WEBVTT

NOTE duration: ”00:21:03.872”

NOTE Confidence: 0.9839767

00:00:00.000 --> 00:00:01.220 We have Maria Martinez,

NOTE Confidence: 0.9942697

00:00:01.760 --> 00:00:02.399 from the,

NOTE Confidence: 0.92699486

00:00:02.879 --> 00:00:04.980 the data's department of, biomedical

NOTE Confidence: 0.92699486

00:00:05.040 --> 00:00:06.580 informatics and data science.

NOTE Confidence: 0.96224076

00:00:06.960 --> 00:00:08.480 So Maria just joined Yale

NOTE Confidence: 0.96224076

00:00:08.480 --> 00:00:10.080 recently, a few months ago.

NOTE Confidence: 0.96224076

00:00:10.400 --> 00:00:12.000 She initially, again, came to

NOTE Confidence: 0.96224076

00:00:12.000 --> 00:00:13.325 biology from physics,

NOTE Confidence: 0.93388927

00:00:13.705 --> 00:00:15.385 and completed a a postdoc

NOTE Confidence: 0.93388927

00:00:15.385 --> 00:00:16.825 in computational biology as the

NOTE Confidence: 0.93388927

00:00:16.825 --> 00:00:18.525 Weizmann Institute and then Columbia

NOTE Confidence: 0.93388927

00:00:18.585 --> 00:00:19.085 University.

NOTE Confidence: 0.96782345

00:00:19.785 --> 00:00:20.904 And then before she came

NOTE Confidence: 0.96782345

00:00:20.904 --> 00:00:22.445 to Yale for a decade,

NOTE Confidence: 0.96782345

00:00:22.744 --> 00:00:23.704 she led a group at

NOTE Confidence: 0.96782345

00:00:23.704 --> 00:00:25.545 IBM Research in Switzerland where

NOTE Confidence: 0.96782345

00:00:25.545 --> 00:00:27.305 she focuses on developing machine

NOTE Confidence: 0.96782345

00:00:27.305 --> 00:00:29.480 learning approaches for cancer personalized

NOTE Confidence: 0.96782345

00:00:29.620 --> 00:00:30.120 medicine.

NOTE Confidence: 0.9512908

00:00:30.900 --> 00:00:32.340 And, and I know Marie

NOTE Confidence: 0.9512908

00:00:32.340 --> 00:00:33.620 has been now working on,

NOTE Confidence: 0.9512908

00:00:33.780 --> 00:00:35.140 the immune system focusing on

NOTE Confidence: 0.9512908

00:00:35.140 --> 00:00:36.500 looking at molecular recognition and

NOTE Confidence: 0.9512908

00:00:36.500 --> 00:00:38.120 TCR specificity prediction.

NOTE Confidence: 0.98553604

00:00:38.500 --> 00:00:39.000 So

NOTE Confidence: 0.7589726

00:00:39.540 --> 00:00:40.040 yeah.

NOTE Confidence: 0.92473376

00:00:41.905 --> 00:00:43.185 Thank you, John. It's great

NOTE Confidence: 0.92473376

00:00:43.185 --> 00:00:44.465 to be here. And, John

NOTE Confidence: 0.92473376

00:00:44.465 --> 00:00:45.425 said, I just joined,

NOTE Confidence: 0.9714835

00:00:45.985 --> 00:00:47.265 Yale five months ago, so

NOTE Confidence: 0.9714835

00:00:47.265 --> 00:00:48.385 it's a great opportunity to

NOTE Confidence: 0.9714835

00:00:48.385 --> 00:00:49.105 be here and tell you

NOTE Confidence: 0.9714835

00:00:49.105 --> 00:00:50.225 a little bit about what

NOTE Confidence: 0.9714835

00:00:50.225 --> 00:00:51.025 the work my group has

NOTE Confidence: 0.9714835

00:00:51.025 --> 00:00:51.765 been doing.

NOTE Confidence: 0.854697

00:00:53.025 --> 00:00:53.525 So,

NOTE Confidence: 0.97287065

00:00:54.590 --> 00:00:55.550 the thing the part I'm

NOTE Confidence: 0.97287065

00:00:55.550 --> 00:00:56.430 gonna be telling you here

NOTE Confidence: 0.97287065

00:00:56.430 --> 00:00:57.309 is about the work we

NOTE Confidence: 0.97287065

00:00:57.309 --> 00:00:58.350 have been doing for the

NOTE Confidence: 0.97287065

00:00:58.350 --> 00:00:59.629 last years to model the

NOTE Confidence: 0.97287065

00:00:59.629 --> 00:01:01.010 binding of T cell receptors.

NOTE Confidence: 0.9112998

00:01:01.390 --> 00:01:02.269 And I guess here I

NOTE Confidence: 0.9112998

00:01:02.269 --> 00:01:03.070 don't need to make a

NOTE Confidence: 0.9112998

00:01:03.070 --> 00:01:04.670 big introduction about T cells.

NOTE Confidence: 0.9112998

00:01:04.670 --> 00:01:06.190 There is a huge interest

NOTE Confidence: 0.9112998

00:01:06.190 --> 00:01:07.390 now in modeling the the

NOTE Confidence: 0.9112998

00:01:07.390 --> 00:01:08.509 predicting the binding of T

NOTE Confidence: 0.9112998

00:01:08.509 --> 00:01:10.425 cell receptor receptors because of

NOTE Confidence: 0.9112998

00:01:10.425 --> 00:01:12.284 the huge medi biomedical applications

NOTE Confidence: 0.9182085

00:01:12.905 --> 00:01:15.245 from developing better cancer immunotherapies,

NOTE Confidence: 0.9807825

00:01:15.705 --> 00:01:17.725 from understanding many autoimmune diseases

NOTE Confidence: 0.9807825

00:01:17.784 --> 00:01:19.245 that are driven by autoreactive

NOTE Confidence: 0.86685604

00:01:19.625 --> 00:01:21.805 t cells, from vaccine development,

NOTE Confidence: 0.86685604

00:01:21.944 --> 00:01:22.444 etcetera.

NOTE Confidence: 0.9219354

00:01:23.600 --> 00:01:24.640 So this has been a

NOTE Confidence: 0.9219354

00:01:24.640 --> 00:01:25.680 lot of work done.

NOTE Confidence: 0.94813275

00:01:26.080 --> 00:01:27.520 This is this plot is,

NOTE Confidence: 0.94813275

00:01:27.680 --> 00:01:29.200 it's from a recent review

NOTE Confidence: 0.94813275

00:01:29.200 --> 00:01:30.640 we wrote, where we are

NOTE Confidence: 0.94813275

00:01:30.640 --> 00:01:32.880 surveying the recent computational models

NOTE Confidence: 0.94813275

00:01:32.880 --> 00:01:34.240 to predict T cell receptor

NOTE Confidence: 0.94813275

00:01:34.240 --> 00:01:35.600 binding. And as you can

NOTE Confidence: 0.94813275

00:01:35.600 --> 00:01:36.560 see, just in two thousand

NOTE Confidence: 0.94813275

00:01:36.560 --> 00:01:37.680 twenty three, there were more

NOTE Confidence: 0.94813275

00:01:37.680 --> 00:01:39.415 than twenty papers, yes, for

NOTE Confidence: 0.94813275

00:01:39.415 --> 00:01:41.034 t cell receptor binding prediction.

NOTE Confidence: 0.86144394

00:01:41.735 --> 00:01:42.875 There was interesting

NOTE Confidence: 0.9408023

00:01:43.175 --> 00:01:44.215 trends, and the trend is

NOTE Confidence: 0.9408023

00:01:44.215 --> 00:01:45.495 not slowing down. I'm not,

NOTE Confidence: 0.9408023

00:01:45.895 --> 00:01:47.015 checking what's going on. There's

NOTE Confidence: 0.9408023

00:01:47.015 --> 00:01:47.975 still a lot of papers

NOTE Confidence: 0.9408023

00:01:47.975 --> 00:01:49.034 published this year.

NOTE Confidence: 0.8884192

00:01:50.135 --> 00:01:50.635 So,

NOTE Confidence: 0.79344565

00:01:52.170 --> 00:01:52.970 there a point. Is there

NOTE Confidence: 0.79344565

00:01:52.970 --> 00:01:53.630 a pointer

NOTE Confidence: 0.6969429

00:01:54.970 --> 00:01:55.630 or not?

NOTE Confidence: 0.9587407

00:01:57.450 --> 00:01:58.810 Or the mouse? Okay. Yeah.

NOTE Confidence: 0.9587407

00:01:58.810 --> 00:02:00.250 The mouse works. Yeah. Thank

NOTE Confidence: 0.9587407

00:02:00.250 --> 00:02:01.690 you. Yeah. So the other

NOTE Confidence: 0.9587407

00:02:01.690 --> 00:02:02.970 interesting trend that we can

NOTE Confidence: 0.9587407

00:02:02.970 --> 00:02:03.850 see in this plot is

NOTE Confidence: 0.9587407

00:02:03.850 --> 00:02:04.729 that there seem to be

NOTE Confidence: 0.9587407

00:02:04.729 --> 00:02:06.490 a switch. So until maybe

NOTE Confidence: 0.9587407

00:02:06.490 --> 00:02:07.945 two thousand twenty one, most

NOTE Confidence: 0.9587407

00:02:07.945 --> 00:02:09.385 models were using some sort

NOTE Confidence: 0.9587407

00:02:09.385 --> 00:02:11.465 of supervised machine learning. But

NOTE Confidence: 0.9587407

00:02:11.465 --> 00:02:13.065 in recent years, very recently,

NOTE Confidence: 0.9587407

00:02:13.065 --> 00:02:14.105 we are seeing the emergence

NOTE Confidence: 0.9587407

00:02:14.105 --> 00:02:16.025 of protein language models. And,

NOTE Confidence: 0.9587407

00:02:16.025 --> 00:02:16.525 well,

NOTE Confidence: 0.9540296

00:02:17.225 --> 00:02:18.665 Naveep already introduced a little

NOTE Confidence: 0.9540296

00:02:18.665 --> 00:02:19.465 bit, and I will tell

NOTE Confidence: 0.9540296

00:02:19.465 --> 00:02:20.185 you a bit more in

NOTE Confidence: 0.9540296

00:02:20.185 --> 00:02:20.845 a few,

NOTE Confidence: 0.95211893

00:02:21.385 --> 00:02:22.540 a little bit. But as

NOTE Confidence: 0.95211893

00:02:22.540 --> 00:02:23.100 you can see, there is

NOTE Confidence: 0.95211893

00:02:23.100 --> 00:02:24.060 a lot of interest now

NOTE Confidence: 0.95211893

00:02:24.060 --> 00:02:25.419 in applying the latest deep

NOTE Confidence: 0.95211893

00:02:25.419 --> 00:02:26.320 learning technologies.

NOTE Confidence: 0.9591252

00:02:27.500 --> 00:02:28.780 Now the other thing to

NOTE Confidence: 0.9591252

00:02:28.780 --> 00:02:29.580 tell you that most of

NOTE Confidence: 0.9591252

00:02:29.580 --> 00:02:30.540 these models, well, at least

NOTE Confidence: 0.9591252

00:02:30.540 --> 00:02:31.419 the ones in this in

NOTE Confidence: 0.9591252

00:02:31.419 --> 00:02:32.700 this plot were based only

NOTE Confidence: 0.9591252

00:02:32.700 --> 00:02:33.820 in sequence. That means we

NOTE Confidence: 0.9591252

00:02:33.820 --> 00:02:34.780 take the sequence of the

NOTE Confidence: 0.9591252

00:02:34.780 --> 00:02:36.060 cell receptors, the sequence of

NOTE Confidence: 0.9591252

00:02:36.060 --> 00:02:37.100 the epitope, and we try

NOTE Confidence: 0.9591252

00:02:37.100 --> 00:02:39.075 to predict binding with more

NOTE Confidence: 0.9591252

00:02:39.075 --> 00:02:40.135 or less success.

