

On the application of TCR- epitope prediction models

Pieter Meysman
28 September 2023
AIRR seminar

Epitope specificity

Determine the epitope target of TCR sequences

- Train a machine learning model to identify those TCR sequences that target a specific epitope

TRBV4-2*01

V region

CASSQRRANSPLHF

CDR3 region

TRBJ1-6*02

J region

+

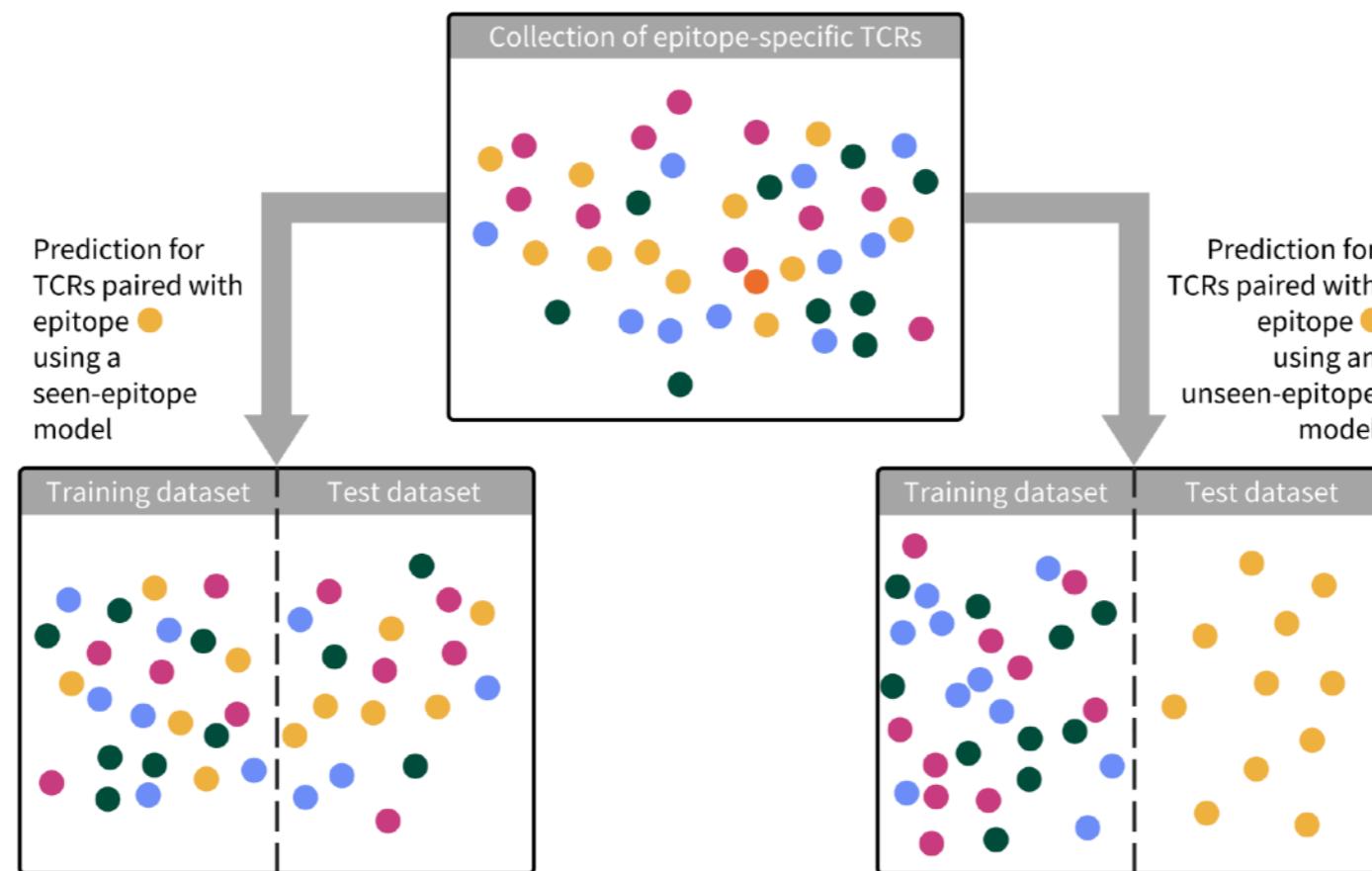
IIFLFILLLCLIFLL

Hepatitis B surface antigen epitope



Binding?

TCR-epitope prediction



- Seen epitope model

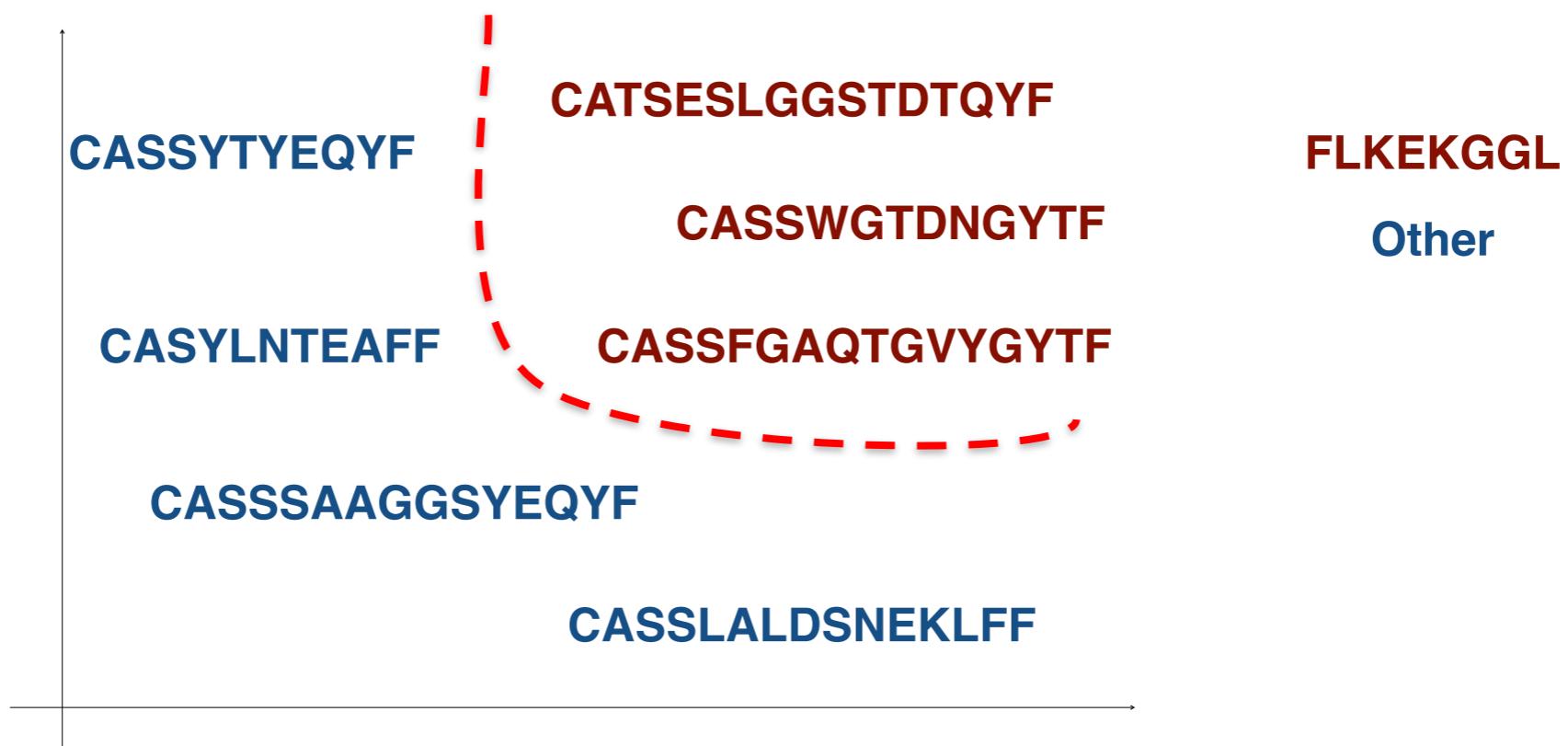
E.g. **TCRex**

- Unseen epitope model

E.g. **ImRex**

Seen epitope prediction

Binary prediction: identify those TCRs that are sufficiently similar to TCRs matched with the epitope in the training data set.

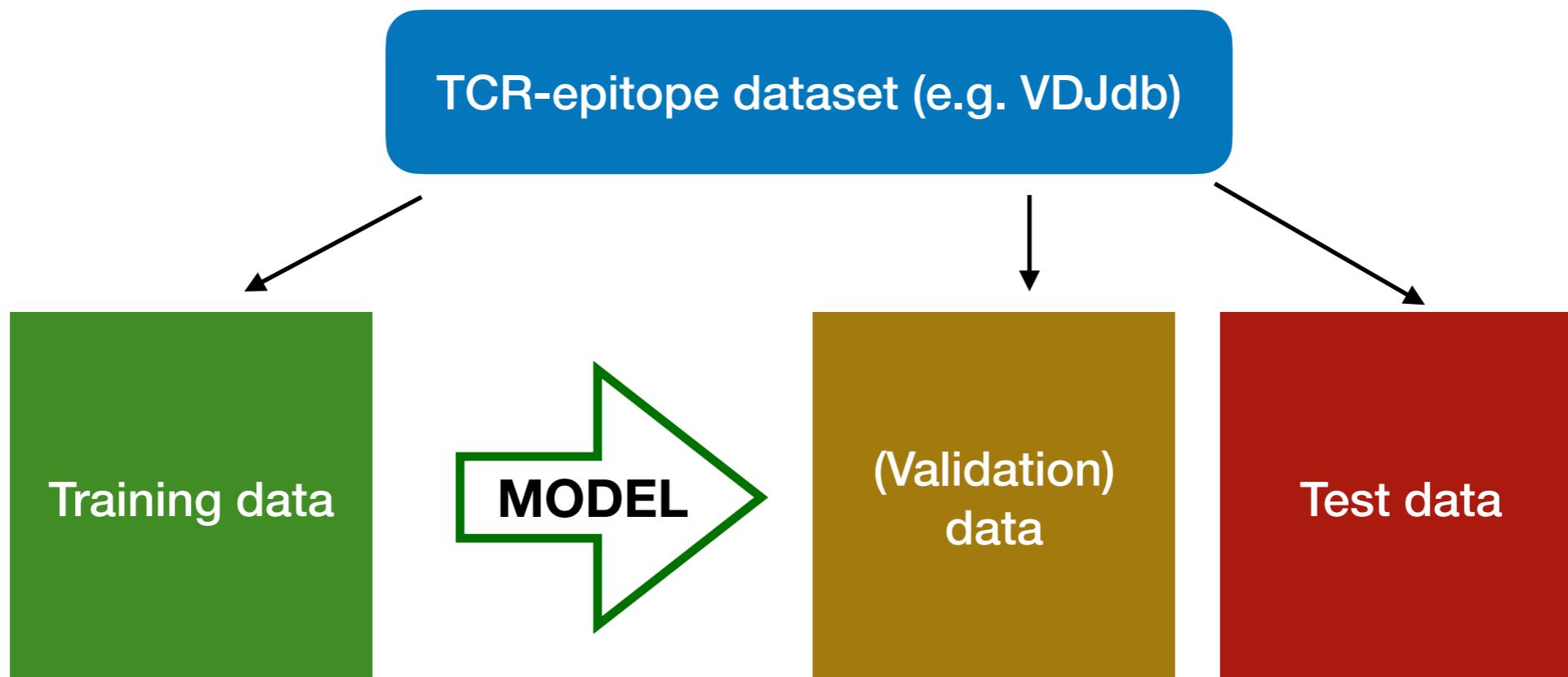


Bold claim #1:

We are all (including myself) evaluating the performance of TCR-epitope prediction models wrong!

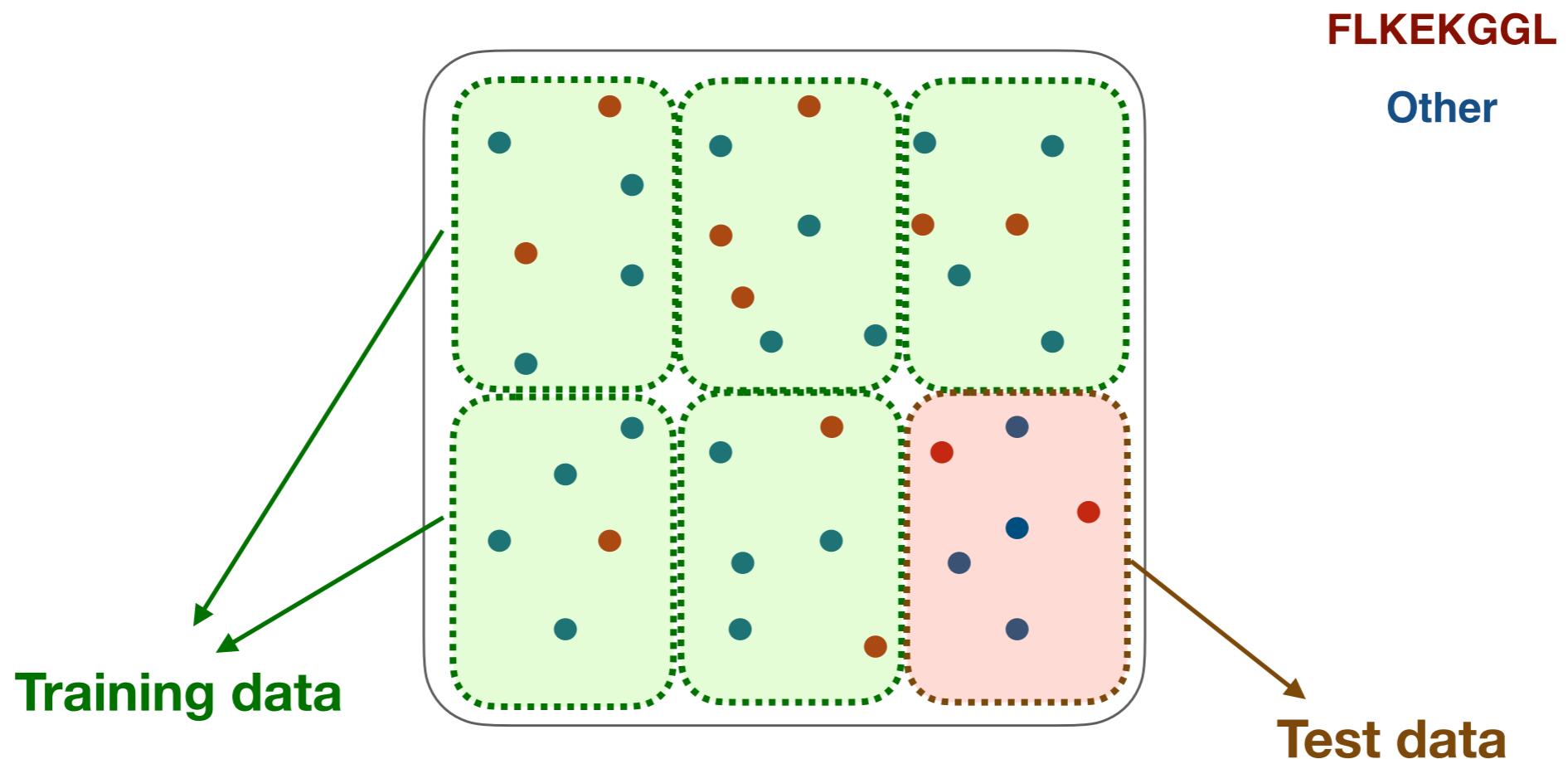
Evaluation

- Machine learning model evaluation



Evaluation

- Typical setup cross validation



- But where do the negatives come from?

