

Bioinformatics for Precision Medicine

Prediction of T cell epitopes



Agenda for today

Day 4: Neo epitope prediction and vaccine development (on-line)

Where: [Zoom link](#)

Program:

9.00 - 9.30: MHC binding and T cell epitopes (Carolina Barra Quaglia)

Exercise: Making Sequence logos

10.00 - 10.30: MHC binding predictions (Carolina Barra Quaglia)

Exercise: Calculate the weights of a neural network

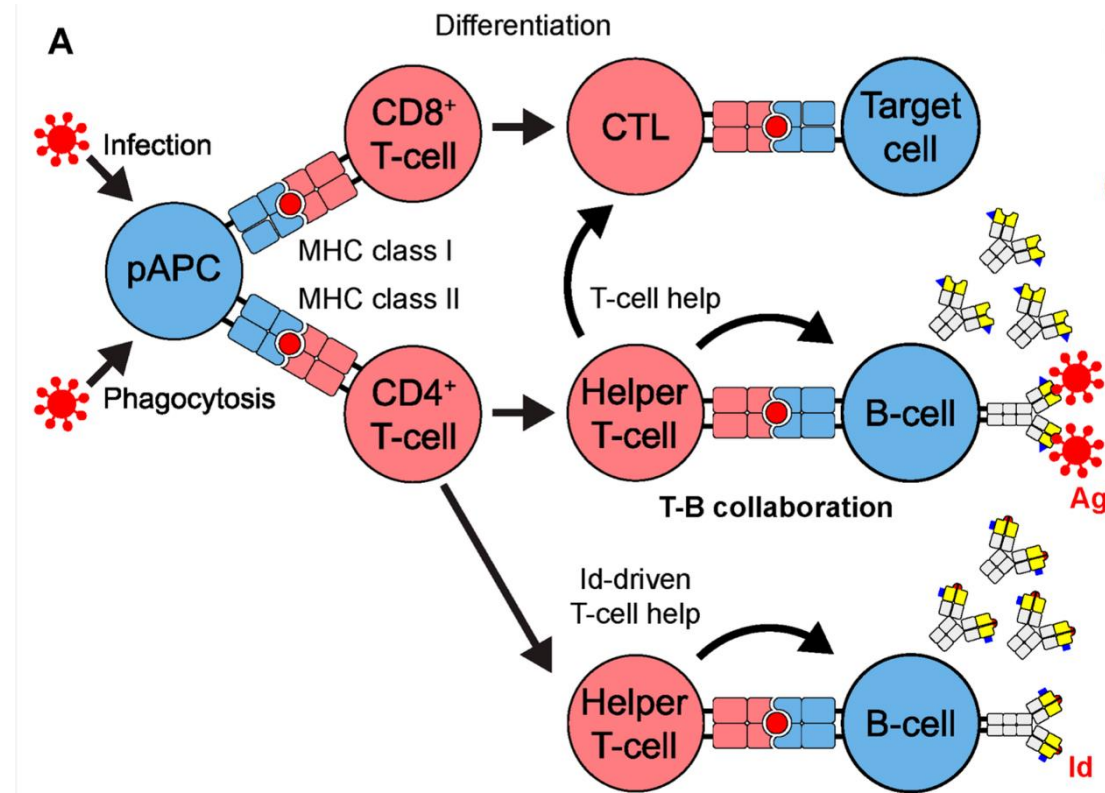
11.00 - 11.30: MHC binding predictions and medical applications (Carolina Barra Quaglia)

Exercise: IEDB database search

12.00 - 14.00: Lunch + Exercise T cell Epitopes prediction **(time off-line)**

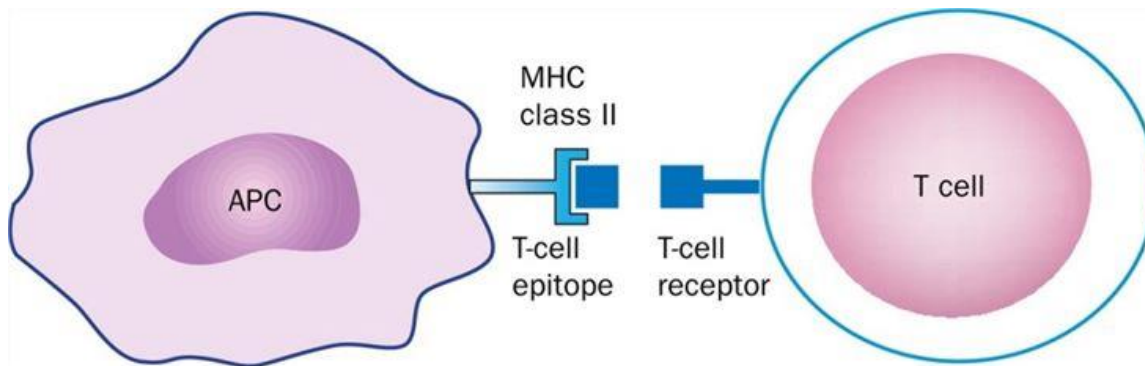
14.00 - 15.00: Walk-through Exercise T cell Epitopes prediction (Carolina Barra Quaglia)

First of all a few notes on T and B cell epitopes

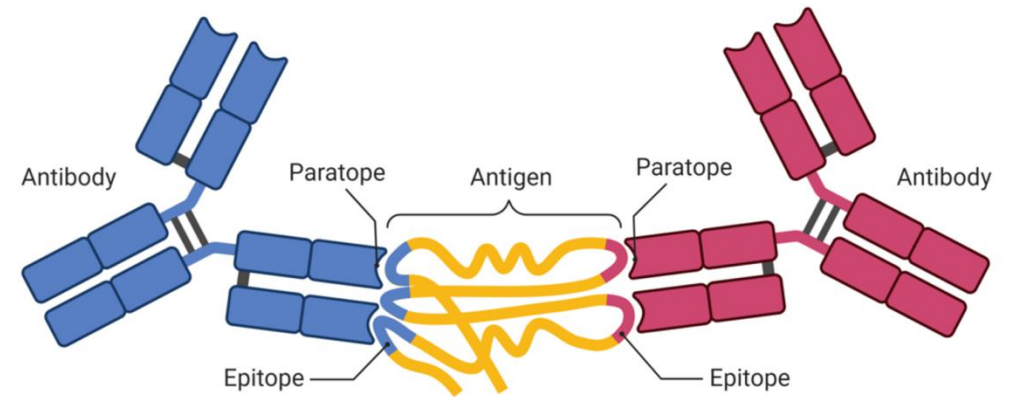


First of all a few notes on T and B cell epitopes

T cell epitopes

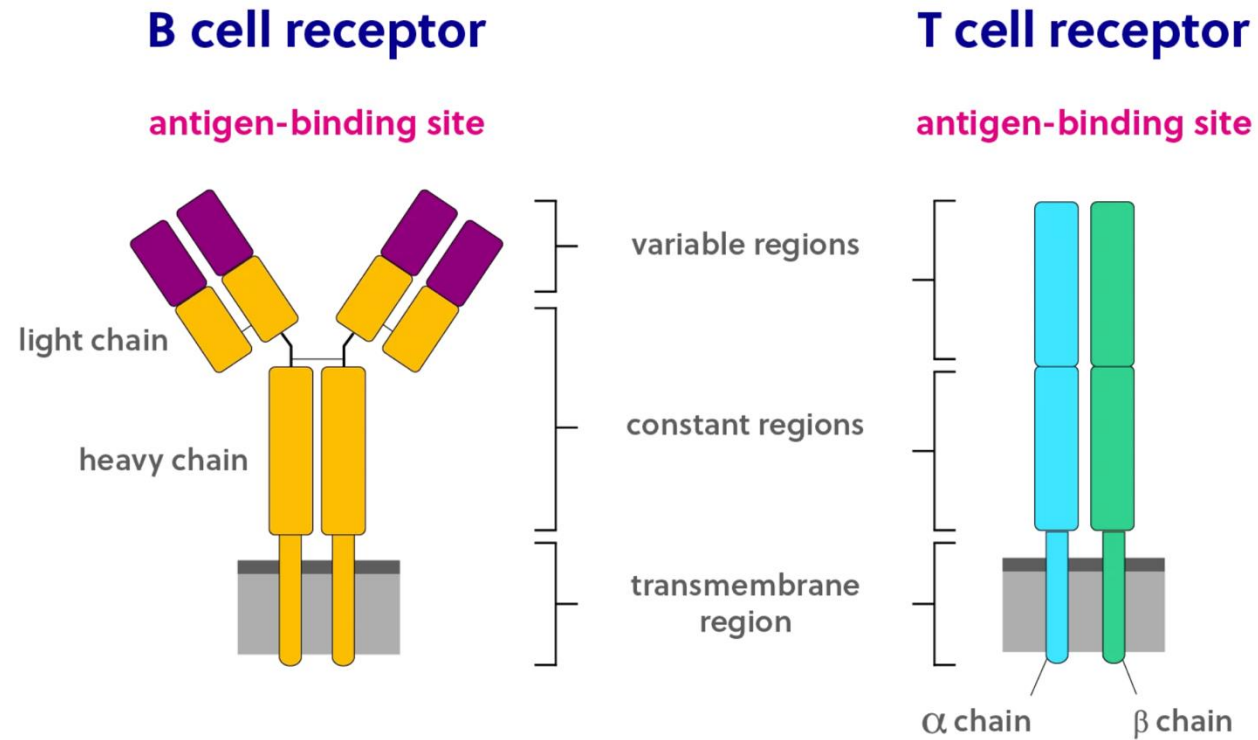


B cell epitopes



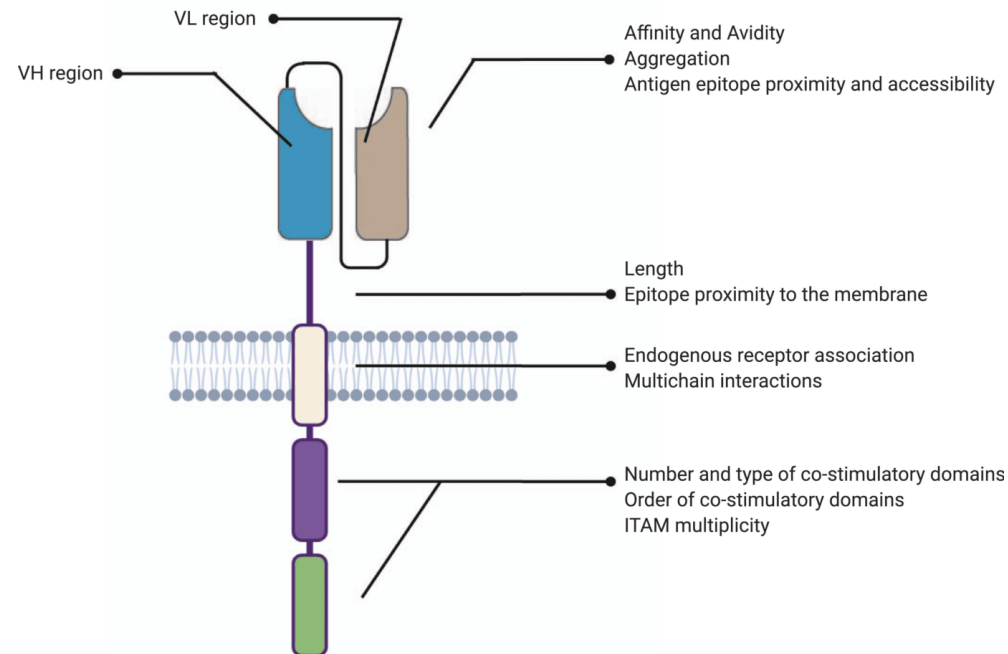
<https://www.creative-biolabs.com/cremap-t-cell-epitope-discovery-service.html>

First of all a few notes on T and B cell epitopes



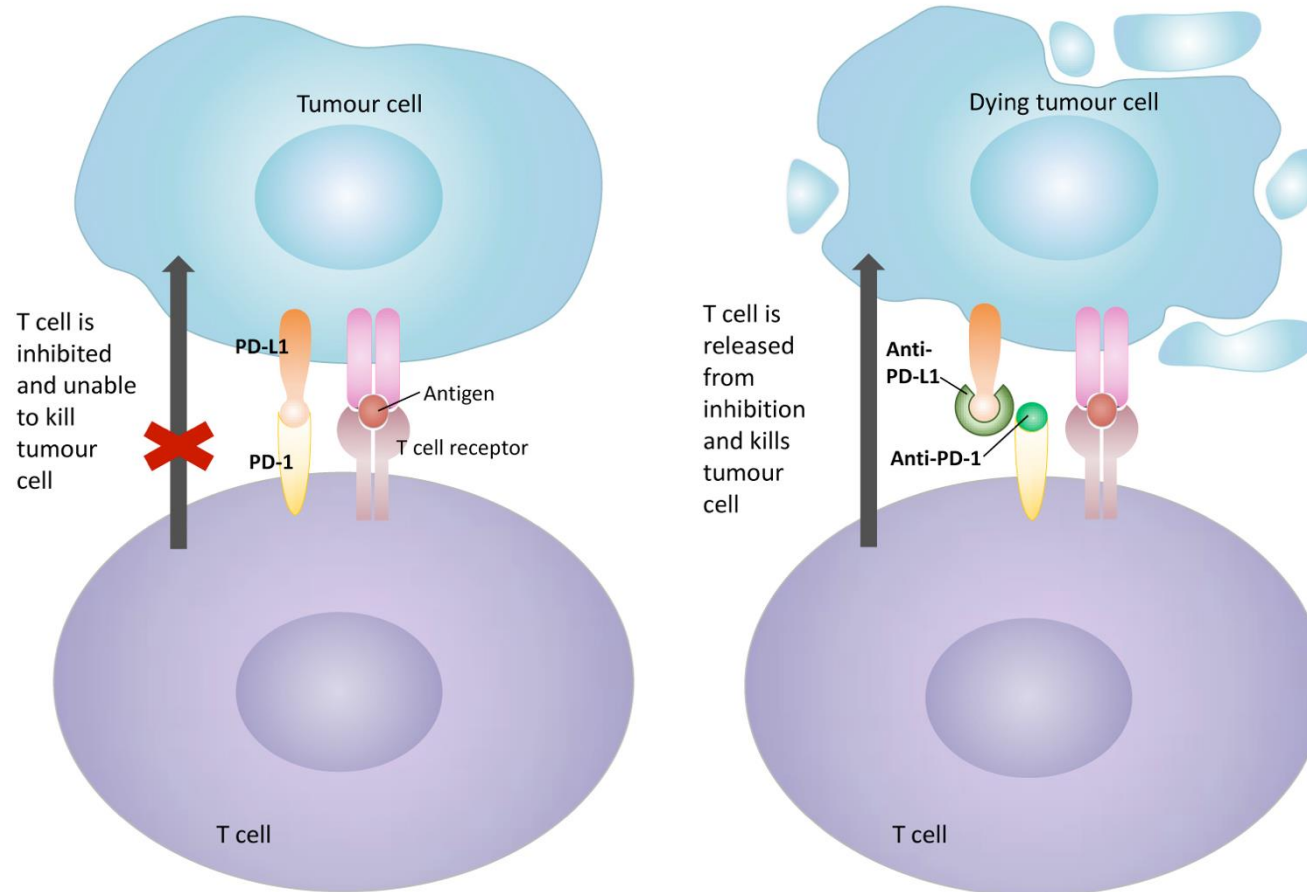
First of all a few notes on T and B cell epitopes

CAR – T cell receptor



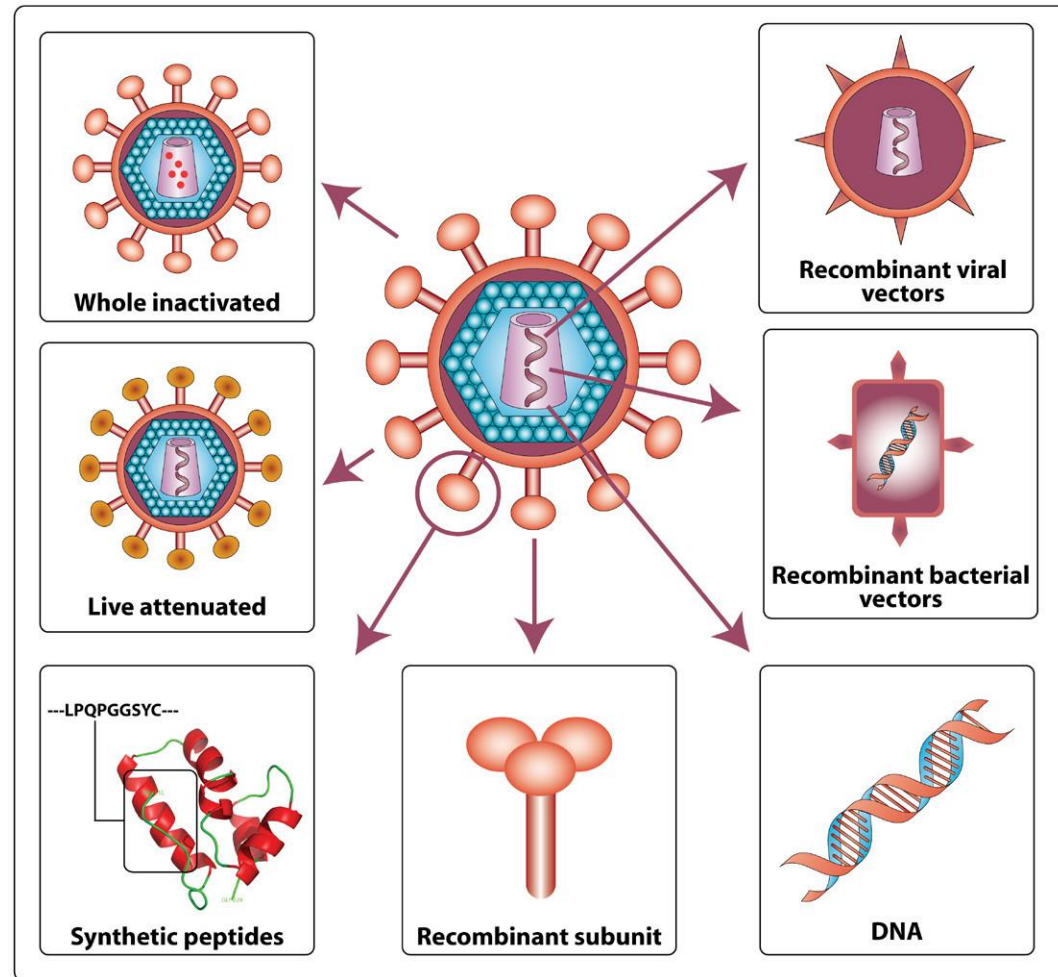
Why are we interested on T cell epitopes?

Cancer immunotherapeutics



Credit: Dr Koh Shimin Grace, Department of Paediatrics, NUS

Peptide vaccines



MHC binding predictions, a historical overview - From a few to all in a decade



Morten Nielsen
Immunoinformatics and Machine learning
Bioinformatics Section
DTU Health Tech

- 1994-1997, Bimas HLA-A2, B27 motif, SYFPEITHI

- **2003, NetMHC-1.0**

HLA-A0204, H-2Kk

- **2007, NetMHCII-1.0**

Prediction for prevalent HLA-DR molecules

- **2007, NetMHCpan-1.0, 2008 NetMHCIIpan-1.0**

Pan-specific prediction for any HLA-I and HLA-DR molecule with known protein sequence

...

- **2020, NetMHCpan-4.1 and NetMHCIIpan-4.0**

Pan-specific prediction to any MHC molecule with known protein sequence. Predictions of binding for 8-13mer peptides. Pan-specific predictions to any HLA class II molecule (DR, DP and DQ) with known protein sequence. Integration of multi-allele MS immunopeptidomics data

MHC binding motifs and information content

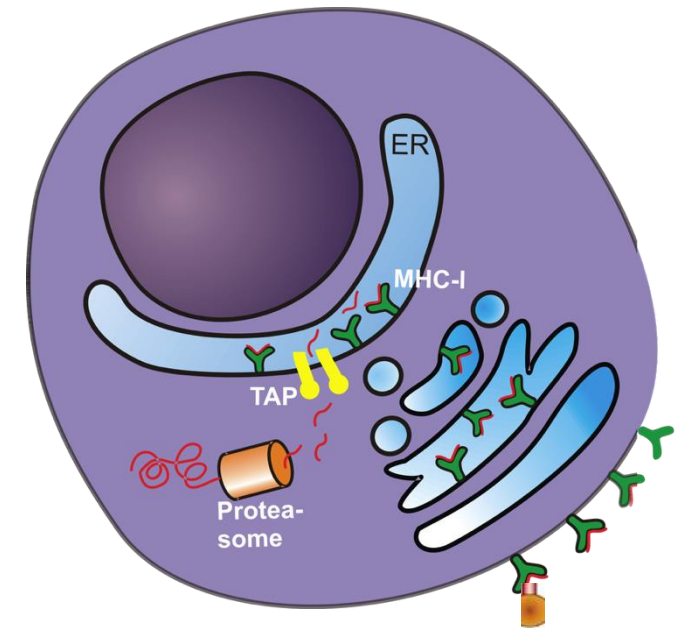
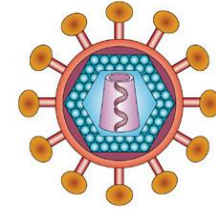
Learning Objectives

Upon completion of this lecture you should be able to:

- Describe the process of MHC binding peptides and T cell epitopes
- Interpret and generate binding motifs by the construction of sequence logos
- Understand the concepts of weight matrix construction

What defines a T cell epitope?

- Antigen Processing (MHC-I: Proteasomal cleavage, TAP; MHC-II: cathepsins)
- MHC binding
- MHC:peptide complex stability
- T cell repertoire
- T cell recognition
- Source protein abundance, cellular location and function
- ???