NOTE Confidence: 0.9339432

00:02:40.675 --> 00:02:41.475 It's true that we are

NOTE Confidence: 0.9339432

00:02:41.475 --> 00:02:42.595 at this point, most people

NOTE Confidence: 0.9339432

00:02:42.595 --> 00:02:44.275 were still neglecting a structure.

NOTE Confidence: 0.9339432

00:02:44.275 --> 00:02:45.315 And, certainly, this is a

NOTE Confidence: 0.9339432

00:02:45.315 --> 00:02:46.535 big, oversimplification.

NOTE Confidence: 0.9651075

00:02:47.235 --> 00:02:48.675 Structure is important to predict

NOTE Confidence: 0.9651075

00:02:48.675 --> 00:02:49.175 binding.

NOTE Confidence: 0.93287855

00:02:49.475 --> 00:02:50.675 The reason until now, most

NOTE Confidence: 0.93287855

00:02:50.675 --> 00:02:51.794 people are leaving a structure

NOTE Confidence: 0.93287855

00:02:51.794 --> 00:02:52.770 on the side is just

NOTE Confidence: 0.93287855

00:02:52.770 --> 00:02:53.970 because T cell receptors are

NOTE Confidence: 0.93287855

00:02:53.970 --> 00:02:54.930 extremely difficult,

NOTE Confidence: 0.9167473

00:02:55.410 --> 00:02:57.010 proteins to model. The part

NOTE Confidence: 0.9167473

00:02:57.010 --> 00:02:58.130 that binds the anti the

NOTE Confidence: 0.9167473

00:02:58.130 --> 00:02:59.330 the epitope is a flexible

NOTE Confidence: 0.9167473

00:02:59.330 --> 00:03:01.010 loop. And flexible loops and

NOTE Confidence: 0.9167473

00:03:01.010 --> 00:03:02.530 other so their proteins cannot

NOTE Confidence: 0.9167473

00:03:02.530 --> 00:03:04.450 be effectively model with alpha

NOTE Confidence: 0.9167473

00:03:04.450 --> 00:03:05.675 fold, not even in the

NOTE Confidence: 0.9167473

00:03:05.675 --> 00:03:07.675 last version and similarity level

NOTE Confidence: 0.9167473

00:03:07.675 --> 00:03:09.595 of models. So people are

NOTE Confidence: 0.9167473

00:03:09.595 --> 00:03:10.875 neglecting a structure not because

NOTE Confidence: 0.9167473

00:03:10.875 --> 00:03:11.995 it's not important, it's just

NOTE Confidence: 0.9167473

00:03:11.995 --> 00:03:13.675 because it's challenging. But we

NOTE Confidence: 0.9167473

00:03:13.675 --> 00:03:15.115 are still starting to see

NOTE Confidence: 0.9167473

00:03:15.115 --> 00:03:16.395 the new wave of papers

NOTE Confidence: 0.9167473

00:03:16.395 --> 00:03:17.915 that are somehow trying to

NOTE Confidence: 0.9167473

00:03:17.915 --> 00:03:19.294 to to look into that.

NOTE Confidence: 0.93385124

00:03:20.500 --> 00:03:21.459 For the time being, let

NOTE Confidence: 0.93385124

00:03:21.459 --> 00:03:22.739 me take one step back,

NOTE Confidence: 0.93385124

00:03:22.739 --> 00:03:23.620 and let me tell you

NOTE Confidence: 0.93385124

00:03:23.620 --> 00:03:25.400 the work with how chronologically

NOTE Confidence: 0.930811

00:03:25.700 --> 00:03:26.500 the work we have been

NOTE Confidence: 0.930811

00:03:26.500 --> 00:03:28.099 doing in this topic starting

NOTE Confidence: 0.930811

00:03:28.099 --> 00:03:29.940 from the work, still focus

NOTE Confidence: 0.930811

00:03:29.940 --> 00:03:31.139 on sequence that we did,

NOTE Confidence: 0.930811

00:03:31.379 --> 00:03:32.580 when we started thinking about

NOTE Confidence: 0.930811

00:03:32.580 --> 00:03:33.319 this problem.

NOTE Confidence: 0.9085432

00:03:33.780 --> 00:03:34.924 So, basically, this was like

NOTE Confidence: 0.9085432

00:03:34.924 --> 00:03:35.805 a few years ago, like,

NOTE Confidence: 0.9085432

00:03:35.805 --> 00:03:36.924 I think four years ago,

NOTE Confidence: 0.9085432

00:03:36.924 --> 00:03:38.364 and we're still start thinking

NOTE Confidence: 0.9085432

00:03:38.364 --> 00:03:39.885 about predicting the cell receptor

NOTE Confidence: 0.9085432

00:03:39.885 --> 00:03:41.485 binding. And we're got wanted

NOTE Confidence: 0.9085432

00:03:41.485 --> 00:03:42.784 to use we have expertise

NOTE Confidence: 0.9085432

00:03:42.845 --> 00:03:43.965 on machine learning, and we

NOTE Confidence: 0.9085432

00:03:43.965 --> 00:03:45.504 wanted to develop a multimodal

NOTE Confidence: 0.9085432

00:03:45.724 --> 00:03:47.165 deep learning approach. That means

NOTE Confidence: 0.9085432

00:03:47.165 --> 00:03:48.990 you give, the the model

NOTE Confidence: 0.9085432

00:03:49.150 --> 00:03:49.950 the sequence of the T

NOTE Confidence: 0.9085432

00:03:49.950 --> 00:03:50.690 cell receptor,

NOTE Confidence: 0.98724514

00:03:50.990 --> 00:03:52.030 the sequence of the epitope,

NOTE Confidence: 0.98724514

00:03:52.030 --> 00:03:52.830 and you train it to

NOTE Confidence: 0.98724514

00:03:52.830 --> 00:03:53.650 predict binding.

NOTE Confidence: 0.89333683

00:03:54.270 --> 00:03:55.390 And another topic I will

NOTE Confidence: 0.89333683

00:03:55.390 --> 00:03:56.270 tell you a little bit,

NOTE Confidence: 0.89333683

00:03:56.270 --> 00:03:57.230 we have in my team

NOTE Confidence: 0.89333683

00:03:57.230 --> 00:03:58.670 at IBM were very interested

NOTE Confidence: 0.89333683

00:03:58.670 --> 00:03:59.810 in inter interpretability.

NOTE Confidence: 0.96990603

00:04:00.110 --> 00:04:01.070 So that means we wanted

NOTE Confidence: 0.96990603

00:04:01.070 --> 00:04:02.325 just to build a model,

NOTE Confidence: 0.96990603

00:04:02.325 --> 00:04:03.525 but also to get some

NOTE Confidence: 0.96990603

00:04:03.525 --> 00:04:04.965 sort of insight of why

NOTE Confidence: 0.96990603

00:04:04.965 --> 00:04:05.945 the model predicts

NOTE Confidence: 0.9993367

00:04:06.245 --> 00:04:07.545 binding or not binding.

NOTE Confidence: 0.93049675

00:04:09.045 --> 00:04:10.325 So now the challenges well,

NOTE Confidence: 0.93049675

00:04:10.325 --> 00:04:11.285 one of the main challenge

NOTE Confidence: 0.93049675

00:04:11.285 --> 00:04:12.585 of this problem in biology,

NOTE Confidence: 0.93049675

00:04:12.645 --> 00:04:14.505 typically, most models are underpowered,

NOTE Confidence: 0.93049675

00:04:14.645 --> 00:04:16.185 but especially in this particular

NOTE Confidence: 0.9246081

00:04:16.485 --> 00:04:18.960 regime. So the theoretical diversity

NOTE Confidence: 0.9246081

00:04:19.100 --> 00:04:20.779 of potential disal receptor that

NOTE Confidence: 0.9246081

00:04:20.779 --> 00:04:22.560 a person can harbor is

NOTE Confidence: 0.9010193

00:04:22.940 --> 00:04:23.440 astronomical.

NOTE Confidence: 0.9858167

00:04:24.699 --> 00:04:24.940 But,

NOTE Confidence: 0.94537234

00:04:25.660 --> 00:04:26.860 we have some what? We

NOTE Confidence: 0.94537234

00:04:26.860 --> 00:04:27.979 have, quite a lot of

NOTE Confidence: 0.94537234

00:04:27.979 --> 00:04:29.355 data and public databases, but

NOTE Confidence: 0.94537234

00:04:29.435 --> 00:04:30.475 if you compare these two

NOTE Confidence: 0.94537234

00:04:30.475 --> 00:04:32.075 numbers, certainly you see quite

NOTE Confidence: 0.94537234

00:04:32.075 --> 00:04:33.355 a few orders of magnitude

NOTE Confidence: 0.94537234

00:04:33.355 --> 00:04:33.935 of indifference.

NOTE Confidence: 0.87444186

00:04:34.395 --> 00:04:35.515 So it's clear that although

NOTE Confidence: 0.87444186

00:04:35.515 --> 00:04:36.955 the the numbers these numbers

NOTE Confidence: 0.87444186

00:04:36.955 --> 00:04:38.315 by the world was towards

NOTE Confidence: 0.87444186

00:04:38.315 --> 00:04:39.275 the numbers we use when

NOTE Confidence: 0.87444186

00:04:39.275 --> 00:04:40.795 we train this model, by

NOTE Confidence: 0.87444186

00:04:40.795 --> 00:04:42.415 now, these numbers has increased

NOTE Confidence: 0.87444186

00:04:42.555 --> 00:04:43.055 substantially.

NOTE Confidence: 0.98633194

00:04:43.520 --> 00:04:44.720 But still, certainly, we are

NOTE Confidence: 0.98633194

00:04:44.720 --> 00:04:46.320 not even sampling a fraction

NOTE Confidence: 0.98633194

00:04:46.320 --> 00:04:47.860 of the huge potential diversity.

NOTE Confidence: 0.9531769

00:04:48.240 --> 00:04:49.040 So this is something to

NOTE Confidence: 0.9531769

00:04:49.040 --> 00:04:50.160 keep in mind. Models are

NOTE Confidence: 0.9531769

00:04:50.160 --> 00:04:51.140 still under power.

NOTE Confidence: 0.9330418

00:04:51.760 --> 00:04:53.200 And especially when we come

NOTE Confidence: 0.9330418

00:04:53.200 --> 00:04:54.720 to the EBITDAF side so,

NOTE Confidence: 0.9330418

00:04:54.720 --> 00:04:55.920 again, the number has increased

NOTE Confidence: 0.9330418

00:04:55.920 --> 00:04:56.640 by now. I think by

NOTE Confidence: 0.9330418

00:04:56.640 --> 00:04:57.600 now, there might be state

NOTE Confidence: 0.9330418

00:04:57.600 --> 00:04:57.920 about,

NOTE Confidence: 0.91088754

00:04:58.564 --> 00:04:59.925 less than one thousand. That's

NOTE Confidence: 0.91088754

00:04:59.925 --> 00:05:01.525 that's still it's still quite

NOTE Confidence: 0.91088754

00:05:01.525 --> 00:05:02.745 small amount of data.

NOTE Confidence: 0.96083325

00:05:03.444 --> 00:05:04.245 So as you can see

NOTE Confidence: 0.96083325

00:05:04.245 --> 00:05:05.125 already, I can I don't

NOTE Confidence: 0.96083325

00:05:05.125 --> 00:05:06.085 need to tell you that

NOTE Confidence: 0.96083325

00:05:06.085 --> 00:05:07.044 we might be able to

NOTE Confidence: 0.96083325

00:05:07.044 --> 00:05:08.245 predict something about T cell

NOTE Confidence: 0.96083325

00:05:08.245 --> 00:05:09.764 receptors, but certainly, we'll not

NOTE Confidence: 0.96083325

00:05:09.764 --> 00:05:10.564 have a lot of power

NOTE Confidence: 0.96083325

00:05:10.564 --> 00:05:11.685 to predict binding to new

NOTE Confidence: 0.96083325

00:05:11.685 --> 00:05:12.884 epitopes, which is one of

NOTE Confidence: 0.96083325

00:05:12.884 --> 00:05:14.550 the has the most interesting

NOTE Confidence: 0.96083325

00:05:14.550 --> 00:05:15.770 biomedical applications?

