Machine learning in immunology

A brief overview

Vadim Nazarov

Genomics of Adaptive Immunity Lab, IBCH RAS
National Research University Higher School of Economics

Table of contents

Introduction to immunology

Immune system

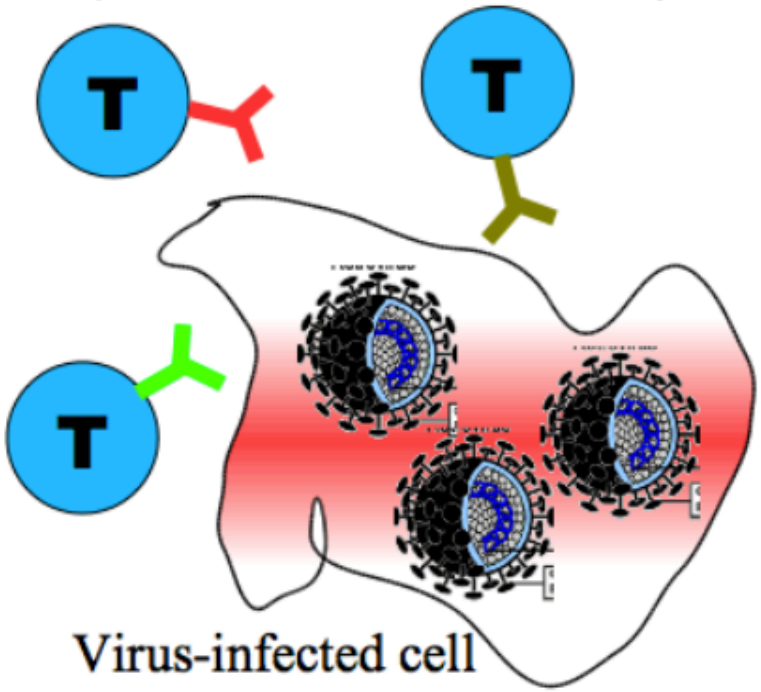
- Recognizes foreign / dangerous substances from the environment (mainly microbes).
- Is envolved in elimintation of old and damaged cells of the body.
- Attacks tumor and virus-infected celss.

Two branches of immune system

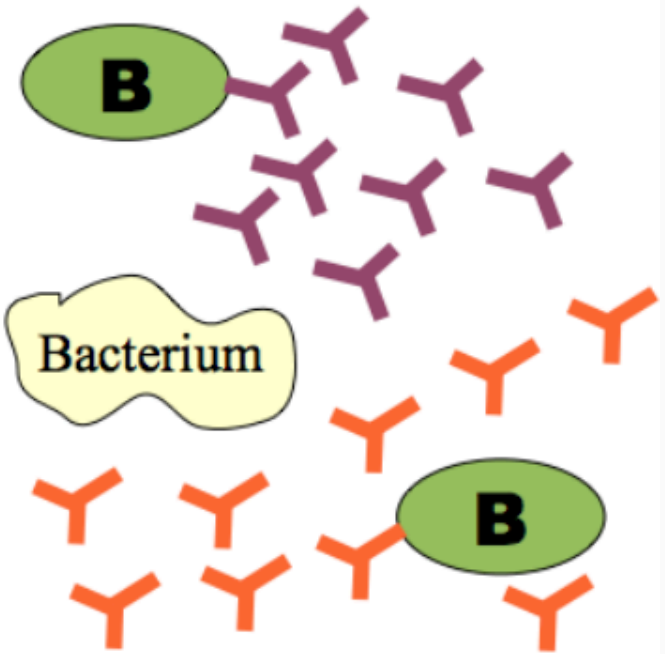
- Innate, nonspecific – very quickly recognizes most foreign substances and eliminates them. No memory or learning.
- Adaptive, specific – high degree of specificity in distinction between self and non-self. The reaction takes several days to be effectively triggered. It learns and memorizes the pathogen landscape.

Adaptive immune system

T cells destroy infected cells
to eradicate intracellular pathogens.
(Some bacteria, all viruses)



B cells secrete antibodies
to attack extracellular pathogens
(Most bacteria)



The colors of the receptors indicate specificity: each can bind to one specific antigen. Adaptive immunity can only attack targets that it has prepared for.

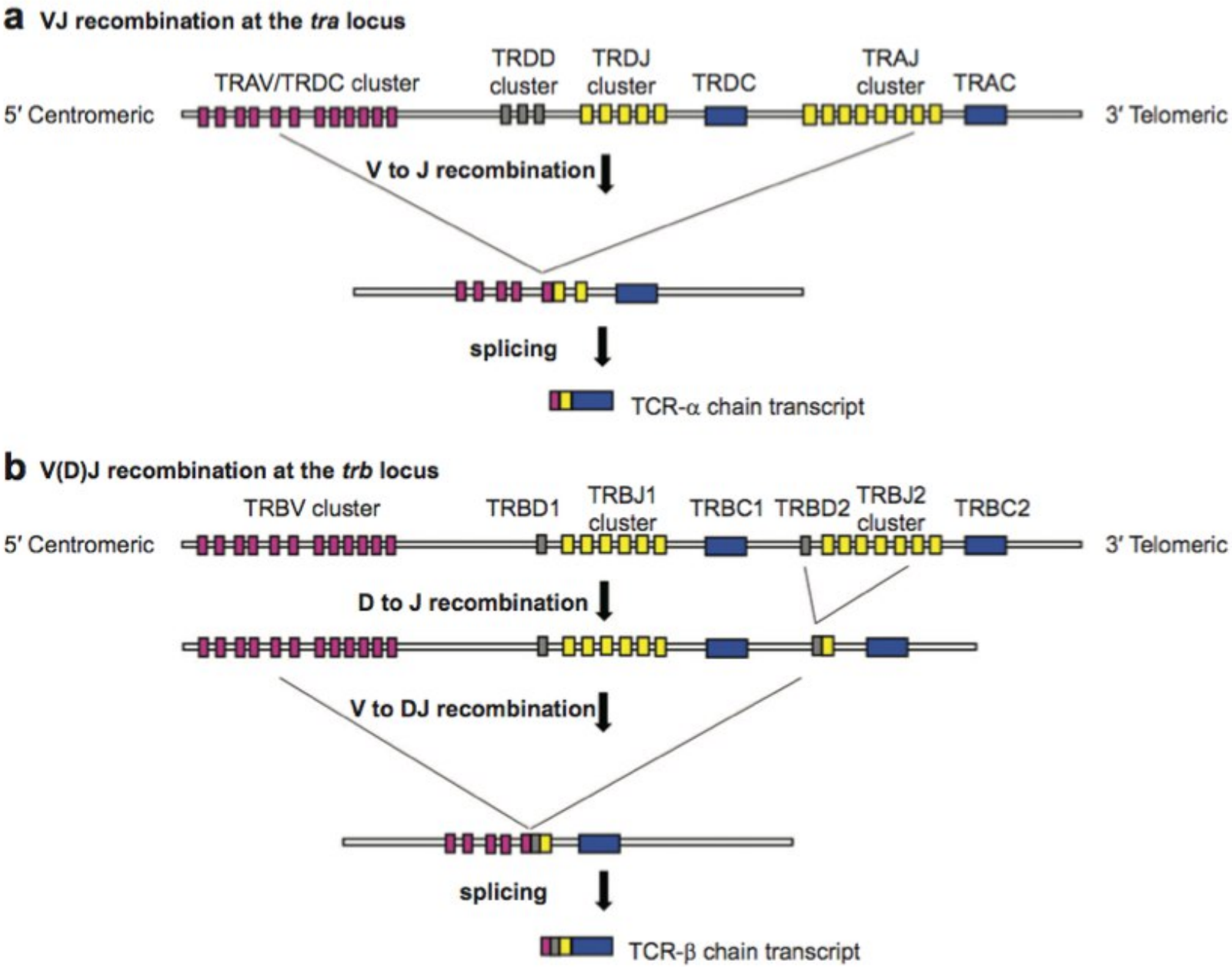
TCR chains

$\alpha\beta$ chain - ”classic” adaptive immunity (virus detection)

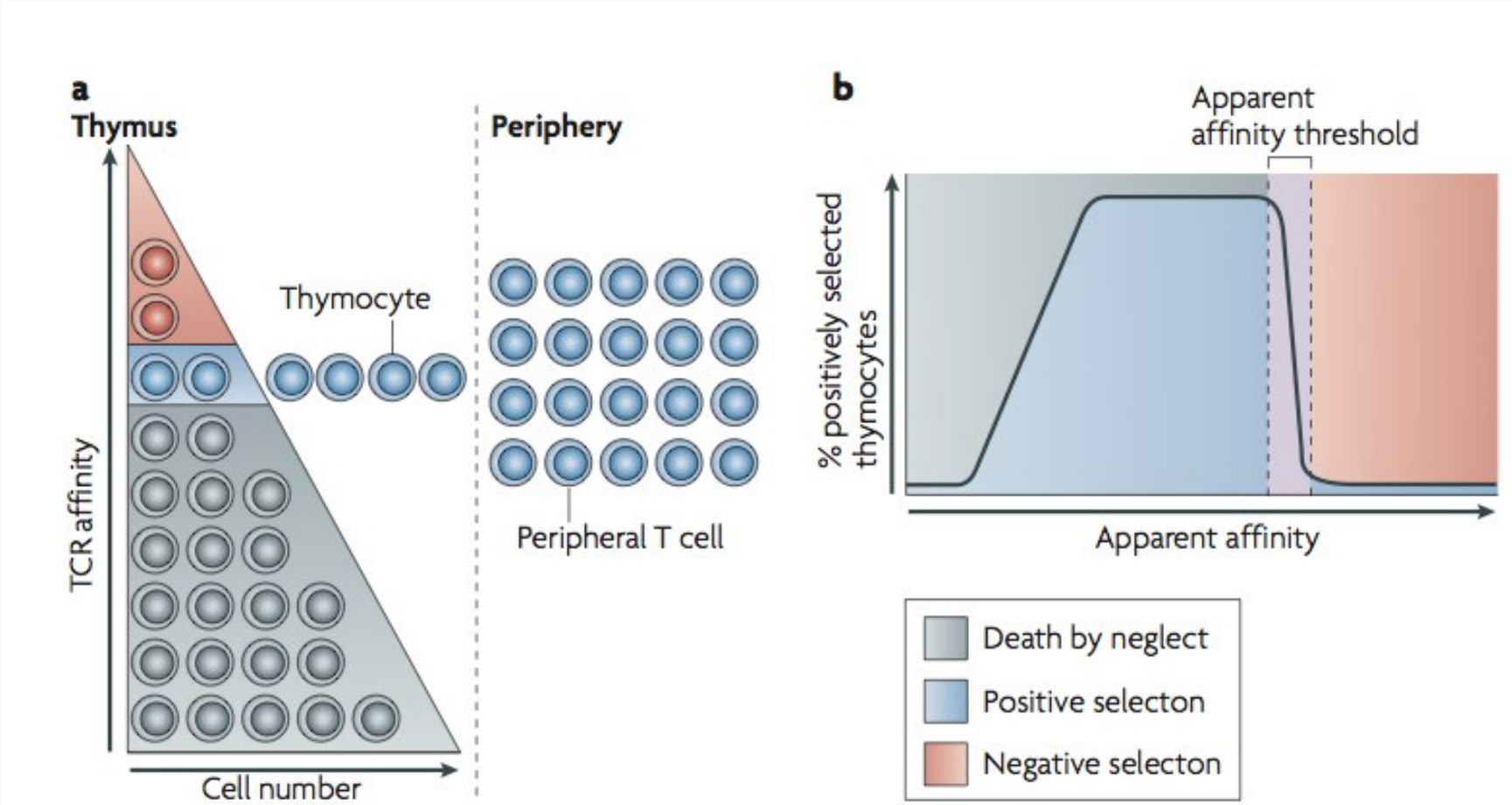
$\gamma\delta$ chain - terra incognita (phagocytosis, invariant cells)

Different generation processes!

