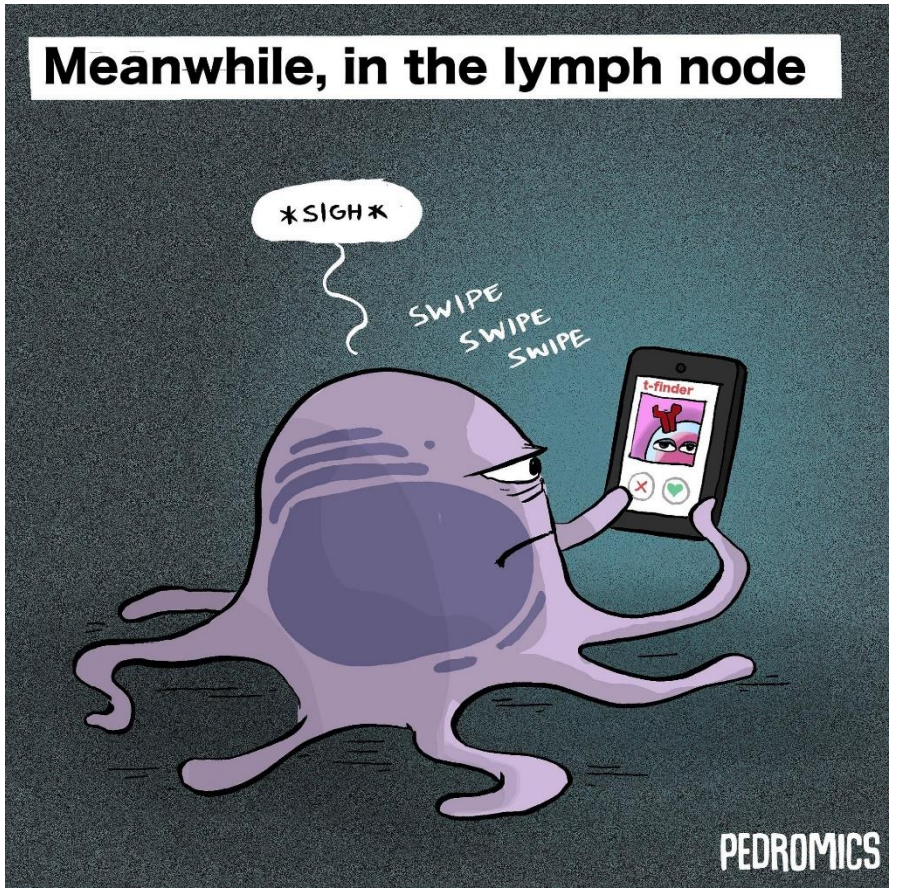
Andreas Tiffeau-Mayer
Division of Infection and Immunity
Institute for the Physics of Livings Systems

Kavli Institute for Theoretical Physics
October 11th, 2024

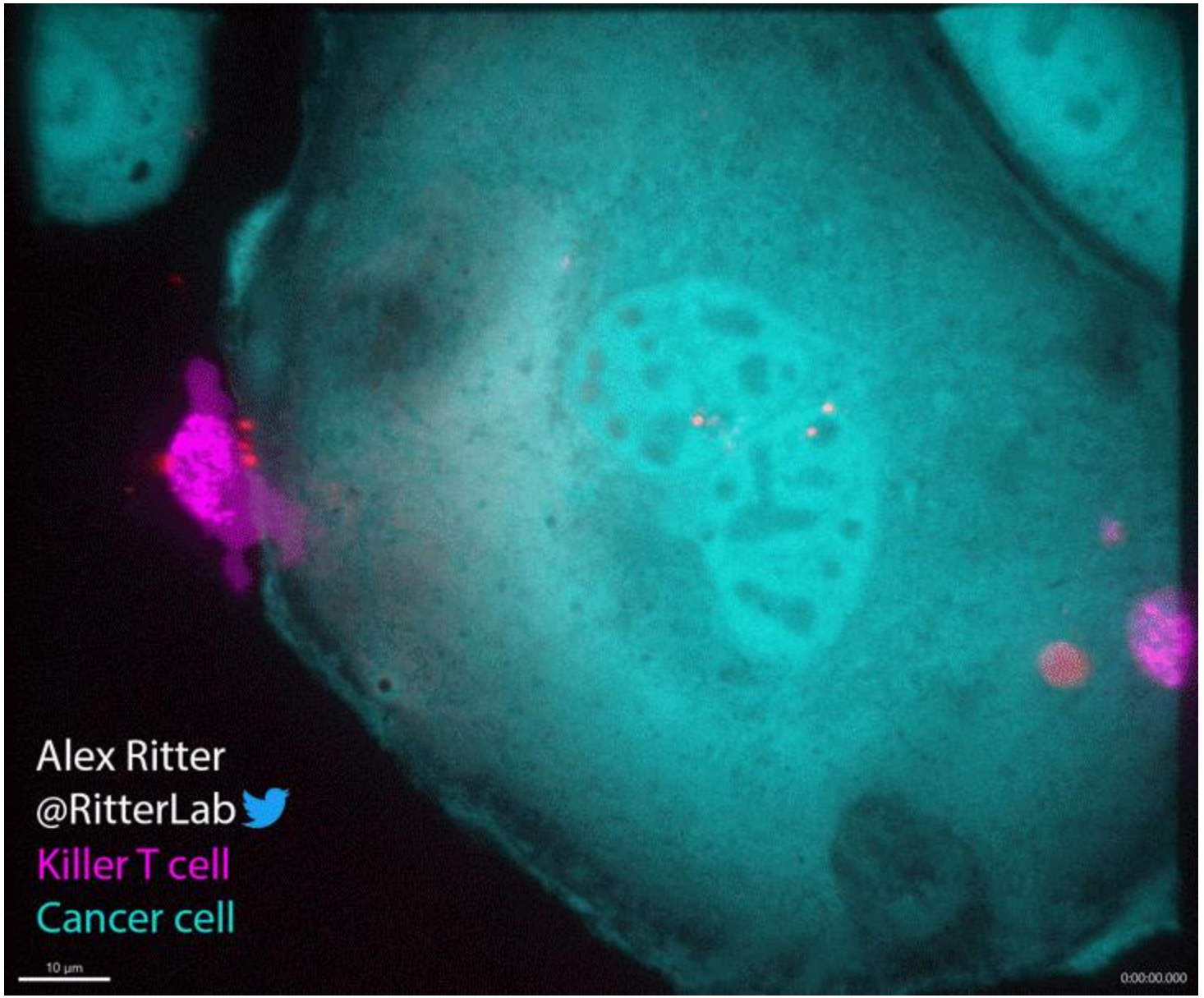
Statistical insights into the immune receptor code



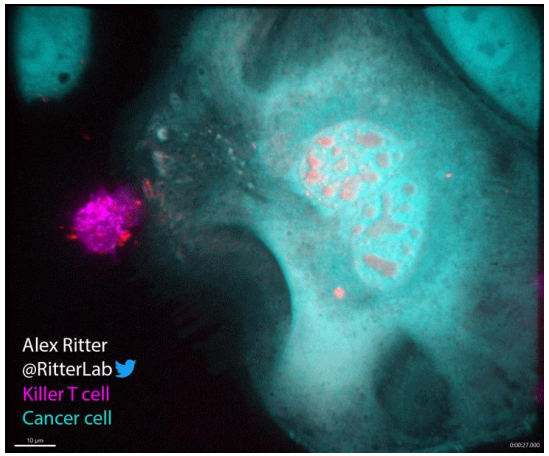
LDN QIMMUNO

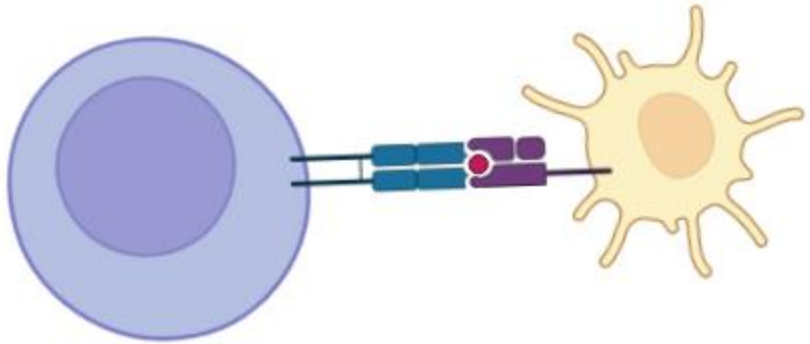
@andimscience

www.qimmuno.com


Alex Ritter
@RitterLab
Killer T cell
Cancer cell
10 µm
0:00:00.000

How do lymphocytes know what to attack?

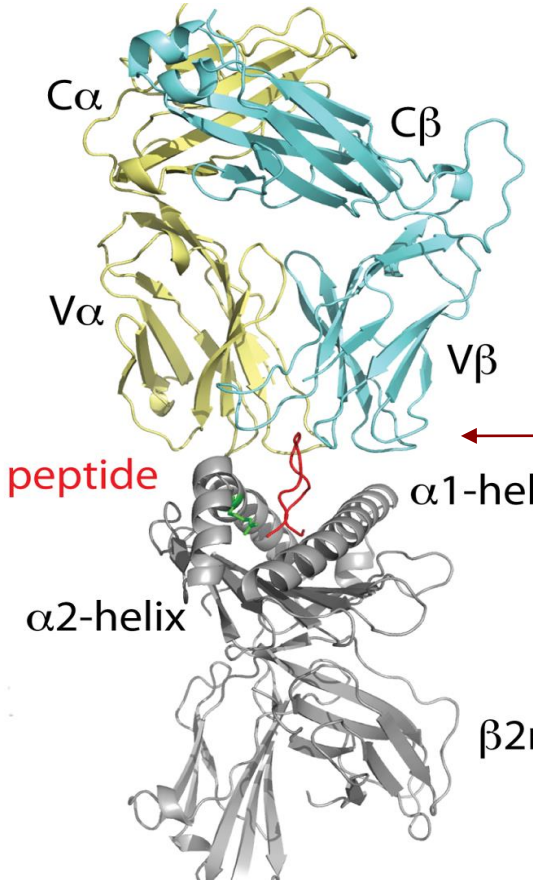




Heroes of today's talk: T cell receptors

T cell receptor (TCR)

Peptide presented on target cell



hyper-variable

Outline

1. Why reading the immune receptor code is hard and why it matters
   Sureshchandra, Henderson, …, ATM*, Saligrama*, Wagar* bioRxiv 2024
2. The state of *in silico* TCR-pMHC binding prediction
   Nielsen … ATM … Barton Immunoinformatics 2024
   Nagano, Pyo, …, Chain*, ATM* arXiv 2024
3. Statistical physics of sequence-function maps
   ATM, Callan PNAS 2023
4. Insights into the biophysical determinants TCR specificity
   Henderson, Nagano, Milighetti, ATM PNAS 2024
   Nagano, Pyo, …, Chain*, ATM* arXiv 2024

Which question would you ask if we could predict antigen-specificity from antibody/TCR sequence?

- **Basic human immunology**
  - ✓ Dynamics of immune response and memory at the level of single clones?
  - ✓ How stochastic or biased are cell fate decisions taken by antigen-specific populations *in vivo*?
  - ✓ …?
- **Vaccine design and protein engineering**
  - ✓ What clones and functional states are associated with protection from disease?
  - ✓ Can we use the epitopes they target to elicit such responses using vaccines?
  - ✓ Rational protein engineering of TCR T-cells / antibodies ?
  - ✓ …?

Heroes of today's talk (cont.)