NOTE Confidence: 0.9680107

00:05:17.110 --> 00:05:18.550 Nevertheless, with this caveat, let

NOTE Confidence: 0.9680107

00:05:18.550 --> 00:05:19.190 me tell you a little

NOTE Confidence: 0.9680107

00:05:19.190 --> 00:05:20.310 bit what we did. So

NOTE Confidence: 0.9680107

00:05:20.310 --> 00:05:21.750 here, we are focusing first

NOTE Confidence: 0.9680107

00:05:21.750 --> 00:05:22.710 on the site where we

NOTE Confidence: 0.9680107

00:05:22.710 --> 00:05:23.750 can predict. That means we

NOTE Confidence: 0.9680107

00:05:23.750 --> 00:05:24.650 fix the epitope,

NOTE Confidence: 0.89283115

00:05:24.950 --> 00:05:26.070 and we try to we

NOTE Confidence: 0.89283115

00:05:26.070 --> 00:05:27.770 build models with the predict

NOTE Confidence: 0.89283115

00:05:27.830 --> 00:05:28.950 the the binding to know

NOTE Confidence: 0.89283115

00:05:28.950 --> 00:05:29.770 what to new,

NOTE Confidence: 0.93038535

00:05:30.074 --> 00:05:31.294 to to different decelerate.

NOTE Confidence: 0.9112181

00:05:32.714 --> 00:05:34.235 So with these constraints, we

NOTE Confidence: 0.9112181

00:05:34.235 --> 00:05:35.514 train for models that are

NOTE Confidence: 0.9112181

00:05:35.514 --> 00:05:37.595 different variations because different choices

NOTE Confidence: 0.9112181

00:05:37.595 --> 00:05:38.474 you can make in what

NOTE Confidence: 0.9112181

00:05:38.474 --> 00:05:39.995 you encode, so the full

NOTE Confidence: 0.9112181

00:05:39.995 --> 00:05:41.375 sequence or only the variable

NOTE Confidence: 0.9112181

00:05:41.435 --> 00:05:42.714 part, and also the ways

NOTE Confidence: 0.9112181

00:05:42.714 --> 00:05:43.835 in which you encode the

NOTE Confidence: 0.9112181

00:05:43.835 --> 00:05:45.010 the epitope. And already and

NOTE Confidence: 0.9112181

00:05:45.010 --> 00:05:46.770 Nadif already explained. So an

NOTE Confidence: 0.9112181

00:05:46.770 --> 00:05:48.690 amino acid is, an epitope

NOTE Confidence: 0.9112181

00:05:48.690 --> 00:05:49.810 is a string of amino

NOTE Confidence: 0.9112181

00:05:49.810 --> 00:05:51.810 acids, nine to ten, fifteen

NOTE Confidence: 0.9112181

00:05:51.810 --> 00:05:53.110 amino acids, but it's relatively

NOTE Confidence: 0.9112181

00:05:53.250 --> 00:05:53.750 small.

NOTE Confidence: 0.930249

00:05:54.529 --> 00:05:56.290 There are typically called as

NOTE Confidence: 0.930249

00:05:56.290 --> 00:05:57.650 amino acids, but there are

NOTE Confidence: 0.930249

00:05:57.650 --> 00:05:58.850 some tricks that you can

NOTE Confidence: 0.930249

00:05:58.850 --> 00:05:59.634 do here and I'll take

NOTE Confidence: 0.930249

00:05:59.634 --> 00:06:00.455 you in a second.

NOTE Confidence: 0.9134565

00:06:01.314 --> 00:06:02.354 So first, we train this

NOTE Confidence: 0.9134565

00:06:02.354 --> 00:06:03.094 for models.

NOTE Confidence: 0.952211

00:06:03.875 --> 00:06:05.314 We were, of course, you

NOTE Confidence: 0.952211

00:06:05.314 --> 00:06:06.194 need to compare with a

NOTE Confidence: 0.952211

00:06:06.194 --> 00:06:06.694 baseline.

NOTE Confidence: 0.9756438

00:06:07.235 --> 00:06:08.275 So as a baseline, we

NOTE Confidence: 0.9756438

00:06:08.275 --> 00:06:10.534 develop a super simplistic classifier.

NOTE Confidence: 0.89061356

00:06:11.154 --> 00:06:12.810 So is this, based on

NOTE Confidence: 0.89061356

00:06:12.889 --> 00:06:14.010 k nearest neighbors is one

NOTE Confidence: 0.89061356

00:06:14.010 --> 00:06:15.450 of the simplest machine learning

NOTE Confidence: 0.89061356

00:06:15.450 --> 00:06:16.669 approaches you can develop.

NOTE Confidence: 0.9866027

00:06:17.210 --> 00:06:18.669 And we were disappointing

NOTE Confidence: 0.87714034

00:06:19.050 --> 00:06:20.970 disappointed, but not completely surprised

NOTE Confidence: 0.87714034

00:06:20.970 --> 00:06:21.930 when see that that will

NOTE Confidence: 0.87714034

00:06:21.930 --> 00:06:23.450 be the simple classifier with

NOTE Confidence: 0.87714034

00:06:23.450 --> 00:06:24.330 a much better and deep

NOTE Confidence: 0.87714034

00:06:24.330 --> 00:06:26.195 learning models. Again, this is

NOTE Confidence: 0.87714034

00:06:26.195 --> 00:06:27.815 a problem of data limitation.

NOTE Confidence: 0.87714034

00:06:27.955 --> 00:06:28.915 We do not have enough

NOTE Confidence: 0.87714034

00:06:28.915 --> 00:06:29.415 data.

NOTE Confidence: 0.89816046

00:06:30.275 --> 00:06:31.475 Now the trick that was

NOTE Confidence: 0.89816046

00:06:31.475 --> 00:06:32.675 really interesting here is is

NOTE Confidence: 0.89816046

00:06:32.675 --> 00:06:33.635 that coming back to this

NOTE Confidence: 0.89816046

00:06:33.635 --> 00:06:35.555 representation of epitopes. So an

NOTE Confidence: 0.89816046

00:06:35.555 --> 00:06:37.095 epitope, the the the more,

NOTE Confidence: 0.9227947

00:06:39.020 --> 00:06:40.540 normal representation will be to

NOTE Confidence: 0.9227947

00:06:40.540 --> 00:06:41.900 represent each amino acid and

NOTE Confidence: 0.9227947

00:06:41.900 --> 00:06:42.779 then just have a string

NOTE Confidence: 0.9227947

00:06:42.779 --> 00:06:44.720 of amino acids. But, however,

NOTE Confidence: 0.9227947

00:06:44.779 --> 00:06:46.220 if you consider an epitope

NOTE Confidence: 0.9227947

00:06:46.220 --> 00:06:47.420 is kind of rather small

NOTE Confidence: 0.9227947

00:06:47.420 --> 00:06:49.180 molecule, so you couldn't call

NOTE Confidence: 0.9227947

00:06:49.180 --> 00:06:50.205 it as if it was

NOTE Confidence: 0.9227947

00:06:50.284 --> 00:06:51.884 a compound using the small

NOTE Confidence: 0.9227947

00:06:51.884 --> 00:06:52.384 representations,

NOTE Confidence: 0.9365179

00:06:53.044 --> 00:06:54.705 and, Nadeep was talking about.

NOTE Confidence: 0.9365179

00:06:54.925 --> 00:06:55.805 Again, this is not a

NOTE Confidence: 0.9365179

00:06:55.805 --> 00:06:57.324 chemical compound, but from a

NOTE Confidence: 0.9365179

00:06:57.324 --> 00:06:58.844 computational point of view, they

NOTE Confidence: 0.9365179

00:06:58.844 --> 00:07:00.145 admit similar representations.

NOTE Confidence: 0.97624224

00:07:01.164 --> 00:07:02.284 And if you do that,

NOTE Confidence: 0.97624224

00:07:02.284 --> 00:07:03.805 now you can use transfer

NOTE Confidence: 0.97624224

00:07:03.805 --> 00:07:05.100 learning because you can now

NOTE Confidence: 0.97624224

00:07:05.100 --> 00:07:06.540 pretrain your model on a

NOTE Confidence: 0.97624224

00:07:06.540 --> 00:07:07.839 much larger database.

NOTE Confidence: 0.9324661

00:07:08.380 --> 00:07:09.660 So binding DB is like

NOTE Confidence: 0.9324661

00:07:09.660 --> 00:07:11.820 a database of compound protein

NOTE Confidence: 0.9324661

00:07:11.820 --> 00:07:13.500 compound interactions where we have

NOTE Confidence: 0.9324661

00:07:13.500 --> 00:07:15.260 around one million pairs. So

NOTE Confidence: 0.9324661

00:07:15.260 --> 00:07:16.160 this is a substantially

NOTE Confidence: 0.92722476

00:07:16.620 --> 00:07:18.060 larger amount of data to

NOTE Confidence: 0.92722476

00:07:18.060 --> 00:07:20.125 pretend our model. So, basically,

NOTE Confidence: 0.92722476

00:07:20.125 --> 00:07:20.925 what we did here is

NOTE Confidence: 0.92722476

00:07:20.925 --> 00:07:22.285 pretrain the model on binding

NOTE Confidence: 0.92722476

00:07:22.285 --> 00:07:23.005 d v and then, of

NOTE Confidence: 0.92722476

00:07:23.005 --> 00:07:24.285 course, fine tune on the

NOTE Confidence: 0.92722476

00:07:24.285 --> 00:07:25.985 cell receptor specific data.

NOTE Confidence: 0.9263091

00:07:26.365 --> 00:07:27.245 And as you can see,

NOTE Confidence: 0.9263091

00:07:27.245 --> 00:07:28.205 this is an ex this

NOTE Confidence: 0.9263091

00:07:28.205 --> 00:07:30.205 this, transfer learning exercise, it

NOTE Confidence: 0.9263091

00:07:30.205 --> 00:07:31.824 was assisted led a substantial,

NOTE Confidence: 0.9611415

00:07:32.365 --> 00:07:33.425 boost in performance.

NOTE Confidence: 0.9362621

00:07:33.919 --> 00:07:34.720 This, by the way, is

NOTE Confidence: 0.9362621

00:07:34.720 --> 00:07:35.680 done all the time in

NOTE Confidence: 0.9362621

00:07:35.680 --> 00:07:36.960 in in in different areas

NOTE Confidence: 0.9362621

00:07:36.960 --> 00:07:38.400 of machine learning. In biology,

NOTE Confidence: 0.9362621

00:07:38.400 --> 00:07:39.360 we are still getting used

NOTE Confidence: 0.9362621

00:07:39.360 --> 00:07:40.320 to it, yeah. But you

NOTE Confidence: 0.9362621

00:07:40.320 --> 00:07:41.360 can see here, we could

NOTE Confidence: 0.9362621

00:07:41.360 --> 00:07:42.580 really boost the performance.