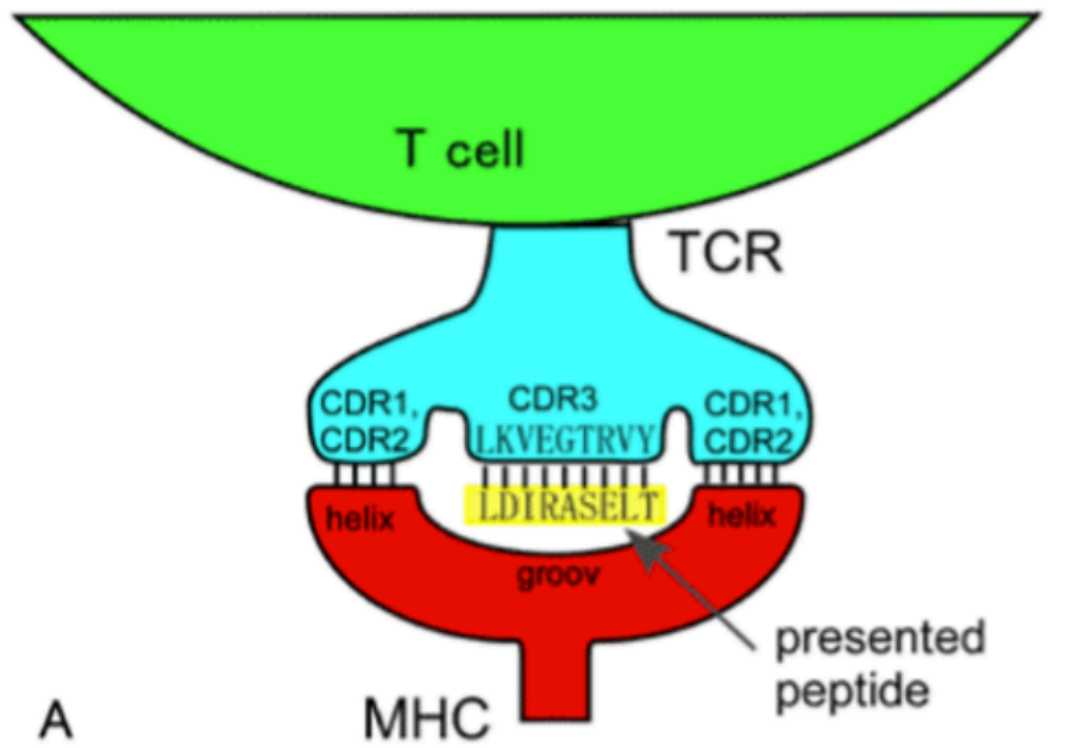
V(D)J recombination

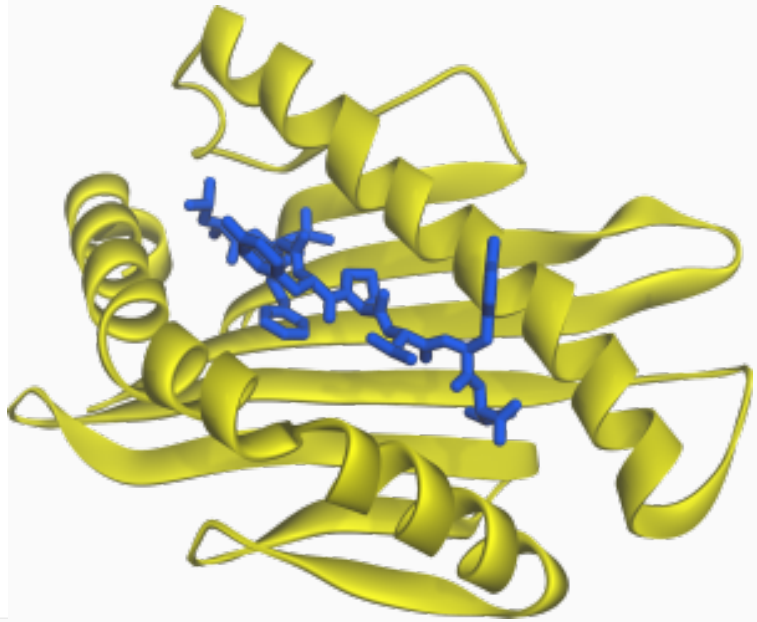


TCR selection



TCR:peptide:MHC interaction





TCR data example

Count	Proportion	CDR3.nucleotide.sequence	CDR3.amino.acid.sequence	V.gene	J.gene
9959.760753	7.416466e-02	TGTGCCAGCAGCCAAGCTCTAGCGGGAGCAGATACGC...	CASSQALAGADTQYF	TRBV4-2	TRBJ2-3
4425.389760	3.295335e-02	TGTGCCAGCAGCTTAGGCCCCAGGAACACCGGGGAGC...	CASSLGPRNTGELFF	TRBV13	TRBJ2-2
3890.686845	2.897173e-02	TGTGCCAGCAGTTATGGAGGGGCGGCAGATACGCAGT...	CASSYGGAADTQYF	TRBV12-4, TRBV12-3	TRBJ2-3
221.330500	1.648122e-03	TGCAGTGCTGGAGGGATTGAAACCTCCTACAATGAGCA...	CSAGGIETSYNEQFF	TRBV20-1	TRBJ2-1
1799.436602	1.339938e-02	TGTGCCAGCTCACCCATCTTAGGGGAGCAGTTCTTC	CASSPILGEQFF	TRBV18	TRBJ2-1
1316.984630	9.806834e-03	TGTGCCAGCAAAAAAGACAGGGACTATGGCTACACCTTC	CASKKDRDYGYTF	TRBV6-5	TRBJ1-2
2309.863250	1.720023e-02	TGTGCCAGCAGCCAACAGGGATCTGGAAACACCATATA...	CASSQQGSGNTIYF	TRBV7-2	TRBJ1-3
3339.582627	2.486797e-02	TGTGCCAGCAGTTTAGGTCTTCACTACGAGCAGTACTTC	CASSLGLHYEQYF	TRBV28	TRBJ2-7

Introduction to deep learning

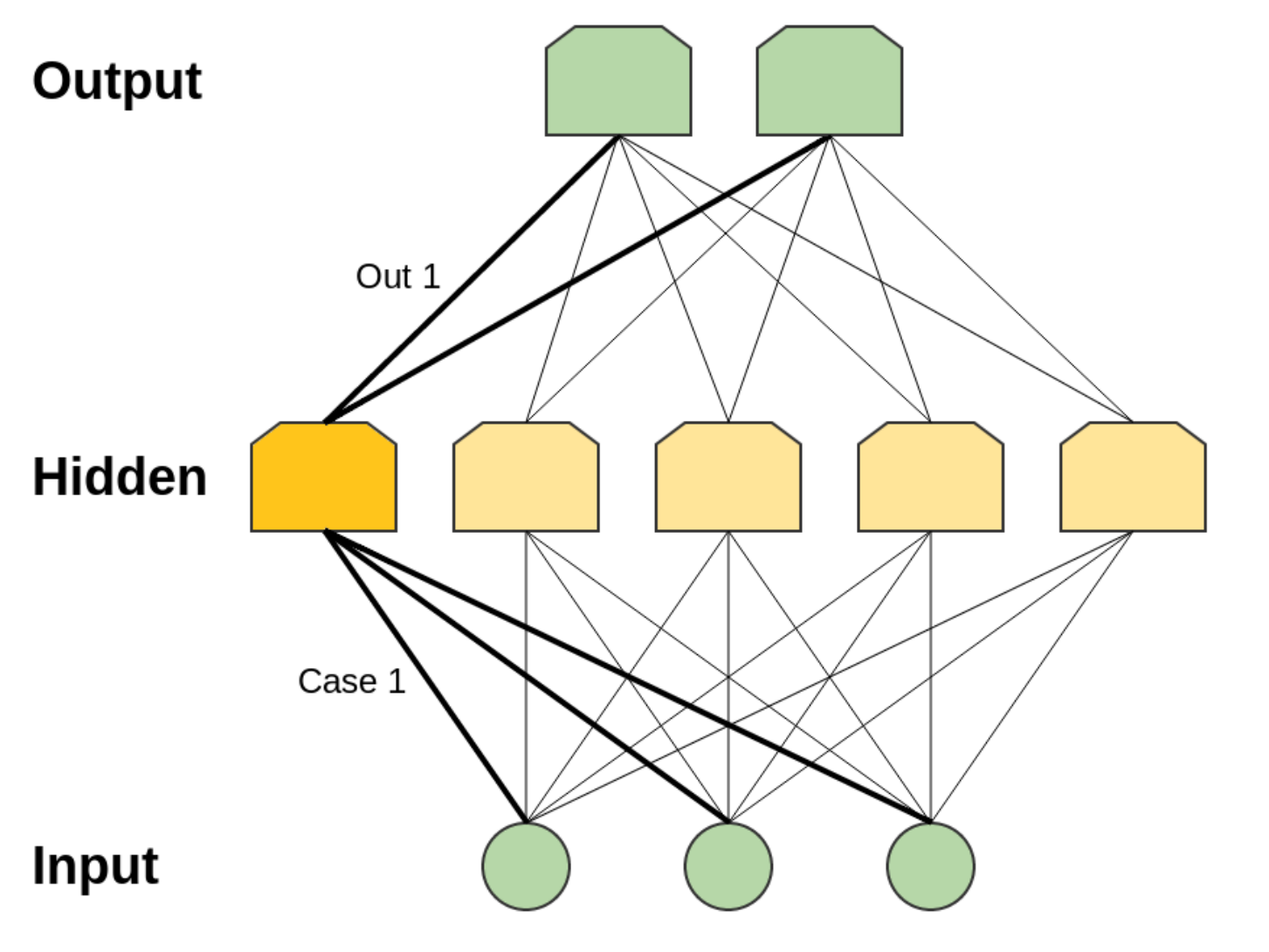
Deep network architecture ideas

Fully connected / dense networks (DNN)

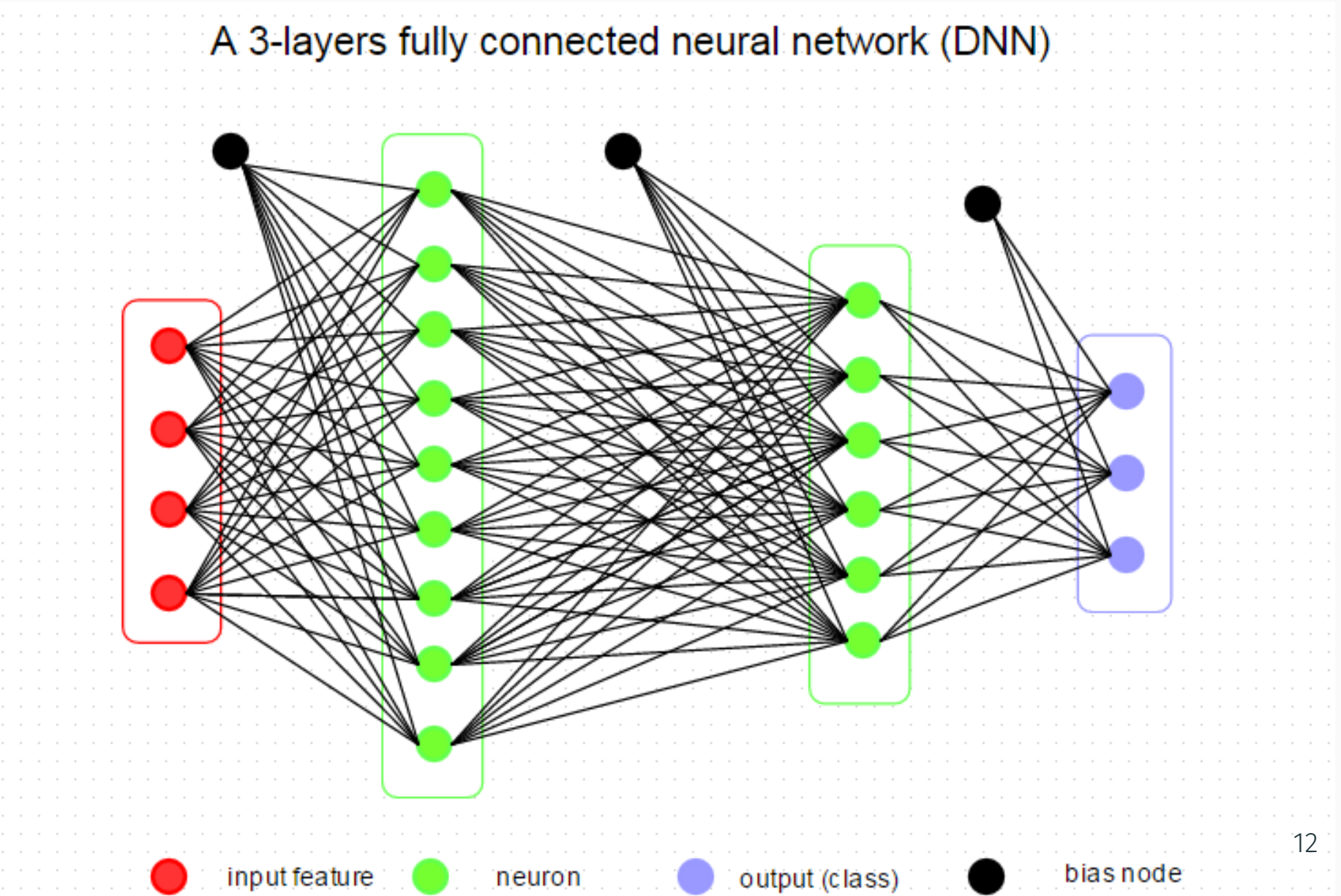
Convolutional neural networks (CNN)