Yuta Nagano

Martina Milighetti

Andrew Pyo

James Henderson

UCL

Benny Chain
John Shawe-Taylor

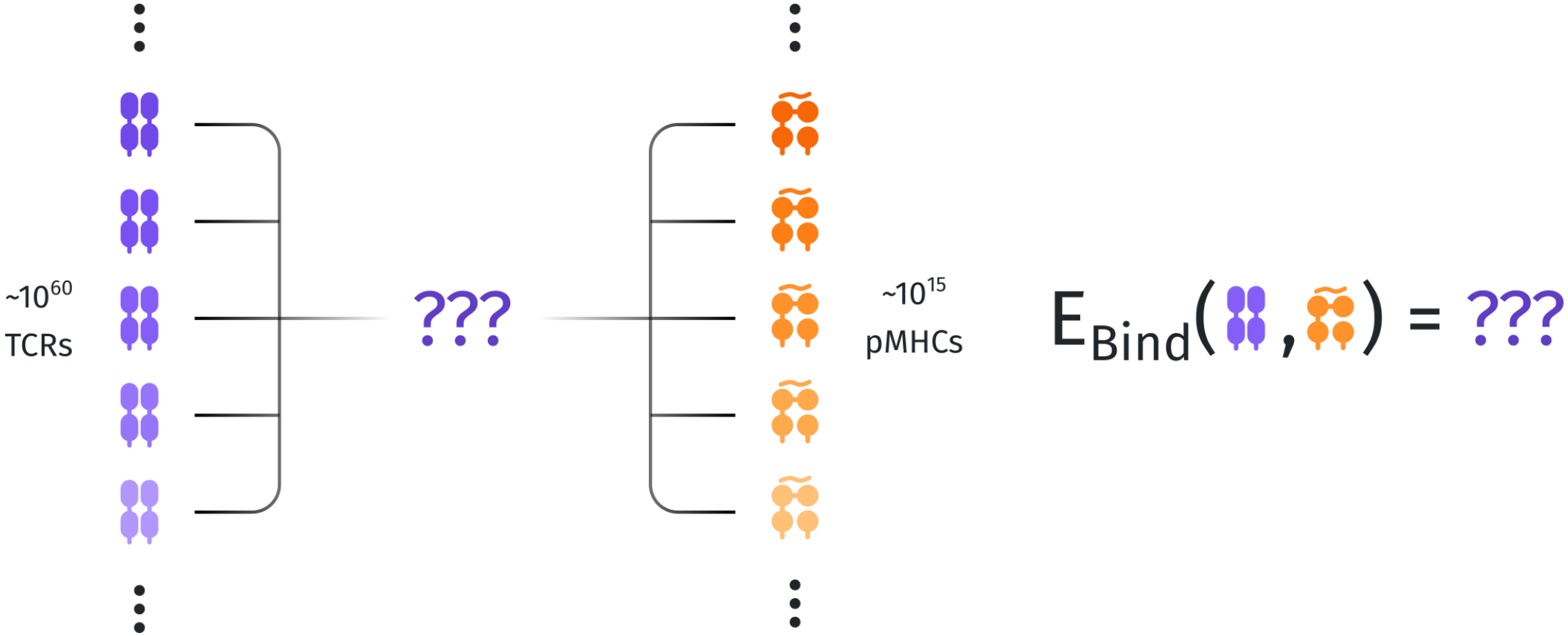
Sankalan Bhattacharyya
Rishika Saxena
Touchchai Chotisorayuth
Ursule Demaël

Curtis Callan (Princeton)
Ned Wingreen (Princeton)
Lisa Wagar (UC Irvine)
Naresha Saligrama (WUSTL)

Why reading the immune receptor code is hard and why it matters

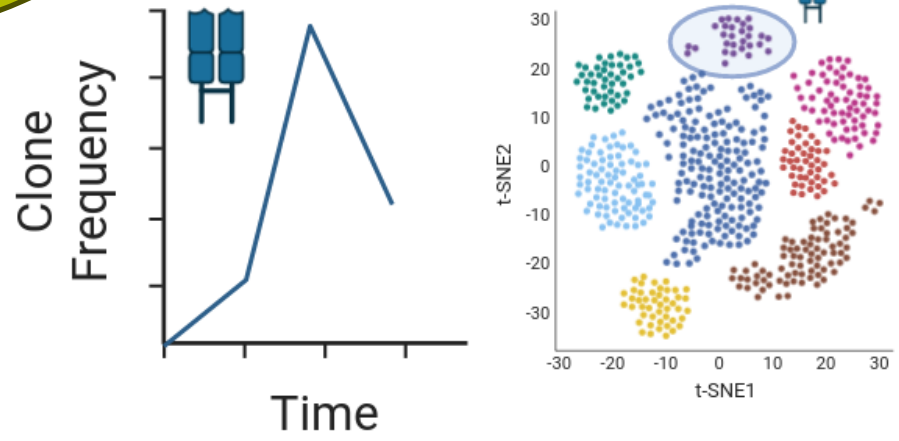
Sureshchandra, Henderson, …, ATM*, Saligrama*, Wagar*, bioRxiv 2024

Receptors and ligands are immensely diverse



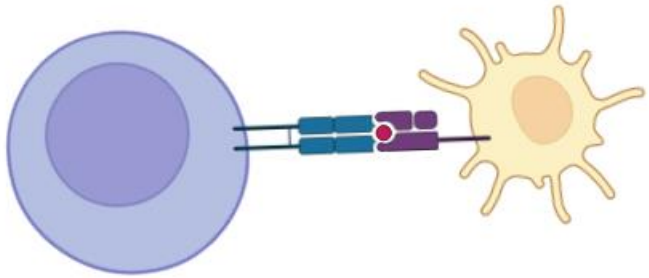
What does this mean for TCRseq?

TCRs make great barcodes



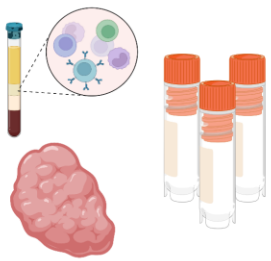
Lineage Tracing

Annotating antigens is hard

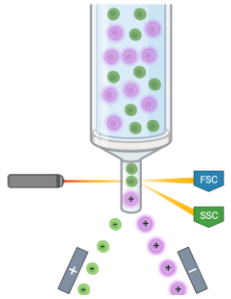


Antigen Deconvolution

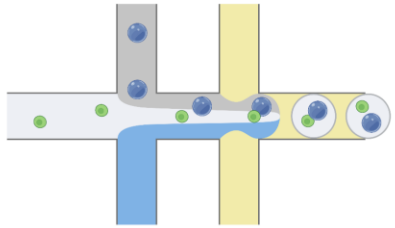
DeepTCR study: Profiling the human blood/tonsil repertoire of 5.6 million single T cells



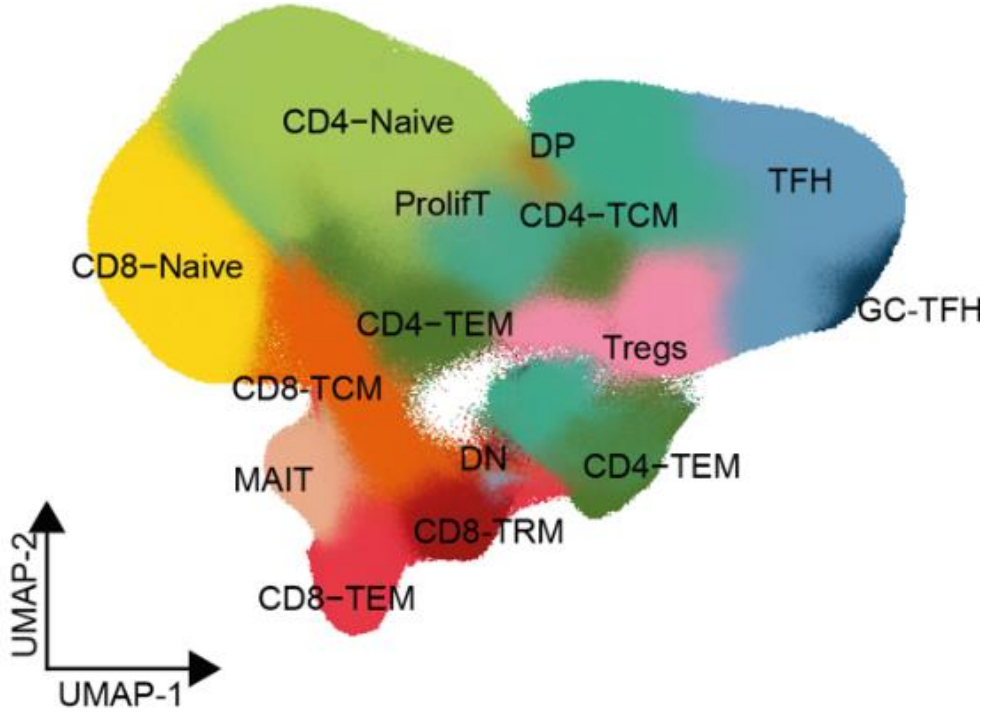
Cryopreserved tonsil/PBMC from 10 donors (Age 3-30 yo)