MHC class I antigen presentation pathway

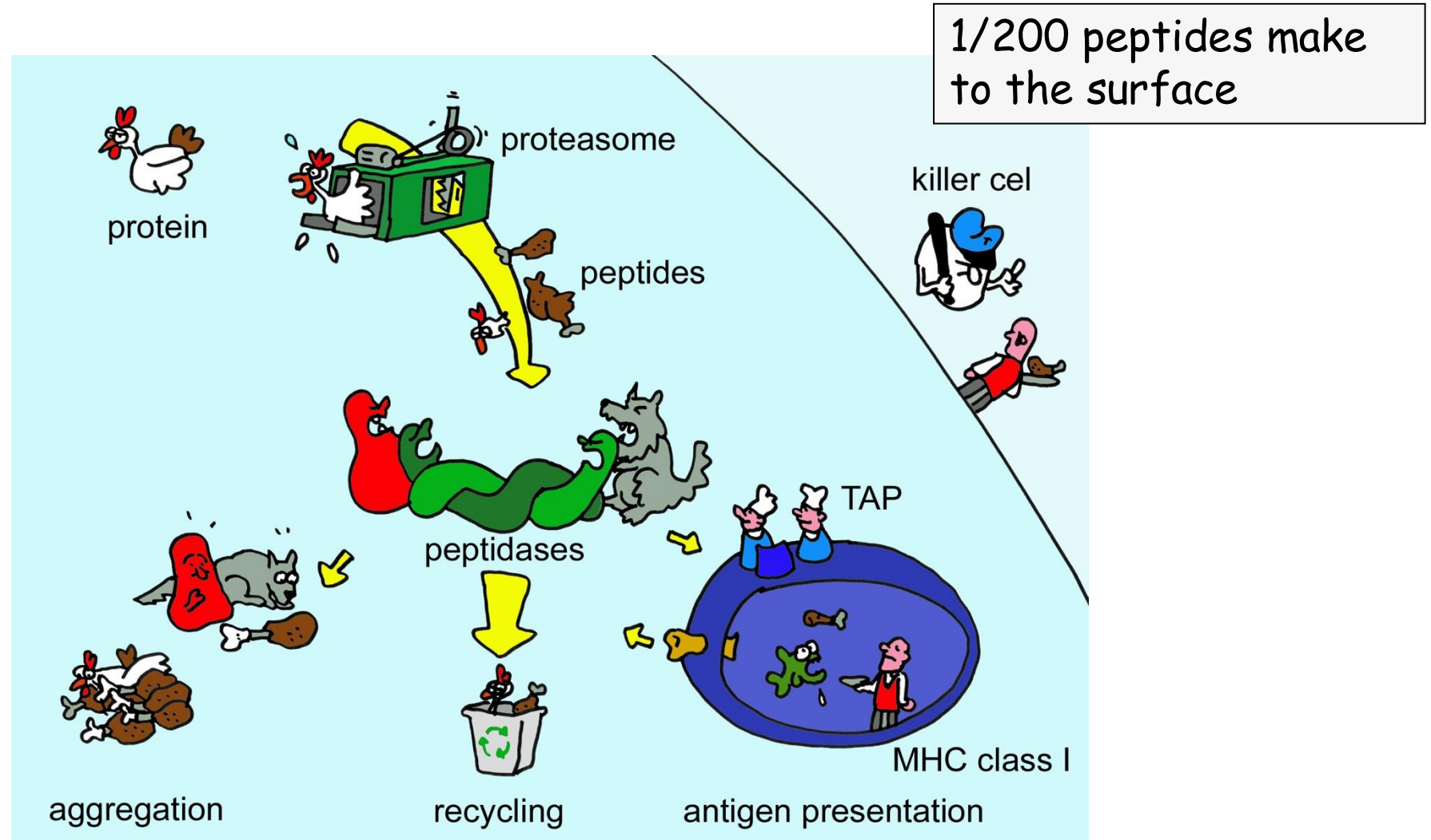
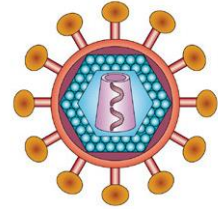


Figure by Eric A.J. Reits

The challenge of identifying the target of T cells

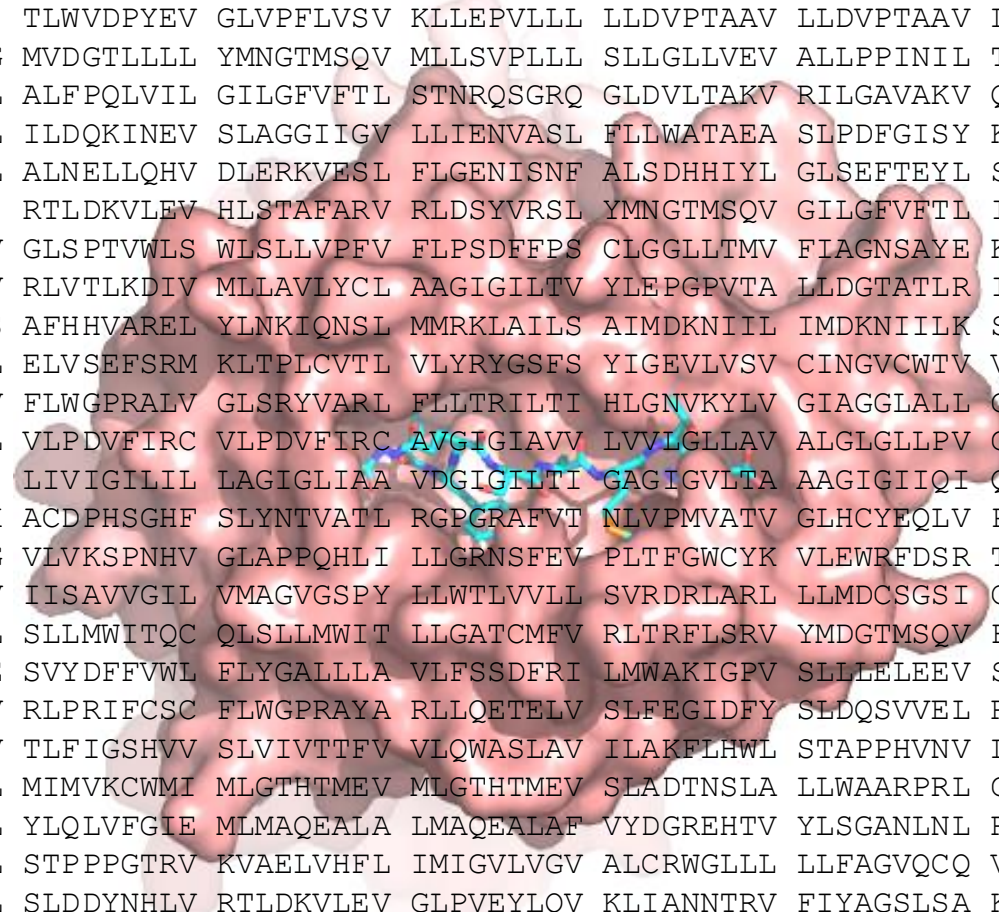
- A pathogen encodes ~10,000 unique peptides
- In a given population, we have ~25-50 different prevalent HLA-I alleles
- This gives all-together ~ 500,000 different HLA:peptide combinations
- Only 10-20 (<0.002%) of these are immunogenic in a given individual



Can we understand why this is, and more importantly can we learn the rules and predict these immunogenic peptides?

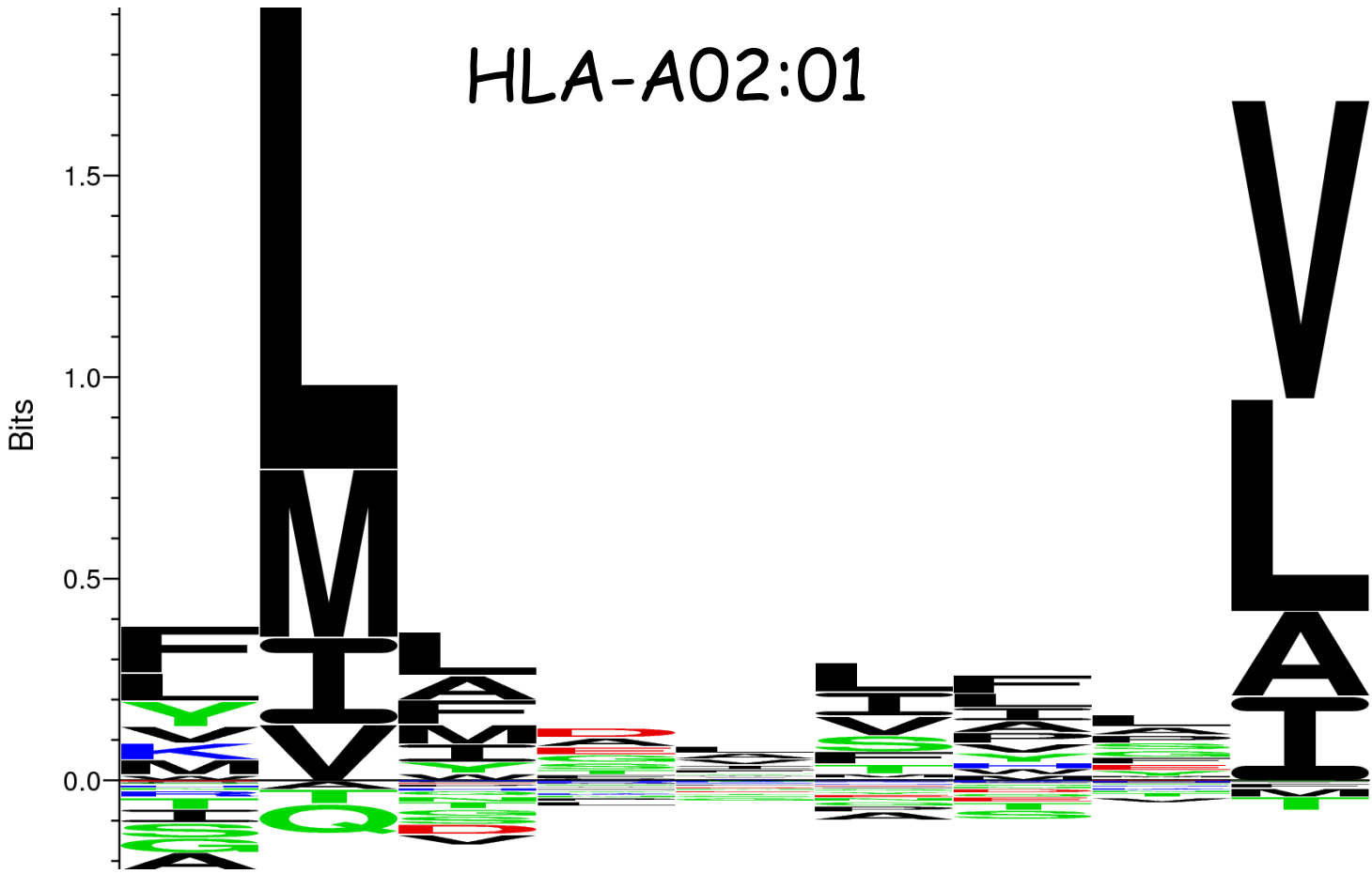


Sequence information and binding motifs



SLLPAIVEL YLLPAIVHI TLWVDPYEV GLVPFLVSV KLLPEVLLL LLDVPTAAV LLDVPTAAV LLDVPTAAV
LLDVPTAAV VLFRGGPRG MVDGTL LLL YMNGTMSQV MLLSVPLLL SLLGLLVEV ALLPPINIL TLIKIQHTL
HLIDYLVTS ILAPPVVKL ALFPQLVIL GILGFVFTL STNRQSGRQ GLDVLTAKV RILGAVAKV QVCERIPTI
ILFGHENRV ILMEHHLK ILDQKINEV SLAGGIIGV LLIENVASL FLLWATAEA SLPDFGISY KKREEAPSL
LERPGGNEI ALSNLEVKL ALNELLQHV DLERKVESL FLGENISNF ALSDHHIYL GLSEFTEYL STAPPAHGV
PLDGEYFTL GVLVGVALI RTLDKVLEV HLSTAFARV RLDSYVRS L YMNGTMSQV GILGFVFTL ILKEPVHGV
ILGFVFTLT LLFGYPVYV GLSPTVWLS WLSLLVPFV FLPSDFFPS CLGGLLTMV FIAGNSAYE KLGEFYNQM
KLVALGINA DLMGYIPLV RLVTLKDIV MLLAVLYCL AAGIGILTV YLEPGPVTA LLDGTATLR ITDQVPFSV
KTWGQYWQV TITDQVPFS AFHHVAREL YLNKIQNSL MMRKLAILS AIMDKNIIL IMDKNIILK SMVGNWAKV
SLLAPGAKQ KIFGSLAFL ELVSEFSRM KLTPLCVTL VLYRYGSFS YIGEVLSV CINGVCWTV VMNILLQYV
ILTVILGVL KVLEYVIKV FLWGPRALV GLSRYVARL FLLTRILTI HLGNVKYLVI GIAGGLALL GLQDCTMLV
TGAPVTYST VIYQYMDDL VLPDVFIRC VLPDVFIRC AVGIGIAVV LVVLGLLAV ALGLGLLPV GIGIGVLA
GAGIGVAVL IAGIGILAI LIVIGILIL LAGIGLIAA VDGIGILTI GAGIGVLT AAGIGI IQI QAGIGILLA
KARDPHSGH KACDPHSGH ACDPHSGHF SLYNTVATL RGPGRFVT NLVPMVATV GLHCYEQLV PLKQHFQIV
AVFDRKSDA LLDFVRFMG VLVKSPNHV GLAPPQH LI LLGRNSFEV PLTFGW CYK VLEWRFD SR TLNAWVKV
GLCTLVAML FIDSYICQV IISAVVGIL VMAGVGS PY LLWTLVLL SVRDR LARL LLMDCSGSI CLTSTVQLV
VLHDDLLEA LMWITQCFL SLLMWITQC QLSLLMWIT LLGATCMFV RLTRFLSRV YMDGTMSQV FLTPKKLQC
ISNDVCAQV VKTDGNPPE SVYDFVWL FLYGALLA VLFSSDFRI LMWAKIGPV SLLLELEE V SLSRFSWGA
YTAFTIPSI RLMKQDFSV RLPRIFCSC FLWGPRAYA RLLQETELV SLFEGIDFY SLDQSVVEL RLMNFTPYI
NMFTPYIGV LMI IPLINV TLF IGSHV SLVIVTTFV VLQWASLAV ILAKFLHWL STAPPHVNV LLLLTVLTV
VVLGVVFGI ILHNGAYSL MIMVKCMMI MLGTHTMEV MLGTHTMEV SLADTNSLA LLWAARPR L GVALQTMKQ
GLYDGM EHL KMVELVHFL YLQLVFGIE MLMAQEALA LMAQEALAF VYDGREHTV YLSGANLNL RMFPNAPYL
EAAGIGILT TLDSQVMSL STPPPGTRV KVAELVHFL IMIGVLVGV ALCRWGLLL LLFAGVQCQ VLLCESTAV
YLSTAFARV YLLEMLWRL SLDDYNHLV RTLDKVLEV GLPVEYLQV KLIANNTRV FIYAGSLSA KLVANNTRL
FLDEFMEGV ALQPGTALL VLDGLDVLL SLYSFPEPE ALYVDSLFF SLLQHLIGL ELTLGEFLK MINAYLDKL
AAGIGILTV FLPSDFFPS SVRDR LARL SLREWLLRI LLSAWILTA AAGIGILTV AVPDEIPPL FAYDGKDYI
AAGIGILTV FLPSDFFPS AAGIGILTV FLPSDFFPS AAGIGILTV FLWGPRALV ETVSEQSNV ITLWQRPLV

HLA binding motifs



<https://services.healthtech.dtu.dk/service.php?Seq2Logo-2.0>

Sequence Information

- Say that a peptide must have L at P_2 in order to bind, and that A, F, W, and Y are found at P_1 . Which position has most information?
- How many questions do I need to ask to tell if a peptide binds looking at only P_1 or P_2 ?

ALAKAAAAM
ALAKAAAAN
ALAKAAAAR
ALAKAAAAT
ALAKAAAAV
WLNERPILT
YLLGFVFTM
FLNAWVKVV
YLNEPVLLL
ALVPFIVSV

Sequence Information

- Say that a peptide must have L at P_2 in order to bind, and that A, F, W, and Y are found at P_1 . Which position has most information?
- How many questions do I need to ask to tell if a peptide binds looking at only P_1 or P_2 ?
- P_1 : 4 questions (at most)
- P_2 : 1 question (L or not)
- P_2 has the most information

```
ALAKAAAAM  
ALAKAAAAN  
ALAKAAAAR  
ALAKAAAAT  
ALAKAAAAV  
WLNERPILT  
YLLGFVFTM  
FLNAWVKVV  
YLNEPVLLL  
ALVPFIVSV
```

Sequence Information

- Say that a peptide must have L at P₂ in order to bind, and that A, F, W, and Y are found at P₁. Which position has most information?

- How many questions do I need to ask to tell if a peptide binds looking at only P₁ or P₂?

- P1: 4 questions (at most)
- P2: 1 question (L or not)
- P2 has the most information

- Calculate p_a at each position
- Entropy

$$S = -\sum_a p_a \log(p_a)$$

- Information content

$$I = \log(20) + \sum_a p_a \log(p_a)$$

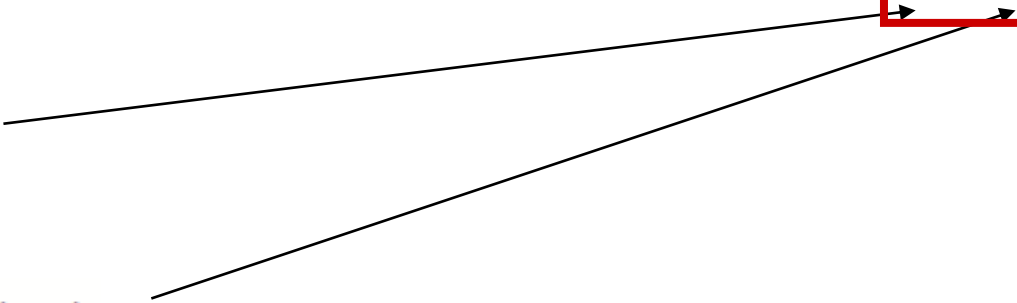
- Conserved positions
 - P_L=1, P_{!L}=0 => S=0, I=log(20)
- Mutable positions
 - P_{aa}=1/20 => S=log(20), I=0

Information content

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	S	I
1	0.10	0.06	0.01	0.02	0.01	0.02	0.02	0.09	0.01	0.07	0.11	0.06	0.04	0.08	0.01	0.11	0.03	0.01	0.05	0.08	3.96	0.37
2	0.07	0.00	0.00	0.01	0.01	0.00	0.01	0.01	0.00	0.08	0.59	0.01	0.07	0.01	0.00	0.01	0.06	0.00	0.01	0.08	2.16	2.16
3	0.08	0.03	0.05	0.10	0.02	0.02	0.01	0.12	0.02	0.03	0.12	0.01	0.03	0.05	0.06	0.06	0.04	0.04	0.04	0.07	4.06	0.26
4	0.07	0.04	0.02	0.11	0.01	0.04	0.08	0.15	0.01	0.10	0.04	0.03	0.01	0.02	0.09	0.07	0.04	0.02	0.00	0.05	3.87	0.45
5	0.04	0.04	0.04	0.04	0.01	0.04	0.05	0.16	0.04	0.02	0.08	0.04	0.01	0.06	0.10	0.02	0.06	0.02	0.05	0.09	4.04	0.28
6	0.04	0.03	0.03	0.01	0.02	0.03	0.03	0.04	0.02	0.14	0.13	0.02	0.03	0.07	0.03	0.05	0.08	0.01	0.03	0.15	3.92	0.40
7	0.14	0.01	0.03	0.03	0.02	0.03	0.04	0.03	0.05	0.07	0.15	0.01	0.03	0.07	0.06	0.07	0.04	0.03	0.02	0.08	3.98	0.34
8	0.05	0.09	0.04	0.01	0.01	0.05	0.07	0.05	0.02	0.04	0.14	0.04	0.02	0.05	0.05	0.08	0.10	0.01	0.04	0.03	4.04	0.28
9	0.07	0.01	0.00	0.00	0.02	0.02	0.02	0.01	0.01	0.08	0.26	0.01	0.01	0.02	0.00	0.04	0.02	0.00	0.01	0.38	2.78	1.55