Recurrent neural networks (RNN)

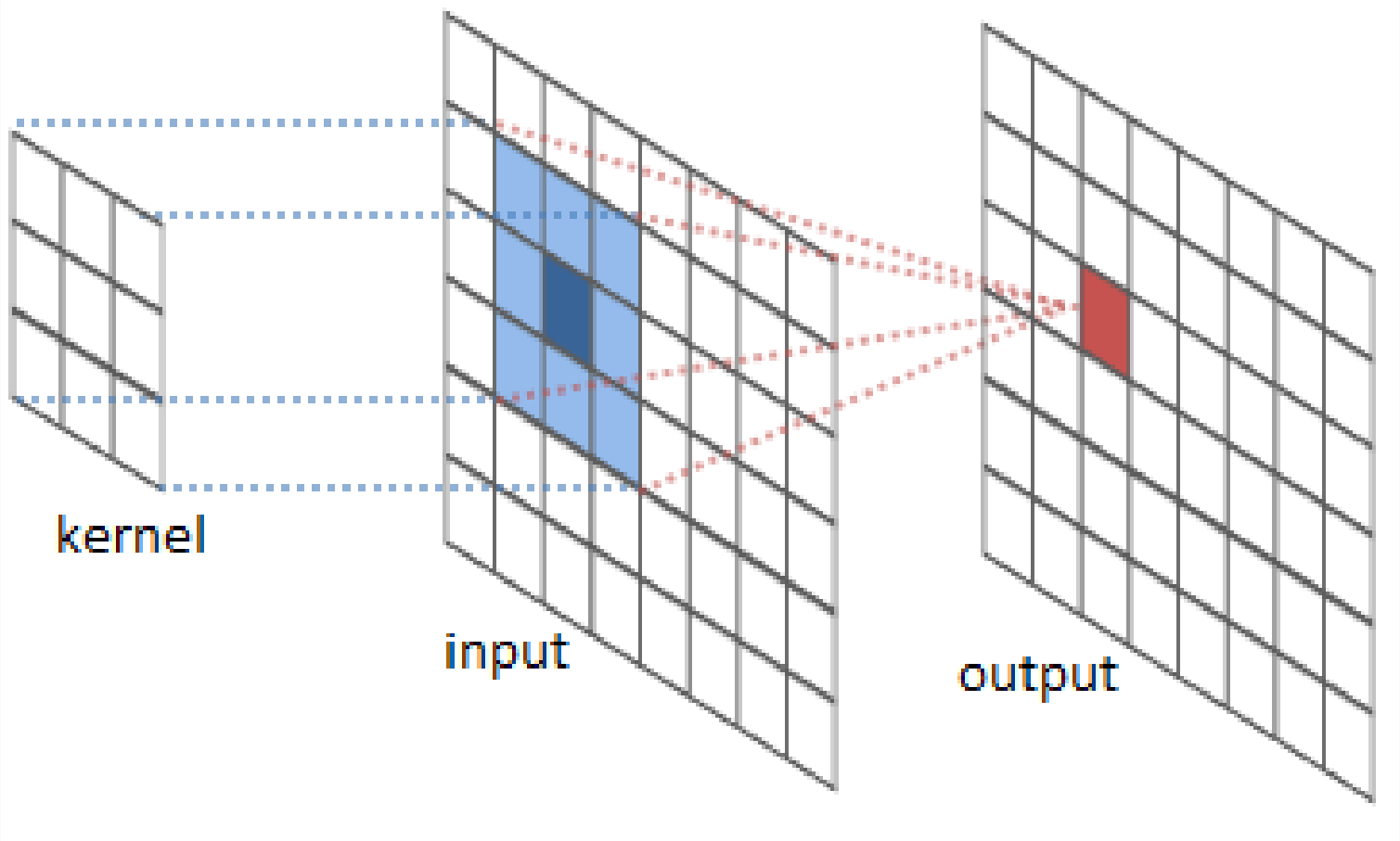
Fully connected networks 1



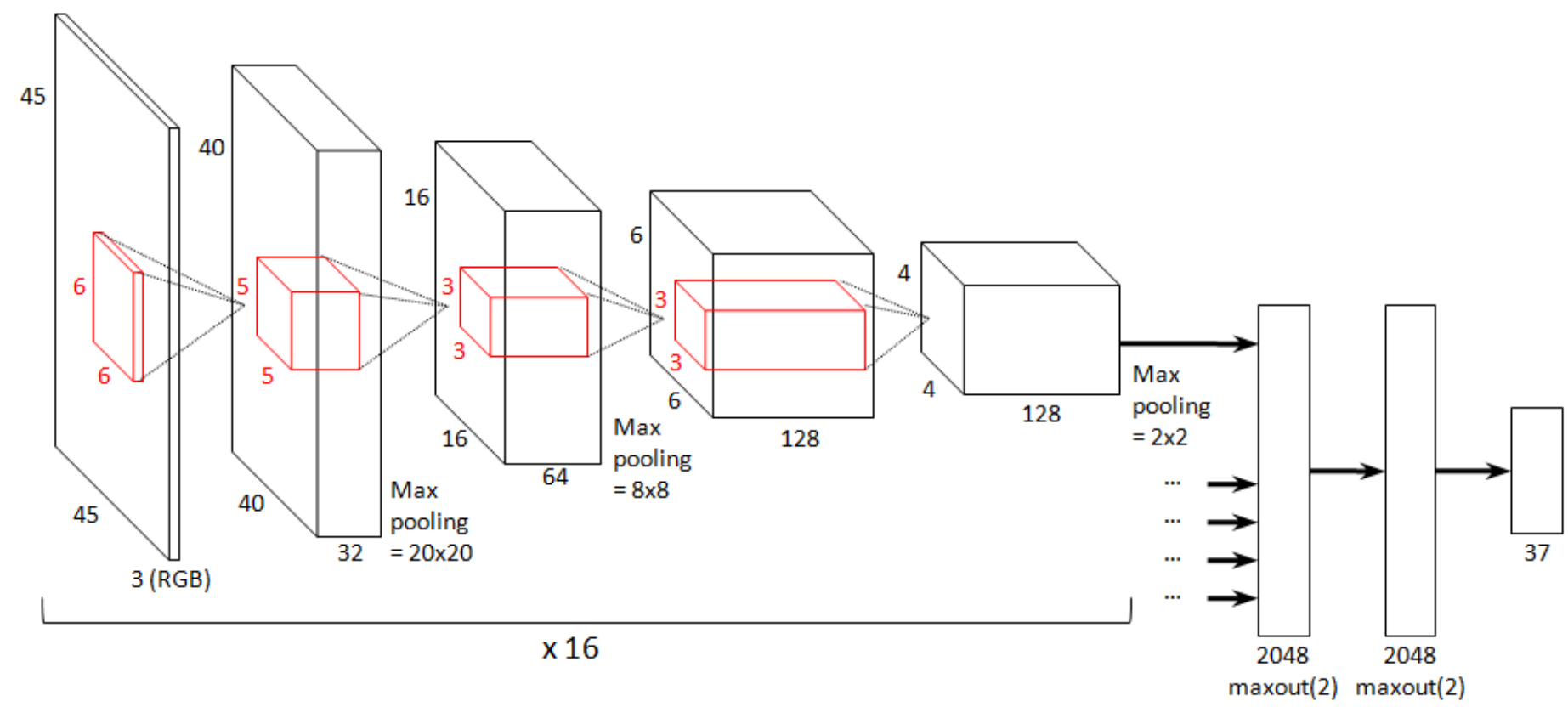
Fully connected networks 2



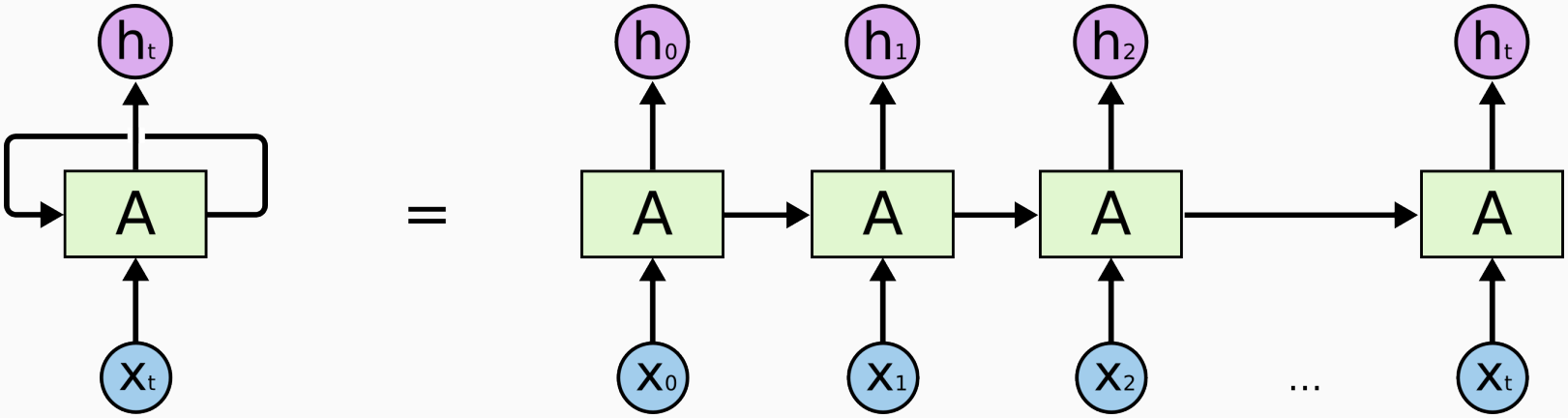
Convolutions



Convolutional neural networks



Recurrent neural networks



MHC:peptide binding affinity prediction

Problem

Prediction of strong / weak binders (immunotherapy, etc.)