NOTE Confidence: 0.88581777

00:07:44.720 --> 00:07:45.600 Now let me talk a

NOTE Confidence: 0.88581777

00:07:45.600 --> 00:07:47.199 little bit about interpretability and

NOTE Confidence: 0.88581777

00:07:47.199 --> 00:07:48.320 again, so another concept that

NOTE Confidence: 0.88581777

00:07:48.320 --> 00:07:49.975 Navif already explained, introduce a

NOTE Confidence: 0.88581777

00:07:49.975 --> 00:07:51.835 little bit. So we wanted,

NOTE Confidence: 0.89933324

00:07:52.295 --> 00:07:53.415 what when we we working

NOTE Confidence: 0.89933324

00:07:53.415 --> 00:07:54.695 here, we want we wanted

NOTE Confidence: 0.89933324

00:07:54.695 --> 00:07:55.895 not only to develop the

NOTE Confidence: 0.89933324

00:07:55.895 --> 00:07:57.415 model, but also just get

NOTE Confidence: 0.89933324

00:07:57.415 --> 00:07:59.095 insights about which amino acid

NOTE Confidence: 0.89933324

00:07:59.095 --> 00:08:00.375 the model the model thought

NOTE Confidence: 0.89933324

00:08:00.375 --> 00:08:01.995 were important to predict binding

NOTE Confidence: 0.89933324

00:08:02.055 --> 00:08:02.715 or nonbinding.

NOTE Confidence: 0.9615398

00:08:03.800 --> 00:08:05.160 So here we use attention

NOTE Confidence: 0.9615398

00:08:05.160 --> 00:08:06.600 mechanisms, which is one of

NOTE Confidence: 0.9615398

00:08:06.600 --> 00:08:07.960 the most common, not the

NOTE Confidence: 0.9615398

00:08:07.960 --> 00:08:08.920 only one, of course, but

NOTE Confidence: 0.9615398

00:08:08.920 --> 00:08:10.440 one of the popular approaches

NOTE Confidence: 0.9615398

00:08:10.440 --> 00:08:11.180 for interpretability.

NOTE Confidence: 0.9118585

00:08:11.560 --> 00:08:12.920 They are using transformers and

NOTE Confidence: 0.9118585

00:08:12.920 --> 00:08:14.040 in language models, so they

NOTE Confidence: 0.9118585

00:08:14.040 --> 00:08:15.260 are becoming very popular.

NOTE Confidence: 0.9644656

00:08:16.035 --> 00:08:17.475 The idea with with potential

NOTE Confidence: 0.9644656

00:08:17.475 --> 00:08:18.515 layers that you can get

NOTE Confidence: 0.9644656

00:08:18.515 --> 00:08:19.715 after you train your model

NOTE Confidence: 0.9644656

00:08:19.715 --> 00:08:20.935 and you get your prediction,

NOTE Confidence: 0.9644656

00:08:21.075 --> 00:08:21.955 you can get this type

NOTE Confidence: 0.9644656

00:08:21.955 --> 00:08:23.475 of heat maps that tell

NOTE Confidence: 0.9644656

00:08:23.475 --> 00:08:24.995 you that dark color means

NOTE Confidence: 0.9644656

00:08:24.995 --> 00:08:26.675 how much attention the model

NOTE Confidence: 0.9644656

00:08:26.675 --> 00:08:27.795 paid to each feature. In

NOTE Confidence: 0.9644656

00:08:27.795 --> 00:08:29.335 that case, each amino acid.

NOTE Confidence: 0.970371

00:08:29.850 --> 00:08:31.050 So for instance, these three

NOTE Confidence: 0.970371

00:08:31.050 --> 00:08:32.970 sequences correspond to three the

NOTE Confidence: 0.970371

00:08:32.970 --> 00:08:34.590 variable part of three sequences.

NOTE Confidence: 0.970371

00:08:34.730 --> 00:08:36.330 These are receptor sequences that

NOTE Confidence: 0.970371

00:08:36.330 --> 00:08:37.530 were predicted to be to

NOTE Confidence: 0.970371

00:08:37.530 --> 00:08:38.750 bind the same epitope.

NOTE Confidence: 0.9131957

00:08:39.530 --> 00:08:40.970 And, again, the the the

NOTE Confidence: 0.9131957

00:08:40.970 --> 00:08:42.090 the trick the problem here

NOTE Confidence: 0.9131957

00:08:42.090 --> 00:08:42.730 is we don't have the

NOTE Confidence: 0.9131957

00:08:42.730 --> 00:08:43.915 ground truth because, I mean,

NOTE Confidence: 0.9131957

00:08:43.915 --> 00:08:44.795 we need to have the

NOTE Confidence: 0.9131957

00:08:44.795 --> 00:08:45.295 crystallographical

NOTE Confidence: 0.9738994

00:08:45.755 --> 00:08:46.955 structures to see whether these

NOTE Confidence: 0.9738994

00:08:46.955 --> 00:08:48.075 amino acids are involved in

NOTE Confidence: 0.9738994

00:08:48.075 --> 00:08:49.195 the binding, which we don't

NOTE Confidence: 0.9738994

00:08:49.195 --> 00:08:49.695 have.

NOTE Confidence: 0.9672602

00:08:50.395 --> 00:08:51.595 But we can do some

NOTE Confidence: 0.9672602

00:08:51.595 --> 00:08:53.675 sort of sanity check. So

NOTE Confidence: 0.9672602

00:08:53.675 --> 00:08:55.035 if they bind the SNP

NOTE Confidence: 0.9672602

00:08:55.035 --> 00:08:56.715 top, it's logical to expect

NOTE Confidence: 0.9672602

00:08:56.715 --> 00:08:57.915 that they share a similar

NOTE Confidence: 0.9672602

00:08:57.915 --> 00:08:59.470 binding motif. And as you

NOTE Confidence: 0.9672602

00:08:59.470 --> 00:09:00.510 can see here, actually, the

NOTE Confidence: 0.9672602

00:09:00.510 --> 00:09:01.870 amino acids that were highlighted

NOTE Confidence: 0.9672602

00:09:01.870 --> 00:09:03.809 with high attention are conserved

NOTE Confidence: 0.9672602

00:09:03.950 --> 00:09:05.010 on the three sequences.

NOTE Confidence: 0.92669743

00:09:05.390 --> 00:09:06.350 So again, this is not

NOTE Confidence: 0.92669743

00:09:06.350 --> 00:09:07.550 a validation. It's all only

NOTE Confidence: 0.92669743

00:09:07.550 --> 00:09:08.990 a sanity check. But from

NOTE Confidence: 0.92669743

00:09:08.990 --> 00:09:10.110 this point of view, this

NOTE Confidence: 0.92669743

00:09:10.110 --> 00:09:12.030 prediction, this attention prediction makes

NOTE Confidence: 0.92669743

00:09:12.030 --> 00:09:12.530 sense.

NOTE Confidence: 0.9563044

00:09:13.804 --> 00:09:14.605 Now I want to show

NOTE Confidence: 0.9563044

00:09:14.605 --> 00:09:15.965 you now a negative example.

NOTE Confidence: 0.9563044

00:09:15.965 --> 00:09:17.645 Another case where attention actually

NOTE Confidence: 0.9563044

00:09:17.645 --> 00:09:18.445 tell you that something is

NOTE Confidence: 0.9563044

00:09:18.445 --> 00:09:19.505 wrong in your model.

NOTE Confidence: 0.94775224

00:09:20.205 --> 00:09:21.405 So when I was telling

NOTE Confidence: 0.94775224

00:09:21.405 --> 00:09:22.445 you before that we have

NOTE Confidence: 0.94775224

00:09:22.445 --> 00:09:23.485 very little data to try

NOTE Confidence: 0.94775224

00:09:23.485 --> 00:09:25.405 to predict the initialization bending

NOTE Confidence: 0.94775224

00:09:25.405 --> 00:09:26.740 to unseen epitopes. This is

NOTE Confidence: 0.94775224

00:09:26.740 --> 00:09:28.420 an unsolved problem even today,

NOTE Confidence: 0.94775224

00:09:28.420 --> 00:09:29.240 forty years

NOTE Confidence: 0.94930726

00:09:29.780 --> 00:09:31.059 later. So we train our

NOTE Confidence: 0.94930726

00:09:31.059 --> 00:09:32.500 models also in this direction.

NOTE Confidence: 0.94930726

00:09:32.500 --> 00:09:33.860 The performance was quite bad

NOTE Confidence: 0.94930726

00:09:33.860 --> 00:09:34.600 as expected,

NOTE Confidence: 0.9891363

00:09:35.140 --> 00:09:36.100 but we wanted to look

NOTE Confidence: 0.9891363

00:09:36.100 --> 00:09:37.160 at attention maps.

NOTE Confidence: 0.9608789

00:09:38.020 --> 00:09:38.900 So this is what we

NOTE Confidence: 0.9608789

00:09:38.900 --> 00:09:40.455 got. And, actually, that will

NOTE Confidence: 0.9608789

00:09:40.615 --> 00:09:42.375 we were kind of, thinking

NOTE Confidence: 0.9608789

00:09:42.375 --> 00:09:43.335 about this for quite some

NOTE Confidence: 0.9608789

00:09:43.335 --> 00:09:44.554 time. We could not understand

NOTE Confidence: 0.9608789

00:09:44.615 --> 00:09:45.815 these maps for quite some

NOTE Confidence: 0.9608789

00:09:45.815 --> 00:09:46.315 time.

NOTE Confidence: 0.9887455

00:09:46.695 --> 00:09:48.214 Eventually, what we realized because,

NOTE Confidence: 0.9887455

00:09:48.214 --> 00:09:49.895 basically, for all two hundred

NOTE Confidence: 0.9887455

00:09:49.895 --> 00:09:51.510 epitopes in our database, the

NOTE Confidence: 0.9887455

00:09:51.590 --> 00:09:53.190 model was always selecting with

NOTE Confidence: 0.9887455

00:09:53.190 --> 00:09:54.550 the same attention or with

NOTE Confidence: 0.9887455

00:09:54.550 --> 00:09:55.910 the same three epitopes in

NOTE Confidence: 0.9887455

00:09:55.910 --> 00:09:56.730 the same positions.

NOTE Confidence: 0.98800117

00:09:57.110 --> 00:09:58.410 So that was quite puzzling.

NOTE Confidence: 0.94050014

00:09:59.110 --> 00:10:00.550 But eventually, we realized that

NOTE Confidence: 0.94050014

00:10:00.550 --> 00:10:02.470 the model actually realizes quite

NOTE Confidence: 0.94050014

00:10:02.470 --> 00:10:03.670 early that there's not sufficient

NOTE Confidence: 0.94050014

00:10:03.670 --> 00:10:04.710 data to try to make

NOTE Confidence: 0.94050014

00:10:04.710 --> 00:10:06.330 a prediction to unseen epitopes.

NOTE Confidence: 0.94050014

00:10:06.550 --> 00:10:08.204 So instead, just it's looking

NOTE Confidence: 0.94050014

00:10:08.204 --> 00:10:10.045 at it selects three random

NOTE Confidence: 0.94050014

00:10:10.045 --> 00:10:11.964 positions, highly variable positions, but

NOTE Confidence: 0.94050014

00:10:11.964 --> 00:10:12.464 random.

NOTE Confidence: 0.9201678

00:10:13.084 --> 00:10:14.125 And this is enough to

NOTE Confidence: 0.9201678

00:10:14.125 --> 00:10:15.084 give a label to each

NOTE Confidence: 0.9201678

00:10:15.084 --> 00:10:16.045 epitope. And this is a

NOTE Confidence: 0.9201678

00:10:16.045 --> 00:10:17.665 way of memorizing your data.

NOTE Confidence: 0.9201678

00:10:17.885 --> 00:10:19.165 So now you have instead

NOTE Confidence: 0.9201678

00:10:19.165 --> 00:10:20.865 of just training to generalize,

NOTE Confidence: 0.9201678

00:10:20.925 --> 00:10:22.204 you're just transforming this into

NOTE Confidence: 0.9201678

00:10:22.204 --> 00:10:23.425 a classification problem.

NOTE Confidence: 0.94517666

00:10:23.750 --> 00:10:24.550 Of course, you're not able

NOTE Confidence: 0.94517666

00:10:24.550 --> 00:10:25.990 to generalize to new epitopes.