Sort αβ T cells

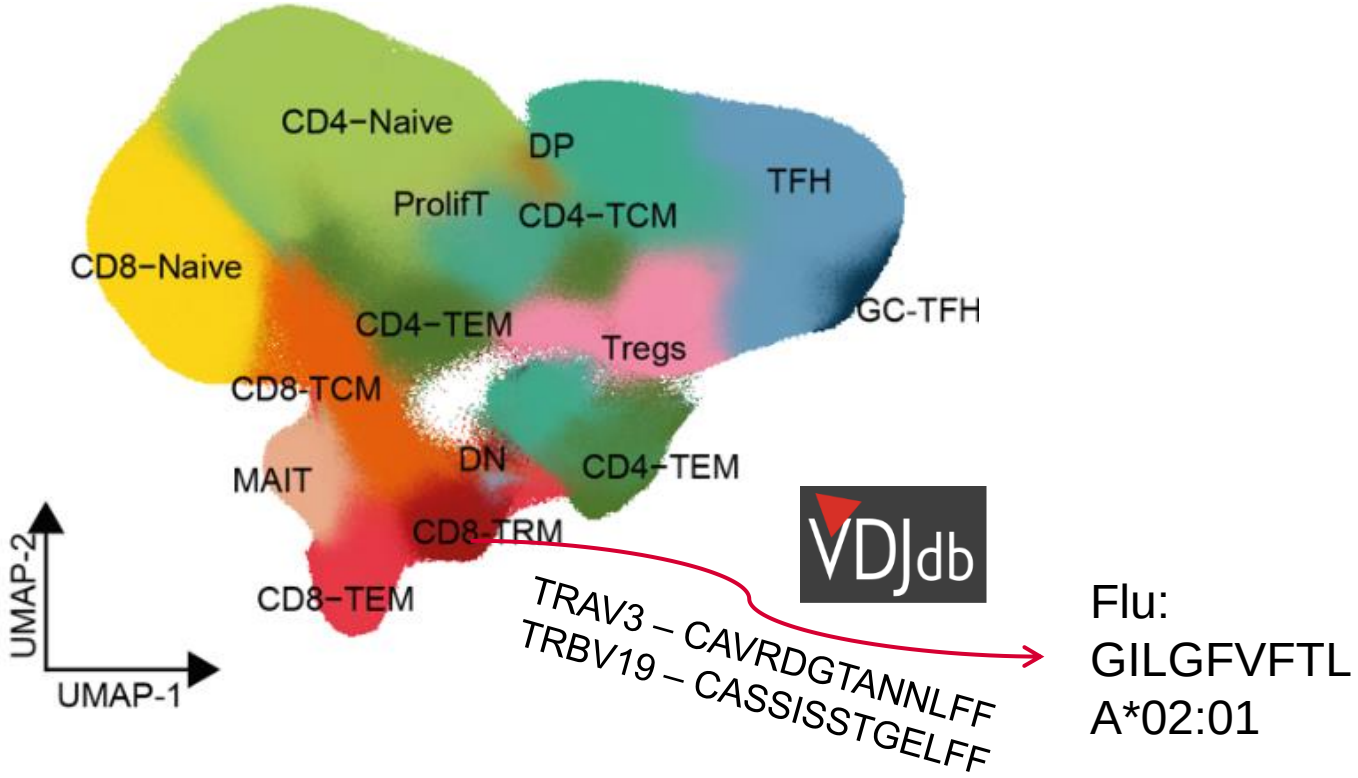


Droplet based scRNA/TCR-Seq



Sureshchandra, Henderson, …, ATM*, Saligrama*, Wagar*, bioRxiv 2024

Linking T cell function to antigen specificity



Only 24 TCRs could be mapped this way

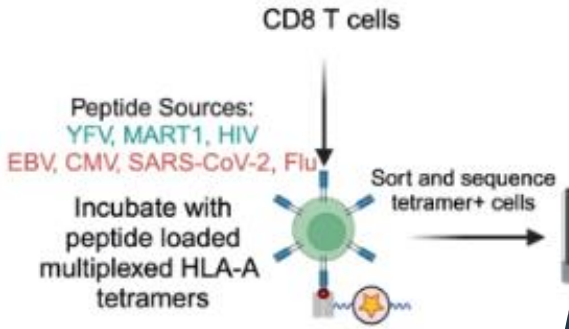
>99.99% of TCRs are orphan

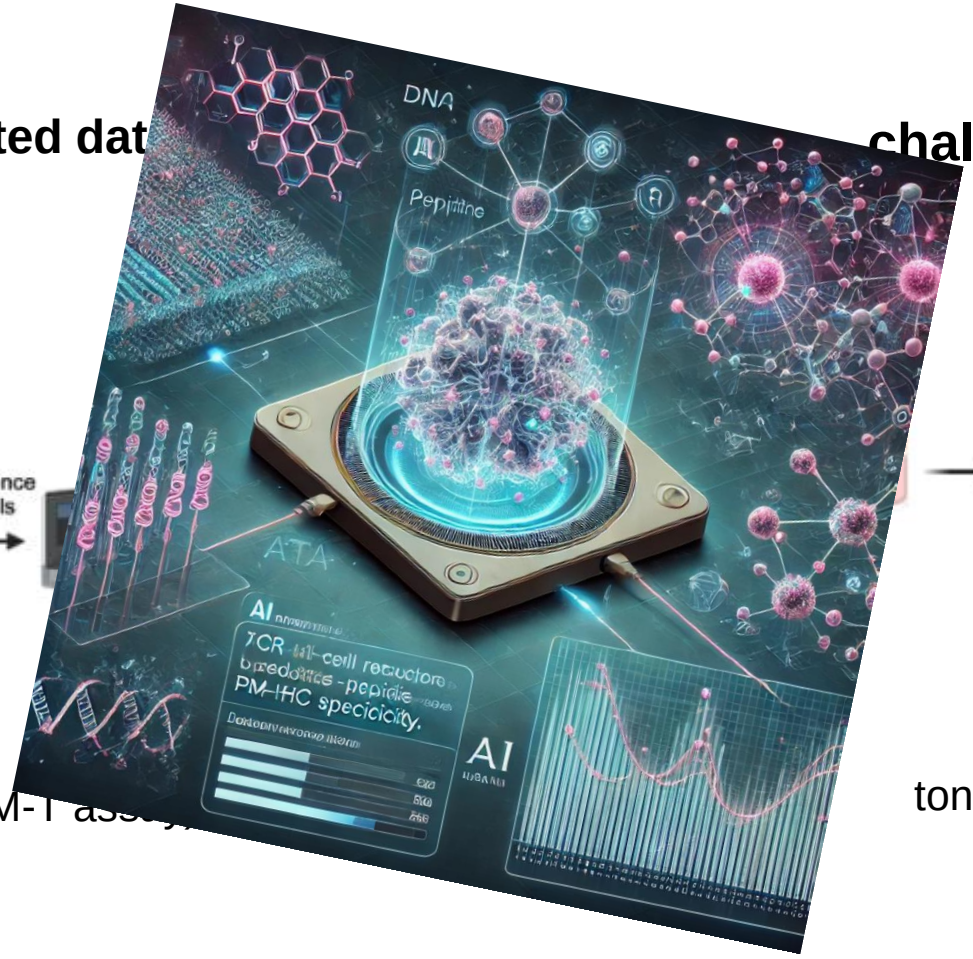
Sureshchandra, Henderson, …, ATM*, Saligrama*, Wagar*, bioRxiv 2024

How can we bridge the annotation gap?

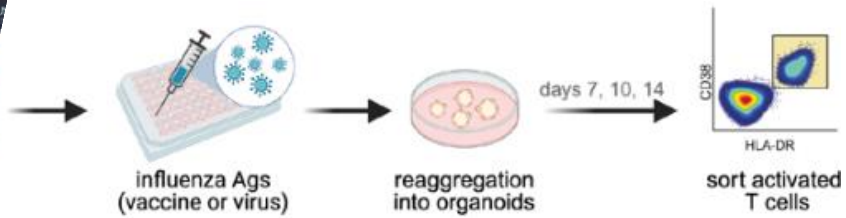
>99.99% of TCRs are orphan

generate more annotated dat





challenge experiments



antigen-specific sort (BEAM-T ass

tonsil organoids challenged with flu

Sureshchandra, Henderson, …, ATM*, Saligrama*, Wagar*, bioRxiv 2024

The state of *in silico* TCR-pMHC binding prediction

Nielsen … ATM … Barton Immunoinformatics 2024

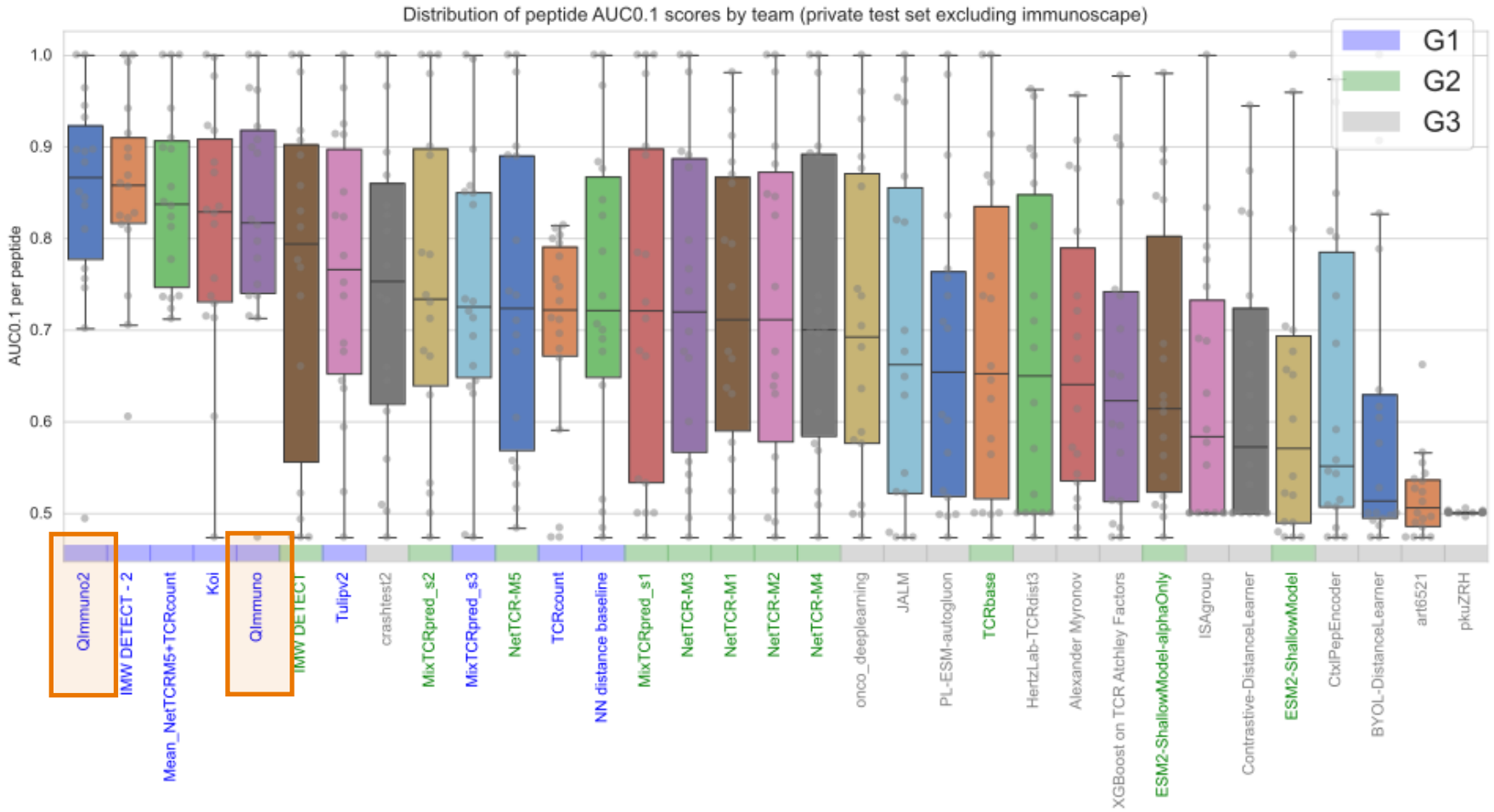
Nagano, Pyo, Milighetti, Henderson, Shawe-Taylor, Chain*, ATM* arXiv 2024

AI to the rescue? Predict TCR-pMHC binding from sequence

Immrep23 benchmark

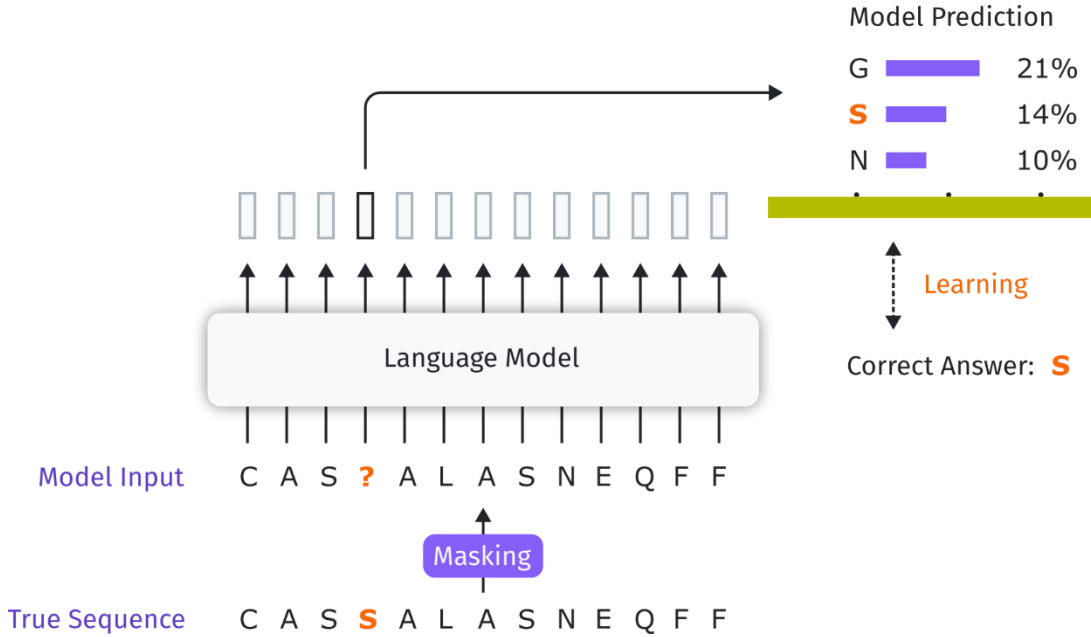
For each TCR determine most likely target among 20 pMHCs