Data

140,000 pairs of MHC-peptide for training

30,000 pairs of MHC-peptide for testing

species	mhc	peptide_length	cv	sequence	inequality	meas
cow	BoLA-HD6	9	TBD	ALFYKDGKL	=	1.0
cow	BoLA-HD6	9	TBD	ALYEKKLAL	=	1.0
cow	BoLA-HD6	9	TBD	AMKDRFQPL	=	4.52170583277
cow	BoLA-HD6	9	TBD	AQRELFFTL	=	1.0
cow	BoLA-HD6	9	TBD	FMKVKFEAL	=	1.57674703262
cow	BoLA-HD6	9	TBD	FQHERLGQF	=	1.0
cow	BoLA-HD6	9	TBD	FQRAIMNAM	=	1.0
cow	BoLA-HD6	9	TBD	GQFLSFASL	=	1.0
cow	BoLA-HD6	9	TBD	GQFNRYAAM	=	1.0

NetMHCpan

Paper: just google ”netMHCpan paper”

Features:

- Onehot encoding
- Blosum encoding
- Lengths
- Indels

Pseudo-sequences – pan-allele approach

Model: DNN with 60 hidden neurons

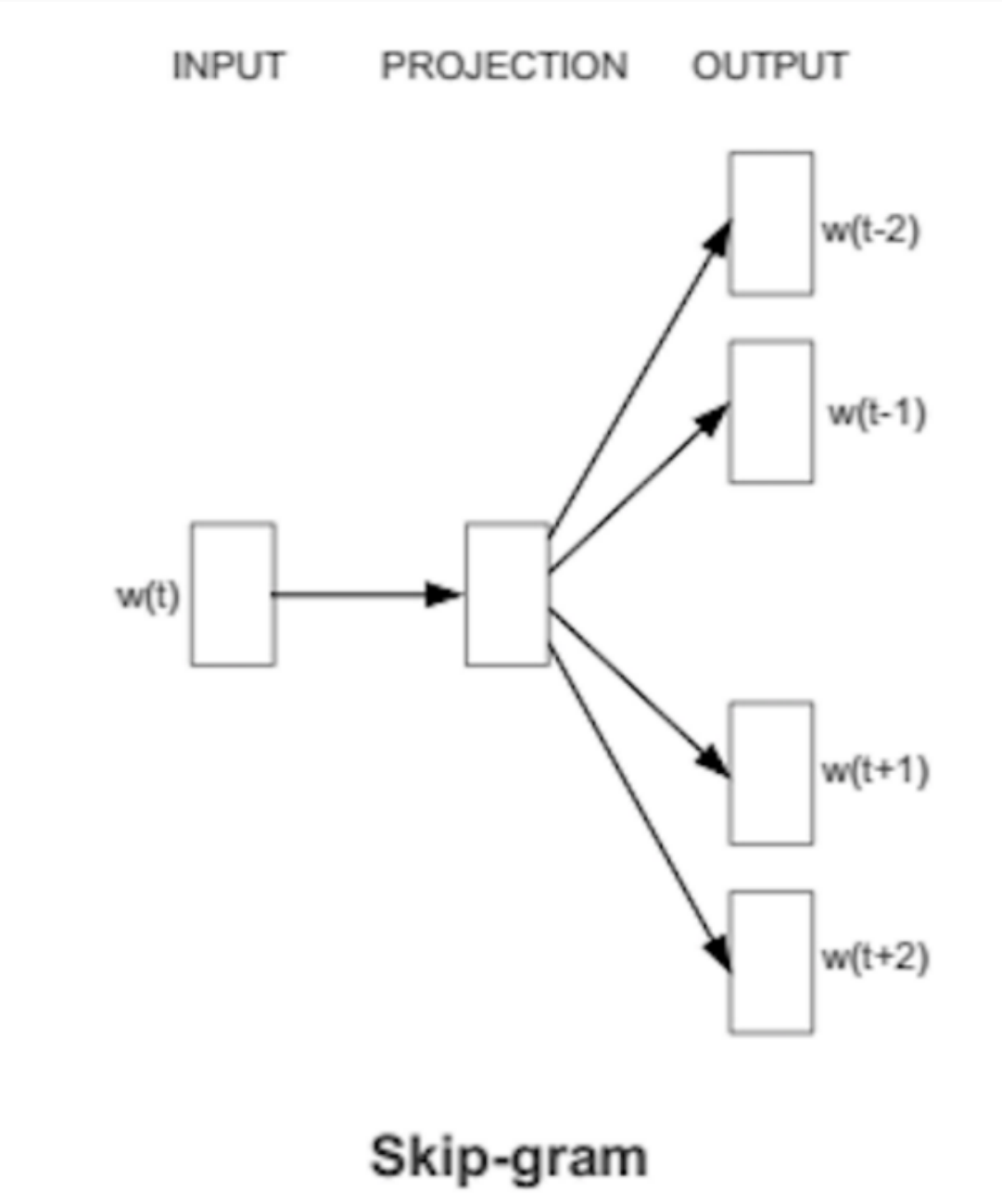
F1 score - 0.8

$$F1 = 2 * precision * recall/(precision + recall)$$

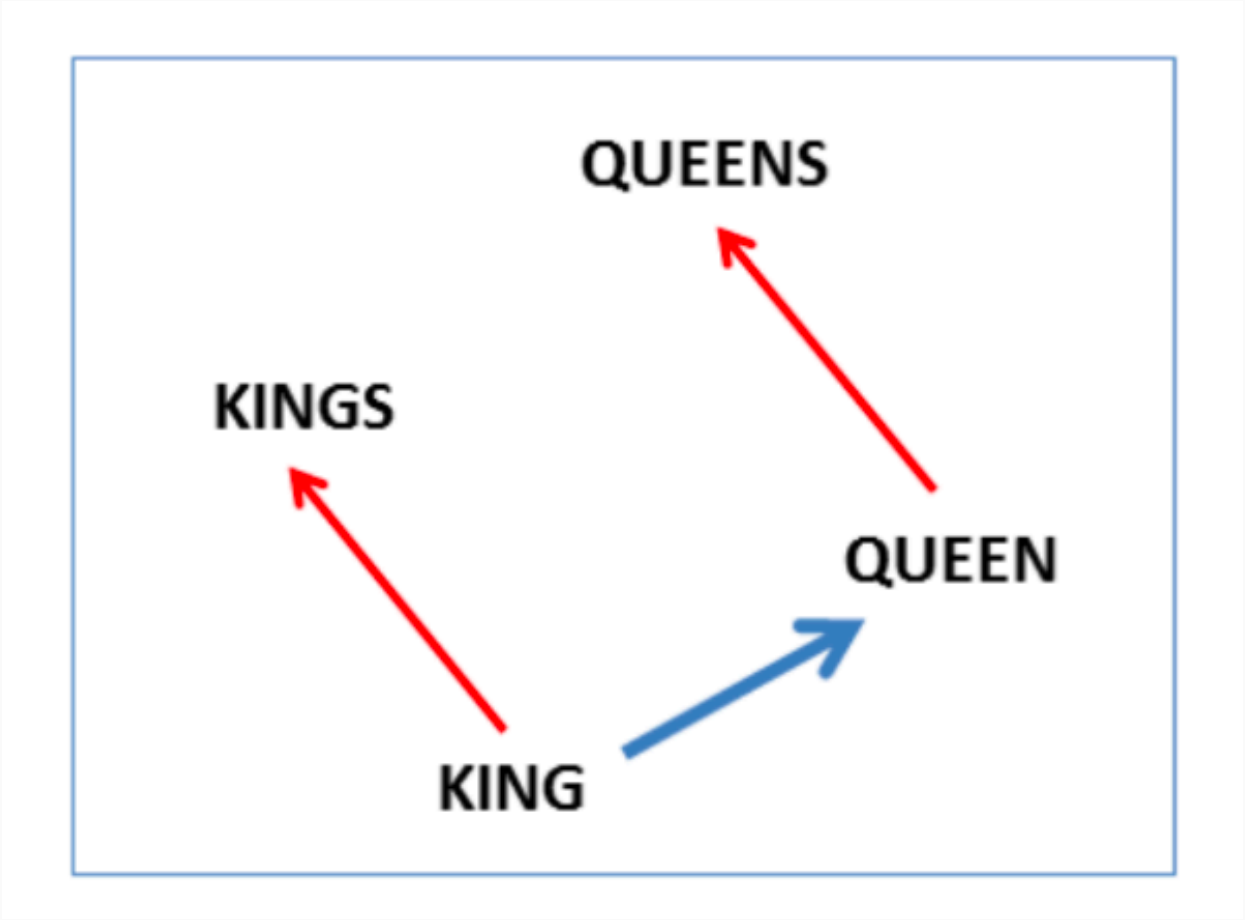
$$precision = TP/(TP + FP)$$

$$recall = TP/(TP + FN)$$

word2vec

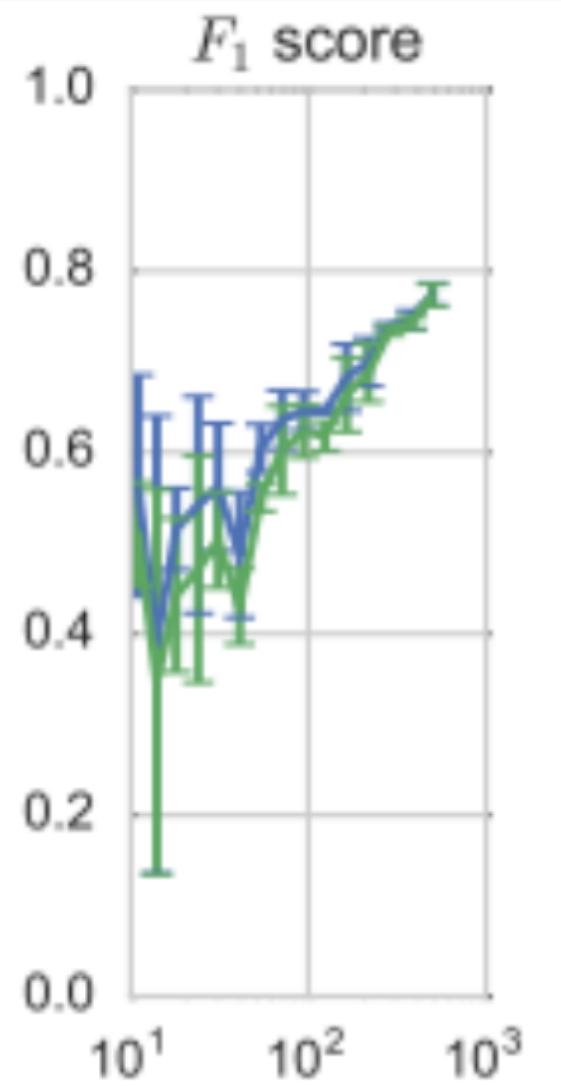


word2vec vectors



Imputation

MICE: average multiple imputations generated using Gibbs sampling from the joint distribution of columns.



mhcflurry

Paper: http://biorxiv.org/content/biorxiv/early/2016/05/22/054775.full.pdf

Features:

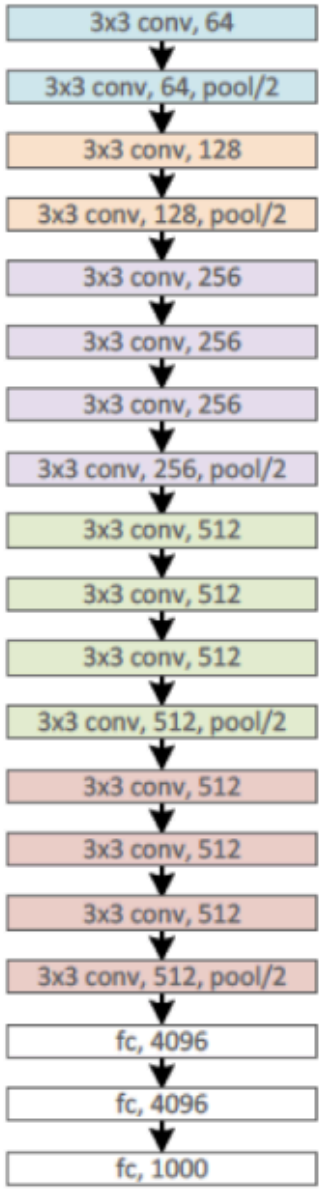
- Embeddings (per-pseudo-sequence!)