Negative TCRs

- Every TCR (from a non-naive cell) binds something...

V gene	J gene	CDR3	Epitope
TRBV4-1	TRBJ1-5	CASSQATGGGQPQHF	TPINLVRDL
TRBV19-1	TRBJ1-2	CASKGGSFYGYTF	KLPDDFTGCV
TRBV12-3/12-4	TRBJ2-3	CASQRGYTDTQYF	SLVKPSFYV
TRBV20	TRBJ2-3	CSARGSADTQYF	MPASWVMRI
TRBV19-1	TRBJ1-5	CASSISGGQPQHF	KPLEFGATSAAL
TRBV5-6	TRBJ2-1	CASSLRLAGGVDEQFF	KPLEFGATSAAL
TRBV18-1	TRBJ2-7	CASSQPGQGIHEQYF	RQLLFVVEV
TRBV9-1	TRBJ2-5	CASSVGTGETQYF	FLPRVFSAV
TRBV7-9	TRBJ1-2	CASSLETHGYTF	FLNGSCGSV
TRBV7-2	TRBJ1-5	CASSSPRTGNQPQHF	FVDGVPFVV
TRBV7-3	TRBJ2-3	CASSPSLGTDTQYF	TLVPQEHYV
TRBV27-1	TRBJ2-7	CASSSLAEGYEQYF	LEPLVLDPI
TRBV5-4	TRBJ2-1	CASSRDYPNEQFF	KLPDDFTGCV
TRBV4-1	TRBJ2-7	CASSQGLRGEQYF	LPAADLDDF

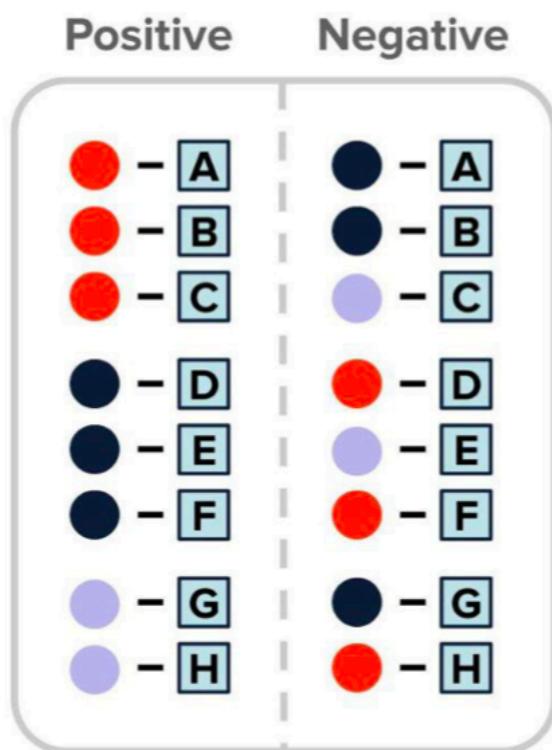
Negative TCRs

- Every TCR (from a non-naive cell) binds something...
we just don't always know it

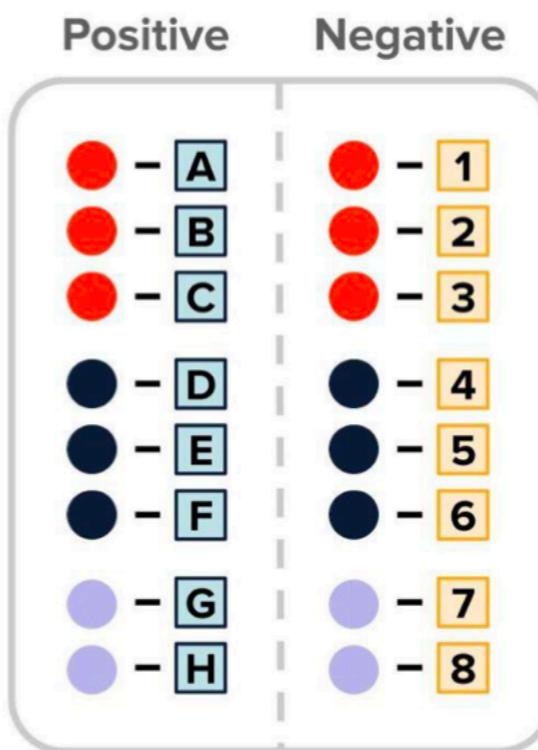
V gene	J gene	CDR3	Epitope
TRBV4-1	TRBJ1-5	CASSQATGGGQPQHF	
TRBV19-1	TRBJ1-2	CASKGGSFYGYTF	
TRBV12-3/12-4	TRBJ2-3	CASQRGYTDTQYF	
TRBV20	TRBJ2-3	CSARGSADTQYF	
TRBV19-1	TRBJ1-5	CASSISGGQPQHF	KPLEFGATSAAL
TRBV5-6	TRBJ2-1	CASSLRLAGGVDEQFF	
TRBV18-1	TRBJ2-7	CASSQPGQGIHEQYF	
TRBV9-1	TRBJ2-5	CASSVGTGETQYF	
TRBV7-9	TRBJ1-2	CASSLETHGYTF	
TRBV7-2	TRBJ1-5	CASSSPRTGNQPQHF	
TRBV7-3	TRBJ2-3	CASSPSLGTDTQYF	TLVPQEHYV
TRBV27-1	TRBJ2-7	CASSSLAEGYEQYF	
TRBV5-4	TRBJ2-1	CASSRDYPNEQFF	
TRBV4-1	TRBJ2-7	CASSQGLRGEQYF	

Negative data

Negatives by shuffling



Negatives from background data

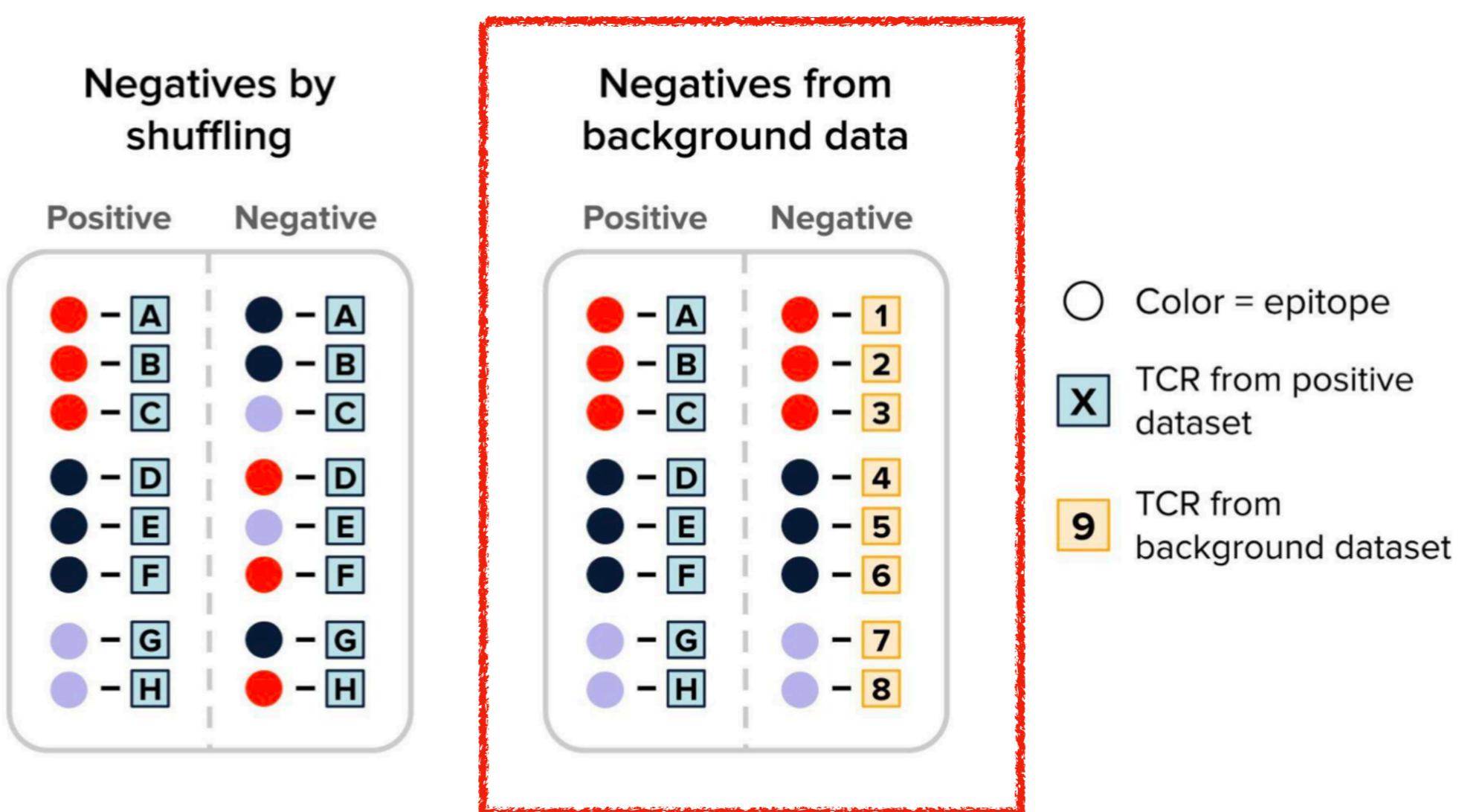


- Color = epitope
- ☒ TCR from positive dataset
- ▣ TCR from background dataset



Dens C et al. (2023). The pitfalls of negative data bias for the T-cell epitope specificity challenge. Nature Machine Intelligence (in press).

Negative data



Machine learning methods learn the difference between “background” and tetramer-sorted TCRs



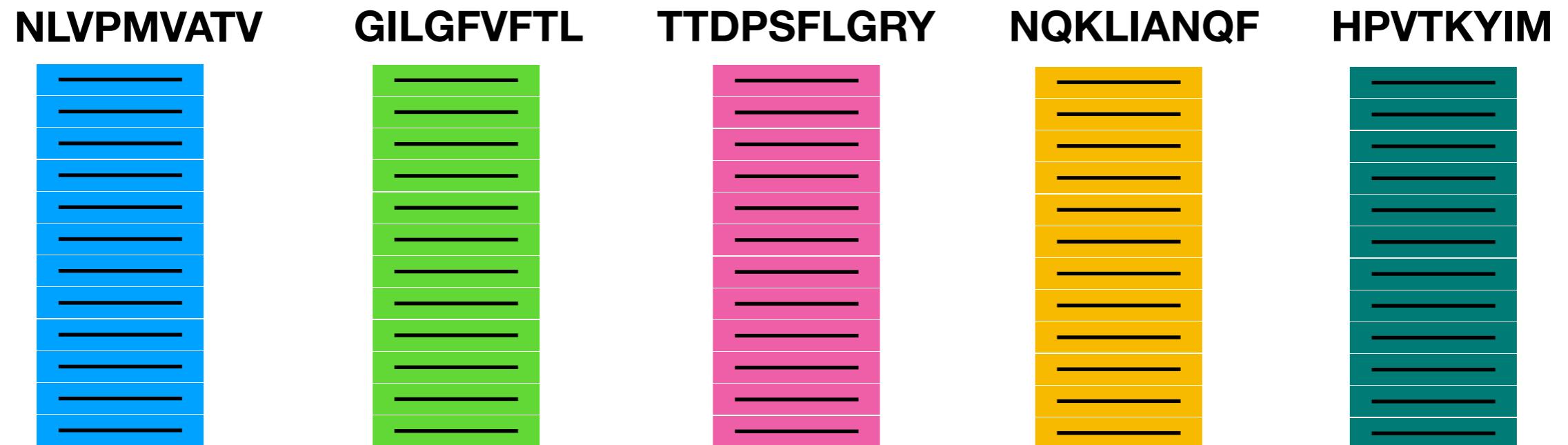
Dens C et al. (2023). The pitfalls of negative data bias for the T-cell epitope specificity challenge. Nature Machine Intelligence (in press).

Bold claim #2:

Epitope-shuffled negatives is preferred for
training.

Performant models should work on test
datasets from both strategies.

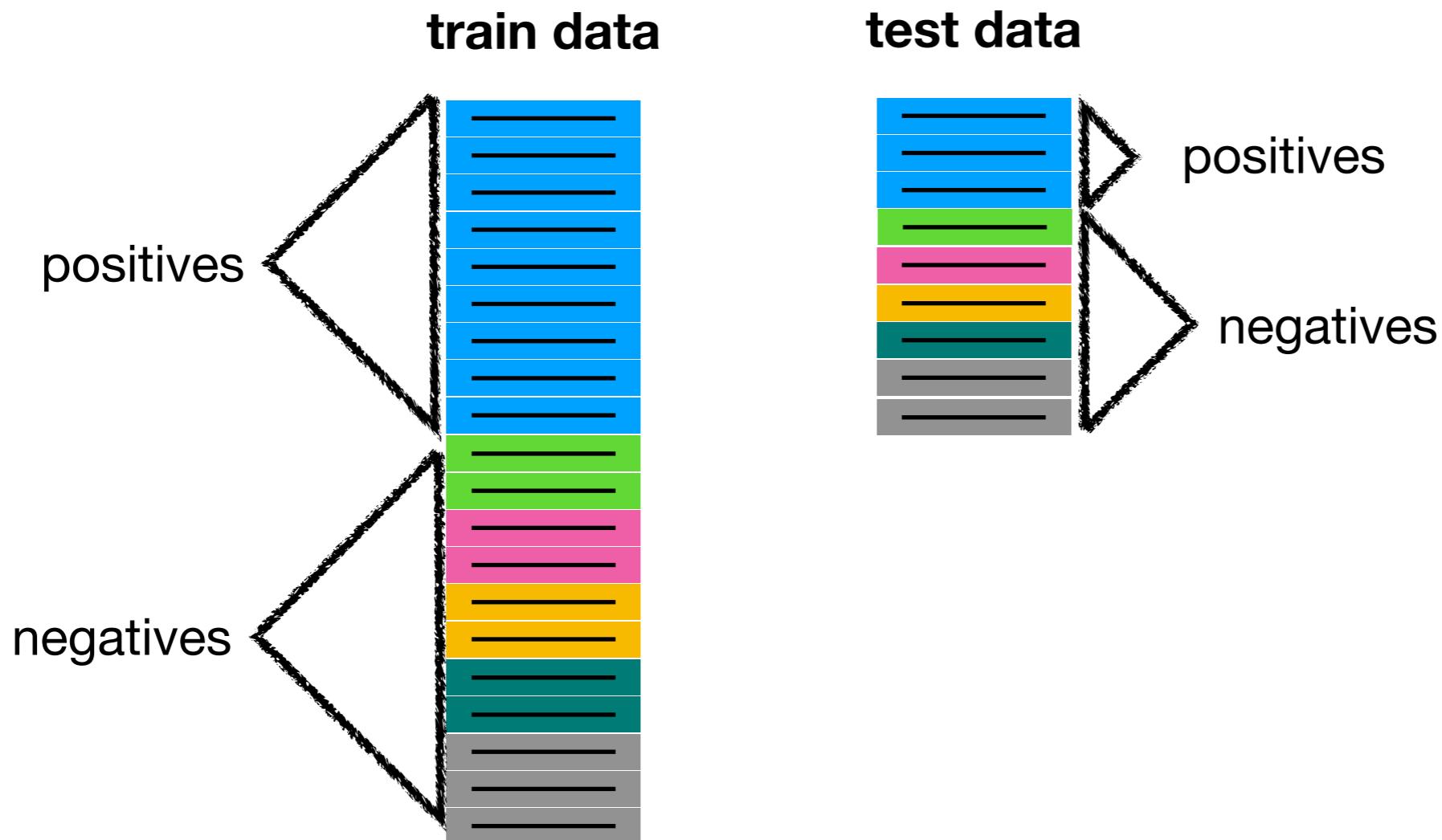
One example: the IMMREP benchmark



...

