Rufus podcast - to discuss paper 'CellTypeAI Automated cell identification for scRNA-seq using local generative-AI'-20260603_133934-Meeting Recording

3 June 2026, 12:39pm

20m 39s

Paul 0:08

Great. Well, welcome to another edition of To Immunity and Beyond. And today I'm talking to Rufus Daw, who's a senior bioinformatician at Manchester, but most importantly, part of the MRC CoRE in Exposome Immunology, otherwise known as Moxie, which is a very important new initiative between Manchester and Oxford. So welcome, Rufus.

Rufus Daw 0:36

Thanks very much. Thanks for having me on.

Paul 0:39

So great. So you've got a very nice post as part of this MRC CoRE, which is really kind of, it's a very big initiative. But perhaps you could explain a little bit about first about your background and, you know, what sort of things you've been doing and what led up to this. That would be great for everybody to hear.

Rufus Daw 0:57

Yeah, sure. I mean, I mean, to really start off with, my PhD was actually in immunology. So I was a wet lab scientist for quite a long time. But then I've always been sort of interested in data science and my role started to transition more into sort of bioinformatics and data science.

I got involved with a very large consortium called the Stroke Impact Consortium between Manchester and Stanford, and that's really where I started like getting my teeth into real data science, using sort of big flow cytometry data sets to sort of understand cognitive outcomes in stroke.

That project finished and now here I am working for the MRC core.

Paul 1:44

Great. And what's the sort of general area? I mean, if you can describe it broadly, what sort of, what is the role essentially?

Rufus Daw 1:52
Yeah, so.
In part, as part of the MRC core, the role is primarily a sort of additive one to the laboratory research that's happening within the core. But then similarly, we're using kind of new nascent AI machine learning techniques to do digital twins, digital cells type techniques, and then also building what we like to call foundation models, which are these big AI models that can be used to predict loads of different things. But the big one is kind of like single cell RNA sequencing data.

Paul 2:32
Yeah, well, that takes us straight on to your paper. So that's this paper called Cell Type AI, which is your, essentially a pipeline for using AI in a single cell kind of approach. So perhaps you could just give us a flavour of what you were trying to do with the paper and why you chose that approach.

Rufus Daw 2:53
Yeah, absolutely. I mean, so as many bioinformaticians kind of know, the hardest bit is annotating cells, like identifying what cells are what in your UMAP. And often, you know, right now the like the best or the gold standard is when a human would annotate it. So you'll get like a list of genes that are up-regulated or differentially expressed in a cluster and then you'll talk with immunologists or domain experts and you'll kind of come to a conclusion as to what cell type is present within your data set. The problem is, is that these data sets are getting really big now. You know, everything's a lot cheaper. We can really push the envelope when it comes to the size of these data sets.
So it becomes quite a complicated task. You know, maybe five years ago, we were looking at, I don't know, 9000 cells, and that was kind of easy. You know, there wasn't too much complexity there. But now we're looking at huge Atlas sized data sets of 200,000 cells, 300,000 cells, upwards, right? And that becomes a really difficult problem for someone to annotate by themselves or even in a group. So the idea here is basically, take the human aspect out of it and use an LLM to do the prediction for you. And it ends up being nearly as accurate and much quicker, basically.

Paul 4:21

Great. So that was a very nice nutshell. So the, I mean, everybody who does any single cell will have come across this. And then I, sometimes you get these annotations which seem not to get the best out of the data to me, whereas you've got like such rich data and then you'll just say, oh, it's a CD8 memory cell, something like that. So it's, I guess there's a lot to be gained as well as the accuracy bit as you described. So just, and I think, I guess one issue to overcome is, as you sort of say in the paper, you don't want to just chip this all off to ChatGPT or one of these other ones on the cloud. So what was your approach then to kind of keep it all in-house?

Rufus Daw 4:54

Yeah, it's kind of a tricky one, isn't it? And you know, when we're dealing with sort of sensitive data sets, we're not really in a position where we want to send even something that's reminiscent of those data sets to cloud-based AI providers like ChatGPT or Gemini, Claude, etc. Because at the end of the day, we don't really know where that data is going and what's being done with that data. So there is a kind of contingent, I guess you would call it, of local large language models or local AI that is kind of been around since AI has been around, but certainly hasn't had the same sort of, it hasn't been as famous as perhaps your ChatGPTs. And so what we did is we essentially leveraged these AIs that can be run on your laptop or perhaps your like your high-powered computer and use them to do the inference, so the sort of guessing, rather than sending this off to ChatGPT.

Paul 6:08

Okay, and you've got a bunch of them. There's a sort of, it sounded like there was a sort of overarching system for using them, and then you could plug in different ones of different scales. And how do they compare really to something that people are more familiar with, like ChatGPT or Claude and things?

