AIRR Community

ANTIBODY SOCIETY

UiO : University of Oslo

Machine learning for the analysis of adaptive immune receptors and repertoires

Maria Chernigovskaya
mariiac@uio.no

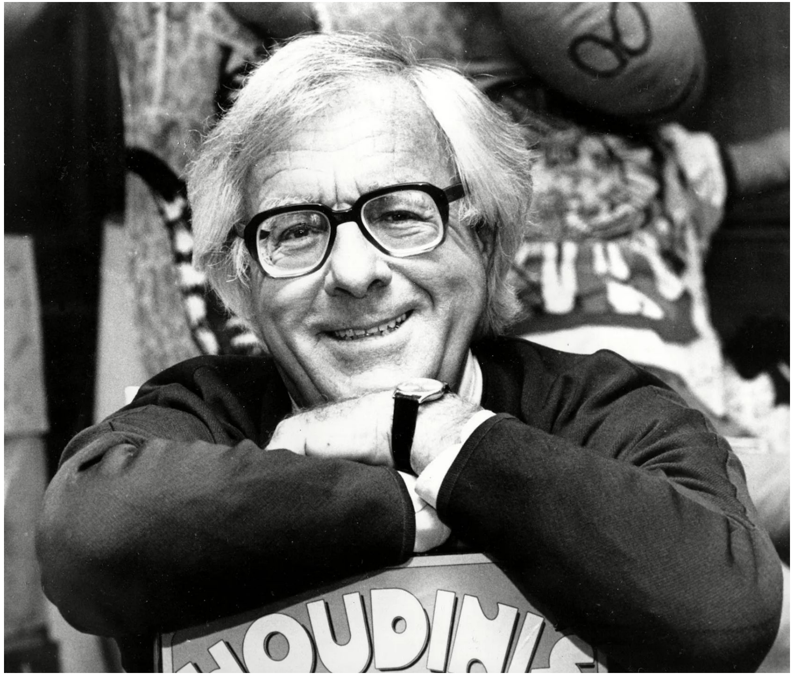
Milena Pavlović
milenpa@uio.no

AIRR Community Webinar

November 15, 2022

“Life is trying things to see if they work.”

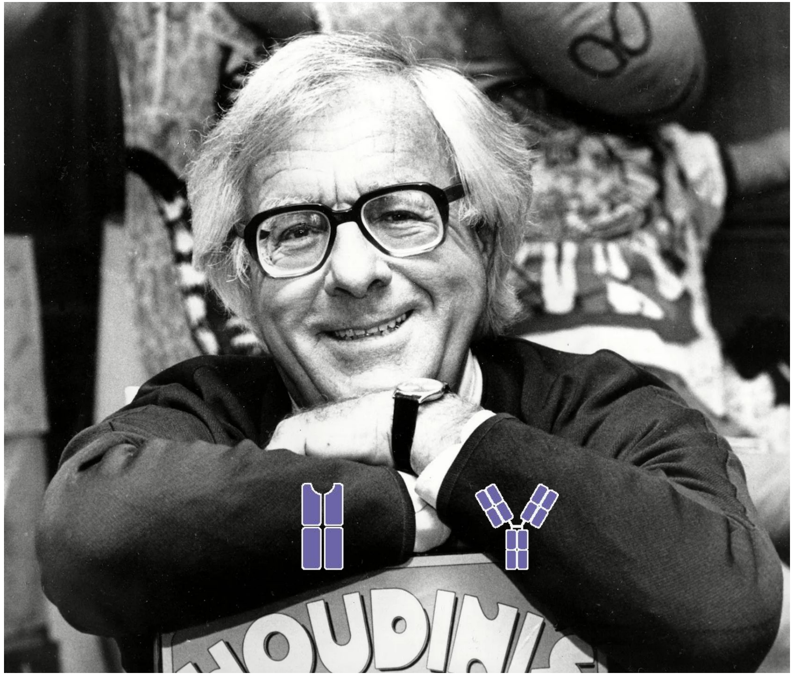
- Ray Bradbury



Machine learning for the analysis of adaptive immune receptors and repertoires

“~~Life~~ is trying things to see if they work.”

- AIRR researches

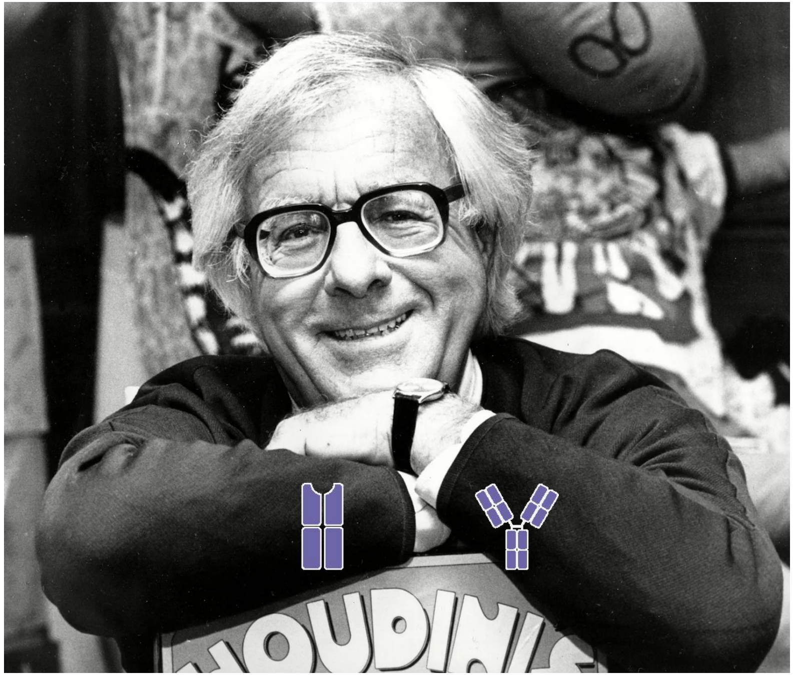


Machine learning for the analysis of adaptive immune receptors and repertoires

“~~Life~~ is trying things to see if they work.”

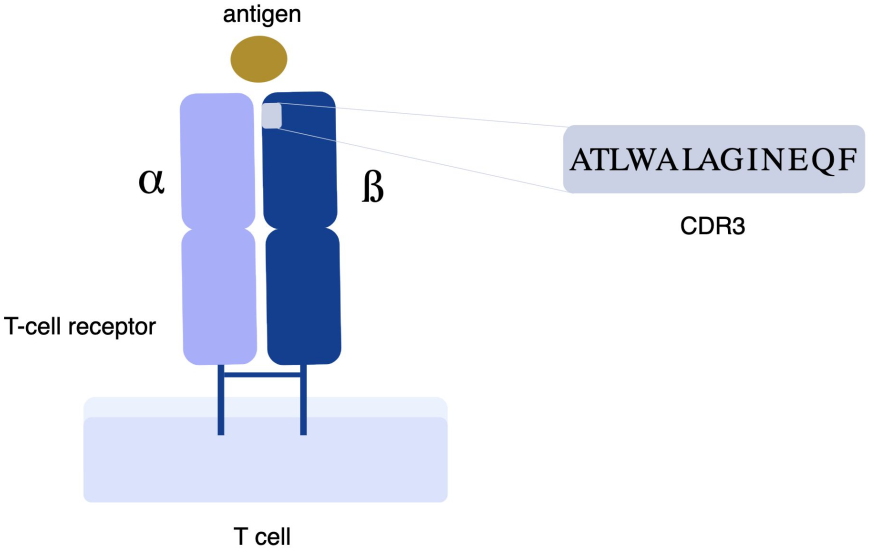
- AIRR researches

Or not?

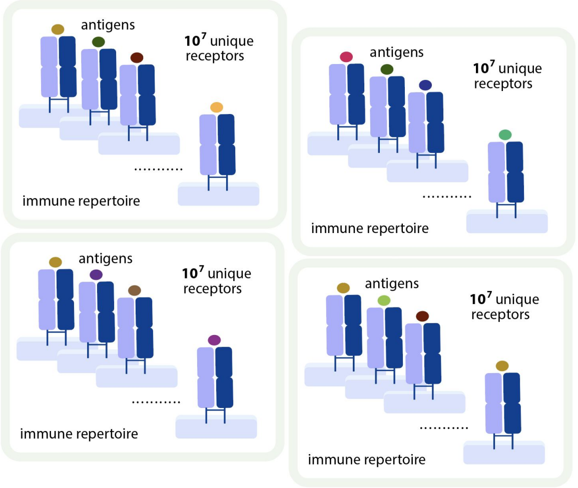


Adaptive immune receptors (AIRs) and repertoires (AIRRs)

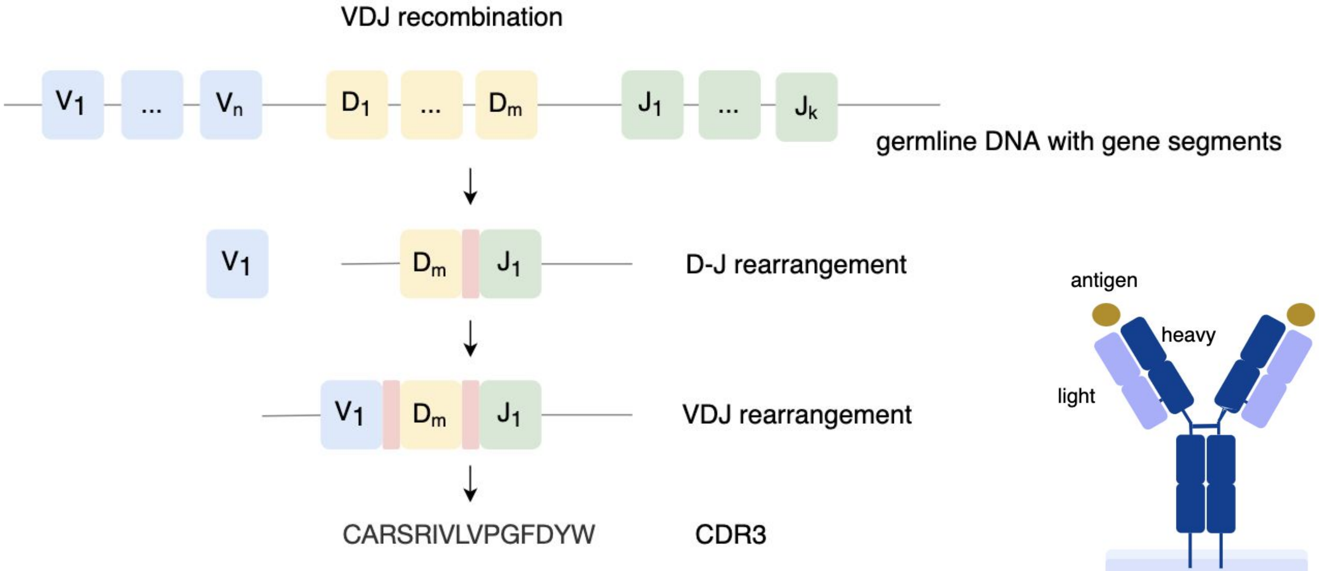
❏ Adaptive immune receptors (AIRs)

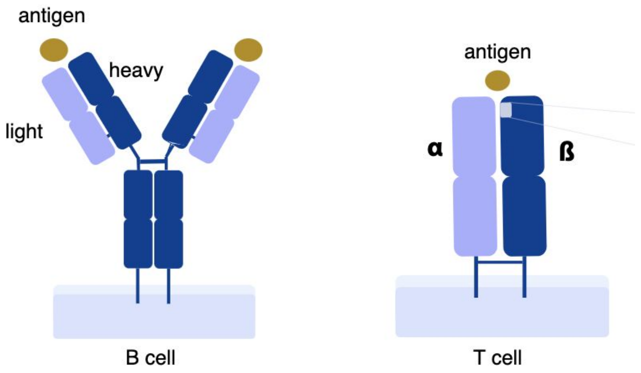


❏ Adaptive immune repertoires (AIRRs)



V(D)J recombination assembles AIRs (BCRs or TCRs)



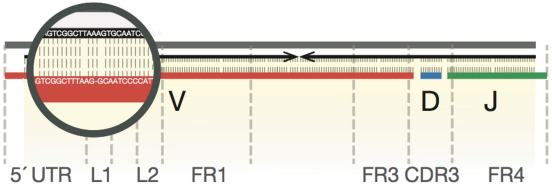


Alt et al. 1980, 1984, 1992, Bassing et al. 2000, 2002, Bareto et al. 2000, Schatz et al. 2012

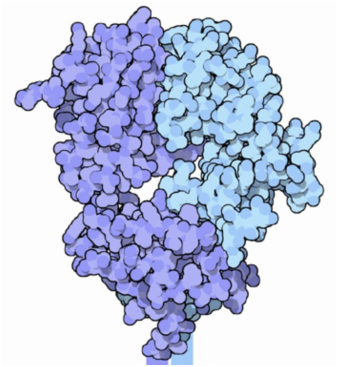
Overview of AIRR data on the receptor level

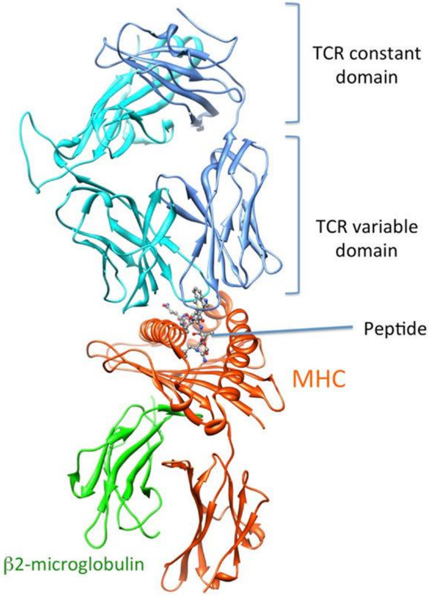
TCR/BCR sequence ⟶ TCR/BCR structure ⟶ TCR/BCR function (binding)

- □ Nt or aa
- □ Full-length or CDR3



```
TRBV7-3 + CASSDRHQPQHF + TRBJ2-7
```

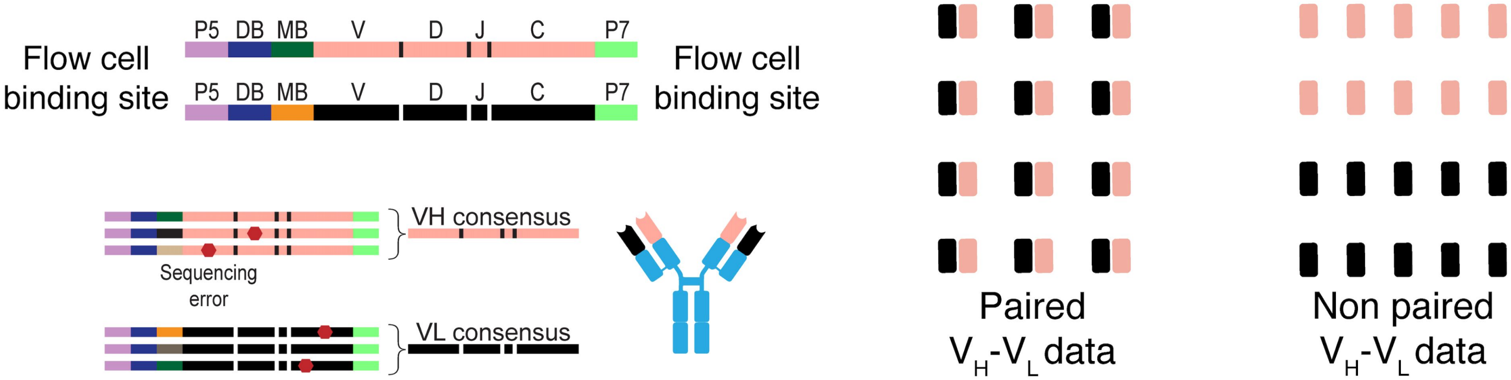




Bolotin et al. 2015, Zoete et al. 2013

Overview of AIRR data on the repertoire level

- AIR-seq (Bulk or Single cell, with or without UMIs)

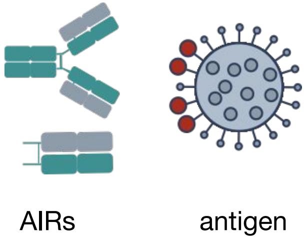


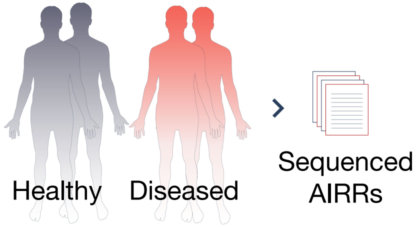
Benichou et al. 2012, Yaari et al. 2015, Brown et al. 2019

- Antibody repertoire proteomics (Cheung et al. 2012, Sato et al. 2012, Wine et al. 2015, Snapkov et al. 2021)
- Paired AIRs mapped to Ag specificity (Setliff et al. 2019, “A new way of exploring immunity” 2020)
- Paired AIRs + gene expression (Tu et al. 2019, Mathew et al. 2021, Shlesinger et al. 2022, Gao et al. 2022, Stephenson et al. 2021)

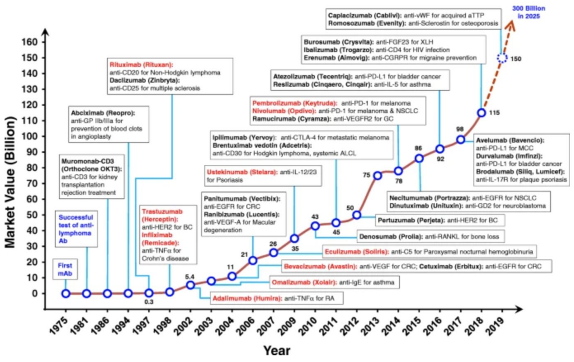
What immunological questions can we ask using AIRR data?

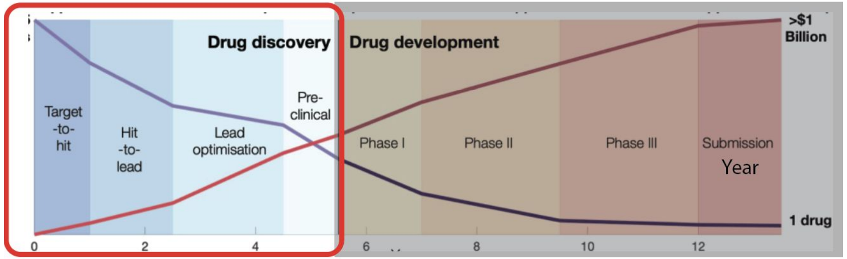
- ❑ Can we use an immune repertoire for disease diagnostics?





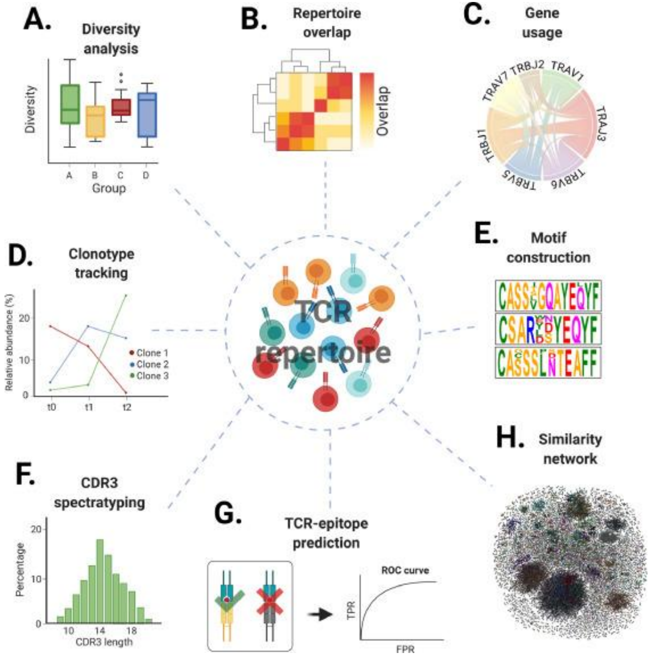
- ❑ How can we improve vaccines/CAR-T/mAbs/other therapeutics design?



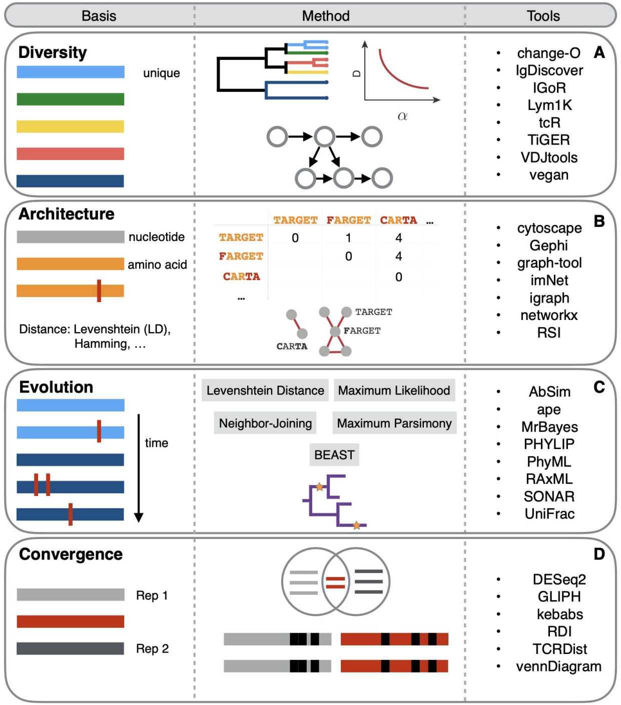


Lu et al. 2020

Some of these questions can be answered with various AIRR computational tools



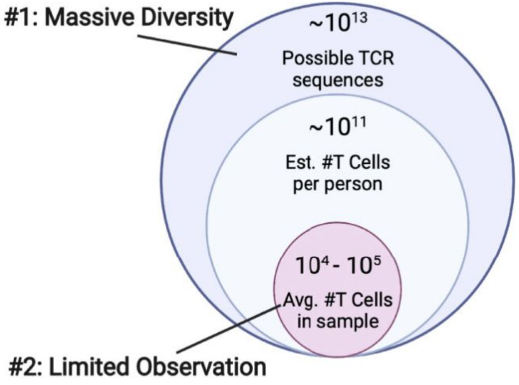
Valkiers et al. 2022

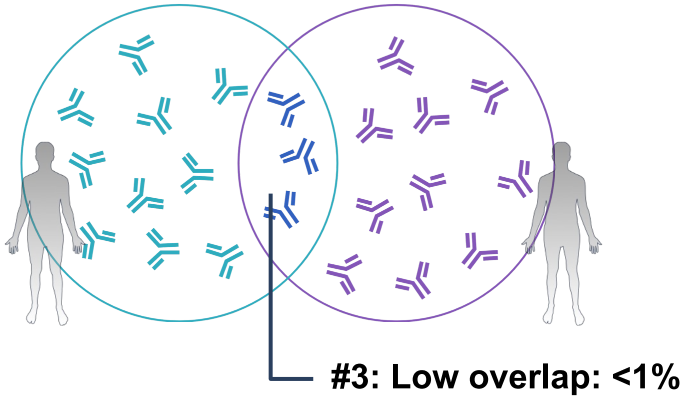


Basis	Method	Tools
Diversity		change-O IgDiscover IGoR Lym1K tcR TiGER VDJtools vegan
Architecture Distance: Levenshtein (LD), Hamming, …		TARGET	FARGET	CARTA	… | TARGET 0 1 4 FARGET 0 4 CARTA 0 …	cytoscape Gephi graph-tool imNet igraph networkx RSI
Evolution	Levenshtein Distance Maximum Likelihood Neighbor-Joining Maximum Parsimony BEAST	AbSim ape MrBayes PHYLIP PhyML RAxML SONAR UniFrac
Convergence		DESeq2 GLIPH kebabs RDI TCRDist vennDiagram

Miho et al. 2018

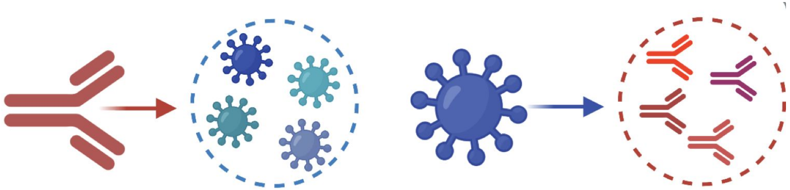
Challenges in computational analysis on AIRR data



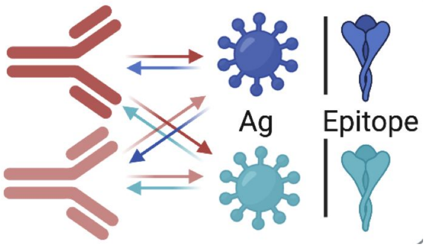


#3: Low overlap: <1%

#4: Many-to-many binding



#5: Sequence similarity ≠ similar binding



Katayama et al. 2022, Greiff et al. 2020

Machine learning (ML) provides various approaches to detect signals in complex high-dimensional data

What is machine learning?

