Episode 4 - David Gfeller

[00:00:00] Welcome back to Immunity by Design, where we bring together leading scientists, biotech innovators, and policy representatives to explore how emerging technologies and AI are reshaping our understanding of immunology. I'm Hashem Koohy, an associate professor of systems immunology at the University of Oxford, and I am delighted to be hosting this series.

Today, I'm truly honoured to be joined by Professor David Gfeller from the Ludwig Institute for Cancer Research in Lausanne. David leads a remarkable group working at the intersection of computational immunology, structural biology, and machine learning, with a particular focus on understanding how T cells recognize their targets at the molecular level David and his team [00:01:00] recently posted three preprints, and that sits right at the heart of what this podcast series is about.

In the first, they introduced TCR specificity profile, or TSPs, a fully interpretable probabilistic framework that captures the key determinants of TCR epitope recognition across hundreds of epitopes. Rather than chasing incremental gains in black box prediction manner, TSPs ask deeper question, what can we actually understand about why a T cell recognizes one epitope but not the other?

The answer turns out to reshape some long-hold assumptions about the relative contribution of V and J gene usage versus CDR3 sequences and amino acid composition that actually forms the, the content of their second preprint, [00:02:00] where they present a statistical platform to quantify the impact of V and J gene usage on CDR3 length and amino acid composition.

And in the third preprint, building directly on these insights, they demonstrate that native TCR alpha and beta chain pairing carries surprisingly little additional information for training machine learning predictions. And that bulk sequences of alpha and beta chain separately at a fraction of cost of single cell approaches is sufficient to match or even exceed the performance of models trained on paired data.

So together, these three studies form a coherent and timely argument about where the field's attention and resources should be. Though today we will be mostly focusing on just the TSP [00:03:00] preprint because of the time. So David, it's wonderful to have you with us today. Welcome to Immunity by Design. Thank you, Ashem.

It's a pleasure to be here. To get started, David, would you like to introduce yourself, tell us about your research interest and what drives the questions your group is pursuing?

Yes. So I'm David Khella. I'm an associate professor here at University of Lausanne at the Department of Oncology, and one of the key questions we have here is how and what can immune cells, and especially T cell, see on the surface of cancer cells? And this has been a longstanding question in the lab and with many collaborators here, we have a very collaborative environment of understanding which are the cancer epitopes

that are visible to the immune system and how these immune cells, especially T cell, can recognize their [00:04:00] epitope.

And over the last two or three years, we got much even more interested into this TCR epitope recognition specificity question, not only in terms of being able to predict, but as Ashton mentioned, in terms of understanding what characterizes and where specificity is encoded in these interactions. Yeah. So let's start by putting one step back and look a little bit into the gaps in the knowledge that led you to this tree preference.

And for that, the field has had sequence-based TCR epitope predictors for years. Tools like NetTCR or MixedTCRPred or other tools. They work reasonably well when there is plenty of training data, but as soon as you move to unseen epitopes, those that we have either none or very limited number of TCRs in the training data, [00:05:00] performance drops to almost random.

That's a major generalization problem. So some people argue it's mainly about data, too little or too noisy, while others think black box models make it harder to understand what's going on. So from your point of view, what is the underlying problem? Is it the data? Is it the model? Is it our perception to this or combination of everything?

If you had asked me this question five years ago, I would have said, "Oh, it's a data problem." Because we have data for very few epitopes, roughly one hundred or two hundred, depending on how you count. Well, we have, you know, a bit more than ten TCRs that bind to it. But a few years ago, I, I went back and I tried to do some very simple calculation about the theoretical diversity of the epitope space.

And if you do so, uh, the current estimates, if you take the peptide and the [00:06:00] MHC diversity, it's in the order of ten to the power fourteen different possibilities. On top of this, the epitopes are not much related to each other from an evolutionary point of view because they come from different pathogens, from cancer neoantigens, self-antigens.

So there's no clear structure in the epitope space besides the one imposed by the restriction to the MHC. Now, if you step back, you think we have data for roughly one hundred epitopes, and we want to build a model that generalizes to a space that is, let's say, ten to the power fourteen. So there's a ten to the power twelve orders of magnitude difference.

And I think this has been really underappreciated, that the epitope space is so vast and so complex that this idea of generalizing across all epitopes makes actually very little sense from a machine learning point of view. It's basically [00:07:00] learning out of hundred data points, a manifold in a space that has a theoretical diversity in the order of ten to the power fourteen.

And I think it's important because in my view, I may be wrong, maybe the future will prove me wrong. Even if we come up with ten thousand new epitopes, each of them with one thousand TCRs, I'm not at all sure we're gonna be able to generalize using this sequence-based algorithm. And that's the reason why in my lab, from the beginning, we decided to build epitope-specific, and we make only predictions for epitopes for which we have data.

So basically my understanding from what you said is that yes, we should accept the reality of the data, but it is basically important to put more emphasis on mechanistic understanding rather than pushing black box prediction. Is that a, a correct conclusion from your point of view? That's one [00:08:00] correct conclusion.

A second one is that we should just take the task seriously and generate data for most epitopes which are clinically relevant. And this- Right ... number of clinically relevant epitope is not so huge so far. We're gonna discover maybe new ones, but my message that I give to the community is that if you want to do a sequence-based prediction- Mm

you should generate epitope-specific TCR for as many clinically relevant epitope as possible. Then the second question is now can we start understanding specificity based on this data in terms also of visual, visually amenable models? Okay, perfect. So thinking about your TSP preprint, a key idea in that preprint is that what looks like a specificity in TCR data set is often confounded by the huge variability in baseline repertoires.