Rufus Daw 6:25

Yeah, so the scale really is in how big they are, how many, we call it parameters that they have. And the bigger the parameters, the kind of harder it is to run. So these chat GPTs, et cetera, they'll have really massive, like on the trillion scale of parameters. Whereas usually when you're running it on your laptop, it's in the low

billions. But when we did our comparisons, essentially between these different sort of tiers, these different levels, we were basically able to get an accuracy output that was as good as or better than Google's Gemini. So that is to say, with the right computational power locally, you can get the same output as if you gave it to Google Gemini.

Paul 7:13
Okay, well, that's very encouraging. There was something in the technical side which sounded important, but I didn't really, I might not be the only one, but it was retrieval augmented generation or RAG. You said that was very important, but could you explain to me and anybody else listening what that means?

Rufus Daw 7:27
Yeah, of course. So I guess how people often kind of interact with an LLM like ChatGPT or whatever, they'll type in just the question that they want to ask. But the way that LLMs are kind of moving is that we'll have this particular question and then we'll have additional information that's relevant to it. And this is what RAG essentially is. So it's basically a mechanism that inserts important, relevant information that can be associated to the answer to your question, basically. So in this instance, it's markers that delineate a cell type. In many other instances, it can be any sort of example, right? So, you know, you could ask a question about football teams and the rag input would be a list of all Premier League football teams, right? It's just this additional context, this additional information to help have a more accurate, more reliable answer.

Paul 8:18
Okay, so you're basically giving it a bit of a framework to work on and that's okay. I guess in many ways there's some bits of the sort of annotation which are actually, you know, fairly easy because they always come out a bit the same. It's just these more complicated subsets, I suppose.
So you're giving it somewhat a head start to, okay, that makes sense. Right, so you set the thing up, you had this idea that you could use AI and you could use these local models and then you tested it. Perhaps you could explain how you did that and what the findings were with that.

Rufus Daw 9:11

Yeah, it's funny, actually, the testing took about 10 times as long as building the thing. But essentially what we did was we took known data sets, so ones that had been previously published, and then similarly we took a data set that we had generated in-house, and we assessed the kind of conventional methods of cell type annotation that exist in the literature, and then brought in Cell Type AI and compared it to those conventional methods. And then of course, comparing it to things like Gemini as well to kind of assess cloud-based AI, whether it's equally as good, etc.

And so we saw across the board of tissue types, we limited it to humans just because of the scope of the work, and we were able to essentially see that, yeah, we have pretty much comparable or in many most cases, to be honest, compared to the conventional annotation methods, better annotation accuracy compared to human annotation. It was slightly tricky because with so often in like machine learning and AI we'll compare, we'll do accuracy metrics, and these are matching accuracy metrics. So, you know, did the machine learning pipeline, did the AI give you the answer A, when the answer should be A? That's sort of an accuracy metric, right? So that would obviously give 100% accuracy, like that would be accurate, that would match, right? But the problem with LLMs is there's this degree of sort of ambiguity or sort of this degree of proportion where you can have the same answer that's biologically relevant, but it won't be exactly the same word. So an example of this would be a tissue resident macrophage in the liver versus a Kupffer cell, right? Essentially they're the same thing, essentially the same annotation, but the words are different.

So we had to change this sort of accuracy metric slightly in order to be able to properly or fairly perform this accuracy metric. So we had an additional testing metric, basically, that enabled us to basically match biologically relevant labelling, as opposed to accuracy matching labelling, if that makes sense.

Paul 11:48

Yes, I think so. And but that does that mean if you put the same data in twice, it might one time give you Kupffer cell and one time tissue resident microphage?

Rufus Daw 11:58

Yep, so there's always going to be this degree of possibility or this range of

possibility. But there are metrics, or sorry, not metrics, there are methods in which you can circumvent that and they're actually baked into Cell Type AI. So there's this thing called ensemble prompting where you, it's quite simple, you just send multiple prompts of the same question, and then you take the mode response from that. And so that's baked into Cell Type AI, and you generally see a slightly better annotation accuracy when that's implemented.

Paul 12:31
Good, and and it, I guess this is a sort of not a great question, but just to get it out there, because we get hallucination all the time with the classical models. Does this hallucinate at all or is it, it won't make things up for you?

Rufus Daw 12:45
So when I first started building it, the big problem was hallucination, but there are ways that you can avoid hallucination via prompt engineering, which is the process of basically tailoring the prompt that you give the LLM to give you the response that you need. So you can actually go into the source code and see how it's engineered, and see the way in which we've done it. But we've made it very deterministic, essentially, so that there is very little hallucination. Now, of course, sometimes hallucination does occur, but in those instances, ensemble prompting should save the day. So, yeah.

Paul 13:24
Okay, because it'll just essentially take the aggregate of a number of tries. Yeah.

Rufus Daw 13:28
Yeah.

Paul 13:29
And did you get any pushback at all on this? I mean, there's a sort of, in the, you know, obviously there's great concern over the use of AI in science and how it's regulated and proper and improper use of it. So how did you kind of deal with that kind of issue?

Rufus Daw 13:46

It didn't get any pushback on it. I think just because it's a tooling system, it's quite sort of limited in its reach. Similarly, we didn't have any sort of pushback from this sort of cloud AI-based sort of area of concern because everything's being kept local. I think the only major concern was reproducibility and we just essentially had to code in methods to enable it to be reproducible.