No evidence of generalization of current deep learning models

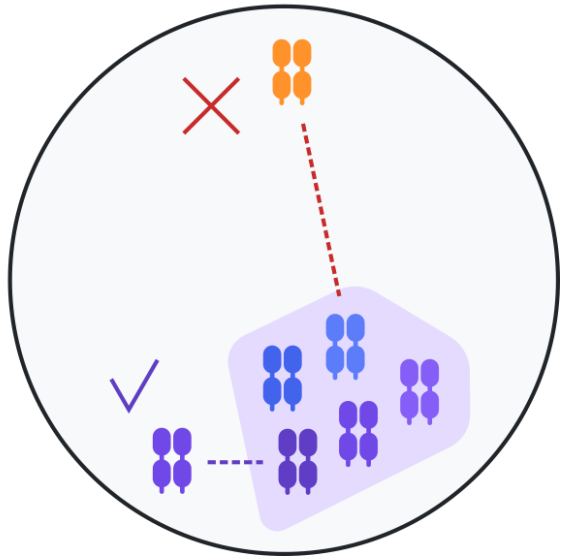


The premise of using large language models

pre-training



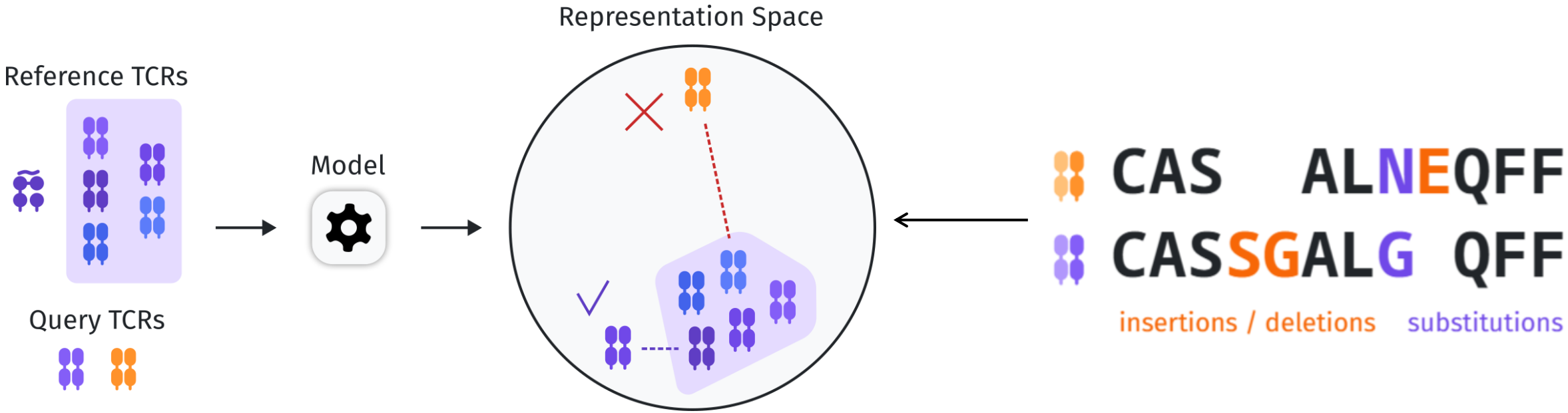
transfer learning



representation

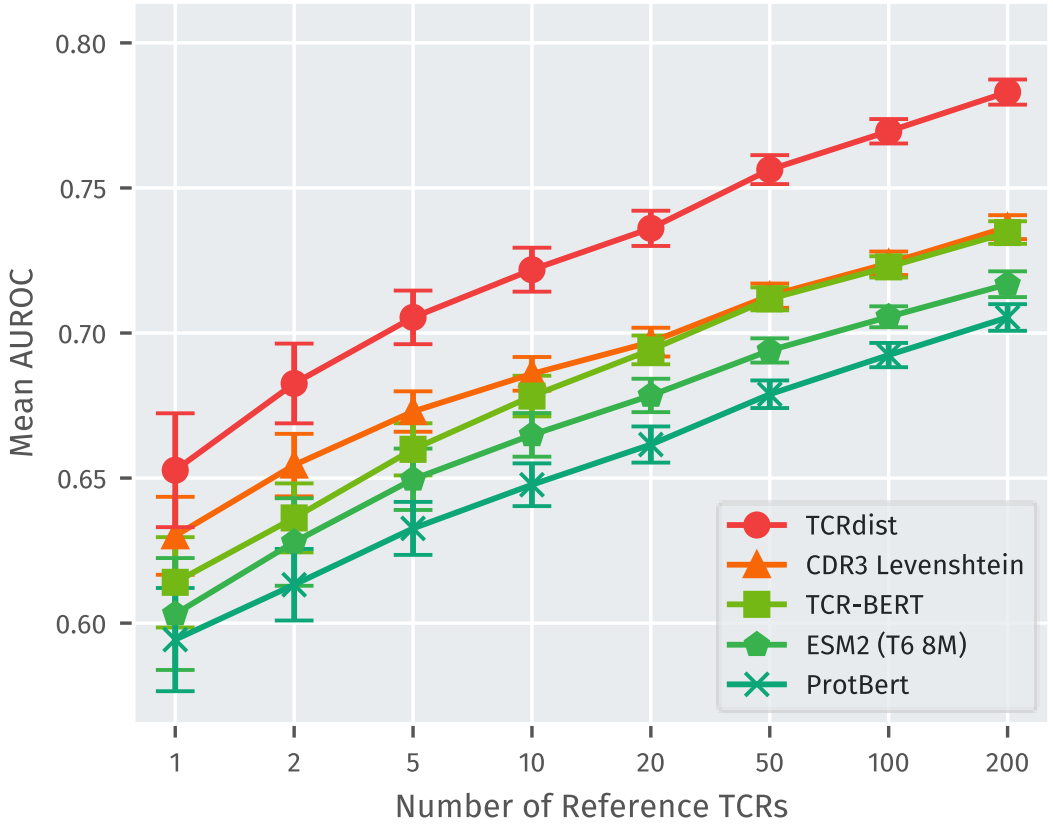
Nagano, Pyo, Milighetti, Henderson, Shawe-Taylor, Chain*, ATM* arXiv 2024

Benchmarking language models against sequence alignment



Nagano, Pyo, Milighetti, Henderson, Shawe-Taylor, Chain*, ATM* arXiv 2024

Language embeddings are inferior to sequence alignment for TCR-pMHC prediction



Nearest neighbour prediction

Positives: binders not used as reference
Negatives: all remaining TCRs

Data: 6 pMHC with > 300 αβTCRs
Average over pMHC / data splits

Nagano, Pyo, Milighetti, Henderson, Shawe-Taylor, Chain*, ATM* arXiv 2024

A statistical physics view on sequence-function maps

ATM, Callan PNAS 2023

Pyo, Nagano, Milighetti, Henderson, Callan, Chain, Wingreen, ATM (in preparation)

Each ligand defines a selection landscape

Repertoire

$P(\boldsymbol{\sigma})$

CASSWNGPTYEQYF
CSAPLGGGEQFF
CADPFRDRGSNQPQHF
CASSPGQGSYEQYF
CAISGQGTGEKYQPQHF
CSARDGTGNGYTF
CASSTIEGQGGRHTQYF
....

$Q(\boldsymbol{\sigma})$

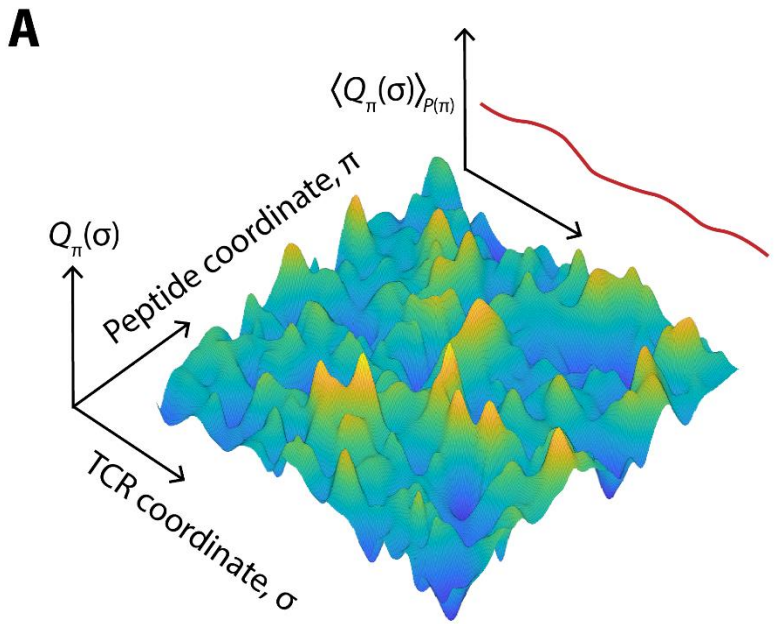
Selection factors

pMHC-specific repertoire

$Q(\boldsymbol{\sigma})P(\boldsymbol{\sigma})$

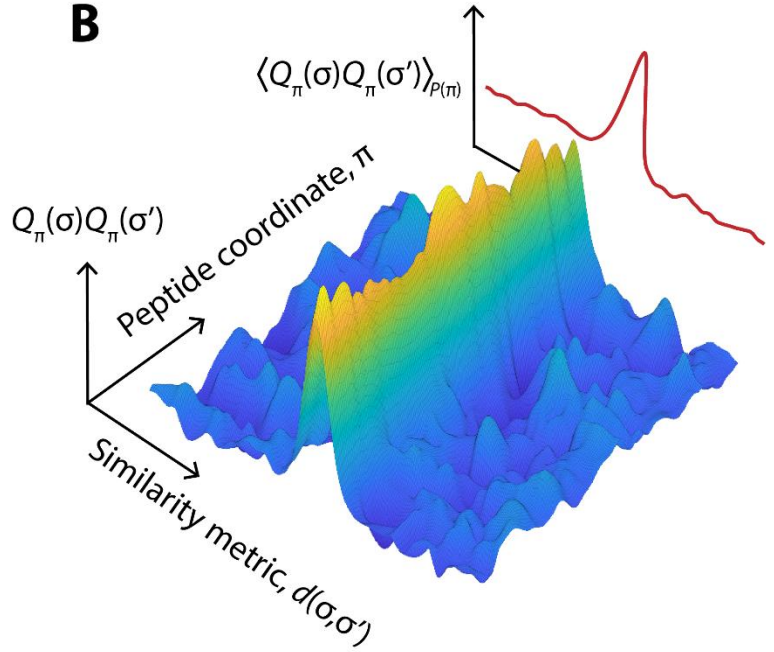
CSARDRVGNGYTF
CASSPGQGSYEQYF
CSARDGTGNGYTF
CSARDGTGNGYTF
...