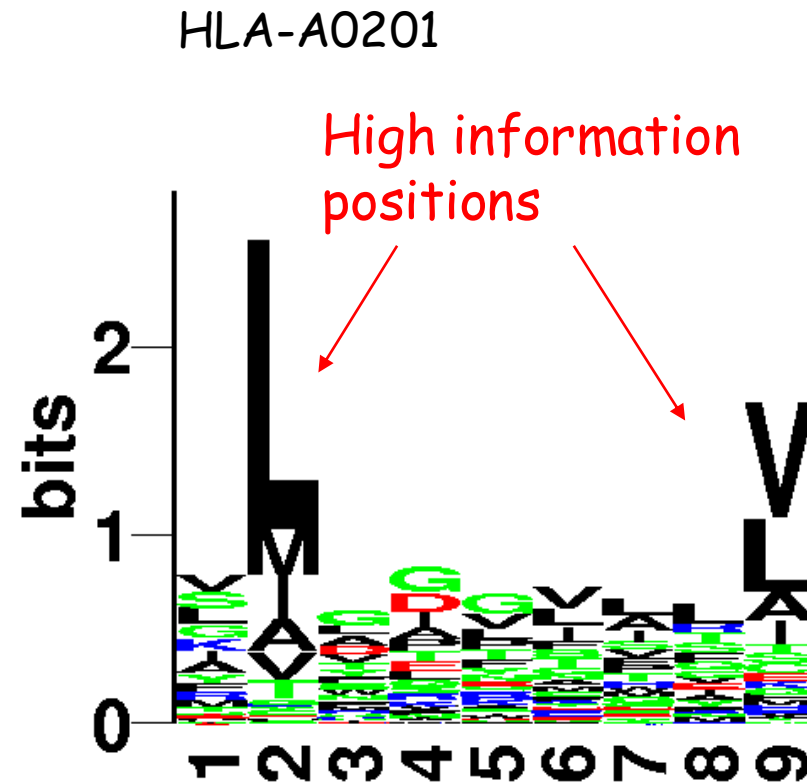
$$S = -\sum_a p_a \log(p_a)$$

$$I = \log(20) + \sum_a p_a \log(p_a)$$



Sequence logos

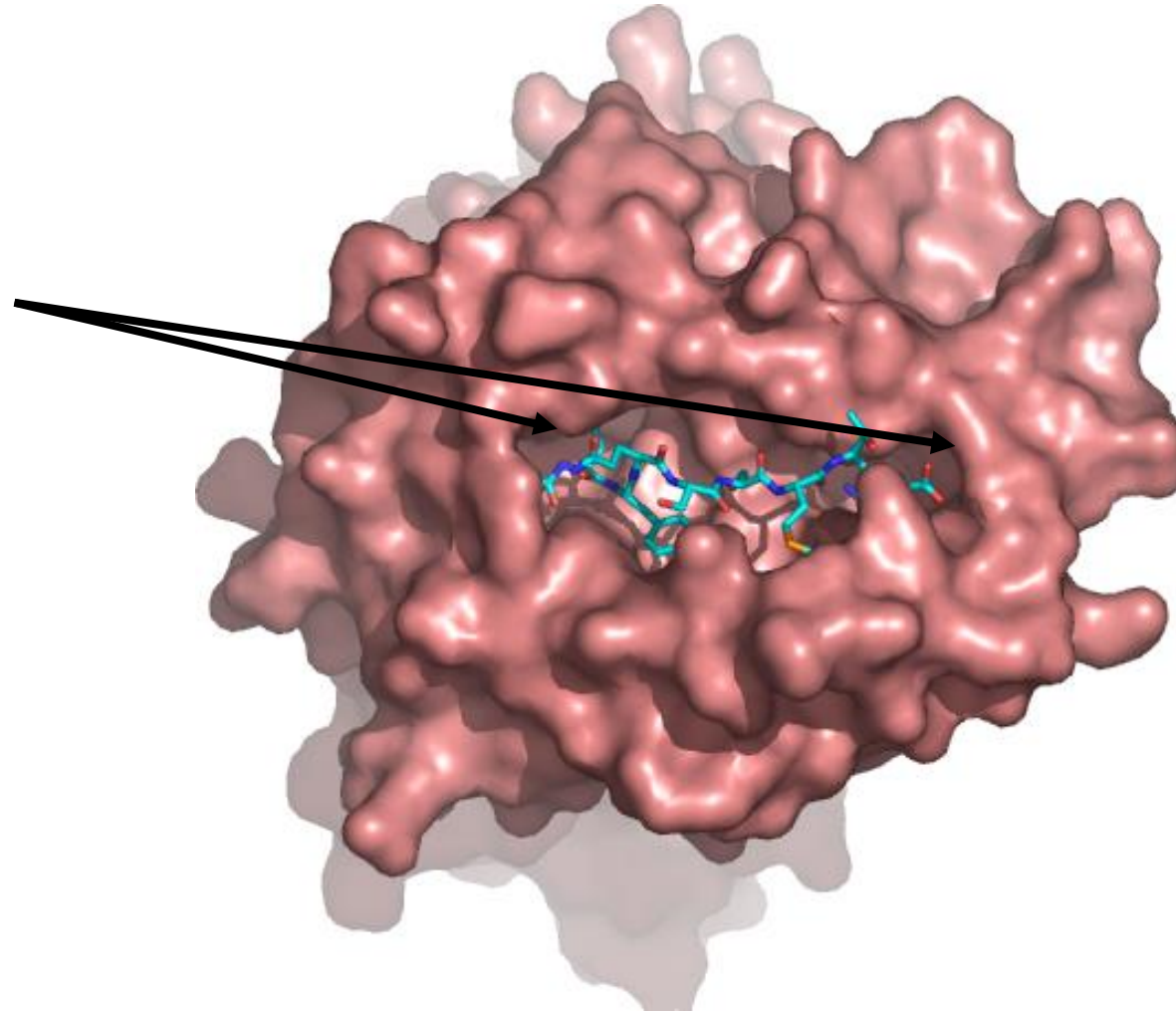
- Height of a column equal to I
- Relative height of a letter is p
- Highly useful tool to visualize sequence motifs



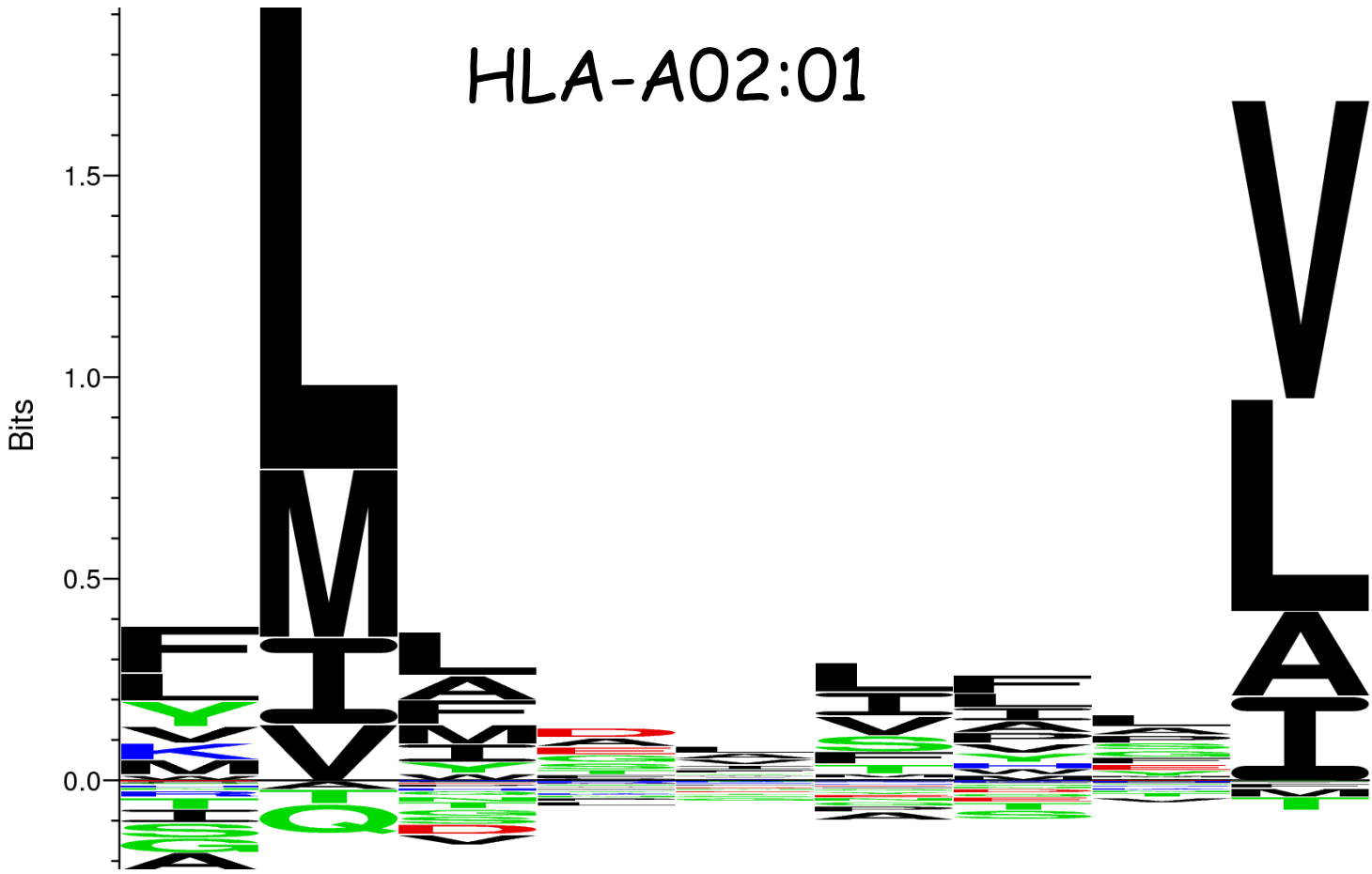
Now, do it yourself!

Binding Motif of MHC class I with peptide

Anchor positions



HLA binding motifs



<https://services.healthtech.dtu.dk/service.php?Seq2Logo-2.0>

Characterizing a binding motif from small data sets

10 MHC restricted peptides

ALAKAAAAM
ALAKAAAAN
ALAKAAAAR
ALAKAAAAT
ALAKAAAAV
GMNERPILT
GILGFVFTM
TLNAWVKVV
KLNEPVLLL
AVVPFIVSV

What can we learn?

A at P1 favors binding?
I is not allowed at P9?

Characterizing a binding motif from small data sets



Sequence weighting

- Poor or biased sampling of sequence space
- Example P1

$$P_A = 2/6$$

$$P_G = 2/6$$

$$P_T = P_K = 1/6$$

$$P_C = P_D = \dots P_V = 0$$

ALAKAAAAM
ALAKAAAAN
ALAKAAAAR
ALAKAAAAT
ALAKAAAAV

GMNERPILT

GILGFVFTM

TLNAWVKVV

KLNEPVLLL

AVVPFIVSV

} Similar sequences
Weight 1/5

RLLDDTPEV 84 nM

GLLGNVSTV 23 nM

ALAKAAAAL 309 nM

Sequence weighting



Pseudo counts

- **I** is not found at position P9. Does this mean that **I** is forbidden ($P(I)=0$)?
- No! Use Blosom substitution matrix to estimate pseudo frequency of **I** at P9

ALAKAAAAM
ALAKAAAAN
ALAKAAAAR
ALAKAAAAT
ALAKAAAAV
GMNERPILT
GILGFVFTM
TLNAWVKVV
KLNEPVLLL
AVVPFIVSV

The Blosum (substitution frequency) matrix

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	0.29	0.03	0.03	0.03	0.02	0.03	0.04	0.08	0.01	0.04	0.06	0.04	0.02	0.02	0.03	0.09	0.05	0.01	0.02	0.07
R	0.04	0.34	0.04	0.03	0.01	0.05	0.05	0.03	0.02	0.02	0.05	0.12	0.02	0.02	0.02	0.04	0.03	0.01	0.02	0.03
N	0.04	0.04	0.32	0.08	0.01	0.03	0.05	0.07	0.03	0.02	0.03	0.05	0.01	0.02	0.02	0.07	0.05	0.00	0.02	0.03
D	0.04	0.03	0.07	0.40	0.01	0.03	0.09	0.05	0.02	0.02	0.03	0.04	0.01	0.01	0.02	0.05	0.04	0.00	0.01	0.02
C	0.07	0.02	0.02	0.02	0.48	0.01	0.02	0.03	0.01	0.04	0.07	0.02	0.02	0.02	0.02	0.04	0.04	0.00	0.01	0.06
Q	0.06	0.07	0.04	0.05	0.01	0.21	0.10	0.04	0.03	0.03	0.05	0.09	0.02	0.01	0.02	0.06	0.04	0.01	0.02	0.04
E	0.06	0.05	0.04	0.09	0.01	0.06	0.30	0.04	0.03	0.02	0.04	0.08	0.01	0.02	0.03	0.06	0.04	0.01	0.02	0.03
G	0.08	0.02	0.04	0.03	0.01	0.02	0.03	0.51	0.01	0.02	0.03	0.03	0.01	0.02	0.02	0.05	0.03	0.01	0.01	0.02
H	0.04	0.05	0.05	0.04	0.01	0.04	0.05	0.04	0.35	0.02	0.04	0.05	0.02	0.03	0.02	0.04	0.03	0.01	0.06	0.02
I	0.05	0.02	0.01	0.02	0.02	0.01	0.02	0.02	0.01	0.27	0.17	0.02	0.04	0.04	0.01	0.03	0.04	0.01	0.02	0.18
L	0.04	0.02	0.01	0.02	0.02	0.02	0.02	0.02	0.01	0.12	0.38	0.03	0.05	0.05	0.01	0.02	0.03	0.01	0.02	0.10
K	0.06	0.11	0.04	0.04	0.01	0.05	0.07	0.04	0.02	0.03	0.04	0.28	0.02	0.02	0.03	0.05	0.04	0.01	0.02	0.03
M	0.05	0.03	0.02	0.02	0.02	0.03	0.03	0.03	0.02	0.10	0.20	0.04	0.16	0.05	0.02	0.04	0.04	0.01	0.02	0.09
F	0.03	0.02	0.02	0.02	0.01	0.01	0.02	0.03	0.02	0.06	0.11	0.02	0.03	0.39	0.01	0.03	0.03	0.02	0.09	0.06
P	0.06	0.03	0.02	0.03	0.01	0.02	0.04	0.04	0.01	0.03	0.04	0.04	0.01	0.01	0.49	0.04	0.04	0.00	0.01	0.03
S	0.11	0.04	0.05	0.05	0.02	0.03	0.05	0.07	0.02	0.03	0.04	0.05	0.02	0.02	0.03	0.22	0.08	0.01	0.02	0.04
T	0.07	0.04	0.04	0.04	0.02	0.03	0.04	0.04	0.01	0.05	0.07	0.05	0.02	0.02	0.03	0.09	0.25	0.01	0.02	0.07
W	0.03	0.02	0.02	0.02	0.01	0.02	0.02	0.03	0.02	0.03	0.05	0.02	0.02	0.06	0.01	0.02	0.02	0.49	0.07	0.03
Y	0.04	0.03	0.02	0.02	0.01	0.02	0.03	0.02	0.05	0.04	0.07	0.03	0.02	0.13	0.02	0.03	0.03	0.03	0.32	0.05
V	0.07	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.01	0.16	0.13	0.03	0.03	0.04	0.02	0.03	0.05	0.01	0.02	0.27

Some amino acids are highly conserved (i.e. C),
some have a high change of mutation (i.e. I)

Pseudo count estimation

- Calculate observed amino acid frequencies f_a
- Pseudo frequency for amino acid b

- Example
$$g_b = \sum_a f_a \cdot q_{b|a}$$

$$g_I = 0.2 \cdot q_{I|M} + 0.1 \cdot q_{I|R} + \dots + 0.3 \cdot q_{I|V} + 0.1 \cdot q_{I|L}$$

$$g_I = 0.2 \cdot 0.1 + 0.1 \cdot 0.02 + \dots + 0.3 \cdot 0.16 + 0.1 \cdot 0.12 = 0.094$$

ALAKAAAAM
ALAKAAAAN
ALAKAAAAR
ALAKAAAAT
ALAKAAAAV
GMNERPILT
GILGFVFTM
TLNAWVKVV
KLNEPVLLL
AVVPFIVSV

Weight on pseudo count

- Example

$$p_a = \frac{\alpha \cdot f_a + \beta \cdot g_a}{\alpha + \beta}$$

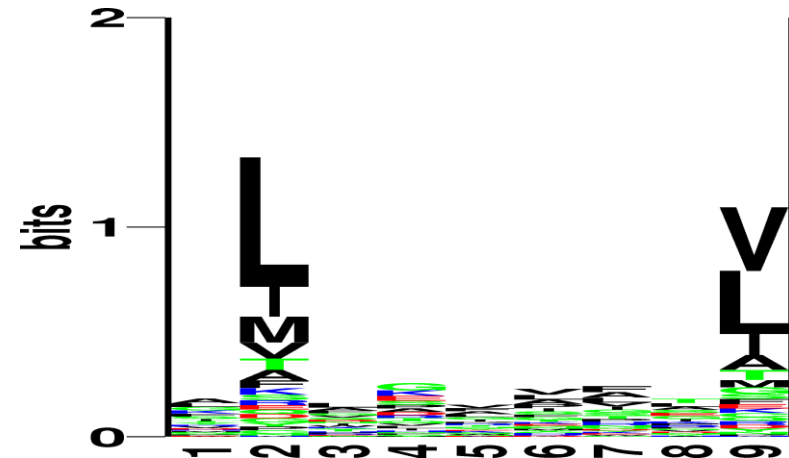
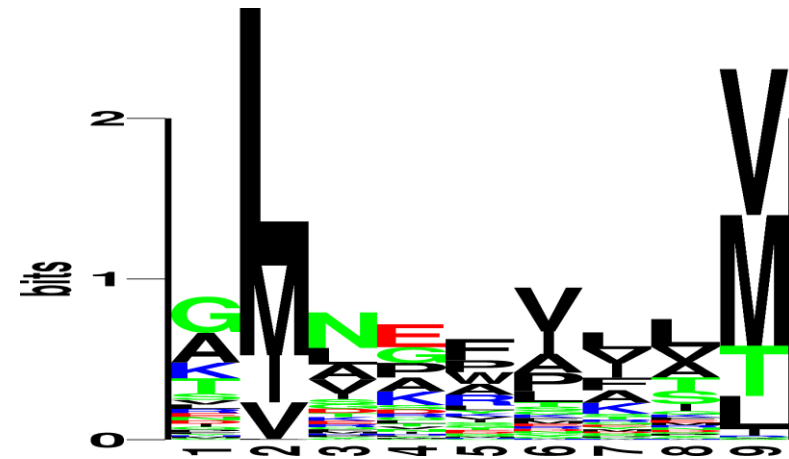
- If α large, $p \approx f$ and only the observed data defines the motif
- If α small, $p \approx g$ and the pseudo counts (or prior) defines the motif
- β is [50-200] normally

ALAKAAAAM
ALAKAAAAN
ALAKAAAAR
ALAKAAAAT
ALAKAAAAV
GMNERPILT
GILGFVFTM
TLNAWVKVV
KLNEPVLLL
AVVPFIVSV

Position specific weighting

- We know that positions 2 and 9 are anchor positions for most MHC binding motifs
 - Increase weight on high information positions

- Motif found on large data set



Weight matrices

- Estimated amino acid frequencies from alignment

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
1	0.08	0.06	0.02	0.03	0.02	0.02	0.03	0.08	0.02	0.08	0.11	0.06	0.04	0.06	0.02	0.09	0.04	0.01	0.04	0.08
2	0.04	0.01	0.01	0.01	0.01	0.01	0.02	0.02	0.01	0.11	0.44	0.02	<u>0.06</u>	0.03	0.01	0.02	0.05	0.00	0.01	<u>0.10</u>
3	0.08	0.04	0.05	0.07	0.02	0.03	0.03	0.08	0.02	0.05	0.11	0.03	0.03	0.06	0.04	0.06	0.05	0.03	0.05	0.07
4	0.08	0.05	0.03	0.10	0.01	0.05	0.08	0.13	0.01	0.05	0.06	0.05	0.01	0.03	0.08	0.06	0.04	0.02	0.01	0.05
5	0.06	0.04	0.05	0.03	0.01	0.04	0.05	0.11	0.03	0.04	0.09	0.04	0.02	0.06	0.06	0.04	0.05	0.02	0.05	0.08
6	0.06	0.03	0.03	0.03	0.03	0.03	0.04	0.06	0.02	0.10	0.14	0.04	0.03	0.05	0.04	0.06	0.06	0.01	0.03	0.13
7	0.10	0.02	0.04	0.04	0.02	0.03	0.04	0.05	0.04	0.08	0.12	0.02	0.03	0.06	0.07	0.06	0.05	0.03	0.03	0.08
8	0.05	0.07	0.04	0.03	0.01	0.04	0.06	0.06	0.03	0.06	0.13	0.06	0.02	0.05	0.04	0.08	0.07	0.01	0.04	0.05
9	0.08	0.02	0.01	0.01	0.02	0.02	0.03	0.02	0.01	0.10	0.23	0.03	0.02	0.04	0.01	0.04	0.04	0.00	0.02	0.25

- What do the numbers mean?
 - $P_2(V) > P_2(M)$. Does this mean that V enables binding more than M.
 - In nature not all amino acids are found equally often
 - In nature V is found more often than M, so we must somehow rescale with the background
 - $q_M = 0.025$, $q_V = 0.073$
 - Finding 7% V is hence not significant, but 7% M highly significant

Weight matrices

- A weight matrix is given as

$$W_{ij} = \log(p_{ij}/q_j)$$

- where i is a position in the motif, and j an amino acid. q_j is the background frequency for amino acid j

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
1	0.6	0.4	-3.5	-2.4	-0.4	-1.9	-2.7	0.3	-1.1	1.0	0.3	0.0	1.4	1.2	-2.7	1.4	-1.2	-2.0	1.1	0.7
2	-1.6	-6.6	-6.5	-5.4	-2.5	-4.0	-4.7	-3.7	-6.3	1.0	5.1	-3.7	<u>3.1</u>	-4.2	-4.3	-4.2	-0.2	-5.9	-3.8	<u>0.4</u>
3	0.2	-1.3	0.1	1.5	0.0	-1.8	-3.3	0.4	0.5	-1.0	0.3	-2.5	1.2	1.0	-0.1	-0.3	-0.5	3.4	1.6	0.0
4	-0.1	-0.1	-2.0	2.0	-1.6	0.5	0.8	2.0	-3.3	0.1	-1.7	-1.0	-2.2	-1.6	1.7	-0.6	-0.2	1.3	-6.8	-0.7
5	-1.6	-0.1	0.1	-2.2	-1.2	0.4	-0.5	1.9	1.2	-2.2	-0.5	-1.3	-2.2	1.7	1.2	-2.5	-0.1	1.7	1.5	1.0
6	-0.7	-1.4	-1.0	-2.3	1.1	-1.3	-1.4	-0.2	-1.0	1.8	0.8	-1.9	0.2	1.0	-0.4	-0.6	0.4	-0.5	-0.0	2.1
7	1.1	-3.8	-0.2	-1.3	1.3	-0.3	-1.3	-1.4	2.1	0.6	0.7	-5.0	1.1	0.9	1.3	-0.5	-0.9	2.9	-0.4	0.5
8	-2.2	1.0	-0.8	-2.9	-1.4	0.4	0.1	-0.4	0.2	-0.0	1.1	-0.5	-0.5	0.7	-0.3	0.8	0.8	-0.7	1.3	-1.1
9	-0.2	-3.5	-6.1	-4.5	0.7	-0.8	-2.5	-4.0	-2.6	0.9	2.8	-3.0	-1.8	-1.4	-6.2	-1.9	-1.6	-4.9	-1.6	4.5

Scoring a sequence to a weight matrix

- Score sequences to weight matrix by looking up and adding L values from the matrix

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
1	0.6	0.4	-3.5	-2.4	-0.4	-1.9	-2.7	0.3	-1.1	1.0	0.3	0.0	1.4	1.2	-2.7	1.4	-1.2	-2.0	1.1	0.7
2	-1.6	-6.6	-6.5	-5.4	-2.5	-4.0	-4.7	-3.7	-6.3	1.0	5.1	-3.7	3.1	-4.2	-4.3	-4.2	-0.2	-5.9	-3.8	0.4
3	0.2	-1.3	0.1	1.5	0.0	-1.8	-3.3	0.4	0.5	-1.0	0.3	-2.5	1.2	1.0	-0.1	-0.3	-0.5	3.4	1.6	0.0
4	-0.1	-0.1	-2.0	2.0	-1.6	0.5	0.8	2.0	-3.3	0.1	-1.7	-1.0	-2.2	-1.6	1.7	-0.6	-0.2	1.3	-6.8	-0.7
5	-1.6	-0.1	0.1	-2.2	-1.2	0.4	-0.5	1.9	1.2	-2.2	-0.5	-1.3	-2.2	1.7	1.2	-2.5	-0.1	1.7	1.5	1.0
6	-0.7	-1.4	-1.0	-2.3	1.1	-1.3	-1.4	-0.2	-1.0	1.8	0.8	-1.9	0.2	1.0	-0.4	-0.6	0.4	-0.5	-0.0	2.1
7	1.1	-3.8	-0.2	-1.3	1.3	-0.3	-1.3	-1.4	2.1	0.6	0.7	-5.0	1.1	0.9	1.3	-0.5	-0.9	2.9	-0.4	0.5
8	-2.2	1.0	-0.8	-2.9	-1.4	0.4	0.1	-0.4	0.2	-0.0	1.1	-0.5	-0.5	0.7	-0.3	0.8	0.8	-0.7	1.3	-1.1
9	-0.2	-3.5	-6.1	-4.5	0.7	-0.8	-2.5	-4.0	-2.6	0.9	2.8	-3.0	-1.8	-1.4	-6.2	-1.9	-1.6	-4.9	-1.6	4.5

RLLDDTPEV

GLLGNVSTV

ALAKAAAAL

Which peptide is most likely to bind?
Which peptide second?

