to Twitter. I worked on Twitter video.

**MAX:** And then I worked on developer tools and that's how I got involved in big data. I've just been incredibly fortunate.

**CRAIG:** A lot of my listeners are practitioners, but not PhDs. A lot of the people I interview are PhD candidates or postdocs or whatever, doing research.

**CRAIG:** And there is a. A trend where people are deciding not to go on and to do PhDs and instead, go into industry because there's plenty of learning to be had in industry without having to go through the process of getting a PhD and a lot of industry like Google is no longer really looking for PhD.

**CRAIG:** They're more interested in looking at a portfolio of work and that sort of thing. Are you part of that or representative of that at all?

**MAX:** Yeah. That's me. I chose not to go to get my PhD. I think that there's a deep and important question, which I don't have the answer to, but where does private industry belong in doing research? And there have been ebbs and flows in our country's history, but also in the broader history beyond the U S. And for machine learning research, you have to look at where the people are, where the data is and where the funding is. And a lot of the most famous folks in machine learning either work in industry or have positions in the industry. And there's obviously a lot of data and the funding is actually a quite important thing. Quite different from academia where we have to go after grants on a periodic cycle. So yeah, I think an important question that we have to think about is where do we want research to happen?

**CRAIG:** So, tell us about the project that you're working on now.

**MAX:** Sure. So, this project looks at using machine learning to annotate data, which comes from all sorts of different organisms. So, these are proteins.

**MAX:** And proteins for those listeners who don't know are these little molecules that perform tremendous tasks inside your body and out in all organisms. And the way I like to think about it is we're in this factory. And we're surrounded by all these little machines, and we don't know what a lot of them are doing.

**MAX:** And it's quite expensive to take one of these little machines into a wet lab and measure it and see what it's doing. And so, we need computational approaches to be able to annotate this data and this idea of using computers to annotate this data is not new. It is many decades old and has a lot of success.

**MAX:** And the thing that we're working on is leveraging some of the newer techniques to add new information to the vast treasure trove that we already have about how these machines work and how we might be able to use them.

**CRAIG:** Are you looking at the morphology or are you looking at the makeup of the molecules, of the proteins? And the reason I ask is I just spoke with Pushmeet Kohli. Who's the head of science research at DeepMind and was the leads on AlphaFold and so how does your work relate to what they've been doing, and do you use their work?

**MAX:** That's a great question.

**MAX:** So, we have, we have the chemical formula of a protein, and I like to think of this as a blueprint for if you built this machine, how would you build it? And what DeepMind focuses on is the shape of the machine once it's built. And this is incredibly important for determining how it might be used.

**MAX:** What binds to all sorts of different things. We are focusing on a different aspect of that same problem, which is what is the function of that. So, trying to go directly to how might this be used and both approaches are really trying to get at this core of the problem that I was saying earlier which is there's so much value here and there are these new techniques which can leverage much more data than previous techniques.

**MAX:** Can we annotate more of the protein universe with shape or with structure or function? And how far can we go?

**CRAIG:** And so, you're annotating shape as well as the amino acid string or sequence.

**MAX:** Yeah. So we are, we do not predict the shape. We do not predict the structure. We predict the function. And so, for those of your listeners who might be familiar with ImageNet or something like this, it's a large classification problem.

**MAX:** So, if you list out a coarse grain classification of what all proteins might do, it might come up with about 20,000 classes. And this is what we did is we want to, given this sub sequence of a protein, predict its function by classifying it into one of 20,000-ish buckets

**CRAIG:** and based on the amino acid sequence or based on the shape or based on something.

**MAX:** Yeah. So, we, we just take the amino acid sequence, and this is in contrast to. Other approaches, including DeepMind that use something called a multiple sequence alignment. And so, the bedrock of bioinformatics is this thing called sequence alignment or multiple sequence alignment. And this idea comes from the fact that evolution conserves things that work well and doesn't conserve things that don't.

**MAX:** And so, if two sequences of DNA or two protein sequences are related evolutionally, they should be able to be lined up. And so, when I say lined up, the characters should be mostly conserved. And one of the things that we focused on is can you annotate what a protein does without whining it up to its evolutional ancestors or relatives

**MAX:** So how do you do that? The method that we're using is the dilated convolutional neural network. So we took a model out of the vision literature, and what it does is it looks at a context window and tries to, it tries to look at the surrounding context in a way that is agnostic or invariant to translations, much in the same way that if you have a picture of somebody and the person's face is in the top left or the bottom right corner, it still should be classified as a picture of somebody. That's why we chose this particular neural network architecture for our problem.