Model: DNN with 60 neurons

F1 score - 0.79

ResNet - old networks

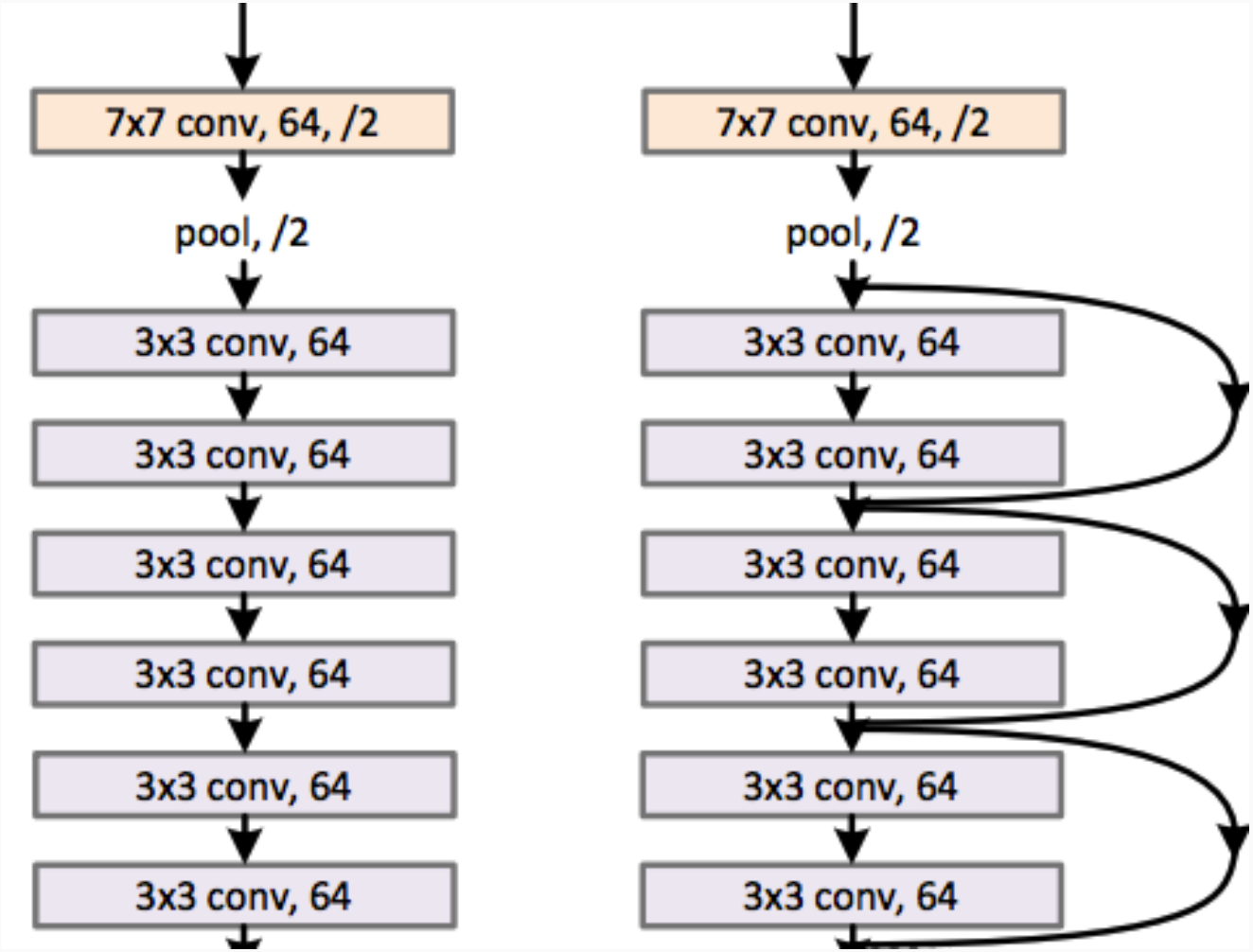
VGG, 19 layers
(ILSVRC 2014)



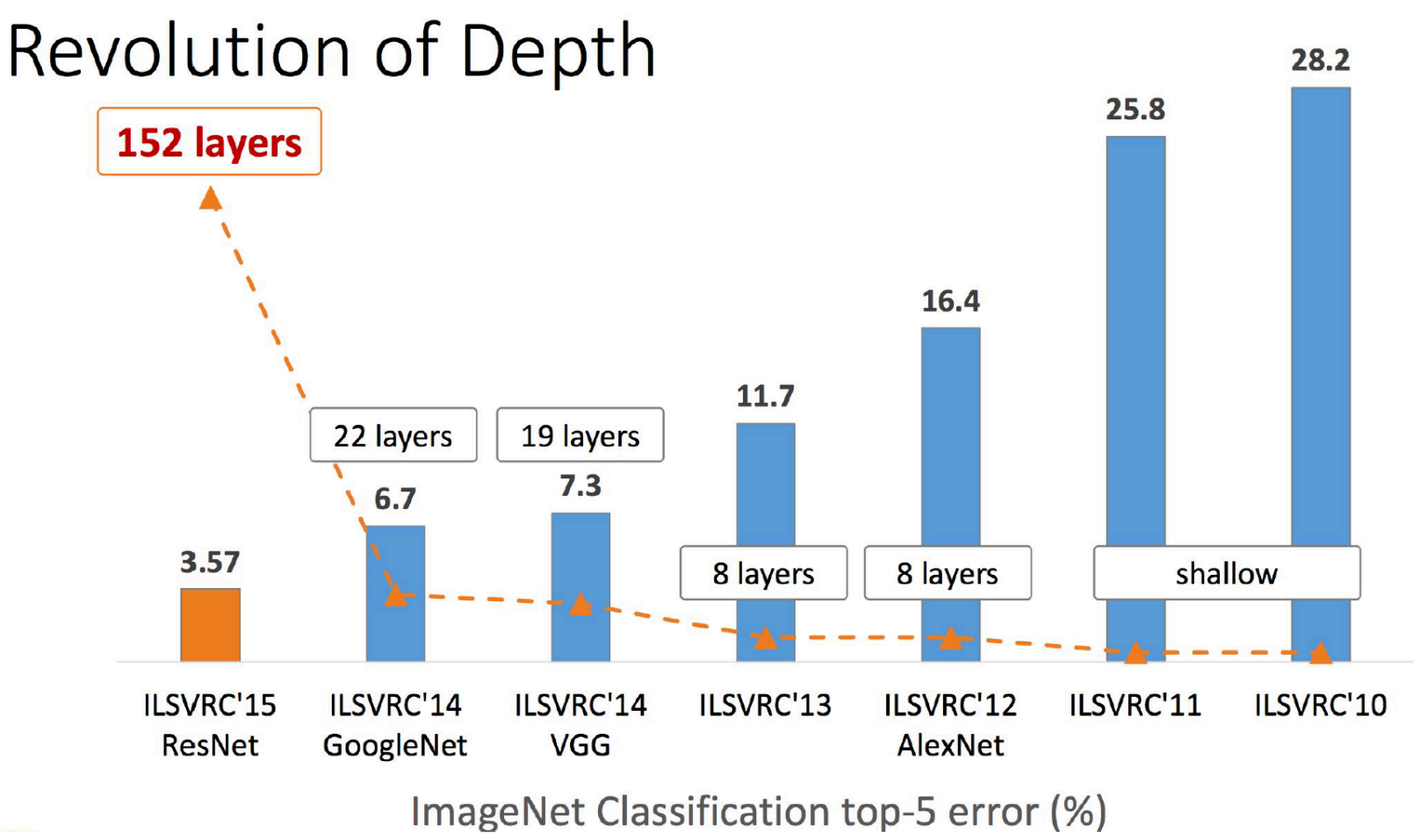
ResNet - old networks' problems

- Gradient vanishing
- Large number of parameters
- Shallowness

ResNet - proposed model

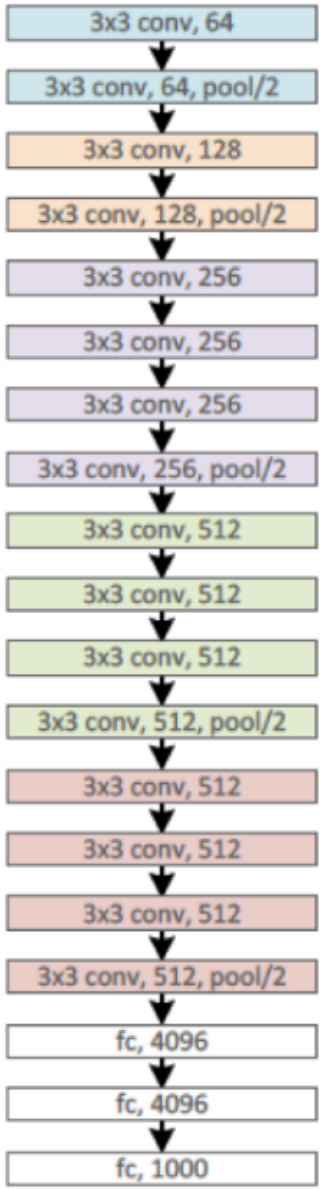


ResNet - results



ResNet - old networks

VGG, 19 layers
(ILSVRC 2014)