- Machine learning (ML): a set of pattern recognition and function approximation techniques that find patterns within groups in (large amounts of) data

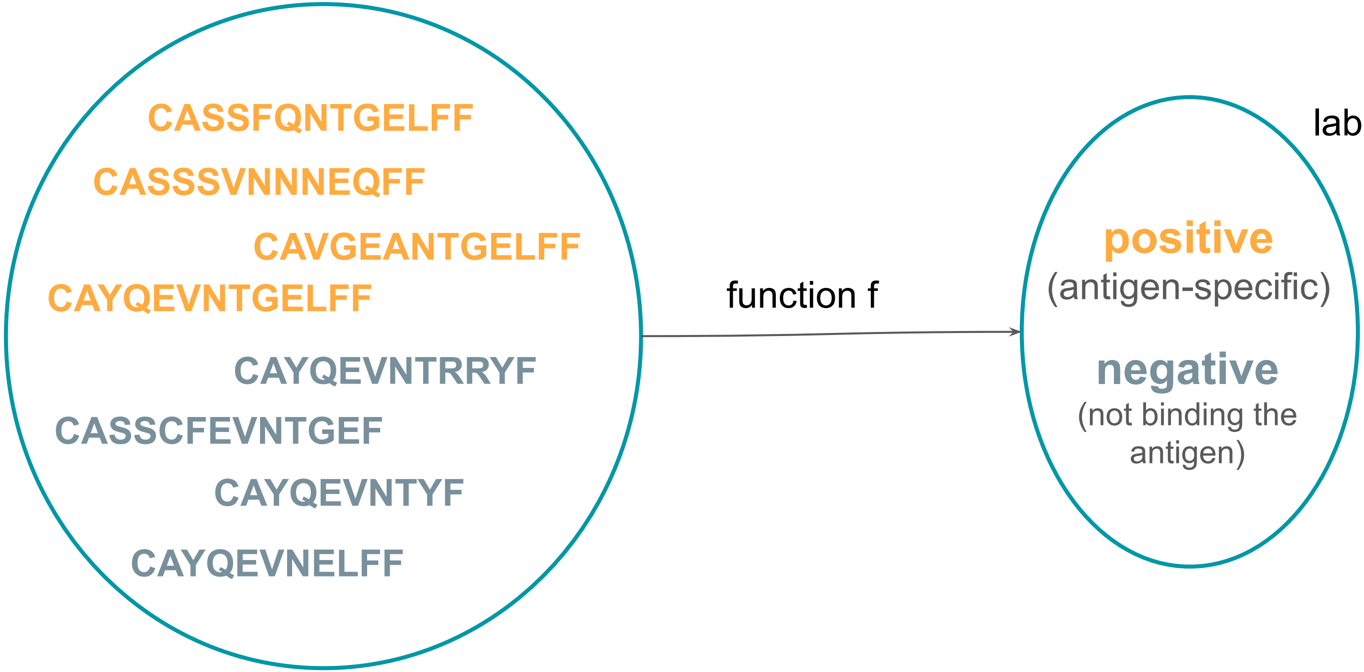
What is machine learning?

❏ Machine learning (ML): a set of pattern recognition and function approximation techniques that find patterns within groups in (large amounts of) data

❏ A set of methods that allow for making inferences about the data

Machine learning as a function approximation task

examples represented by features: immune receptor sequences



CASSFQNTGELFF

CASSSVNNNEQFF

CAVGEANTGELFF

CAYQEVNTGELFF

CAYQEVNTRRYF

CASSCFEVNTGEF

CAYQEVNTYF

CAYQEVNELFF

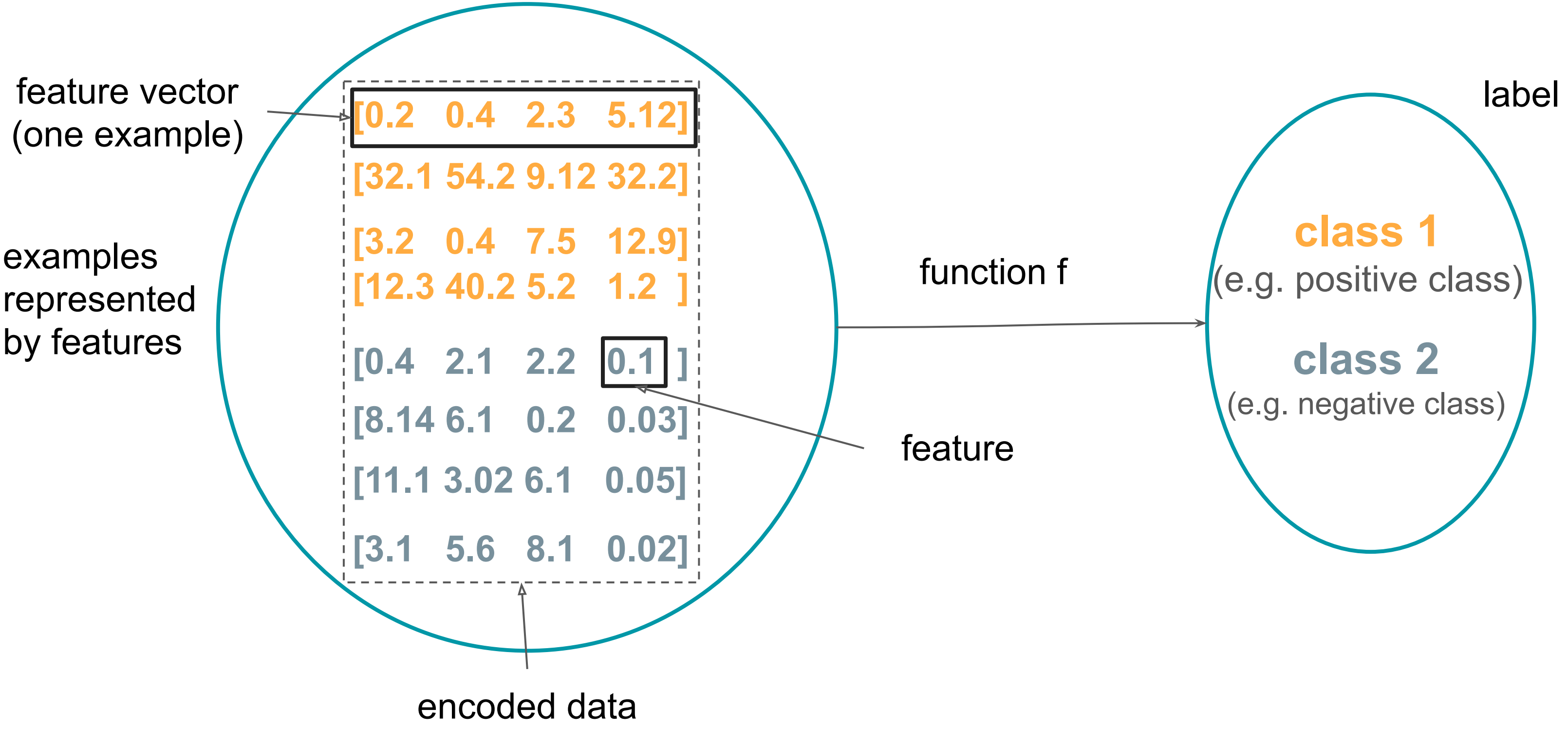
function f

positive
(antigen-specific)

negative
(not binding the antigen)

label: COVID-19 specificity

Machine learning as a function approximation task



feature vector
(one example)

examples
represented
by features

[0.2 0.4 2.3 5.12]
[32.1 54.2 9.12 32.2]
[3.2 0.4 7.5 12.9]
[12.3 40.2 5.2 1.2]
[0.4 2.1 2.2 0.1]
[8.14 6.1 0.2 0.03]
[11.1 3.02 6.1 0.05]
[3.1 5.6 8.1 0.02]

encoded data

function f

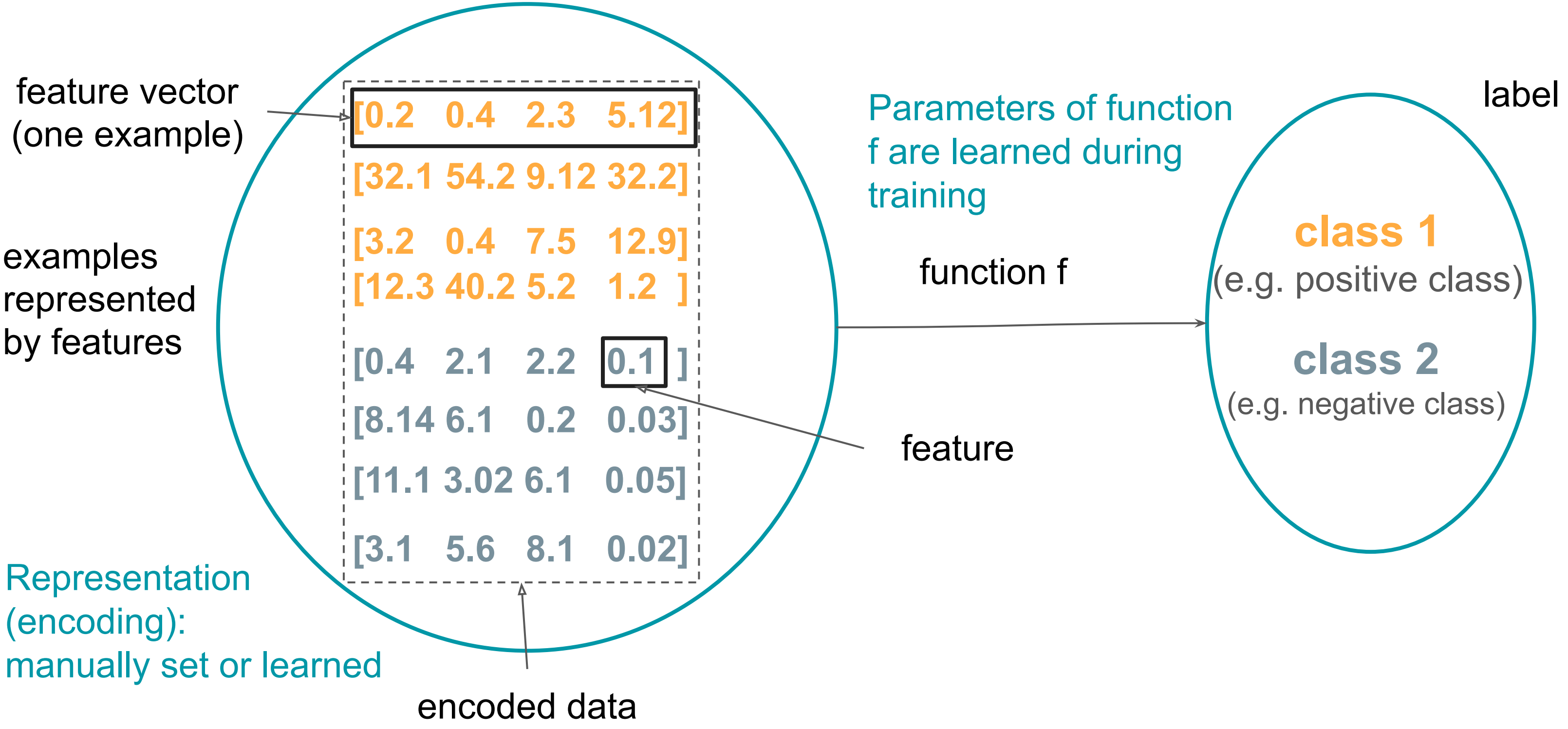
feature

label

class 1
(e.g. positive class)

class 2
(e.g. negative class)

Machine learning as a function approximation task



feature vector (one example)

examples represented by features

[0.2 0.4 2.3 5.12]

[32.1 54.2 9.12 32.2]

[3.2 0.4 7.5 12.9]

[12.3 40.2 5.2 1.2]

[0.4 2.1 2.2 0.1]

[8.14 6.1 0.2 0.03]

[11.1 3.02 6.1 0.05]

[3.1 5.6 8.1 0.02]

Representation (encoding): manually set or learned

encoded data

Parameters of function f are learned during training

function f

feature

label

class 1

(e.g. positive class)

class 2

(e.g. negative class)

Building predictive models

❏ Machine learning discovers statistical associations in the data
→ these associations enable good prediction

Building predictive models

- Machine learning discovers statistical associations in the data
  → these associations enable good prediction

- Aim: get a good predictive model, but also get biological insight
  - This is why we want the ML models to be interpretable

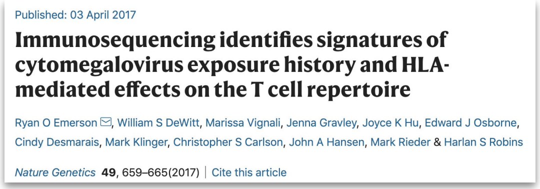
Building predictive models

- ❏ Machine learning discovers statistical associations in the data
→ these associations enable good prediction

- ❏ Aim: get a good predictive model, but also get biological insight
  - ❏ This is why we want the ML models to be interpretable

- ❏ Not causal relations, but starting points for further analyses

There is a surge in (AIRR) ML studies

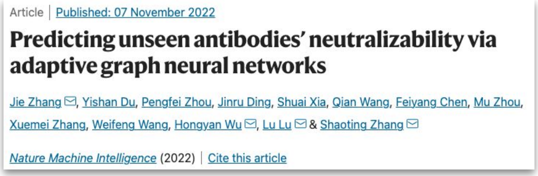


Published: 03 April 2017

Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire

Ryan O Emerson, William S DeWitt, Marissa Vignali, Jenna Gravley, Joyce K Hu, Edward J Osborne, Cindy Desmarais, Mark Klinger, Christopher S Carlson, John A Hansen, Mark Rieder & Harlan S Robins

Nature Genetics **49**, 659–665(2017) | Cite this article

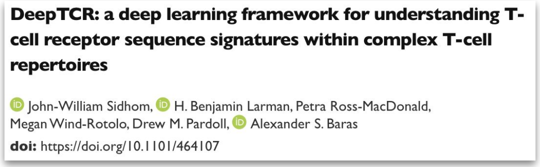


Article | Published: 07 November 2022

Predicting unseen antibodies' neutralizability via adaptive graph neural networks

Jie Zhang, Yishan Du, Pengfei Zhou, Jinru Ding, Shuai Xia, Qian Wang, Feiyang Chen, Mu Zhou, Xuemei Zhang, Weifeng Wang, Hongyan Wu, Lu Lu & Shaoting Zhang

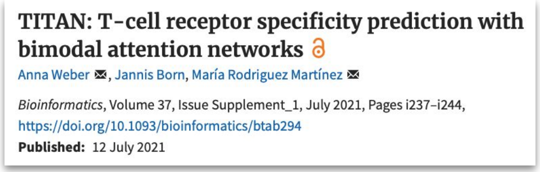
Nature Machine Intelligence (2022) | Cite this article



DeepTCR: a deep learning framework for understanding T-cell receptor sequence signatures within complex T-cell repertoires

John-William Sidhom, H. Benjamin Larman, Petra Ross-MacDonald, Megan Wind-Rotolo, Drew M. Pardoll, Alexander S. Baras

doi: https://doi.org/10.1101/464107

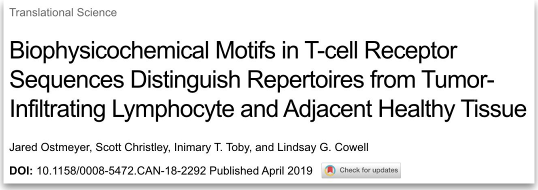


TITAN: T-cell receptor specificity prediction with bimodal attention networks

Anna Weber, Jannis Born, María Rodriguez Martínez

Bioinformatics, Volume 37, Issue Supplement_1, July 2021, Pages i237–i244, https://doi.org/10.1093/bioinformatics/btab294

Published: 12 July 2021

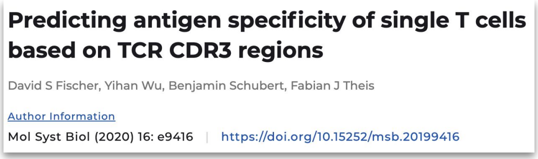


Translational Science

Biophysicochemical Motifs in T-cell Receptor Sequences Distinguish Repertoires from Tumor-Infiltrating Lymphocyte and Adjacent Healthy Tissue

Jared Ostmeyer, Scott Christley, Inimary T. Toby, and Lindsay G. Cowell

DOI: 10.1158/0008-5472.CAN-18-2292 Published April 2019 Check for updates

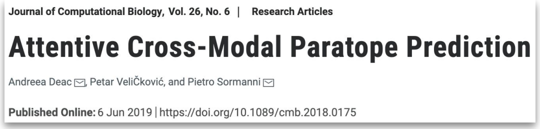


Predicting antigen specificity of single T cells based on TCR CDR3 regions

David S Fischer, Yihan Wu, Benjamin Schubert, Fabian J Theis

Author Information

Mol Syst Biol (2020) 16: e9416 | https://doi.org/10.15252/msb.20199416

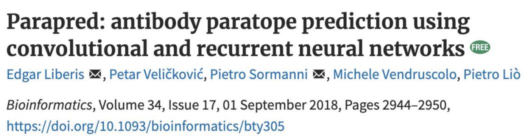


Journal of Computational Biology, Vol. 26, No. 6 | **Research Articles**

Attentive Cross-Modal Paratope Prediction

Andreea Deac, Petar Veličković, and Pietro Sormanni

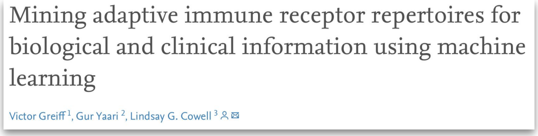
Published Online: 6 Jun 2019 | https://doi.org/10.1089/cmb.2018.0175



Parapred: antibody paratope prediction using convolutional and recurrent neural networks FREE

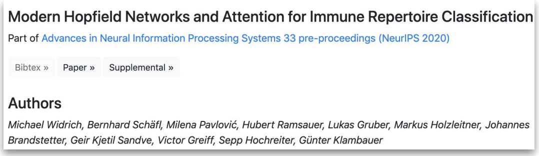
Edgar Liberis, Petar Veličković, Pietro Sormanni, Michele Vendruscolo, Pietro Liò

Bioinformatics, Volume 34, Issue 17, 01 September 2018, Pages 2944–2950, https://doi.org/10.1093/bioinformatics/bty305



Mining adaptive immune receptor repertoires for biological and clinical information using machine learning

Victor Greiff [1], Gur Yaari [2], Lindsay G. Cowell [3]

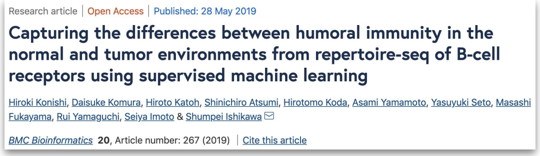


Modern Hopfield Networks and Attention for Immune Repertoire Classification

Part of Advances in Neural Information Processing Systems 33 pre-proceedings (NeurIPS 2020)

Authors

Michael Widrich, Bernhard Schäfl, Milena Pavlović, Hubert Ramsauer, Lukas Gruber, Markus Holzleitner, Johannes Brandstetter, Geir Kjetil Sandve, Victor Greiff, Sepp Hochreiter, Günter Klambauer

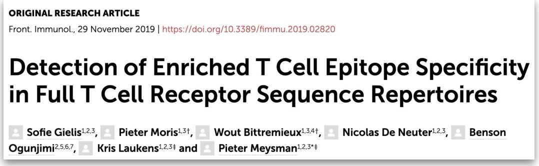


Research article | Open Access | Published: 28 May 2019

Capturing the differences between humoral immunity in the normal and tumor environments from repertoire-seq of B-cell receptors using supervised machine learning

Hiroki Konishi, Daisuke Komura, Hiroto Katoh, Shinichiro Atsumi, Hirotomo Koda, Asami Yamamoto, Yasuyuki Seto, Masashi Fukayama, Rui Yamaguchi, Seiya Imoto & Shumpei Ishikawa

BMC Bioinformatics **20**, Article number: 267 (2019) | Cite this article

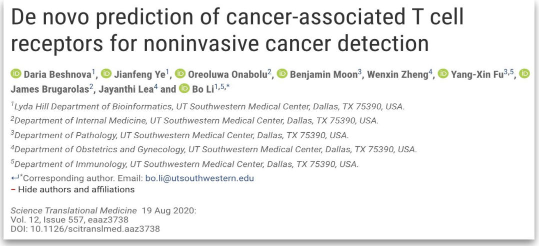


ORIGINAL RESEARCH ARTICLE

Front. Immunol., 29 November 2019 | https://doi.org/10.3389/fimmu.2019.02820

Detection of Enriched T Cell Epitope Specificity in Full T Cell Receptor Sequence Repertoires

Sofie Gielis[1,2,3], Pieter Moris[1,3†], Wout Bittremieux[1,3,4†], Nicolas De Neuter[1,2,3], Benson Ogunjimi[2,5,6,7], Kris Laukens[1,2,3‡] and Pieter Meysman[1,2,3*‡]



De novo prediction of cancer-associated T cell receptors for noninvasive cancer detection

Daria Beshnova[1], Jianfeng Ye[1], Oreoluwa Onabolu[2], Benjamin Moon[3], Wenxin Zheng[4], Yang-Xin Fu[3,5], James Brugarolas[2], Jayanthi Lea[4] and Bo Li[1,5,*]

[1] *Lyda Hill Department of Bioinformatics, UT Southwestern Medical Center, Dallas, TX 75390, USA.*

[2] *Department of Internal Medicine, UT Southwestern Medical Center, Dallas, TX 75390, USA.*

[3] *Department of Pathology, UT Southwestern Medical Center, Dallas, TX 75390, USA.*

[4] *Department of Obstetrics and Gynecology, UT Southwestern Medical Center, Dallas, TX 75390, USA.*

[5] *Department of Immunology, UT Southwestern Medical Center, Dallas, TX 75390, USA.*

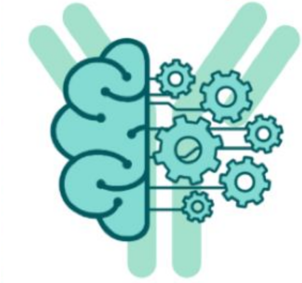
↵*Corresponding author. Email: bo.li@utsouthwestern.edu

– Hide authors and affiliations

Science Translational Medicine 19 Aug 2020:
Vol. 12, Issue 557, eaaz3738
DOI: 10.1126/scitranslmed.aaz3738

+ Many more!