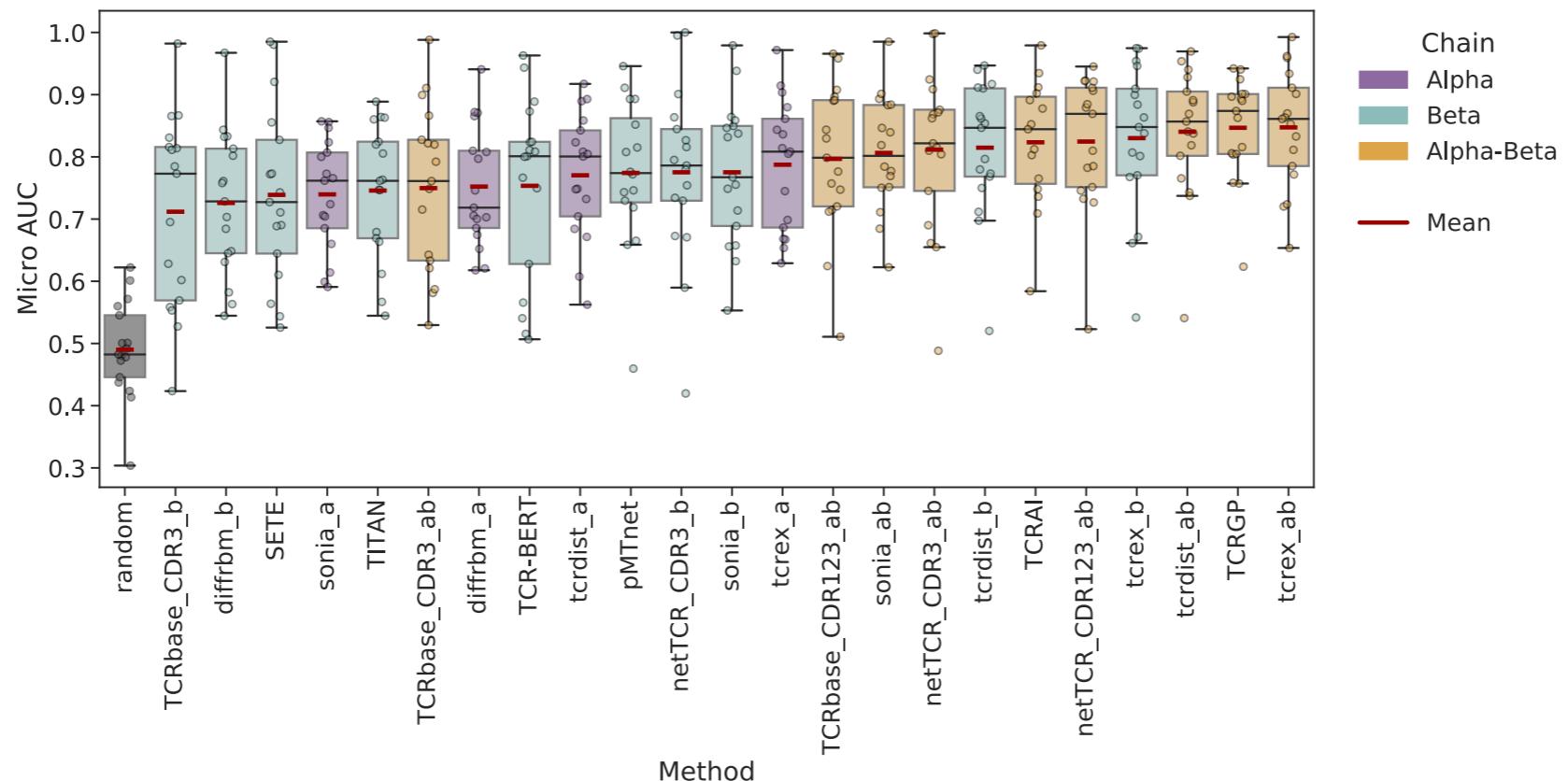
One example: the IMMREP benchmark

NLVPMVATV

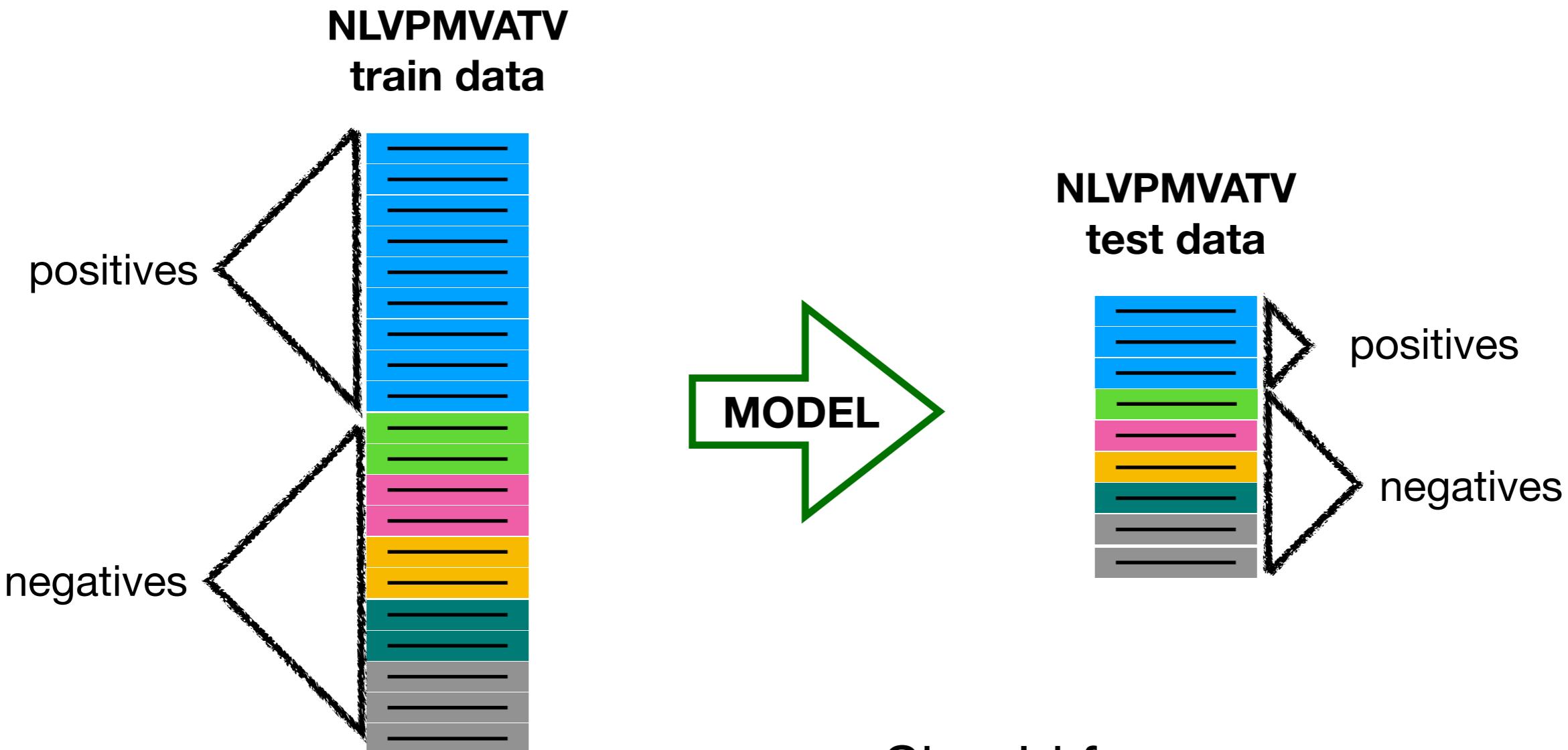


Evaluation

- Four best performing models are all ‘single epitope models’

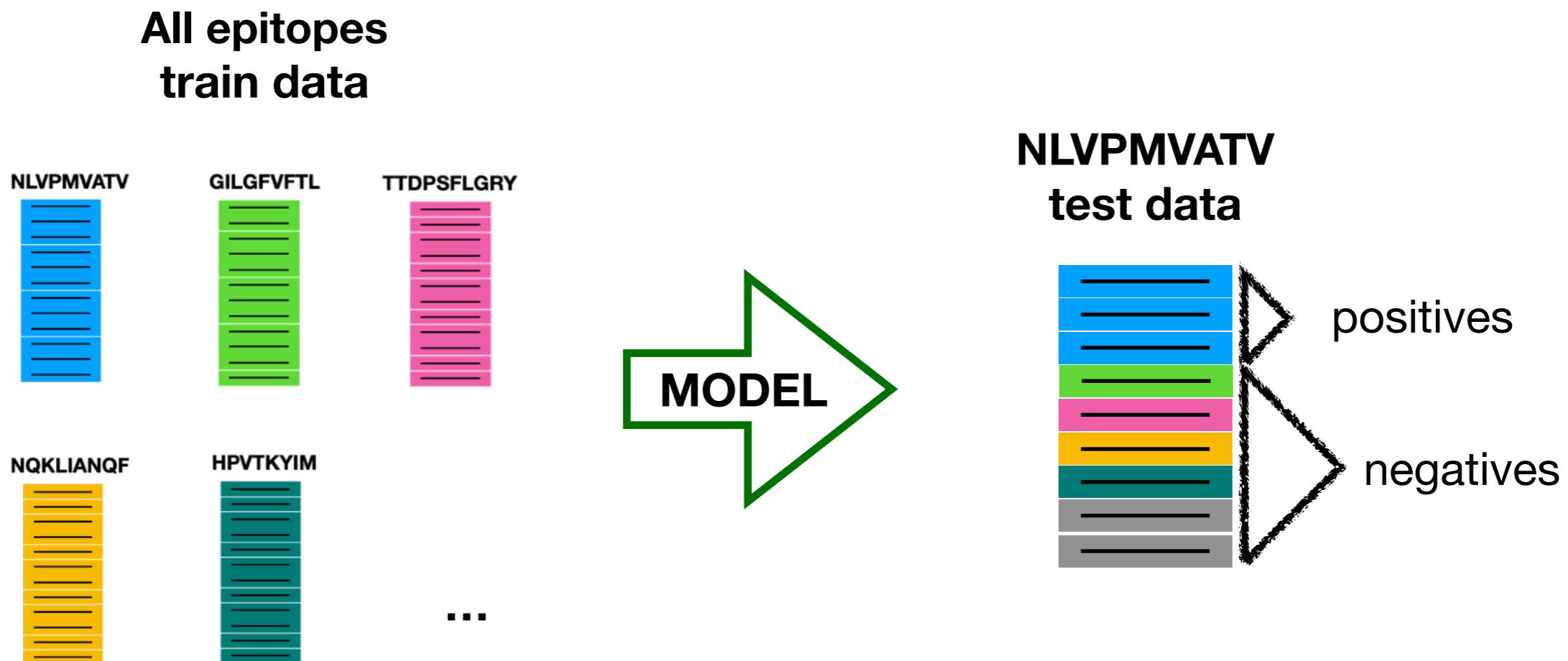


Single-epitope model



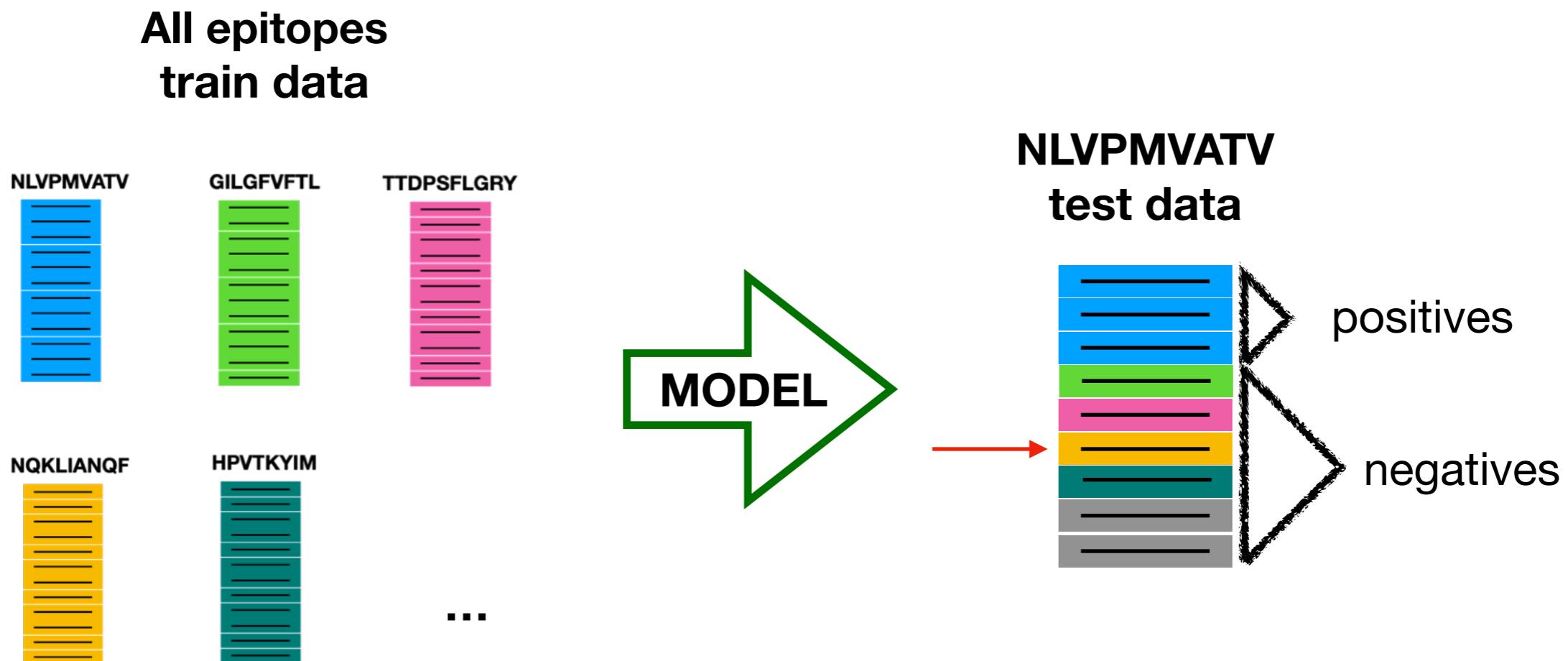
- Should focus on common patterns in the **positives**.

Multi-epitope model



- Negatives are from other epitopes (that they have seen!)
- But predictions are worse?