Things [00:09:00] like, you know, sequencing depth, perhaps sequencing method, patient's genetic and immune history. This seems obvious in hindsight, yet it wasn't systematically accessible before. Why do you think the field works around this problem rather than tackling it directly? What made it possible for you to address it now?

Yeah, that's a good question. The TCR specificity profile, as you correctly said, Hashan, is indeed based on this fundamental idea of comparing epitope-specific TCRs versus what we call baseline TCR repertoires, which are TCR repertoires that you find from blood without any sorting procedure. In my view, this is the correct way of computing enrichment or differences in epitope-specific TCRs, and it also mimics the realistic situation where you are making predictions out of a repertoires, and you want to find something that [00:10:00] characterizes the TCRs that recognize your favourite epitope.

Now, why has the field not looked at baseline TCR repertoire? I think this is a result of many publications that are absolutely correct that realize there's lots of batch effects in TCR repertoires. And they illustrated this, that depending on the sequencing protocols, you have different VJ usage, and this is the main batch effect actually.

You have slight difference in VJ usage depending on the age of the patient, depending on immune history, but it is actually very small. The biggest one is technical aspect of the sequencing protocols. And unfortunately, there's been a few studies that were not

fully honest that capitalize on this batch effect to learn batch effect in the-- especially in the negatives, and claim amazing performance power.

And this has been correctly pinpointed, and this has discouraged people to look at baseline repertoire. And I think it was [00:11:00] retrospectively a mistake because even if batch effect exists, there is still lots of things that are conserved in TCR-- baseline TCR repertoires. For instance, we have some V genes that are very common, five percent of the TCR repertoire, plus or minus two or three percent of the TCRs have these V genes, and we have other V genes that are very rare.

Again, similarly, the length of the CDR3, the distribution of the CDR3 lengths is something extremely well conserved across many patient sequencing protocols, different age, different disease that use different HLA background. All this is actually very conserved. So I think people were a bit obsessed by this variability in the TCR repertoires, and they missed the fact that there is actually lots of conservation, which means we can make a model of baseline repertoire that is reasonably correct.

And in the TCR specificity profile, for those who have seen the pictures, we also add these error bars which tell us a bit of the variability. We think this is something very important to consider when we [00:12:00] discuss about enrichment in V usage, J usage, or CDR3 motifs. And great. I'm talking about V and J gene usage and their importance.

One of the most striking findings in the TSP preprint is that V and J gene usage at CDR3 sequence is the dominant determinant of epitope specificity, if my understanding is correct. This seems to challenge some of the previous works where they show CDR3 or not CDR one or two is the key loop in the T cell antigen specificity.

So what do you say about that? Yeah, this is the key of the TCR specificity profile. This is really something extremely important. I should still emphasize that there have been some studies in the past that studied one specific epitope and reported enrichment in some VJ usage. So it's not that we completely discovered [00:13:00] from scratch.

The problem is that many of these studies were scattered in different places. There was absolutely no framework to vi-visualize this V and J gene usage enrichment compared to the baseline, and this goes back to the previous question, is that enrichment in VJ usage should be quantified compared to a baseline.

Otherwise, it's very difficult to know if there is an enrichment. Now, why we came up to this was some ideas a few years ago that I remember this was Christmas break. I-I had been working with the mixed TCR thread, which is a very good tool, but a black box, a deep network story, and I was extremely frustrated not to be able to understand.

And I thought, really, there must be a way we can visualize TCR specificity. And then I started working on VJ usage and developing this predictor, and the power of the TEMPO

predictor is that you can split, because it's a linear predictor, you can split to different input. And I suddenly realized that if I was to train my predictor [00:14:00] just on VJ usage instead of VJ usage plus the CDR3, the difference in AUC was statistically significant, so it's better to include the CDR3.

That was actually very modest in absolute values. And this came as a really striking demonstration that most of the specificity is actually encoded in VJ usage. I want to emphasize that it doesn't mean that CDR3 are not important. What it means is that much of the CDR3 are determined by V and J usage Now what really convinced me, and I think this is again an important contribution of the work we've done, is that we went back to existing crystal structures.

And then instead of annotating residue of the TCR as CDR1, CDR2, and CDR3, we said it's either V or J usage or the junction part. And we did this for all crystal structures, and then we counted the number of interaction with the either the epitope, so the peptide MHC or the peptide only. [00:15:00] And what came out, and this is again figure two in the preprint, very, very striking, is that the number of interaction with the epitope, whether it's only the, whether it's a peptide MHC or only the peptide, is much higher for residues that are encoded in the V and J genes than in the junction of the CDR.

So I think this is a very strong evidence that actually specificity is encoded in VJ usage because if you look from a structural point of view at the TCR epitope interface on the TCR side, the part encoded by the V and J usage makes up most of the interactions. Thank you. Beautiful. Turning to chain pairing manuscript, the motivation here is partly practical.

Pairing SL TCR sequences is, for example, expensive, low throughput, and noisy, and widely inaccessible or with some [00:16:00] limitations. Yet the idea of how much specificity information is encoded in pairing was a gray area. So what made you question that assumption, and was there any resistance to the idea that unpaired data might actually be enough?

Thank you for the good question. Uh, I, I should still start by emphasizing that if you can generate paired data, you have the money, if you manage to have the sequencing depth also, it's great. It's the gold standard. It's the best you can do Also, for many application, you need paired data. If you want to track clones, you need paired data.