Paul 14:17
Great. And so it's on Bioarchive. We'll put the link on for anybody who wants to read it. And how do people then use it? What's the kind of process for accessing it?

Rufus Daw 14:30
Yeah, super simple. So it's, if you know what this means, it's pip installable. So it's on PyPi. So you just type in in your command line, pip install Cell Type AI, and then there's documentation to show you how to use it, but it's what we call a one-liner. So it's not a complicated programme to get going, basically.

Paul 14:52
And then you do have a choice of which of the models that you would use on your local computer or installation, basically. OK, so people can make choices.

Rufus Daw 15:01
Yeah, absolutely. Yeah, so it's kind of enveloped around this service called Ollama, which is a library of open source LLMs. So you can use anything in the library of Ollama, which is constantly being updated as well. So we kind of expect actually the accuracy metrics to get better over time as these models progress in sort of complexity and capability.

Paul 15:24
Great, and as people started to use it already, have you got...

Rufus Daw 15:36
I think maybe when the bioinformatics paper comes out. For now, we've got like a couple of stars, but it's sort of slow going. The papers had a lot of reads though, so that's good.

Paul 15:45
Okay, yeah, well, maybe, maybe if the people listen to the podcast, it will spread the word. So, and so it seems like a very cool approach just to kind of blend the two things together where we've got a sort of essentially a, you know, traffic jam of data coming through that we don't always deal with very well and bring in the AI to kind of, as carefully as you can, help us through that log jam. Can you see what sort of where are you moving on to next with it? Because there could be lots of further downstream uses. I'm sure you've got lots of ideas.

Rufus Daw 16:19
Yeah, so I've had a lot of people come up to me and ask to implement it into spatial sequencing, to implement it into spectral flow cytometry as well. Now that panels are getting much larger with spectral flow cytometry, this is sort of a technique that can also be integrated.
So yeah, there's these sort of additional omic avenues that people are interested in me kind of expanding into.

Paul 16:43
Yeah, that sounds a lot very sensible. Again, similar problems and yeah, the spatial stuff is just going to explode anyway. So imagine it'd be really, you might, do you have to develop a separate tool for that or could you use more or less the same kind of structure?

Rufus Daw 16:54
Yeah, it pretty much stays the same. I think just a kind of a couple of, maybe a new module, a couple of tweaks here and there, but not too much work.

Paul 17:10
Oh, you make it all sound so kind of straightforward. I mean, a lot of the time people talk about these large language models as if nobody could ever understand it, and it's a kind of mysterious, but I mean, do you think that's, I mean, how, if you've made the transition from a wet lab scientist to a bioinformatician, and I've seen that happen a few times, it's, you know, incredibly valuable role, but for those of us who

are still kind of a bit stone age, how far into this can we get without having to kind of commit enormous amounts of time to retrain ourselves?

Rufus Daw 17:43
Yeah, it's interesting.
I've had a few discussions about this lately, and I think I don't necessarily think people need to retrain, but like a basic understanding of the principles is really helpful because the kind of direction in which AI integration into the laboratory is going to go, most likely, is this kind of what we call human in the loop system, where you'll have multiple AI agents doing kind of technical data management or data processing stuff, and then you'll have an overseer or someone who knows the biology, who knows the immunology, kind of ensuring that those processes are doing reasonable things, so I don't think people have to retrain, but I think that an understanding of kind of the bare bones how it works, so essentially it's a prediction model in many cases, is helpful in being able to kind of recognise why something's gone wrong when using these tools, but really in terms of like picking up and integrating it into your workflows, it's getting easier and easier just to do it without any sort of technical expertise. There's a kind of term called no code, which is becoming more and more popular, where people essentially are implementing these AIs to do different things for them. And they don't have to have any Python or R knowledge or C++ knowledge at all. They just ask sort of normal language questions. Can you do this for me? And the AIs will go away and do that for you.

Paul 19:17
Okay, so yeah, I've heard that. But you'd still, but the way you describe it, you've carefully, for example, incorporated this ensemble set to kind of minimise hallucination. So you still need to be really careful, I suppose, even if it can do a lot. But as you describe that, it sounds like you should either write a kind of the manual or, you know, a kind of some kind of training seminar for the rest of us to just get us a little bit more confident in this area. I mean, probably like a lot of things, once you get started, it's kind of actually becomes much easier.
Amazing.

Rufus Daw 20:11
Yeah, We just did a workshop, actually, so there re links in YouTube of a talk I did

introducing wet lab scientists to AI techniques, I can send you that link if that's helpful.

Paul 20:12

Fantastic. We'll attach it to the podcast. Even better. You've done it already. So that's absolutely amazing. OK, well, it's been a pleasure talking to you. And I'm so excited you're part of the MRC core / Moxie team. That's great. I'm sure there'll be plenty of questions and lots for you to do. So thank you very much.

Rufus Daw 20:16

Thanks for having me on. It's been great.