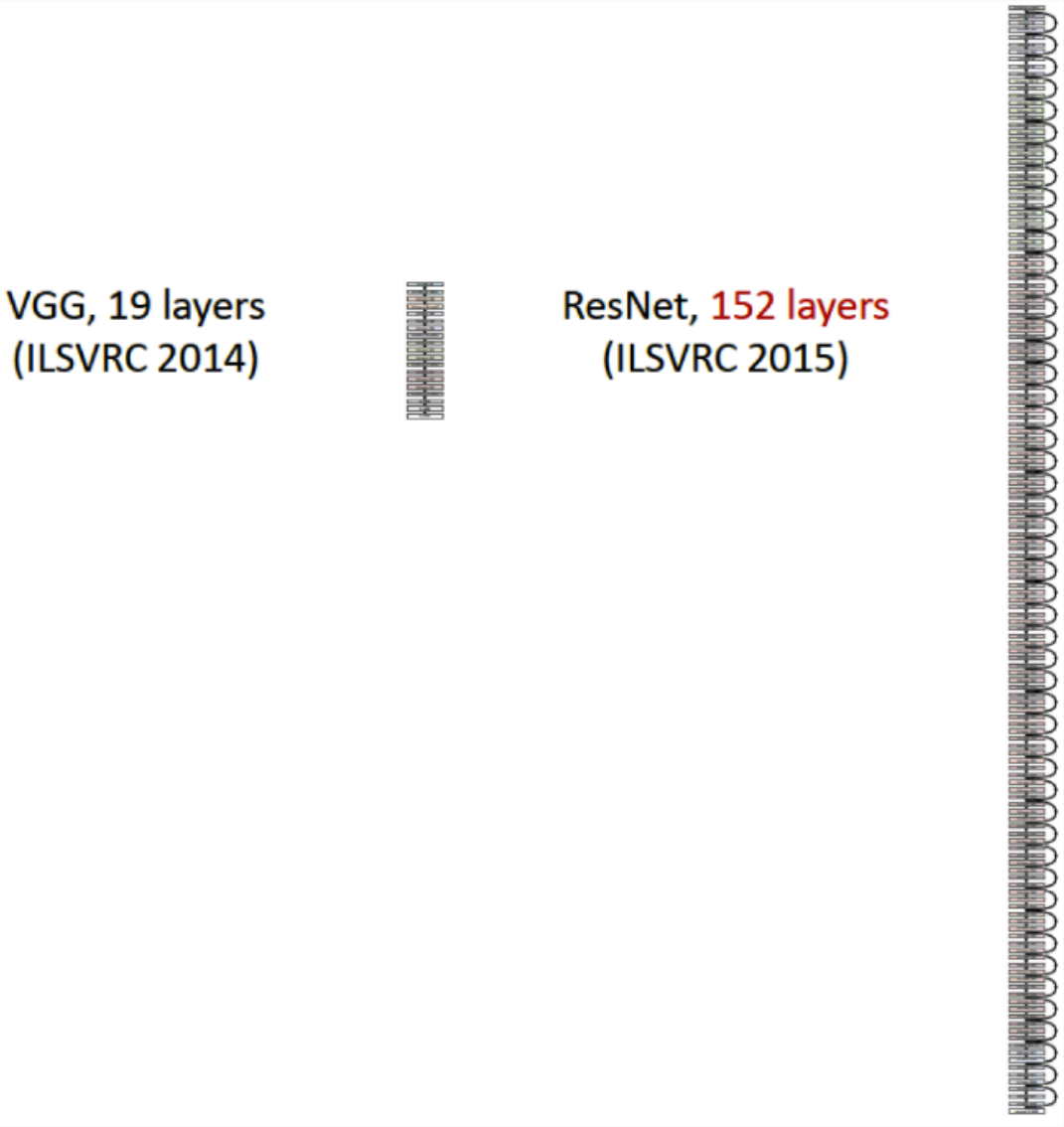


ResNet - current deep networks

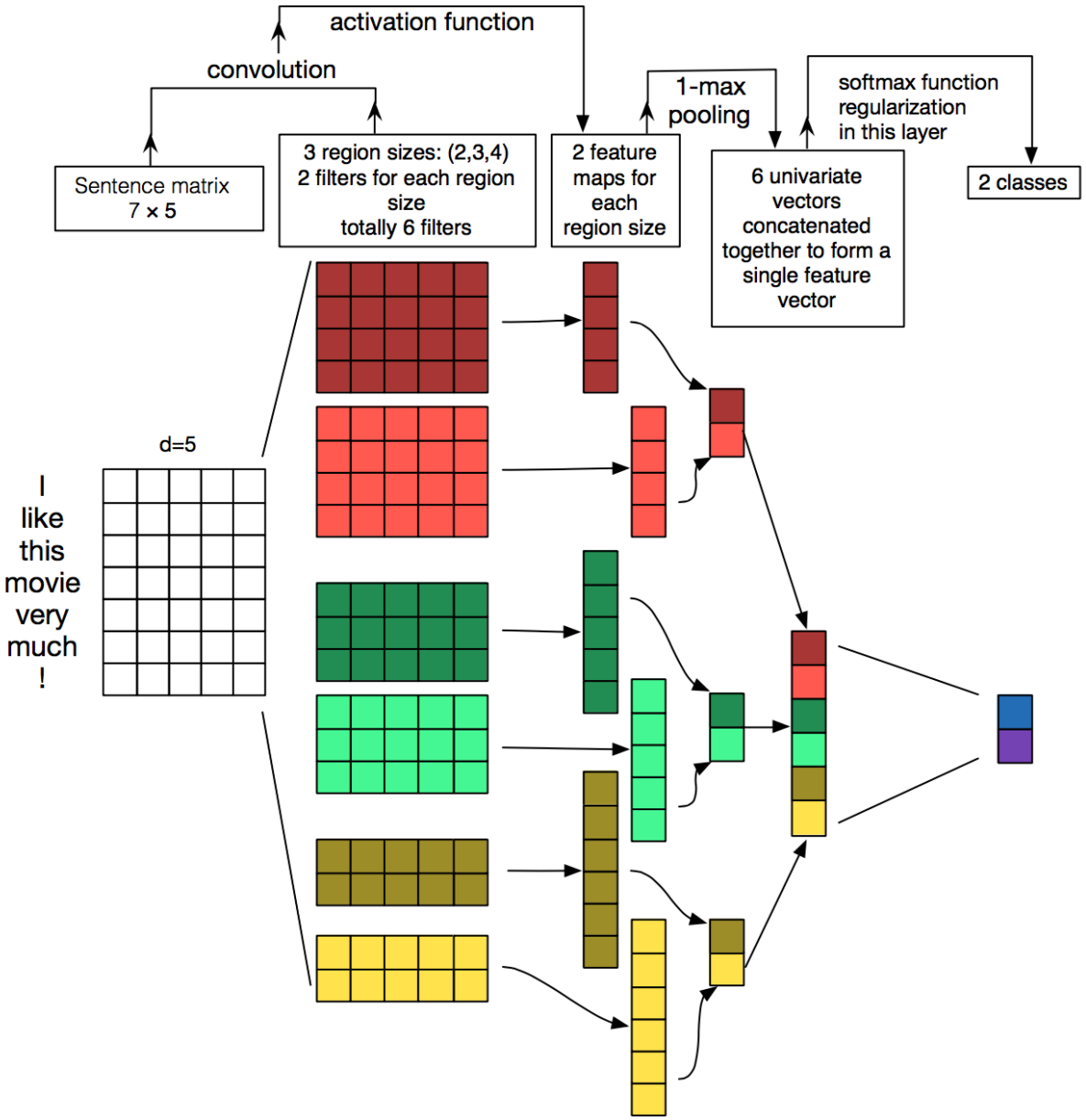
VGG, 19 layers
(ILSVRC 2014)

ResNet, 152 layers
(ILSVRC 2015)

ResNet - current deep networks



CNN for NLP


activation function
convolution
1-max pooling
softmax function regularization in this layer
Sentence matrix 7 × 5
3 region sizes: (2,3,4) 2 filters for each region size totally 6 filters
2 feature maps for each region size
6 univariate vectors concatenated together to form a single feature vector
2 classes
d=5
I
like
this
movie
very
much
!

Our approach

Our approach - results

- F1 0.81 (on a subset of the dataset)
- Global models – prediction of binding affinities for unseen MHCs (mean F1 0.72)
- Better models for the per-pseudo-sequence approach.

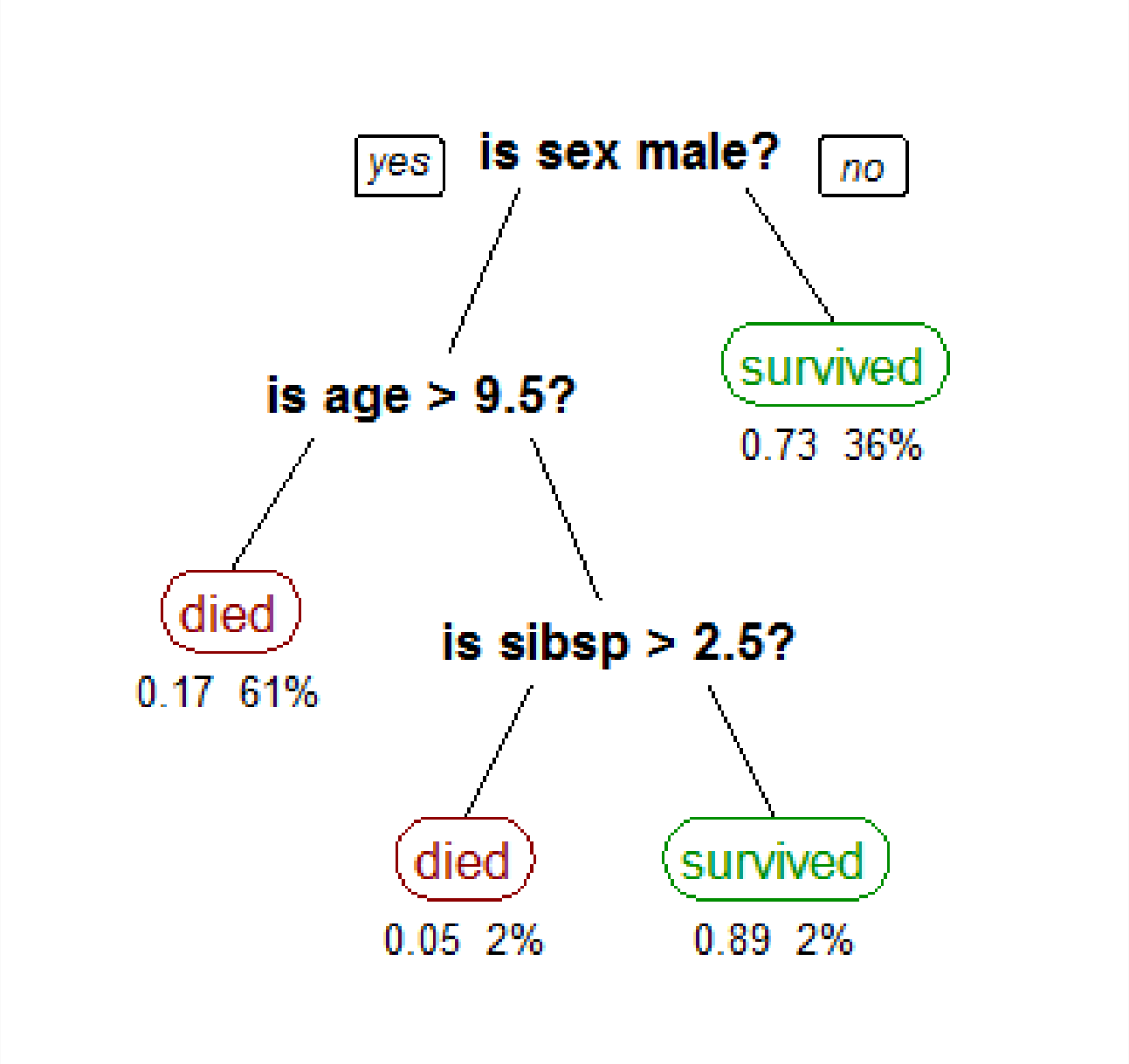
TCR-peptide binding prediction

Problem

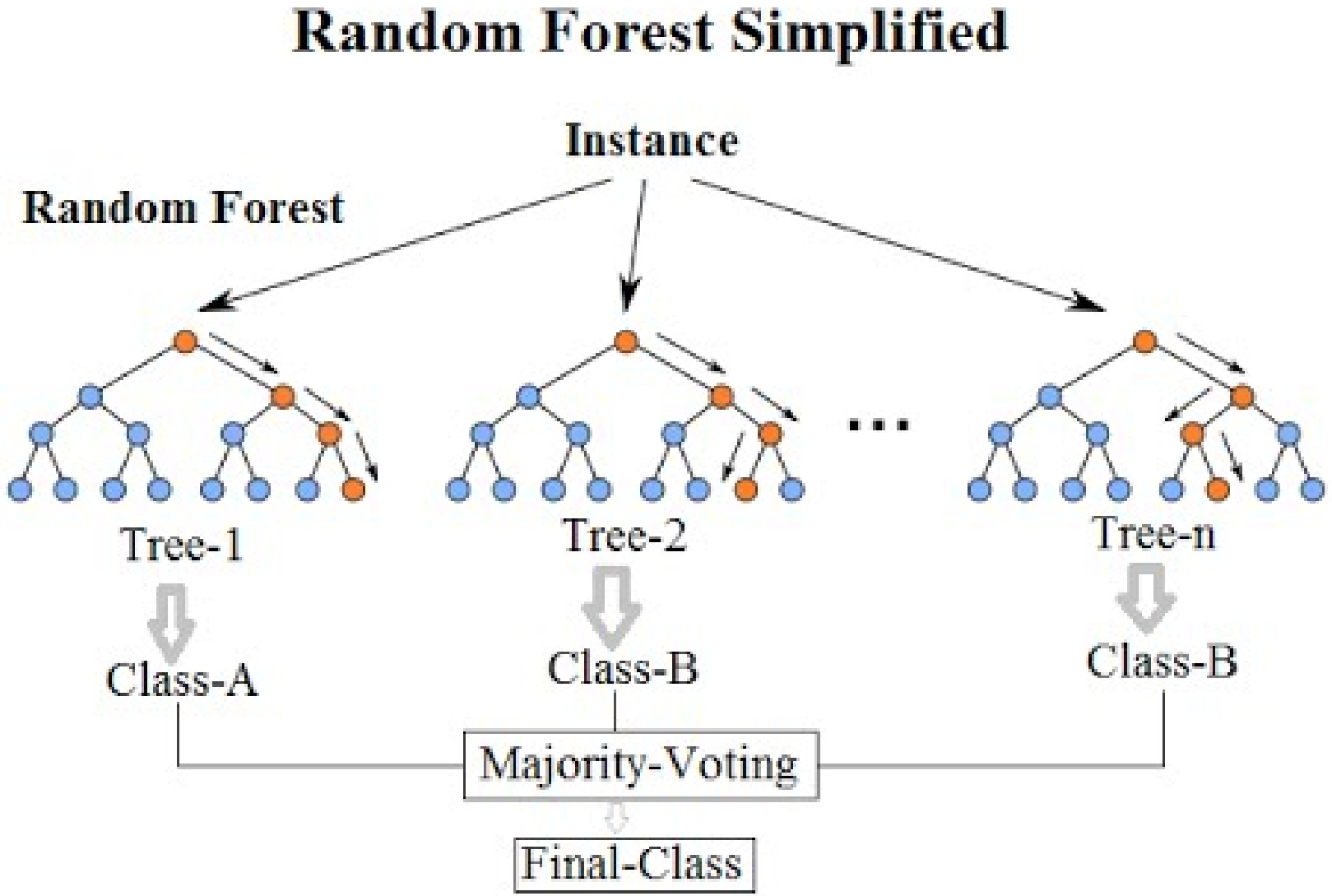
Paper:
http://biorxiv.org/content/early/2017/03/20/118539.full.pdf+htm

Immunogenicity prediction.