Averaging selection landscapes is destructive



Expect flat marginal distribution

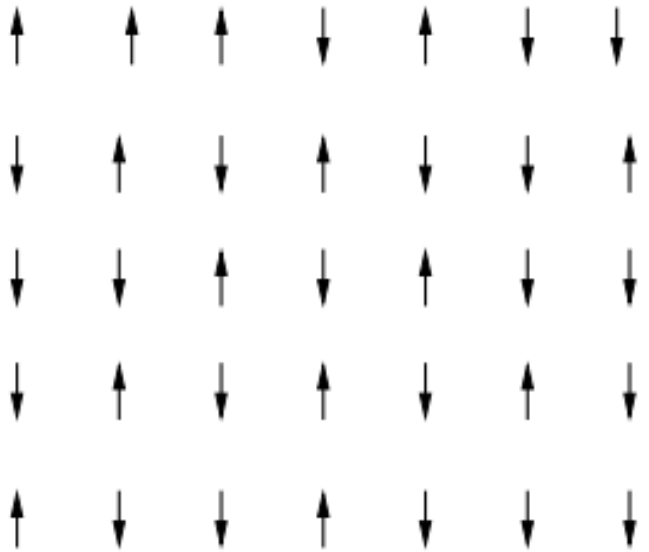
Modest selection relative to Pgen (Elhanati PNAS 2014)



Intuition:
similarity-relationships are generalizable

ATM, Callan PNAS 2023

Analogy with spin glasses



$$H = \sum_{<i,j>} \mathrm{J}_{ij}\sigma_i\sigma_j$$

Zero magnetization

$$\mathrm{m} = \langle\sigma_i\rangle = 0$$

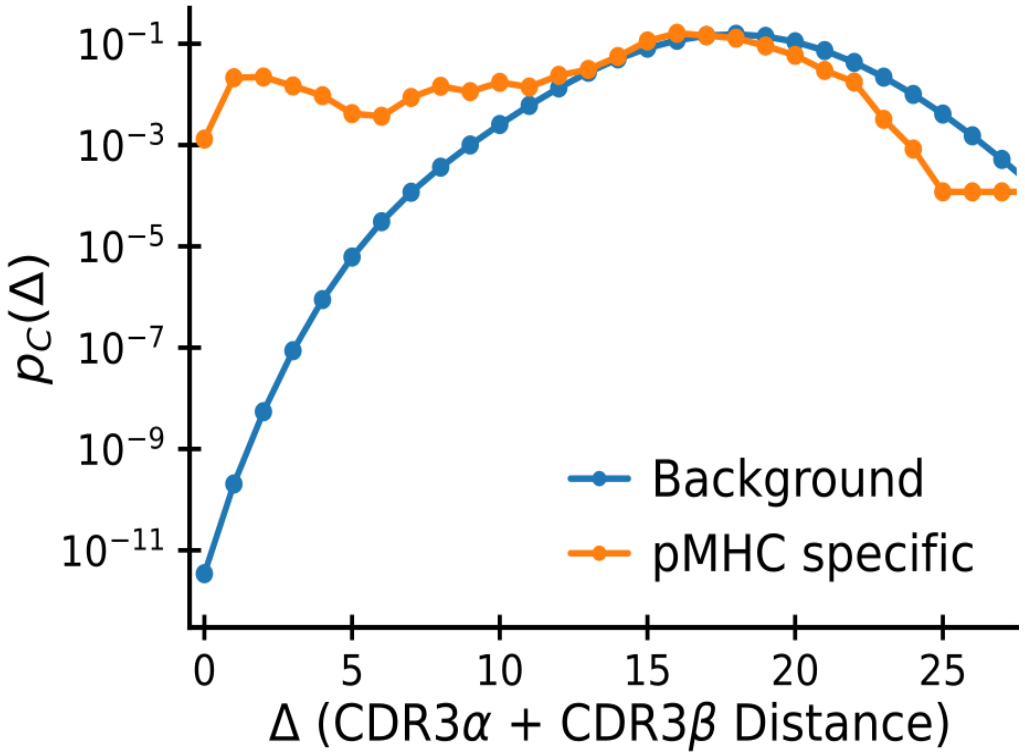
Non-trivial order parameter

$$q_{EA} = \lim_{t\to\infty}\langle\sigma_i(0)\sigma_i(t)\rangle \neq 0$$

Edwards, Anderson J Phys F 1975

Convergent TCRs in pMHC-specific repertoires

Data: Dash … Thomas Nature 2017



Histogram of pairwise sequence distances

$$p_C(\Delta) = \sum_{\{\boldsymbol{\sigma},\boldsymbol{\sigma}'\}} P(\boldsymbol{\sigma})P(\boldsymbol{\sigma}')\boldsymbol{I}_{d(\boldsymbol{\sigma},\boldsymbol{\sigma}')=\Delta}$$

Sequences at distance Δ

Sequence (near-) coincidences are highly enriched

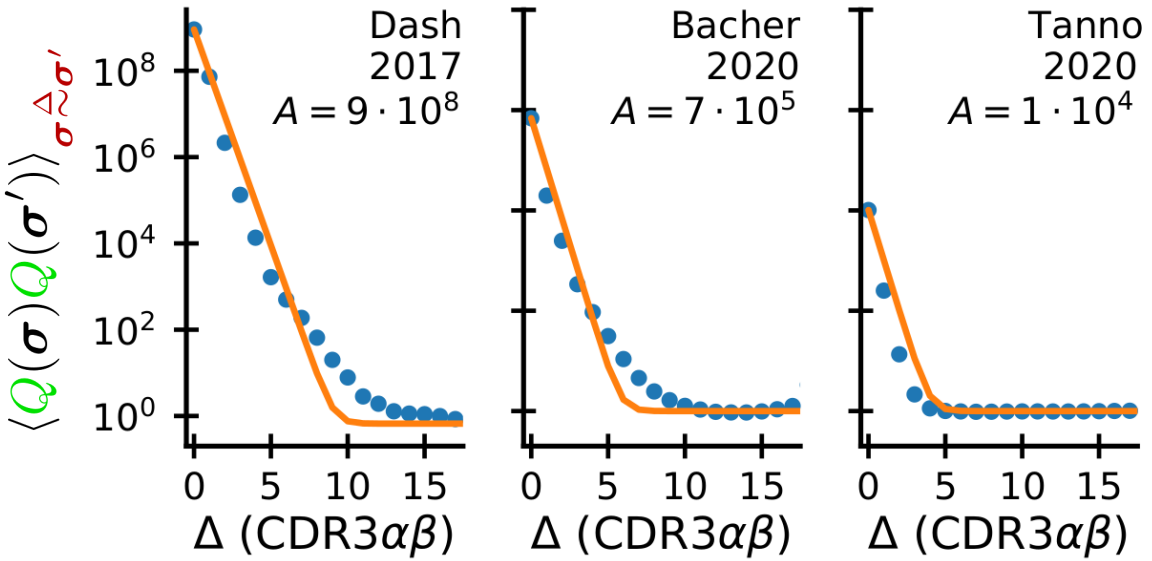
ATM Callan PNAS 2023

Sequence similarity relates to co-specificity in a generalizable manner

$$\frac{p_C[QP](\Delta)}{p_C[P](\Delta)} = \langle Q(\boldsymbol{\sigma})Q(\boldsymbol{\sigma}')\rangle_{\boldsymbol{\sigma}\overset{\Delta}{\sim}\boldsymbol{\sigma}'}$$

Pairs at d=Δ

Near-coincidence ratios
= how selection co-varies with sequence



Some insights into the biophysical determinants TCR co-specificity

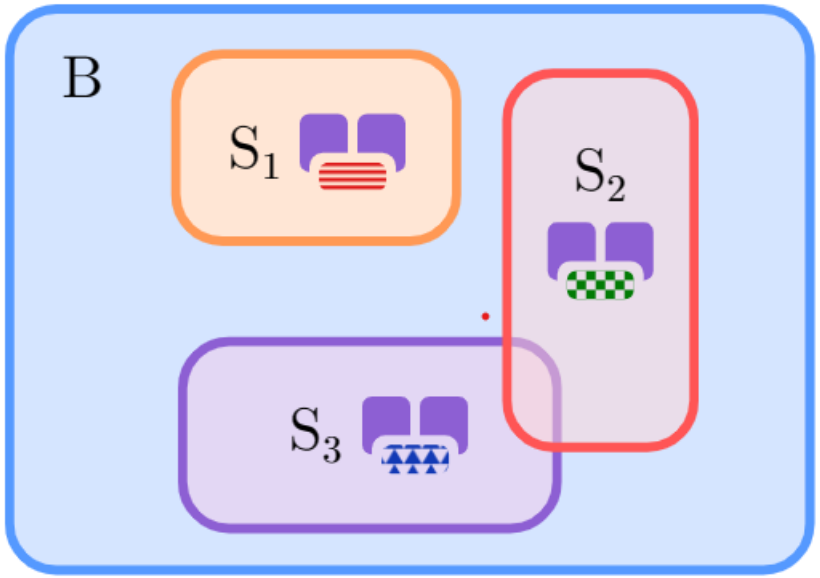
ATM PRE 2024

Henderson, Nagano, Milighetti, ATM PNAS 2024

Nagano, Pyo, Milighetti, Henderson, Shawe-Taylor, Chain*, ATM* arXiv 2024

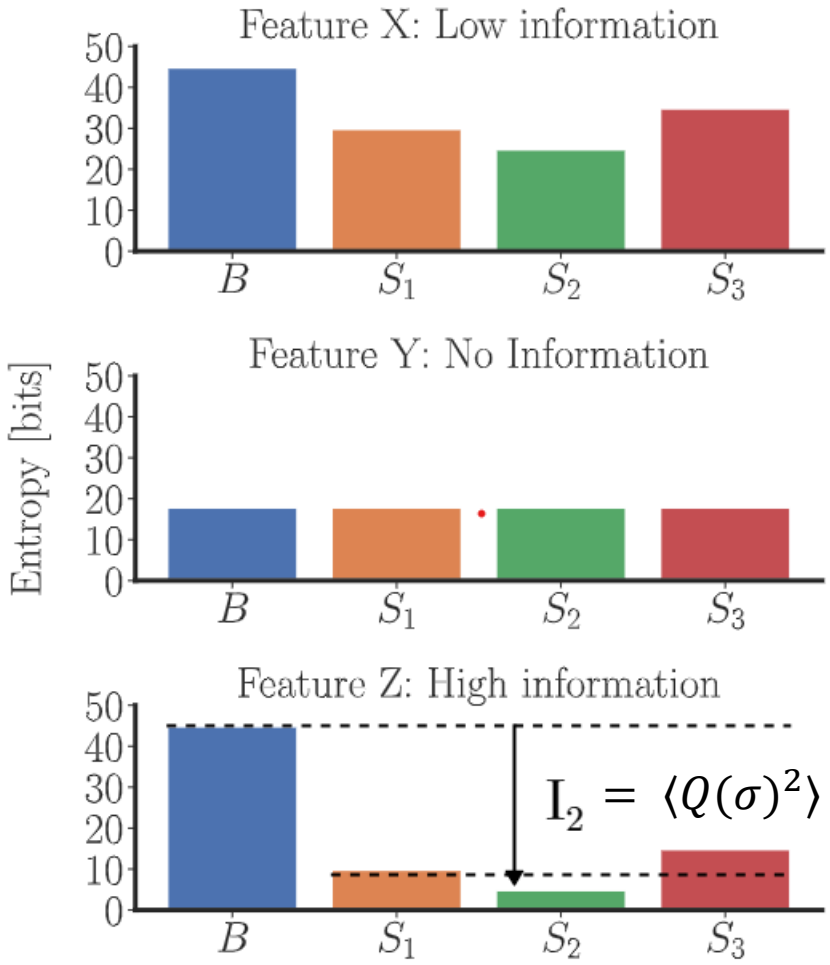
Pyo, Nagano, Milighetti, Henderson, Callan, Chain, Wingreen, ATM (in preparation)