**CRAIG:** Yeah. And I understand how a convolutional neural network works. But in your case, how is the data structured and that frame of the convolutional neural network, as it's moving in an image, you would be moving across groups of pixels. What is it moving across in your case?

**MAX:** So, we take in the amino acids, you can think of this as a one-dimensional image. It only has left and right. And so, we have a one-dimensional convolutional neural network, whereas images you'd use two or more, depending on if you have other inputs

**CRAIG:** and this one-dimensional string, is that coming from a database that already exists?

**MAX:** Yes. So, as I mentioned, this is not a new problem and people have recognized how important this problem is for a long time. We used a database called the Pfam database. This is run by experts at EBI, which is the European Bioinformatics Institute. And we set up training tests to benchmark methods.

**MAX:** And they have a number of different sets from hand curated to computationally annotated. So, we decided to stick with the smaller set. It's the one that a human has been in the loop for. And we wanted to see, compared to other methods that use multiple sequence alignment, can we correctly classify these sub sequences of proteins into their corresponding buckets?

**MAX:** We worked with the Pfam database, we published a pre-print, and we got a lot of feedback from the community. And though there was initial skepticism, the curators of the Pfam database ended up being our partners and are now our collaborators, the head of the Pfam database is a co-author on our paper and we have worked together to annotate even more proteins for the Pfam database.

**CRAIG:** And when you say annotating, what are the labels that you're giving?

**MAX:** Yeah. So, let's pick a protein in all of our bodies. Hemoglobin has one sub sequence that's really well conserved across a lot of different organisms. And so, we'd say, hey, this chunk of your hemoglobin protein looks like it's a globin protein.

**MAX:** And this globin is one of 20,000 that I told you before they have about 20,000 different classes of proteins. so, hemoglobin is a particular property that we predict. There are a number of proteins that bind to your DNA. Ones that as DNA is a helix, you can unwind the DNA. That's important for copying your DNA. There are a number of proteins that sort of enhance or turn down particular genes.

**MAX:** All of these are types of things that we would annotate.

**CRAIG:** You have all of these amino acids sequences in the database, there are known properties for all of those sequences or some of those sequences, or, if all of those sequences that is the training set, and then you do a test set and then once you've trained the convolutional neural network. You can, come up with a new amino acid sequence that no one has ever seen in nature or, and feed them through the CNN. And it will predict the properties of that sequence. Is, am I getting it right at all?

**MAX:** Yeah, you're getting it like 95%. The only change that I would make. So

**MAX:** We have this database of human curated, annotated proteins. We have a fairly good idea of what they do. We cut off a piece of that. Save it for our test set. We train our neural network, make sure that we're doing a good job on our test set and now there are a lot of proteins that aren't human annotated, and they aren't even computationally annotated. We want to take our neural network and see if we can predict, reliably predict, the function of these other naturally occurring proteins.

**MAX:** And so, these are not manufactured necessarily in a lab. These occur in some species. We just got a sample from Colorado Groundwater Springs, and we want to figure out what's going on in these organisms. How do they work? And this is what our neural network does.

**CRAIG:** Are you only looking at annotating proteins that occur naturally?

**MAX:** For right now, our focus has been on naturally occurring proteins. So, there are billions of proteins and for about a third, and it's hard to really put your finger on that number exactly, for about a third, we really don't know what they do. And that coupled with the knowledge that naturally occurring proteins produce a lot of very valuable things, things like penicillin, leads you to believe that there is tremendous value in solving this problem of, hey, can I predict just given the formula, the blueprint for this machine, can I predict what it's going to do?

**MAX:** And our method doesn't answer that question for all proteins, but it does answer this question confidently and reliably for quite a few.

**CRAIG:** You're focused just on annotation. You're not drilling down on any individual protein to follow up with how your work affects research or applications down the line.

**MAX:** So, the next step for our research is prospective hypothesis validation. So, we have a ton of predictions. There is no other computational support for our prediction. You just have to believe that our network is right and then test it.

**MAX:** In terms of following up with finer grain predictions, one of the things that we see is that there might be a slight modification, let's say between a gorilla’s hemoglobin and ours and if you did want to optimize or make a new enzyme that is more efficient, that gives you some information about the things that you can change without breaking the protein.

**MAX:** And while we're not specifically focused on that finer grain stuff those sorts of insights can be derived from our research.