Multi-epitope model



- Last time it saw something like ———, it was a positive sample

Model viewpoint

Training data



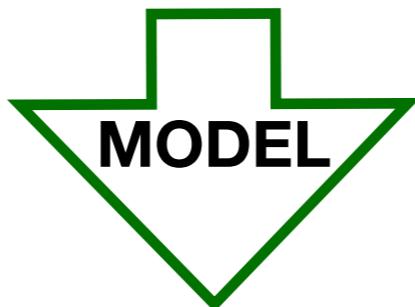
NQKLIANQF

Positive



NLVPMVATV

Positive



Test data



NLVPMVATV

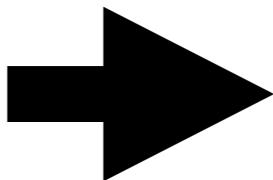
? (Negative)

Novel generation

TCRex



IMW DETECT



- Released in 2017
- Single-epitope model
- Developed in 2021
- Multi-epitope model
- + Many changes

IMW DETECT



Given the information in this TCR sequence:

TRBV4-2*01

V region

CASSQRRANSPLHF

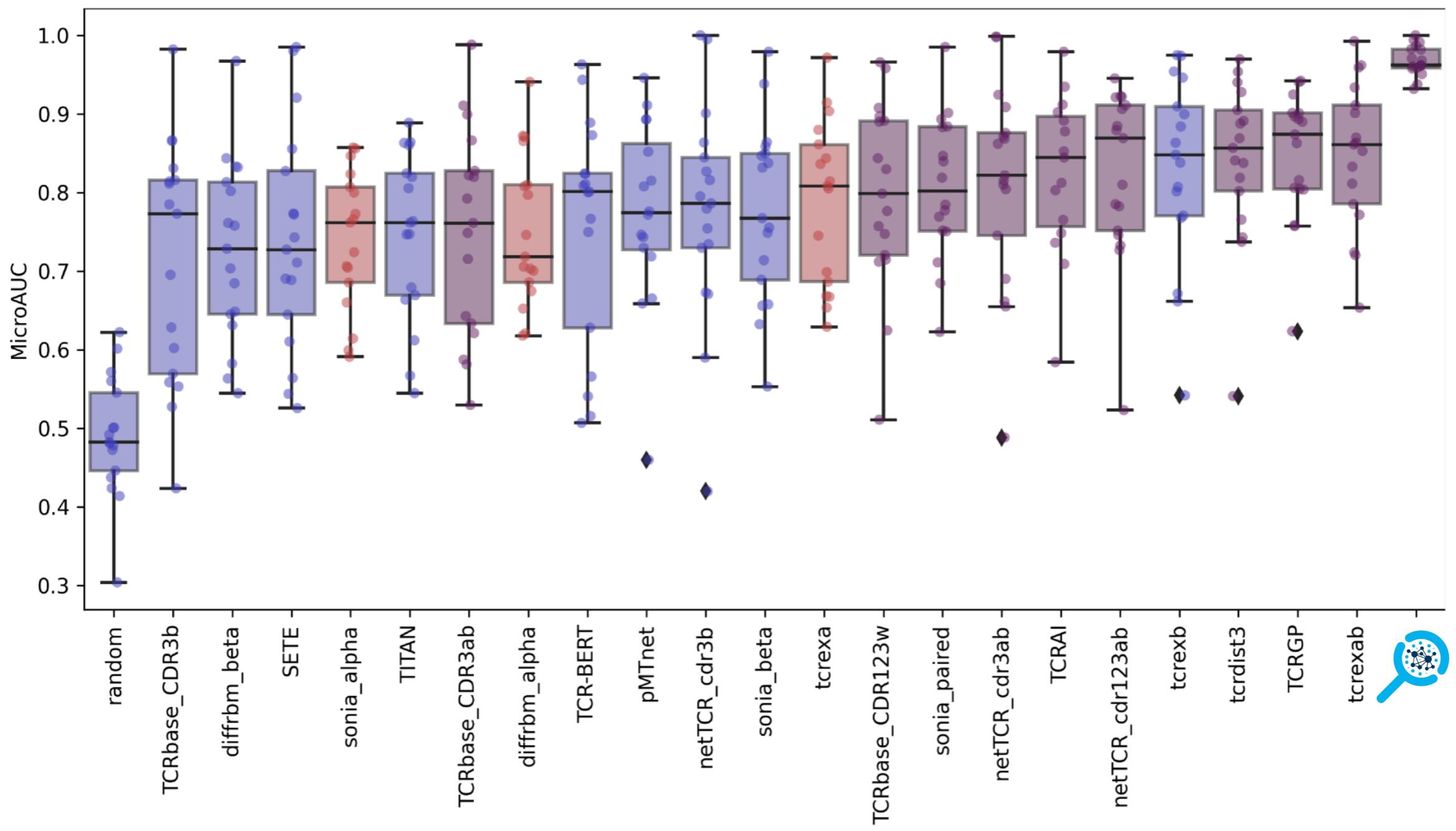
CDR3 region

TRBJ1-6*02

J region

What is the most likely epitope (if any)?

DETECT on benchmark



Zoom in on NLVPMVATV

- Positive samples

True label



NLVPMVATV	0.86	0.14
NLVPMVATV	0.205	0.00
NLVPMVATV	0.295	-0.08
NLVPMVATV	0.205	-0.19
NLVPMVATV	0.665	0.02
NLVPMVATV	0.76	0.11
NLVPMVATV	0.915	0.23
NLVPMVATV	0.155	-0.09
NLVPMVATV	0.125	-0.10
NLVPMVATV	0.84	0.12

Zoom in on NLVPMVATV

- Negative samples

True label



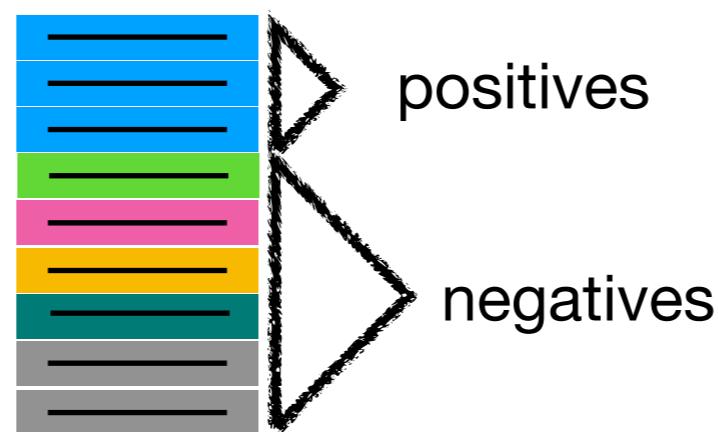
YLQPRTFLL	0.205	-0.28
GILGFVFTL	0.13	-0.38
LLWNGPMAV	0.205	-0.41
CINGVCWTV	0.13	-0.38
LLWNGPMAV	0.17	-0.23
GLCTLVAML	0.065	-0.41
HPVTKYIM	0.095	-0.31
GILGFVFTL	0.035	-0.24
TTDPSFLGRY	0.22	-0.34
GILGFVFTL	0.03	-0.12

Zoom in on NLVPMVATV

- Negative samples

True label			DETECT predicted label
YLQPRTFLL	0.205	-0.28	YLQPRTFLL
GILGFVFTL	0.13	-0.38	GILGFVFTL
LLWNGPMAV	0.205	-0.41	LLWNGPMAV
CINGVCWTV	0.13	-0.38	CINGVCWTV
LLWNGPMAV	0.17	-0.23	YLQPRTFLL
GLCTLVAML	0.065	-0.41	GLCTLVAML
HPVTKYIM	0.095	-0.31	HPVTKYIM
GILGFVFTL	0.035	-0.24	GILGFVFTL
TTDPSFLGRY	0.22	-0.34	TTDPSFLGRY
GILGFVFTL	0.03	-0.12	GILGFVFTL

Performance evaluation?



Not realistic

- The positives are even less frequent
- The diversity of negatives is even higher

Bold claim #3:

Many known TCR-epitope pairs are incorrect due to false positives.

ImmuneWatch database

Curated **367,072** T-cell receptors from the literature



Stringent filtering

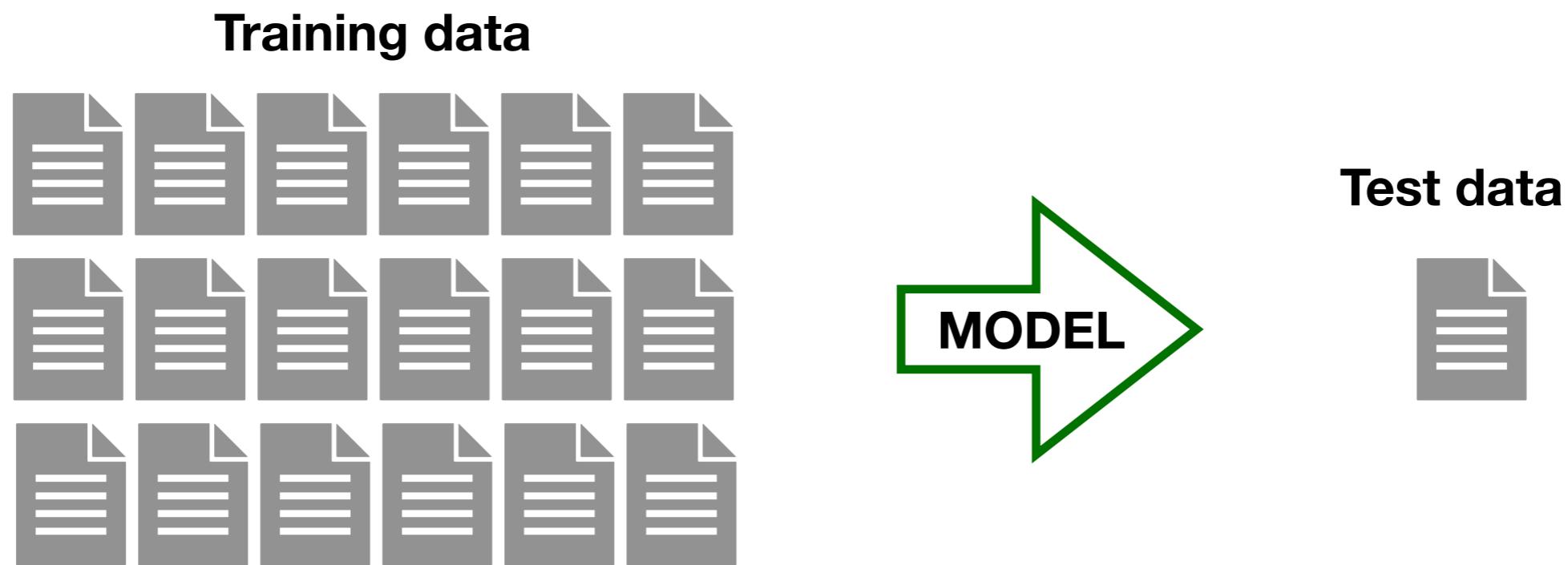
114,943 “maybe correct” TCRs covering **1742** epitopes

+

Proprietary data

Evaluation

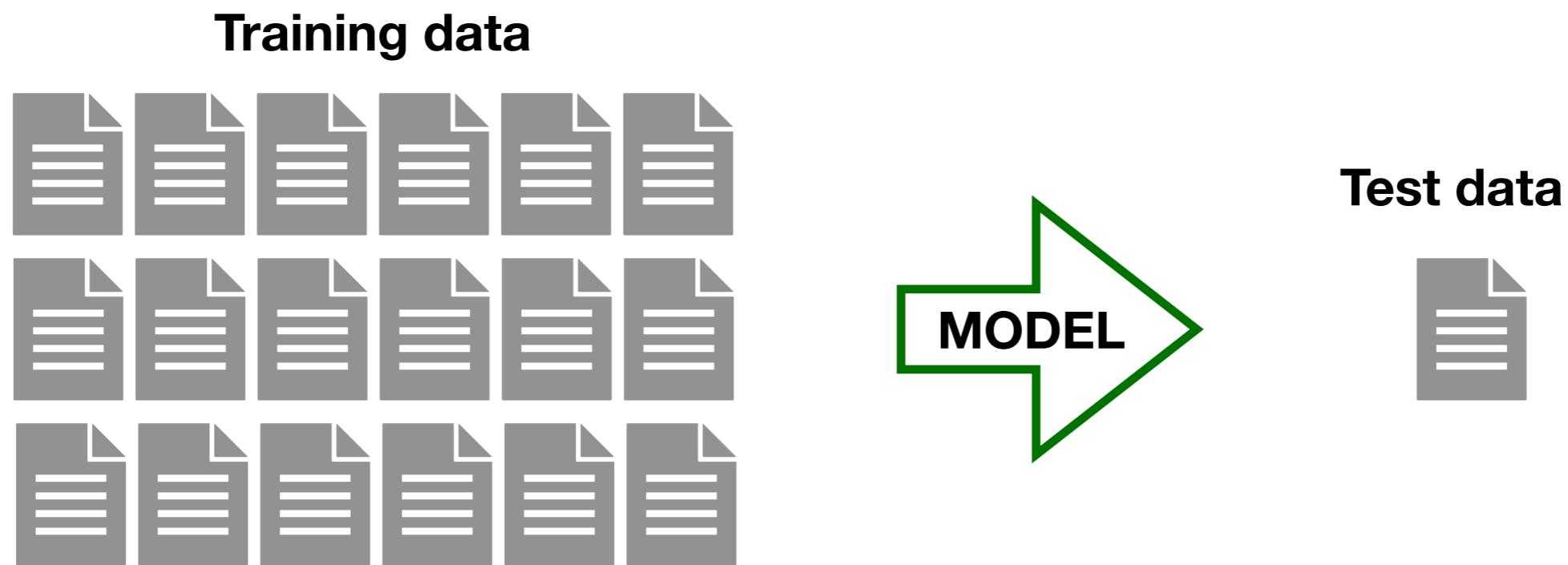
Leave-one-study-out cross validation



Try to annotate TCRs of a left-out study,
without specifying which epitopes are being tested

Evaluation

Leave-one-study-out cross validation



Global AUROC = 0.92

Evaluation

Leave-one-study-out cross validation

back to NLVPMVATV

**TCR in dataset
and ImmRep benchmark
for NLVPMVATV**

TRAV1-2*01 CAVLDSNYQLIW TRAJ33*01
TRBV6-1*01 CASSEDSGGYNEQFF TRBJ2-1*01

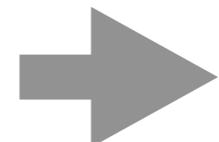
Evaluation

Leave-one-study-out cross validation

back to NLVPMVATV

**TCR in dataset
and ImmRep benchmark
for NLVPMVATV**

TRAV1-2*01 CAVLDSNYQLIW TRAJ33*01
TRBV6-1*01 CASSEDSGGYNEQFF TRBJ2-1*01



MR1
5-(2-oxopropylideneamino)-6-d-ribitylaminouracil

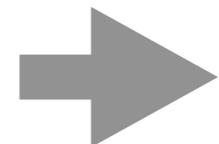
Evaluation

Leave-one-study-out cross validation

back to NLVPMVATV

**TCR in dataset
and ImmRep benchmark
for NLVPMVATV**

(invariant TCR for MAIT cells)
TRAV1-2*01 CAVLDSNYQLIW TRAJ33*01
TRBV6-1*01 CASSEDSGGYNEQFF TRBJ2-1*01



MR1
5-(2-oxopropylideneamino)-6-d-ribitylaminouracil

Performance evaluation?