If you want to use your data as a benchmarking set, as a test set for benchmarking algorithm, you need paired data. If you want to, to do many other things like TC-- if you want to extract TCR for TCR T cell therapy or for [00:17:00] therapeutic application, you need paired data. So I'm not claiming that paired data are useless.

Now, the question comes, if you plan to generate data for training predictors, can you afford to have unpaired data? And what we've demonstrated is that the performance on tools that are trained on paired or unpaired data is basically the same. Which means that if you want to generate data for training your tools, you can use unpaired data,

which is much cheaper and comes with another benefit that the sequencing depth is much higher.

Why have people failed to realize this? I've been talking to many people about this story, and many people were very sceptical. I agree. One of the confusions that people have is that they believe that if you do unpaired alpha and beta sequencing, you're actually training your model on only one chain. And this is a very often, very-- a confusion that I find very often.

The-- People tell us, "Oh, but you're only considering one chain. We know specificity is [00:18:00] encoded in both chains." And this is a very unfortunate misunderstanding that has been raised many times, so that's why I'm bringing it here, that we s-- we-- Obviously, the work on the TSP, the TCR specificity profile, also demonstrate in a crystal clear manner that specificity is encoded in both chains.

Some pe- some epitope have a bit more specificity in alpha chain, others have a bit more specificity in beta chain, but both chains are absolutely critical. What, in my view, is not completely essential is this pairing, which means the information that goes beyond what is encoded in each chain. So typically, this would be a frequency of the V alpha and the V beta, specific V alpha and V beta, that is much higher than what you expect from the frequency of the V alpha itself and the V beta.

These are these cases where specificity could be encoded in the pairing. What we've seen is that such cases exist, and I'm not saying there is no specificity in the [00:19:00] pairing. What I'm saying is that it does not help you much for the predictions. And there is a very simple way to understand this that again goes back to some theoretical calculation of the T-- the diversity of the TCRs or the size, the theoretical size of the TCR space.

If we take a lower bound, we are in the order of ten to the power of sixteen diversity. So it's really important to realize that the TCR space is enormous And this has a major concept that if you give me 10,000 TCRs, the likelihood that I have a TCR that has a very good alpha chain and a very good beta chain, but it happens that these two chains in complex do not bind to the epitope, is actually very small.

And I, I was running this week some calculations based on all the evidence we can find, this-- the probability to have such cases which are not described, whi- which could, uh, be wrongly described by the unpaired data, namely the [00:20:00] presence of a good alpha chain and a good beta chain, which together do not bind to the epitope, is less than zero point one percent of the TCR.

Which means if you give me 10,000 TCRs, I'm making 10, roughly less than 10 mistakes. Now, if you think of the states of predictions, we are making many more mistakes. So that's the main reason why learning information in the pairing has not-- is not helping much for the predictions. It may at some point, once we reach AUC of point ninety-eight, maybe the paired information can bring us to point ninety-nine.

But we are now at point seventy-five, and there are many other things that we need to solve before we can really truly benefit from paired data. So again, paired data is the best you can do. I love them, but sometimes it's better to generate a lot of unpaired data, which is much cheaper than limited amount of paired data.

Thank you. [00:21:00] Very clear. So these, uh, three studies are highly interconnected. The Tsv study shows the importance of baseline data. The second sheds light on how CDR3 length and amino acid composition is shaped by V and J gene usage. And the third looks into the amount of information encoded in chain pairing.

So the question here is that was this three planned as a part of a bigger project, or they just came along when you were working on one of them? So how was the story? Each, each study has a initial story. What is it? Yeah, that's a good question. It's always interesting to, to redo the, the history. Mm.

There's been different things that crossed my mind, but I think the key to everything was this ability to visualize, to develop this TCR specificity profile, which can be thought of as a motif TCR. [00:22:00] And suddenly, when you start visualizing things, and if you train yourself at reading these motifs, you can ask plenty of questions.

And for instance, the fact that the predictor that is built on this motif had similar predictive value as, as our, as the deep learning method, including our own mixed TCR pred. So in this comparison, we were quite sure we were not unfair to the others because we were beating ourselves. Uh, thi- this resulted in the second question that actually the information in the pairing is not so important, and this I realized very early on actually, and it, it actually saved us a huge amount of money.

I mean, at least $100,000, actually much more, I think, in, in the different experiment we've done. So this was a critical step to, to realize this. Thank you. Now let's move on to the study. There are a number of findings and observations that you made, although we wouldn't have time to go through all of them, but just a [00:23:00] few that over time earned us.

So to begin with, the TSP framework relies on several key design choices. For example, building baseline repertoires across many donors and studies, modeling CDR3 probabilities, condition on the J gene usage, and separating a specificity signal layer by layer. So these feel very natural in hindsight, but I imagine there were alternative approaches you explored and discarded.

Could you walk us through the decisions that were most critical for making TSP work? And in particular, I am interested in the decisions behind the dataset you developed Yes. So wh- when I-- so first of all, to really put this in a historic, uh, historical perspective, we started working on TCR peptide recognition six years ago, and we went like everyone, CDR1, CDR2, [00:24:00] and CDR3, build some kind of deep network, attend autoencoder, and try to, to, to, to train these models.

What really then made the TSP possible was the, this obsession that I had two and a half years ago that we must be able to understand why and what makes TCR, makes a TCR recognizing a specific epitope. So I somehow this to put myself a challenge and said, "I want to develop a framework to model TCR peptide recognition that is fully interpretable.