**CRAIG:** Is this methodology for annotation - as you said, it's a problem that's existed for a long time - have convolutional neural nets been used in this way before

**MAX:** they have been used in this way in a piecemeal way.

**MAX:** One of the focuses in our research was to bridge the two communities, the bioinformatics community, and the machine learning community. And so, we really wanted to make sure that when we were testing our methods, we were first, we were testing against what is used as state-of-the-art in the bioinformatics community.

**MAX:** And as well, we were getting input from that community. And this is one of the reasons that our partnership with the Pfam database itself is so integral to the contribution that our research has made.

**CRAIG:** And how has the annotation been done in the past for proteins or amino acid sequences?

**MAX:** Yeah. So based on alignment.

**MAX:** So again, this idea of evolution, things that are related are very similar. The sequences can be lined up and there are computational tools to figure out whether or not two sequences can be aligned and the way that Pfam and a lot of other databases work is there is a subset of proteins that a human looks at.

**MAX:** So, they'll run the alignments, or they'll say, I'm sure based on reading some paper, that this is the function of this protein. They'll look at the two aligning sequences, say yes, that make sense and then apply some sort of evolutionary search based on that alignment. And what we see is that there's this huge growth in the availability of genomic and protein data, and it's ever more difficult to keep pace in your computational annotations of all this new data. And so, we see this kind of stagnation in the ability of the human curators. They don't have the tools that they need to be able to keep up with this deluge of data. And that's one of the stories here is why we think that machine learning can help

**CRAIG:** This convolutional neural networking along a one-dimensional vector, along a vector, to make classifications, you're applying that to proteins or amino acid sequences. That methodology has been used for annotation of other vectors. Is that right?

**MAX:** Yes. Natural language processing. It has been used before. I believe it's also been used in audio.

**CRAIG:** That's interesting. So, once you get this set up is it all automated. You have this massive database, and did you say not all of the database is annotated, is that right? So, is it just running 24/7 and spitting out classifications for each string? How does it work practically?

**MAX:** Yeah. So, there is a human in the loop in terms of quality control, right? And so, we worked with the Pfam database, we made all these predictions. You can even think of our method as a booster.

**MAX:** If the current methods aren't really sure. And our method is really sure then that sort of gets flagged and it gets put into this potentially promising set of predictions. And we worked with the curator to figure out settings to make sure that while we were not making a ton of mistakes, but also annotating a lot more proteins.

**MAX:** They're not looking at every prediction, but they are involved in the process of designing the parameters and in quality control.

**CRAIG:** And what percentage of the database is annotated when you started this?

**MAX:** About 75% of the database is annotated. The database does not include all proteins, but it is what's called a representative set of proteins.

**CRAIG:** You've already validated this method. So, you're past the training and test phase, right?

**MAX:** Yes. And so, in addition to training and test and making sure that all the benchmarks looked great, we also applying to a number of different circumstances in a number of different unseen test sets.

**MAX:** We also have gotten the buy-in of a number of players, a number of leading figures that our predictions are accurate.

**CRAIG:** Yeah. And so, 75% were already annotated. You're basically refining or validating that 75%, is that right? And then you move on to the other 25% that have not been annotated at all.

**MAX:** That's right.

**MAX:** And so, we annotated about 4% more and there are a number of different ways to measure this. So, each protein has just how an image might have a number of different objects in it. A protein might have a number of different sub sequences that do different things. And so, this is not just only a question of, do you have one annotation for a protein, but you have a complete picture of the different parts of this machine and how they interact. And so, in both of those ways, our method adds value.

**CRAIG:** How long would it take to complete the annotation of the entire database?

**MAX:** This method runs quite quickly. So that is another thing that we're excited to see through to fruition and.

**MAX:** The reason it runs quickly is because you can leverage GPU's. So, you might be able to annotate a thousand sequences per second, and that is just not possible with the methods before.

**CRAIG:** In a matter of months or a year or something, you will have annotated the entire database. And then what do you do?

**MAX:** Yeah, so we have annotated all that we can confidently with that particular method with those particular settings. However, there's new data coming in daily. And there are a number of refinements to our method that are quite promising. We're working on releasing the next large increase to the Pfam database sometime in the next couple of months.

**CRAIG:** And then looking ahead, do you see how this is going to be used? What the changes are in biochemistry or biomedicine.

**MAX:** When I think about the impact of this, I think that there is a core biological understanding impact, and then there is a tool building impact. And so, for the biological understanding saying something like, this organism can process this particular metabolite and it got that from this evolutionary ancestor.

**MAX:** And now we have a mechanistic understanding of what genes are involved and how this process works.

**MAX:** Now that's core basic science research. And on the other side, there's the tool building research can we take this understanding and turn it into medicines therapeutics? Or can we degrade plastics in the atmosphere?