ML application areas in AIRR analyses



Predicting receptor specificity

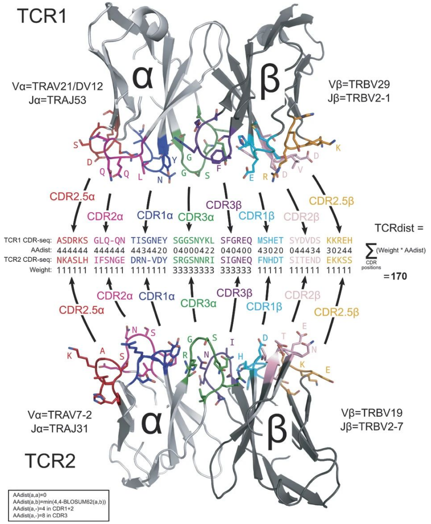
- ❏ Examining sequence similarity:

  Find the similarity between known positive and negative sequences, and predict the specificity to be the same as the sequences in the closest proximity

Predicting receptor specificity

❑ Examining sequence similarity:

Find the similarity between known positive and negative sequences, and predict the specificity to be the same as the sequences in the closest proximity



TCR1 CDR-seq:	ASDRKS	GLQ-QN	TISGNEY	SGGSNYKL	SFGREQ	MSHET	SYDVDS	KKREH
AAdist:	444444	444444	4434420	04000422	040400	43020	044434	30244
TCR2 CDR-seq:	NKASLH	IFSNGE	DRN-VDY	SRGSNNRI	SIGNEQ	FNHDT	SITEND	EKKSS
Weight:	111111	111111	1111111	33333333	333333	11111	111111	11111

TCRdist = $\sum_{\text{CDR positions}} (\text{Weight} * \text{AAdist})$ = **170**

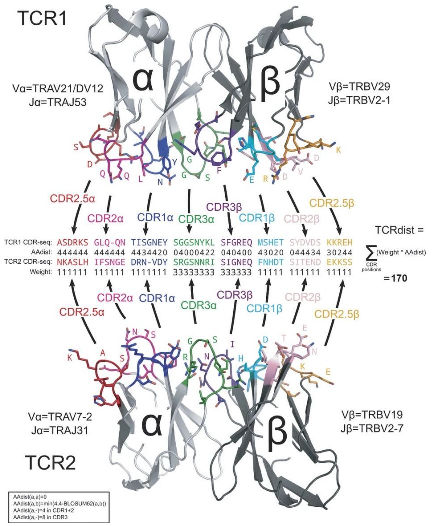
AAdist(a,a)=0
AAdist(a,b)=min(4,4-BLOSUM62(a,b))
AAdist(a,-)=4 in CDR1+2
AAdist(a,-)=8 in CDR3

TCRdist, Dash et al. 2017

Predicting receptor specificity

- Examining sequence similarity:

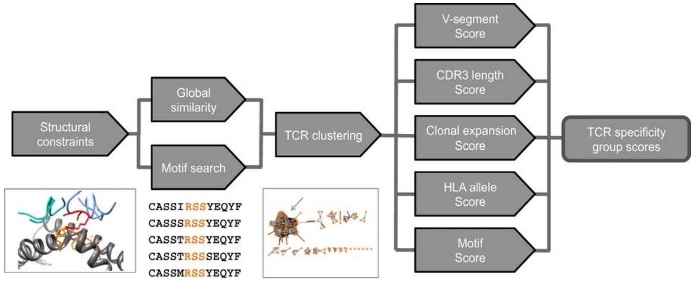
  Find the similarity between known positive and negative sequences, and predict the specificity to be the same as the sequences in the closest proximity



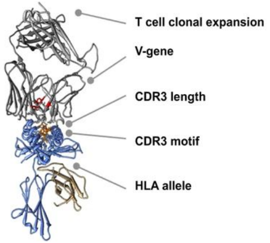
TCR1 CDR-seq:	ASDRKS	GLQ-QN	TISGNEY	SGGSNYKL	SFGREQ	MSHET	SYDVDS	KKREH
AAdist:	444444	444444	4434420	04000422	040400	43020	044434	30244
TCR2 CDR-seq:	NKASLH	IFSNGE	DRN-VDY	SRGSNNRI	SIGNEQ	FNHDT	SITEND	EKKSS
Weight:	111111	111111	1111111	33333333	333333	11111	111111	11111

AAdist(a,a)=0
AAdist(a,b)=min(4,4-BLOSUM62(a,b))
AAdist(a,-)=4 in CDR1+2
AAdist(a,-)=8 in CDR3

TCRdist, Dash et al. 2017



CASSIRSSYEQYF
CASSSRSSYEQYF
CASSTRSSYEQYF
CASSTRSSSEQYF
CASSMRSSYEQYF
CASSIRSSDTQYF
CASSIRSAYEQYF

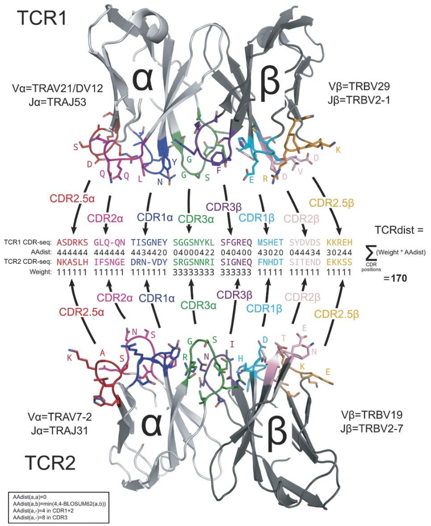


GLIPH, Glanville et al. 2017

Predicting receptor specificity

❑ Examining sequence similarity:

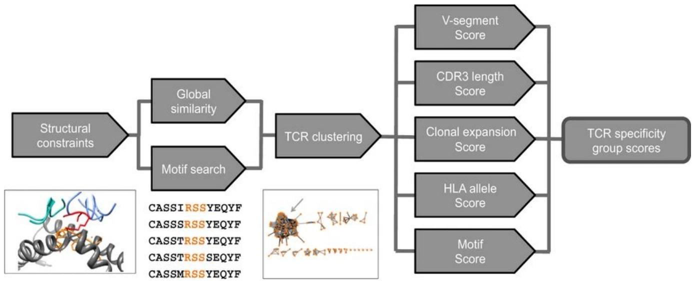
Find the similarity between known positive and negative sequences, and predict the specificity to be the same as the sequences in the closest proximity

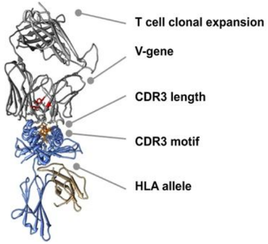


TCR1 CDR-seq:	ASDRKS	GLQ-QN	TISGNEY	SGGSNYKL	SFGREQ	MSHET	SYDVDS	KKREH
AAdist:	444444	444444	4434420	04000422	040400	43020	044434	30244
TCR2 CDR-seq:	NKASLH	IFSNGE	DRN-VDY	SRGSNNRI	SIGNEQ	FNHDT	SITEND	EKKSS
Weight:	111111	111111	1111111	33333333	333333	11111	111111	11111

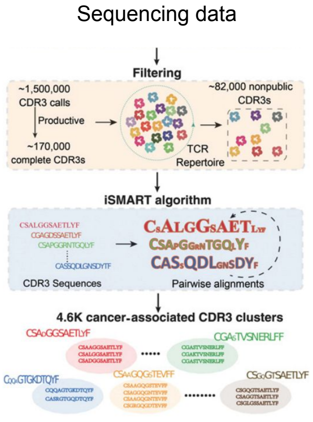
AAdist(a,a)=0
AAdist(a,b)=min(4,4-BLOSUM62(a,b))
AAdist(a,-)=4 in CDR1+2
AAdist(a,-)=8 in CDR3

TCRdist, Dash et al. 2017





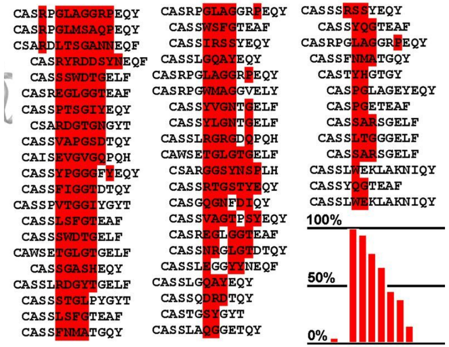
GLIPH, Glanville et al. 2017



iSMART, Zhang et al. 2020

Predicting receptor specificity

- ❏ Discovering short motifs in the sequence that are indicative of its specificity
- ❏ Predictions made based on physicochemical properties of receptors



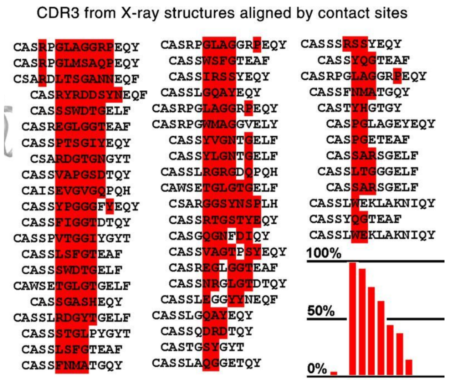
CDR3 from X-ray structures aligned by contact sites

Ostmeyer et al. 2019

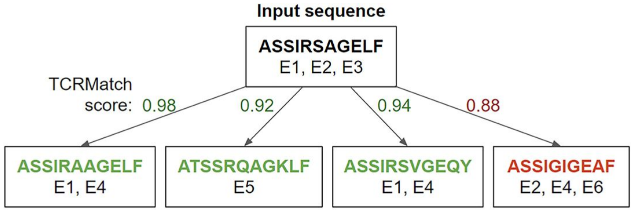
Predicting receptor specificity

❑ Discovering short motifs in the sequence that are indicative of its specificity

❑ Predictions made based on physicochemical properties of receptors



Ostmeyer et al. 2019

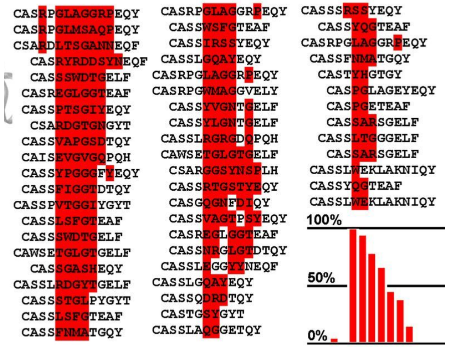


Chronister et al. 2021

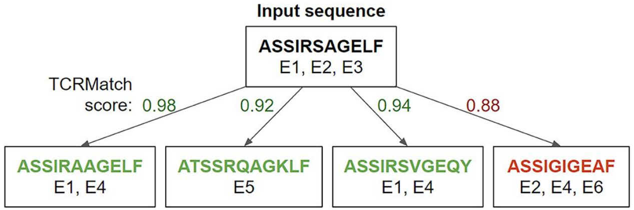
Predicting receptor specificity

- ❏ Discovering short motifs in the sequence that are indicative of its specificity
- ❏ Predictions made based on physicochemical properties of receptors

CDR3 from X-ray structures aligned by contact sites



Ostmeyer et al. 2019


Input sequence

ASSIRSAGELF
E1, E2, E3

TCRMatch score: 0.98 0.92 0.94 0.88

ASSIRAAGELF
E1, E4

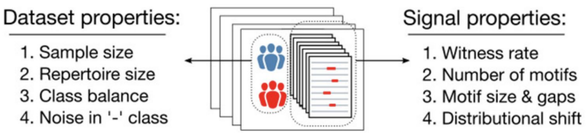
ATSSRQAGKLF
E5

ASSIRSVGEQY
E1, E4

ASSIGIGEAF
E2, E4, E6

Chronister et al. 2021

AIRR datasets with varying properties


Dataset properties:

1. Sample size
2. Repertoire size
3. Class balance
4. Noise in '-' class

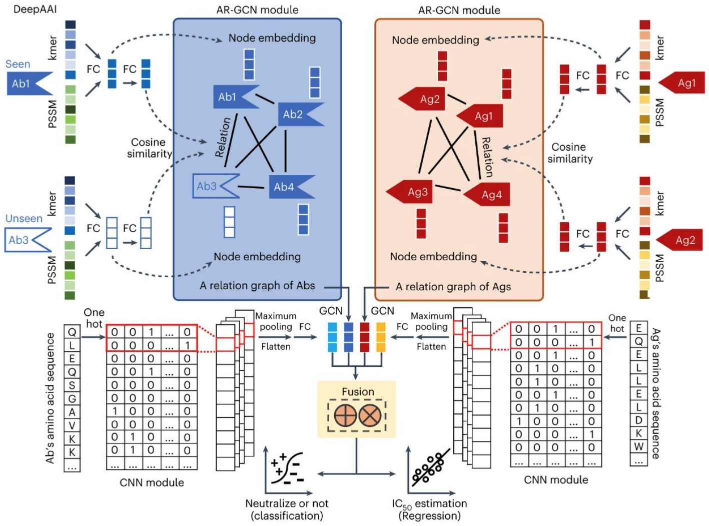
Signal properties:

1. Witness rate
2. Number of motifs
3. Motif size & gaps
4. Distributional shift

Kanduri et al. 2022

Predicting receptor specificity

- ❑ Modeling antibody-antigen interactions

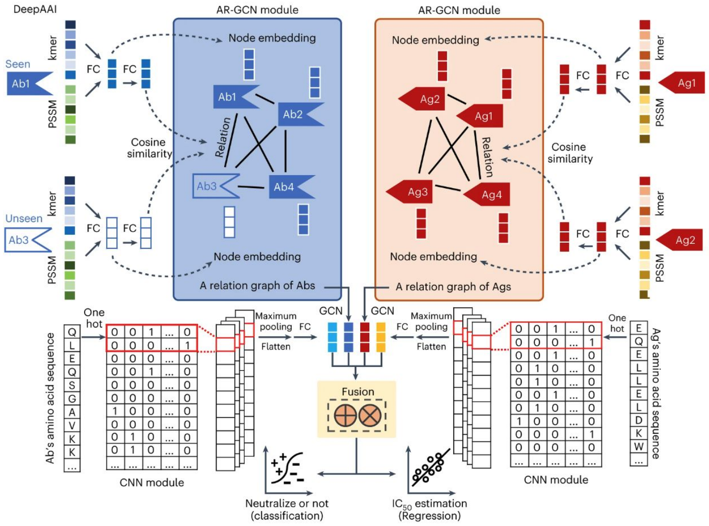


Zhang et al. 2022

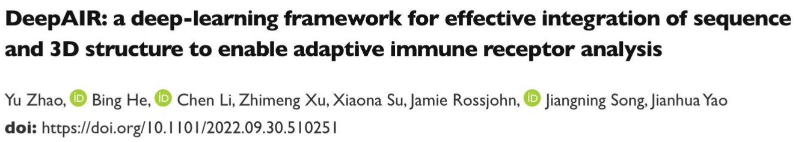
Predicting receptor specificity

❑ Modeling antibody-antigen interactions

❑ Using structural information



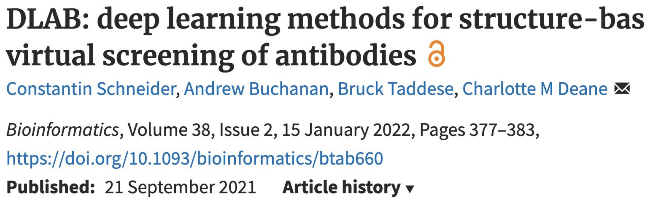
Zhang et al. 2022



DeepAIR: a deep-learning framework for effective integration of sequence and 3D structure to enable adaptive immune receptor analysis

Yu Zhao, Bing He, Chen Li, Zhimeng Xu, Xiaona Su, Jamie Rossjohn, Jiangning Song, Jianhua Yao

doi: https://doi.org/10.1101/2022.09.30.510251



DLAB: deep learning methods for structure-based virtual screening of antibodies

Constantin Schneider, Andrew Buchanan, Bruck Taddese, Charlotte M Deane

Bioinformatics, Volume 38, Issue 2, 15 January 2022, Pages 377–383, https://doi.org/10.1093/bioinformatics/btab660

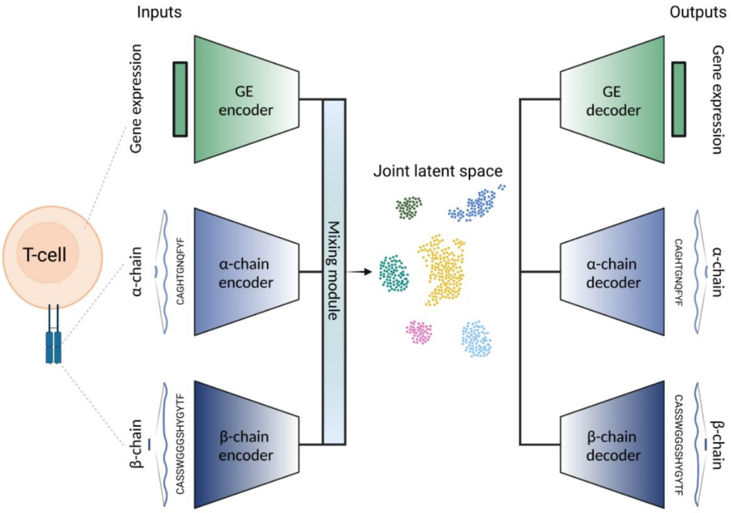
Published: 21 September 2021 **Article history ▾**

Analysis of AIR data

❏ Learning a latent representation using sequence and gene expression data

Analysis of AIR data

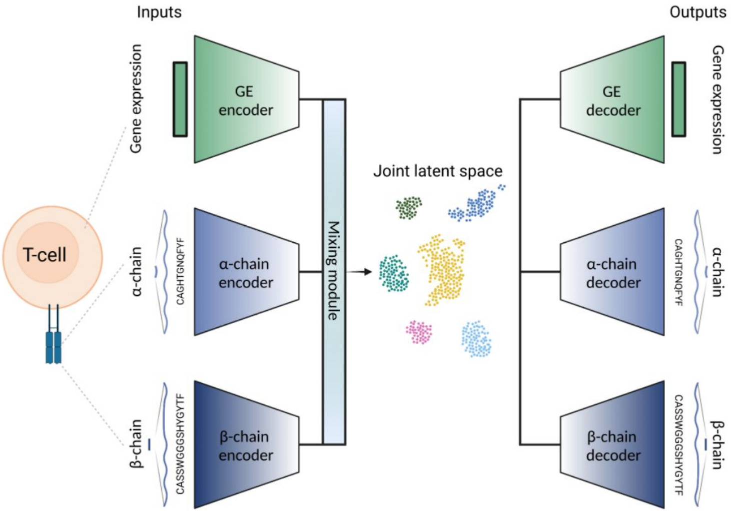
- Learning a latent representation using sequence and gene expression data



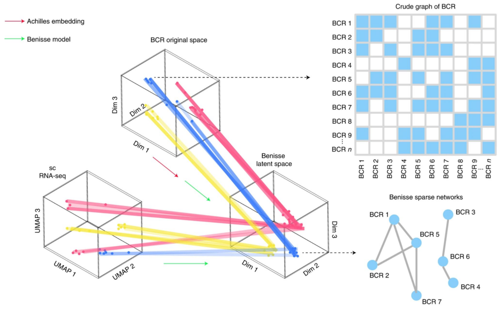
mvTCR, Drost et al. 2022

Analysis of AIR data

- ❑ Learning a latent representation using sequence and gene expression data



mvTCR, Drost et al. 2022

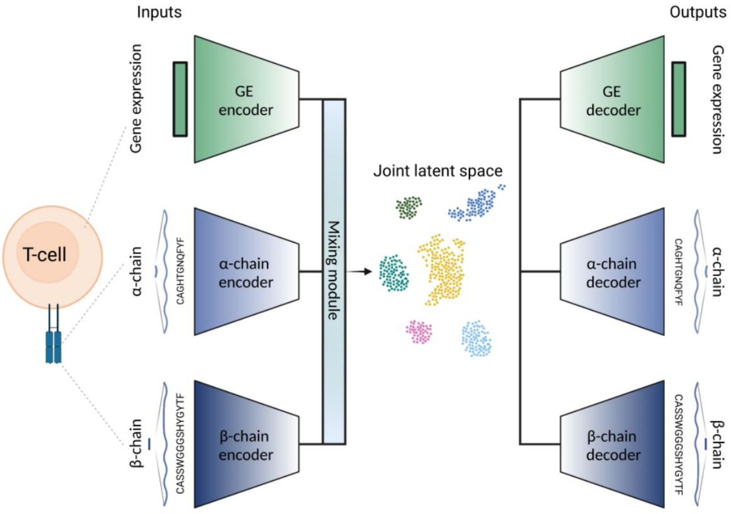


	BCR 1	BCR 2	BCR 3	BCR 4	BCR 5	BCR 6	BCR 7	BCR 8	BCR 9	BCR *n*
BCR 1	🟦	🟦	🟦			🟦	🟦		🟦	
BCR 2	🟦	🟦			🟦	🟦				
BCR 3	🟦		🟦		🟦	🟦	🟦			
BCR 4				🟦					🟦	🟦
BCR 5		🟦	🟦		🟦		🟦		🟦	🟦
BCR 6	🟦	🟦	🟦			🟦	🟦			🟦
BCR 7	🟦		🟦		🟦	🟦	🟦		🟦	🟦
BCR 8								🟦	🟦	🟦
BCR 9	🟦			🟦	🟦		🟦	🟦	🟦	
BCR *n*				🟦	🟦	🟦	🟦	🟦		🟦

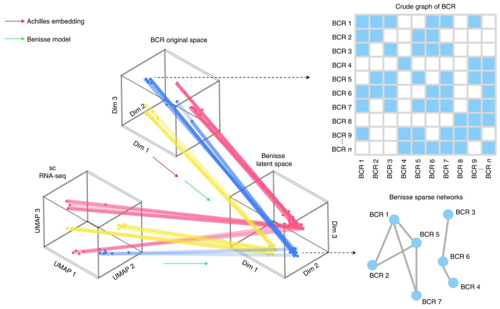
Benisse, Zhang et al. 2022

Analysis of AIR data

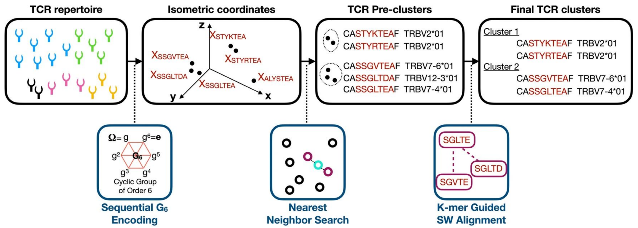
- ❑ Learning a latent representation using sequence and gene expression data



mvTCR, Drost et al. 2022



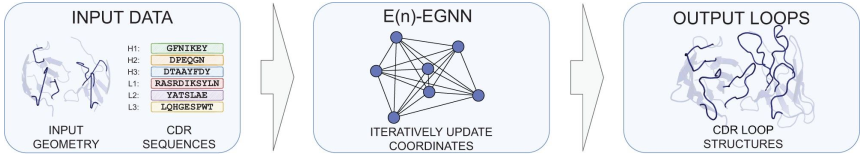
Benisse, Zhang et al. 2022



GIANA, Zhang et al. 2021

Predicting the 3D structure of AIRs

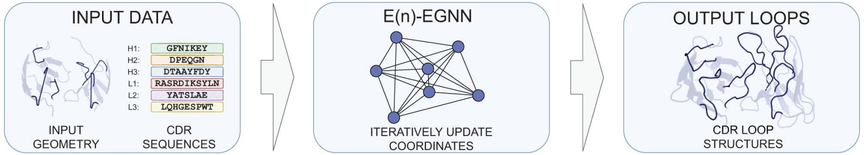
❑ Antibody-specific methods achieve better prediction performance than generic protein structure prediction tools