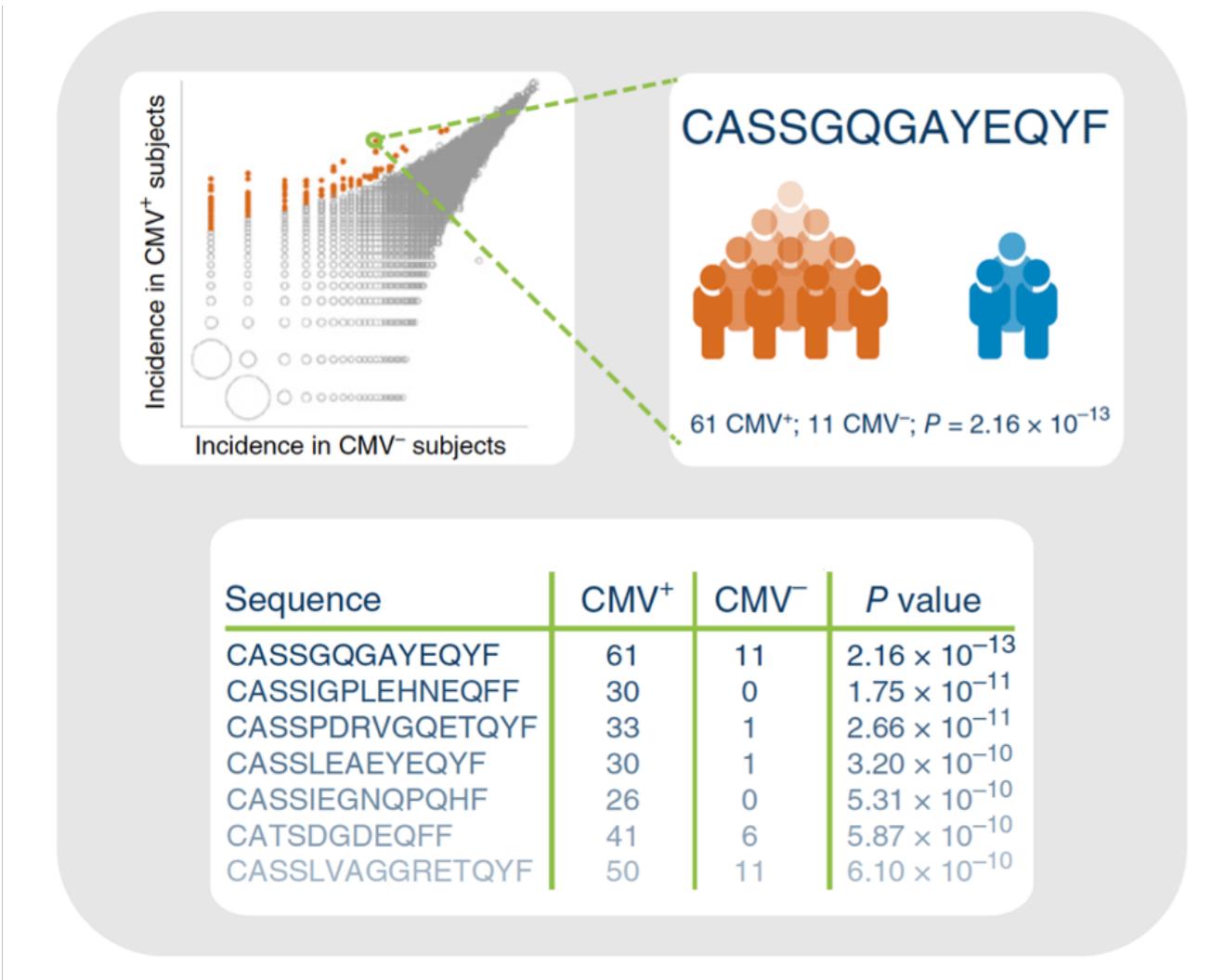
Not realistic

- Still only limited amount of epitopes

What do we want to use these models for?

- Likely not the identification of multimer-sorted cells

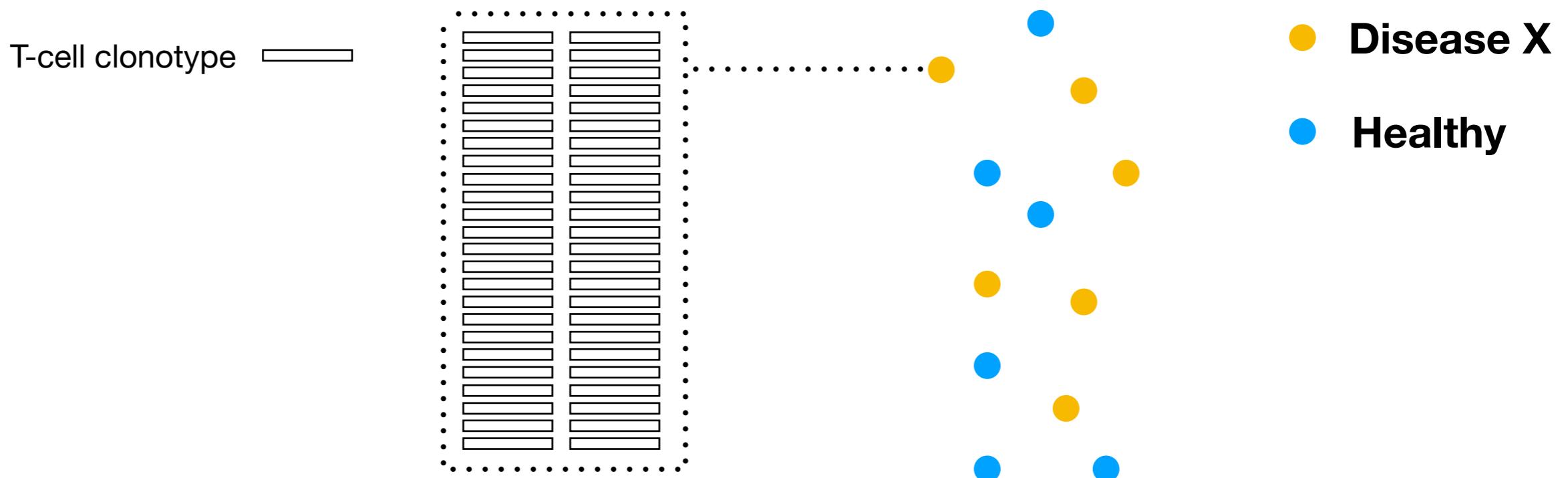
Diagnostics



- Identify enriched public clonotypes for diagnostics.
- Requires large cohorts in a top-down approach.

Bottom-up?

- Classify samples based on TCR sequencing

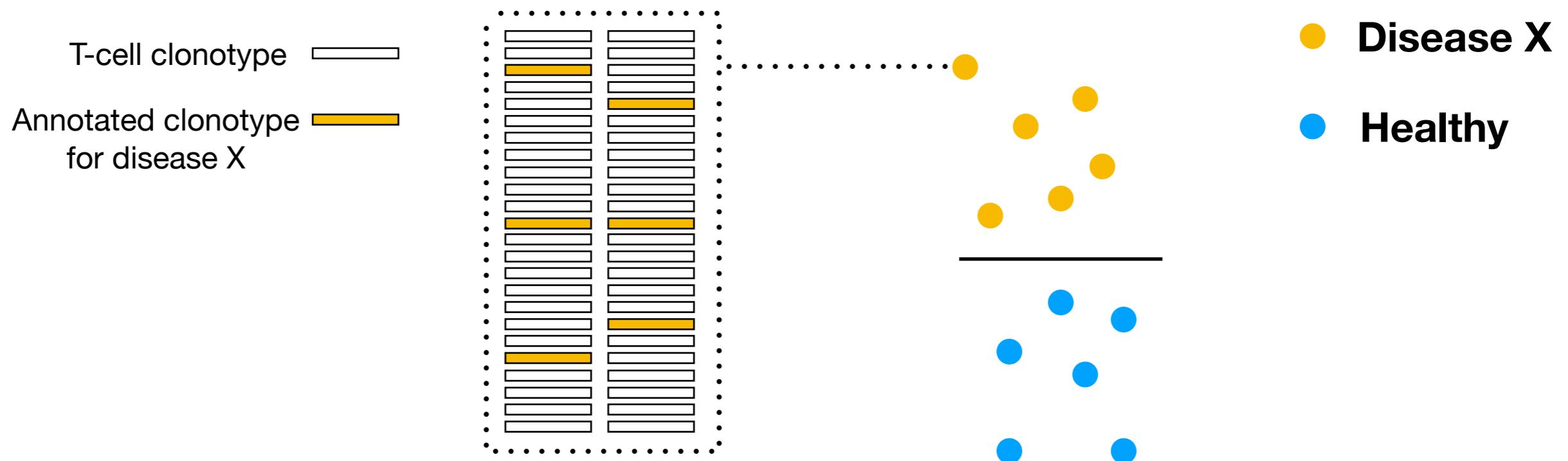


Bold claim #4:

Performant epitope annotation models should be able to easily stratify full repertoires based on disease history.

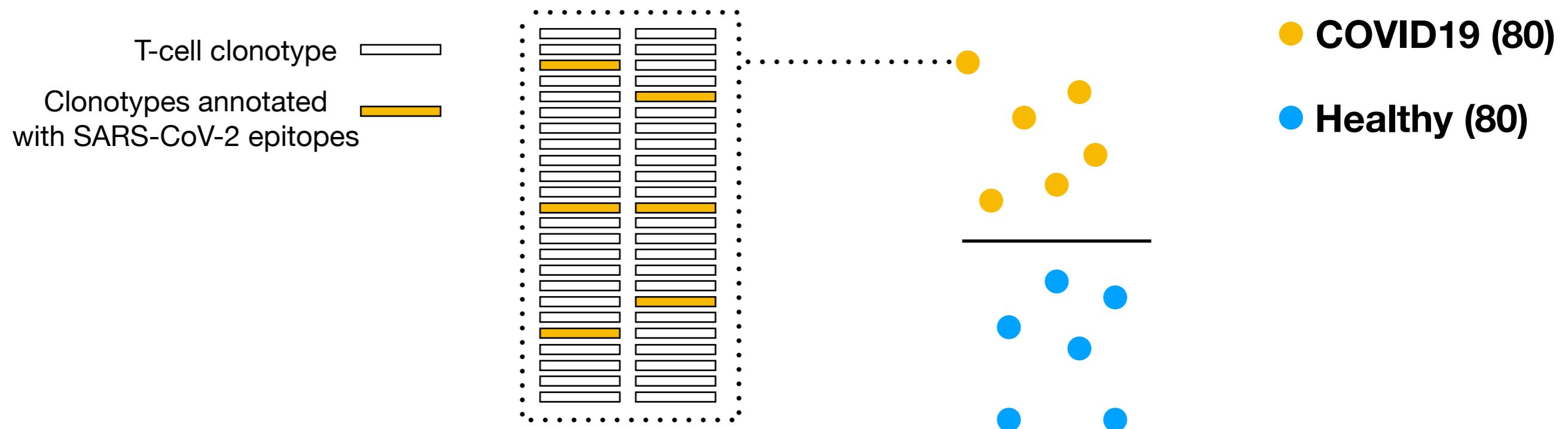
Bottom-up?

- Repertoires from individuals with disease X should have more X-reactive T-cells.