**MAX:** Can we sequester carbon? All of these things I think are extremely exciting and we can basically leverage billions of years of experimentation by evolution to try to solve some of the biggest problems we have today. And when doing that tool building, there are two ways to go about it. One's to look into the database, say, ‘Hey, can you please give me a protein that binds to the SARS-CoV-2 spike protein, and also is nontoxic.’

**MAX:** So that's one way to go about it is just to mine a large database. And the other way is to evolve proteins or, make these changes that, that make enzymes much more efficient. And I think that in both cases, having a strong understanding and the ability to rely on classification methods is going to be a core component.

**CRAIG:** So again, since I've spoken to DeepMind about AlphaFold. You're approaching this through classification, which creates a database that people can search for particular characteristics or functions of a protein. AlphaFold is looking at the shape of the protein and predicting the shapes of protein based on the amino sub sequence, is there some point at which those two come together?

**MAX:** Yes. So, I can give you two examples. The first is when we make a prediction, and it looks kind of out there. And nobody else is corroborating it, the experts really don't know, the other computational methods don't know. One of the things that we do is we use AlphaFold to predict the structure.

**MAX:** And when we get a predicted structure from AlphaFold we can say, ‘Hey, I do know that a lot of proteins with this structure have this function.’ And so that's one way that the two come together. The other way is how might we provide value to structure prediction methods like AlphaFold.

**MAX:** One of the things they rely on is they want to look at all of the evolutionary relatives for a particular protein to understand the sites of variation. And if you can dramatically increase your ability to recall the evolutionary relatives by doing classification, that can also improve their performance.

**CRAIG:** Does this fit into a larger effort at Google? What is Google doing?

**MAX:** So, Google and alphabet is a large place, where, if a problem is important enough, it's okay to try to take a couple of different approaches to solve it. And we do know that across Alphabet, above the Google level - so alphabet is a holding company of Google, DeepMind - there also is this new called Isomorphic Labs, which aims to use some of DeepMind's technology and AlphaFold to develop therapeutics.

**MAX:** So that is not what I do, but there is interest across a number of different Alphabet properties in solving this problem.

**CRAIG:** It's basic research heading in one direction, but there's no application attached to it at this point.

**MAX:** No, but Google's mission is to organize information and there is a ton of information in proteins and there are a ton of people who care about it on the scientist level, but also my grandma had rheumatoid arthritis and proteins are the number one drug to treat severe rheumatoid arthritis.

**MAX:** So, this organization of information is important on a humanitarian level, I believe. And that is how I see my work.

**CRAIG:** How long has this effort been going on and how long do you think it'll continue?

**MAX:** It's been about four years since it started

**MAX:** I don't know how long it will continue. I think that we're moving into a new phase moving into the wet lab, but in also trying to make the information yet more accessible. Like how DeepMind released the AlphaFold results to the world and made it the website jointly actually with EBI who is our partner as well.

**MAX:** And, making a search tool on top of that, I think that all of those are promising potential future directions, but unfortunately, I don't have a crystal ball.

**CRAIG:** I'm interested in the annotation generally. Are there problems that this method hasn't been applied to, that it could be applied to that you're aware of?

**MAX:** There is work in applying these methods to DNA, instead of proteins. It is quite successful there. And predicting things like does this particular sequence of DNA, is this part of a transcriptional regulator? So, what I mean by that is, does it encourage or repress a particular gene function? I think that there is a lot of genomic information, and we will see tremendous changes over the next 10 years, even five years in terms of using the latest, greatest machine learning methods to annotate those sorts of things as well.

**MAX:** One of the things I think is really important is that it's not the case that everyone needs to be a machine learning expert in order to contribute to a project. There are chemists on this project. There are a number of different backgrounds that all contributed to making this particular effort successful and not all of them require a PhD in machine learning.

**MAX:** This was not a one-person effort. There were nine people on the paper and even more that helped us behind the scenes and across multiple different institutions. And their effort is incredibly appreciated.

**CRAIG:** That’s it for this week’s episode. I want to thank [ClearML](https://t.clear.ml/eye_on_ai) for their support and I urge you to take a moment and visit [clear.ml](https://t.clear.ml/eye_on_ai) to see what they offer. If you’re a Data Engineer, ML engineer, DevOps, or a Data Scientist, you may find that they are the solution you’re looking for.

I want to thank Max for his time. If you want to read a transcript of this or any episode, you can find a link to one at eye-on.ai. Check the archives under the podcast section for all past episodes.

In the meantime, remember, AI is about to change your world, so pay attention.