Scoring a sequence to a weight matrix

- Score sequences to weight matrix by looking up and adding L values from the matrix

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
1	0.6	0.4	-3.5	-2.4	-0.4	-1.9	-2.7	0.3	-1.1	1.0	0.3	0.0	1.4	1.2	-2.7	1.4	-1.2	-2.0	1.1	0.7
2	-1.6	-6.6	-6.5	-5.4	-2.5	-4.0	-4.7	-3.7	-6.3	1.0	5.1	-3.7	3.1	-4.2	-4.3	-4.2	-0.2	-5.9	-3.8	0.4
3	0.2	-1.3	0.1	1.5	0.0	-1.8	-3.3	0.4	0.5	-1.0	0.3	-2.5	1.2	1.0	-0.1	-0.3	-0.5	3.4	1.6	0.0
4	-0.1	-0.1	-2.0	2.0	-1.6	0.5	0.8	2.0	-3.3	0.1	-1.7	-1.0	-2.2	-1.6	1.7	-0.6	-0.2	1.3	-6.8	-0.7
5	-1.6	-0.1	0.1	-2.2	-1.2	0.4	-0.5	1.9	1.2	-2.2	-0.5	-1.3	-2.2	1.7	1.2	-2.5	-0.1	1.7	1.5	1.0
6	-0.7	-1.4	-1.0	-2.3	1.1	-1.3	-1.4	-0.2	-1.0	1.8	0.8	-1.9	0.2	1.0	-0.4	-0.6	0.4	-0.5	-0.0	2.1
7	1.1	-3.8	-0.2	-1.3	1.3	-0.3	-1.2	-1.4	2.1	0.6	0.7	-5.0	1.1	0.9	1.3	-0.5	-0.9	2.9	-0.4	0.5
8	-2.2	1.0	-0.8	-2.9	-1.4	0.4	0.1	-0.4	0.2	-0.0	1.1	-0.5	-0.5	0.7	-0.5	0.8	0.8	-0.7	1.3	-1.1
9	-0.2	-3.5	-6.1	-4.5	0.7	-0.8	-2.5	-4.0	-2.6	0.9	2.8	-3.0	-1.8	-1.4	-6.2	-1.9	-1.6	-4.9	-1.6	4.5

RLLDDTPEV	11.9	84nM
GLLGNVSTV	14.7	23nM
ALAKAAAAL	4.3	309nM

Which peptide is most likely to bind?
Which peptide second?

Neural networks and MHC binding predictions

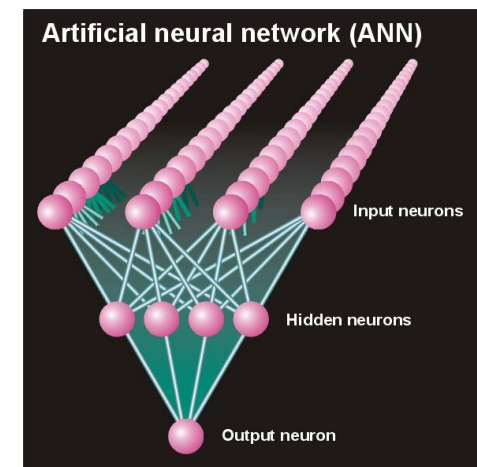
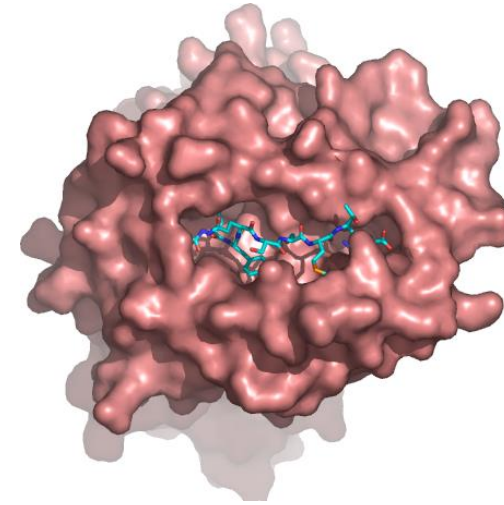
Learning Objectives

Upon completion of this module the student should be able to:

- Summarize the limitations of linear functions such as weight matrices
- Identify and interpret the components of a neural network and calculate its weights
- Discuss the challenges of model fitting: overfitting
- Discuss the solutions to overfitting: test-set model, cross-validation

Is there anything beyond weight matrices

- The effect on the binding affinity of having a given amino acid at one position can be influenced by the amino acids at other positions in the peptide (sequence correlations).
 - Two adjacent amino acids may for example compete for the space in a pocket in the MHC molecule.
- Artificial neural networks (ANN) are ideally suited to take such correlations into account



HUMAN

Higher order sequence correlations

Neural networks can learn higher order correlations!

- What does this mean?

Say that the peptide needs one and only one large amino acid in the positions P3 and P4 to fill the binding cleft

How would you formulate this to test if a peptide can bind?

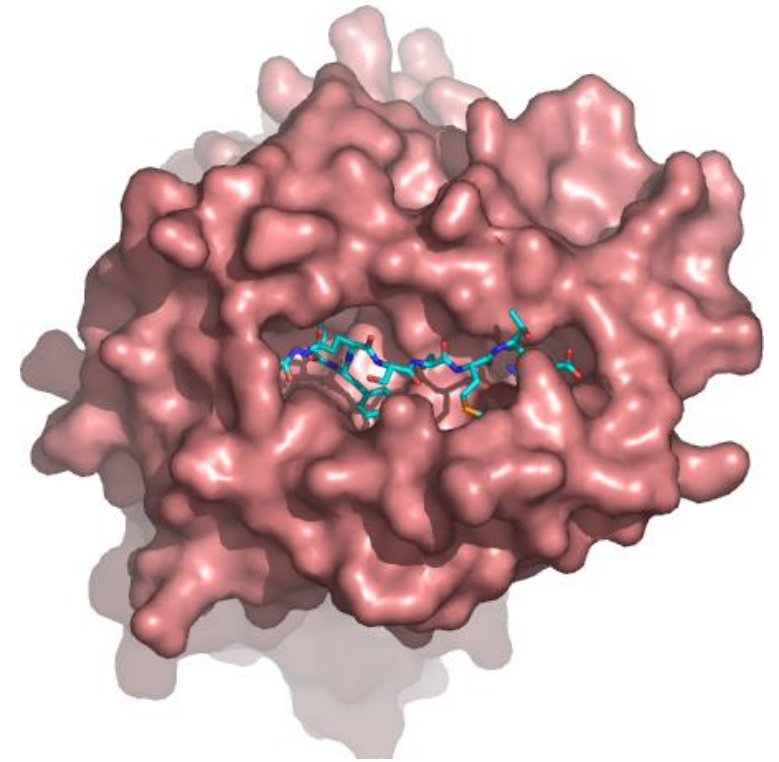
$S S \Rightarrow 0$

$L S \Rightarrow 1$

$S L \Rightarrow 1$

$L L \Rightarrow 0$

No linear function
can learn this
(XOR) pattern



Linear functions (like PSSM's) cannot learn higher order signals

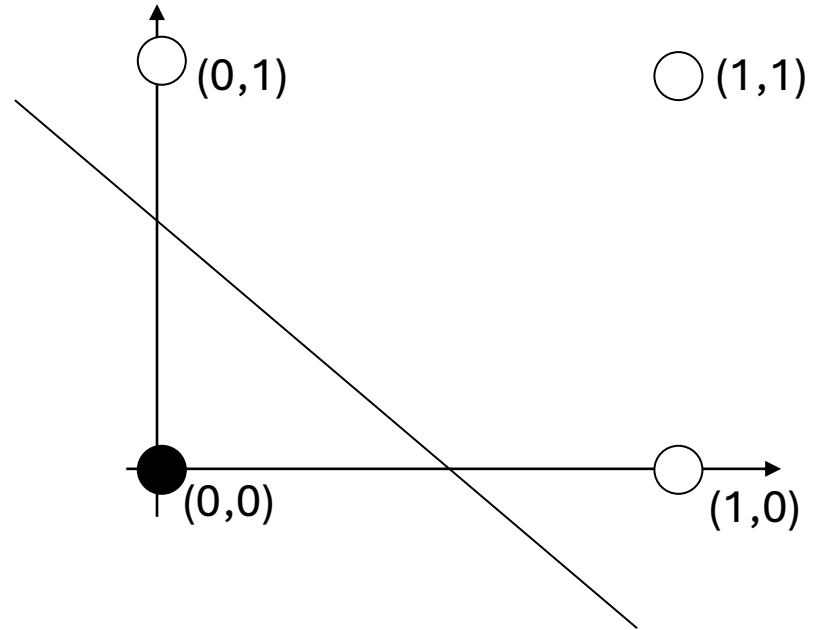
XOR function:

0 0 \Rightarrow 0

1 0 \Rightarrow 1

0 1 \Rightarrow 1

1 1 \Rightarrow 0



Linear functions (like PSSM's) cannot learn higher order signals

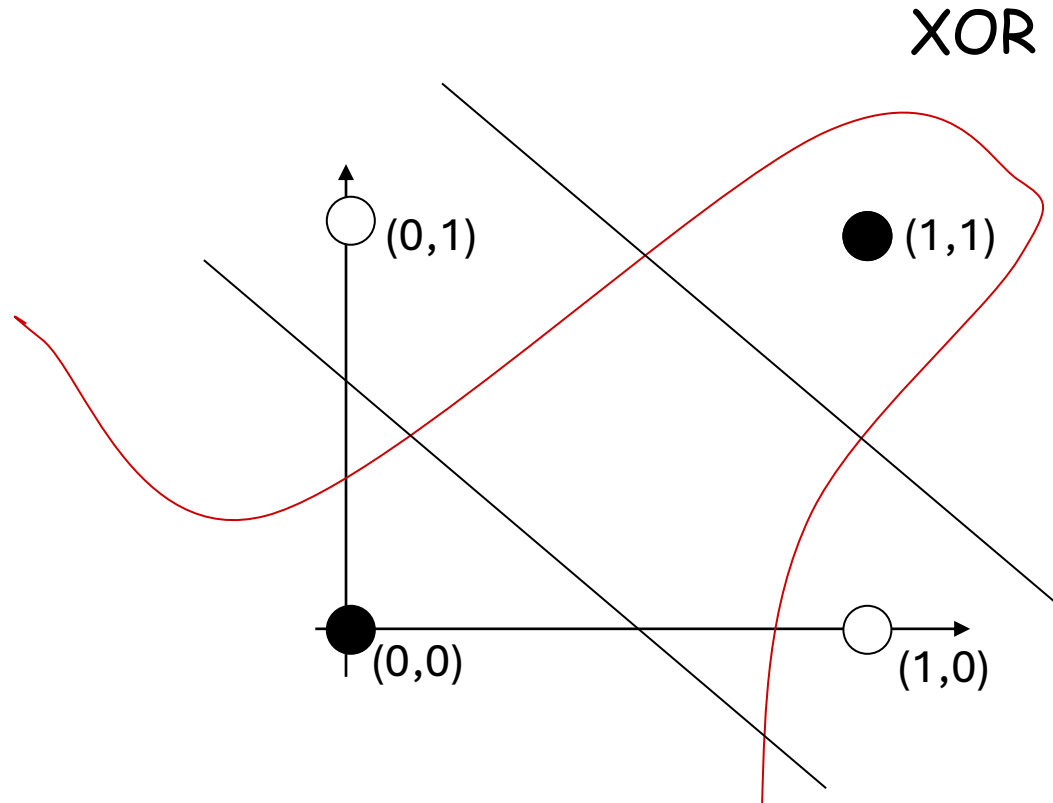
XOR function:

$0\ 0 \Rightarrow 0$

$1\ 0 \Rightarrow 1$

$0\ 1 \Rightarrow 1$

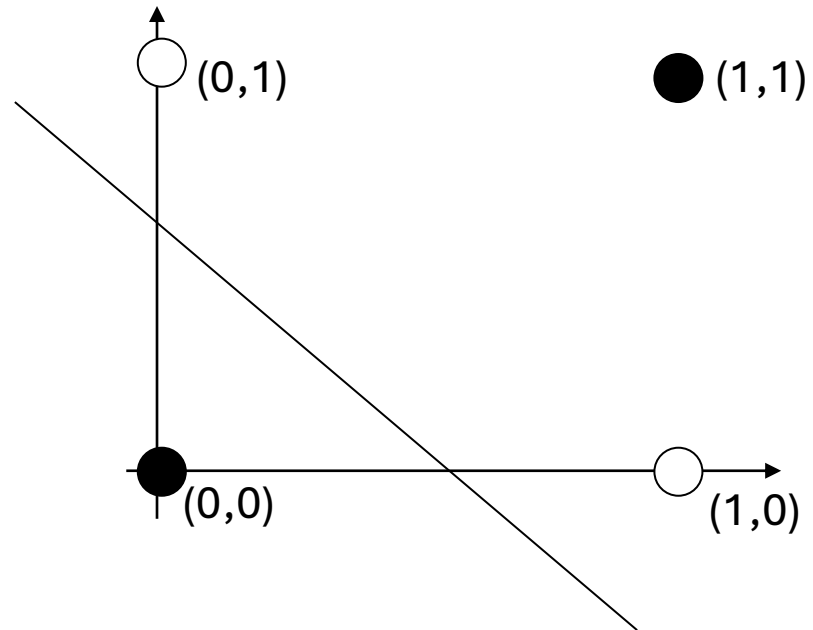
$1\ 1 \Rightarrow 0$



No linear function can separate the points

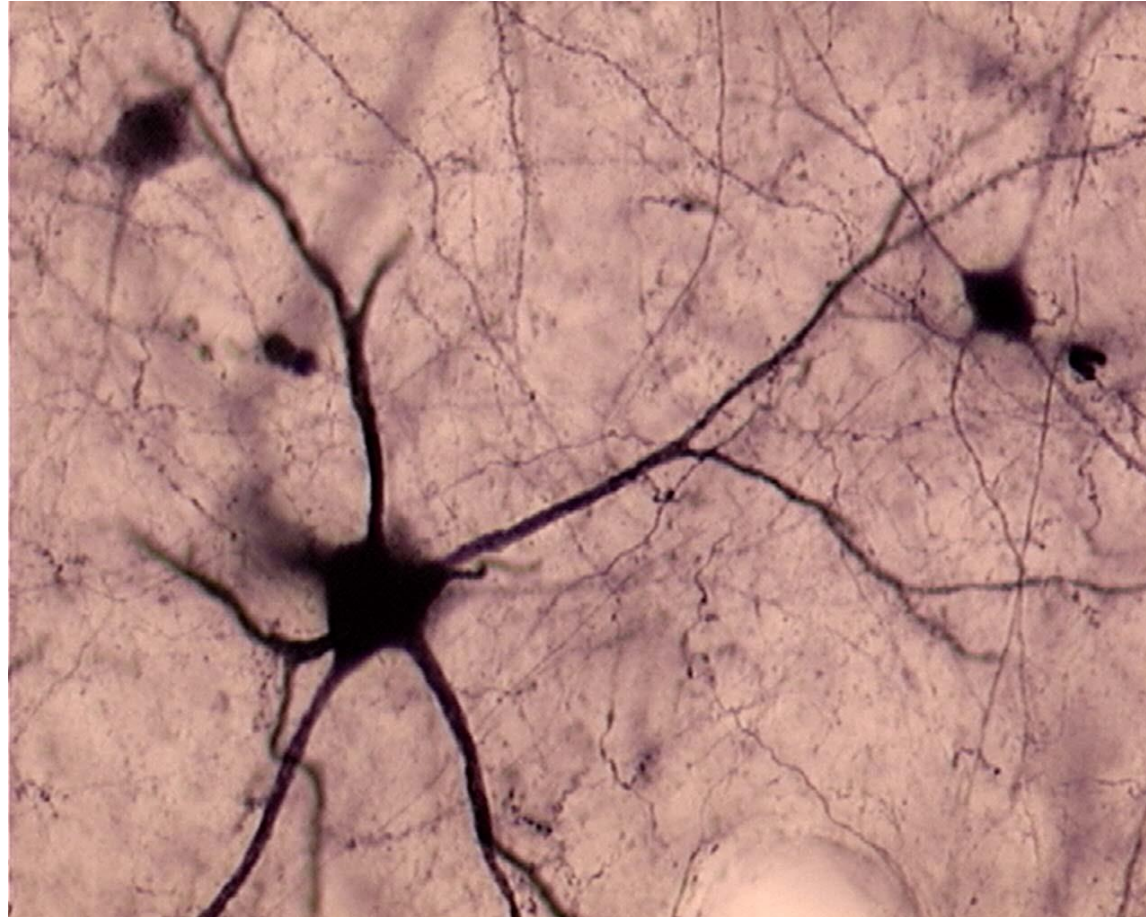
Error estimates

XOR	Predict	Error
0 0 \Rightarrow 0	0	0
1 0 \Rightarrow 1	1	0
0 1 \Rightarrow 1	1	0
1 1 \Rightarrow 0	1	1