Every coefficient must be linked to something that we can visualize. And then it must accurately capture the specificity, since when we look at it, it should not be wrong, and it must enable comparison to a baseline repertoire." This comparison to the baseline repertoire dates back to work that we did on peptide MHC interaction prediction, where I had realized a long time ago that it's absolutely critical to have an idea of what your baseline is.

And this is even more [00:25:00] critical when you work with TCR because the baseline is very heterogeneous. As I said, we have some V segments that are very common. More than five percent of the TCR have these V segments, and we have other V segments that are very rare. Now, this is absolutely key to consider when you model TCR epitope recognition.

So these were the key ideas. And then there's also this idea to, to just think a bit theoretically of where information is encoded. And if you think of CDR1 and CDR2 and CDR3, which is what most people are looking at, it's a complex way of representing V-usage. For instance, CDR1 and CDR2 are fully determined by V-usage.

So instead of having, I don't know, 15 residues positioned with each 20 amino acids, so this becomes a very complex space, why not having only one variable, V-usage, which has an alphabet of size 20-- of size 50, roughly, and that's it, and it captures everything. So this was really this idea of reductionist, and certainly my, my training in physics contributed a lot to this, and this idea of first order [00:26:00] approximation of a log likelihood to make something as simple as possible.

And then somehow it was a big intuition, I think, on my head, as I said, two and a half years ago on a Christmas break, that at some point we can understand TCR epitope recognition. And since then, it's been really a passion that I've had. Actually, the TSP code is something that I wrote entirely myself.

Excellent. That's, that's wonderful that you are still hands-on work along with your group members. So the next question is, in the TSP paper, you also introduced TEMPO, the probabilistic or statistical model. Firstly, I would like you to intuitively introduce TEMPO. What is it? And then my understanding from TEMPO is that it's essentially a log linear model where every coefficient maps directly back to TSP-- to a TSP feature.

So yet it matches or outperform tools like Mix TCR Pred or Net TCR or tool across several [00:27:00] benchmarks. From a machine learning perspective, that is quite striking. So the question is that, what does this tell us about the nature of signal in

current TCR epitope data sets? Is it mainly data limitation again or feature engineering issue or something more fundamental about the biology?

Yeah, good question. For us, Tempo was very important, and the first reason is that we wanted to make sure we can trust the TSPs. And I think the best evidence that you can trust what's on your screen is if you train a predictor that where every coefficient, as you said, Ashan, is actually the ratio between the frequency and the epitope specific versus your baseline repertoires, and if this predictor has very similar predictive power as other machine learning tools.

Because if we had much lower predictive power, this would have been a clear indication that we are missing something in the TSPs, and that you should be careful. Now, it turned out, [00:28:00] and this is very established now, that Tempo was at least as good, if not better than other tools. We can discuss why it's better than many other tools.

Many other tools are only considering some, some part of the TCRs, like only the CDR3, and this creates lots of issues. One of the other big advantage of Tempo is about the negatives, because instead of trying to give some negatives, as you do in a standard neural network, in deep network, in all existing tools, with Tempo, we actually have an analytical model, probabilistic model of the entire TCR space.

And I think this is a major advantage because again, think of the size of the TCR space, ten to the power of sixteen at least. I personally think it's ten to the power of twenty-six. How can you imagine that a few hundreds of negatives can correctly capture this huge diversity of your negatives? And the big advantage of Tempo, and this is the way why I design it, this is the reason why I design it this way, is that instead of providing negatives when we train, we only provide [00:29:00] positive, and then we use our analytical, mathematical probabilistic model of TCR repertoire to normalize the coefficient in Tempo.

And in this way, we have a much better sampling of the-- or much better representation of the huge, enormous size of the negatives. And I think this is an important aspect of Tempo, of much better dealing with the negatives. And then the final reason why it works as well as other deep learning is that most of the information is about TCR epitope recognition, recognition specificity is encoded in V usage, J usage, CDR3 lengths, and CDR3 amino acid.

And even if you step back, it's actually not so easy to think where else would specificity be encoded. If it's not V, J usage, CDR3 lengths, and CDR3 amino acid, where is specificity encoded? It's not so easy to think about it. There are some aspect, and with the pairing we can discuss, but it turns out that these parameters, which correspond to [00:30:00] the first order approximation of the log likelihood, are very good.

And this is something we've seen in many cases in, in, in more physics-based system that the first approximation is, is not bad Very good. One of the very likable and

interesting part of your study from me, for example, a reader, is the cross-react-- the patterns of cross-reactivity using yellow fever variants, where you build a very rich data set around yellow fever epitopes, covering many sequence variants, length variants, MHC variants, and show that Tsp dissimilarity strongly correlates with cross-reactivity.

The resulting mixed TCR cross model is quite simple, if I understand it correctly again, and that means conserve non-anchor residues from position, positions from three to five preserve MHC TCR interface. So was that simplicity [00:31:00] something you expected or was it something-- What other people say about that?

Yeah. A little bit elaboration about that part. Yeah. So for the first question about the variant of the yellow fever, thi- this goes back to this question, can we generalize this TCR epitope prediction? And as I said in the beginning, I personally think we'll never be able to generalize using only sequence-based approaches to the whole space of epitopes.