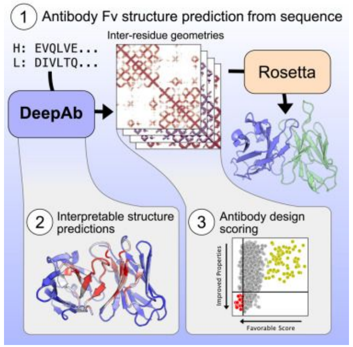
ABlooper, Abanades et al. 2022

Predicting the 3D structure of AIRs

- Antibody-specific methods achieve better prediction performance than generic protein structure prediction tools



ABlooper, Abanades et al. 2022

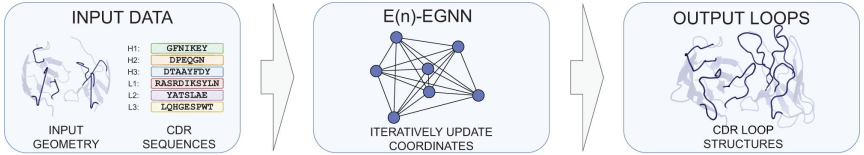


1 Antibody Fv structure prediction from sequence
Inter-residue geometries
H: EVQLVE...
L: DIVLTQ...
DeepAb
Rosetta
2 Interpretable structure predictions
3 Antibody design scoring
Improved Properties
Favorable Score

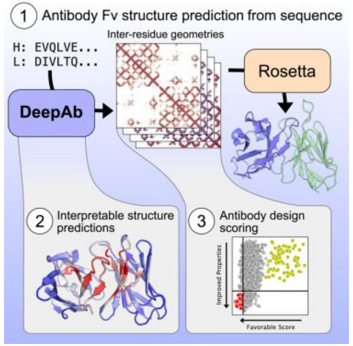
DeepAb, Ruffolo et al. 2022

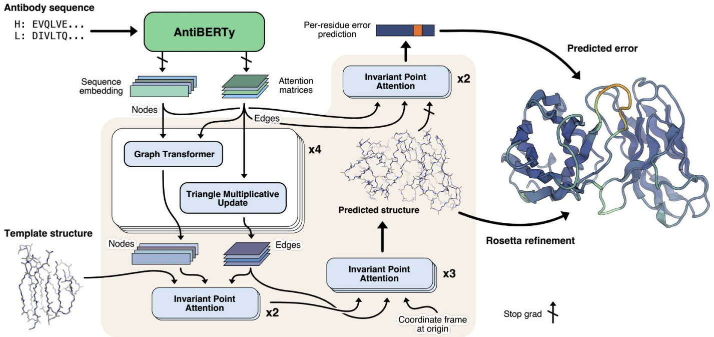
Predicting the 3D structure of AIRs

- Antibody-specific methods achieve better prediction performance than generic protein structure prediction tools



ABlooper, Abanades et al. 2022


1 Antibody Fv structure prediction from sequence
Inter-residue geometries
H: EVQLVE...
L: DIVLTQ...
DeepAb
Rosetta
2 Interpretable structure predictions
3 Antibody design scoring
Improved Properties
Favorable Score

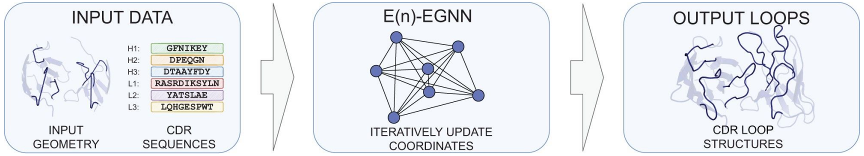


DeepAb, Ruffolo et al. 2022

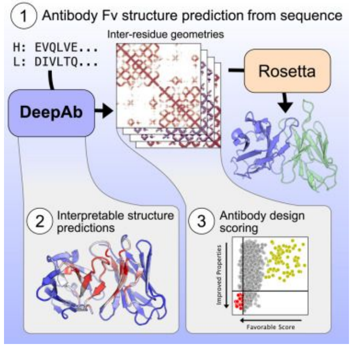
IgFold, Ruffolo et al. 2022

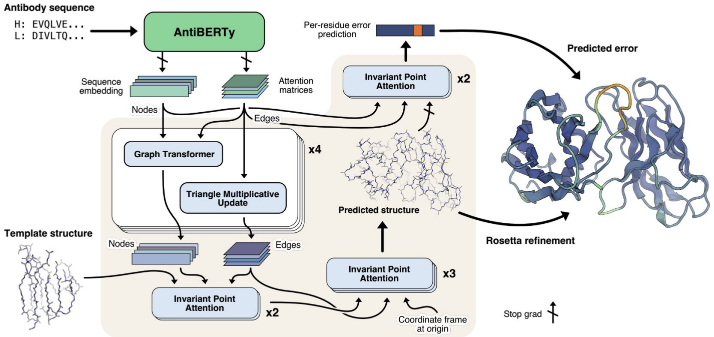
Predicting the 3D structure of AIRs

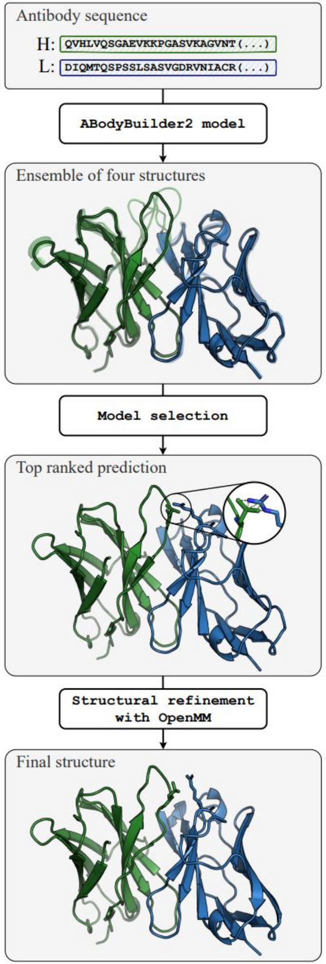
- ❑ Antibody-specific methods achieve better prediction performance than generic protein structure prediction tools



ABlooper, Abanades et al. 2022


1 Antibody Fv structure prediction from sequence
Inter-residue geometries
H: EVQLVE...
L: DIVLTQ...
DeepAb
Rosetta
2 Interpretable structure predictions
3 Antibody design scoring
Improved Properties
Favorable Score




Antibody sequence
H: QVHLVQSGAEVKKPGASVKAGVNT(...)
L: DIQMTQSPSSLSASVGDRVNIACR(...)
ABodyBuilder2 model
Ensemble of four structures
Model selection
Top ranked prediction
Structural refinement with OpenMM
Final structure

DeepAb, Ruffolo et al. 2022

IgFold, Ruffolo et al. 2022

ImmuneBuilder, Abanades et al. 2022

Language models for antibody sequences

- Some of the previous models are based on language models

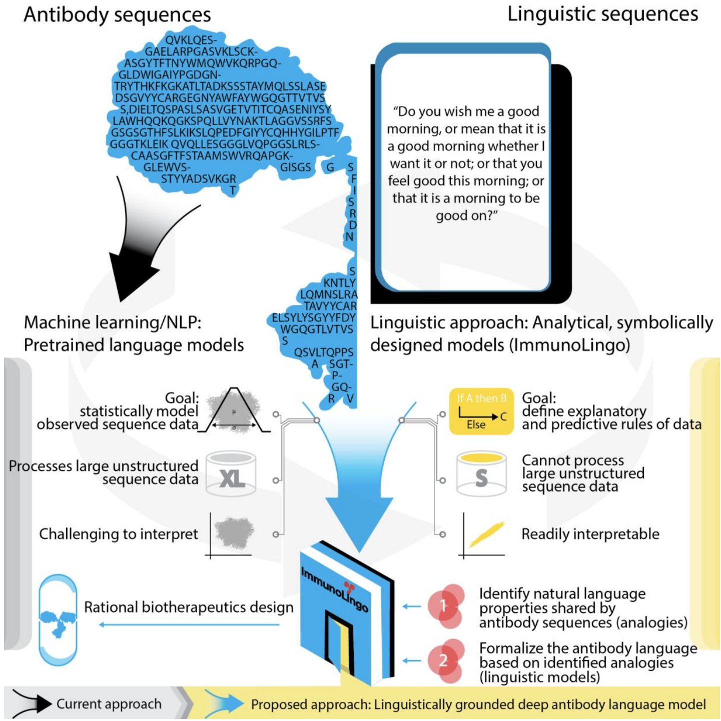
Language models for antibody sequences

❏ Some of the previous models are based on language models

❏ Necessary to formalize the “antibody language”

Language models for antibody sequences

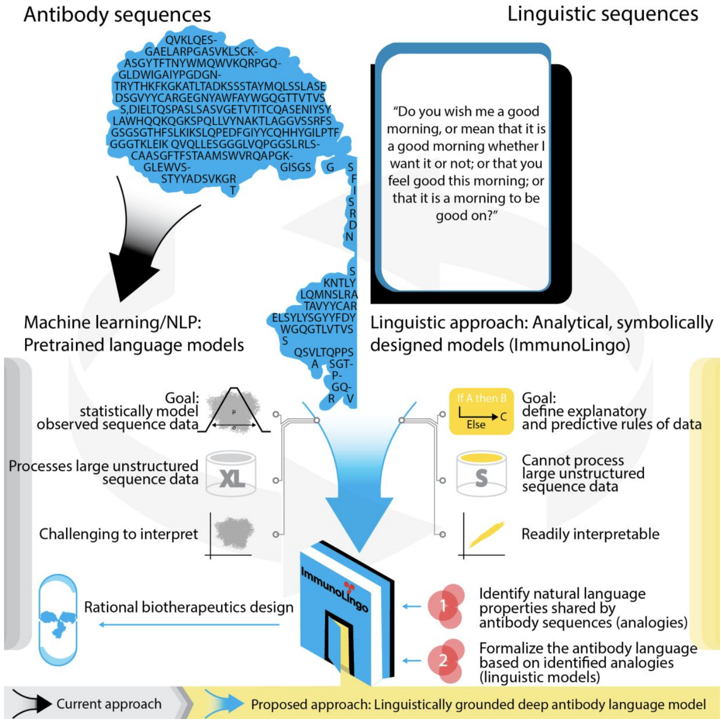
- ❑ Some of the previous models are based on language models
- ❑ Necessary to formalize the “antibody language”


Antibody sequences
Linguistic sequences
QVKLQES-
GAELARPGASVKLSCK-
ASGYTFTNYWMQWVKQRPGQ-
GLDWIGAIYPGDGN-
TRYTHKFKGKATLTADKSSSTAYMQLSSLASE
DSGVYYCARGEGNYAWFAYWGQGTTVTVS
S,DIELTQSPASLSASVGETVTITCQASENIYSY
LAWHQQKQGKSPQLLVYNAKTLAGGVSSRFS
GSGSGTHFSLKIKSLQPEDFGIYYCQHHYGILPTF
GGGTKLEIK QVQLLESGGGLVQPGGSLRLS-
CAASGFTFSTAAMSWVRQAPGK-
GLEWVS-
GISGS G S
STYYADSVKGR
T
F I S R D N
S
KNTLY
LQMNSLRA
TAVYYCAR
ELSYLYSGYYFDY
WGQGTLVTVS
S
QSVLTQPPS
A SGT-
P-
GQ-
R V
“Do you wish me a good morning, or mean that it is a good morning whether I want it or not; or that you feel good this morning; or that it is a morning to be good on?”
Machine learning/NLP:
Pretrained language models
Linguistic approach: Analytical, symbolically designed models (ImmunoLingo)
Goal:
statistically model
observed sequence data
μ
σ
Processes large unstructured
sequence data
XL
Challenging to interpret
If A then B
C
Else
Goal:
define explanatory
and predictive rules of data
S
Cannot process
large unstructured
sequence data
Readily interpretable
ImmunoLingo
Rational biotherapeutics design
1
Identify natural language properties shared by antibody sequences (analogies)
2
Formalize the antibody language based on identified analogies (linguistic models)
Current approach
Proposed approach: Linguistically grounded deep antibody language model

ImmunoLingo, Vu et al. 2022

Language models for antibody sequences

- ❑ Some of the previous models are based on language models
- ❑ Necessary to formalize the “antibody language”
- ❑ Improved interpretability through formalization

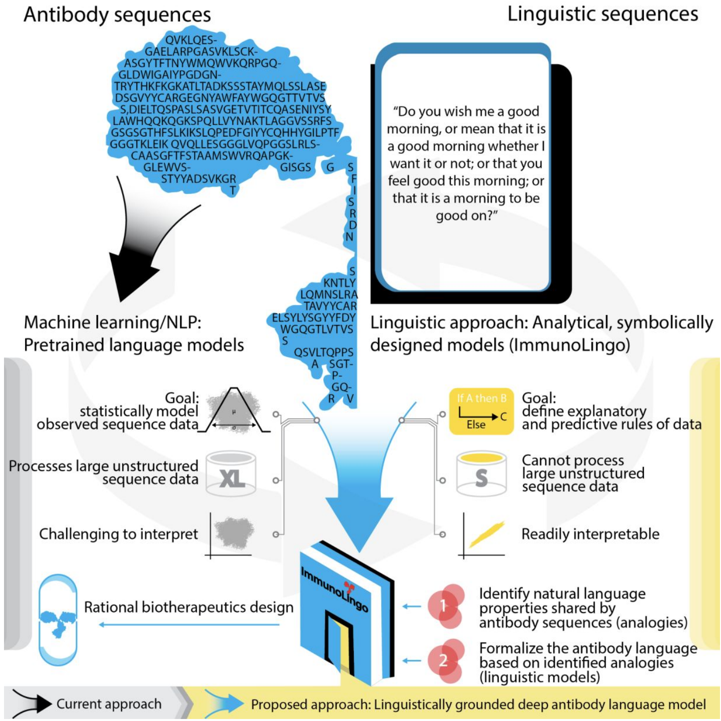


Antibody sequences
Linguistic sequences
QVKLQES-
GAELARPGASVKLSCK-
ASGYTFTNYWMQWVKQRPGQ-
GLDWIGAIYPGDGN-
TRYTHKFKGKATLTADKSSSTAYMQLSSLASE
DSGVYYCARGEGNYAWFAYWGQGTTVTVS
S,DIELTQSPASLSASVGETVTITCQASENIYSY
LAWHQQKQGKSPQLLVYNAKTLAGGVSSRFS
GSGSGTHFSLKIKSLQPEDFGIYYCQHHYGILPTF
GGGTKLEIK QVQLLESGGGLVQPGGSLRLS-
CAASGFTFSTAAMSWVRQAPGK-
GLEWVS-
STYYADSVKGR
T
GISGS
G
SFISRDN
S
KNTLY
LQMNSLRA
TAVYYCAR
ELSYLYSGYYFDY
WGQGTLVTVS
S
QSVLTQPPS
A
SGT-
P-
GQ-
R
V
“Do you wish me a good morning, or mean that it is a good morning whether I want it or not; or that you feel good this morning; or that it is a morning to be good on?”
Machine learning/NLP: Pretrained language models
Linguistic approach: Analytical, symbolically designed models (ImmunoLingo)
Goal: statistically model observed sequence data
μ
σ
Processes large unstructured sequence data
XL
Challenging to interpret
If A then B
C
Else
Goal: define explanatory and predictive rules of data
S
Cannot process large unstructured sequence data
Readily interpretable
ImmunoLingo
Rational biotherapeutics design
1
Identify natural language properties shared by antibody sequences (analogies)
2
Formalize the antibody language based on identified analogies (linguistic models)
Current approach
Proposed approach: Linguistically grounded deep antibody language model

ImmunoLingo, Vu et al. 2022

Language models for antibody sequences

- ❏ Some of the previous models are based on language models
- ❏ Necessary to formalize the “antibody language”
- ❏ Improved interpretability through formalization
- ❏ Potential aim: therapeutics design

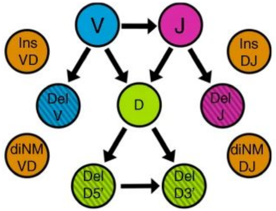

Antibody sequences
Linguistic sequences
QVKLQES-
GAELARPGASVKLSCK-
ASGYTFTNYWMQWVKQRPGQ-
GLDWIGAIYPGDGN-
TRYTHKFKGKATLTADKSSSTAYMQLSSLASE
DSGVYYCARGEGNYAWFAYWGQGTTVTVS
S,DIELTQSPASLSASVGETVTITCQASENIYSY
LAWHQQKQGKSPQLLVYNAKTLAGGVSSRFS
GSGSGTHFSLKIKSLQPEDFGIYYCQHHYGILPTF
GGGTKLEIK QVQLLESGGGLVQPGGSLRLS-
CAASGFTFSTAAMSWVRQAPGK-
GLEWVS-
STYYADSVKGR
T
GISGS G S F I S R D N
S
KNTLY
LQMNSLRA
TAVYYCAR
ELSYLYSGYYFDY
WGQGTLVTVS
S
QSVLTQPPS
A SGT-
P-
GQ-
R V
“Do you wish me a good morning, or mean that it is a good morning whether I want it or not; or that you feel good this morning; or that it is a morning to be good on?”
Machine learning/NLP:
Pretrained language models
Linguistic approach: Analytical, symbolically designed models (ImmunoLingo)
Goal:
statistically model
observed sequence data
μ
σ
If A then B
C
Else
Goal:
define explanatory
and predictive rules of data
Processes large unstructured
sequence data
XL
S
Cannot process
large unstructured
sequence data
Challenging to interpret
Readily interpretable
ImmunoLingo
Rational biotherapeutics design
1
Identify natural language
properties shared by
antibody sequences (analogies)
2
Formalize the antibody language
based on identified analogies
(linguistic models)
Current approach
Proposed approach: Linguistically grounded deep antibody language model

ImmunoLingo, Vu et al. 2022

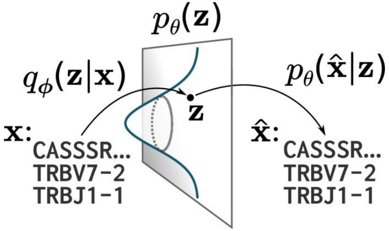
Generative models for AIRs

Generative models for AIRs

❑ Modeling the VDJ recombination process (naive AIRs)



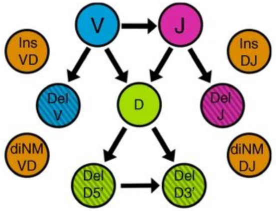
IGoR, Marcou et al. 2018, doi:10.1038/s41467-018-02832-w


$p_\theta(\mathbf{z})$
$q_\phi(\mathbf{z}|\mathbf{x})$
$p_\theta(\hat{\mathbf{x}}|\mathbf{z})$
$\mathbf{z}$
$\mathbf{x}$: CASSSR... TRBV7-2 TRBJ1-1
$\hat{\mathbf{x}}$: CASSSR... TRBV7-2 TRBJ1-1

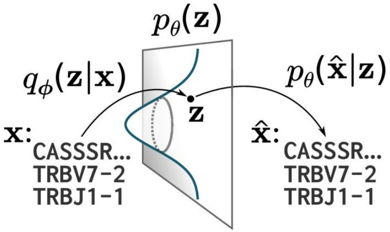
Davidsen et al. 2019, doi: 10.7554/eLife.46935

Generative models for AIRs

- ❑ Modeling the VDJ recombination process (naive AIRs)

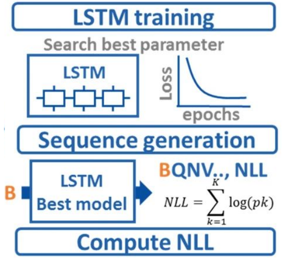


IGoR, Marcou et al. 2018, doi:10.1038/s41467-018-02832-w

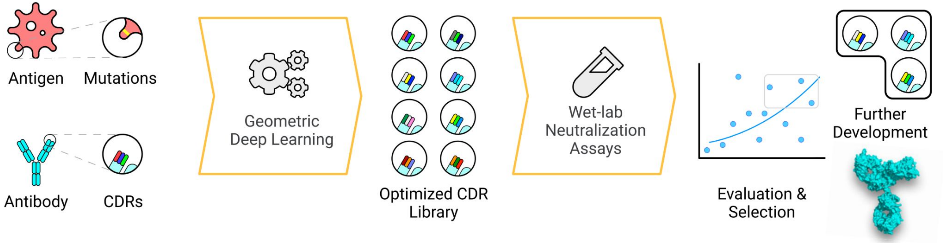

$p_\theta(\mathbf{z})$
$q_\phi(\mathbf{z}|\mathbf{x})$
$p_\theta(\hat{\mathbf{x}}|\mathbf{z})$
z
x: CASSSR... TRBV7-2 TRBJ1-1
x̂: CASSSR... TRBV7-2 TRBJ1-1

Davidsen et al. 2019, doi: 10.7554/eLife.46935

- ❑ Modeling antigen–specific antibodies directly



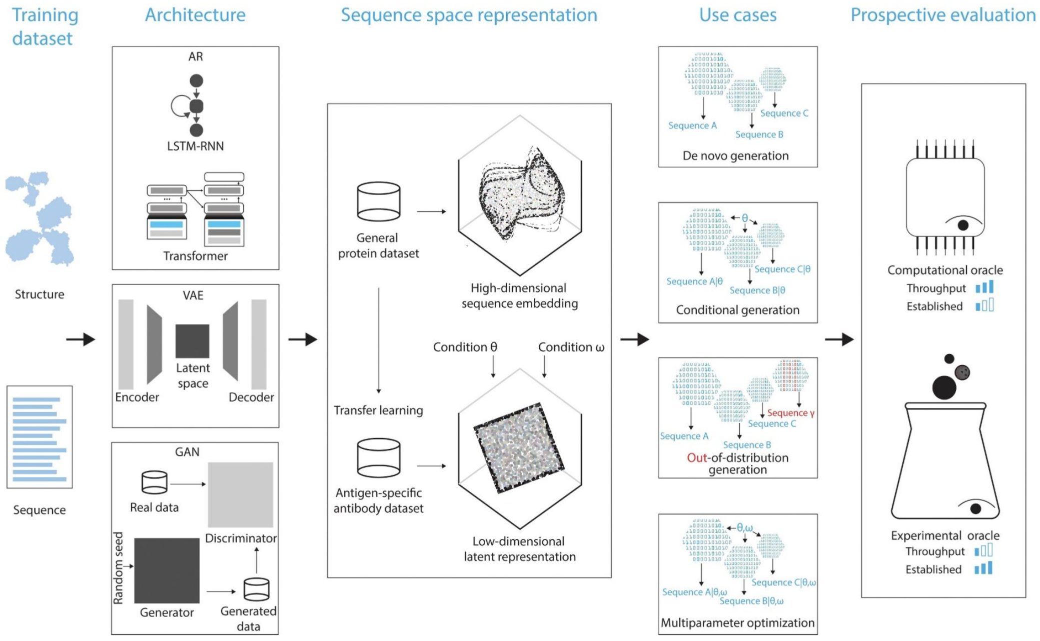
Saka et al. 2021, doi: 10.1038/s41598-021-85274-7



Shan et al. 2022, doi: 10.1073/pnas.2122954119

Antibody design with machine learning

- ❑ Epitope specificity, affinity and developability
- ❑ Public repositories: iReceptor, IEDB, AbDb, AgAbDb
- ❑ Synthetic data: Absolut!