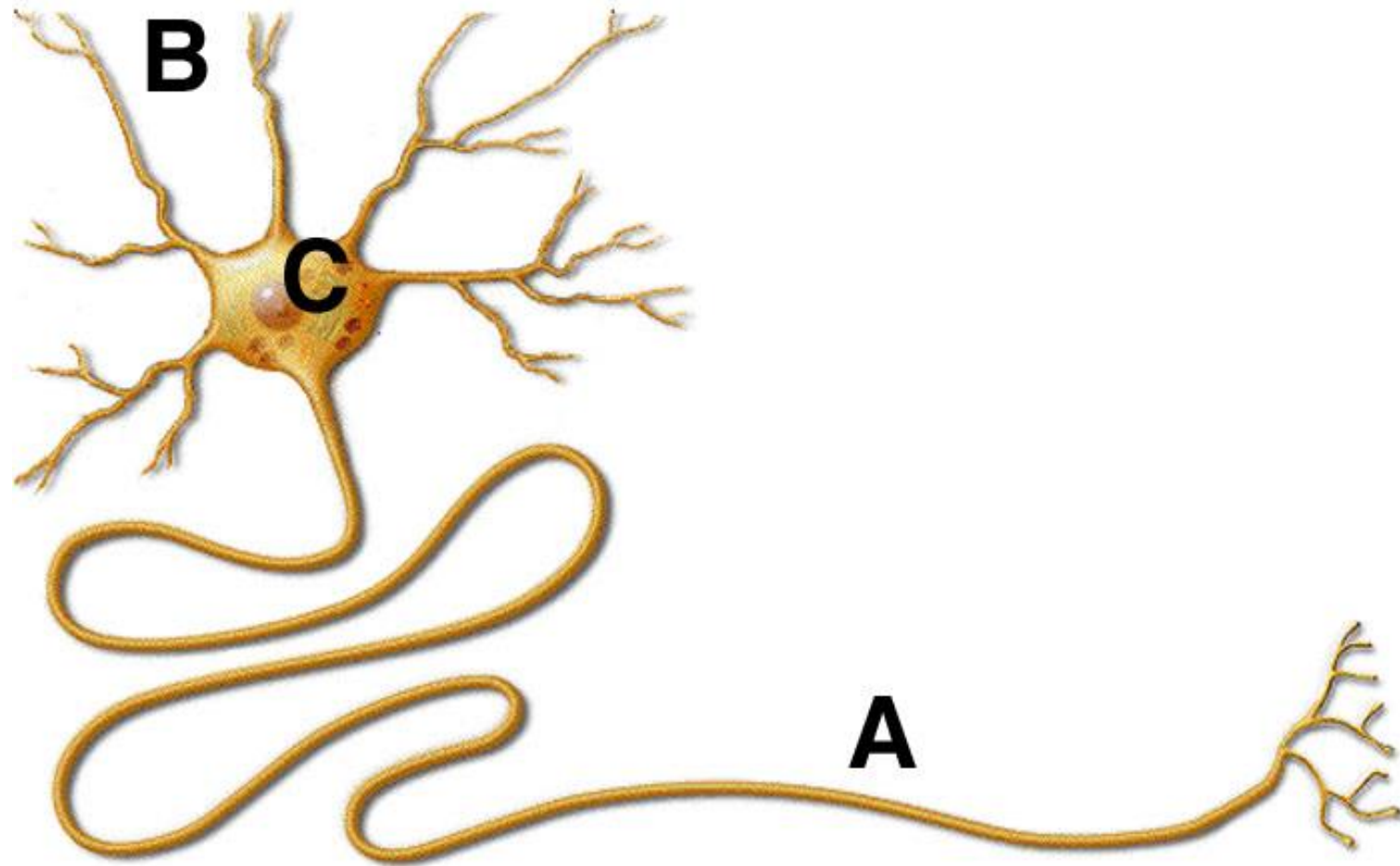


Mean error: $1/4$

Biological neural network



Biological neuron structure



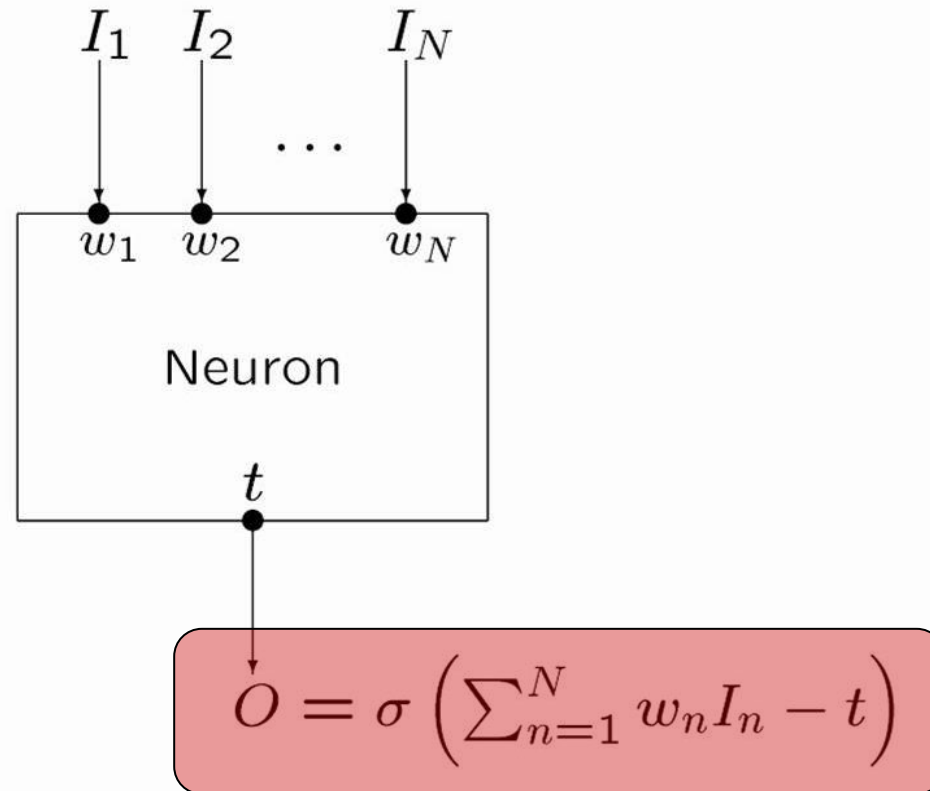
Artificial neural networks

Input signals

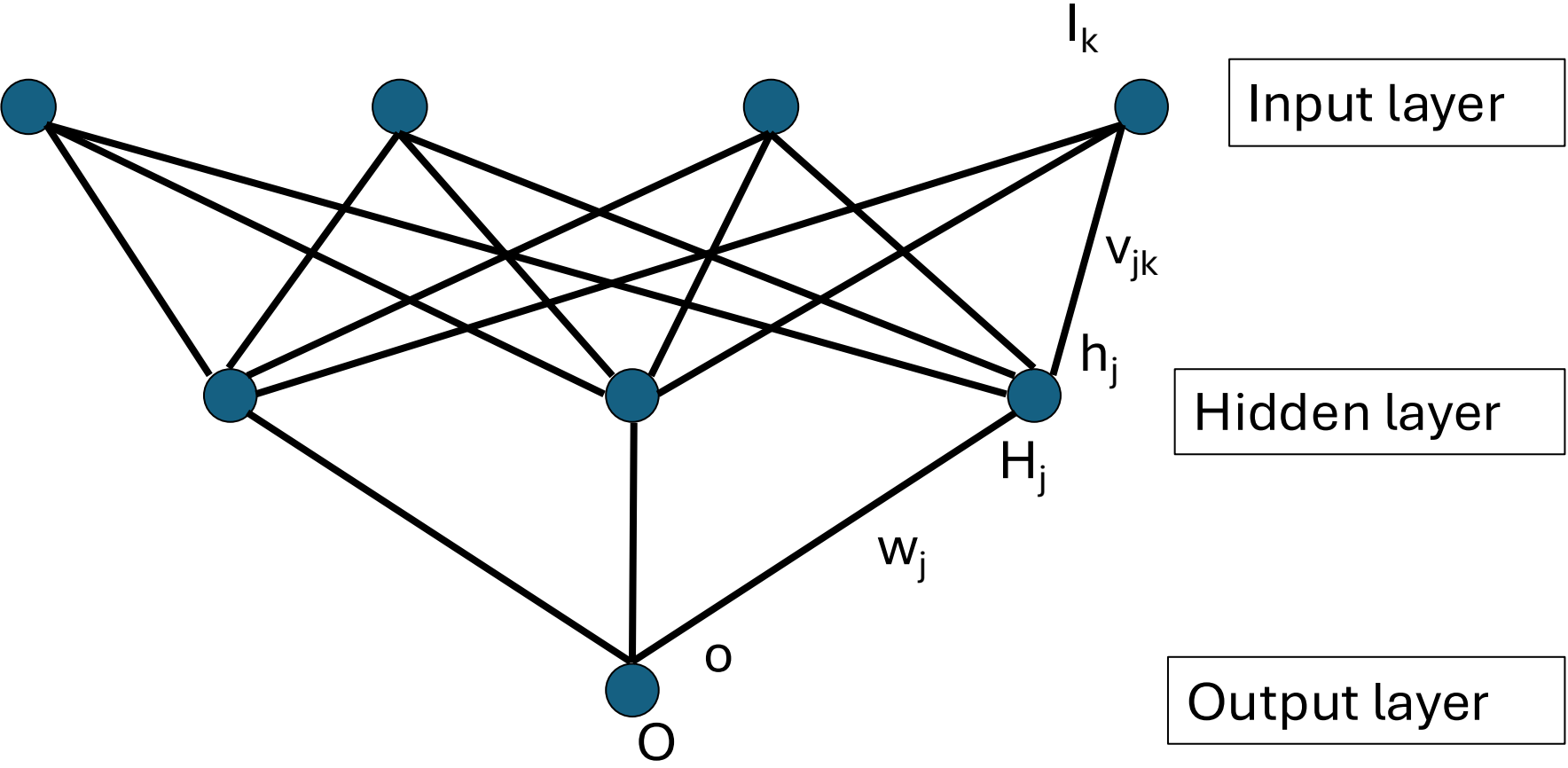
Synaptic weights

Threshold

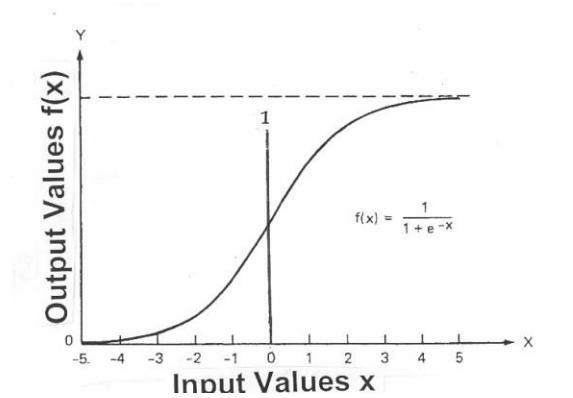
Output signal



Artificial neural network architecture

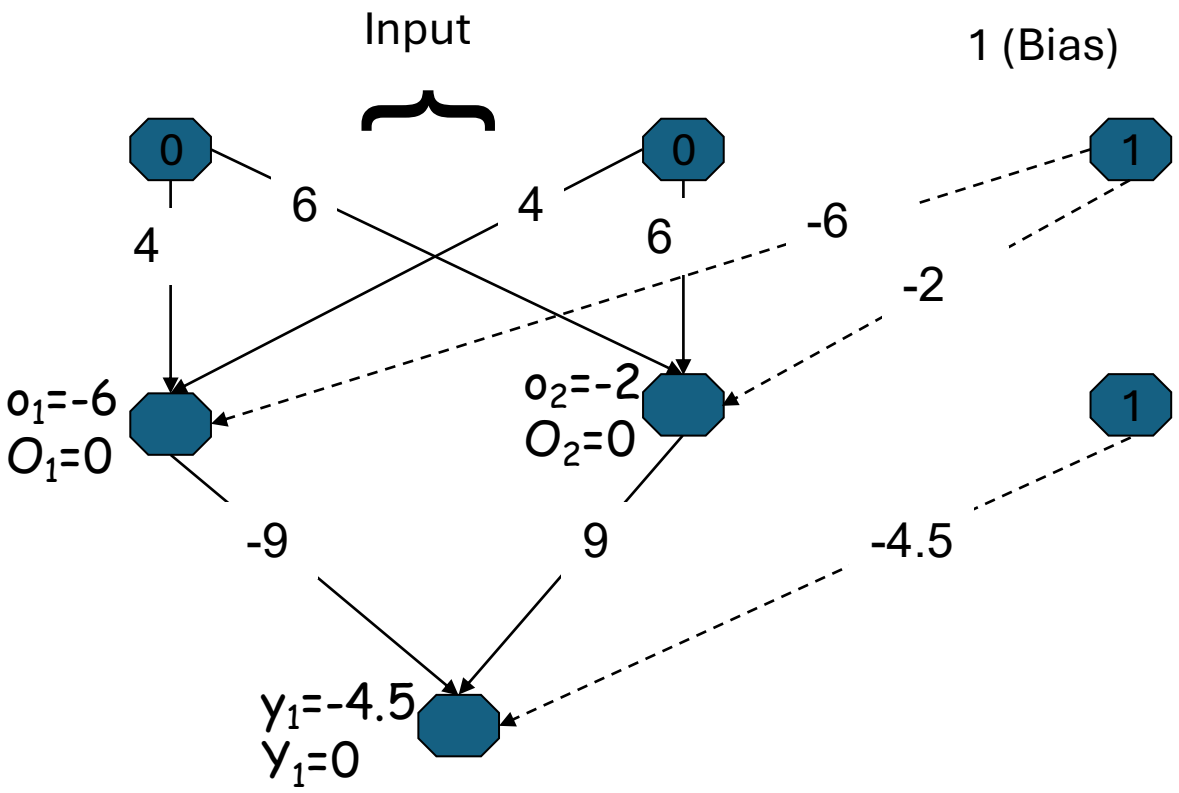


Neural networks: How does it work?



$$O = \frac{1}{1 + \exp(-o)}$$

$$o = \sum x_i \sum w_i$$

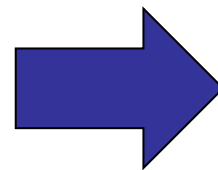
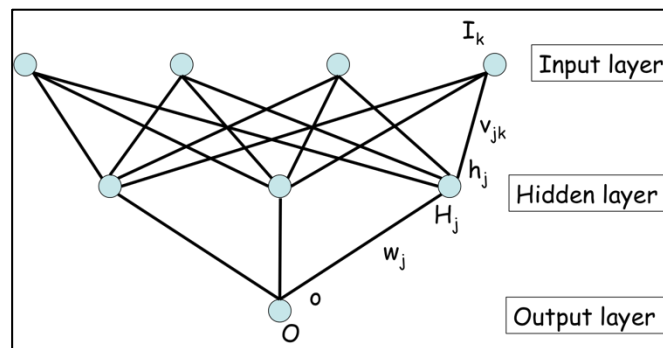


Now, do it yourself!

Data interpretation (fitting mathematical models)

AADFPGIAR	0.085
AAVDLSHFL	0.169
FTFDLTALK	0.085
WVWDTWPLA	0.085
TMMRHRREL	0.085
LLPYPIAGC	0.085
LMFSTSAYL	0.735
KLNENIIRF	0.536
MRVLHLDLK	0.085
GLICGLRQL	0.196
FEFILRYGD	0.085
EFVSANLAM	0.085
RAAHRQSV	0.085
SPLHVAV	0.085
RTFGKLPYR	0.085
GSLFTEQAF	0.197
SYGNANVSF	0.349
CSEVPQSGY	0.085
GSEDRDLIY	0.085

Artificial neural networks,
Support vector machines,
Similarity kernel,
Regression, ..



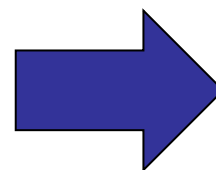
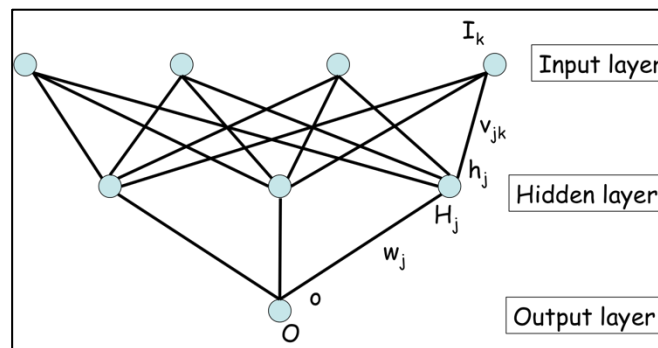
Machine learning

Data interpretation (fitting mathematical models)

AADFPGIAR 0.085
 AAVDLSHFL 0.169
 FTFDLTALK 0.085
 WVWDTWPLA 0.085
 TMMRHRREL 0.085
 LLPYPIAGC 0.085
 LMFSTSAYL 0.735
 KLNENIIRF 0.536
 MRVLHLDLK 0.085
 GLICGLRQL 0.196
 FEFILRYGD 0.085
 EFVSANLAM 0.085
 RAAHRRQSV 0.085
 SPLHV FVAV 0.085
 RTFGKLPYR 0.085
 GSLFTEQAF 0.197
 SYGNANVSF 0.349
 CSEVPQSGY 0.085
 GSEDRDLIY 0.085

Artificial neural networks,
 Support vector machines,
 Similarity kernel,
 Regression, ..

0.036
 0.227
 0.131
 0.147
 0.338
 0.082
 0.713
 0.467
 0.044
 0.239
 0.032
 0.162
 0.126
 0.050
 0.087
 0.392
 0.181
 0.169
 0.187



Machine learning

Training an ANN: Identify weights to get lowest error

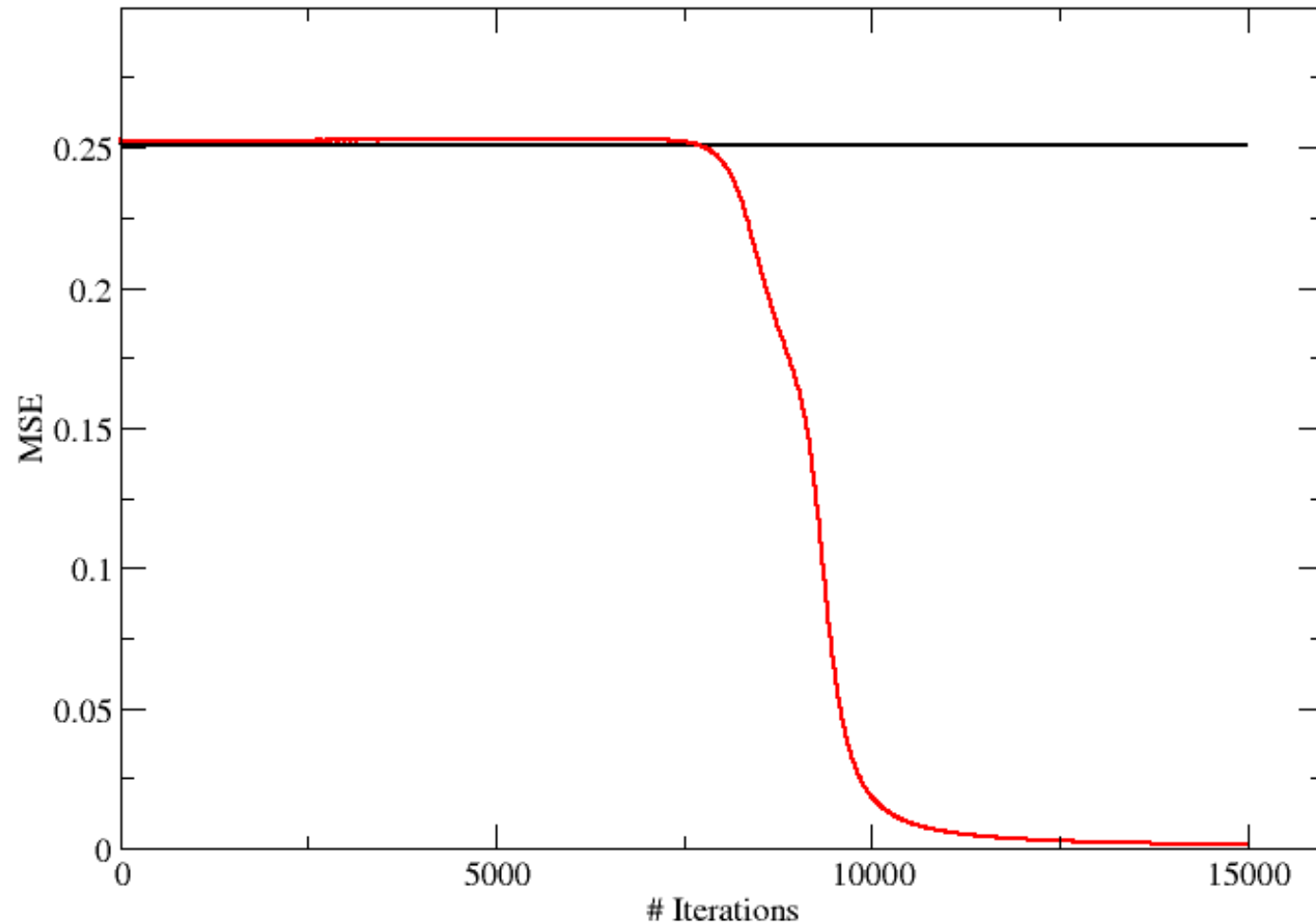
AAAKTPVIV 0.033693
AADFPGIAR 0.084687
ALVARAAVL 0.139013
FILIFNIIIV 0.891622
IMDQVPFSV 0.727865
DEFKLVPEW 0.084687
DEWECTRDD 0.084687
LLFLGVVFL 0.638438
LVFIKPPLI 0.630086
KVDDTFYYV 0.669121
FVDFVIHGL 0.864383
TMDPSRVVL 0.654552
YGPDVEVNV 0.084687
MTAEDMLTV 0.755627
MMVILPDKI 0.530313
APTGDLPRA 0.080705
SLTECPTFL 1.000000

$$w'_i = w_i + \Delta w_i$$

$$\Delta w_i = -\varepsilon \cdot \frac{\partial E}{\partial w_i}$$

$$E = \frac{1}{2} \cdot (O - t)^2$$

Neural networks and the XOR function

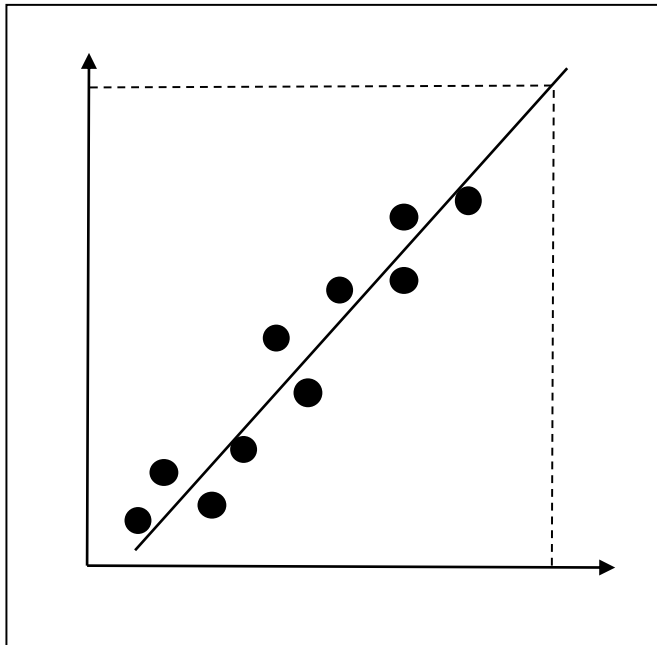


Demo:

<https://playground.tensorflow.org/>

How to train a method?