The variance of selection factors is a measure of information



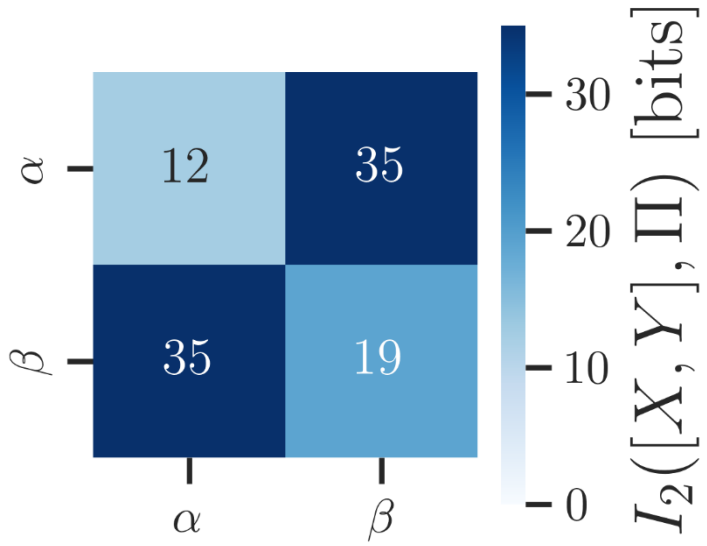
sequencing

$$H_2[P(X)] = -\log p_C[P(X)]$$



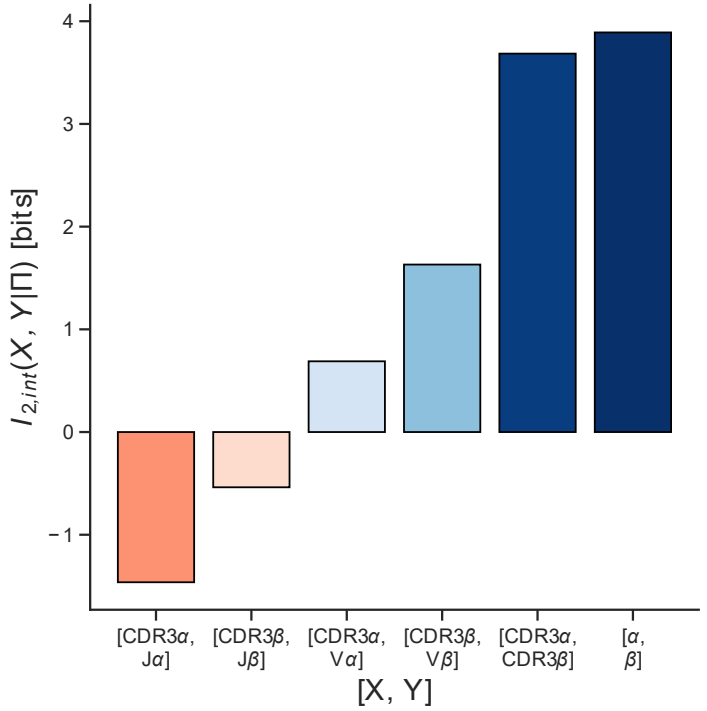
Henderson, Nagano, Milighetti, ATM PNAS 2024

Mapping TCR specificity information



$\beta > \alpha$

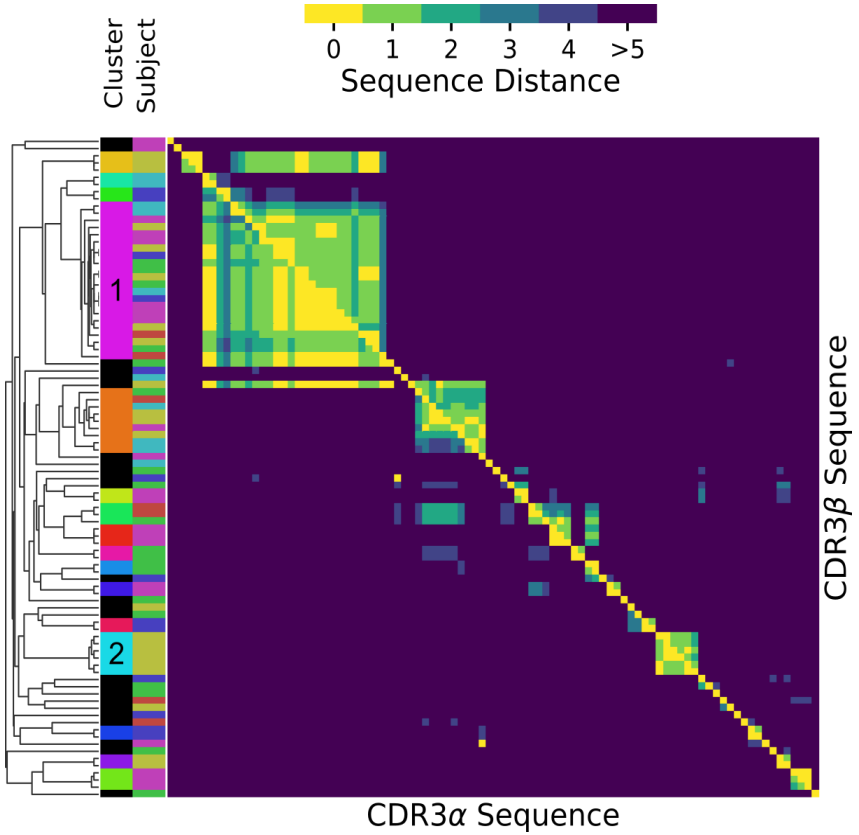
but: both chains matter



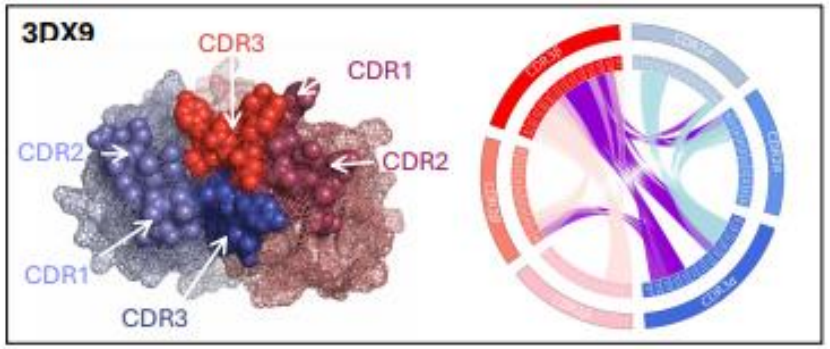
Chains are not independent

Henderson, Nagano, Milighetti, ATM PNAS 2024

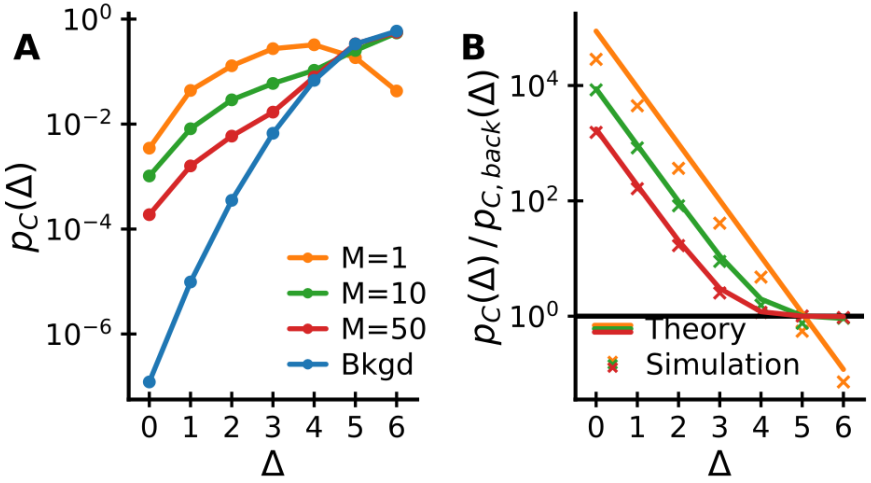
α / β chains have constrained pairing



ATM Callan PNAS 2023



Milighetti … ATM … Chain bioRxiv 2024

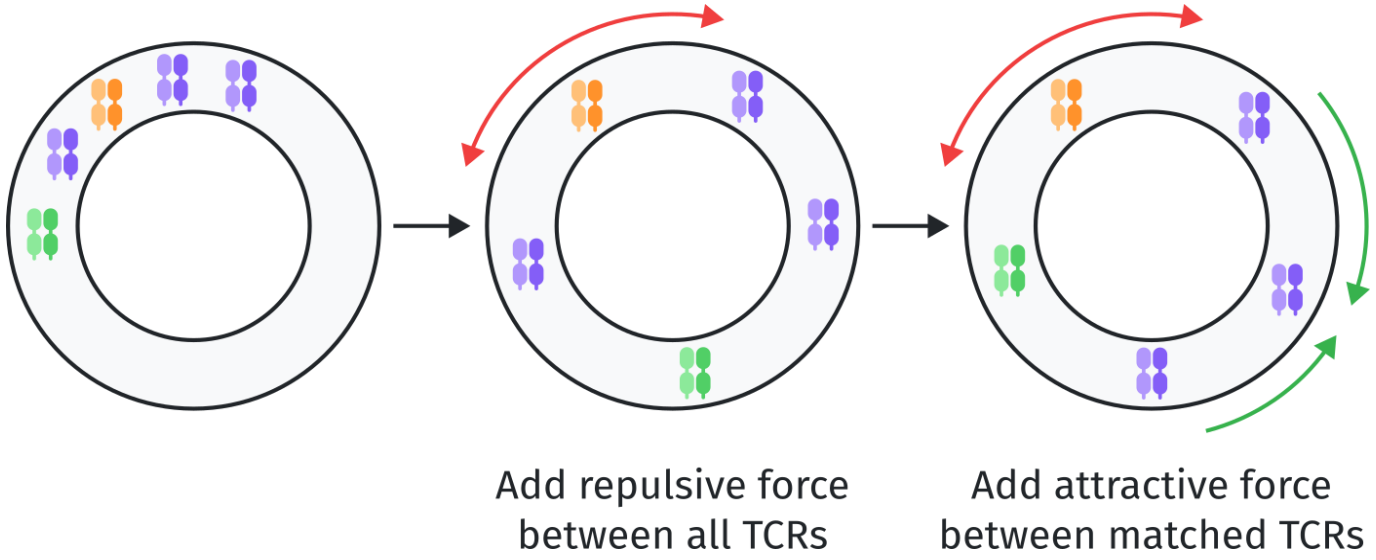


Discover biophysical determinants of co-specificity by contrastive learning

$$Q(\sigma, \sigma') \propto e^{-d(\sigma,\sigma')}$$

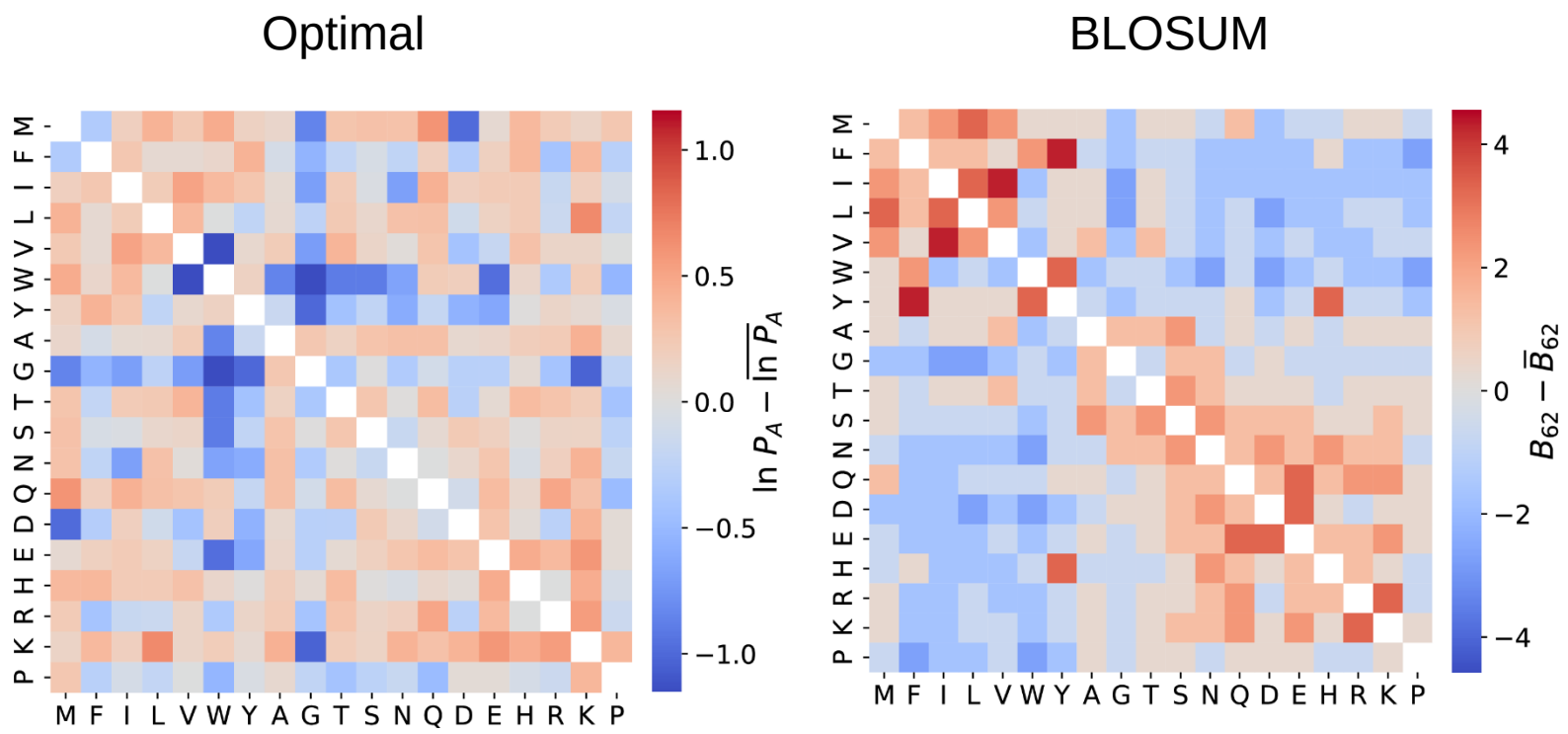
Optimal d? Pseudo-likelihood maximization:

$$\varepsilon(d|\mathcal{P},\mathcal{U}) = \langle d \rangle_{\mathcal{P}} + \log \left\langle e^{-d} \right\rangle_{\mathcal{U}}$$



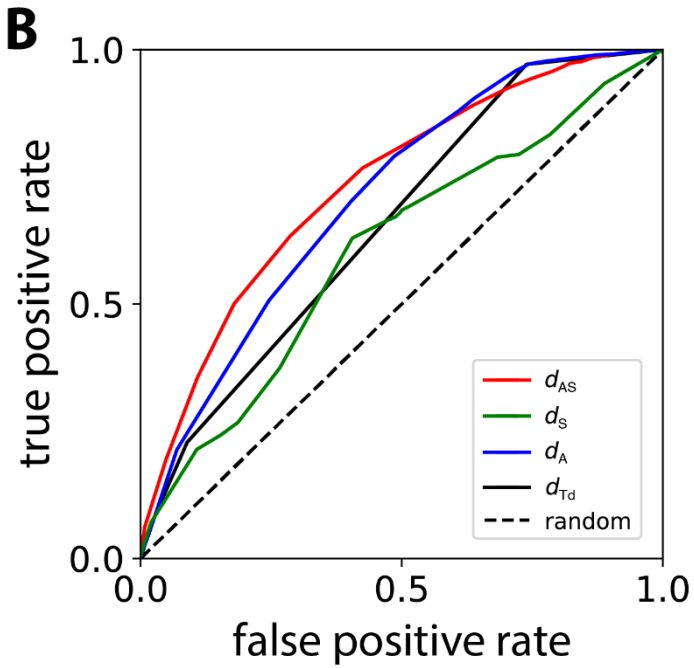
Nagano, Pyo, Milighetti, Henderson, Shawe-Taylor, Chain*, ATM* arXiv 2024
Pyo, Nagano, Milighetti, Henderson, Callan, Chain, Wingreen, ATM (in preparation)

Optimal alignment weights differ from BLOSUM

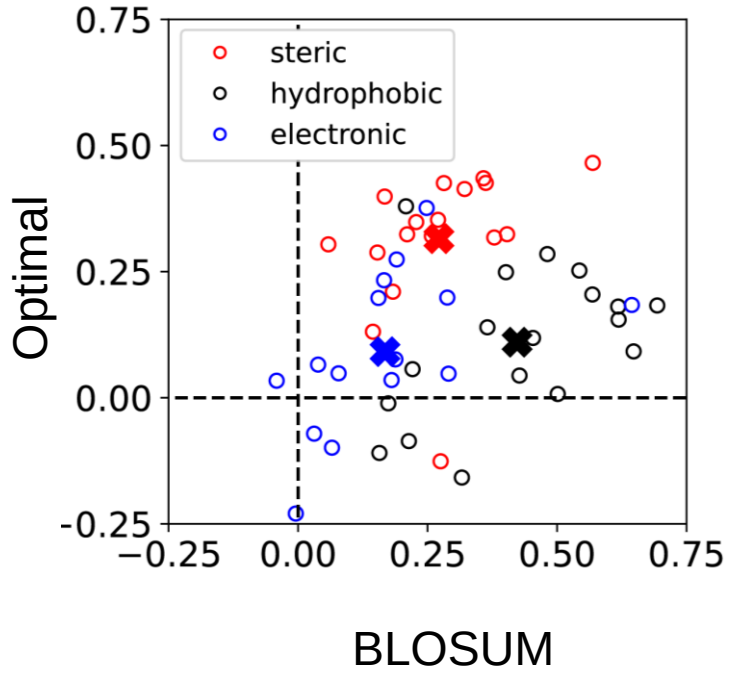


Pyo, Nagano, Milighetti, Henderson, Callan, Chain, Wingreen, ATM (in preparation)

Learned metric generalizes and is interpretable



generalization to pMHCs
not seen during training

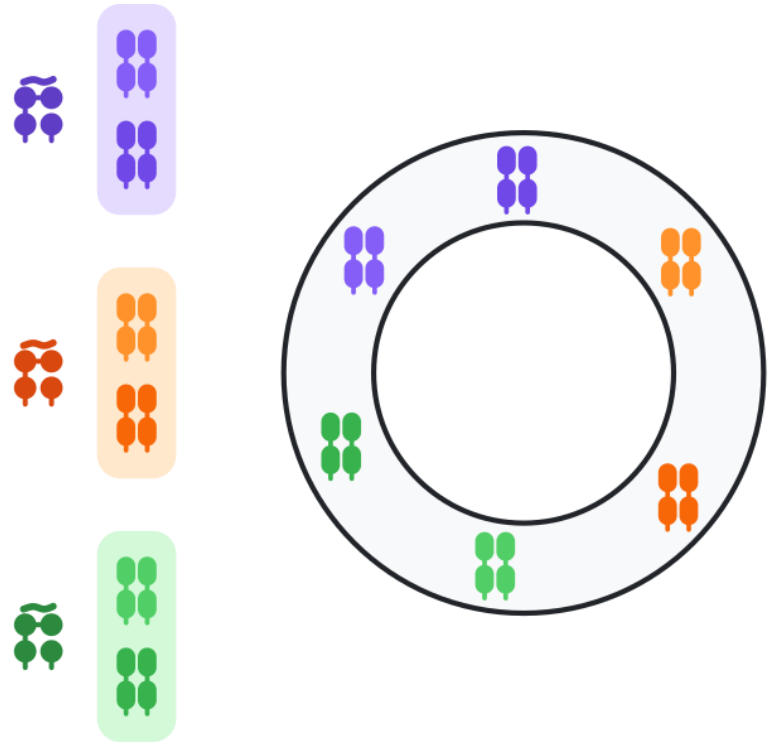


Correlation coefficients of substitution matrix with amino acid characteristics

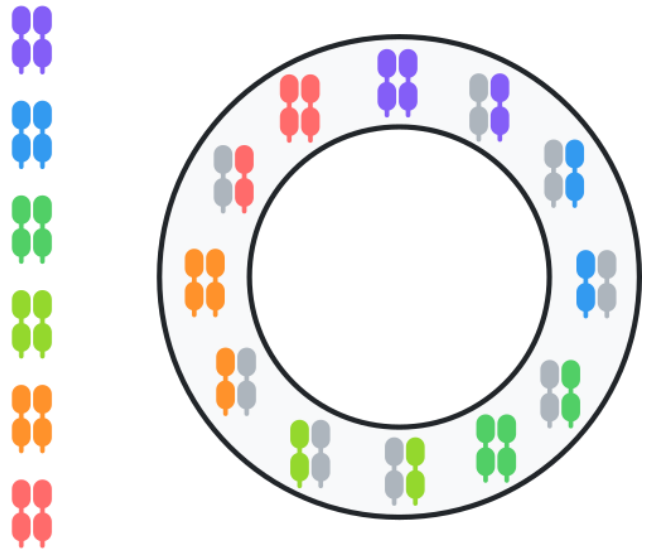
shape matters

Pyo, Nagano, Milighetti, Henderson, Callan, Chain, Wingreen, ATM (in preparation)