NOTE Confidence: 0.94517666

00:10:25.990 --> 00:10:27.290 At least, we already knew.

NOTE Confidence: 0.9715951

00:10:27.910 --> 00:10:29.110 So I'm showing this because,

NOTE Confidence: 0.9715951

00:10:29.110 --> 00:10:31.050 actually, this highlights why interpretability

NOTE Confidence: 0.9715951

00:10:31.350 --> 00:10:31.929 is essential.

NOTE Confidence: 0.9310105

00:10:32.630 --> 00:10:33.910 So it can first tell

NOTE Confidence: 0.9310105

00:10:33.910 --> 00:10:35.350 you when things are working.

NOTE Confidence: 0.9310105

00:10:35.350 --> 00:10:36.230 It can tell you which

NOTE Confidence: 0.9310105

00:10:36.230 --> 00:10:38.054 amino acids might be involved

NOTE Confidence: 0.9310105

00:10:38.054 --> 00:10:39.175 into the grind the the

NOTE Confidence: 0.9310105

00:10:39.175 --> 00:10:40.535 binding prediction and give you

NOTE Confidence: 0.9310105

00:10:40.535 --> 00:10:41.894 ideas to just build new

NOTE Confidence: 0.9310105

00:10:41.894 --> 00:10:42.394 hypothesis.

NOTE Confidence: 0.9208386

00:10:43.175 --> 00:10:44.615 But equally important, it tells

NOTE Confidence: 0.9208386

00:10:44.615 --> 00:10:45.415 you when things are not

NOTE Confidence: 0.9208386

00:10:45.415 --> 00:10:47.175 working. And especially in biology,

NOTE Confidence: 0.9208386

00:10:47.175 --> 00:10:48.375 with very often when very

NOTE Confidence: 0.9208386

00:10:48.375 --> 00:10:49.735 often we're working with really

NOTE Confidence: 0.9208386

00:10:49.735 --> 00:10:51.254 underpowered datasets and we are

NOTE Confidence: 0.9208386

00:10:51.254 --> 00:10:52.774 overfitting, it's good to get

NOTE Confidence: 0.9208386

00:10:52.774 --> 00:10:53.710 hints when things are not

NOTE Confidence: 0.9208386

00:10:53.710 --> 00:10:54.210 working.

NOTE Confidence: 0.9644453

00:10:55.870 --> 00:10:57.230 Okay. So after saying that,

NOTE Confidence: 0.9644453

00:10:57.230 --> 00:10:58.030 let me tell you what

NOTE Confidence: 0.9644453

00:10:58.030 --> 00:10:59.309 we did after we published

NOTE Confidence: 0.9644453

00:10:59.309 --> 00:11:00.590 this paper, and it was

NOTE Confidence: 0.9644453

00:11:00.590 --> 00:11:01.470 more or less at the

NOTE Confidence: 0.9644453

00:11:01.470 --> 00:11:03.470 time where, protein language models

NOTE Confidence: 0.9644453

00:11:03.470 --> 00:11:04.670 started to become a hit

NOTE Confidence: 0.9644453

00:11:04.670 --> 00:11:05.410 in biology.

NOTE Confidence: 0.93746495

00:11:05.870 --> 00:11:06.704 So I think the first

NOTE Confidence: 0.93746495

00:11:06.704 --> 00:11:08.065 protein language model was published

NOTE Confidence: 0.93746495

00:11:08.065 --> 00:11:09.445 in two thousand twenty one.

NOTE Confidence: 0.93746495

00:11:09.505 --> 00:11:10.865 So, yes, well, we're introduced

NOTE Confidence: 0.93746495

00:11:10.865 --> 00:11:11.825 by Nadiv, but in case,

NOTE Confidence: 0.93746495

00:11:12.144 --> 00:11:13.265 for those of who just

NOTE Confidence: 0.93746495

00:11:13.265 --> 00:11:15.184 came recently and have not

NOTE Confidence: 0.93746495

00:11:15.184 --> 00:11:16.725 heard about protein language models.

NOTE Confidence: 0.93746495

00:11:16.865 --> 00:11:17.904 So the idea is that

NOTE Confidence: 0.93746495

00:11:17.904 --> 00:11:18.865 they use the same type

NOTE Confidence: 0.93746495

00:11:18.865 --> 00:11:19.559 of models, a

NOTE Confidence: 0.95363426

00:11:33.765 --> 00:11:34.485 And if you do that,

NOTE Confidence: 0.95363426

00:11:34.485 --> 00:11:35.684 then you are learning the

NOTE Confidence: 0.95363426

00:11:35.684 --> 00:11:37.365 language of proteins, and this

NOTE Confidence: 0.95363426

00:11:37.365 --> 00:11:38.325 is what is called protein

NOTE Confidence: 0.95363426

00:11:38.325 --> 00:11:39.145 language models.

NOTE Confidence: 0.90936

00:11:40.325 --> 00:11:41.525 So there by now quite

NOTE Confidence: 0.90936

00:11:41.525 --> 00:11:42.405 a few of this model,

NOTE Confidence: 0.90936

00:11:42.405 --> 00:11:43.845 this is one of, early

NOTE Confidence: 0.90936

00:11:43.845 --> 00:11:45.445 ones, was published in Nature

NOTE Confidence: 0.90936

00:11:45.445 --> 00:11:46.929 Biotech, and this one was

NOTE Confidence: 0.90936

00:11:46.929 --> 00:11:47.809 shown to be able to

NOTE Confidence: 0.90936

00:11:47.809 --> 00:11:49.910 predict secondary and tertiary structure.

NOTE Confidence: 0.92781514

00:11:50.370 --> 00:11:51.650 Later on, there was other

NOTE Confidence: 0.92781514

00:11:51.650 --> 00:11:52.690 models that were able to

NOTE Confidence: 0.92781514

00:11:52.690 --> 00:11:54.610 design the novel proteins with

NOTE Confidence: 0.92781514

00:11:54.610 --> 00:11:55.970 particular with good,

NOTE Confidence: 0.91653496

00:11:57.570 --> 00:11:59.330 faulting capabilities. So that was

NOTE Confidence: 0.91653496

00:11:59.330 --> 00:12:00.150 quite impressive.

NOTE Confidence: 0.92221594

00:12:01.425 --> 00:12:02.705 And, also, as Steve mentioned

NOTE Confidence: 0.92221594

00:12:02.705 --> 00:12:03.505 in a bit, by now,

NOTE Confidence: 0.92221594

00:12:03.505 --> 00:12:04.945 we have a bunch of

NOTE Confidence: 0.92221594

00:12:04.945 --> 00:12:06.804 different, protein language models.

NOTE Confidence: 0.9691284

00:12:07.184 --> 00:12:08.645 Some of them are general

NOTE Confidence: 0.9691284

00:12:08.705 --> 00:12:09.905 like ESM. This is one

NOTE Confidence: 0.9691284

00:12:09.905 --> 00:12:11.184 of the earlier models. Well,

NOTE Confidence: 0.9691284

00:12:11.184 --> 00:12:11.905 now we are in the

NOTE Confidence: 0.9691284

00:12:11.905 --> 00:12:13.105 third version already, but this

NOTE Confidence: 0.9691284

00:12:13.105 --> 00:12:13.905 is one of the first

NOTE Confidence: 0.9691284

00:12:13.905 --> 00:12:14.885 models that appeared.

NOTE Confidence: 0.91682905

00:12:15.920 --> 00:12:17.200 This is the particularity of

NOTE Confidence: 0.91682905

00:12:17.200 --> 00:12:18.399 these models is this is

NOTE Confidence: 0.91682905

00:12:18.559 --> 00:12:19.679 this is probably trained on

NOTE Confidence: 0.91682905

00:12:19.679 --> 00:12:21.279 the largest collections of I

NOTE Confidence: 0.91682905

00:12:21.279 --> 00:12:22.899 mean, collection of amino acids.

NOTE Confidence: 0.9263223

00:12:23.679 --> 00:12:24.880 This is a super large,

NOTE Confidence: 0.9263223

00:12:24.880 --> 00:12:26.399 but also very heterogeneous. So

NOTE Confidence: 0.9263223

00:12:26.399 --> 00:12:27.600 you have proteins from all

NOTE Confidence: 0.9263223

00:12:27.600 --> 00:12:29.535 sort of organisms, tissues, conditions,

NOTE Confidence: 0.8401746

00:12:32.815 --> 00:12:33.694 opposed to that, you now

NOTE Confidence: 0.8401746

00:12:33.694 --> 00:12:34.975 will start also seeing models

NOTE Confidence: 0.8401746

00:12:34.975 --> 00:12:36.514 that are specific to particular

NOTE Confidence: 0.8401746

00:12:36.654 --> 00:12:38.654 families like TCR BERT, only

NOTE Confidence: 0.8401746

00:12:38.654 --> 00:12:40.274 trained on this TCR sequences

NOTE Confidence: 0.7401085

00:12:40.815 --> 00:12:42.254 and up lung trained on

NOTE Confidence: 0.7401085

00:12:42.254 --> 00:12:43.394 amino acid sequences.

NOTE Confidence: 0.98066074

00:12:44.550 --> 00:12:45.750 So thinking about it, we

NOTE Confidence: 0.98066074

00:12:45.750 --> 00:12:47.190 wanted not to train just

NOTE Confidence: 0.98066074

00:12:47.190 --> 00:12:48.470 a new protein language model

NOTE Confidence: 0.98066074

00:12:48.470 --> 00:12:49.270 because there were quite a

NOTE Confidence: 0.98066074

00:12:49.270 --> 00:12:50.470 few already, but we wanted

NOTE Confidence: 0.98066074

00:12:50.470 --> 00:12:51.990 to test how good this

NOTE Confidence: 0.98066074

00:12:51.990 --> 00:12:53.750 model were to represent immune

NOTE Confidence: 0.98066074

00:12:53.750 --> 00:12:54.950 receptors, which are, as I

NOTE Confidence: 0.98066074

00:12:54.950 --> 00:12:56.410 said before, they're quite particular

NOTE Confidence: 0.98066074

00:12:56.710 --> 00:12:57.530 and they are

NOTE Confidence: 0.9952147

00:12:57.990 --> 00:12:59.530 different than regular proteins.

NOTE Confidence: 0.9701966

00:13:00.365 --> 00:13:01.325 So we are looking both

NOTE Confidence: 0.9701966

00:13:01.325 --> 00:13:02.365 at T cell receptors and

NOTE Confidence: 0.9701966

00:13:02.365 --> 00:13:03.725 B cell receptors. So let

NOTE Confidence: 0.9701966

00:13:03.725 --> 00:13:04.684 me show you some of

NOTE Confidence: 0.9701966

00:13:04.684 --> 00:13:06.065 the things that we tested

NOTE Confidence: 0.9701966

00:13:06.204 --> 00:13:07.584 starting with T cell receptors.

NOTE Confidence: 0.9828829

00:13:08.684 --> 00:13:09.584 So for TCRs,

NOTE Confidence: 0.82053727

00:13:09.964 --> 00:13:12.125 we built, here we focused

NOTE Confidence: 0.82053727

00:13:12.365 --> 00:13:13.965 we exploited ESM, this general

NOTE Confidence: 0.82053727

00:13:13.965 --> 00:13:14.535 heterogeneous model that trained on

NOTE Confidence: 0.82053727

00:13:14.535 --> 00:13:16.320 the heterogeneous model that train

NOTE Confidence: 0.82053727

00:13:16.320 --> 00:13:17.440 on the largest collection of

NOTE Confidence: 0.82053727

00:13:17.440 --> 00:13:17.940 proteins.