Decision tree (Titanic survival prediction



Random forest



Methods and results

Features:

- One-hot encoding of V/J
- The average CDR3 basicity, hydrophobicity, helicity, isoelectric point
- The asolute count of each individual amino acid in the CDR3 sequence
- The total mass of the 258 amino acids in the CDR3 sequence
- Numerical features encoding individual amino acid basicity, hydrophobicity, helicity, 269 isoelectric point, and mutation stability were also created for each position

Accuracy: 75.90%

Analysis of feature importances

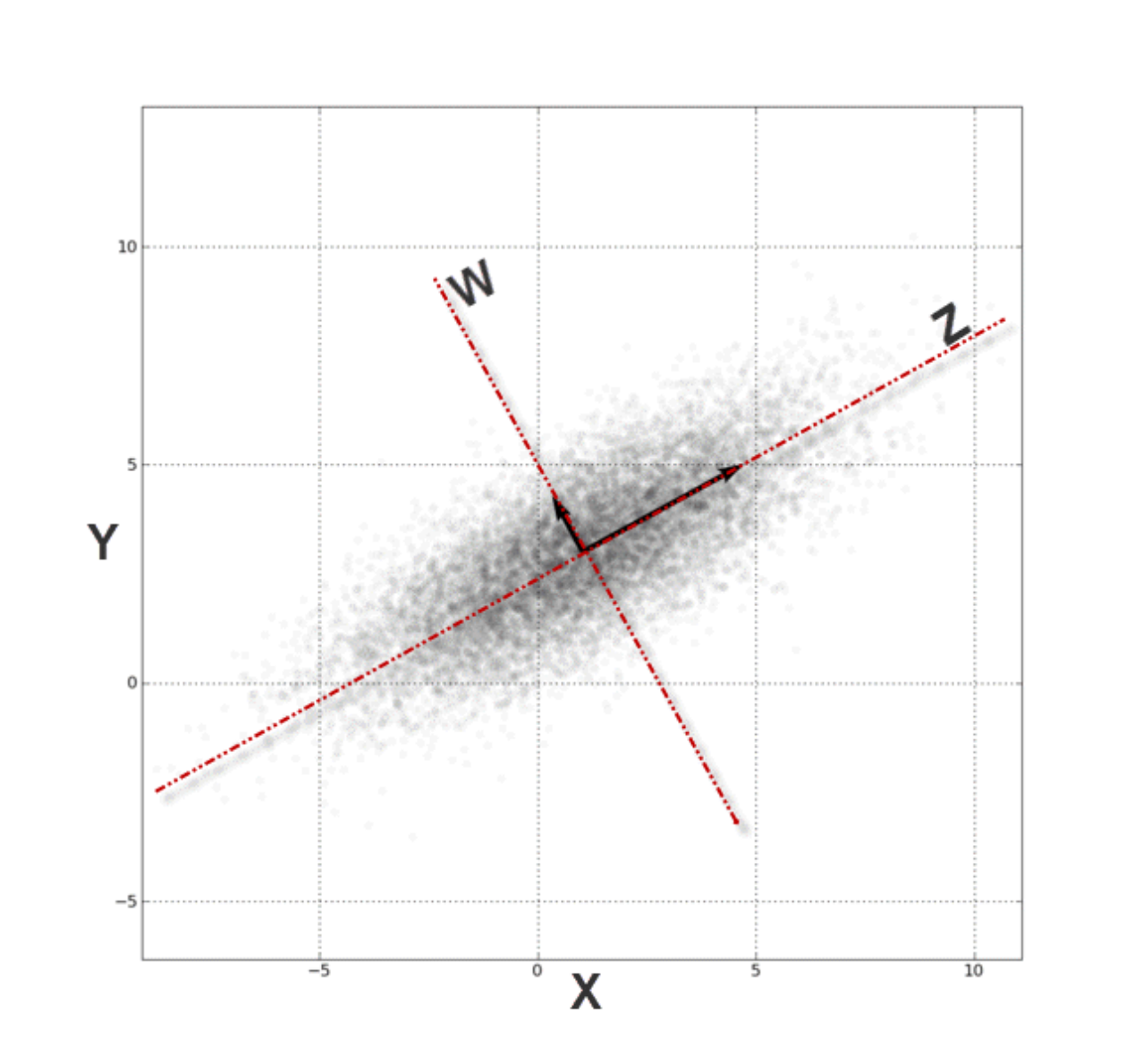
TCR CD4/CD8 classification

Problem

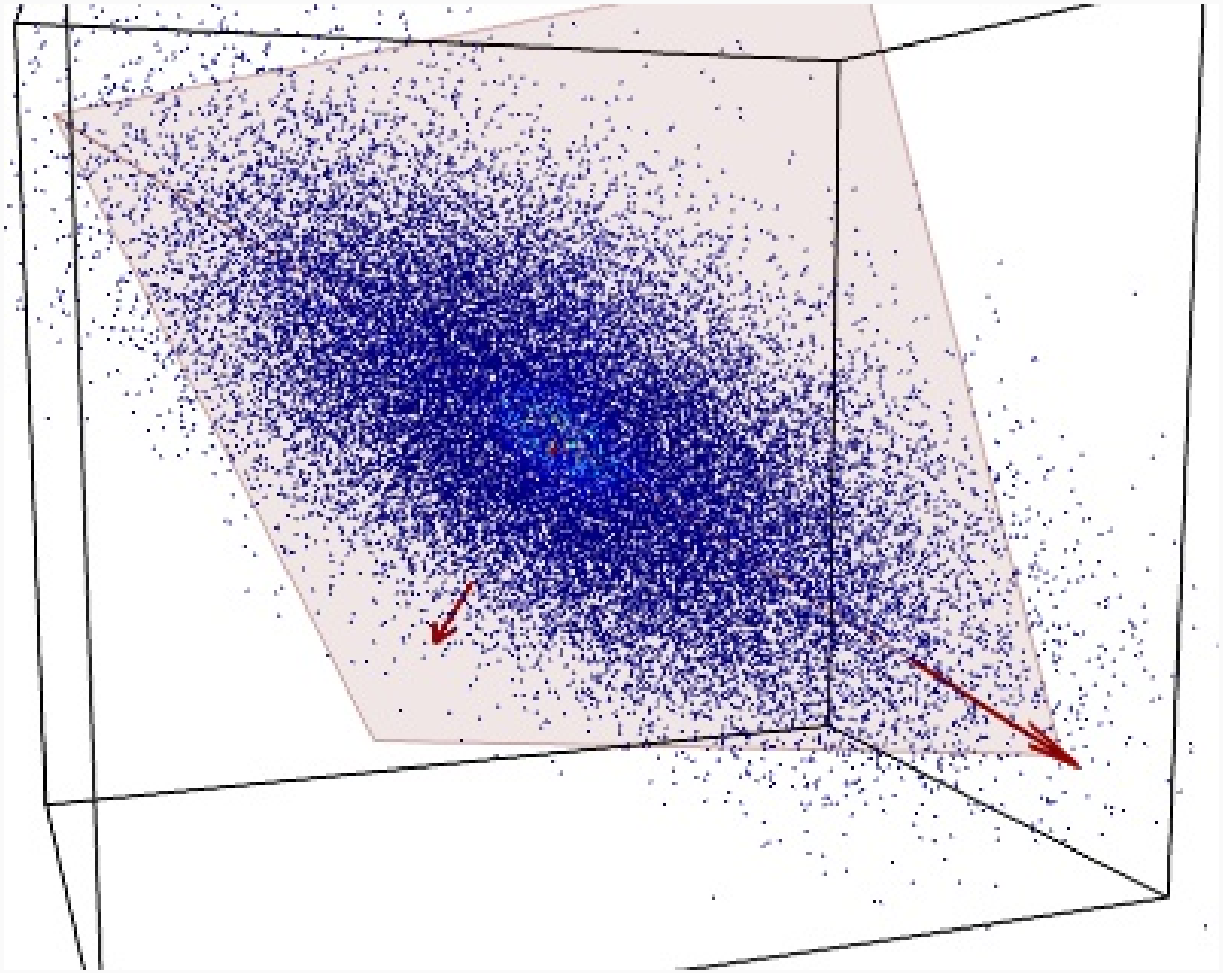
Paper: http://www.jleukbio.org/content/99/3/505.short

In-silico detection of CD4 / CD8 TCRs. Exploratory pre-analysis.

PCA 2D



PCA 3D



Kidera / Atchley factors

PCA on biophysical properties, explains ~80% variability in the data

amino.acid	f1	f2	f3	f4	f5	f6	f7	f8	f9	f10
A	-1.56	-1.67	-0.97	-0.27	-0.93	-0.78	-0.2	-0.08	0.21	-0.48
R	0.22	1.27	1.37	1.87	-1.7	0.46	0.92	-0.39	0.23	0.93
N	1.14	-0.07	-0.12	0.81	0.18	0.37	-0.09	1.23	1.1	-1.73
D	0.58	-0.22	-1.58	0.81	-0.92	0.15	-1.52	0.47	0.76	0.7
C	0.12	-0.89	0.45	-1.05	-0.71	2.41	1.52	-0.69	1.13	1.1
Q	-0.47	0.24	0.07	1.1	1.1	0.59	0.84	-0.71	-0.03	-2.33
E	-1.45	0.19	-1.61	1.17	-1.31	0.4	0.04	0.38	-0.35	-0.12
G	1.46	-1.96	-0.23	-0.16	0.1	-0.11	1.32	2.36	-1.66	0.46
H	-0.41	0.52	-0.28	0.28	1.61	1.01	-1.85	0.47	1.13	1.63

Methods and results

- CDR3 to kmers, kmers to Atchley factors
- Support Vector Machines classifier (different lengths? accuracy?)

TCR repertoire comparison using high-dimensional features

Problem

Paper:
http://biorxiv.org/content/early/2017/04/20/128025

Comparison of repertoires of TCRs and detection the subrepertoires with the most contribution to the inter-sample differences in-silico.

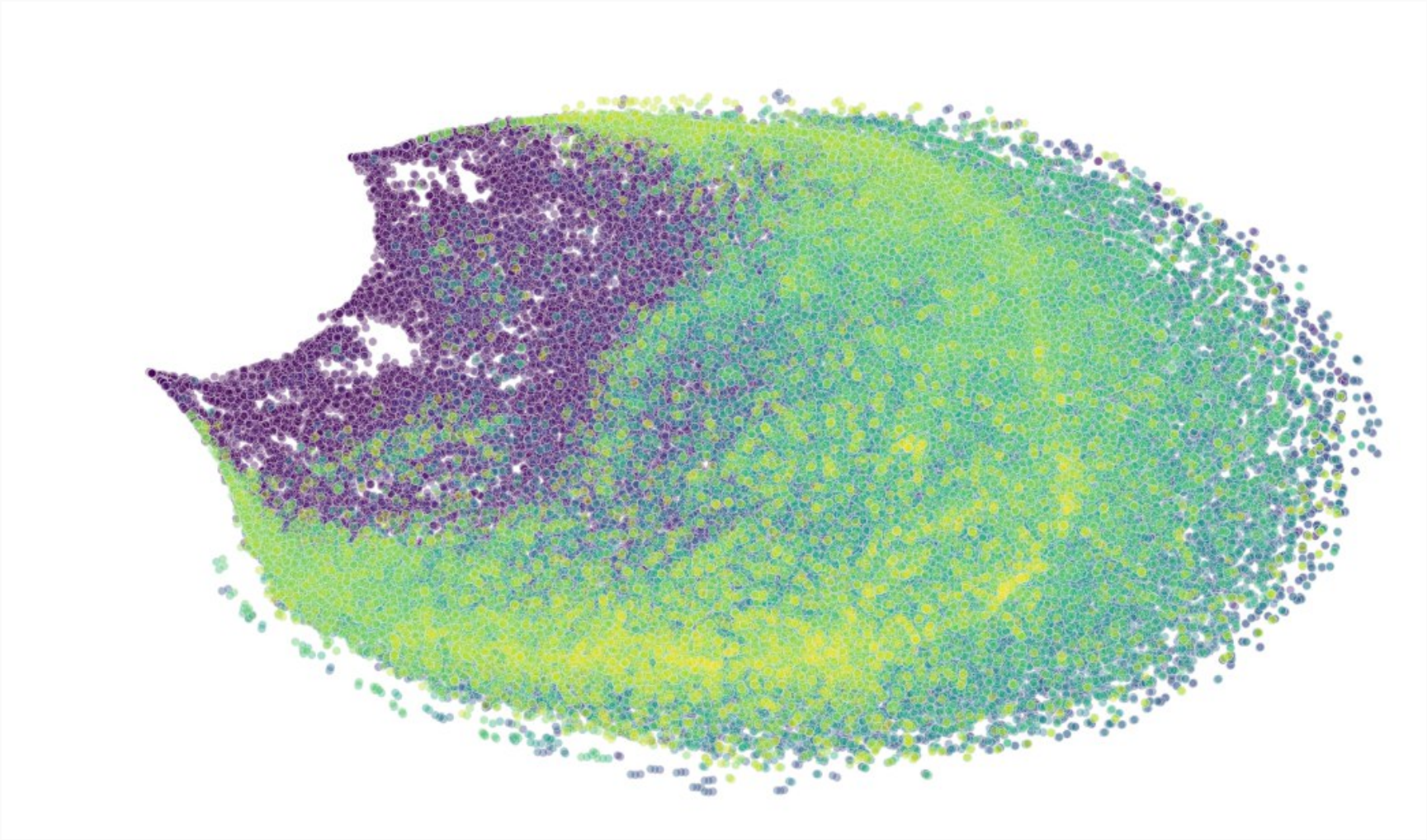
Data

- 8 repertoires
- Repertoire - table with CDR nuc/aa sequence, V gene, J gene, abundance columns.