A simple statistical method: Linear regression



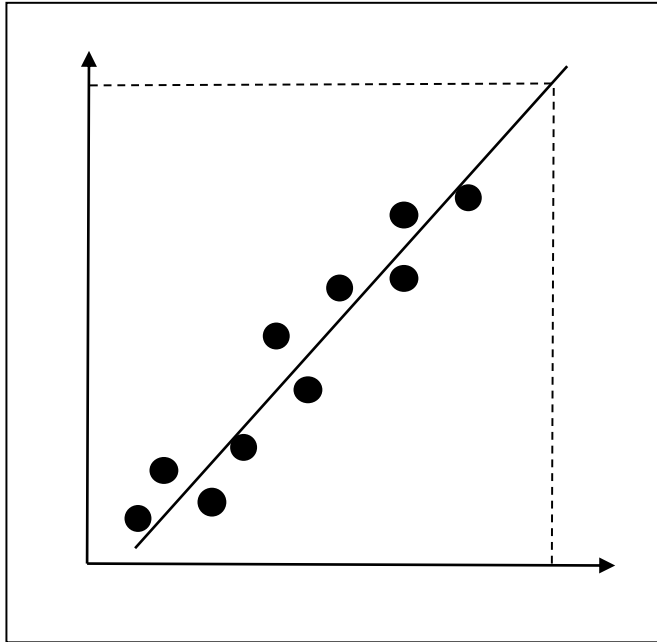
Observations (*training data*): a set of x values (*input*) and y values (*output*).

Model: $y = ax + b$ (2 *parameters*, which are estimated from the training data)

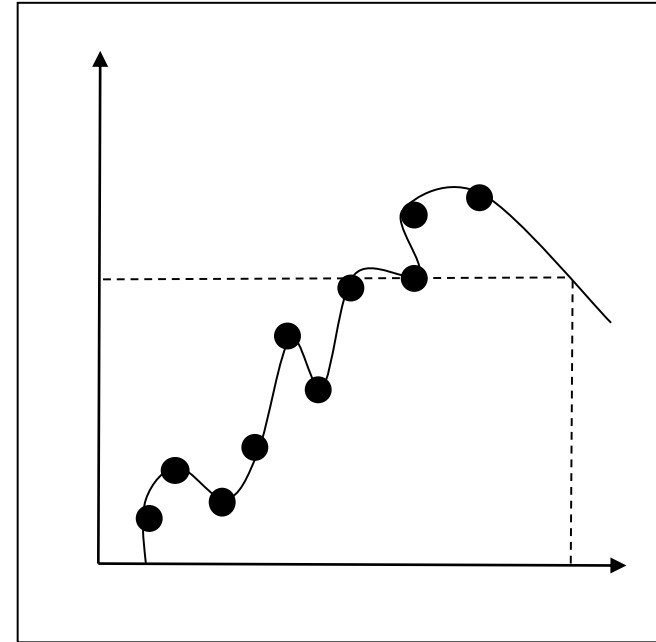
Prediction: Use the model to calculate a y value for a *new* x value

Note: the model does not fit the observations exactly. Can we do better than this?

Overfitting



$y = ax + b$
2 parameter model
Good description, poor fit



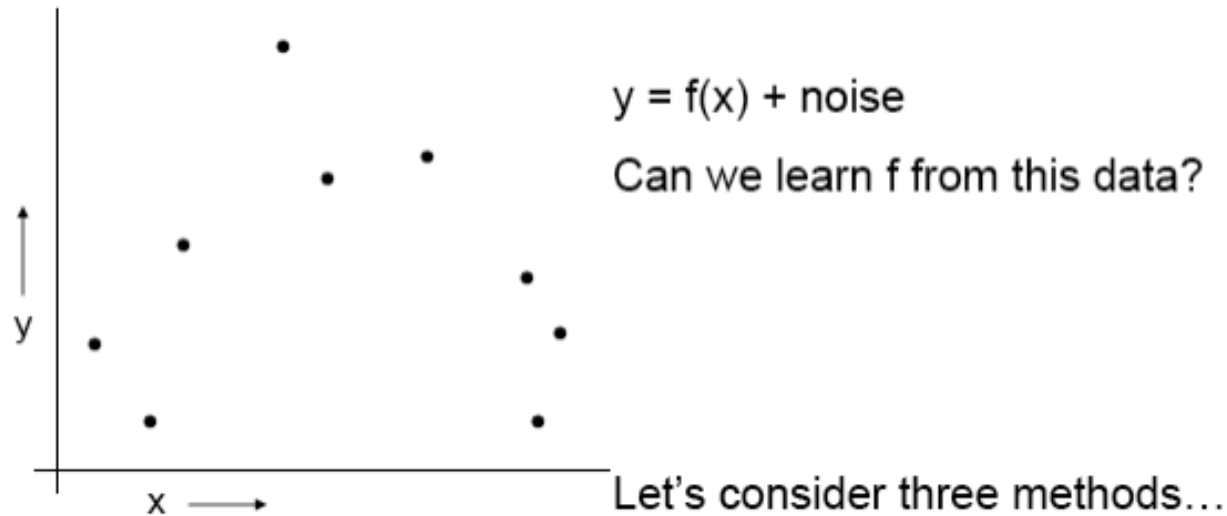
$y =$
 $ax^6 + bx^5 + cx^4 + dx^3 + ex^2 + fx + g$
7 parameter model
Poor description, good fit

Note: It is *not interesting* that a model can fit its observations (training data) exactly.

To function as a prediction method, a model must be able to *generalize*, i.e. produce sensible output on *new* data.

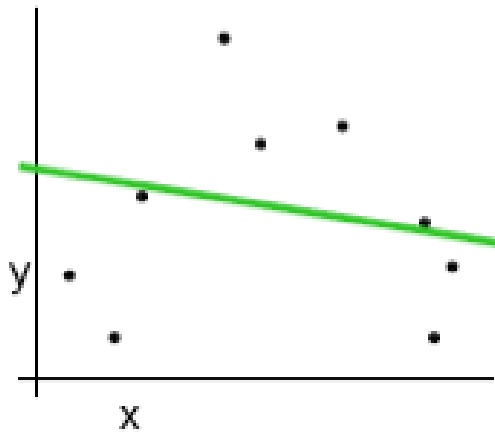
How to estimate parameters for prediction?

A Regression Problem

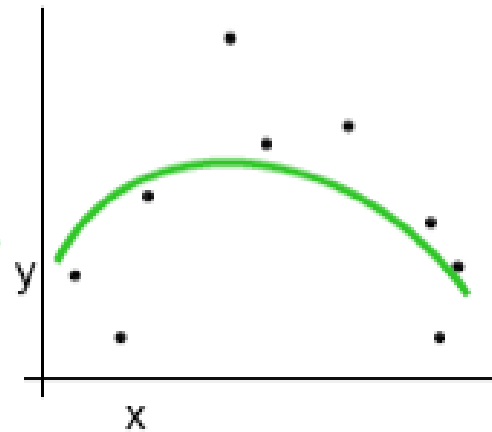


Model selection

Which is best?



Linear Regression

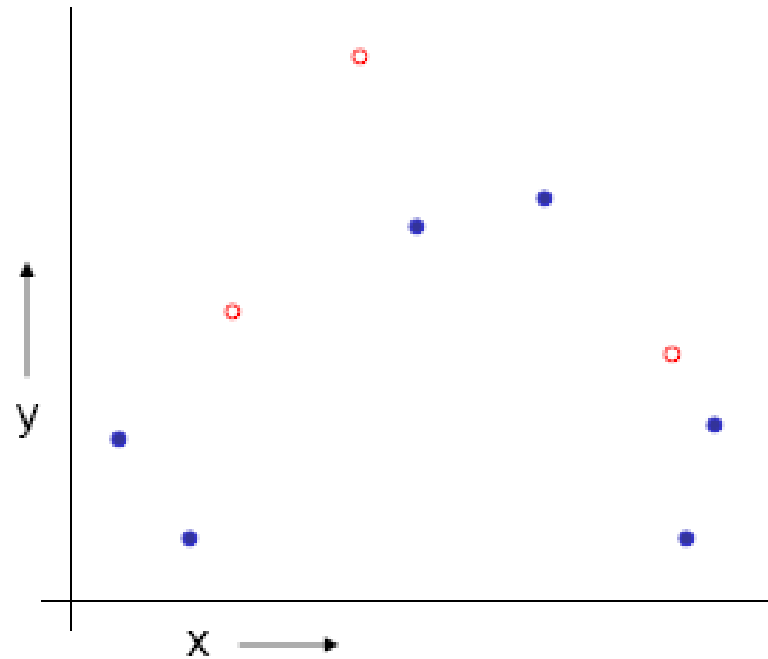


Quadratic Regression



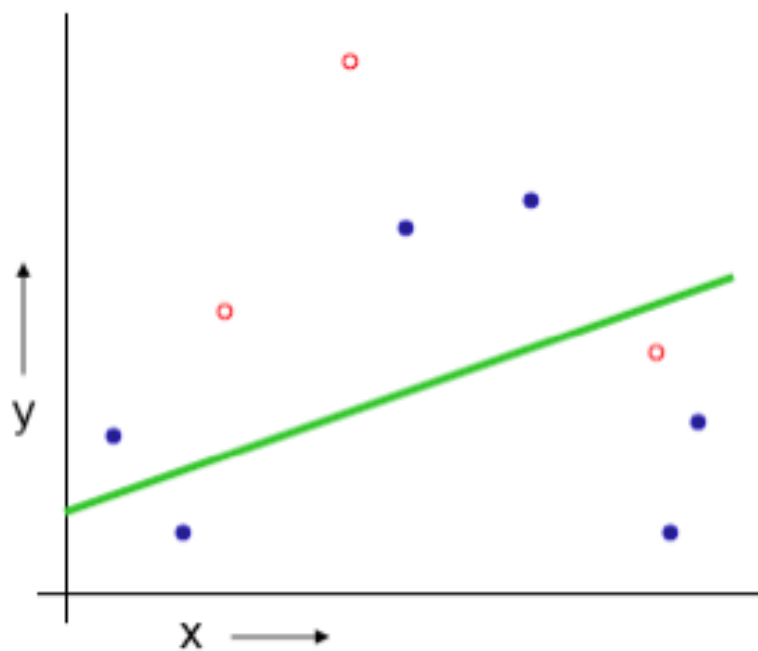
Join-the-dots

The test set method



1. Randomly choose 30% of the data to be in a **test set**
2. The remainder is a **training set**

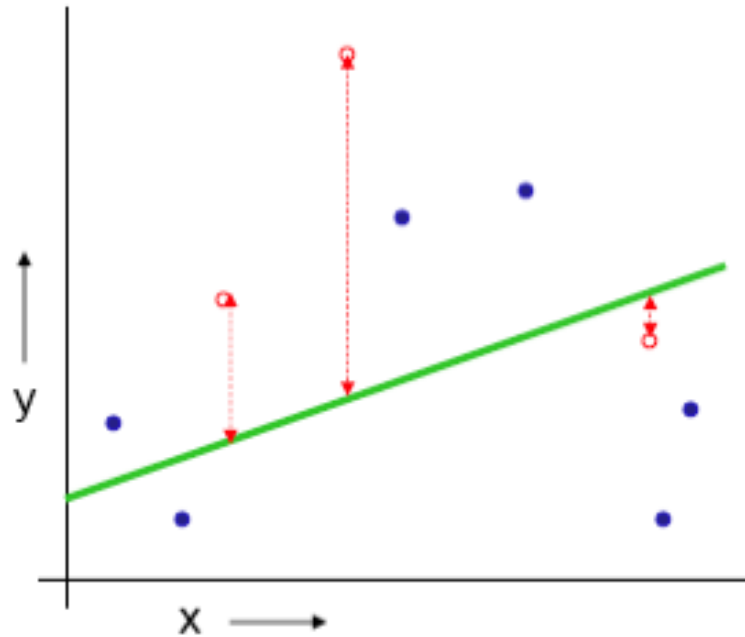
The test set method



(Linear regression example)

1. Randomly choose 30% of the data to be in a **test set**
2. The remainder is a **training set**
3. Perform your regression on the training set

The test set method

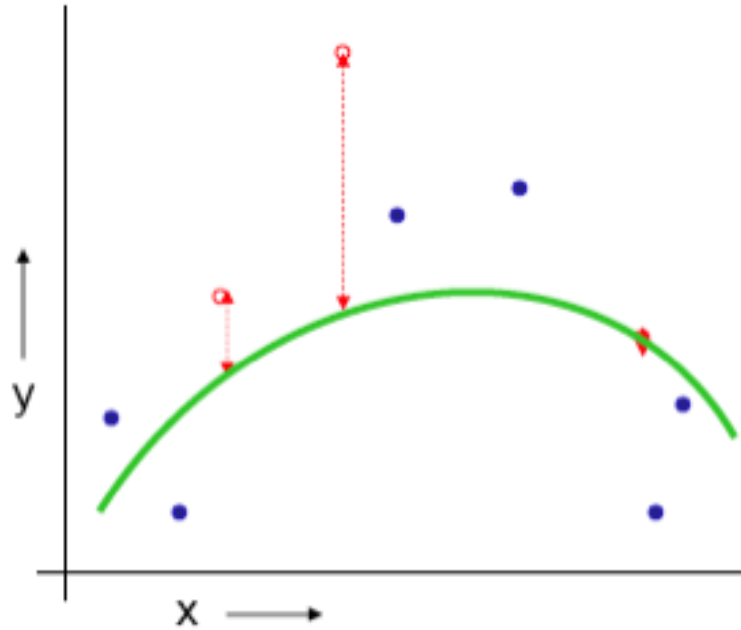


(Linear regression example)

Mean Squared Error = 2.4

1. Randomly choose 30% of the data to be in a **test set**
2. The remainder is a **training set**
3. Perform your regression on the training set
4. **Estimate your future performance with the test set**

The test set method

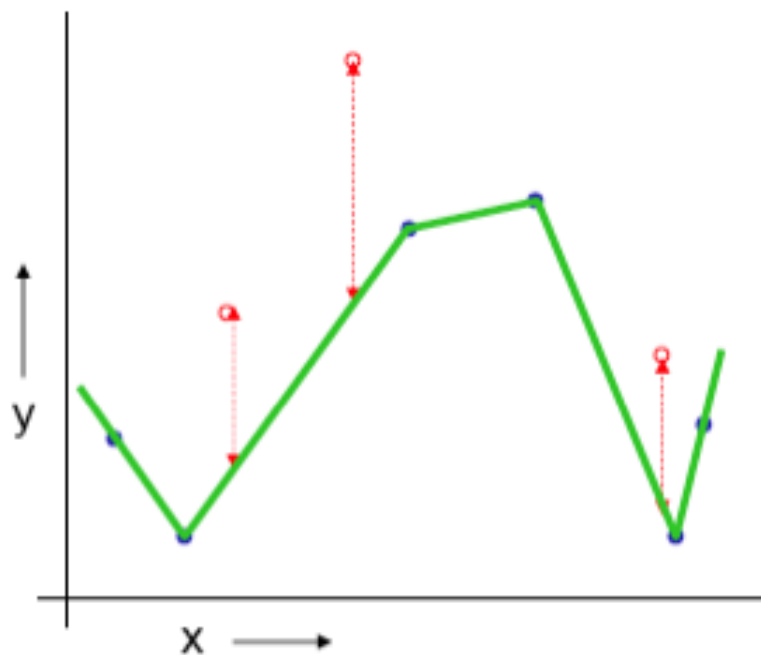


(Quadratic regression example)

Mean Squared Error = 0.9

1. Randomly choose 30% of the data to be in a **test set**
2. The remainder is a **training set**
3. Perform your regression on the training set
4. **Estimate your future performance with the test set**

The test set method



(Join the dots example)

Mean Squared Error = 2.2

1. Randomly choose 30% of the data to be in a **test set**

2. The remainder is a **training set**

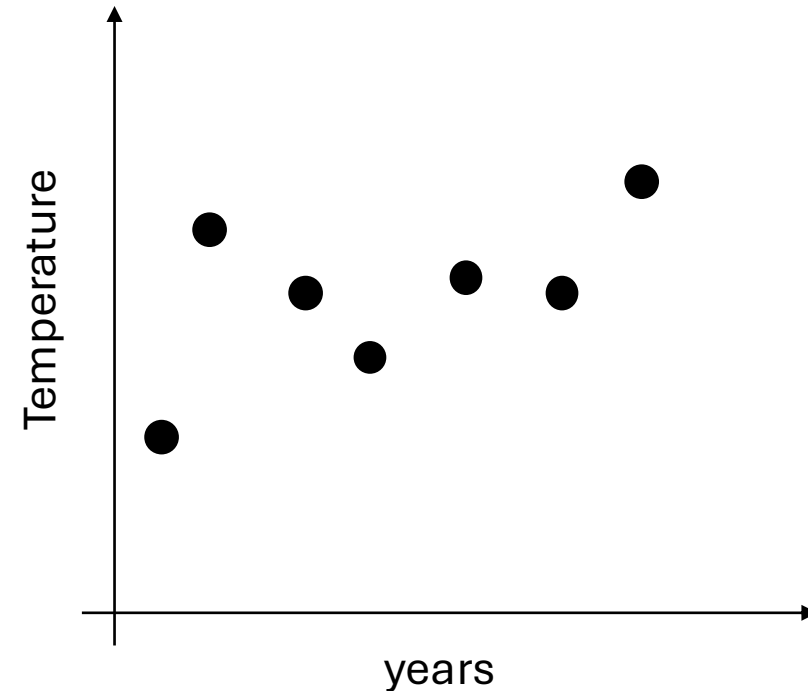
3. Perform your regression on the training set

4. Estimate your future performance with the **test set**

So quadratic function is best

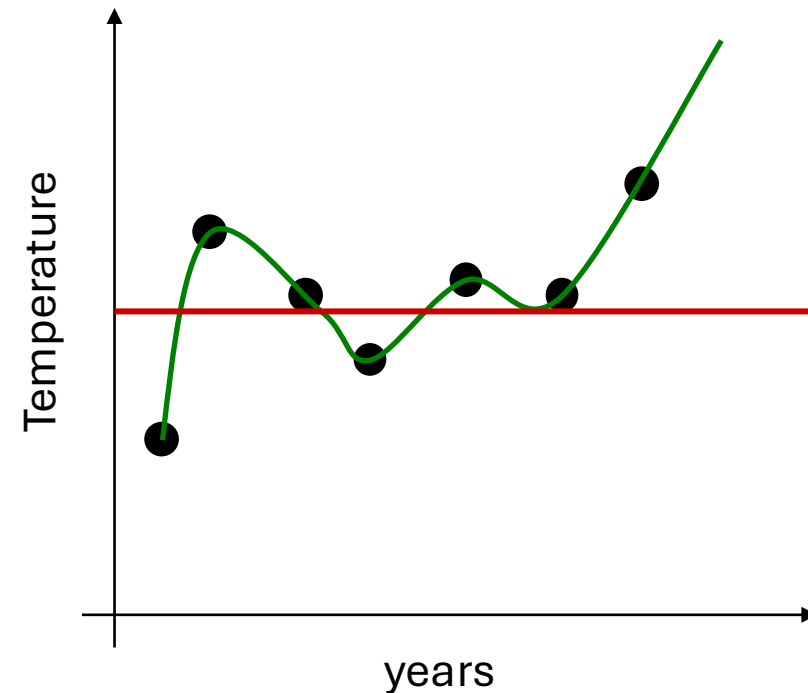
Neural network training

- A Network contains a very large set of parameters
 - A network with 5 hidden neurons predicting binding for 9meric peptides has more than $9 \times 20 \times 5 = 900$ weights
- Overfitting is a problem
- Stop training when test performance is optimal



Neural network training

- A Network contains a very large set of parameters
 - A network with 5 hidden neurons predicting binding for 9meric peptides has more than $9 \times 20 \times 5 = 900$ weights
- Overfitting is a problem
- Stop training when test performance is optimal



Neural network training. Cross validation

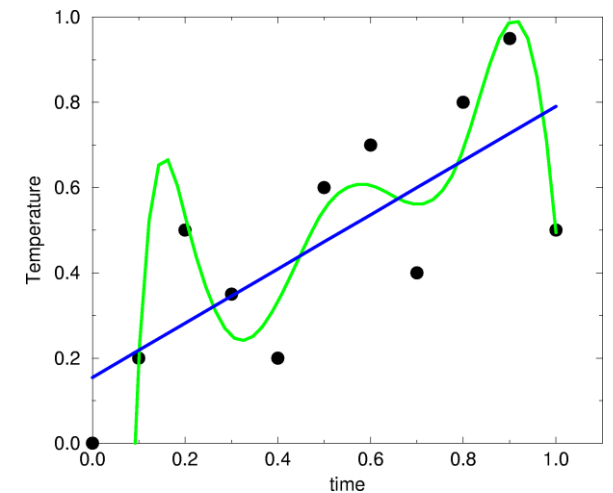
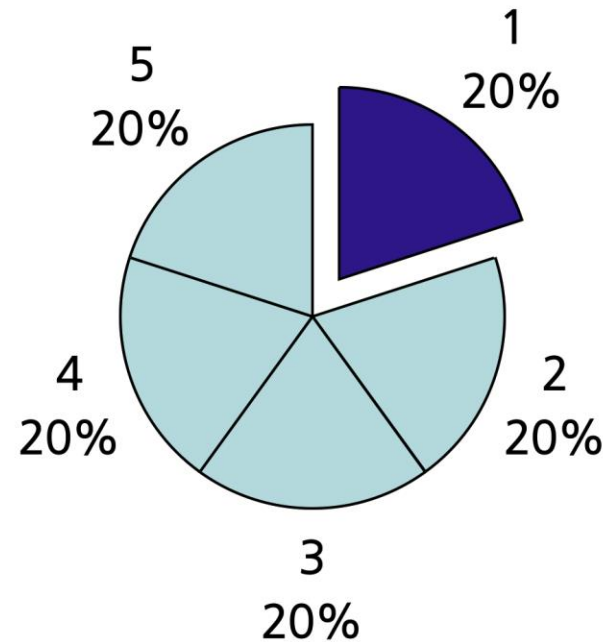
Cross validation