But still, I want-- I thought my-- I thought let, let's try to see if we can at least learn something how to generalize across small-- across, yeah, throughout small changes in the epitope. And that was the motivation to, to study these variants where we modify the amino acid of the peptide, where we change the length of the peptide, where we change the MHC restriction because these are the important parameters to start exploring a bit how we can extrapolate or we can map the TCR specificity profile across small variants.

And we really wanted to do it [00:32:00] using the TCR specificity profile, not using a single TCR, because whatever you learn from a single TCR, you never know if it applies to the next TCR that binds to the same epitope. So we thought we now with that we have these TSPs, we can do it, and we did plenty of variants, and we see that whenever there was cross-reactivity, it also means that the TSP was conserved.

And then we stepped back and we thought, okay, now we've done all these variants, and this is almost fifty different variants, length, amino acid, re-- changing residues or changing the HLA restriction. Do we find some general pattern that maybe are conserved across many epitopes? And this is really the motivation for the mixed TCR cross, because what we observed is that, for instance, in the yellow fever variants, there were some positions that were extremely sensitive to small changes in amino acid, while other positions were extremely permissive, so position in the epitope.

And we on purpose designed a very general and simple model, [00:33:00] not aiming at having the best predictive power for the yellow fever. And there are ways, and we are now working on it to have better predictive power with yellow fever, but having a model that captures something global. And again, it has to do some with some structural guidance where we know that residues in the middle of the epitope are much more likely to be in contact with the TCRs compared to residues at the beginning or at the end.

So we just put this in a few simple rules And again, I'm stressing, if you have specific data for your favorite epitope, like X scan data, for instance, you can do much better than this simple model. What this simple model has as an advantage is that it captures some common patterns that are likely conserved across many epitopes.

And one thing that we realized is that the position four and five in the epitope, at least for class one epitopes, seem to be very important. And then we applied this little rule to other cases like the MEJ3 and the titin, and it turned out that they were fitting very well, even though the titin and [00:34:00] MEJ3, for instance, are restricted to different HLA.

So this rule seems to be applicable across many, uh, different contexts. Uh, thank you. Speaking of mixed TCR cross and cross-reactivity, another very interesting result in your study is that mixed TCR crosses scan of the human proteome for MEJ-003 cross-reactive peptide, where the titin peptide known to cause lethal cardiac toxicity in a TCR therapy trial comes up as a third, third hit out of literally millions of candidates.

So was that analysis done prospectively as a validation or retrospectively once the model was in place? And how do you think about the translational and regulatory implications of having such a tool in hand? So the full story between-- behind the MA-MAJ3 and titin is interesting because we had developed [00:35:00] this, or we were developing th-this model based on the yellow fever.

And initially, I thought maybe it applies only to class one, to HLA-A2. But anyway, w-- I had not looked at the titin and the MAJ3 case, and I was sitting in a seminar, and someone was showing these sequences, and then I realized, oh, that's interesting, because the homology or the... It's not a homology, but it's just the conservation between MAJ3 and titin is this three amino acid DPI, so aspartic, proline, and isoleucine at position three to five.

And this was matching exactly what we saw in the yellow fever data, that position three to five are actually very important for TCR epitope recognition. And if you change amino acid there, it turns out that this will break TCR recognition. While if you change amino acid at position eight, for instance, in many cases it will not break TCR recognition.

And then I did this little exercise of checking, starting from the MAJ3 and checking which are the peptide that [00:36:00] have this DPI motif in the middle, and they also predicted to bind to HLA-A zero one, zero one, which i- which is the HLA. And it turns out that titin is among the very few. There's only two or three other peptide that have this DPI and have amino acid that are compatible with binding to HLA-A zero one.

And this was really impressive that rules that we were learning from yellow fever, A zero two zero one epitope applied strikingly well for this case, which is a different HLA restriction and a cancer antigen. Thank you. Yeah, we can actually speak quite a lot about these findings in these three studies, but I would like to switch the gear and look into a number of implications of these studies.

And with that, to start with, the TCR Motif Atlas now covers hundreds of epitopes, mostly HLA class I, and is still biased toward a relatively small set of [00:37:00] well-studied antigens. With SecTCR providing a cost-effective way to generate training data, you are now in a position to expand this systematically.

So my question is that do you have any prioritization in place? Are you going to start with tumor antigens, autoimmune targets, or infectious disease antigens? What are the plans, uh, looking forward?

Yeah. So we are definitely working on already now the new version of the TCR Motif Atlas. So far, we took the strategy of just collecting, putting epitopes for which we have data that seem reliable. And we've done lots of curation, actually. Even the last two weeks, I spent many hours going through every single motif from every single study for every epitope and, and checking that they are consistent, that we have no evidence of major contaminations.

This has [00:38:00] been so far the driving force. It's feasible if you have something like one hundred or two hundred. I think we're at a hundred and fifty now epitopes. Now, of course, in the future, if we start collecting data for thousands of epitopes, it will become difficult. I think our criteria for pri-prioritization will be the clinical relevance of the epitopes.

And in a sense, if it's a rare neoantigens that you find in only two patient in the entire world, boom, maybe it's not super important to know exactly how TCR recognition works. If it's a dominant shared antigens in cancer, then it's much more interesting because you can have-- you can think of developing therapeutics that apply to many different patients Uh, yeah, thank you.

Also, you basically spoke about image A.3 and Titan example, but I would like to go back to it from more clinical and implication point of view. And for that, the next TCR cross, especially in the image A.3 Titan example, could [00:39:00] potentially be used as a preclinical screening tool for TCR-based therapies, flagging cross-reactive self-peptides before candidates reach to the clinic.