Contrastive training variants



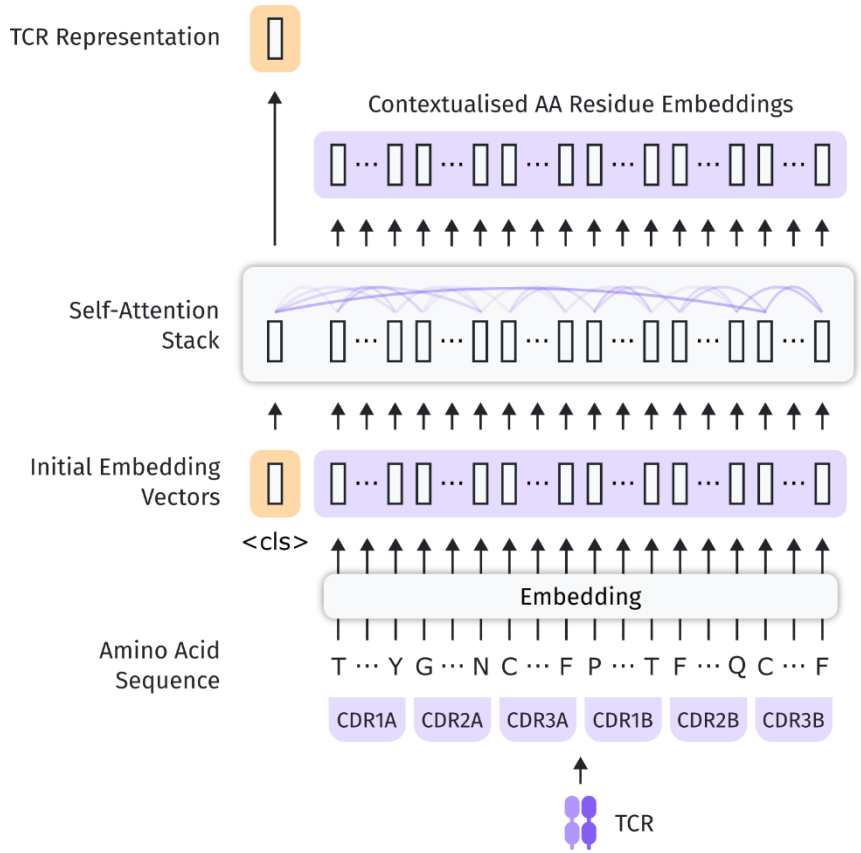
Supervised Contrastive Learning



Autocontrastive Learning

Architecture of SCEPTR

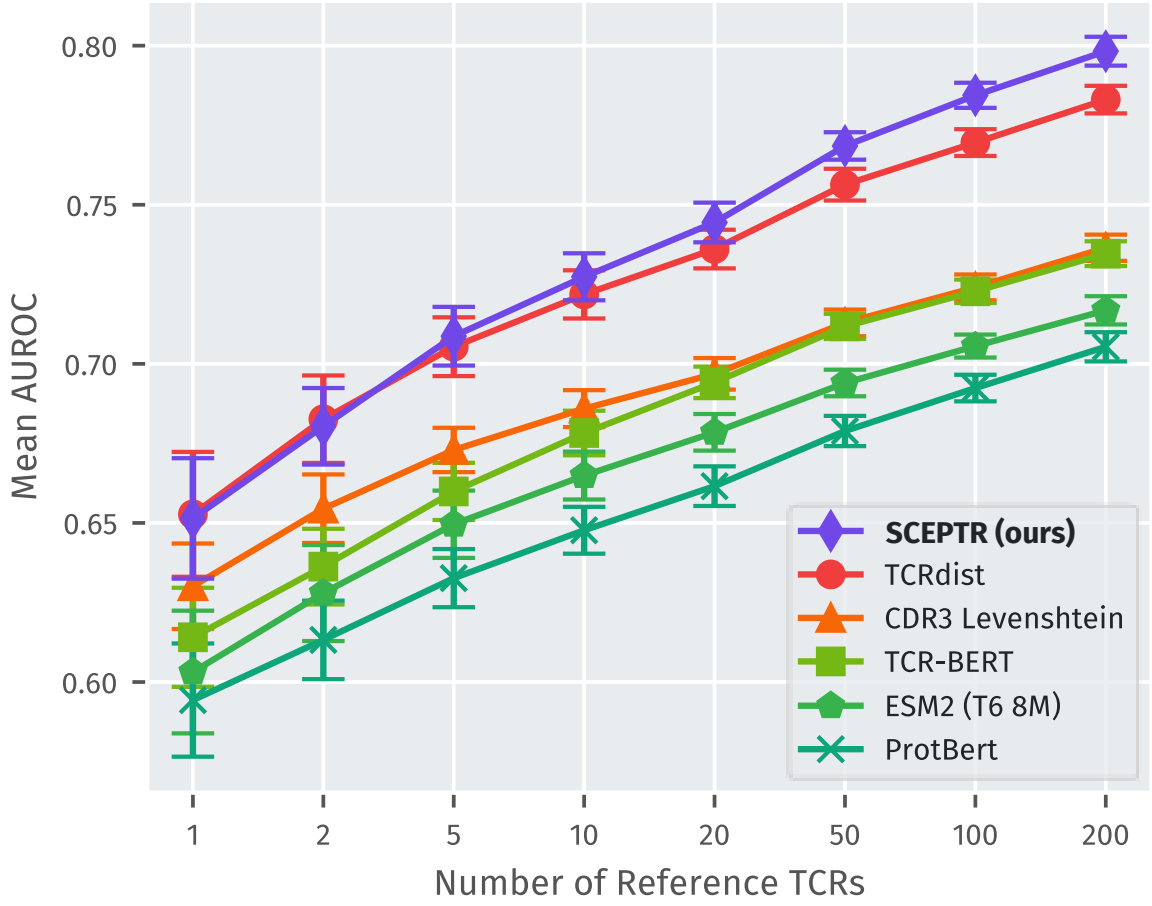
Simple Contrastive Embedding of the Peptide sequence of the T cell Receptor



- **Input**: AA sequence of CDRs 1/2/3 of the α and β TCR
- **Output**: Vector representation of the TCR

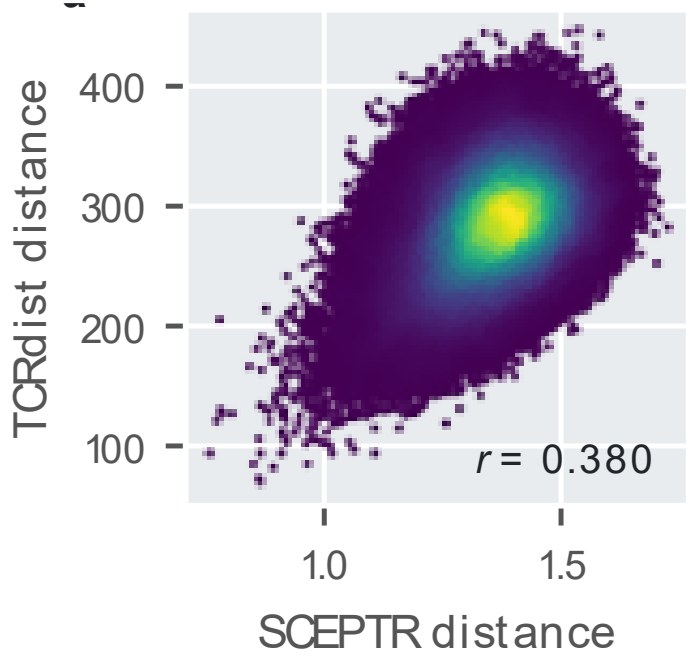
Nagano, Pyo, Milighetti, Henderson, Shawe-Taylor, Chain*, ATM* arXiv 2024

SCEPTR matches alignment accuracy using auto-contrastive pre-training

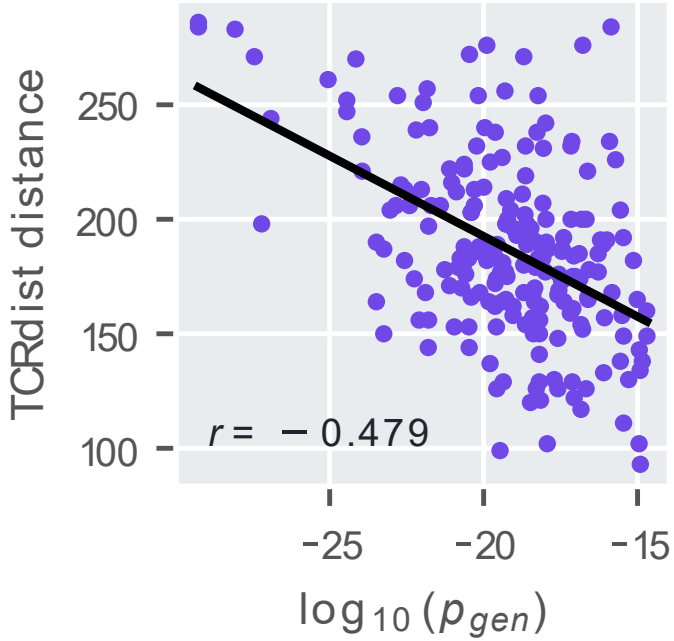


Nagano, Pyo, Milighetti, Henderson, Shawe-Taylor, Chain*, ATM* arXiv 2024

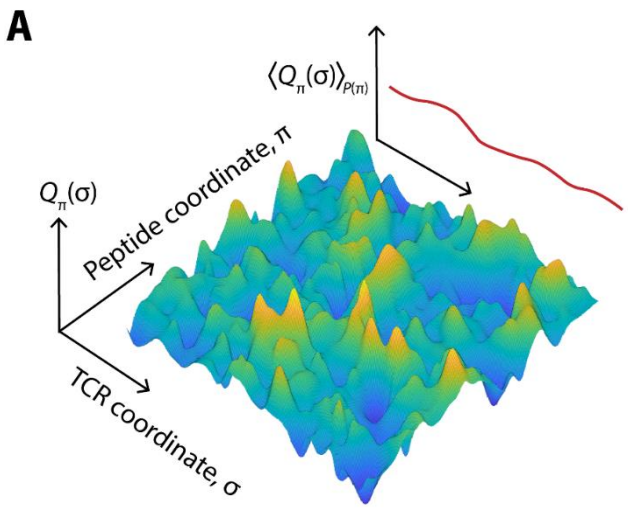
Comparing SCEPTR to TCRdist

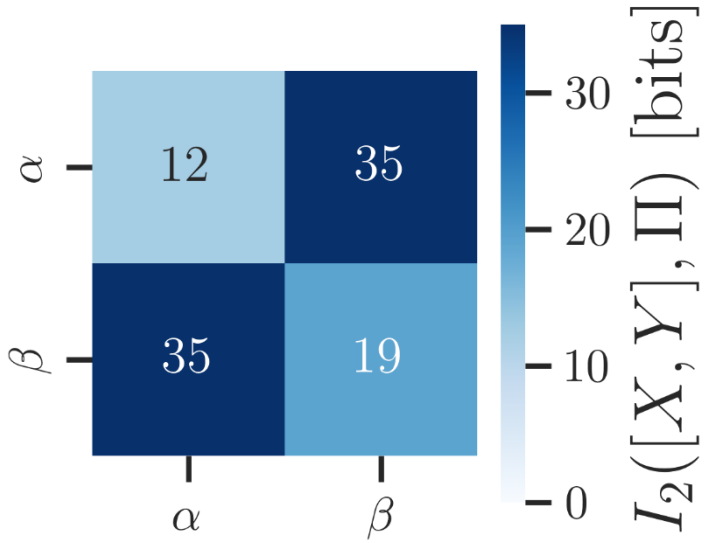


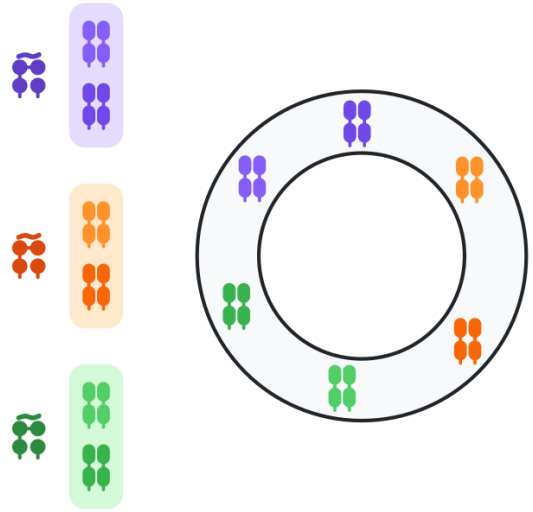
Approximates sequence alignment distances



Corrects for recombination biases

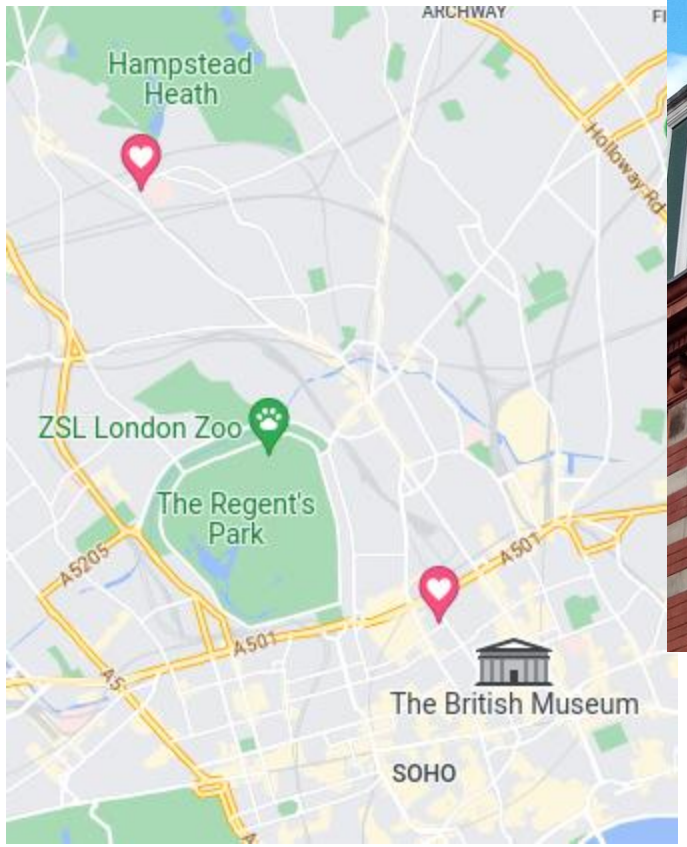






Which question would you ask if we could predict antigen-specificity from antibody/TCR sequence?

Come visit!



@andimscience

Interested in joining the group? Reach out!

www.qimmuno.com