NOTE Confidence: 0.9181985

00:13:18.320 --> 00:13:19.280 And here we train a

NOTE Confidence: 0.9181985

00:13:19.280 --> 00:13:20.640 super simple model. So it

NOTE Confidence: 0.9181985

00:13:20.640 --> 00:13:22.160 was ESM. We started the

NOTE Confidence: 0.9181985

00:13:22.160 --> 00:13:23.860 rep the the embedded representation

NOTE Confidence: 0.9181985

00:13:23.920 --> 00:13:24.740 of the receptors,

NOTE Confidence: 0.97770035

00:13:25.120 --> 00:13:26.160 and then we added a

NOTE Confidence: 0.97770035

00:13:26.160 --> 00:13:26.960 super simple,

NOTE Confidence: 0.9149918

00:13:27.360 --> 00:13:29.040 two neural two layered neural

NOTE Confidence: 0.9149918

00:13:29.040 --> 00:13:30.214 network. So it doesn't get

NOTE Confidence: 0.9149918

00:13:30.214 --> 00:13:31.274 simpler than that.

NOTE Confidence: 0.93689215

00:13:32.295 --> 00:13:33.415 And to test the capability

NOTE Confidence: 0.93689215

00:13:33.415 --> 00:13:34.695 of this model, we use

NOTE Confidence: 0.93689215

00:13:34.695 --> 00:13:36.454 a recent benchmark that on

NOTE Confidence: 0.93689215

00:13:36.454 --> 00:13:38.134 T cell receptor binding prediction

NOTE Confidence: 0.93689215

00:13:38.134 --> 00:13:40.054 models that was published last

NOTE Confidence: 0.93689215

00:13:40.054 --> 00:13:40.954 year in informatics.

NOTE Confidence: 0.9791372

00:13:41.574 --> 00:13:42.774 So in this benchmark so

NOTE Confidence: 0.9791372

00:13:42.774 --> 00:13:44.134 the plot the the dot

NOTE Confidence: 0.9791372

00:13:44.134 --> 00:13:45.360 plot comes from this benchmark,

NOTE Confidence: 0.9791372

00:13:45.519 --> 00:13:46.019 actually.

NOTE Confidence: 0.9867864

00:13:46.320 --> 00:13:47.600 So, basically, this plot here,

NOTE Confidence: 0.9867864

00:13:47.600 --> 00:13:48.720 we are showing two different

NOTE Confidence: 0.9867864

00:13:48.720 --> 00:13:50.640 predictive tasks. On the x

NOTE Confidence: 0.9867864

00:13:50.640 --> 00:13:52.240 axis, the task was a

NOTE Confidence: 0.9867864

00:13:52.240 --> 00:13:52.740 classification.

NOTE Confidence: 0.96617675

00:13:53.120 --> 00:13:54.480 So given a pair, so

NOTE Confidence: 0.96617675

00:13:54.480 --> 00:13:56.240 epitope TCR, does it bind?

NOTE Confidence: 0.96617675

00:13:56.240 --> 00:13:57.975 Yes or no? So, here,

NOTE Confidence: 0.96617675

00:13:57.975 --> 00:13:59.335 good models will have will

NOTE Confidence: 0.96617675

00:13:59.335 --> 00:14:00.714 be be close to this,

NOTE Confidence: 0.8686204

00:14:01.175 --> 00:14:02.295 ax to this side of

NOTE Confidence: 0.8686204

00:14:02.295 --> 00:14:02.955 the plot.

NOTE Confidence: 0.9250624

00:14:03.495 --> 00:14:04.455 And here, the task was

NOTE Confidence: 0.9250624

00:14:04.455 --> 00:14:05.975 an epitope ranking task. So

NOTE Confidence: 0.9250624

00:14:05.975 --> 00:14:07.015 the idea here that you

NOTE Confidence: 0.9250624

00:14:07.015 --> 00:14:08.214 have the for each cell

NOTE Confidence: 0.9250624

00:14:08.214 --> 00:14:09.735 receptor, you will have seventeen

NOTE Confidence: 0.9250624

00:14:09.735 --> 00:14:11.160 different lipidomes, and you have

NOTE Confidence: 0.9250624

00:14:11.160 --> 00:14:12.679 to rank depending according to

NOTE Confidence: 0.9250624

00:14:12.679 --> 00:14:14.279 the likelihood that each one

NOTE Confidence: 0.9250624

00:14:14.360 --> 00:14:15.480 what is the real by

NOTE Confidence: 0.9250624

00:14:15.720 --> 00:14:16.940 epitope that binds.

NOTE Confidence: 0.9824663

00:14:17.399 --> 00:14:19.080 So, basically, good predict good

NOTE Confidence: 0.9824663

00:14:19.080 --> 00:14:20.360 models are in this corner

NOTE Confidence: 0.9824663

00:14:20.360 --> 00:14:21.100 of the plot.

NOTE Confidence: 0.9192751

00:14:21.480 --> 00:14:22.279 And you can see just

NOTE Confidence: 0.9192751

00:14:22.279 --> 00:14:22.920 to to tell you a

NOTE Confidence: 0.9192751

00:14:22.920 --> 00:14:24.060 little bit about the benchmark.

NOTE Confidence: 0.9192751

00:14:24.155 --> 00:14:25.195 So the the there are

NOTE Confidence: 0.9192751

00:14:25.195 --> 00:14:26.715 three colors that depends whether

NOTE Confidence: 0.9192751

00:14:26.715 --> 00:14:27.675 the models were trained on

NOTE Confidence: 0.9192751

00:14:27.675 --> 00:14:28.875 single cell data. So for

NOTE Confidence: 0.9192751

00:14:28.875 --> 00:14:29.995 which we have the alpha

NOTE Confidence: 0.9192751

00:14:29.995 --> 00:14:31.215 beta appearance

NOTE Confidence: 0.942529

00:14:31.835 --> 00:14:32.955 or on bulk data for

NOTE Confidence: 0.942529

00:14:32.955 --> 00:14:34.075 which you have typically the

NOTE Confidence: 0.942529

00:14:34.075 --> 00:14:35.355 beta, sometimes also all of

NOTE Confidence: 0.942529

00:14:35.355 --> 00:14:36.255 the alpha chain.

NOTE Confidence: 0.93008584

00:14:36.630 --> 00:14:37.610 So not surprisingly,

NOTE Confidence: 0.80498755

00:14:38.070 --> 00:14:39.510 the single cell mode model

NOTE Confidence: 0.80498755

00:14:39.510 --> 00:14:40.630 stress of single cell data

NOTE Confidence: 0.80498755

00:14:40.630 --> 00:14:42.330 perform better. So that's expected.

NOTE Confidence: 0.9421453

00:14:42.950 --> 00:14:43.910 The nice thing with this

NOTE Confidence: 0.9421453

00:14:43.910 --> 00:14:45.350 is actually this model we

NOTE Confidence: 0.9421453

00:14:45.350 --> 00:14:46.710 trained. This is the purple,

NOTE Confidence: 0.9516876

00:14:47.030 --> 00:14:48.310 star. This model that we

NOTE Confidence: 0.9516876

00:14:48.310 --> 00:14:50.245 trained starting embedding from ESM

NOTE Confidence: 0.9516876

00:14:50.245 --> 00:14:51.765 and, again, very simple neural

NOTE Confidence: 0.9516876

00:14:51.765 --> 00:14:53.065 network on top of it.

NOTE Confidence: 0.9516876

00:14:53.205 --> 00:14:54.485 Actually, with very little effort,

NOTE Confidence: 0.9516876

00:14:54.485 --> 00:14:55.925 we were performing close to

NOTE Confidence: 0.9516876

00:14:55.925 --> 00:14:58.165 the, best performing models. So,

NOTE Confidence: 0.9516876

00:14:58.165 --> 00:14:59.205 again, this is just one

NOTE Confidence: 0.9516876

00:14:59.205 --> 00:15:00.485 particular task, but that gives

NOTE Confidence: 0.9516876

00:15:00.485 --> 00:15:01.205 you an idea of how

NOTE Confidence: 0.9516876

00:15:01.205 --> 00:15:02.565 much the accuracy you can

NOTE Confidence: 0.9516876

00:15:02.565 --> 00:15:03.690 you can gain with this

NOTE Confidence: 0.9516876

00:15:03.930 --> 00:15:06.010 extracting using exploiting this pretrained

NOTE Confidence: 0.9516876

00:15:06.010 --> 00:15:06.910 language models.

NOTE Confidence: 0.94768184

00:15:09.050 --> 00:15:10.170 We have also looked at

NOTE Confidence: 0.94768184

00:15:10.170 --> 00:15:11.290 B cell receptors in the

NOTE Confidence: 0.94768184

00:15:11.290 --> 00:15:13.230 context of, protein language models.

NOTE Confidence: 0.94768184

00:15:13.290 --> 00:15:15.210 So here, we're also looking

NOTE Confidence: 0.94768184

00:15:15.370 --> 00:15:16.250 wanted to look at the

NOTE Confidence: 0.94768184

00:15:16.250 --> 00:15:17.550 representation of capabilities

NOTE Confidence: 0.9735777

00:15:18.410 --> 00:15:19.310 of these models.

NOTE Confidence: 0.8918405

00:15:19.665 --> 00:15:21.185 Here, we use Uplank. This

NOTE Confidence: 0.8918405

00:15:21.185 --> 00:15:22.945 is a, a model for

NOTE Confidence: 0.8918405

00:15:22.945 --> 00:15:24.305 b cell protein language model

NOTE Confidence: 0.8918405

00:15:24.305 --> 00:15:25.345 trained on b cell receptor

NOTE Confidence: 0.8918405

00:15:25.345 --> 00:15:25.845 sequences.

NOTE Confidence: 0.90482724

00:15:26.305 --> 00:15:27.824 And, well, Steve, again, already

NOTE Confidence: 0.90482724

00:15:27.824 --> 00:15:29.665 introduced the clone evolution within

NOTE Confidence: 0.90482724

00:15:29.665 --> 00:15:31.105 b cells. So here, this

NOTE Confidence: 0.90482724

00:15:31.105 --> 00:15:32.464 is, we're going to just

NOTE Confidence: 0.90482724

00:15:32.464 --> 00:15:34.404 to compare use this exercise

NOTE Confidence: 0.90482724

00:15:34.545 --> 00:15:35.769 to compare how good the

NOTE Confidence: 0.90482724

00:15:35.769 --> 00:15:37.149 world is smallest to represent

NOTE Confidence: 0.90482724

00:15:37.290 --> 00:15:38.110 this grand

NOTE Confidence: 0.9763932

00:15:39.209 --> 00:15:40.730 evolution. So, basically, to explain

NOTE Confidence: 0.9763932

00:15:40.730 --> 00:15:42.570 this plot, this we took

NOTE Confidence: 0.9763932

00:15:42.570 --> 00:15:44.649 data from an experiment where

NOTE Confidence: 0.9763932

00:15:44.649 --> 00:15:46.350 some collaborators have sample,

NOTE Confidence: 0.87158835

00:15:47.690 --> 00:15:49.050 B cell receptor from individual

NOTE Confidence: 0.87158835

00:15:49.050 --> 00:15:50.155 germinal centers. Centers. So we

NOTE Confidence: 0.87158835

00:15:50.155 --> 00:15:51.755 could characterize the germinal center

NOTE Confidence: 0.87158835

00:15:51.755 --> 00:15:53.275 of origin of each B

NOTE Confidence: 0.87158835

00:15:53.275 --> 00:15:53.935 cell receptor.

NOTE Confidence: 0.9187065

00:15:55.275 --> 00:15:56.315 And then we could infer

NOTE Confidence: 0.9187065

00:15:56.395 --> 00:15:57.355 we had to infer high

NOTE Confidence: 0.9187065

00:15:57.355 --> 00:15:58.955 inflate the clonal clonal families

NOTE Confidence: 0.9187065

00:15:58.955 --> 00:15:59.855 of each receptor.