Review: Akbar et al. 2022, doi:10.1080/19420862.2021.2008790

TCRs and peptide-MHC complexes

❏ For TCRs to recognize a peptide, it has to be presented by the MHC complex

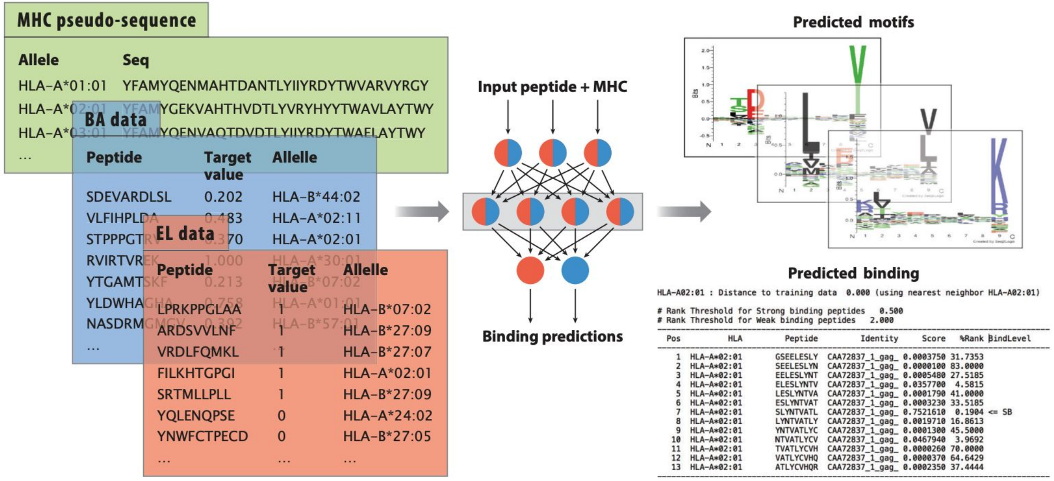
TCRs and peptide-MHC complexes

❏ For TCRs to recognize a peptide, it has to be presented by the MHC complex

❏ Tasks: predicting peptide-MHC binding, predicting the binding of a TCR to pMHC complex

TCRs and peptide-MHC complexes

- ❑ For TCRs to recognize a peptide, it has to be presented by the MHC complex
- ❑ Tasks: predicting peptide-MHC binding, predicting the binding of a TCR to pMHC complex



MHC pseudo-sequence

Allele	Seq
HLA–A*01:01	YFAMYQENMAHTDANTLYIIYRDYTWVARVYRGY
HLA–A*02:01	YFAMYGEKVAHTHVDTLYVRYHYYTWAVLAYTWY
HLA–A*03:01	YFAMYQENVAQTDVDTLYIIYRDYTWAELAYTWY
…	

BA data

Peptide	Target value	Allelle
SDEVARDLSL	0.202	HLA–B*44:02
VLFIHPLDA	0.483	HLA–A*02:11
STPPPGTRV	0.370	HLA–A*02:01
RVIRTVREK	1.000	HLA–A*30:01
YTGAMTSKF	0.213	HLA–B*07:02
YLDWHAGHA	0.758	HLA–A*01:01
NASDRMGMCV	0.392	HLA–B*57:01
…		

EL data

Peptide	Target value	Allelle
LPRKPPGLAA	1	HLA–B*07:02
ARDSVVLNF	1	HLA–B*27:09
VRDLFQMKL	1	HLA–B*27:07
FILKHTGPGI	1	HLA–A*02:01
SRTMLLPLL	1	HLA–B*27:09
YQLENQPSE	0	HLA–A*24:02
YNWFCTPECD	0	HLA–B*27:05
…	…	…

Predicted binding

```
HLA-A02:01 : Distance to training data  0.000 (using nearest neighbor HLA-A02:01)

# Rank Threshold for Strong binding peptides   0.500
# Rank Threshold for Weak binding peptides   2.000
-----------------------------------------------------------------------------------
  Pos          HLA         Peptide       Identity      Score    %Rank BindLevel
-----------------------------------------------------------------------------------
    1  HLA-A*02:01       GSEELESLY  CAA72837_1_gag_ 0.0003750 31.7353
    2  HLA-A*02:01       SEELESLYN  CAA72837_1_gag_ 0.0000100 83.0000
    3  HLA-A*02:01       EELESLYNT  CAA72837_1_gag_ 0.0005480 27.5185
    4  HLA-A*02:01       ELESLYNTV  CAA72837_1_gag_ 0.0357700  4.5815
    5  HLA-A*02:01       LESLYNTVA  CAA72837_1_gag_ 0.0001790 41.0000
    6  HLA-A*02:01       ESLYNTVAT  CAA72837_1_gag_ 0.0003230 33.5185
    7  HLA-A*02:01       SLYNTVATL  CAA72837_1_gag_ 0.7521610  0.1904 <= SB
    8  HLA-A*02:01       LYNTVATLY  CAA72837_1_gag_ 0.0019710 16.8613
    9  HLA-A*02:01       YNTVATLYC  CAA72837_1_gag_ 0.0001300 45.5000
   10  HLA-A*02:01       NTVATLYCV  CAA72837_1_gag_ 0.0467940  3.9692
   11  HLA-A*02:01       TVATLYCVH  CAA72837_1_gag_ 0.0000260 70.0000
   12  HLA-A*02:01       VATLYCVHQ  CAA72837_1_gag_ 0.0000370 64.6429
   13  HLA-A*02:01       ATLYCVHQR  CAA72837_1_gag_ 0.0002350 37.4444
-----------------------------------------------------------------------------------
```

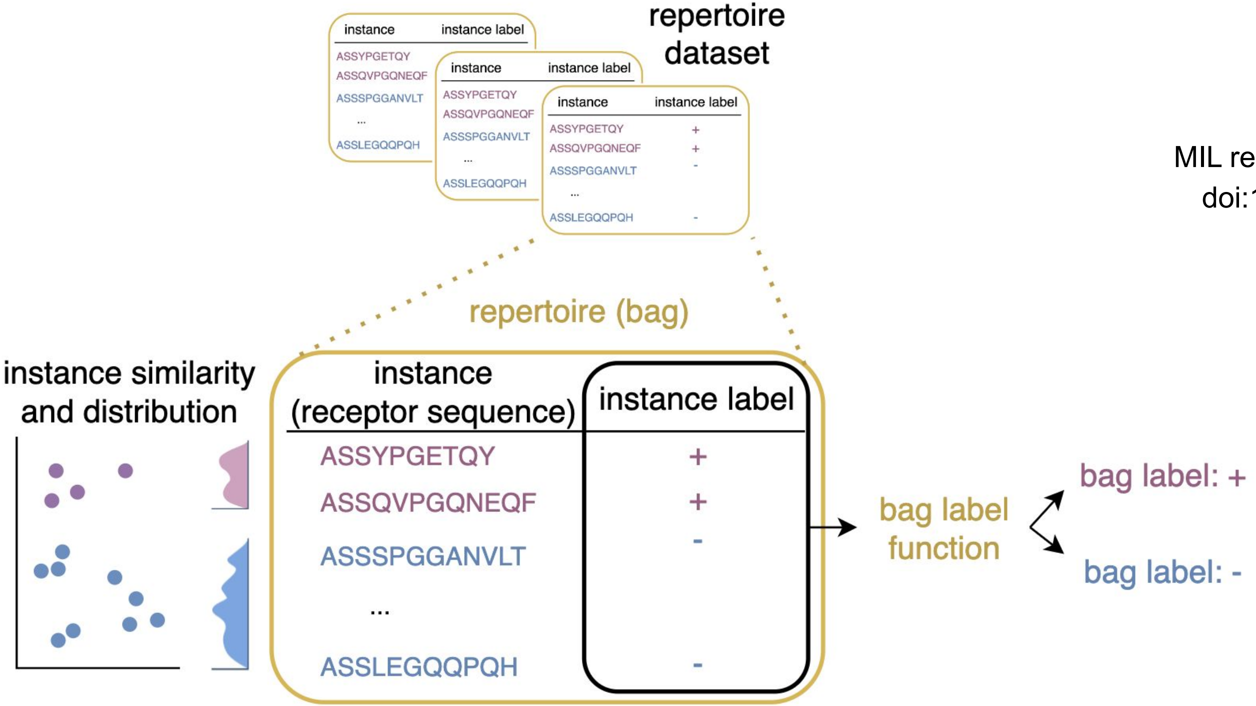
NNAlign in the review by Nielsen et al. 2020,
doi:10.1146/annurev-biodatasci-021920-100259

Diagnosing immune-related diseases with AIRRs

❏ Repertoire classification is a multiple instance learning (MIL) problem

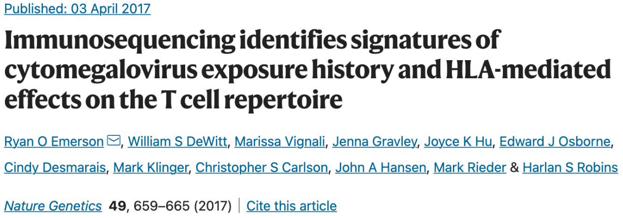
Diagnosing immune-related diseases with AIRRs

- Repertoire classification is a multiple instance learning (MIL) problem



MIL review: Carbonneau et al. 2018
doi:10.1016/j.patcog.2017.10.009

Custom ML approaches for diagnostics

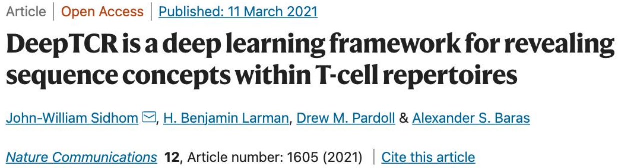


Published: 03 April 2017

Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire

Ryan O Emerson, William S DeWitt, Marissa Vignali, Jenna Gravley, Joyce K Hu, Edward J Osborne, Cindy Desmarais, Mark Klinger, Christopher S Carlson, John A Hansen, Mark Rieder & Harlan S Robins

Nature Genetics **49**, 659–665 (2017) | Cite this article

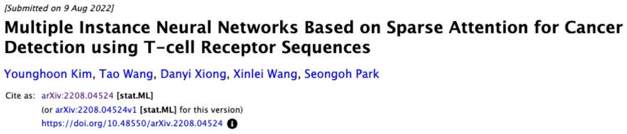


Article | Open Access | Published: 11 March 2021

DeepTCR is a deep learning framework for revealing sequence concepts within T-cell repertoires

John-William Sidhom, H. Benjamin Larman, Drew M. Pardoll & Alexander S. Baras

Nature Communications **12**, Article number: 1605 (2021) | Cite this article

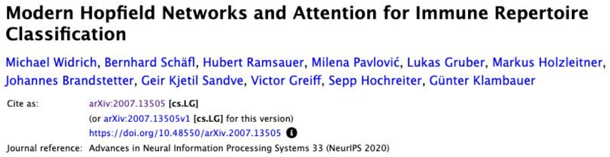


[Submitted on 9 Aug 2022]

Multiple Instance Neural Networks Based on Sparse Attention for Cancer Detection using T-cell Receptor Sequences

Younghoon Kim, Tao Wang, Danyi Xiong, Xinlei Wang, Seongoh Park

Cite as: arXiv:2208.04524 [stat.ML]
(or arXiv:2208.04524v1 [stat.ML] for this version)
https://doi.org/10.48550/arXiv.2208.04524

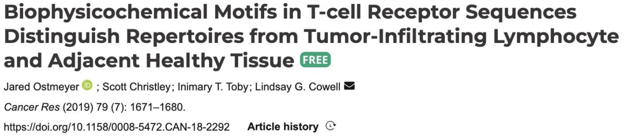


Modern Hopfield Networks and Attention for Immune Repertoire Classification

Michael Widrich, Bernhard Schäfl, Hubert Ramsauer, Milena Pavlović, Lukas Gruber, Markus Holzleitner, Johannes Brandstetter, Geir Kjetil Sandve, Victor Greiff, Sepp Hochreiter, Günter Klambauer

Cite as: arXiv:2007.13505 [cs.LG]
(or arXiv:2007.13505v1 [cs.LG] for this version)
https://doi.org/10.48550/arXiv.2007.13505

Journal reference: Advances in Neural Information Processing Systems 33 (NeurIPS 2020)

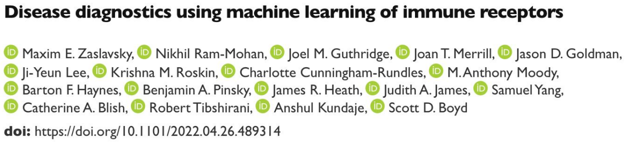


Biophysicochemical Motifs in T-cell Receptor Sequences Distinguish Repertoires from Tumor-Infiltrating Lymphocyte and Adjacent Healthy Tissue FREE

Jared Ostmeyer ; Scott Christley; Inimary T. Toby; Lindsay G. Cowell

Cancer Res (2019) 79 (7): 1671–1680.

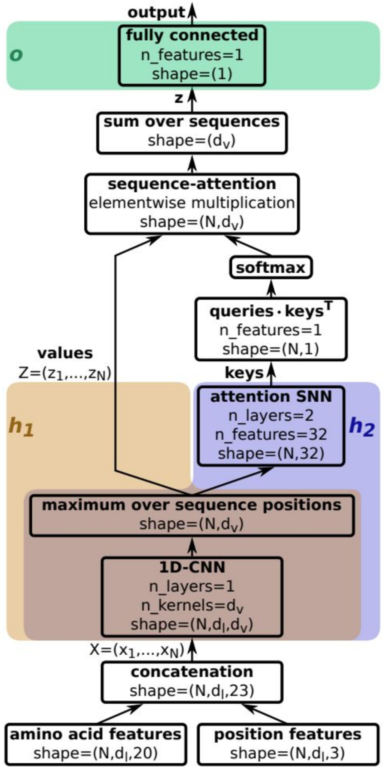
https://doi.org/10.1158/0008-5472.CAN-18-2292 **Article history**



Disease diagnostics using machine learning of immune receptors

Maxim E. Zaslavsky, Nikhil Ram-Mohan, Joel M. Guthridge, Joan T. Merrill, Jason D. Goldman, Ji-Yeun Lee, Krishna M. Roskin, Charlotte Cunningham-Rundles, M. Anthony Moody, Barton F. Haynes, Benjamin A. Pinsky, James R. Heath, Judith A. James, Samuel Yang, Catherine A. Blish, Robert Tibshirani, Anshul Kundaje, Scott D. Boyd

doi: https://doi.org/10.1101/2022.04.26.489314



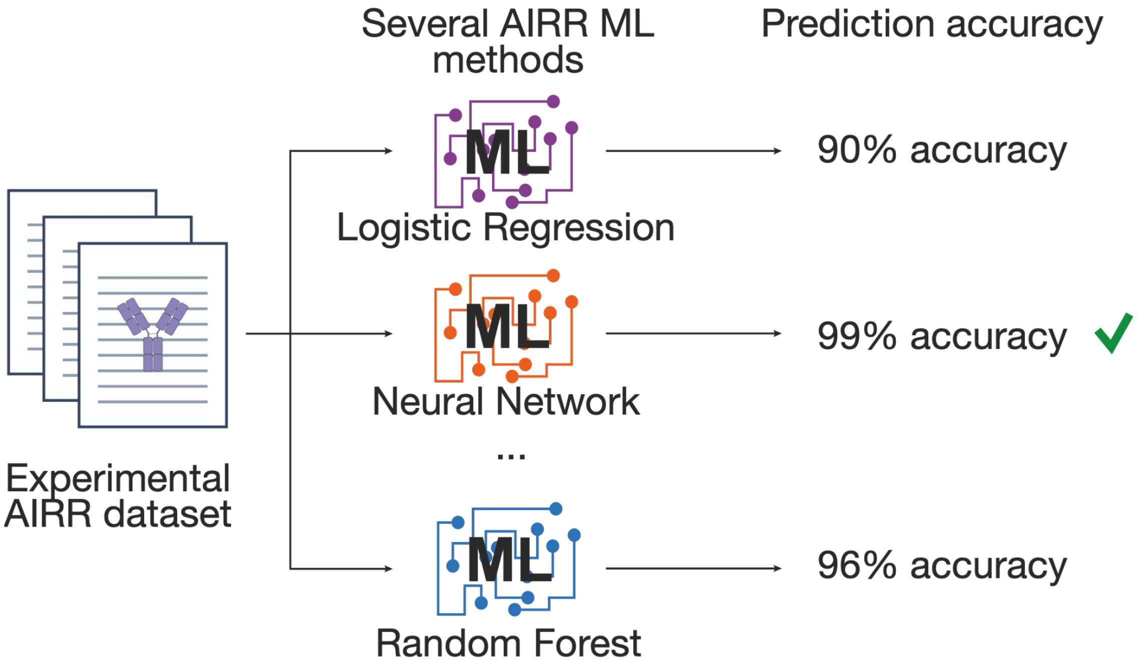
DeepRC, Widrich et al. 2020

Different ML methods have different underlying assumptions – those should be conscious choices to reflect the problem domain

How do we ensure that the method can be applied to unseen receptors or repertoires?

(generalizability of ML methods)

A naive way to perform AIRR ML

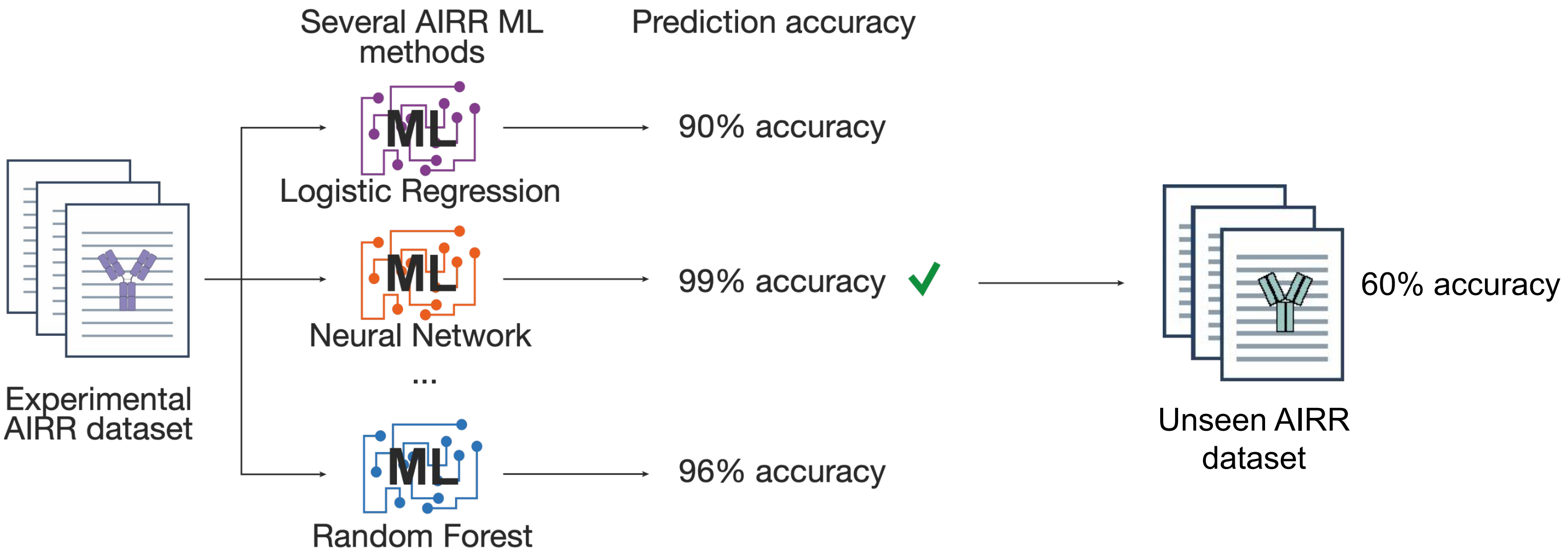


Experimental data challenges in method development

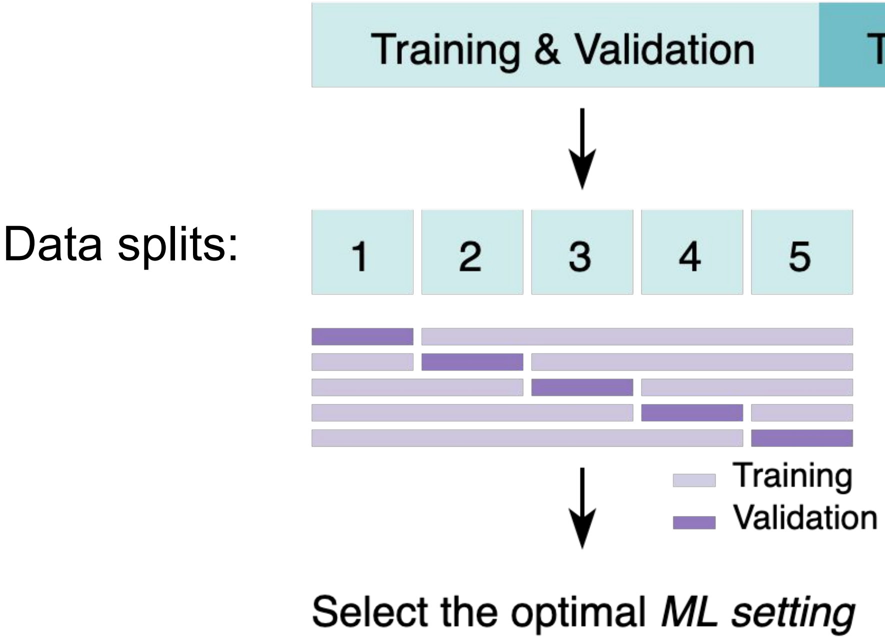
- Usually small dataset size
  - One of the largest AIRR studies: Liu et al., 2019 — 877 repertoires,18 million unique TCRs, Snyder et al., 2020 — 1815 covid TCR repertoires + 3500 controls
  - Is the dataset representative? Performance estimation problems
- Available only for one particular problem setup
  - What if the data was a bit different? Sensitivity estimation problem
- No ground truth information
  - What was learned? Generalizability problem

Sandve and Greiff, 2022,
doi: 10.1093/bioinformatics/btac612

Will our ML method also work good on an unseen data?