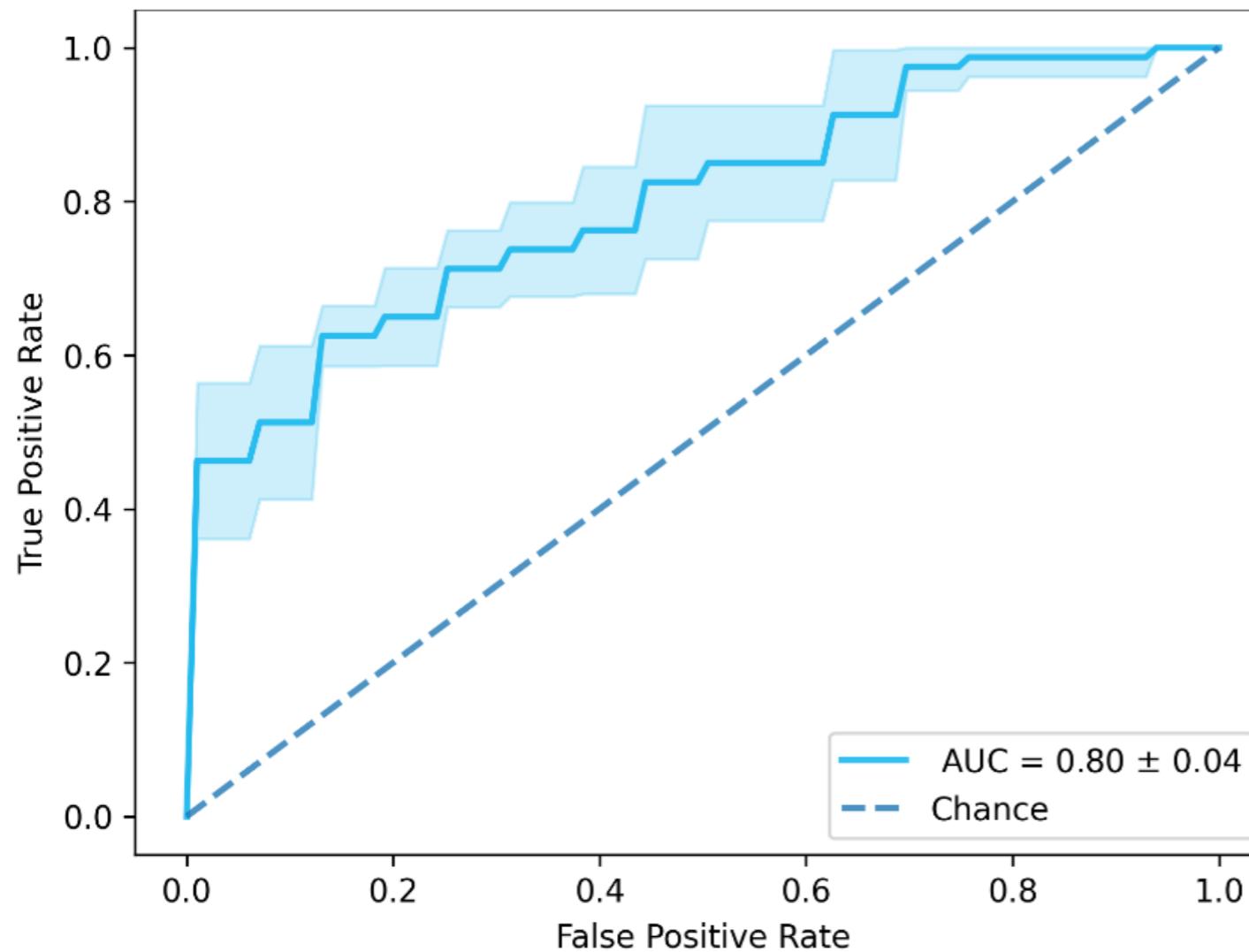


COVID diagnostic

- Repertoires from SARS-CoV-2+ individuals

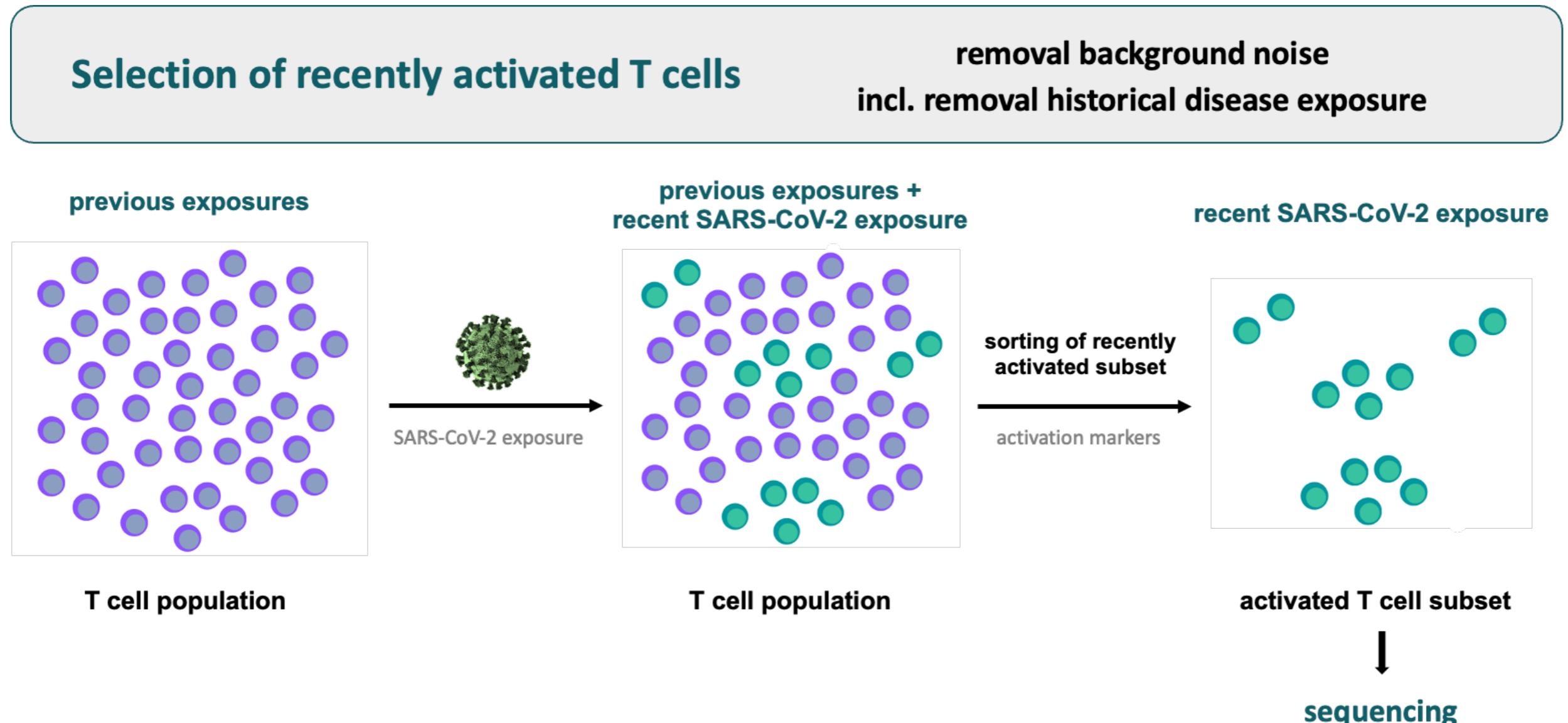


COVID diagnostic



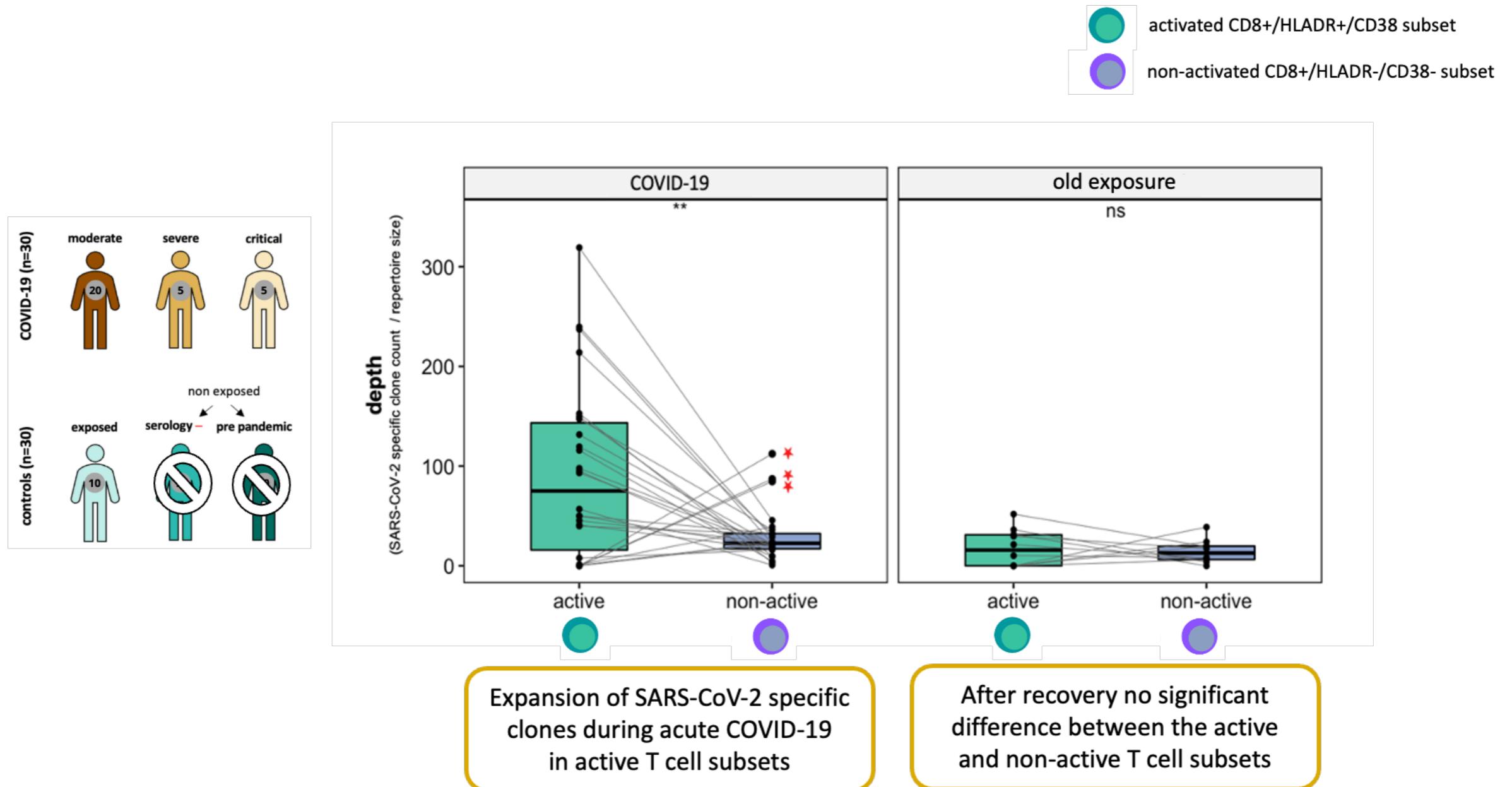
- Only based on 773 epitopes across 10 MHC variants (half mapping to HLA-A02:01)
- Not accounting for cross-reactivity
- Using full T-cell receptor repertoires

Diagnosis with sorting



Vujkovic et al. The diagnostic potential of activated CD8+ T-cells during acute COVID-19. (In press)

Diagnosis with sorting



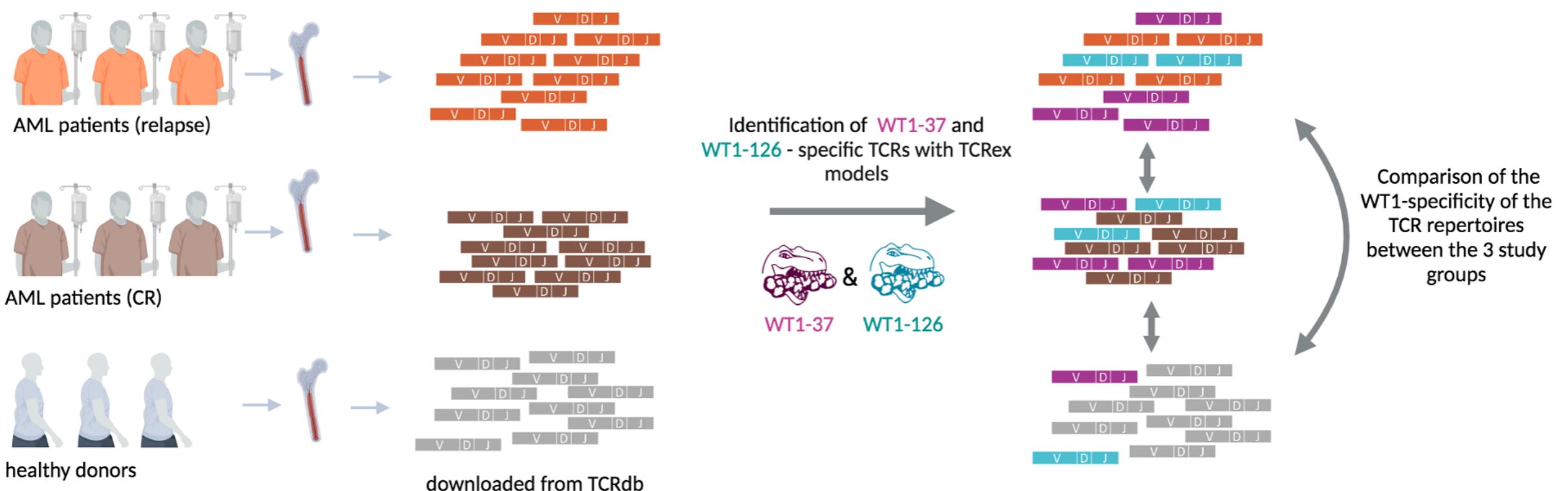
Vujkovic et al. The diagnostic potential of activated CD8+ T-cells during acute COVID-19. (In press)

Bold claim #5:

Epitope-TCR annotations coupled with TCR sequencing has the potential for a powerful diagnostic.

Immunotherapy response

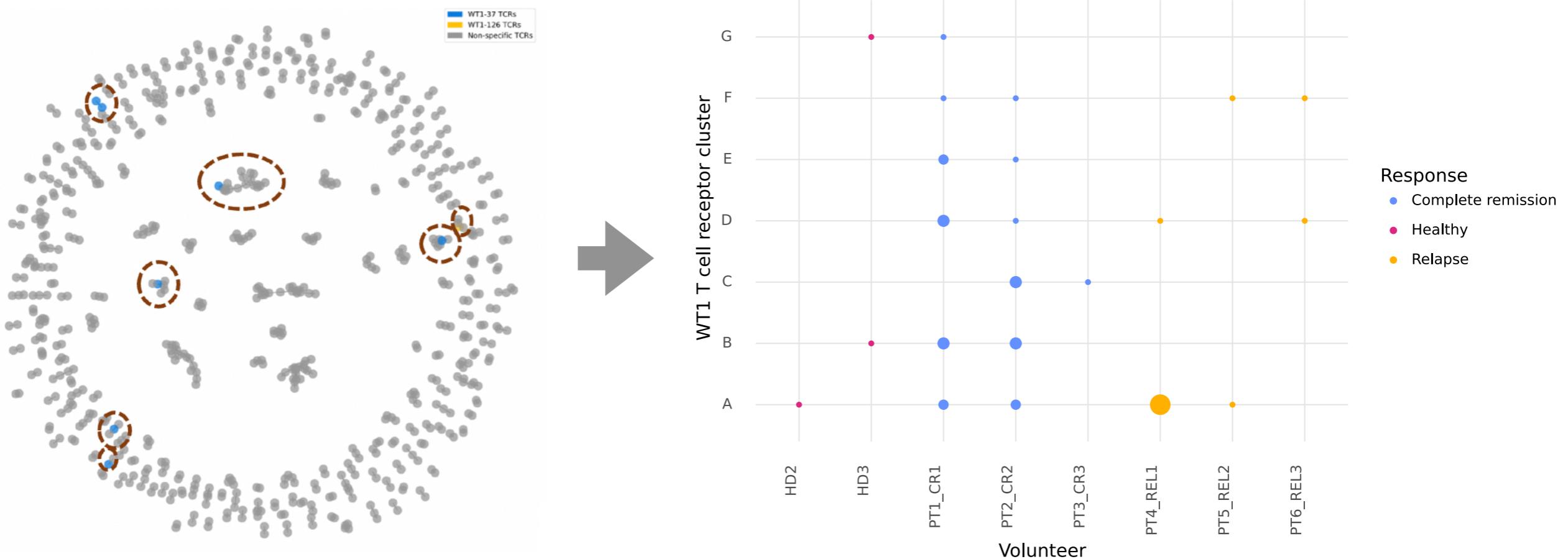
- Discovery of cancer antigens (WT1) in relapsing AML patients



Gielis, S. et al. (2023). Analysis of Wilms' tumor protein 1 specific TCR repertoire in AML patients uncovers higher diversity in patients in remission than in relapsed (Under review)

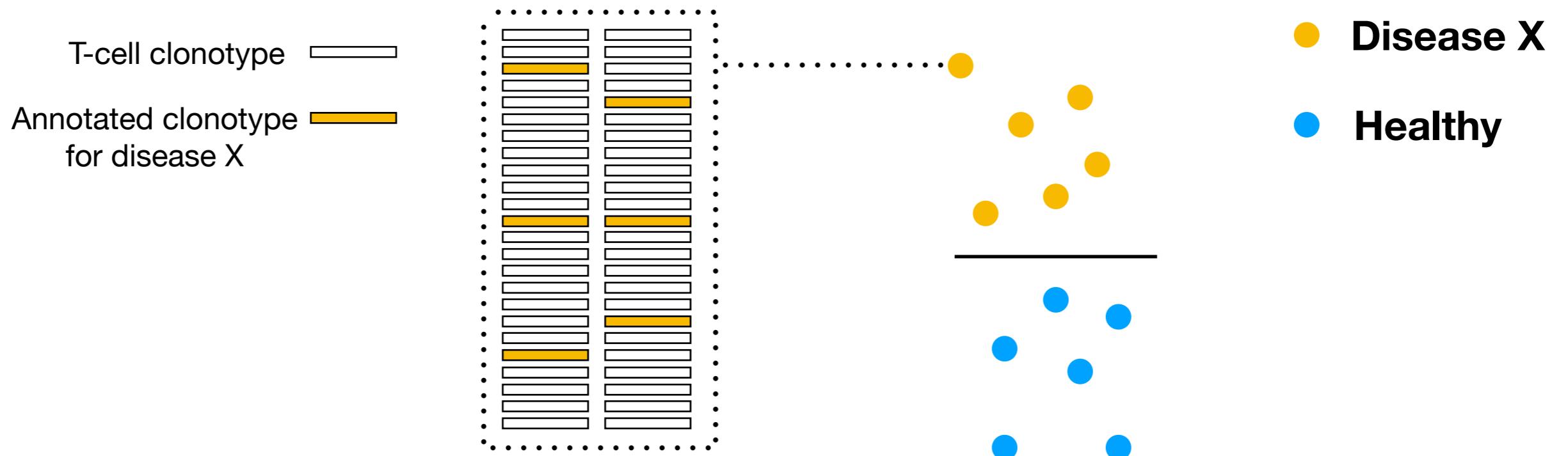
Immunotherapy response

- Patients in remission have broader WT1-specific repertoires



Gielis, S. et al. (2023). Analysis of Wilms' tumor protein 1 specific TCR repertoire in AML patients uncovers higher diversity in patients in remission than in relapsed (Under review)

Epitope-based diagnostics

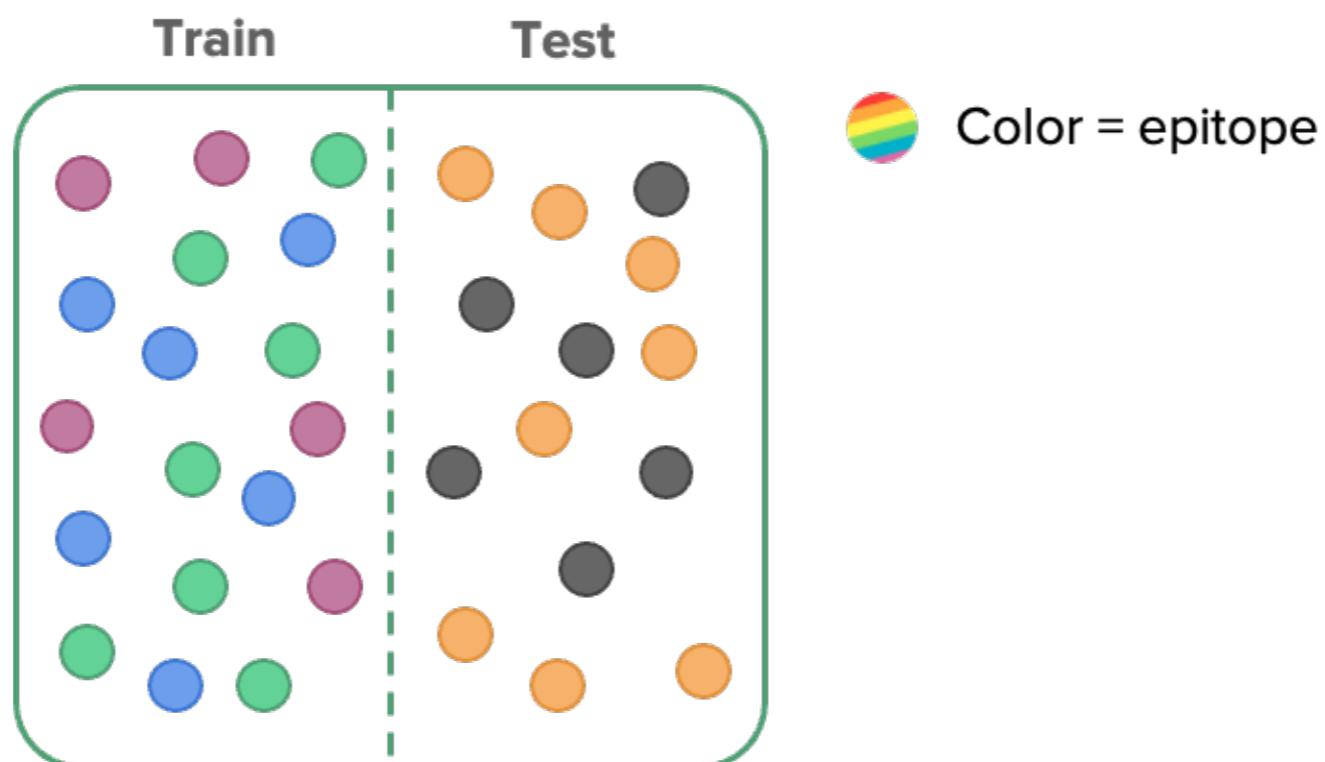


- Diagnostics possible with epitope-TCR predictions, but limited by the epitopes in the database.