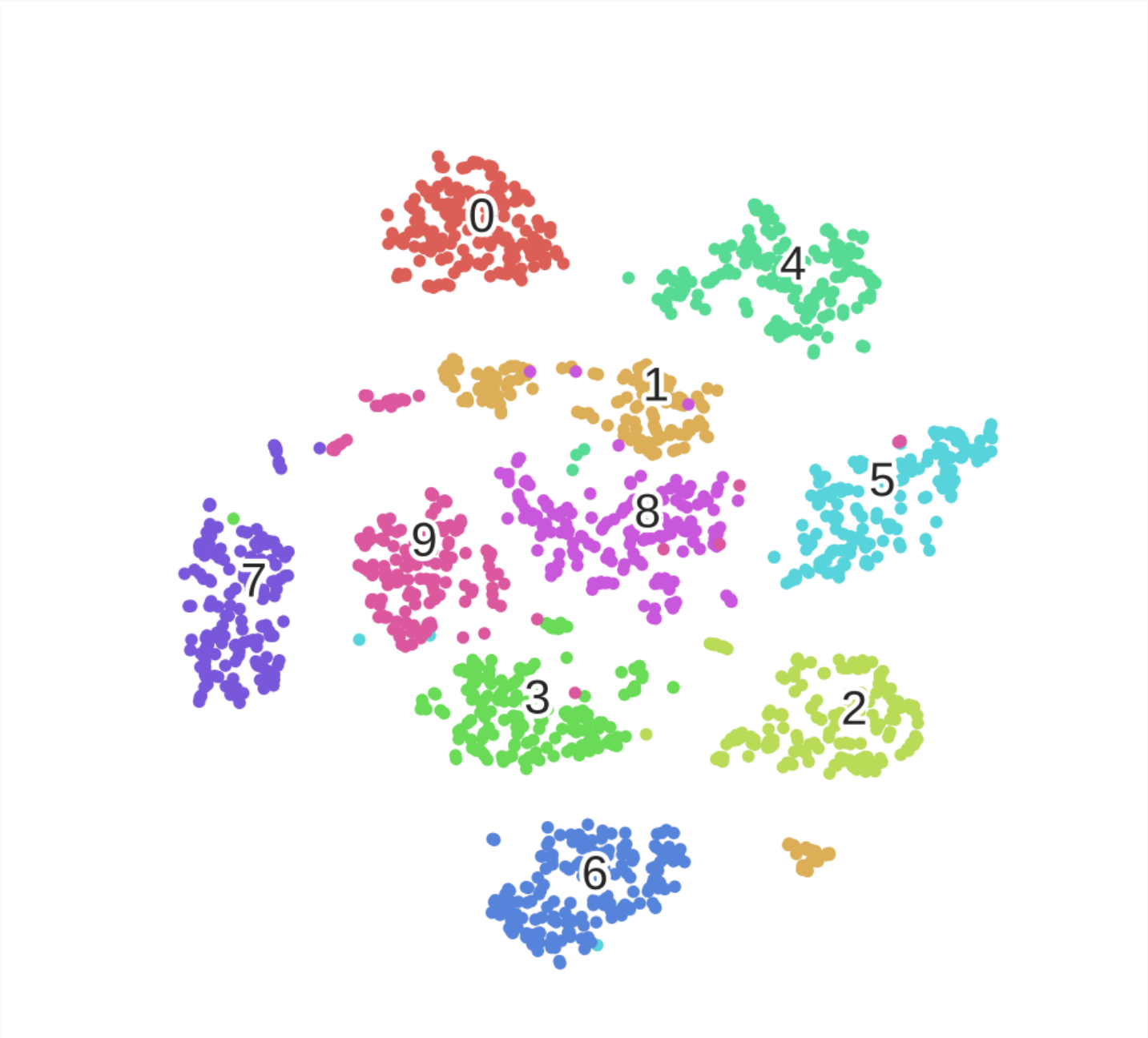
t-SNE

1. Construct a probability distribution over the dataset in such a way that similar objects have a high probability of being ”picked”.
2. Define a similar probability distribution over the points in the low-dimensional map (2-dimensional), and minimize the Kullback–Leibler divergence between the two distributions with respect to the locations of the points in the map.

t-SNE on peptides



t-SNE on MNIST



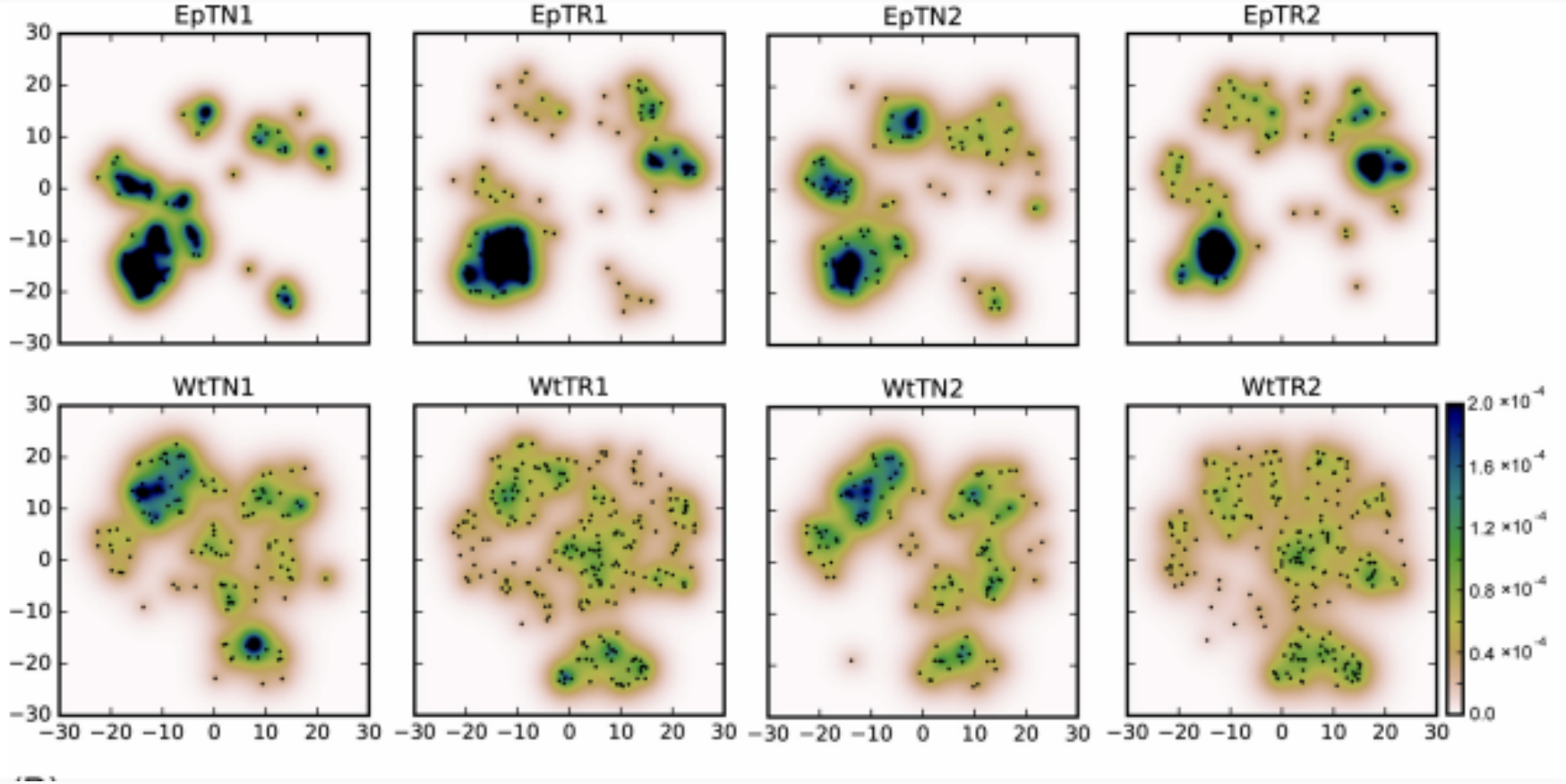
Methods and results

- Smith-Waterman on all pairs of sequences
- Transformation of pairwise similarity matrix into a dissimilarity matrix using:

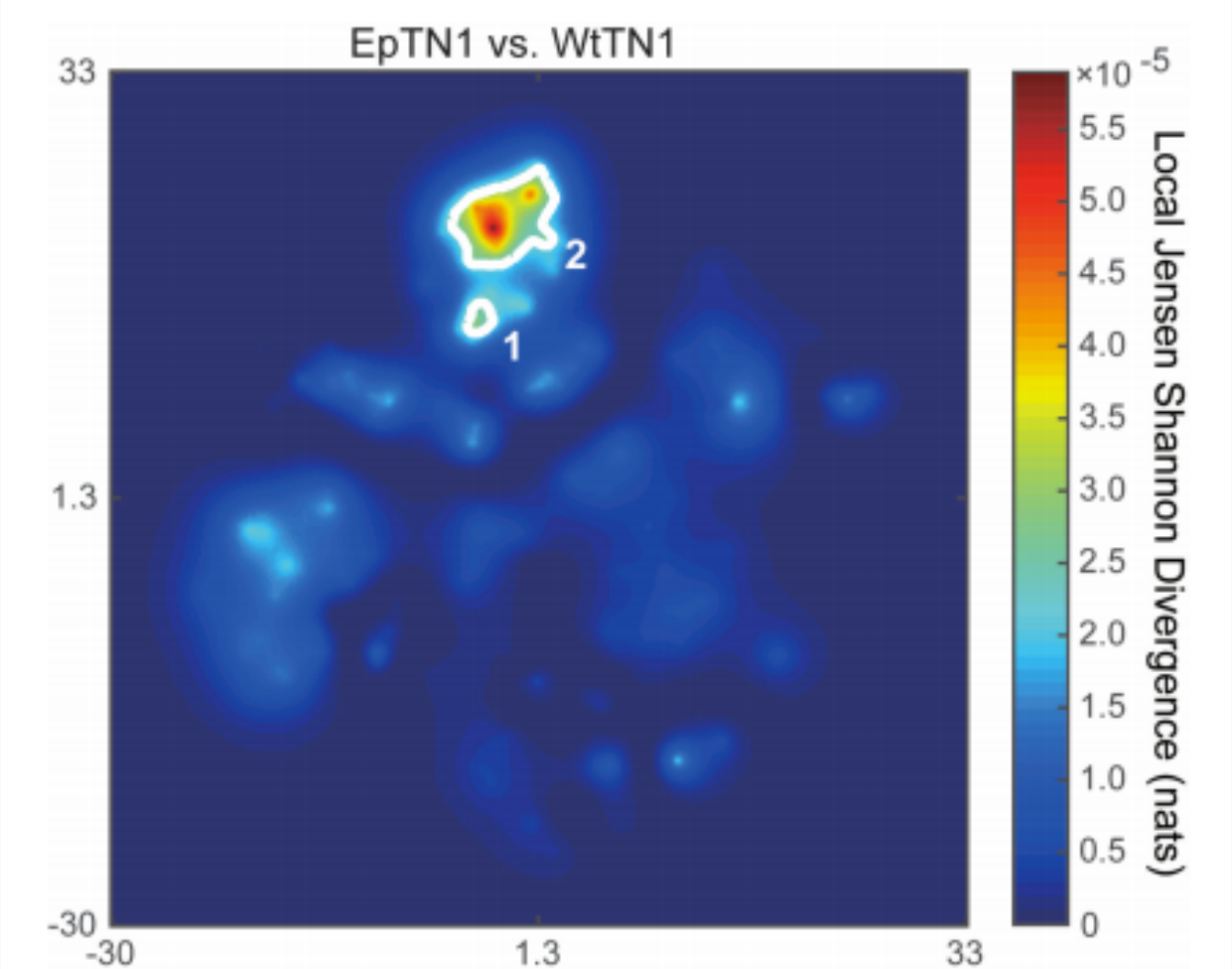
$$S_{i,j} = 1 - 2 * D_{i,j}/(D_{i,i} + D_{j,j})$$

- Apply t-SNE
- Extract subrepertoires and motifs

Methods and results: t-SNE



Methods and results: similarities



Conclusion

Vadim I. Nazarov

Genomics of Adaptive Immunity Lab, IBCH RAS
National Research University Higher School of Economics

email: vdm.nazarov@gmail.com

telegram: @vadimnazarov