Nested cross validation might improve generalizability



Training & Validation

Test

Data splits:

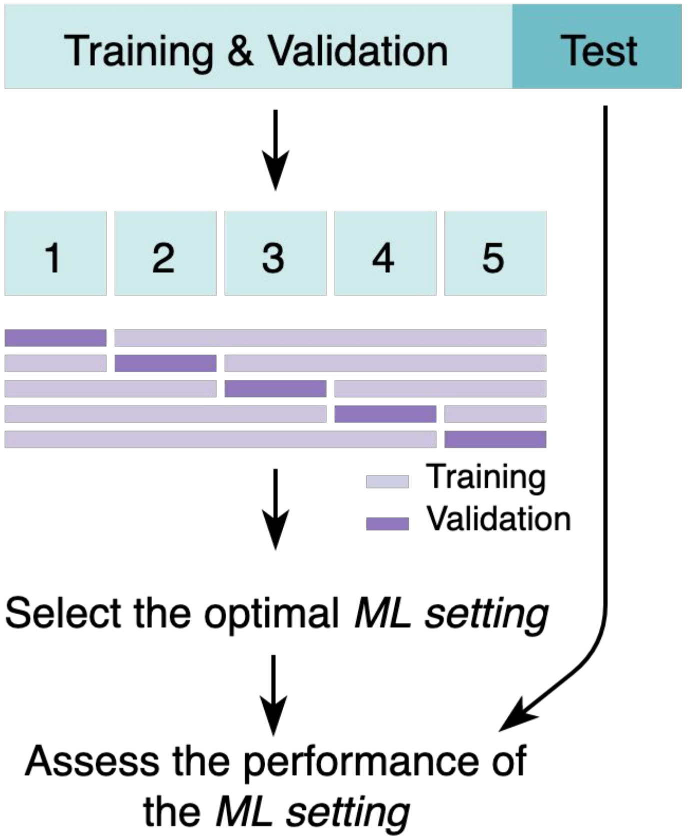
1 2 3 4 5

Training

Validation

Select the optimal *ML setting*

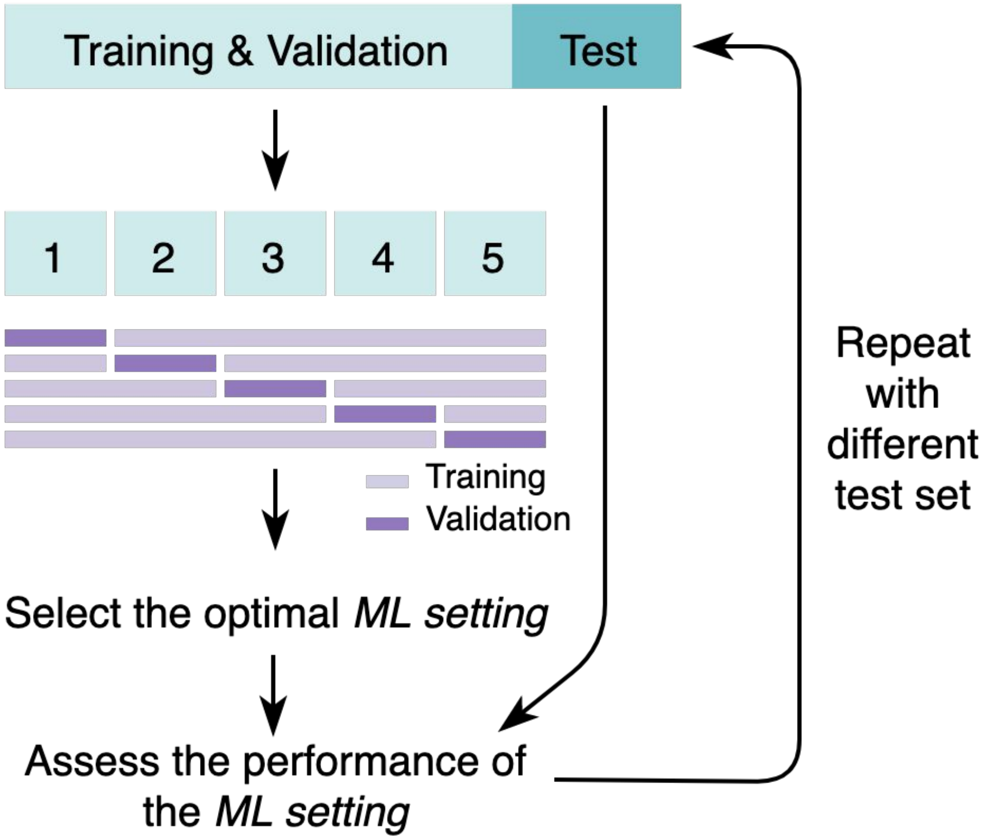
Nested cross validation might improve generalizability



Select the optimal *ML setting*

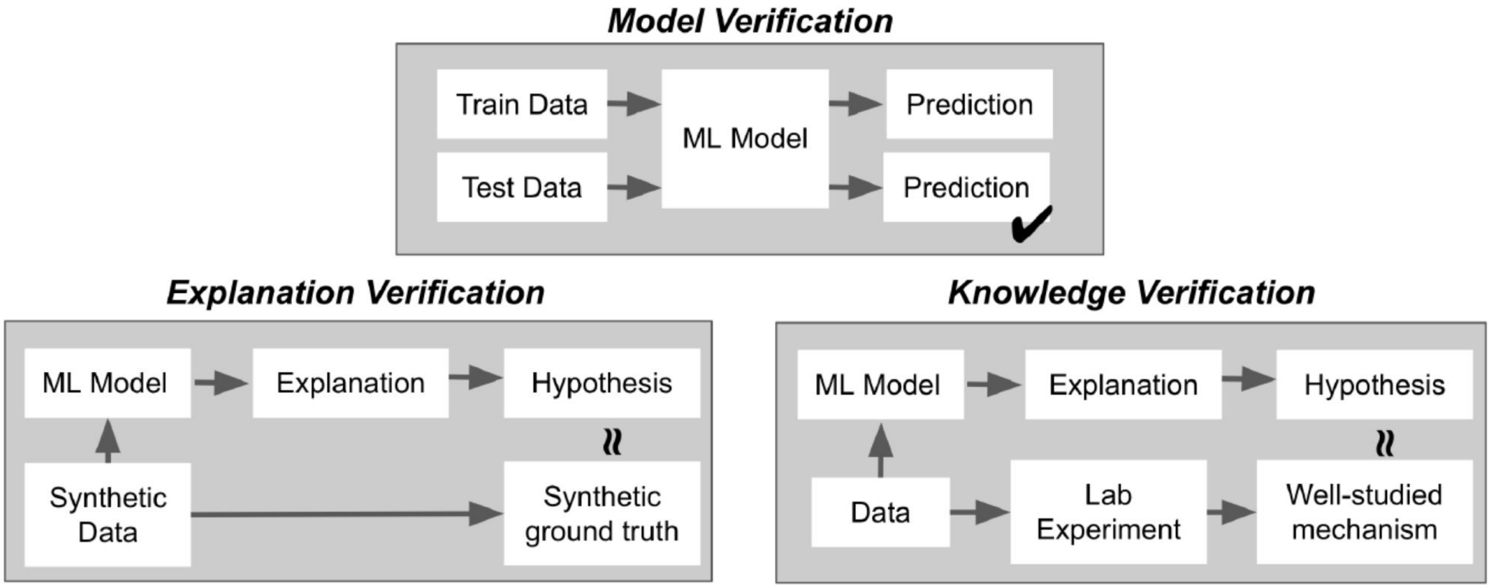
Assess the performance of the *ML setting*

Nested cross validation might improve generalizability


Training & Validation
Test
1
2
3
4
5
Training
Validation
Select the optimal *ML setting*
Assess the performance of the *ML setting*
Repeat with different test set

There are several levels of ML verification

Can we generalise?



Model Verification

Train Data → ML Model → Prediction

Test Data → ML Model → Prediction ✔

Explanation Verification

ML Model → Explanation → Hypothesis

≈

Synthetic Data → ML Model

Synthetic Data → Synthetic ground truth

Knowledge Verification

ML Model → Explanation → Hypothesis

≈

Data → ML Model

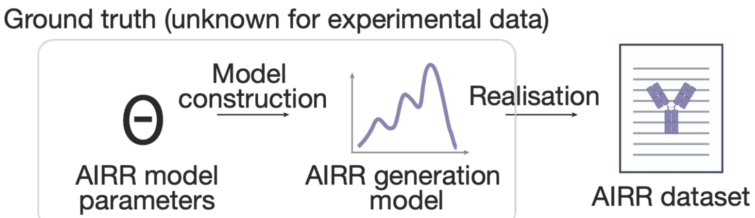
Data → Lab Experiment → Well-studied mechanism

If well-studied prior knowledge is **not** available

If well-studied prior knowledge is available

Chen et al., bioRxiv, 2022

What can be a prior knowledge in AIRR case?



Ground truth (unknown for experimental data)

Θ

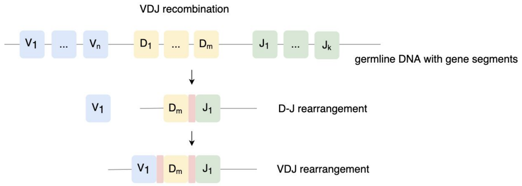
AIRR model parameters

Model construction

AIRR generation model

Realisation

AIRR dataset



VDJ recombination

V1 ... Vn D1 ... Dm J1 ... Jk

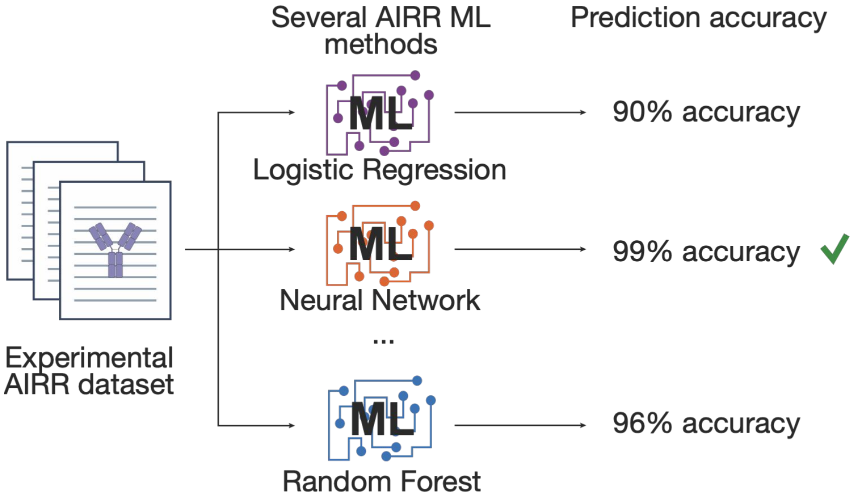
germline DNA with gene segments

V1 Dm J1

D-J rearrangement

V1 Dm J1

VDJ rearrangement



Several AIRR ML methods

Prediction accuracy

Experimental AIRR dataset

ML

Logistic Regression

90% accuracy

ML

Neural Network

99% accuracy ✓

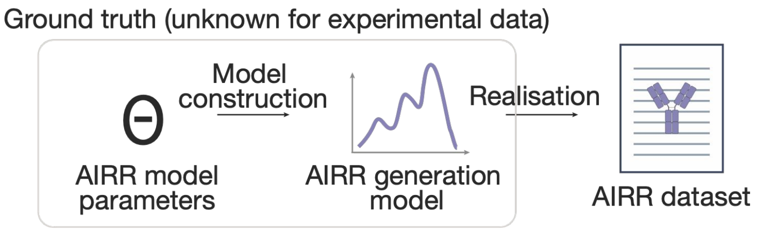
...

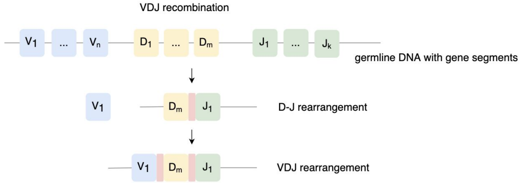
ML

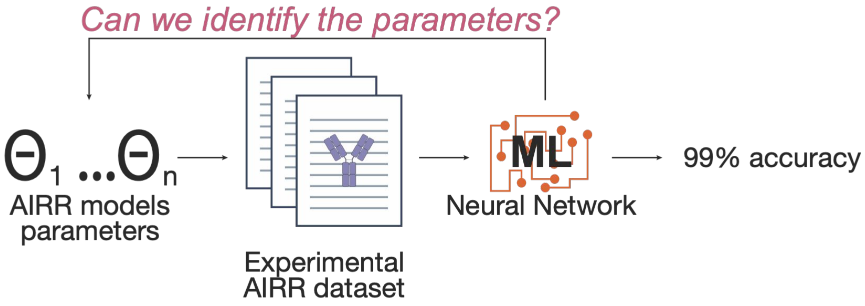
Random Forest

96% accuracy

AIRR ML methods should be also benchmarked on ground truth synthetic AIRR data



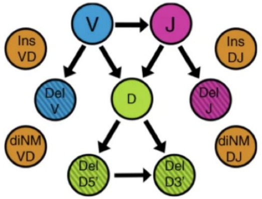




- □ High accuracy
- □ Have we learned the ground truth?

Current VDJ simulation frameworks have pros and cons

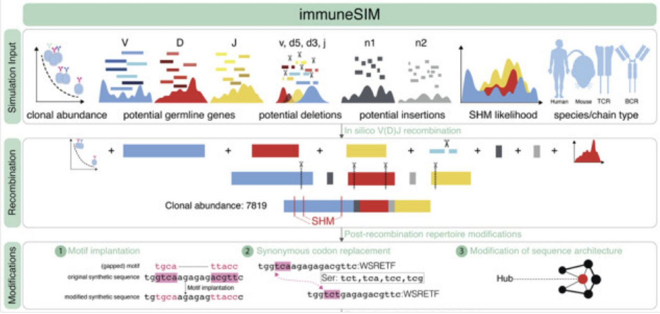
❏ IGoR (nt) / OLGA (aa) (Marcou et al. 2017)



$$
\begin{aligned}
P(\text{scenario}) = {} & P(V)P(J|V)\,P(D|V,J)P(\text{del}V|V) \\
& \times P(\text{del}J|J)\,P(\text{delD5'}|D)\,P(\text{delD3'}|\text{delD5'},D) \\
& \times P(\text{insVD}) \prod_{i}^{\text{InsVD}} P(n_i|n_{i-1}) \\
& \times P(\text{insDJ}) \prod^{\text{InsDJ}} P(n_i|n_{i-1})
\end{aligned}
$$

- + Accurate VDJ recombination model
- + Generation probability evaluation
- + Fast (generates 100k seqs in 5 min)
- - No signal embedding
- - One dataset per one run

❏ immuneSIM (Weber et al. 2020)

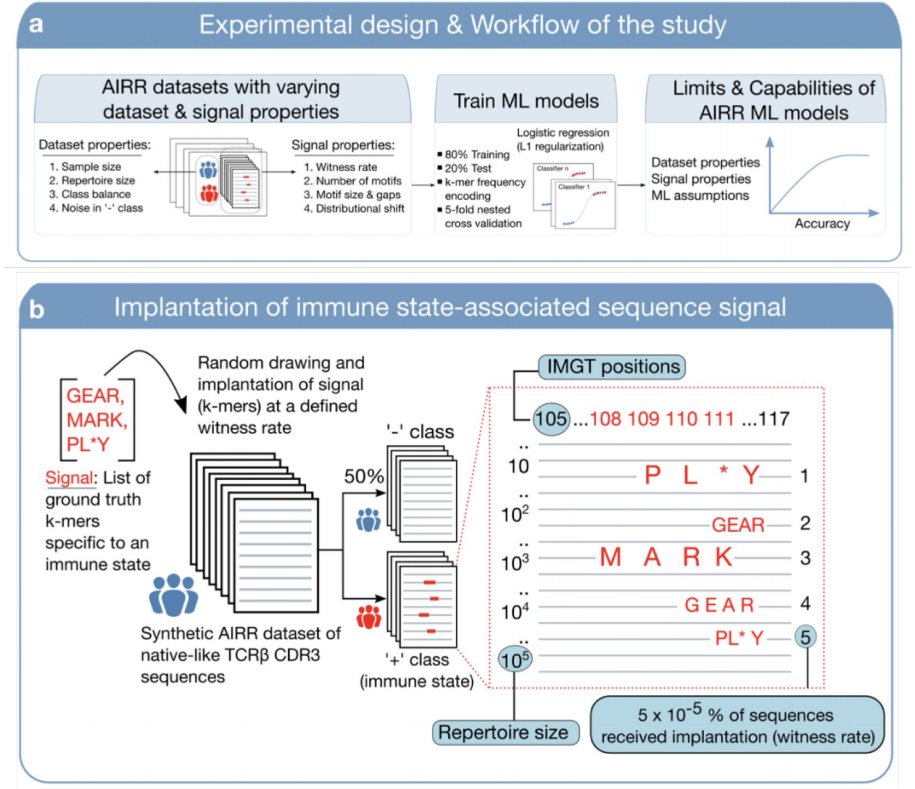


- + Basic signal implantation (gapped k-mers)
- + Simulates clonal abundances
- + Productive receptors
- - Slow (100k seqs in 1 hour)
- - No generation model and generation probabilities
- - One dataset per one run

Profiling AIRR ML models on a range of basic datasets

- ❏ OLGA TCRβ CDR3 sequences (aa) + implanted gapped 2—5-mers
- ❏ Identified parameter boundaries where baseline methods (Logistic Regression) already achieve high accuracy

Immune signal can be more complex!

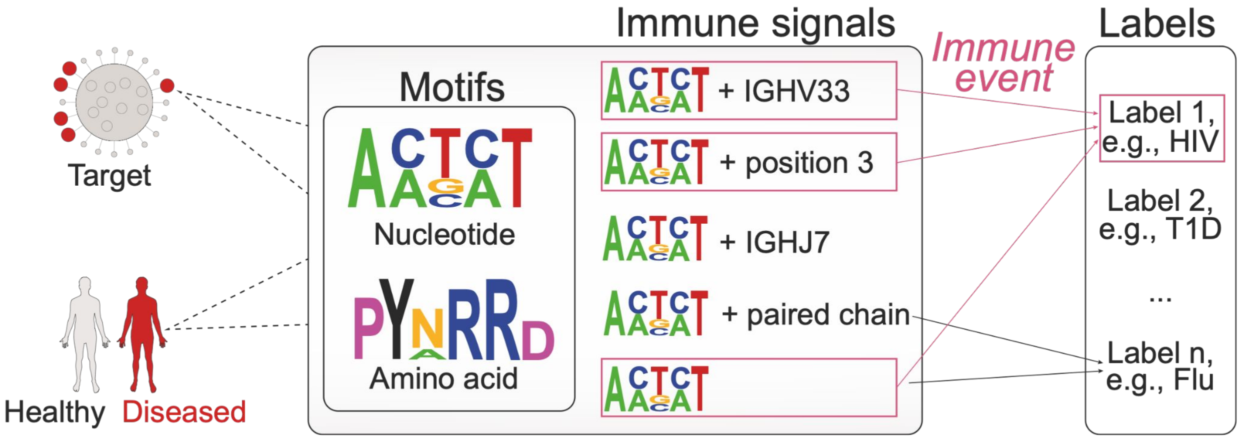

a
Experimental design & Workflow of the study
AIRR datasets with varying dataset & signal properties
Dataset properties:
1. Sample size
2. Repertoire size
3. Class balance
4. Noise in '-' class
Signal properties:
1. Witness rate
2. Number of motifs
3. Motif size & gaps
4. Distributional shift
Train ML models
Logistic regression (L1 regularization)
▪ 80% Training
▪ 20% Test
▪ k-mer frequency encoding
▪ 5-fold nested cross validation
Classifier n
Classifier 1
Limits & Capabilities of AIRR ML models
Dataset properties
Signal properties
ML assumptions
Accuracy
b
Implantation of immune state-associated sequence signal
GEAR, MARK, PL*Y
Signal: List of ground truth k-mers specific to an immune state
Random drawing and implantation of signal (k-mers) at a defined witness rate
Synthetic AIRR dataset of native-like TCRβ CDR3 sequences
50%
'-' class
'+' class (immune state)
IMGT positions
105 ...108 109 110 111 ...117
10
10^2
10^3
10^4
10^5
P L * Y 1
GEAR 2
M A R K 3
G E A R 4
PL* Y 5
Repertoire size
5 x 10^{-5} % of sequences received implantation (witness rate)

Kanduri et al. 2022, doi: 10.1093/gigascience/giac046

Definition of immune event and immune signal

We hypothesise that immune signal should be a substring of the receptor:

- ❏ (Gapped) k-mer (Akbar et al., 2021)
- ❏ Full-length receptor (Emerson et al., 2017)
- ❏ Motif (PWM with a fixed length)
- ❏ The most general definition: immune signal is a function: AIR → True/False

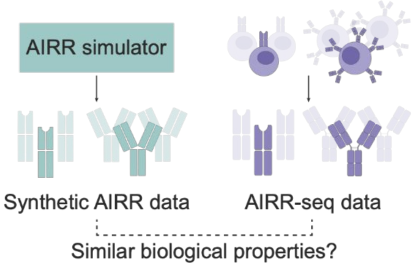


Chernigovskaya, unpublished

A universal AIRR simulator wishlist

❏ Challenges in AIRR simulator development

Synthetic AIRR data nativeness



Overlapping signals

Sequence level

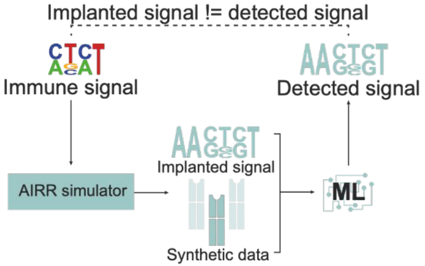
Signal 1 Signal 2 Signal 3

IGHV30 + CASSIMMIEKLAKENEQY + IGHJ5

Repertoire level

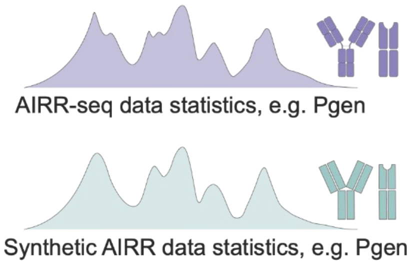
```
CASSIMMIRKKALENEQY  Signal 1
CASSIMMIEAKELANEQY
CASSIMMIEKLAKENEQY  Signal 1 Signal 2
CASVIMMISRLAKENEQY
CASSMCAWEKLAKENEQY  Signal 2
CASSIMCIEKLAKENEQY
CASSCMMCPKLAKENEQY
```

Introducing simulation artifacts

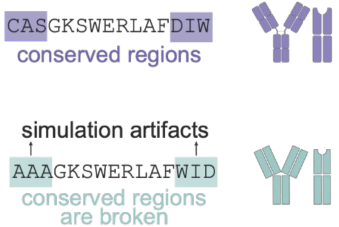


❏ Properties of a universal AIRR simulator

Similar biological statistics distributions

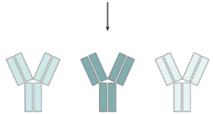


Does not break biological properties of AIRR data



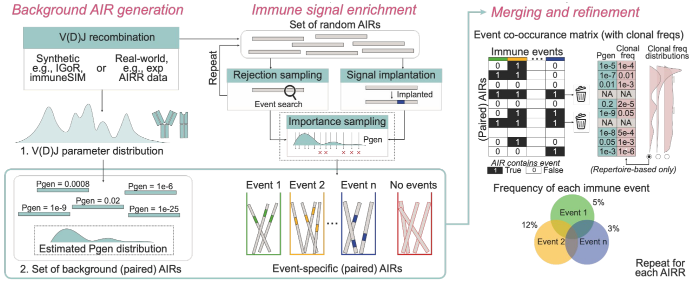
Can simulate signal of varying complexity

```
def signal:
   v_gene = IGHV33;
   CDR3 contains motif (A/C)RRR;
   motif position in the CDR3 = 4 or 6;
   CDR3 length = 12AA;
   ...
```

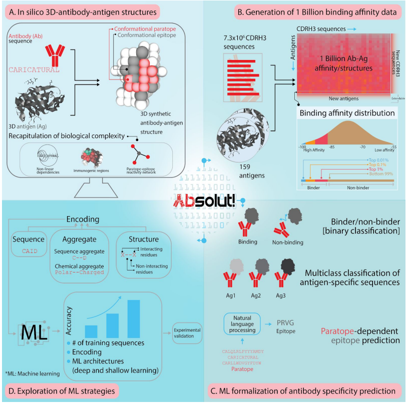


Chernigovskaya, unpublished

Framework for simulating a “native”-like AIR(R) datasets

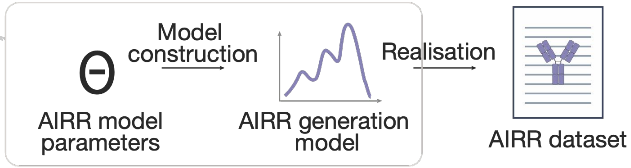


Chernigovskaya, unpublished

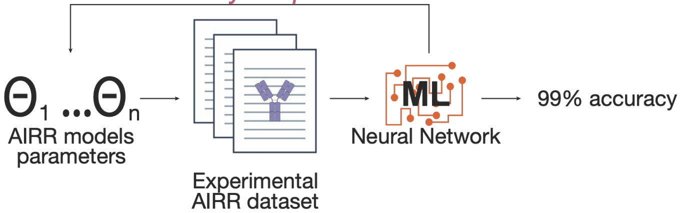

A. In silico 3D-antibody-antigen structures
Antibody (Ab) sequence
CARICATURAL
Conformational paratope
Conformational epitope
3D antigen (Ag)
3D synthetic antibody-antigen structure
Recapitulation of biological complexity
B. Generation of 1 Billion binding affinity data
7.3x10⁶ CDRH3 sequences
CDRH3 sequences
Antigens
1 Billion Ab-Ag affinity/structures
New antigens
159 antigens
Binding affinity distribution
Binder
Non-binder
Absolut!
Encoding
Sequence
Aggregate
Structure
ML
Accuracy
# of training sequences
Encoding
ML architectures (deep and shallow learning)
Experimental validation
*ML: Machine learning
Binder/non-binder [binary classification]
Binding
Non-binding
Multiclass classification of antigen-specific sequences
Ag1
Ag2
Ag3
Natural language processing
PRVG Epitope
Paratope
Paratope-dependent epitope prediction
D. Exploration of ML strategies
C. ML formalization of antibody specificity prediction

Robert et al. 2022

Ground truth (unknown for experimental data)

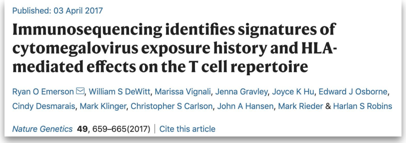


Can we identify the parameters?