Train on 4/5 of data

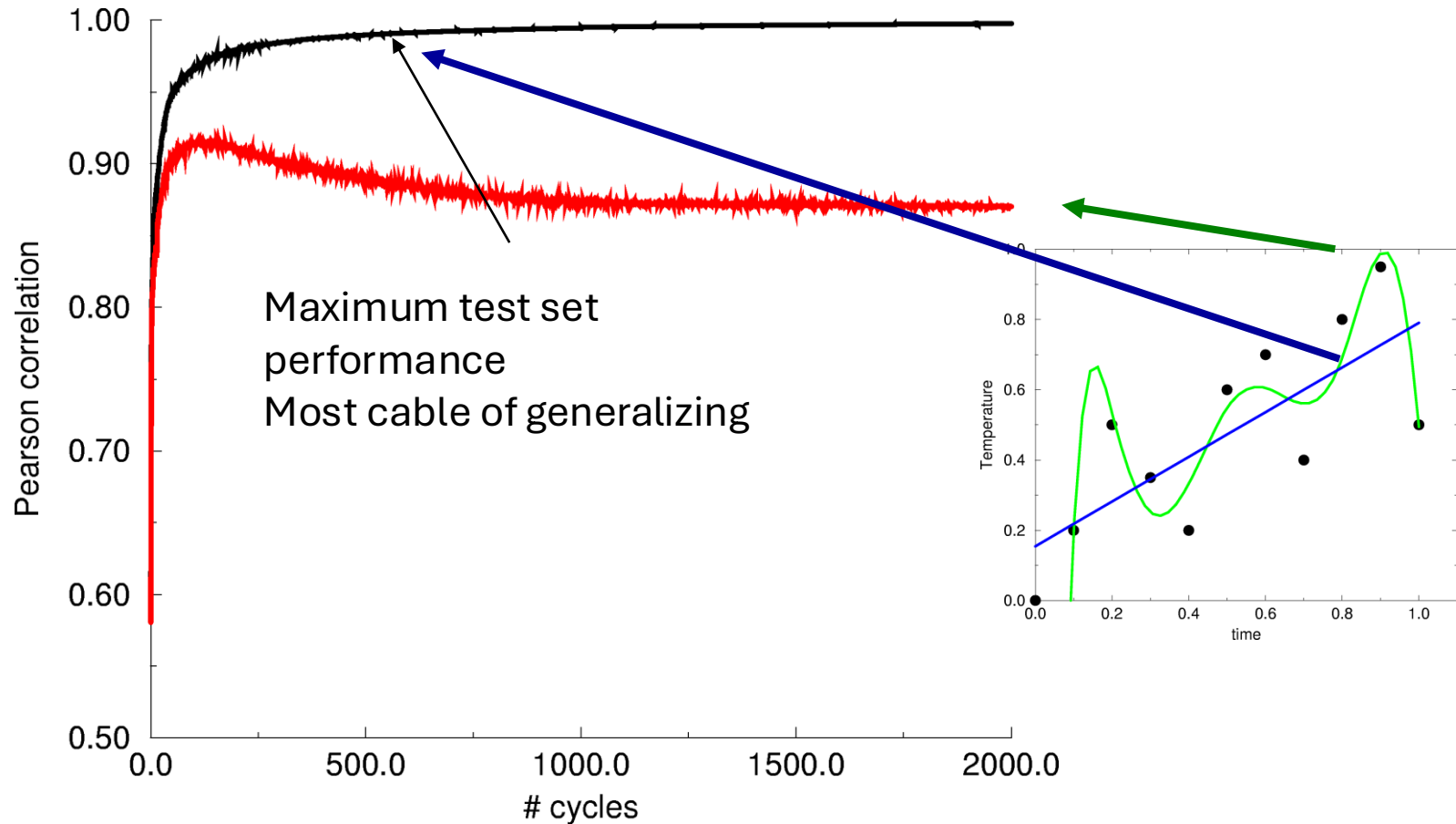
Test on 1/5

=>

Produce 5 different neural networks each with a different prediction focus

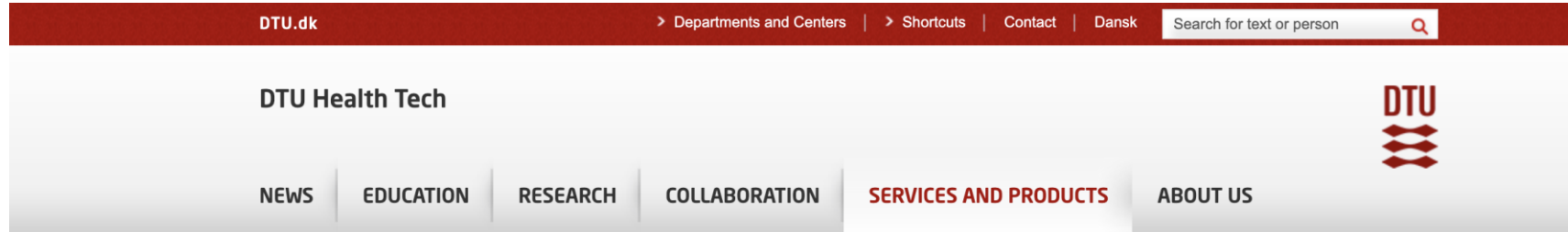


Neural network training curve



NetMHC

services.healthtech.dtu.dk/service.php?NetMHCpan-4.1



Home > [Services and Products](#)

NetMHCpan - 4.1

Pan-specific binding of peptides to MHC class I proteins of known sequence

The NetMHCpan-4.1 server predicts binding of peptides to **any MHC molecule of known sequence** using artificial neural networks (ANNs). The method is trained on a combination of more than 850,000 quantitative Bi 170 MHC molecules from human (HLA-A, B, C, E), mouse (H-2), cattle (BoLA), primates (Patr, Mamu, Gogo), swine (SLA) and equine (Eqca). The EL data covers 177 MHC molecules from human (HLA-A, B, C, E), mo (DLA). Furthermore, the user can obtain predictions to any custom MHC class I molecule by uploading a full length MHC protein sequence. Predictions can be made for peptides of any length.

Note, as of 28/7/2020 the server has been updated (retrained on data resolving a curation error in the IEDB for a single allele (SA) eluted ligand H2-Db/H2-Kb data set. This recuration only affected ~2000 H

To access the earlier version of NetMHCpan-4.1 click here [version 4.1a](#)

Note also, if you have installed the earlier version of NetMHCpan-4.1, click her e to download the updated data file [data.tar.gz](#), and a file with the update test directory [test.tar.gz](#).

The server returns as default the likelihood of a peptide being a natural ligand of the selected MHC(s). If selected, also the predicted binding affinity is rseported.

New in this version: together with Binding Affinity (BA) data, the method has now been trained on EL data from Single Allele (SA, peptides annotated to a single MHC) and Multi Allele (MA, peptides annotated to multip algorithm of NetMHCpan) called NNALign_MA (PMID: 31578220), which enables pseudo-labelling.

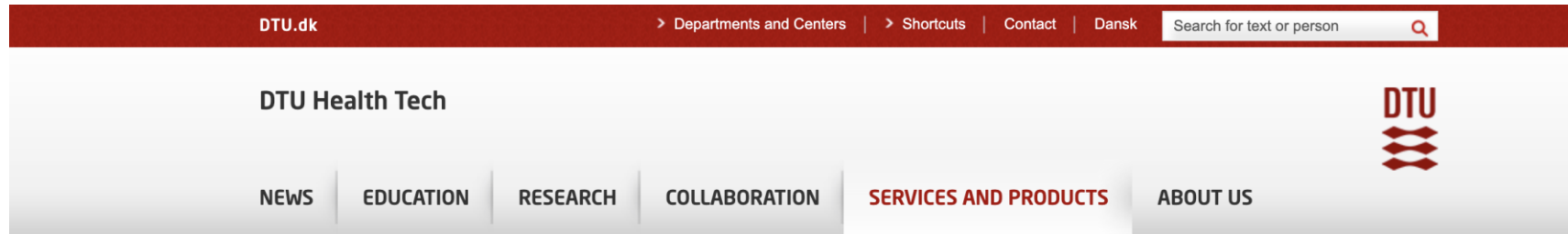
View the version history of this server. All previous versions are available online, for comparison and reference.

The project is a collaboration between CBS, and [LIAI](#).



NetMHC

services.healthtech.dtu.dk/service.php?NetMHCpan-4.1



Home > [Services and Products](#)

NetMHCpan - 4.1

Pan-specific binding of peptides to MHC class I proteins of known sequence

The NetMHCpan-4.1 server predicts binding of peptides to **any MHC molecule of known sequence** using artificial neural networks (ANNs). The method is trained on a combination of more than 850,000 quantitative Bi 170 MHC molecules from human (HLA-A, B, C, E), mouse (H-2), cattle (BoLA), primates (Patr, Mamu, Gogo), swine (SLA) and equine (Eqca). The EL data covers 177 MHC molecules from human (HLA-A, B, C, E), mo (DLA). Furthermore, the user can obtain predictions to any custom MHC class I molecule by uploading a full length MHC protein sequence. Predictions can be made for peptides of any length.

Note, as of 28/7/2020 the server has been updated (retrained on data resolving a curation error in the IEDB for a single allele (SA) eluted ligand H2-Db/H2-Kb data set. This recuration only affected ~2000 H

To access the earlier version of NetMHCpan-4.1 click here [version 4.1a](#)

Note also, if you have installed the earlier version of NetMHCpan-4.1, click here to download the updated data file [data.tar.gz](#), and a file with the update test directory [test.tar.gz](#).

The server returns as default the likelihood of a peptide being a natural ligand of the selected MHC(s). If selected, also the predicted binding affinity is reported.

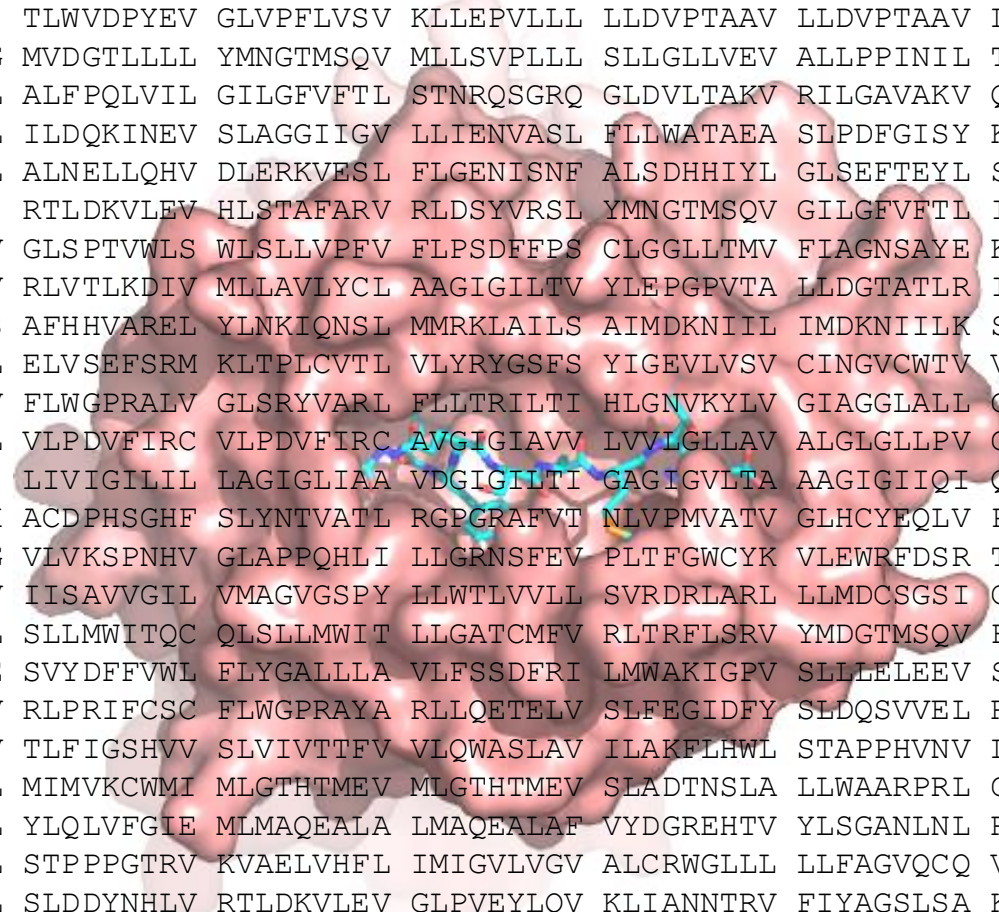
New in this version: together with Binding Affinity (BA) data, the method has now been trained on EL data from Single Allele (SA, peptides annotated to a single MHC) and Multi Allele (MA, peptides annotated to multip algorithm of NetMHCpan) called NNALign_MA (PMID: 31578220), which enables pseudo-labelling.

View the version history of this server. All previous versions are available online, for comparison and reference.

The project is a collaboration between CBS, and [LIAI](#).

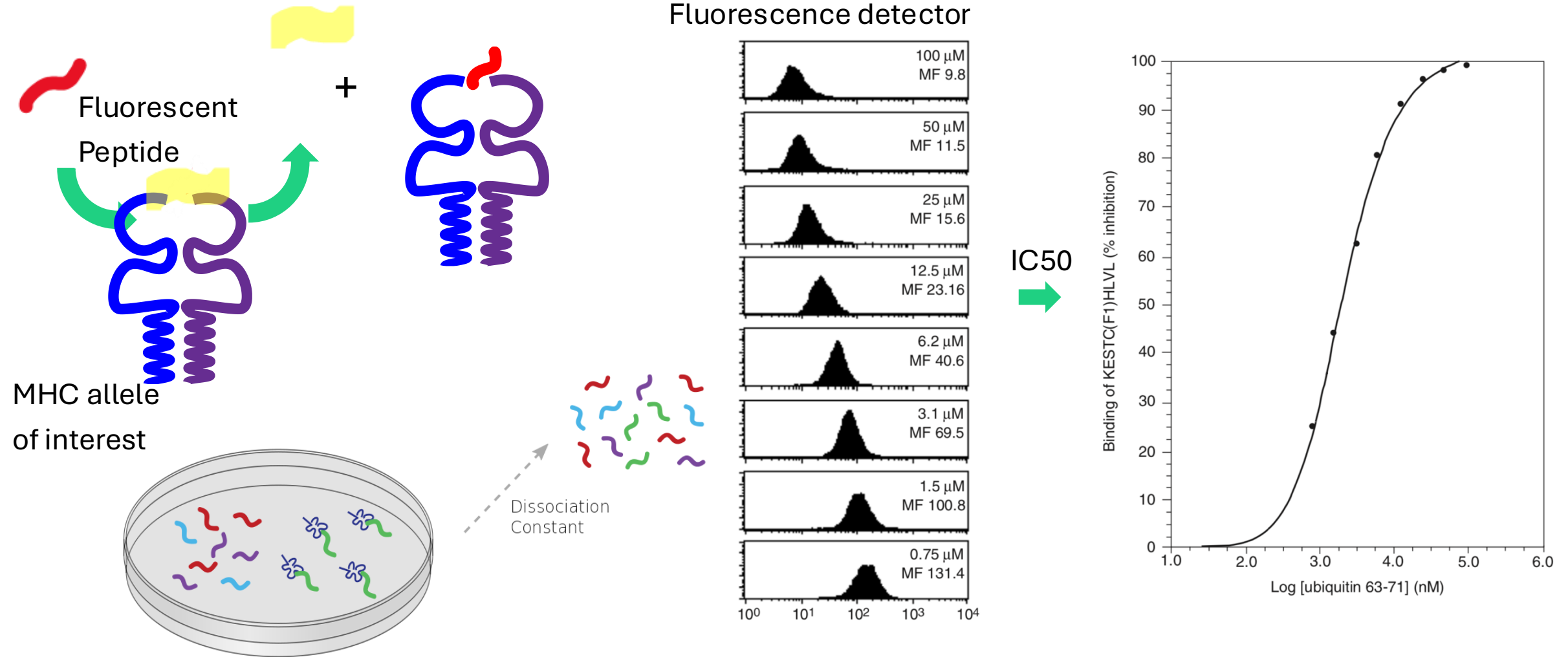


Where is the data coming from?



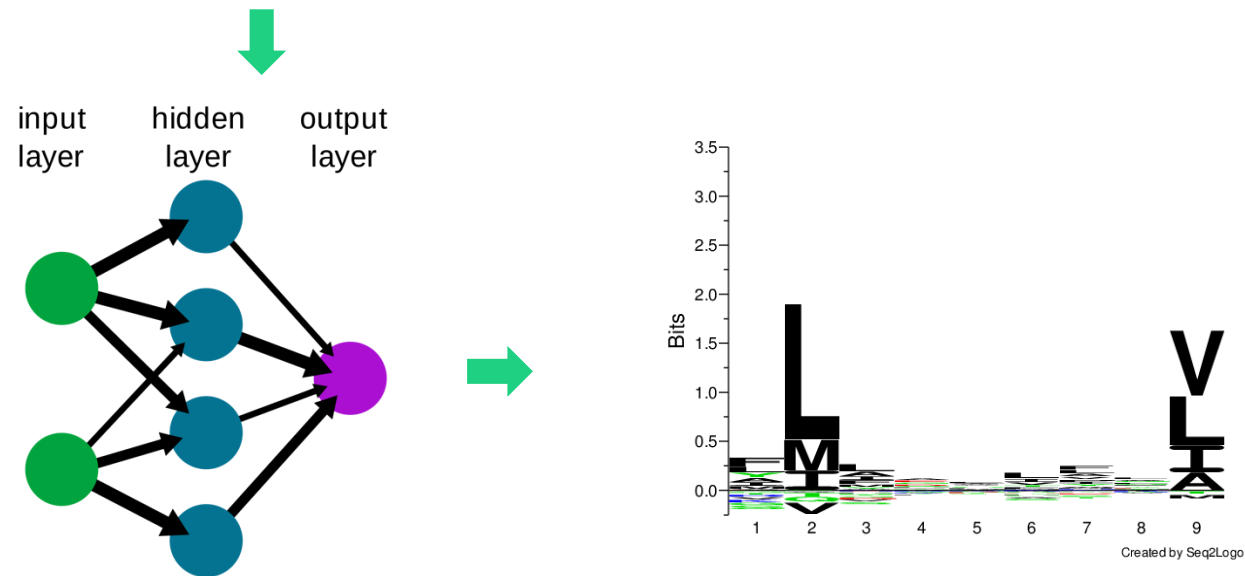
SLLPAIVEL YLLPAIVHI TLWVDPYEV GLVPFLVSV KLLPEVLLL LLDVPTAAV LLDVPTAAV LLDVPTAAV
LLDVPTAAV VLFRGGPRG MVDGTL LLL YMNGTMSQV MLLSVPLLL SLLGLLVEV ALLPPINIL TLIKIQHTL
HLIDYLVTS ILAPPVVKL ALFPQLVIL GILGFVFTL STNRQSGRQ GLDVLTAKV RILGAVAKV QVCERIP TI
ILFGHENRV ILMEH I HKL IL D Q K I N E V S L A G G I I G V L L I E N V A S L F L L W A T A E A S L P D F G I S Y K K R E E A P S L
LERPGGNEI ALSNLEVKL ALNELLQHV DLERKVESL FLGENISNF ALSDHHIYL GLSEFTEYL STAPPAHGV
PLDGEYFTL GVLVGVALI RTLDKVLEV HLSTAFARV RLDSYVRS L Y M N G T M S Q V G I L G F V F T L I L K E P V H G V
ILGFVFTLT LLFGYPVYV GLSPTVWLS WLSLLVPFV FLPSDFFPS CLGGLLTMV FIAGNSAYE KLGEFYNQM
KLVALGINA DLMGYIPLV RLVTLKDIV MLLAVLYCL AAGIGILTV YLEPGPVTA LLDGTATLR ITDQVPFSV
KTWGQYWQV TITDQVPFS AFHHVAREL YLNKIQNSL MMRKLAILS AIMDKNIIL IMDKNIILK SMVGNWAKV
SLLAPGAKQ KIFGSLAFL ELVSEFSRM KLTPLCVTL VLYRYGSFS YIGEVLSV CINGVCWTV VMNILLQYV
ILTVILGVL KVLEYVIKV FLWGPRALV GLSRYVARL FLLTRILTI HLGNVKYL V GIAGGLALL GLQDCTMLV
TGAPV TYST VIYQY MDDL VLPDV FIRC VLPDV FIRC AV G I G I A V V L V V L G L L A V A L G L G L L P V G I G I G V L A A
GAGIGVAVL IAGIGILAI LIVIGILIL LAGIGLIAA VDGIGILTI GAGIGVLT A AAGIGIIQI QAGIGILLA
KARDPHSGH KACDPHSGH ACDPHSGHF SLYNTVATL RGPGRFVT NLVPMVATV GLHCYEQLV PLKQHFQIV
AVFDRKSDA LLDFVRFMG VLVKSPNHV GLAPPQHLI LLGRNSFEV PLTFGW CYK VLEWRFD SR TLNAWVKV V
GLCTLVAML FIDSYICQV IISAVVGIL VMAGVGS PY LLWTLVLL SVRDR LARL LLMDCSGSI CLTSTVQLV
VLHDDLLEA LMWITQCFL SLLMWITQC QLSLLMWIT LLGATCMFV RLTRFLSRV YMDGTMSQV FLTPKKLQC
ISNDVCAQV VKTDGNPPE SVYDFVWL FLYGALLA VLFSSDFRI LMWAKIGPV SLLLELEE V SLSRFSWGA
YTAFTIPSI RLMKQDFSV RLPRIFCSC FLWGPRAYA RLLQETELV SLFEGIDFY SLDQSVVEL RLNMFTPYI
NMFTPYIGV LMI I PLINV TLF IGSHV V SLVIVTTFV VLQWASLAV ILAKFLHWL STAPPHVNV LLLLT V LTV
VVLGVVFGI ILHNGAYSL MIMVKC WMI MLGTH TMEV MLGTH TMEV SLADTNSLA LLWAAR PRL GVALQTMKQ
GLYDGM EHL KMVELVHFL YLQLVFGIE MLMAQEALA LMAQEALAF VYDGREHTV YLSGANLNL RMFPNAPYL
EAAGIGILT TLDSQVMSL STPPGTRV KVAELVHFL IMIGVLVGV ALCRWGLLL L LFAGVQCQ VLLCESTAV
YLSTAFARV YLLEMLWRL SLDDYNHLV RTLDKVLEV GLPVEYLQV KLIANNTRV FIYAGSLSA KLVANNTRL
FLDEFMEGV ALQPGTALL VLDGLDVLL SLYSFPEPE ALYVDSLFF SLLQHLIGL ELTLGEFLK MINAYLDKL
AAGIGILTV FLPSDFFPS SVRDR LARL SLREWLLRI LLSAWILTA AAGIGILTV AVPDEIPPL FAYDGKDYI
AAGIGILTV FLPSDFFPS AAGIGILTV FLPSDFFPS AAGIGILTV FLWGPRALV ETVSEQSNV ITLWQRPLV