NOTE Confidence: 0.95033616

00:16:00.395 --> 00:16:01.435 So in a few cases,

NOTE Confidence: 0.95033616

00:16:01.435 --> 00:16:02.635 we observed, and this has

NOTE Confidence: 0.95033616

00:16:02.635 --> 00:16:04.449 been reported before, that a

NOTE Confidence: 0.95033616

00:16:04.610 --> 00:16:06.290 the same clone was split

NOTE Confidence: 0.95033616

00:16:06.290 --> 00:16:07.990 across different germinal centers.

NOTE Confidence: 0.8997741

00:16:08.370 --> 00:16:09.250 This is, again, this has

NOTE Confidence: 0.8997741

00:16:09.250 --> 00:16:10.850 been observed that happens when

NOTE Confidence: 0.8997741

00:16:10.850 --> 00:16:12.370 one cell is evolving in

NOTE Confidence: 0.8997741

00:16:12.370 --> 00:16:13.649 a germinal center and then

NOTE Confidence: 0.8997741

00:16:13.649 --> 00:16:15.490 migrates onto another germinal center

NOTE Confidence: 0.8997741

00:16:15.490 --> 00:16:17.025 and keeps evolving there. And

NOTE Confidence: 0.8997741

00:16:17.105 --> 00:16:17.665 And this is what this

NOTE Confidence: 0.8997741

00:16:17.665 --> 00:16:19.185 phylogenetic tree is showing you.

NOTE Confidence: 0.8997741

00:16:19.185 --> 00:16:20.145 So there are these three

NOTE Confidence: 0.8997741

00:16:20.145 --> 00:16:21.345 family these clones that are

NOTE Confidence: 0.8997741

00:16:21.345 --> 00:16:23.445 clonally related. So probably they

NOTE Confidence: 0.8997741

00:16:23.585 --> 00:16:24.785 they descend from the same

NOTE Confidence: 0.8997741

00:16:24.785 --> 00:16:25.285 ancestor

NOTE Confidence: 0.91889846

00:16:26.145 --> 00:16:26.645 ancestor,

NOTE Confidence: 0.6084299

00:16:27.185 --> 00:16:27.925 this BCR.

NOTE Confidence: 0.9101569

00:16:28.945 --> 00:16:29.985 And you see that the

NOTE Confidence: 0.9101569

00:16:29.985 --> 00:16:31.025 the blue clone probably is

NOTE Confidence: 0.9101569

00:16:31.025 --> 00:16:31.905 the one that is farther

NOTE Confidence: 0.9101569

00:16:31.905 --> 00:16:33.100 away. So it's once more

NOTE Confidence: 0.9101569

00:16:33.100 --> 00:16:34.860 separated than the others, and

NOTE Confidence: 0.9101569

00:16:34.860 --> 00:16:36.060 the purple and the green

NOTE Confidence: 0.9101569

00:16:36.060 --> 00:16:37.260 are a bit intermixed over

NOTE Confidence: 0.9101569

00:16:37.260 --> 00:16:37.760 here.

NOTE Confidence: 0.9352405

00:16:38.940 --> 00:16:40.780 So compared to this standard

NOTE Confidence: 0.9352405

00:16:40.780 --> 00:16:42.700 flow genetic inference, we just

NOTE Confidence: 0.9352405

00:16:42.700 --> 00:16:44.540 plot the this data with

NOTE Confidence: 0.9352405

00:16:44.540 --> 00:16:45.740 uplank and we just want

NOTE Confidence: 0.9352405

00:16:45.740 --> 00:16:46.700 it to visualize. So this

NOTE Confidence: 0.9352405

00:16:46.700 --> 00:16:48.400 is only a visualization exercise.

NOTE Confidence: 0.9352405

00:16:48.435 --> 00:16:49.895 There's no inference here.

NOTE Confidence: 0.9310355

00:16:50.275 --> 00:16:51.315 And you can see actually

NOTE Confidence: 0.9310355

00:16:51.315 --> 00:16:52.035 that in a at the

NOTE Confidence: 0.9310355

00:16:52.035 --> 00:16:54.035 qualitative level, this this plots

NOTE Confidence: 0.9310355

00:16:54.035 --> 00:16:55.175 have a lot of similarities.

NOTE Confidence: 0.9310355

00:16:55.475 --> 00:16:56.755 So the upland also can

NOTE Confidence: 0.9310355

00:16:56.755 --> 00:16:58.355 recapitulate quite well the three

NOTE Confidence: 0.9310355

00:16:58.355 --> 00:17:00.195 clonal families. The blue one

NOTE Confidence: 0.9310355

00:17:00.195 --> 00:17:00.915 seems to be a bit

NOTE Confidence: 0.9310355

00:17:00.915 --> 00:17:01.654 more separated

NOTE Confidence: 0.9230056

00:17:01.980 --> 00:17:02.940 and there seem to be

NOTE Confidence: 0.9230056

00:17:02.940 --> 00:17:03.920 this good intermixing

NOTE Confidence: 0.97587234

00:17:04.700 --> 00:17:05.740 between the purple and the

NOTE Confidence: 0.97587234

00:17:05.740 --> 00:17:07.419 green family. So again, this

NOTE Confidence: 0.97587234

00:17:07.419 --> 00:17:08.540 is not an inference tool,

NOTE Confidence: 0.97587234

00:17:08.540 --> 00:17:09.600 but as a representation

NOTE Confidence: 0.8624296

00:17:09.900 --> 00:17:10.720 test visualization,

NOTE Confidence: 0.8696842

00:17:11.020 --> 00:17:11.980 this seems that we are

NOTE Confidence: 0.8696842

00:17:11.980 --> 00:17:12.480 capturing

NOTE Confidence: 0.9660538

00:17:12.780 --> 00:17:13.900 a lot of the aspects

NOTE Confidence: 0.9660538

00:17:13.900 --> 00:17:15.179 of the clonal evolution that

NOTE Confidence: 0.9660538

00:17:15.179 --> 00:17:16.799 we'll capture with traditional phylogenetic

NOTE Confidence: 0.9660538

00:17:16.940 --> 00:17:17.955 inference tools.

NOTE Confidence: 0.9545226

00:17:20.034 --> 00:17:20.994 The last thing I would

NOTE Confidence: 0.9545226

00:17:20.994 --> 00:17:22.515 like to mention is, the

NOTE Confidence: 0.9545226

00:17:22.515 --> 00:17:23.635 the the question with which

NOTE Confidence: 0.9545226

00:17:23.635 --> 00:17:25.154 I started this section. So

NOTE Confidence: 0.9545226

00:17:25.154 --> 00:17:26.515 which model is better? If

NOTE Confidence: 0.9545226

00:17:26.515 --> 00:17:27.794 you are interested in predicting,

NOTE Confidence: 0.9545226

00:17:27.794 --> 00:17:30.294 for instance, BCR affinity binding,

NOTE Confidence: 0.9545226

00:17:30.515 --> 00:17:31.619 so what would be better?

NOTE Confidence: 0.9545226

00:17:31.619 --> 00:17:32.660 Should we use a general

NOTE Confidence: 0.9545226

00:17:32.660 --> 00:17:33.700 model or should we use

NOTE Confidence: 0.9545226

00:17:33.700 --> 00:17:35.640 a, BCR specific model?

NOTE Confidence: 0.9880902

00:17:36.020 --> 00:17:36.980 So this is an open

NOTE Confidence: 0.9880902

00:17:36.980 --> 00:17:37.480 question.

NOTE Confidence: 0.9569293

00:17:37.780 --> 00:17:38.900 I see I see many

NOTE Confidence: 0.9569293

00:17:38.900 --> 00:17:40.180 papers in each direction, so

NOTE Confidence: 0.9569293

00:17:40.180 --> 00:17:41.300 I'm not gonna give the

NOTE Confidence: 0.9569293

00:17:41.300 --> 00:17:42.740 final answer here. Probably there's

NOTE Confidence: 0.9569293

00:17:42.740 --> 00:17:44.100 no final answer. I can

NOTE Confidence: 0.9569293

00:17:44.100 --> 00:17:44.980 tell you only about the

NOTE Confidence: 0.9569293

00:17:44.980 --> 00:17:46.340 particular task that we did

NOTE Confidence: 0.9569293

00:17:46.340 --> 00:17:46.840 here.

NOTE Confidence: 0.9453527

00:17:47.355 --> 00:17:48.554 So here, we took a

NOTE Confidence: 0.9453527

00:17:48.554 --> 00:17:50.734 dataset of seventy thousand antibodies.

NOTE Confidence: 0.9161846

00:17:51.355 --> 00:17:52.875 They with with measure binding

NOTE Confidence: 0.9161846

00:17:52.875 --> 00:17:54.155 affinity to a SARS CoV-two,

NOTE Confidence: 0.9161846

00:17:54.475 --> 00:17:54.975 peptide.

NOTE Confidence: 0.9473108

00:17:55.515 --> 00:17:56.635 And then we train again

NOTE Confidence: 0.9473108

00:17:56.635 --> 00:17:58.315 a simple model using both

NOTE Confidence: 0.9473108

00:17:58.315 --> 00:17:59.755 embeddings from the ESM, the

NOTE Confidence: 0.9473108

00:17:59.755 --> 00:18:01.595 general model, and Upland, the

NOTE Confidence: 0.9473108

00:18:01.595 --> 00:18:02.970 VCR specific model.

NOTE Confidence: 0.9681488

00:18:03.450 --> 00:18:04.570 And then the thing here,

NOTE Confidence: 0.9681488

00:18:04.570 --> 00:18:05.850 we wanted to test the

NOTE Confidence: 0.9681488

00:18:05.850 --> 00:18:07.450 accuracy of each model at

NOTE Confidence: 0.9681488

00:18:07.450 --> 00:18:07.850 different,

NOTE Confidence: 0.92148435

00:18:08.490 --> 00:18:09.850 with with different amount of

NOTE Confidence: 0.92148435

00:18:09.850 --> 00:18:10.970 data. So here, we are

NOTE Confidence: 0.92148435

00:18:10.970 --> 00:18:12.270 taking just a

NOTE Confidence: 0.67652506

00:18:13.210 --> 00:18:14.109 piece of the

NOTE Confidence: 0.93405676

00:18:15.095 --> 00:18:16.215 pieces of the data from

NOTE Confidence: 0.93405676

00:18:16.215 --> 00:18:17.335 here for little data to

NOTE Confidence: 0.93405676

00:18:17.335 --> 00:18:18.535 going to largest amount of

NOTE Confidence: 0.93405676

00:18:18.535 --> 00:18:19.035 data.

NOTE Confidence: 0.9880825

00:18:19.494 --> 00:18:20.375 And as you can see

NOTE Confidence: 0.9880825

00:18:20.375 --> 00:18:21.734 here, when the data is

NOTE Confidence: 0.9880825

00:18:21.734 --> 00:18:23.335 small, so the the difference

NOTE Confidence: 0.9880825

00:18:23.335 --> 00:18:24.375 with the model is not

NOTE Confidence: 0.9880825

00:18:24.375 --> 00:18:24.875 significant.

NOTE Confidence: 0.8717214

00:18:25.255 --> 00:18:26.295 But as I the the

NOTE Confidence: 0.8717214

00:18:26.295 --> 00:18:27.755 amount of data starts increasing,

NOTE Confidence: 0.8717214

00:18:27.815 --> 00:18:29.255 in this particular task, we

NOTE Confidence: 0.8717214

00:18:29.255 --> 00:18:30.295 found the SM to be

NOTE Confidence: 0.8717214

00:18:30.295 --> 00:18:30.780 better.

NOTE Confidence: 0.9803791

00:18:31.260 --> 00:18:32.300 So, again, this is an

NOTE Confidence: 0.9803791

00:18:32.300 --> 00:18:33.340 open debate, and I've seen

NOTE Confidence: 0.9803791

00:18:33.340 --> 00:18:34.460 papers in both directions. I

NOTE Confidence: 0.9803791

00:18:34.460 --> 00:18:35.340 don't claim that this is

NOTE Confidence: 0.9803791

00:18:35.340 --> 00:18:36.240 the final answer.