- ❑ High accuracy
- ❑ Have we learned the ground truth?

Conceptual problem: reproducible AIRR ML

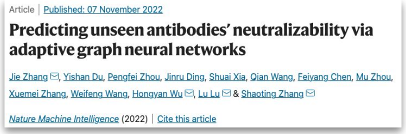


Published: 03 April 2017

Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire

Ryan O Emerson, William S DeWitt, Marissa Vignali, Jenna Gravley, Joyce K Hu, Edward J Osborne, Cindy Desmarais, Mark Klinger, Christopher S Carlson, John A Hansen, Mark Rieder & Harlan S Robins

Nature Genetics **49**, 659–665(2017) | Cite this article



Article | Published: 07 November 2022

Predicting unseen antibodies' neutralizability via adaptive graph neural networks

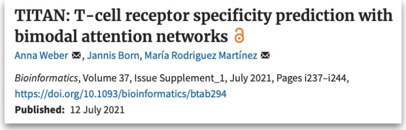
Jie Zhang, Yishan Du, Pengfei Zhou, Jinru Ding, Shuai Xia, Qian Wang, Feiyang Chen, Mu Zhou, Xuemei Zhang, Weifeng Wang, Hongyan Wu, Lu Lu & Shaoting Zhang

Nature Machine Intelligence (2022) | Cite this article

DeepTCR: a deep learning framework for understanding T-cell receptor sequence signatures within complex T-cell repertoires

John-William Sidhom, H. Benjamin Larman, Petra Ross-MacDonald, Megan Wind-Rotolo, Drew M. Pardoll, Alexander S. Baras

doi: https://doi.org/10.1101/464107

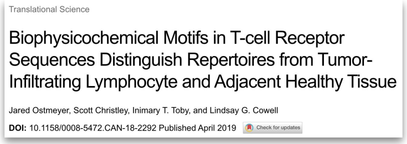


TITAN: T-cell receptor specificity prediction with bimodal attention networks

Anna Weber ✉, Jannis Born, María Rodriguez Martínez ✉

Bioinformatics, Volume 37, Issue Supplement_1, July 2021, Pages i237–i244, https://doi.org/10.1093/bioinformatics/btab294

Published: 12 July 2021



Translational Science

Biophysicochemical Motifs in T-cell Receptor Sequences Distinguish Repertoires from Tumor-Infiltrating Lymphocyte and Adjacent Healthy Tissue

Jared Ostmeyer, Scott Christley, Inimary T. Toby, and Lindsay G. Cowell

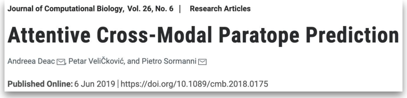
DOI: 10.1158/0008-5472.CAN-18-2292 Published April 2019 Check for updates

Predicting antigen specificity of single T cells based on TCR CDR3 regions

David S Fischer, Yihan Wu, Benjamin Schubert, Fabian J Theis

Author Information

Mol Syst Biol (2020) 16: e9416 | https://doi.org/10.15252/msb.20199416

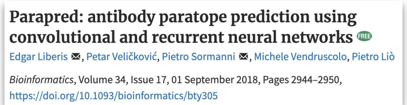


Journal of Computational Biology, Vol. 26, No. 6 | Research Articles

Attentive Cross-Modal Paratope Prediction

Andreea Deac ✉, Petar VeliČković, and Pietro Sormanni ✉

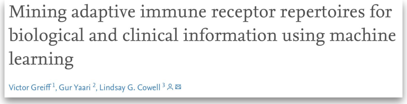
Published Online: 6 Jun 2019 | https://doi.org/10.1089/cmb.2018.0175



Parapred: antibody paratope prediction using convolutional and recurrent neural networks FREE

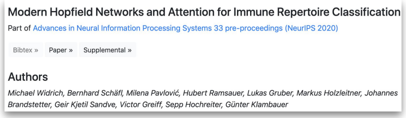
Edgar Liberis ✉, Petar Veličković, Pietro Sormanni ✉, Michele Vendruscolo, Pietro Liò

Bioinformatics, Volume 34, Issue 17, 01 September 2018, Pages 2944–2950, https://doi.org/10.1093/bioinformatics/bty305



Mining adaptive immune receptor repertoires for biological and clinical information using machine learning

Victor Greiff[1], Gur Yaari[2], Lindsay G. Cowell[3] ✉



Modern Hopfield Networks and Attention for Immune Repertoire Classification

Part of Advances in Neural Information Processing Systems 33 pre-proceedings (NeurIPS 2020)

Bibtex » Paper » Supplemental »

Authors

Michael Widrich, Bernhard Schäfl, Milena Pavlović, Hubert Ramsauer, Lukas Gruber, Markus Holzleitner, Johannes Brandstetter, Geir Kjetil Sandve, Victor Greiff, Sepp Hochreiter, Günter Klambauer

Research article | Open Access | Published: 28 May 2019

Capturing the differences between humoral immunity in the normal and tumor environments from repertoire-seq of B-cell receptors using supervised machine learning

Hiroki Konishi, Daisuke Komura, Hiroto Katoh, Shinichiro Atsumi, Hirotomo Koda, Asami Yamamoto, Yasuyuki Seto, Masashi Fukayama, Rui Yamaguchi, Seiya Imoto & Shumpei Ishikawa

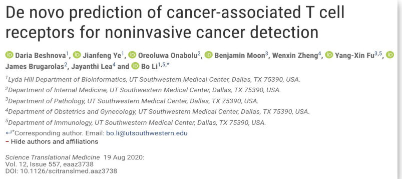
BMC Bioinformatics **20**, Article number: 267 (2019) | Cite this article

ORIGINAL RESEARCH ARTICLE

Front. Immunol., 29 November 2019 | https://doi.org/10.3389/fimmu.2019.02820

Detection of Enriched T Cell Epitope Specificity in Full T Cell Receptor Sequence Repertoires

Sofie Gielis[1,2,3], **Pieter Moris**[1,3†], **Wout Bittremieux**[1,3,4†], **Nicolas De Neuter**[1,2,3], **Benson Ogunjimi**[2,5,6,7], **Kris Laukens**[1,2,3‡] and **Pieter Meysman**[1,2,3*‡]



De novo prediction of cancer-associated T cell receptors for noninvasive cancer detection

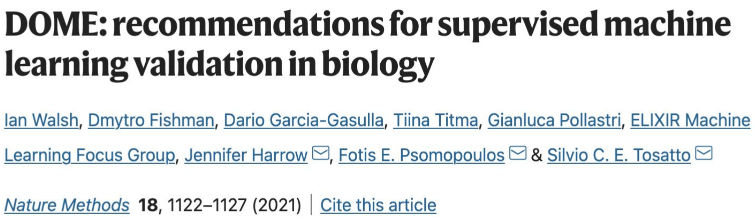
Daria Beshnova[1], **Jianfeng Ye**[1], **Oreoluwa Onabolu**[2], **Benjamin Moon**[3], **Wenxin Zheng**[4], **Yang-Xin Fu**[3,5], **James Brugarolas**[2], **Jayanthi Lea**[4] and **Bo Li**[1,5,*]

[1]Lyda Hill Department of Bioinformatics, UT Southwestern Medical Center, Dallas, TX 75390, USA.
[2]Department of Internal Medicine, UT Southwestern Medical Center, Dallas, TX 75390, USA.
[3]Department of Pathology, UT Southwestern Medical Center, Dallas, TX 75390, USA.
[4]Department of Obstetrics and Gynecology, UT Southwestern Medical Center, Dallas, TX 75390, USA.
[5]Department of Immunology, UT Southwestern Medical Center, Dallas, TX 75390, USA.
↵*Corresponding author. Email: bo.li@utsouthwestern.edu
– Hide authors and affiliations

Science Translational Medicine 19 Aug 2020:
Vol. 12, Issue 557, eaaz3738
DOI: 10.1126/scitranslmed.aaz3738

How can we make all these studies reproducible?

Recommendations for ML in biology



DOME: recommendations for supervised machine learning validation in biology

Ian Walsh, Dmytro Fishman, Dario Garcia-Gasulla, Tiina Titma, Gianluca Pollastri, ELIXIR Machine Learning Focus Group, Jennifer Harrow ✉, Fotis E. Psomopoulos ✉ & Silvio C. E. Tosatto ✉

Nature Methods **18**, 1122–1127 (2021) | Cite this article

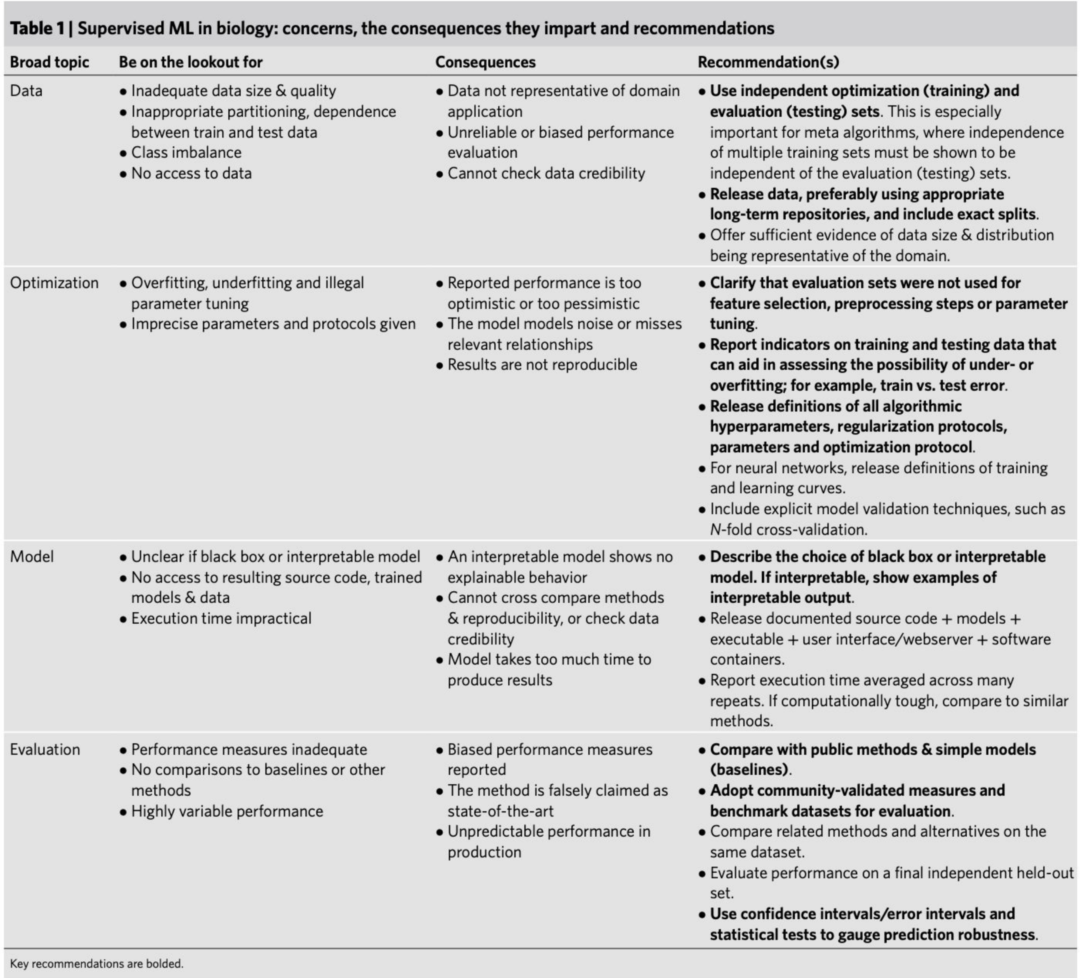
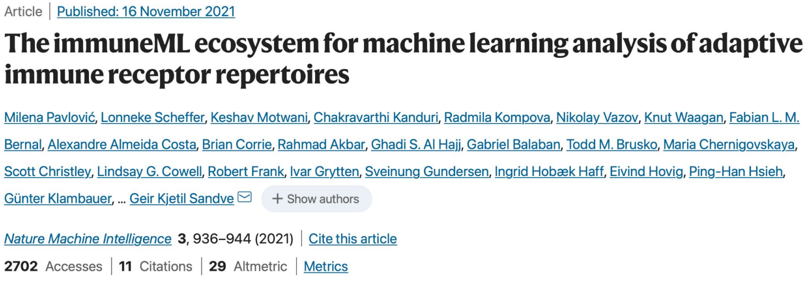


Table 1 | Supervised ML in biology: concerns, the consequences they impart and recommendations

Broad topic	Be on the lookout for	Consequences	Recommendation(s)
Data	• Inadequate data size & quality • Inappropriate partitioning, dependence between train and test data • Class imbalance • No access to data	• Data not representative of domain application • Unreliable or biased performance evaluation • Cannot check data credibility	• **Use independent optimization (training) and evaluation (testing) sets**. This is especially important for meta algorithms, where independence of multiple training sets must be shown to be independent of the evaluation (testing) sets. • **Release data, preferably using appropriate long-term repositories, and include exact splits**. • Offer sufficient evidence of data size & distribution being representative of the domain.
Optimization	• Overfitting, underfitting and illegal parameter tuning • Imprecise parameters and protocols given	• Reported performance is too optimistic or too pessimistic • The model models noise or misses relevant relationships • Results are not reproducible	• **Clarify that evaluation sets were not used for feature selection, preprocessing steps or parameter tuning**. • **Report indicators on training and testing data that can aid in assessing the possibility of under- or overfitting; for example, train vs. test error**. • **Release definitions of all algorithmic hyperparameters, regularization protocols, parameters and optimization protocol**. • For neural networks, release definitions of training and learning curves. • Include explicit model validation techniques, such as *N*-fold cross-validation.
Model	• Unclear if black box or interpretable model • No access to resulting source code, trained models & data • Execution time impractical	• An interpretable model shows no explainable behavior • Cannot cross compare methods & reproducibility, or check data credibility • Model takes too much time to produce results	• **Describe the choice of black box or interpretable model. If interpretable, show examples of interpretable output**. • Release documented source code + models + executable + user interface/webserver + software containers. • Report execution time averaged across many repeats. If computationally tough, compare to similar methods.
Evaluation	• Performance measures inadequate • No comparisons to baselines or other methods • Highly variable performance	• Biased performance measures reported • The method is falsely claimed as state-of-the-art • Unpredictable performance in production	• **Compare with public methods & simple models (baselines)**. • **Adopt community-validated measures and benchmark datasets for evaluation**. • Compare related methods and alternatives on the same dataset. • Evaluate performance on a final independent held-out set. • **Use confidence intervals/error intervals and statistical tests to gauge prediction robustness**.

Key recommendations are bolded.

immuneML is a platform for development and transparent comparative evaluation of AIRR-ML methods



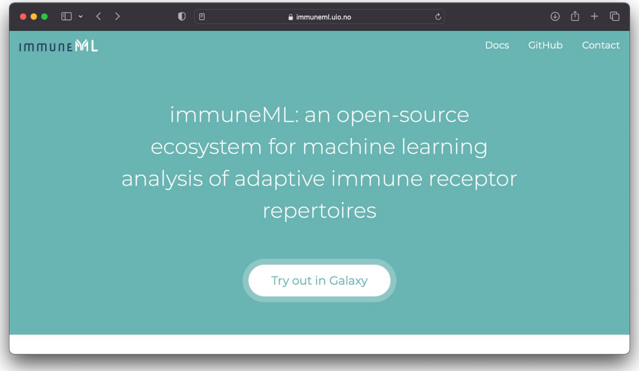
Article | Published: 16 November 2021

The immuneML ecosystem for machine learning analysis of adaptive immune receptor repertoires

Milena Pavlović, Lonneke Scheffer, Keshav Motwani, Chakravarthi Kanduri, Radmila Kompova, Nikolay Vazov, Knut Waagan, Fabian L. M. Bernal, Alexandre Almeida Costa, Brian Corrie, Rahmad Akbar, Ghadi S. Al Hajj, Gabriel Balaban, Todd M. Brusko, Maria Chernigovskaya, Scott Christley, Lindsay G. Cowell, Robert Frank, Ivar Grytten, Sveinung Gundersen, Ingrid Hobæk Haff, Eivind Hovig, Ping-Han Hsieh, Günter Klambauer, … Geir Kjetil Sandve ✉ + Show authors

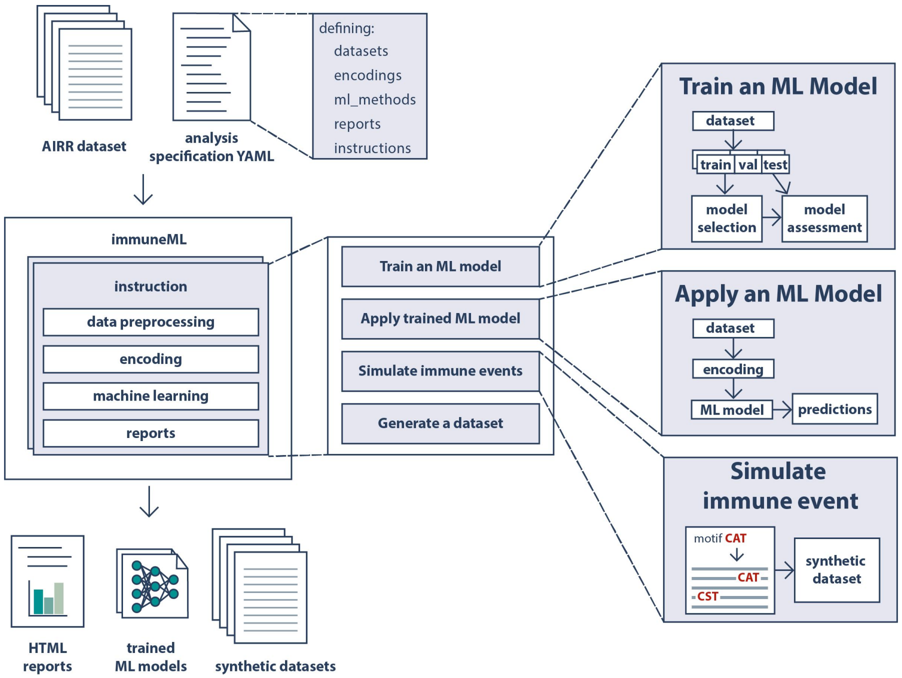
Nature Machine Intelligence **3**, 936–944 (2021) | Cite this article