So what would you-- what would a clinically deployable version of this look like? What kind of validation would be needed to give regulators and clinicians enough confidence to act on it in its predictions? Is there something you are actively pursuing, perhaps with industrial partners? Uh, that's a big question that we so far do at the academic level.

I, I-- we don't have the full answer now. O-one of the big challenges that the human peptidome is big, so you have ten million different peptides. Mm. So even if you have a very good tool that gets you an AUC of zero point nine nine, still in the top ten predictions, you will most often have zero with zero positives.

So that, that's a very big challenge because in order to be deployed to the clinic, we need these tools to reach an [00:40:00] accuracy that is outstanding just because we are screening ten million pep. Uh, so we are working now on models that can be fed with many different types of data, whether it's X scan or phage display.

There's plenty of interesting project, but we are not there yet. I would never claim that any of the tools that we are working on now is ready for clinical deployment. There's a hope that we can at least narrow down the list of potential off-target peptide and exclude them experimentally. I still see it more as a screening tool than a really something you can trust, uh, go in the clinic without even validating by, by experiments.

Okay. And moving to chain pairing preprints. There you have a very strong case that chain pairing information is not currently necessary, partly because datasets are too small to learn interchain correlations. [00:41:00] Is there a threshold data se-- about dataset size beyond which the chain pairing becomes informative?

Yes, that's a very good question. Which could be re-reframed as if we had 10,000 TCR per epitope, maybe then it's much better to have paired data because this amount of data certainly helps learning the correlation. And that's a very important point, is that one of the reasons the chain pairing doesn't help at all, at least in our hands, is that we just don't have enough data to learn the pairing.

Even with 100 TCRs, it's very difficult to learn the pairing and, and some tools they claim to have learned the pairing, but you quickly fall into the overfitting. Is that you see some pairs and you think they're statistically enriched, but actually they're not statistically enriched. They were just by chance paired in this way.

It is one possible solution. I still believe that the explanation about the size of the TCR repertoire is, is very relevant, and [00:42:00] that even if we have a case where there is specificity in the pairing and we are able to learn it correctly, in practical scenarios where you give me 10,000 TCRs, the likelihood that I have one of these One TCR that looks good from each chain, but actually the two, two chains together would not pair is very small because the TCR space, space is so big.

Now, as I've mentioned earlier, maybe the pairing will be what can bring us from AUC of .98 to .99. This is a possibility I'm not excluding. As I've stressed also, I'm never claiming that we should not generate paired data. I'm just claiming that if you have not infinite resources at hand, it's better to go with unpaired data but cover more epitopes and sequence deeper than going with paired data.

But it, it could be that the pairing-- But what I'm saying is that for sure it will not [00:43:00] re- it will not represent a massive step in AUC increase, the pairing. This, I have extensive experience with this. It may bring us from .98 to .99, but it will not bring us from .70 to .98. This is for sure. This I'm very convinced about this.

Thank, thank you, David. My next question is again about TSPs, and my apologies that I'm basically going back and forth about these studies. I should have organized my questions better, but I can only, you know, make an excuse because these studies are very interconnected. So my apologies for that. The question is that TSPs currently rely on experimentally defined epitope-specific TCR data sets.

But in the clinic, a key challenge is predicting immunogenicity for tumor neoantigens where no pa- prior T cell data exists or antigen information exists. Your observation [00:44:00] that AlphaFold 3 can approximate TSPs from ex vivo repertoires suggests possible approach. For example, scoring a patient's own TCR repertoire against candidate neoantigens to build pseudo TSP and infer immunogenicity, if my understanding correct.

Please correct me if I'm wrong. Is this something you are exploring and what do you see as the main obstacle here? So we are exploring this definitely. I can't tell you everything. What I can tell you is that sometimes it gives very nice results, but sometimes it completely fails. And we have unfortunately evidences where old alpha fold prediction were completely in the blue when the experimental data were very robust, done by many different lab, so unlikely to be a problem with the data.

So far it's true that for neoantigen it's challenging because it's very difficult to get TCRs that are epitope specific from donors, [00:45:00] and sometimes you get two or three, and then with two or three it's very difficult to build a TSP. So that's ongoing work, not full solution. I'm not sure anybody has the full solution.

Thank you. Uh, yeah, and I would like to finish off this question regarding three studies with a broad question. The TSP framework is explicitly designed as an interpretable alternative to black box deep learning. In practice, does interpretability change how science is done? Do immunogene- immunologists use TSPs to generate mechanistic hypothesis or most users still primarily focus on performance?

So I'll answer from a very personal point of view. For me, these TSPs have been a, a completely life-changing event in a sense that suddenly we can, in one second look at some TCR data set and [00:46:00] understand, is there a specificity or is it just noise or just no specificity? Uh, you bring two studies that study the same epitope.

Two seconds, you build your TSPs, you check. If the TSP is completely different, then there is something wrong, because you cannot have... And motifs don't lie, as I like to say, in a sense, if the TSPs that are supposed to be the same are completely different, then there's something that happened. We don't know what.

Maybe some study was wrong, maybe there were diff- totally different conditions, but this is a sign that the data suffer from some limited reproducibility. And it's remarkable because good studies, the TSPs are very reproducible. And i- it's, it's-- for me, it's a game changer because instead of sitting in front of these huge Excel files where we

don't understand a single word, we have all these CAS motif everywhere, this trav, TRBV, this is terrible.