Unseen epitopes

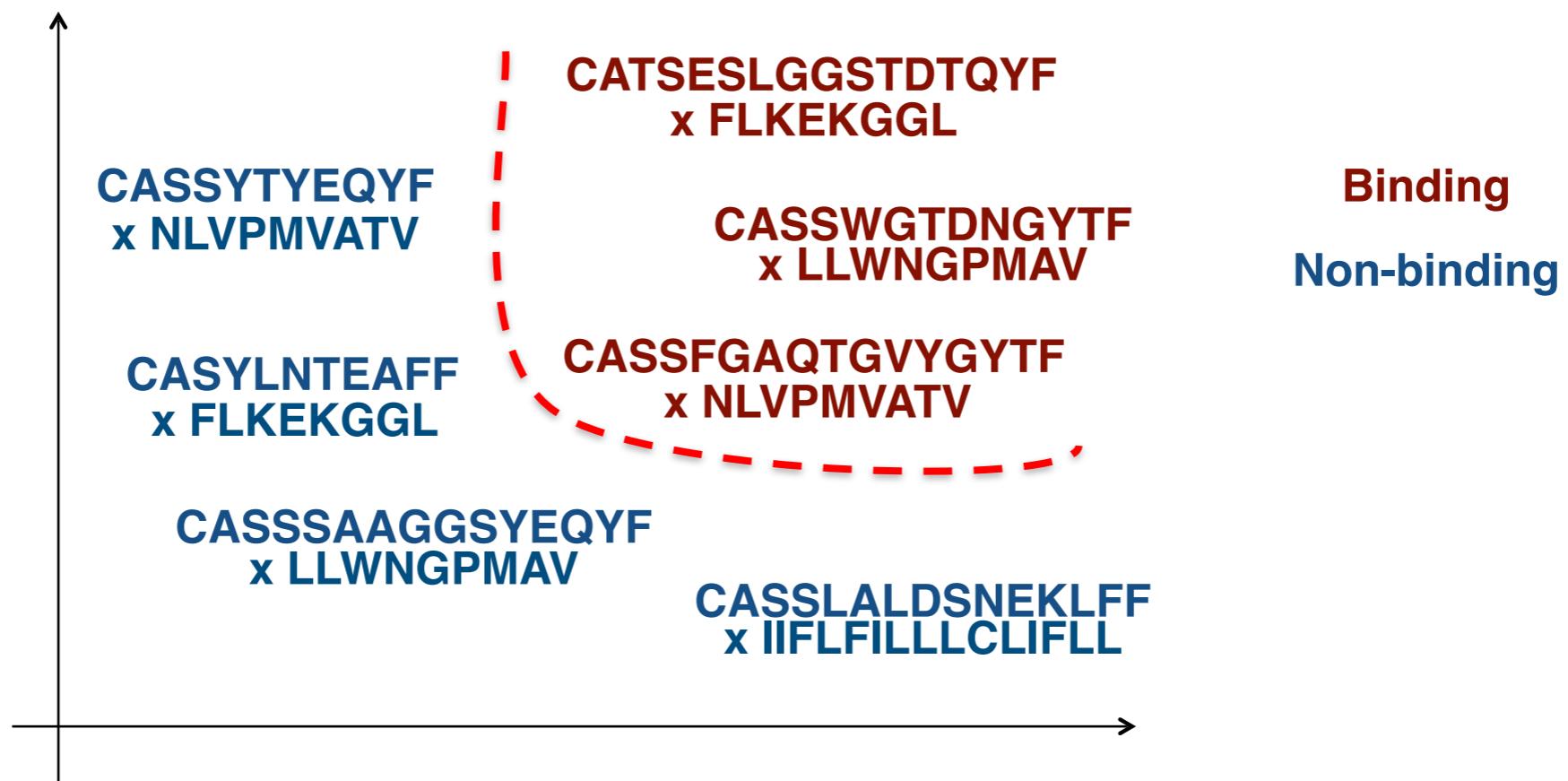
Can we make prioritizations for candidate (unseen) epitopes based on the TCR?

- Only 2000 epitopes with known T-cell receptors, thus most of interest will have no data.
- Use deep learning to learn TCR-epitope relationships, but be smart about it.

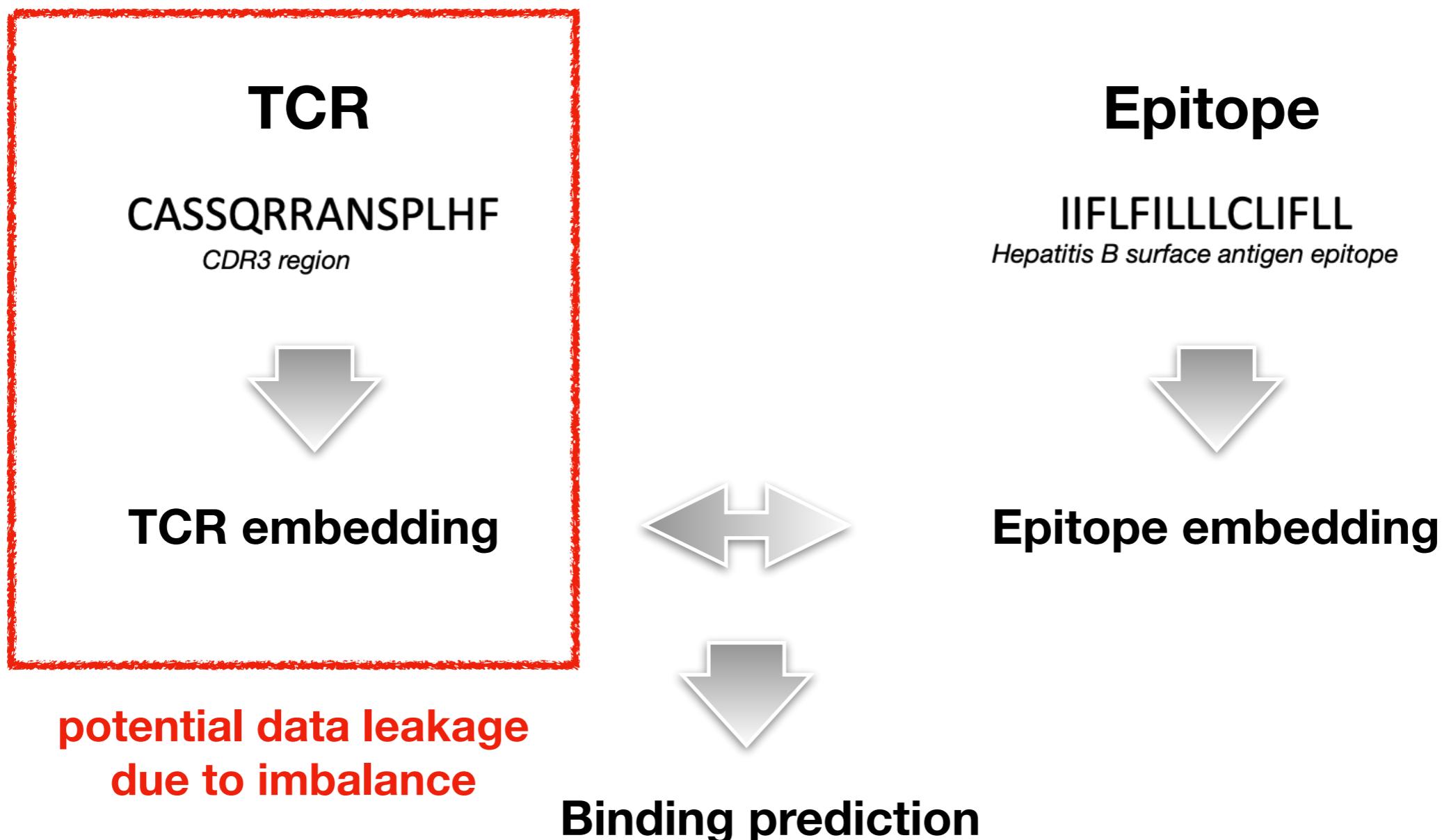


Unseen epitope x TCR

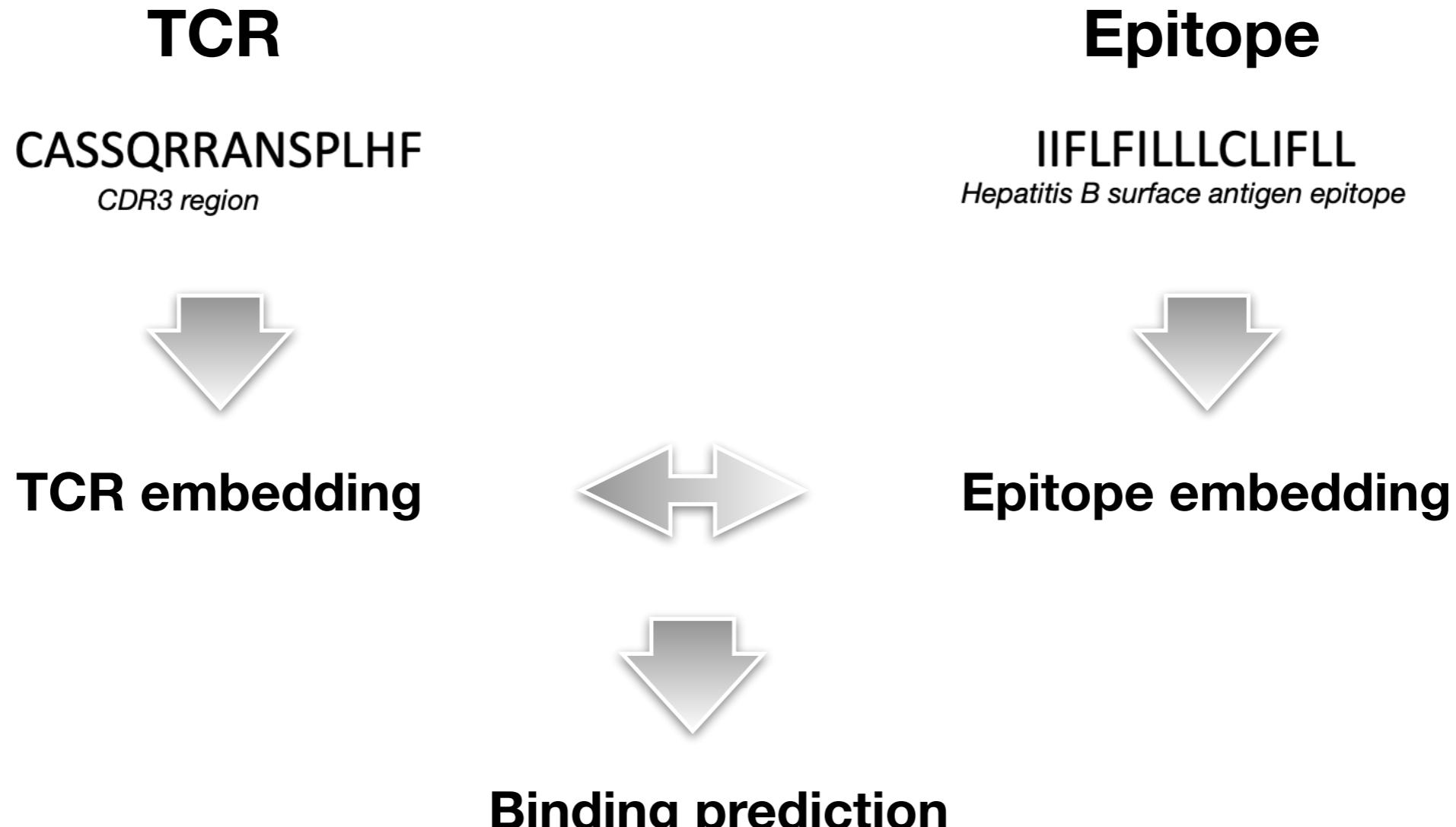
Predict for a pair of one epitope and one TCR if they will bind, by projecting them to a numerical space.