2702 Accesses | **11** Citations | **29** Altmetric | Metrics



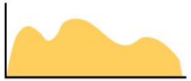
immuneML: an open-source ecosystem for machine learning analysis of adaptive immune receptor repertoires

https://immuneml.uio.no

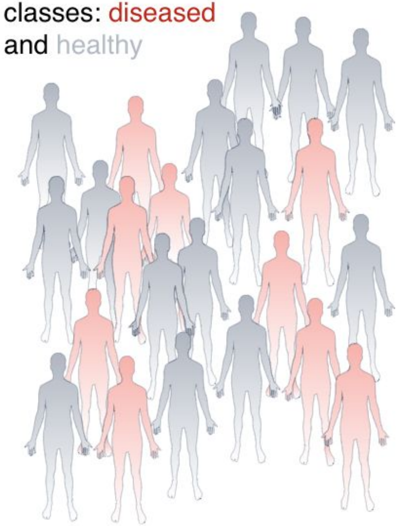


Diagnosing diseases with AIRRs

distribution of
variables of interest

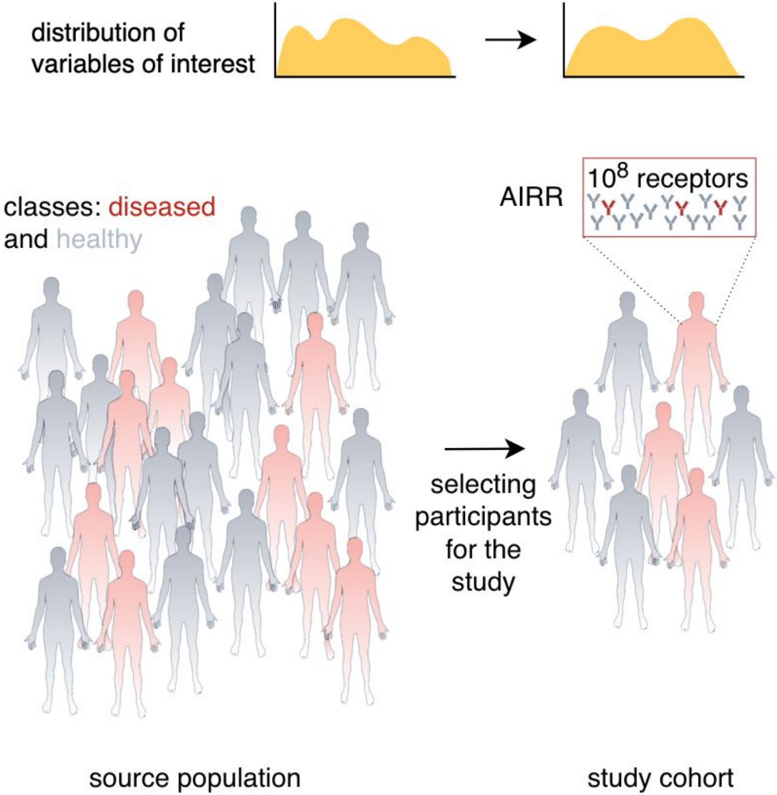


classes: diseased
and healthy

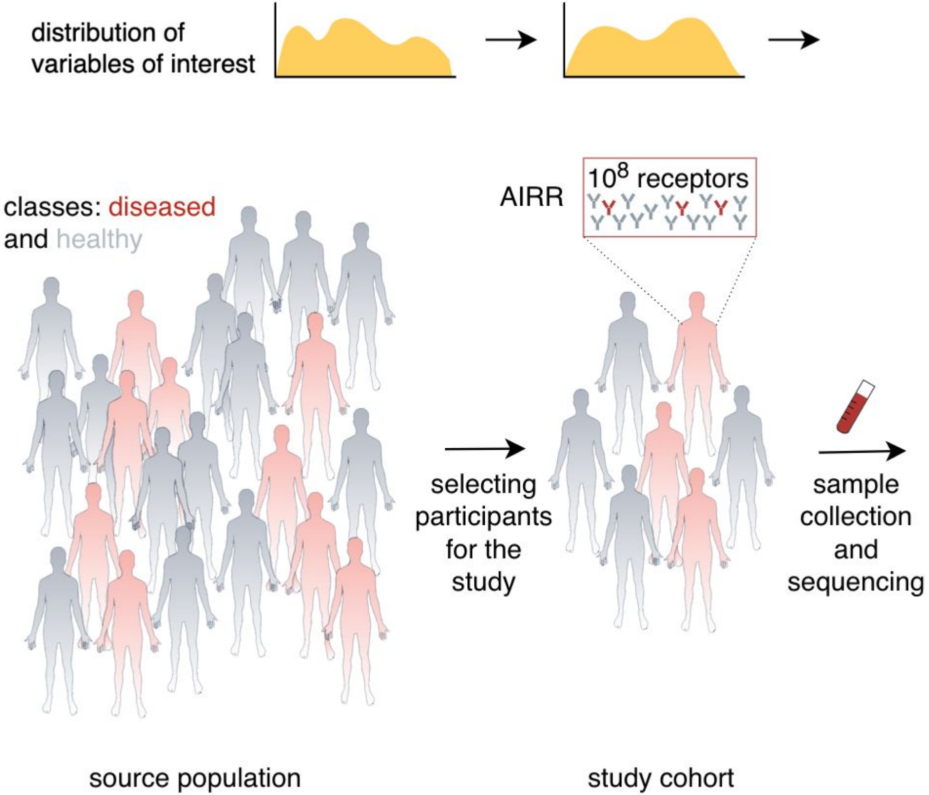


source population

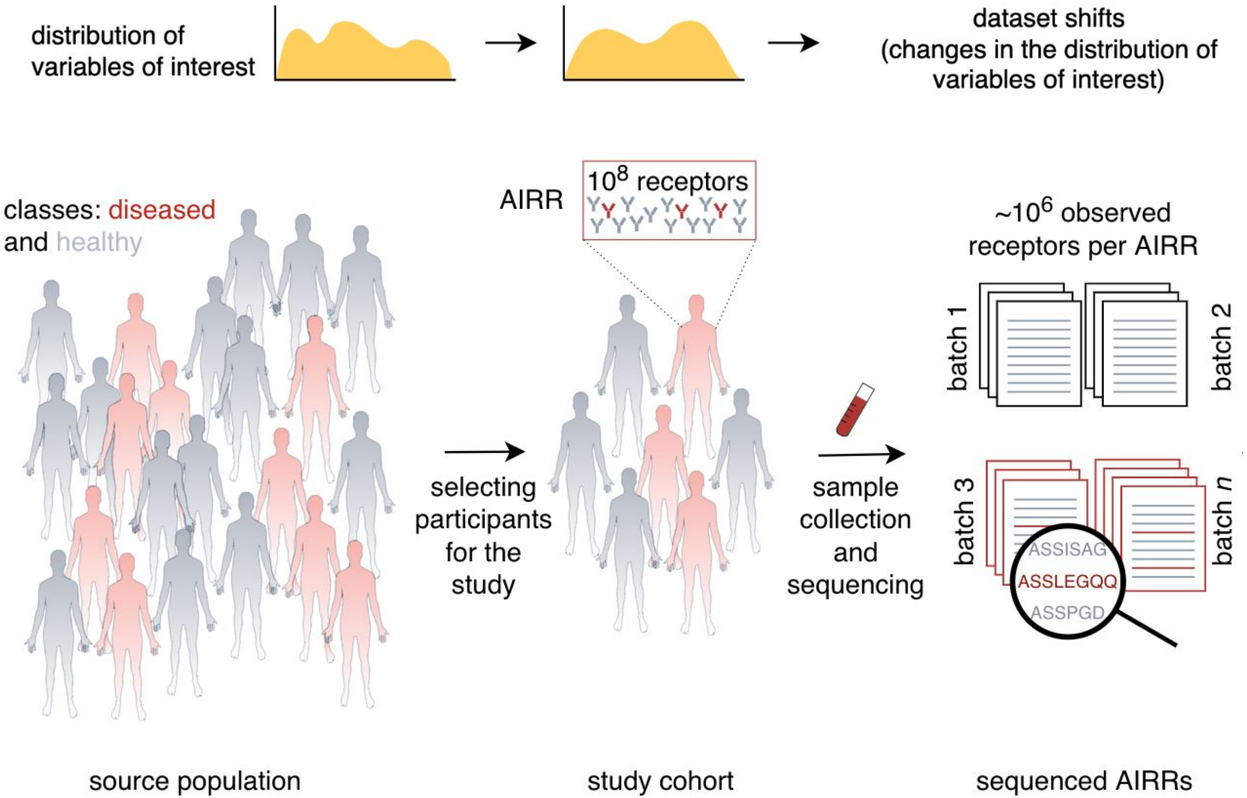
Diagnosing diseases with AIRRs


distribution of variables of interest
classes: diseased and healthy
AIRR
10^8 receptors
selecting participants for the study
source population
study cohort

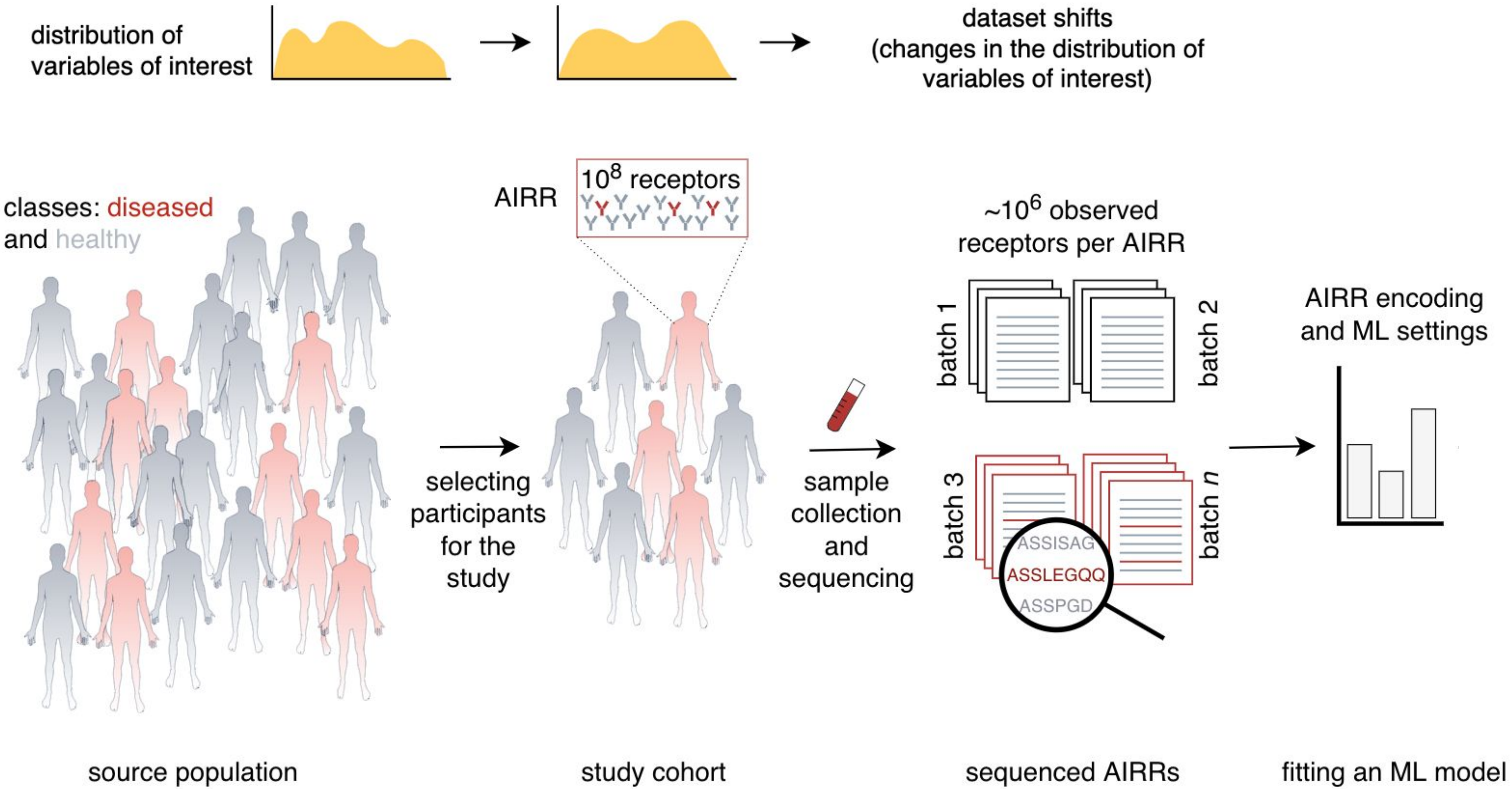
Diagnosing diseases with AIRRs


distribution of variables of interest
classes: diseased and healthy
AIRR
10^8 receptors
selecting participants for the study
sample collection and sequencing
source population
study cohort

Diagnosing diseases with AIRRs

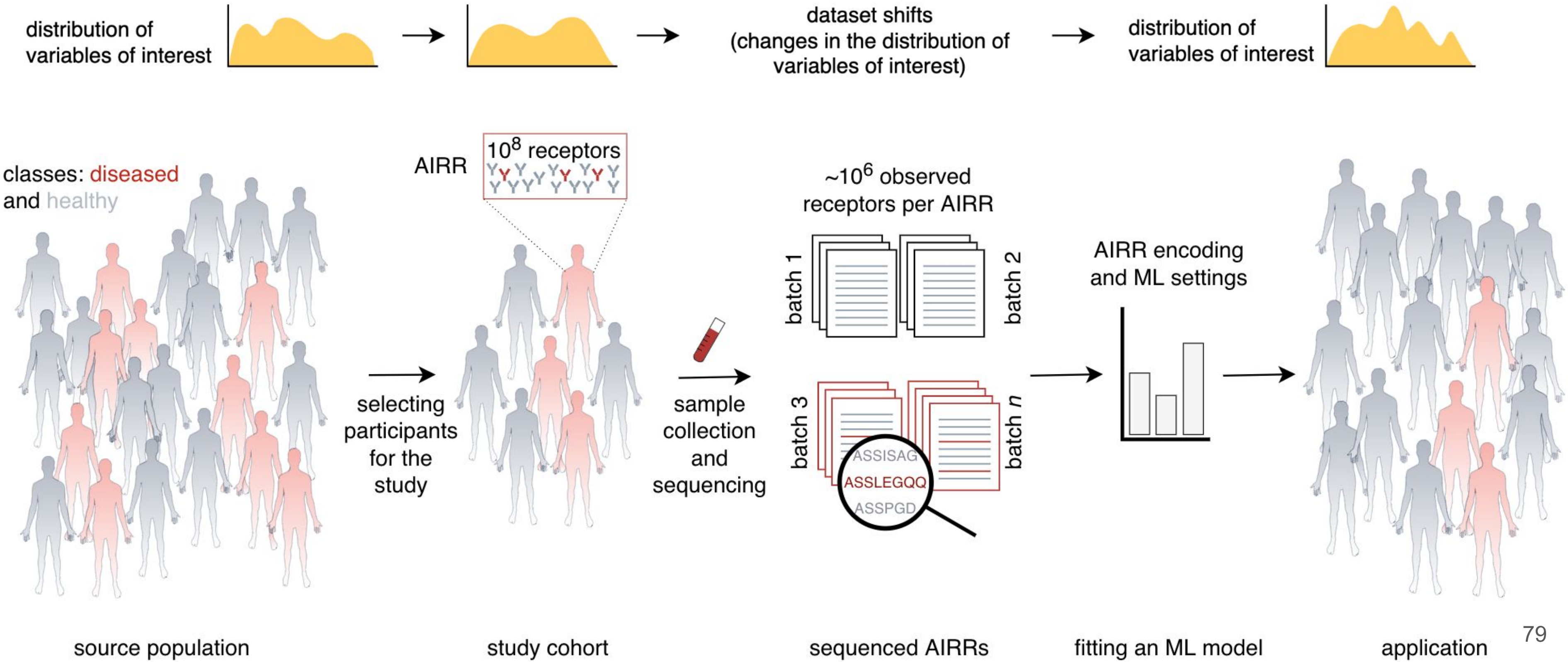

distribution of variables of interest
dataset shifts (changes in the distribution of variables of interest)
classes: diseased and healthy
AIRR
10^8 receptors
selecting participants for the study
sample collection and sequencing
~10^6 observed receptors per AIRR
batch 1
batch 2
batch 3
batch *n*
ASSISAG
ASSLEGQQ
ASSPGD
source population
study cohort
sequenced AIRRs

Diagnosing diseases with AIRRs



distribution of variables of interest
dataset shifts (changes in the distribution of variables of interest)
classes: diseased and healthy
AIRR
10^8 receptors
selecting participants for the study
sample collection and sequencing
~10^6 observed receptors per AIRR
batch 1
batch 2
batch 3
batch *n*
ASSISAG
ASSLEGQQ
ASSPGD
AIRR encoding and ML settings
source population
study cohort
sequenced AIRRs
fitting an ML model

Diagnosing diseases with AIRRs


distribution of variables of interest
dataset shifts (changes in the distribution of variables of interest)
distribution of variables of interest
classes: diseased and healthy
AIRR
10^8 receptors
selecting participants for the study
sample collection and sequencing
~10^6 observed receptors per AIRR
batch 1
batch 2
batch 3
batch *n*
ASSISAG
ASSLEGQQ
ASSPGD
AIRR encoding and ML settings
source population
study cohort
sequenced AIRRs
fitting an ML model
application

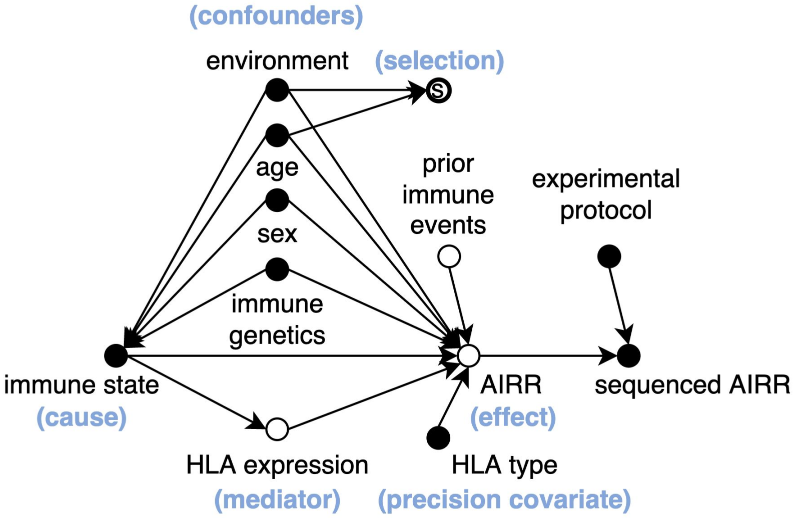
The causal inference framework

- Formally describes the data-generating process to discover causal effects between the variables in the process, under a set of assumptions (Pearl 2009)
- Causal effect of X on Y: the difference in the value of Y while changing X and keeping all other variables and conditions the same

Causality doesn’t matter [too much] for prediction tasks, but when obtaining the data or applying methods to new populations, causality can help formalize and solve challenges even in predictive settings

The causal inference framework in the AIRR field

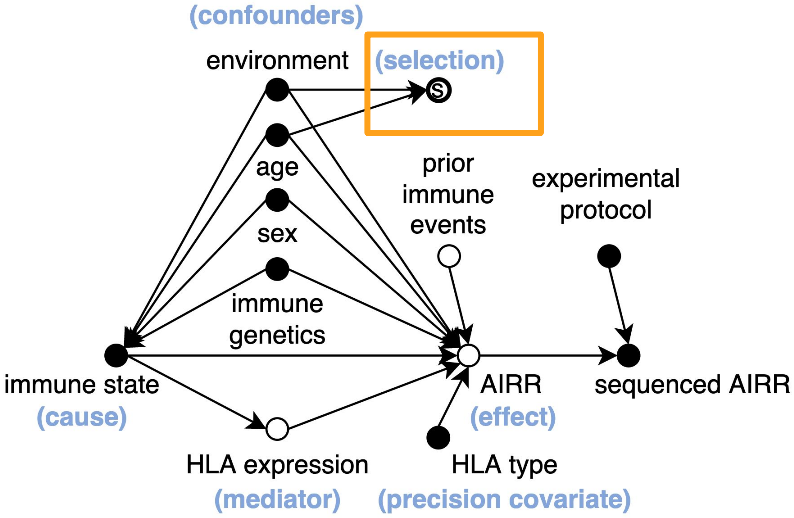
- A causal model for a viral infection (different for different diseases)



Pavlovic et al. 2022

The causal inference framework in the AIRR field

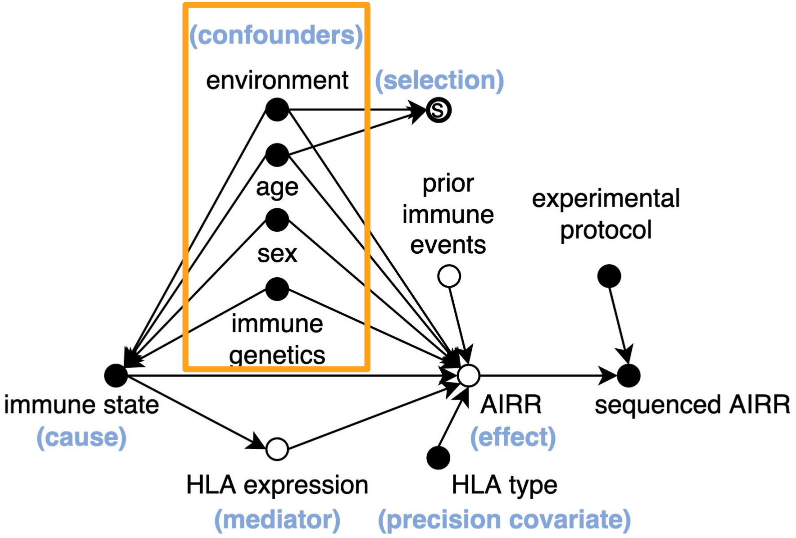
- ❑ A causal model for a viral infection (different for different diseases)
- ❑ Selection bias:
  - ❑ preferential selection of study participants
  - ❑ spurious correlations: introduced, removed, reversed



Pavlovic et al. 2022

The causal inference framework in the AIRR field

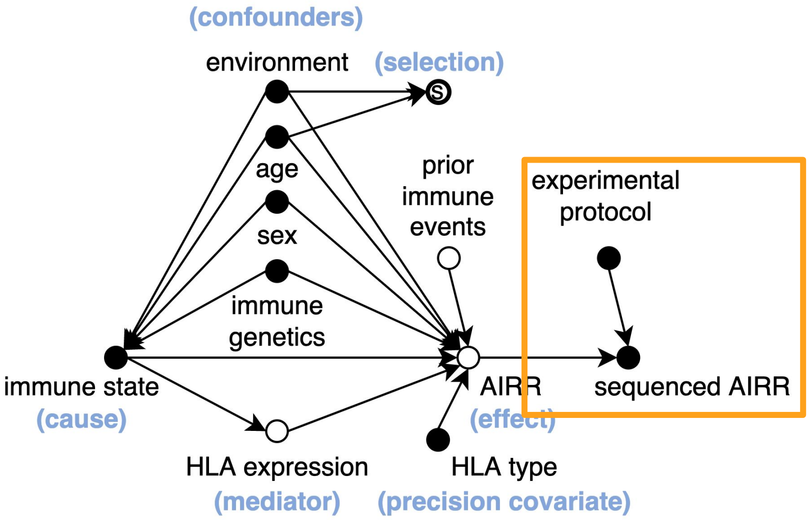
- ❏ A causal model for a viral infection (different for different diseases)
- ❏ Selection bias:
  - ❏ preferential selection of study participants
  - ❏ spurious correlations: introduced, removed, reversed
- ❏ Confounding bias:
  - ❏ influence both the immune state and AIRR



Pavlovic et al. 2022

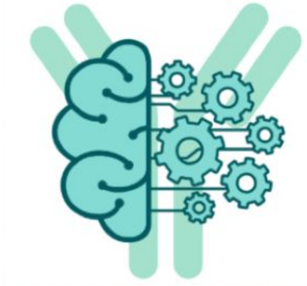
The causal inference framework in the AIRR field

- A causal model for a viral infection (different for different diseases)
- Selection bias:
  - preferential selection of study participants
  - spurious correlations: introduced, removed, reversed
- Confounding bias:
  - influence both the immune state and AIRR
- Batch effects & timing of measurement



Pavlovic et al. 2022

Summary



- ❑ AIRR ML ≠ applying several fancy ML method to AIRR data
    - ❑ Complex biological structure (both receptors and repertoires)
    - ❑ Large variability and sparsity
    - ❑ Causal variables (sex, age, HLA etc)
- ❑ AIRR ML methods should be benchmarked on both experimental and synthetic data with known ground truth
- ❑ Ultimately we need large-scale experimental data with known ground truth

Acknowledgements

UiO Oslo
Prof. Geir Kjetil Sandve
Prof. Victor Greiff
Prof. Dag T.T. Haug
Prof. Ingrid H. Haff
Prof. Ludvig Sollid
Prof. Torbjørn Rognes
Dr. Fridtjof Lund-Johansen
Dr. Rahmad Akbar
Dr. Igor Snapkov
Dr. Philippe Robert
Dr. Chakravarthi Kanduri
Dr. Ivar Grytten
Dr. Knut Rand
Dr. Enrico Riccardi
Dr. Mai Ha Vu
Dr. Brij Bhusan Mehta
Lonneke Scheffer
Andrei Slabodkin
Ghadi Al Hajj
Robert Frank
Thomas Minotto
Khang Le Quy

UFlorida
Keshav Motwani
Prof. Todd Brusko

ETH Zurich
Dr. Cédric R. Weber
Prof. Sai T. Reddy

FHNW
Prof. Enkelejda Miho

JKU Linz
Prof. Günter Klambauer
Prof. Sepp Hochreiter
Michael Widrich

iReceptor+
Prof. Gur Yaari
Prof. Lindsay Cowell
Dr. Scott Christley
Dr. Artur Rocha
Alexandre Almeida Costa

UCSD
Dr. Yana Safonova
Prof. Pavel Pevzner

BGI
Prof. Xiao Liu
Wei Zhang Longlong
Wang Jinghua Wu
Ziyun Wan
Shiyu Wang
Kai Gao

Funding
University of Oslo
Research Council of Norway
Norwegian Cancer Society
Horizon2020
K. G. Jebsen Center