And especially because the, the V gene are named differently across different protocols and like this. So the TSP is really a uniform. It's providing a framework to [00:47:00] unify all this analysis and to immediately realize if there is specificity, if... what the data quality is. And then on top of this, you can ask question, where is specificity encoded?

Is it alpha? Is it beta chain? Is it V usage? Sometimes it's very strong J usage, in other epitopes, no specificity in the J usage. Uh, this is extremely useful to guide structural understanding or hypothesis generation. So from a really understanding, deep understanding of TCR epitope recognition, I believe that the TSPs are a major contribution.

But again, I'm very biased. I'm obsessed by my TSPs. So you, the readers of the preprint need to make up their own mind. No, thank you. It's a great, actually, study, and I personally liked it very much. So David, we can actually talk about each of these studies in one full episode, but I'm afraid we are reaching to the end of episode, and we should move on.

And what I would like to do for the last few minutes of this conversation is to talk a little bit about the [00:48:00] importance of positive research culture and multidisciplinary research. Your group sits at the intersection of computational immunology, structural biology, machine learning, exper- and experimental work.

The breadth looks impressive, but multidisciplinary labs can easily become siloed people exchanging results rather than truly thinking together. So the question here is that from your point of view, what does real collaboration look like in your group day-to-day? And have there been moments where the gap between computational and experimental thinking was genuinely hard to bridge, for example?

Yeah, these are always challenges that we face. For a few things, I don't know if it's the right order to see... to say it, but cultivating good relationship with people that, that have different expertise than you is, is fundamental. [00:49:00] Uh, in my case, nothing of this would have been possible if we didn't have a good relationship with collaborators.

Here in Lausanne, I'm naming some of them, Michal Basanis Sternberg, Alexandre Harari has been incredibly helpful for the whole TSP project. Steve Dunn, Nathalie Huefer, these are all wonderful colleagues. And if you have a chance to be in such a department which includes different expertise, I think really cultivating this close relationship and making efforts to understand the language of how people think, making efforts to build collaborations, to help people, and then also to come with people with, with good ideas to design yourself experiments has been very useful.

Another very important thing is to manage to have a research line in the lab. This, I think, uh, is a game changer. Not that you cannot do anything. Colleagues that work on very different topics, uh, and they're very smart people and they do great work. But I think having a bit of this really research line [00:50:00] helps a lot.

And certainly, the TSPs in my lab became, over the last two or three years, became the main research line, and this helps to make a lab stronger than the sum of the individuals. Yeah, perfect. I, I should actually take this moment and say that hats off to you, your team members and your collaborators because you guys have been very productive, very i-impactful, and I've always enjoyed following what comes out of your labs.

So the next question is that, speaking of collaboration, how do you measure the success? What is the metric for success in a multidisciplinary project? Some people might say that, okay, a publication is actually a good metric for that. But the problem is that because many people are involved, how you are making sure that every person gets enough of credit, uh, for the contribution they have [00:51:00] made

Yeah, this is always a balance between fully collaborative. In my lab, I'm always trying to make sure that at least at the PhD and postdoc level- Mm ... everyone has his or her own project. And sometimes I know that for some PhD students, it's much more tempting to start working on many different projects, collaborate to help everybody, and they get lots of thank you after this.

But as I still am a bit strict on this, that each should have his or her own project, and eventually this means writing a paper about this. And where the paper is published is depends on the quality of the paper, but also on many other factors we will not enter into this very lengthy discussion. But yeah, this is not only the impact factor of the journal, but it must be a standalone paper.

This is very important. Sure. Yeah. Sure. And, uh, talking about the outcome of your team, I have also noticed [00:52:00] that many of your PhD students or postdocs come from either a computational or biological background and have-- They, they have to stretch significantly into other areas during their training. So two interconnected questions.

What are your strategies? What are your supervision, basically secrets, to make successful people in your lab, and how you are basically making sure or providing the training that they understand each other's language and, and collaborate, uh, and, and not rather compete with each other? I don't know if I have a, a secret recipe for this.

Yeah, I think having a, a research line in the group helps. Mm. Because somehow we're all pulling at the same string from different angles maybe, but at least people feel a sense of, of we're working towards the same goal. So th- this can definitely help. In terms of, of training, I guess we should ask my [00:53:00] PhD students if they ever listen to this podcast maybe.

They will disagree with you, Hashem. Yeah, I'm, I'm trying my best. I don't know if I'm always successful. Certainly we have things that we are better off and things that we are more challenged. I, myself, coming from physics, then I moved to computational biology, and now in the Department of Oncology, it's a very experimental department.

I, I certainly learned how to do this transition, and I'm certainly better at accompanying people that come from a similar background as me from, you know, math and physics and computer science towards application in biology. But- Talking about PhD students and PhD programs, each PhD program involves a lot of close ends that may drive the student into a low nothing is working for me mode.

How do you handle this, and what is your advice to students sitting at the interface? Sorry, I didn't get the- Yeah, the question here is that PhD programs is- Sure ... basically [00:54:00] involved with quite a lot of ups and downs. It's not a linear A to B- Path, yeah ... path, and the, the nature of it, the important from my point of view, I think, is that PhD student to understand the whole thing is about the process, not just the outcome.

Yeah. Meaning that there will be quite a lot of failure, but those failures are educational. But those PhD students at their early program, they take sometimes personal, they become low, they think that nothing is working for them. So the question is that, what your advice would be for them? Yeah, this is a tough question, of course.