NOTE Confidence: 0.8934646

00:18:36.700 --> 00:18:37.740 What I would take from

NOTE Confidence: 0.8934646

00:18:37.740 --> 00:18:39.020 here is probably that that

NOTE Confidence: 0.8934646

00:18:39.100 --> 00:18:40.300 probably the answer depends on

NOTE Confidence: 0.8934646

00:18:40.300 --> 00:18:42.000 the task. Probably high specificity

NOTE Confidence: 0.8934646

00:18:42.220 --> 00:18:44.060 task benefit from a specific

NOTE Confidence: 0.8934646

00:18:44.060 --> 00:18:45.600 mod language models.

NOTE Confidence: 0.9655484

00:18:46.065 --> 00:18:47.345 But in any case, when

NOTE Confidence: 0.9655484

00:18:47.345 --> 00:18:48.705 the more data you have

NOTE Confidence: 0.9655484

00:18:48.705 --> 00:18:50.065 to fine tune the general

NOTE Confidence: 0.9655484

00:18:50.065 --> 00:18:51.345 model, the more likely you

NOTE Confidence: 0.9655484

00:18:51.345 --> 00:18:52.465 will be able to extract

NOTE Confidence: 0.9655484

00:18:52.465 --> 00:18:54.065 whatever information is hidden in

NOTE Confidence: 0.9655484

00:18:54.065 --> 00:18:55.585 this large heterogeneous collection of

NOTE Confidence: 0.9655484

00:18:55.585 --> 00:18:56.085 proteins.

NOTE Confidence: 0.8881872

00:18:57.105 --> 00:18:58.144 So, again, this is there

NOTE Confidence: 0.8881872

00:18:58.144 --> 00:18:59.264 were many other variables that

NOTE Confidence: 0.8881872

00:18:59.264 --> 00:19:00.544 we entered in in impacting

NOTE Confidence: 0.8881872

00:19:00.544 --> 00:19:02.190 the the the accuracy, but,

NOTE Confidence: 0.8881872

00:19:02.429 --> 00:19:03.470 this is the the message

NOTE Confidence: 0.8881872

00:19:03.470 --> 00:19:04.530 we got from here.

NOTE Confidence: 0.94733715

00:19:06.190 --> 00:19:07.309 Okay. So I think we're

NOTE Confidence: 0.94733715

00:19:07.309 --> 00:19:08.830 running out of time. So

NOTE Confidence: 0.94733715

00:19:08.830 --> 00:19:09.790 maybe I skip this, and

NOTE Confidence: 0.94733715

00:19:09.790 --> 00:19:10.350 I will tell you a

NOTE Confidence: 0.94733715

00:19:10.350 --> 00:19:11.090 little bit.

NOTE Confidence: 0.98218656

00:19:11.790 --> 00:19:13.005 Yeah. So maybe I will

NOTE Confidence: 0.98218656

00:19:13.085 --> 00:19:13.565 conclude.

NOTE Confidence: 0.96358395

00:19:13.965 --> 00:19:14.925 The last part was about

NOTE Confidence: 0.96358395

00:19:14.925 --> 00:19:16.365 interpretability, but maybe leave it

NOTE Confidence: 0.96358395

00:19:16.365 --> 00:19:17.425 for future discussions,

NOTE Confidence: 0.9995968

00:19:17.805 --> 00:19:19.265 and I will conclude here.

NOTE Confidence: 0.9596949

00:19:19.565 --> 00:19:20.845 So, basically, the the message

NOTE Confidence: 0.9596949

00:19:20.845 --> 00:19:21.565 I would like you to

NOTE Confidence: 0.9596949

00:19:21.565 --> 00:19:22.365 take home is that there

NOTE Confidence: 0.9596949

00:19:22.365 --> 00:19:23.484 is a lot of exciting

NOTE Confidence: 0.9596949

00:19:23.484 --> 00:19:25.744 opportunities to develop, both traditional

NOTE Confidence: 0.9596949

00:19:25.805 --> 00:19:27.484 computational models and machine learning

NOTE Confidence: 0.9596949

00:19:27.484 --> 00:19:29.340 deep learning models to properties

NOTE Confidence: 0.9596949

00:19:29.340 --> 00:19:30.400 of immune receptors.

NOTE Confidence: 0.96002585

00:19:31.420 --> 00:19:32.460 We are we have been

NOTE Confidence: 0.96002585

00:19:32.460 --> 00:19:32.859 developing

NOTE Confidence: 0.8949267

00:19:33.260 --> 00:19:34.140 working a lot on this

NOTE Confidence: 0.8949267

00:19:34.140 --> 00:19:35.660 area. We have developed models

NOTE Confidence: 0.8949267

00:19:35.660 --> 00:19:36.859 such as Taitan that was

NOTE Confidence: 0.8949267

00:19:36.859 --> 00:19:37.520 a multimodal,

NOTE Confidence: 0.95833254

00:19:37.900 --> 00:19:39.580 more traditional deep learning approach

NOTE Confidence: 0.95833254

00:19:39.580 --> 00:19:41.280 to to predict the specificity

NOTE Confidence: 0.85213584

00:19:41.660 --> 00:19:43.040 with, and with interpretability.

NOTE Confidence: 0.94044876

00:19:44.575 --> 00:19:46.015 Then we have been exploring

NOTE Confidence: 0.94044876

00:19:46.015 --> 00:19:47.375 the the capabilities of protein

NOTE Confidence: 0.94044876

00:19:47.375 --> 00:19:48.175 language models. As I said,

NOTE Confidence: 0.94044876

00:19:48.175 --> 00:19:49.135 there is a lot of

NOTE Confidence: 0.94044876

00:19:49.135 --> 00:19:50.015 work that we can do

NOTE Confidence: 0.94044876

00:19:50.015 --> 00:19:51.215 in this area. And when

NOTE Confidence: 0.94044876

00:19:51.215 --> 00:19:51.935 this is the part I

NOTE Confidence: 0.94044876

00:19:51.935 --> 00:19:52.734 did not take did not

NOTE Confidence: 0.94044876

00:19:52.734 --> 00:19:53.535 have time to tell you

NOTE Confidence: 0.94044876

00:19:53.535 --> 00:19:54.575 about, but we have also

NOTE Confidence: 0.94044876

00:19:54.575 --> 00:19:55.795 developed, interpretability

NOTE Confidence: 0.99661666

00:19:56.255 --> 00:19:58.255 pipelines to extract more human

NOTE Confidence: 0.99661666

00:19:58.255 --> 00:19:58.755 understandable

NOTE Confidence: 0.9160669

00:19:59.419 --> 00:20:00.859 explanations of why why a

NOTE Confidence: 0.9160669

00:20:00.859 --> 00:20:02.220 model predict what the model

NOTE Confidence: 0.9160669

00:20:02.220 --> 00:20:02.720 predicts.

NOTE Confidence: 0.93071175

00:20:03.900 --> 00:20:05.100 And maybe the last slide.

NOTE Confidence: 0.93071175

00:20:05.100 --> 00:20:05.899 So I wanted to tell

NOTE Confidence: 0.93071175

00:20:05.899 --> 00:20:06.700 you a little bit about

NOTE Confidence: 0.93071175

00:20:06.700 --> 00:20:08.539 future work. So I I

NOTE Confidence: 0.93071175

00:20:08.539 --> 00:20:09.580 started this talk by telling

NOTE Confidence: 0.93071175

00:20:09.580 --> 00:20:10.940 you that most model until

NOTE Confidence: 0.93071175

00:20:10.940 --> 00:20:12.640 now neglect the structure.

NOTE Confidence: 0.89511865

00:20:13.174 --> 00:20:14.054 I think the way to

NOTE Confidence: 0.89511865

00:20:14.054 --> 00:20:14.855 move forward, I think we

NOTE Confidence: 0.89511865

00:20:14.855 --> 00:20:15.494 are we are seeing now

NOTE Confidence: 0.89511865

00:20:15.494 --> 00:20:17.034 there's a saturation on accuracy.

NOTE Confidence: 0.89511865

00:20:17.095 --> 00:20:18.455 It's difficult to go farther.

NOTE Confidence: 0.89511865

00:20:18.455 --> 00:20:19.494 We have rich model that

NOTE Confidence: 0.89511865

00:20:19.494 --> 00:20:20.695 the models that predict very

NOTE Confidence: 0.89511865

00:20:20.695 --> 00:20:21.815 well what they can predict,

NOTE Confidence: 0.89511865

00:20:21.815 --> 00:20:23.174 but they cannot predict binding

NOTE Confidence: 0.89511865

00:20:23.174 --> 00:20:24.234 to unseen epitopes.

NOTE Confidence: 0.9368536

00:20:24.855 --> 00:20:26.054 One possibility to move forward

NOTE Confidence: 0.9368536

00:20:26.054 --> 00:20:27.174 to is to to integrate

NOTE Confidence: 0.9368536

00:20:27.174 --> 00:20:28.250 a structure. But how to

NOTE Confidence: 0.9368536

00:20:28.250 --> 00:20:29.530 do that is tricky because

NOTE Confidence: 0.9368536

00:20:29.530 --> 00:20:30.570 it's not about the lack

NOTE Confidence: 0.9368536

00:20:30.570 --> 00:20:31.450 of a structure, it's just

NOTE Confidence: 0.9368536

00:20:31.450 --> 00:20:32.810 the flexibility of the loop

NOTE Confidence: 0.9368536

00:20:32.810 --> 00:20:34.170 that is difficult. So certainly,

NOTE Confidence: 0.9368536

00:20:34.170 --> 00:20:35.530 I've seen paper that try

NOTE Confidence: 0.9368536

00:20:35.530 --> 00:20:36.810 to adapt alpha fold in

NOTE Confidence: 0.9368536

00:20:36.810 --> 00:20:38.170 different ways, but I think

NOTE Confidence: 0.9368536

00:20:38.170 --> 00:20:39.290 we have to go beyond

NOTE Confidence: 0.9368536

00:20:39.290 --> 00:20:40.730 alpha fold and integrate models

NOTE Confidence: 0.9368536

00:20:40.730 --> 00:20:41.950 that account the flexibility.

NOTE Confidence: 0.9216092

00:20:42.455 --> 00:20:43.335 And this requires a lot

NOTE Confidence: 0.9216092

00:20:43.335 --> 00:20:44.395 of more more thinking.

NOTE Confidence: 0.90512764

00:20:45.015 --> 00:20:46.855 And, certainly, interpretability is another

NOTE Confidence: 0.90512764

00:20:46.855 --> 00:20:47.895 area that what I'm very

NOTE Confidence: 0.90512764

00:20:47.895 --> 00:20:49.575 interested. If you are interested

NOTE Confidence: 0.90512764

00:20:49.575 --> 00:20:50.855 in, I'm certainly looking for

NOTE Confidence: 0.90512764

00:20:50.855 --> 00:20:52.695 new collaborations. So please contact

NOTE Confidence: 0.90512764

00:20:52.695 --> 00:20:53.275 to me.

NOTE Confidence: 0.9122752

00:20:54.067 --> 00:20:54.946 And with that, I would

NOTE Confidence: 0.9122752

00:20:54.946 --> 00:20:56.467 like to just, thank the

NOTE Confidence: 0.9122752

00:20:56.467 --> 00:20:57.587 people. This work most of

NOTE Confidence: 0.9122752

00:20:57.587 --> 00:20:58.707 the work was not IBM,

NOTE Confidence: 0.9122752

00:20:58.707 --> 00:20:59.827 so thank the people who

NOTE Confidence: 0.9122752

00:20:59.827 --> 00:21:01.207 did the work, funding.

NOTE Confidence: 0.923973

00:21:01.987 --> 00:21:03.127 I'm happy to take questions

NOTE Confidence: 0.923973

00:21:03.186 --> 00:21:03.847 if any.