Typical DL approach

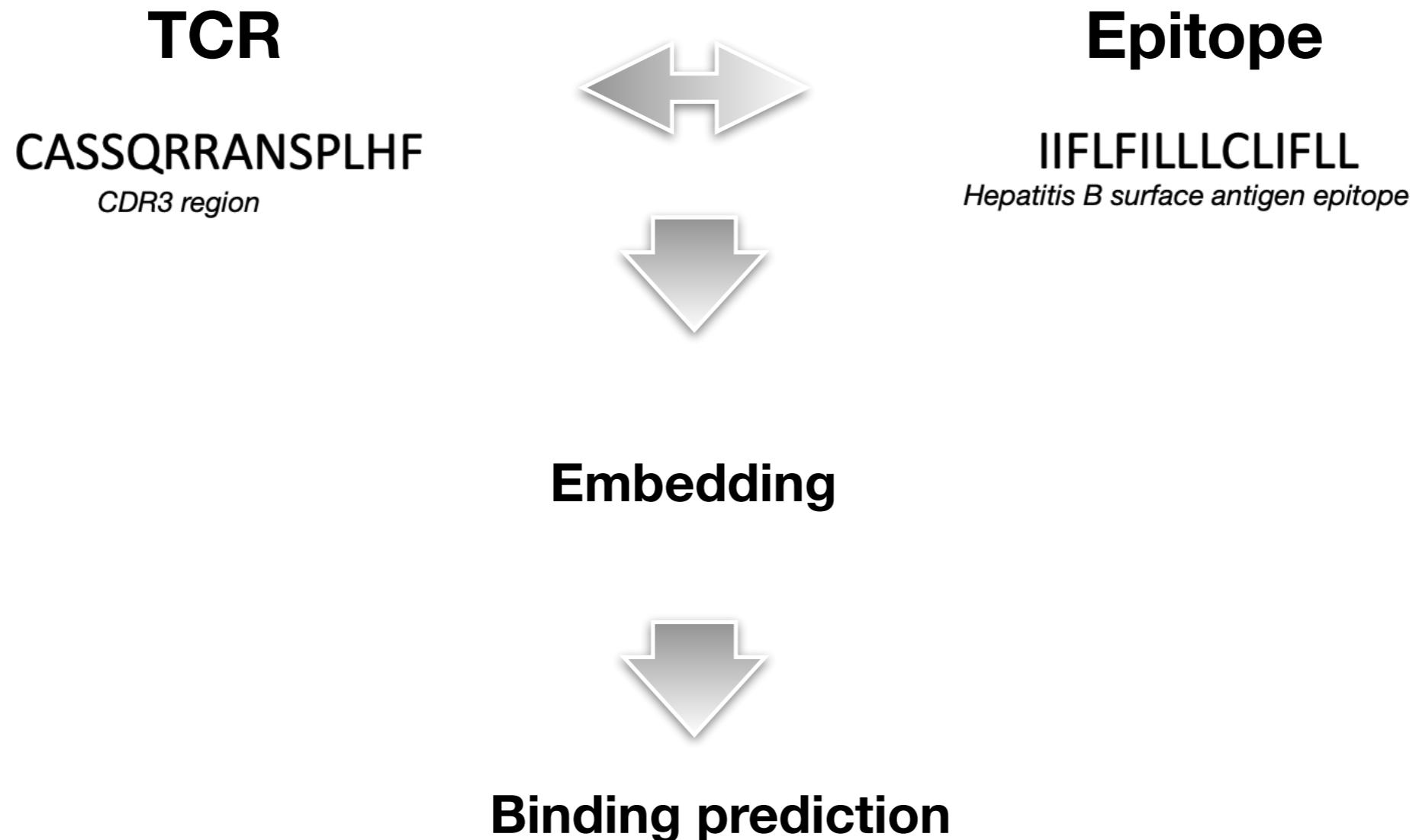


Avoiding bias?



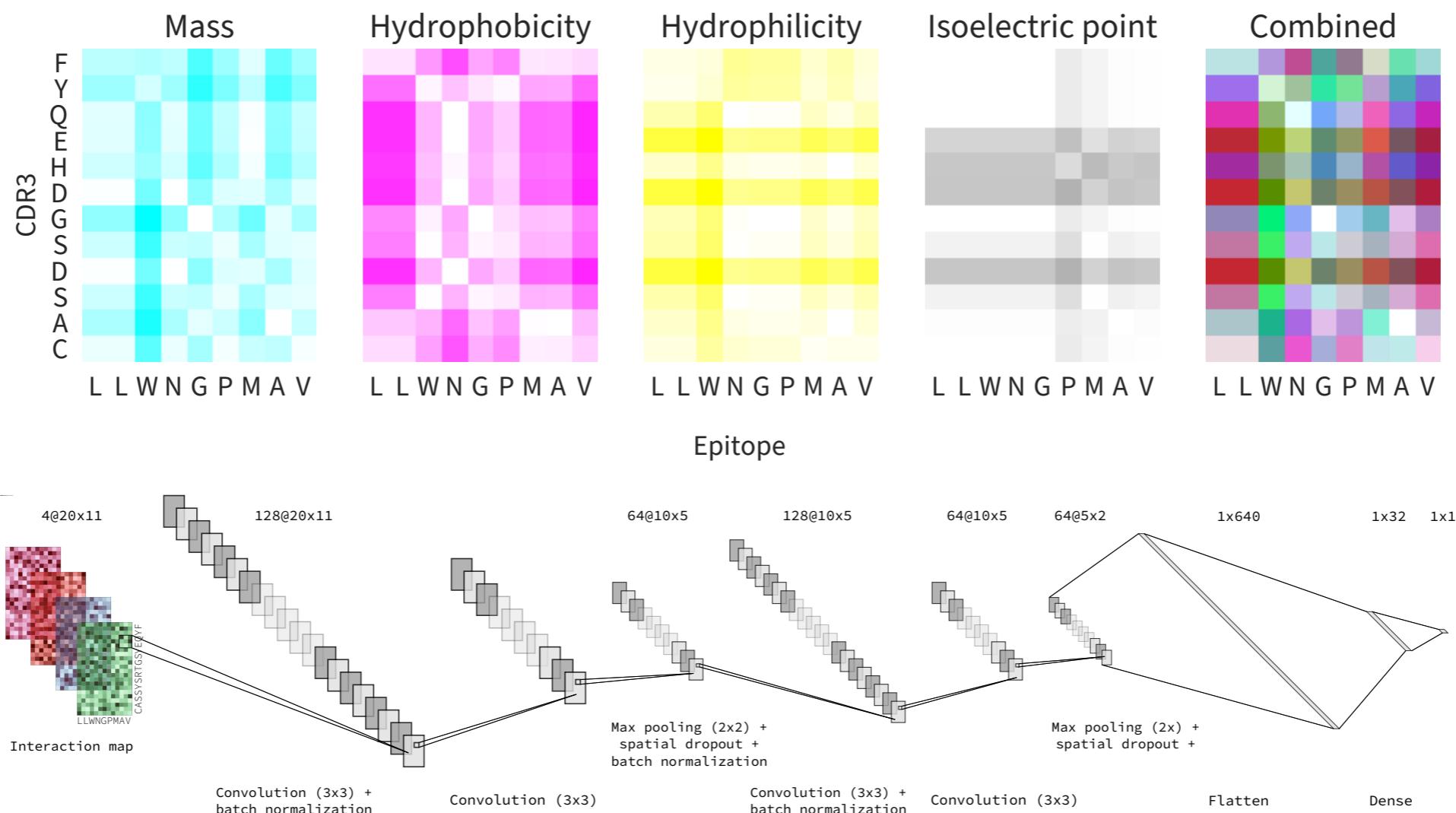
Can we change the architecture reduce focusing on one side?

Avoiding bias?



ImRex

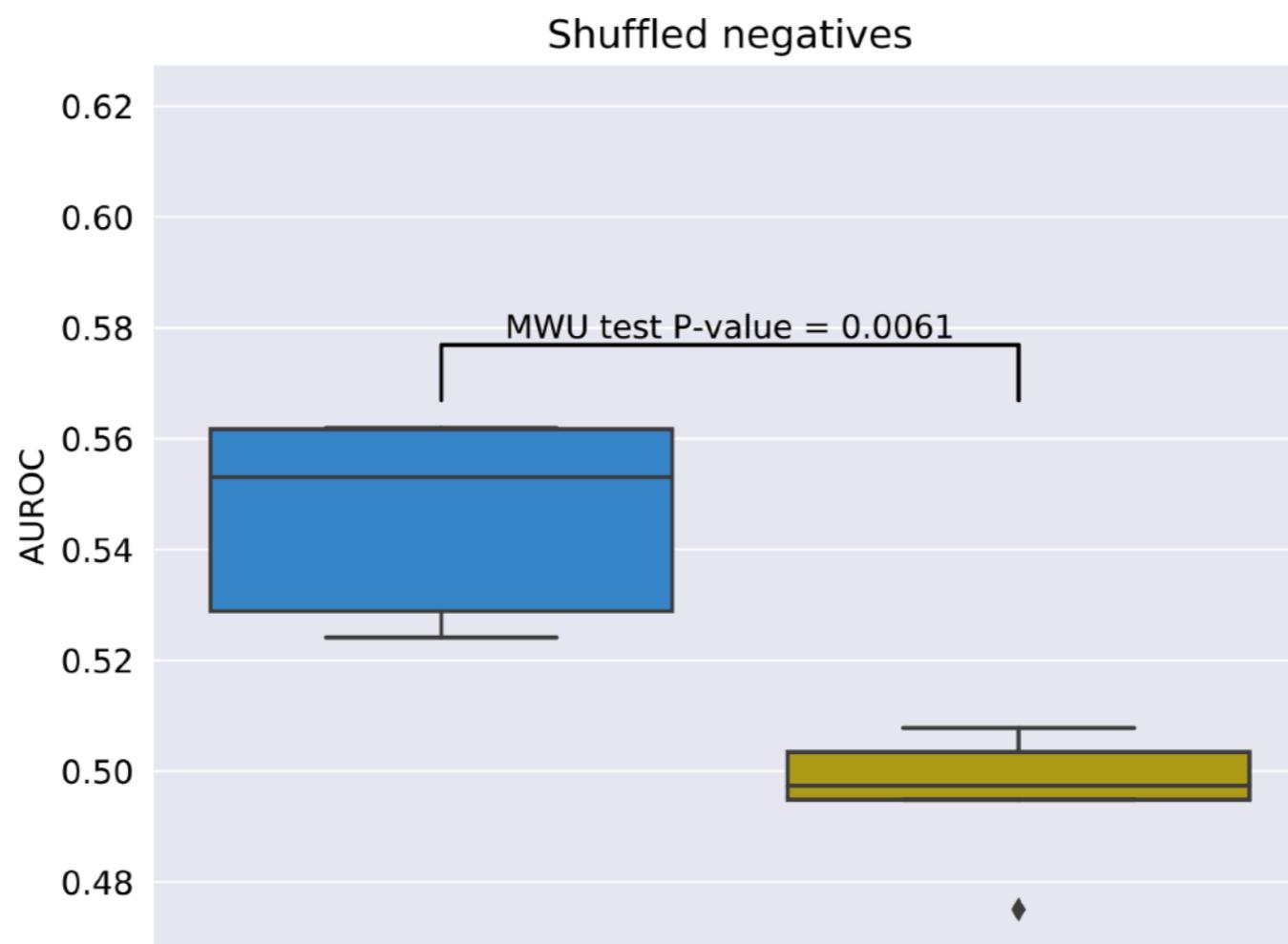
- We converted the TCR-epitope interaction space into an image.



Moris P et al. (2020). Current challenges for unseen-epitope TCR interaction prediction and a new perspective derived from image classification. *Briefings in Bioinformatics*



Shuffled negatives



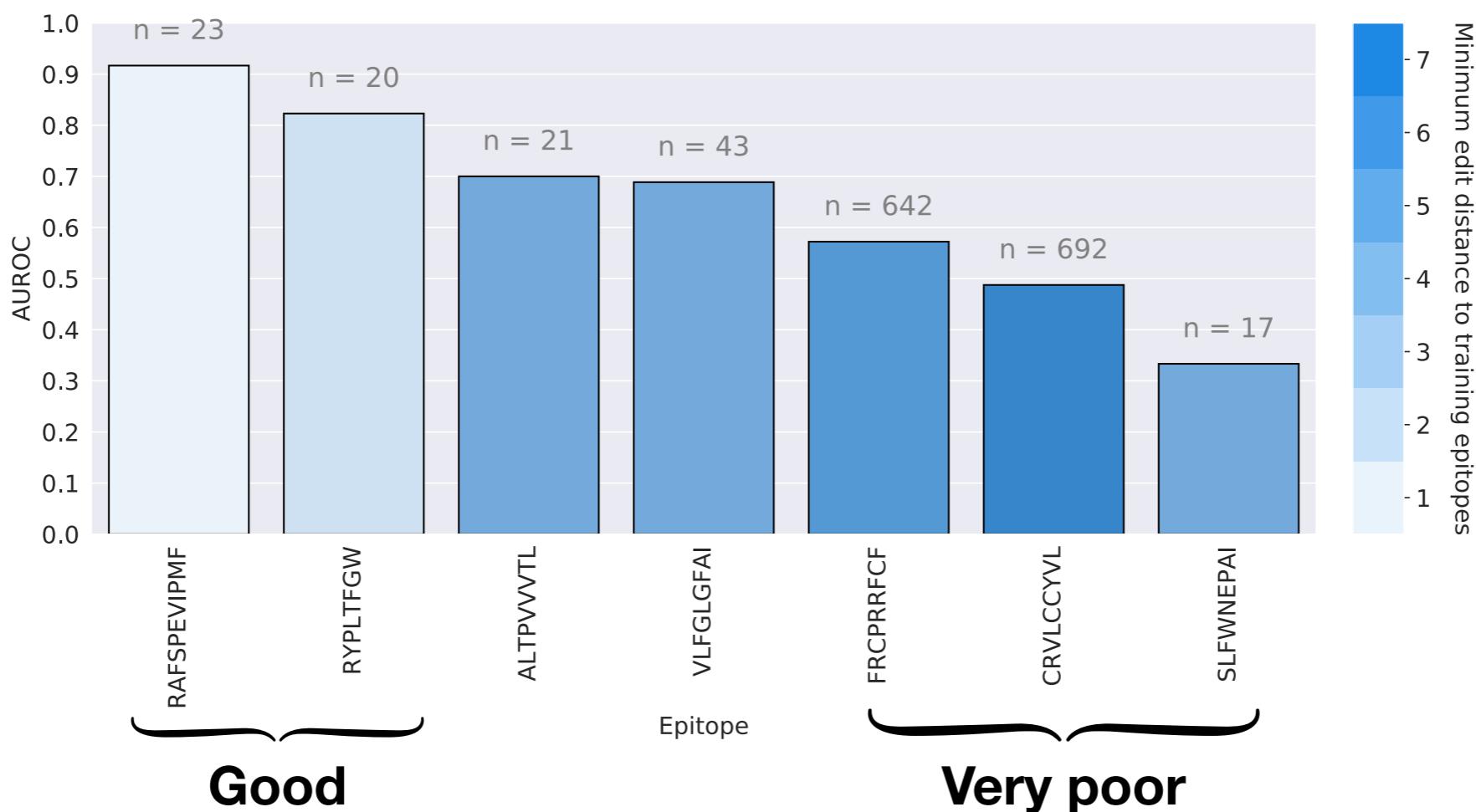
ImRex - TRB - epitope-grouped - shuffled negatives
($\overline{AUROC} = 0.55 \pm 0.02 s$)

ImRex - TRB - epitope-grouped - shuffled negatives - decoy
($\overline{AUROC} = 0.50 \pm 0.01 s$)



ImRex

- Performance barely better than random
- Except when the epitope is close enough to the training examples



Bold claim #6:

We are limited by data, not models, for unseen epitope prediction.

Bold claim #7:

We need about 500 000 TCRs covering at least 5000 epitopes to get good unseen epitope performance.

Bold claim #7:

We need about 500 000 TCRs covering at least 5000 epitopes to get good unseen epitope performance.

This is feasible.

Acknowledgements



- Kris Laukens
- Benson Ogunjimi
- Koen Vercauteren
- Esther Bartholomeus
- My Ha
- Kerry Mullan
- Fabio Affaticati
- Ceder Dens
- Nicky De Vrij
- Sofie Gielis
- Anna Postovskaya
- Sebastiaan Valkiers
- Vincent Van Deuren
- Romi Vandoren
- Alexandra Vujkovic



- Sander Wuyts
- Tom Bosschaerts
- Max Van Houcke



ImmRep benchmark

- Justin Barton
- Barbara Bravi
- Liel Cohen-Lavi
- Vadim Karnaukhov
- Elias Lilleskov
- Alessandro Montemurro
- Morten Nielsen
- Thierry Mora
- Paul Pereira
- María Rodríguez Martínez
- Jorge Fernandez-de-Cossio-Diaz
- Aleksandra M Walczak
- Anna Weber
- Rose Yin
- Anne Eugster
- Virag Sharma