Sometimes if you show some enthusiasm yourself, of course, this can help, uh, to maximize, help boost a bit the motivation of the student. Although, you sometimes have also f- you face situation where people are really discouraged a- and maybe they are not meant for PhD. Uh, that's-- There are plenty of other things in life, and a PhD is a very lonely endeavor.

And you're right, there are ups and downs, [00:55:00] and how you learn to, to go through them is certainly one of the things you should learn in your PhD. And, uh, yeah. So bringing some enthusiasm, being ready also not to harass them and let them, if a week they are not very productive, it's not catastrophic. I, I myself, I'm not productive.

There are many weeks where I'm not productive, and that's how it is. But also setting milestone and deadlines because you can't let people just walk in, in that direction forever. Yeah. Sure. The next question is a slightly different, but it's still under the umbrella of positive research culture. It's about open access data sharing.

All three of these studies that we discussed today, they have been released as preprints with code and data openly available. In a competitive field, that's not trivial choice, especially for early career researchers who rely [00:56:00] on establishing priorities. So how do you navigate that tension, and do you think the culture of open science in immunoinformatics is where it should be?

So for me, it's even an ethical question, and I'll, I'll take it on a even bigger perspective. I'm part of those people that believe science has something to bring to the society. Not every single paper that we publish this idea will bring something to the society, but globally it has something to bring, and we've seen this in history.

Vaccines came from scientists, and we have many other examples like it. If you think what's the budget worldwide of one year of science, this is hundreds of billions of dollars. And if you look at the way we disseminate the knowledge through the publication system, very often it takes a year between the time the experiments and the projects has been finalized and the paper is half-written to the time it becomes [00:57:00] public, it is accepted and available in a journal.

So somehow is it ethical to block all these results that have been paid by taxpayers' money? I'm wondering. So I've decided that go with the preprint. I'm not saying that we put everything on preprint immediately before submitting. Sometimes we also submit and see what happens. But I personally think it's much better.

It will also give a bit less influence to the power of the journals because things are out, and we've seen more and more-- I think if you go to computer science or physics, for instance, the preprint culture is there for many decades. But in biology we see this also, even in the field of TCR, if you think of, of some beautiful studies like the TurtleSeq from the group of Paul Thomas, which is a game changer in terms of sequencing paired TCRs actually.

Very good news that we'll have more paired data. Mm-hmm. This was available as a preprint roughly a year before it got published, and dozens of lab around the world have used the technology, have [00:58:00] implemented it locally, start generating new data, and this can guide, even sometimes have clinical application.

I'm thinking of a, of another very nice work of the MakeTCR from this group in Germany, in Heidelberg. Again, a preprint since a year now. I hope for them it gets published into a very good journal. But I have colleagues here in Lausanne that are already implementing the technology. Now we are testing it.

Soon we'll be able to accelerate science. So I think it's a, it's a very good culture, and the more people take it, the less risky it becomes. Because if you are the only one to put your stuff on preprint, of course, the risk is that it will not be seen and people may even be dishonest and try to steal it.

If half of the people are doing this, and we see this more and more in biology, and this is very good news, then it becomes like a publication True. Very true. And my very final question is about AI. Um, as we both know, AI is reshaping all aspects of our lives, including immunology, at an [00:59:00] unprecedented pace.

Mm-hmm. So the question is that, how are you transforming your lab to keep the pace with AI and the changes that AI introduces into the way we do science?

Yeah, that's a big question, and I'm not sure I'm doing right. What I keep telling people is that if the only thing they do is to implement some kind of AI-based algorithm, PyTorch or Keras, TensorFlow, and to click to validate, it's a bit risky because they will be soon replaced by AI agents. A- and to some extent, if you think of the TCI epitope field, uh, you can dump all the VDG database into ChatGPT, ask to train a predictor, they will not be too bad then.

So for me, that's a risk if you are just driven by, "I'm gonna implement the fanciest AI," because many AI agent will be able to do it [01:00:00] better than us. That was part of the motivation also to really start thinking of interpretability, where I maybe naively still think that we human have something that to do that is maybe not so easy to do by machines.

Maybe I'm wrong, I-- Things are changing so fast. Now, this opens up also many interesting opportunities. We've seen this with AlphaFold, which is, in my view, the best example of extremely successful AI application in biology, and now we start seeing predictive power for these unseen epitopes, at least for some unseen epitopes, and this is very spectacular.

I'm completely baffled by some of the results with AlphaFold. So it's a very nice way where the latest AI developments have been, uh, tremendously useful. Thank you so much. Thank you for your time, David. Is there anything that I haven't covered and you would like to finish off?

Briefly, I've seen many studies as I-- this goes back to what I just said, that, that are, are just delegating everything to, to some deep network or like this. [01:01:00] I made a little call to, to be careful with this and to not fall into this complete cognitive surrender where we delegate all the tasks to, to the machines.

We as human can still use our brain trying to understand, and this has impacts for designing tools also. If I go back to the TSP work, I see dozens of studies that only look at CDR3 when they analyze TCR repertoires, and it's a pity because you are missing, I would say, more than fifty percent of the information.

So it's an encouragement to people to still start try to think carefully and not delegate entirely their work to, to AI, even if their AI is doing a very good job. Thank you so much, David. I hope our audience also enjoyed as much as I did. It was wonderful talking to you. And for the audience, I will be printing the link to these three preprints in the description of this podcast to make them easier to have [01:02:00] access.

Thank you so much, David.