Binding affinity assay



First MHC predictors used binding affinity data

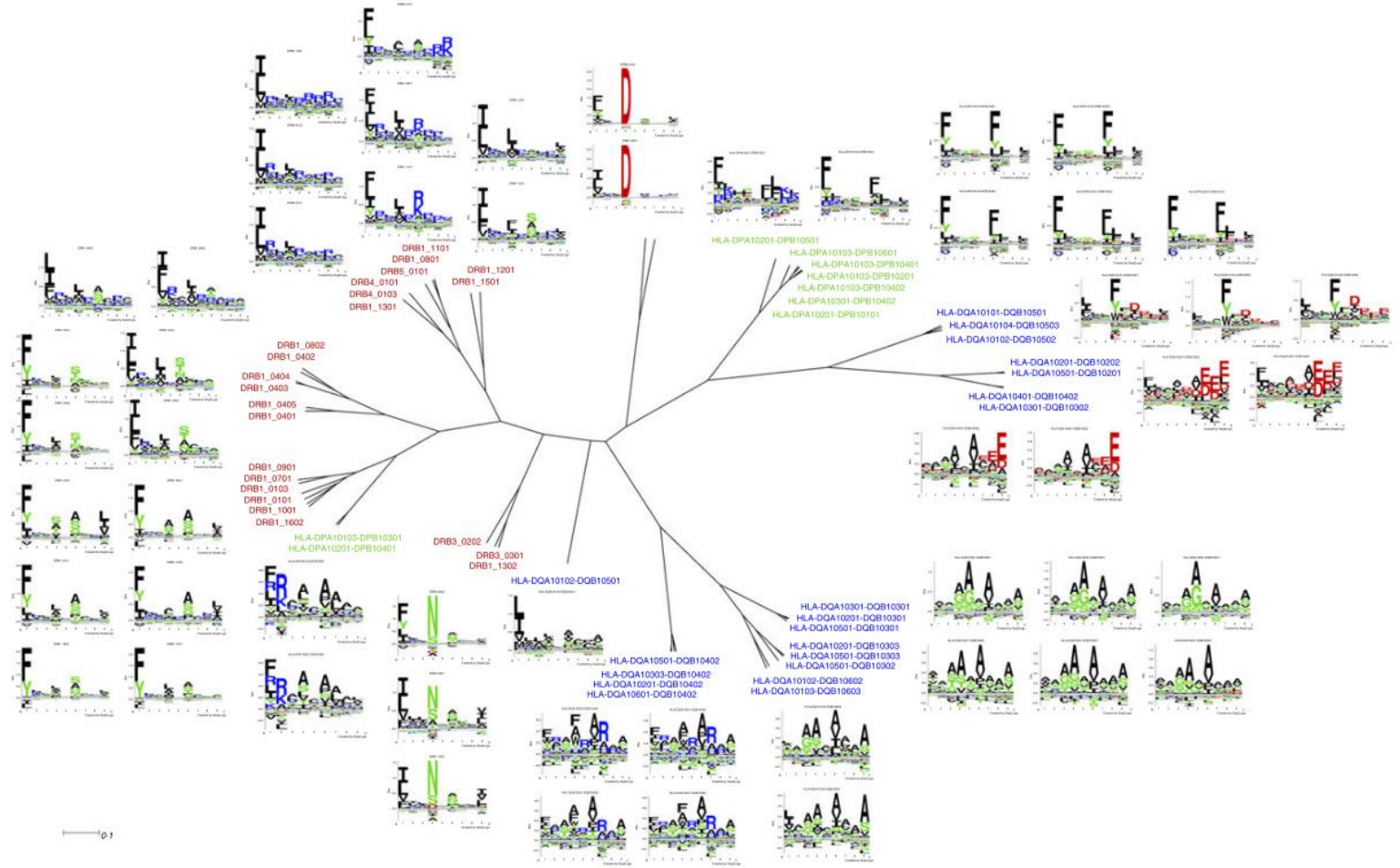
```
FMIDWILDA YFAMYGEKVAHTHVD TLYVRYHYITWAVLAYTWY 0.89 A0201
FMIDWILDA YFAMYQENMAHTDANTLY I IYRDYTWVARVYRGY 0.08 A0101
DSDGSFFLY YFAMYGEKVAHTHVD TLYVRYHYITWAVLAYTWY 0.08 A0201
DSDGSFFLY YFAMYQENMAHTDANTLY I IYRDYTWVARVYRGY 0.85 A0101
```



NetMHCIIpan-3.2 predicts binding affinity to MHC-II



IMMUNE EPITOPE DATABASE
AND ANALYSIS RESOURCE

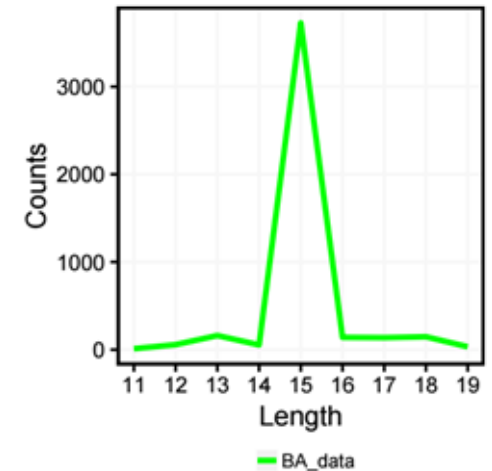
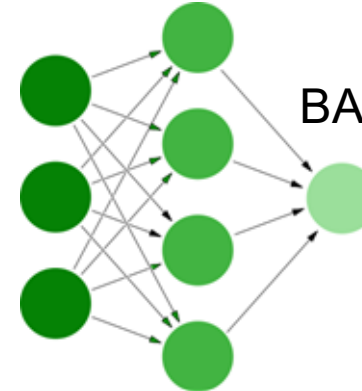


Challenges of binding affinity data

ALAKAAAAM
 ALAKAAAAN
 ALAKAAAAR
 ALAKAAAAT
 ALAKAAAAV
 GMNERPILT
 GILGFVFTM
 TLNAWVKVV
 KLNEPVLLL
 AVVPFIVSV

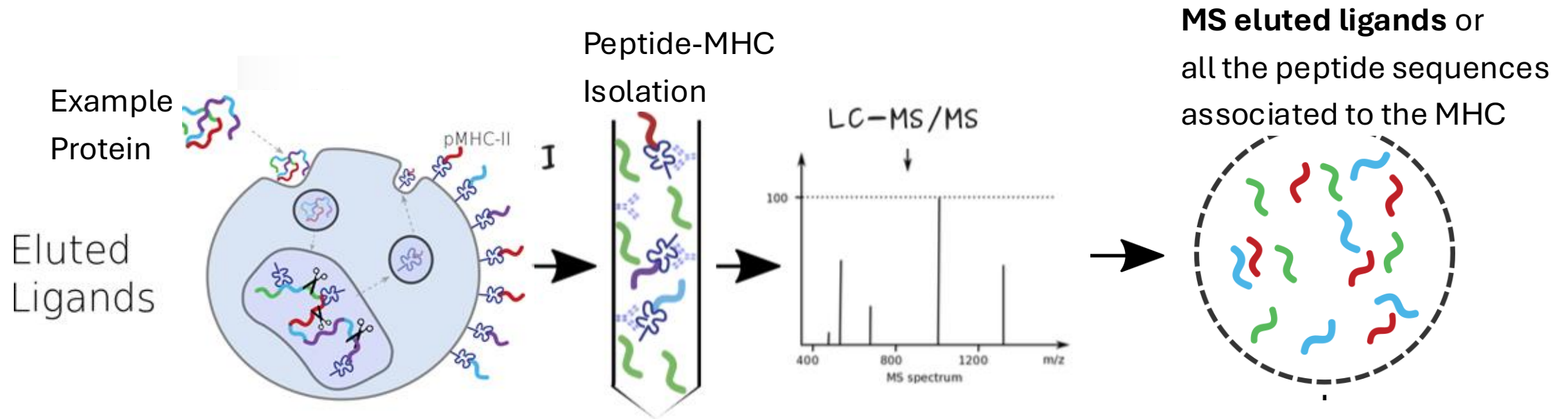
} Similar
 sequences
 Weight 1/5

AAYKLAYKTAEGATP	0.208682
RMMEYGTTMVSYQLV	0.305733
VDKCLELAELYLNII	0.0
PRYFNQLSTGLDMDN	0.357937
TFSSSEIRVGDELLER	0.376583
LSGHAFGAMAKKGDE	0.63762
RPQVPLRPMTYKGAF	0.576544
INAGFKAALAAAAGV	0.84466
IVGILLVLMAVVLAS	0.0
TTLAEMSTPEAT	0.18153
LAATVLLGCTSAKVH	0.588207
FIKVRQYDQILIEIL	0.397951
VSGLSIGTGRAMLGT	0.0
GLLMSRKHKWKLSL	0.0



- Biased sequences -> Alanine
- Over-representation of paradigmatic HLA alleles A*02:01, DRB1*01:01
- Artificial length distribution -> 15mers
- NetMHCpan was used to “design” new synthetic peptides and later feedback loop

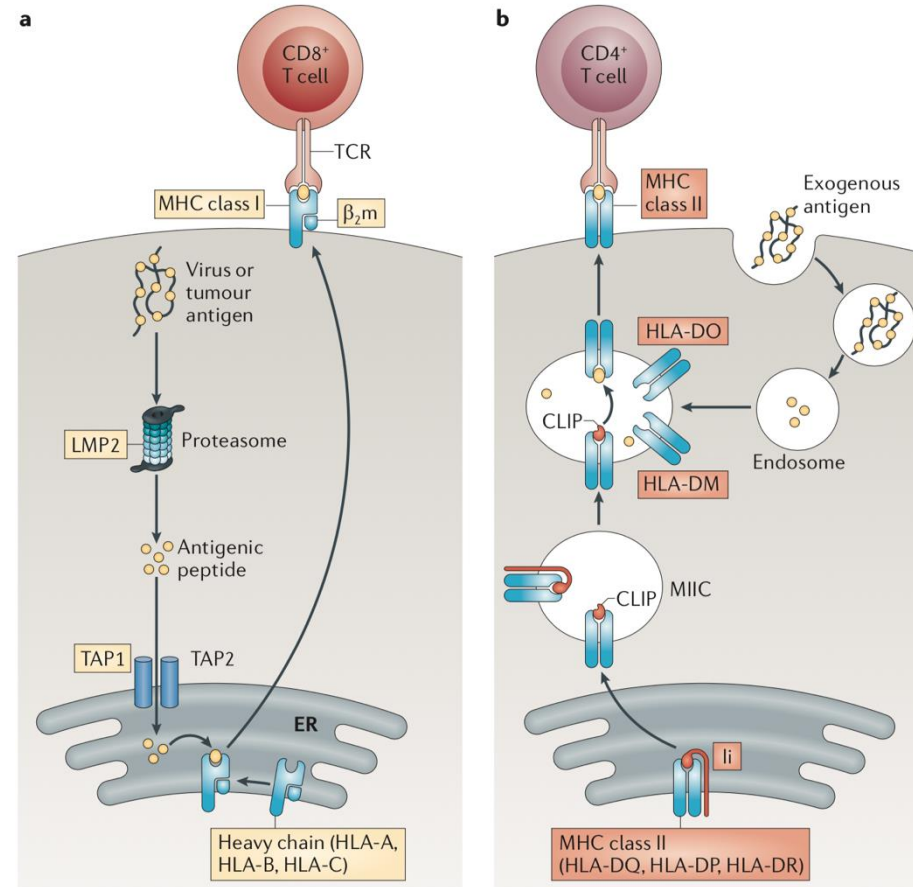
Mass spectrometry assay



MHC Typing or which MHC alleles are expressed by this particular cell

Barra C, et al. *Proteomics*. 2018 Jun;18(12):e1700252.

MS contain antigen processing and natural presentation information

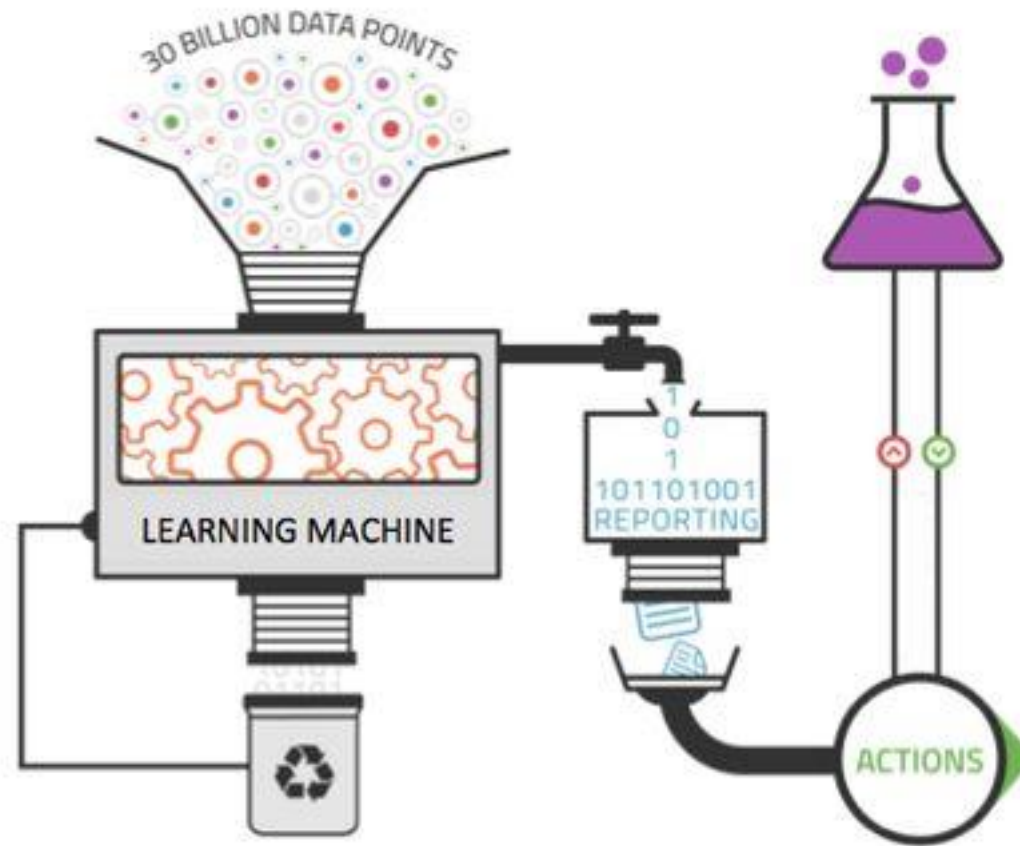


From MacMillan Publishers Ltd 2012. *Nat. Rev. Imm.* Vol 12

MS becomes high-throughput > immunopeptidomics

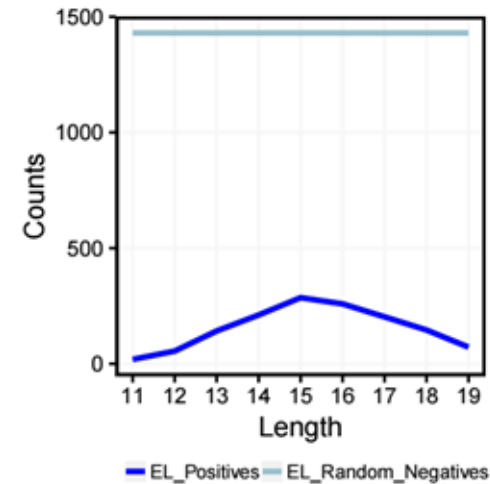
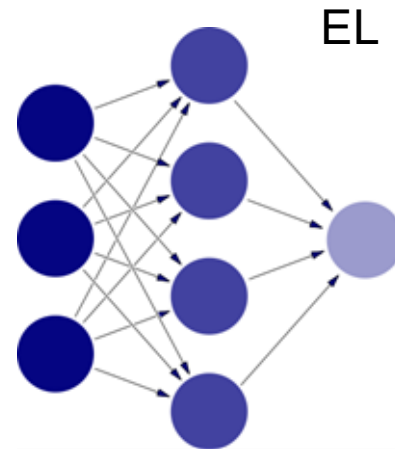


Game changers on immuno-informatics



Now the data follows a natural length distribution

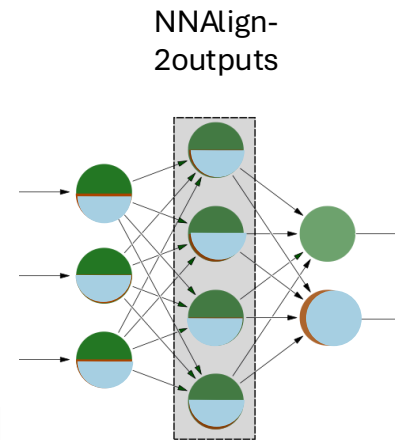
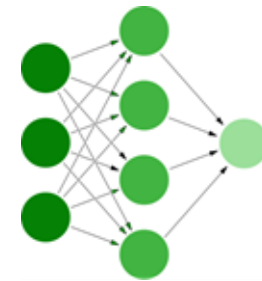
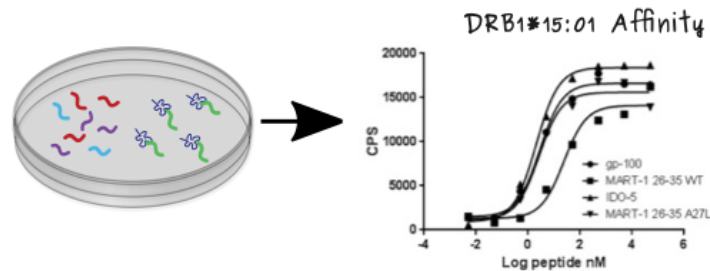
AADTVTQFDNVRLWLG	1
AADVVLIRNDLLD	1
AADVVLIRNDLLDVV	1
AAEAAASHTGTLTGS	0
AAEAIETEAKKRGWWVK	0
DTFENMGPATKKY	0
FENMWEFFPATKKY	0
AADLPQLVGHVPGAVL	0
AADPGATNTDLVGD	0
AADPQFVTVAL	0
AADPQWVPTETDV	0
AADRLNTSNNTKVR	0
AADTFENMGPATKKY	0
AADTRLDDPRSFS	0



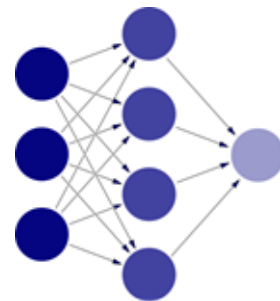
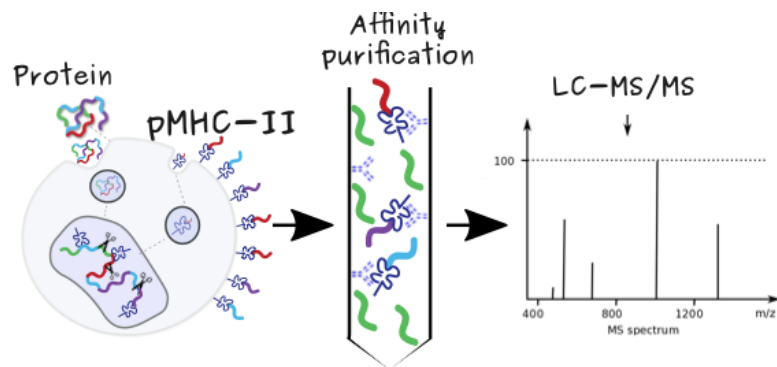
Barra C, et al. *Proteomics*. 2018 Jun;18(12):e1700252.

MS eluted ligands integration to MHC-II predictors

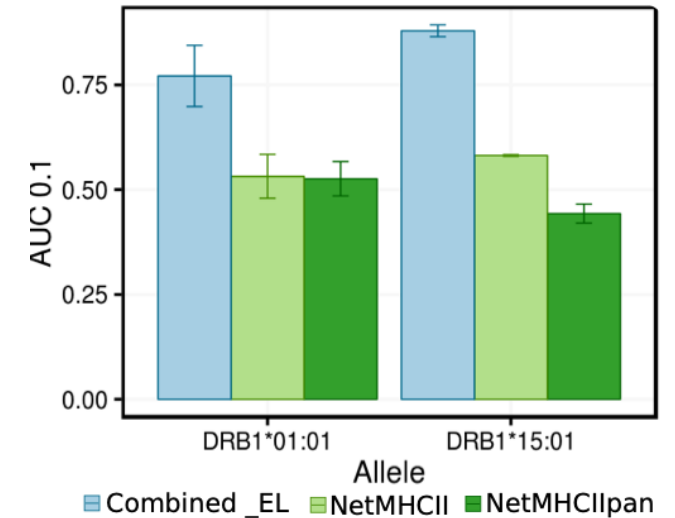
Peptides from binding affinity measurements



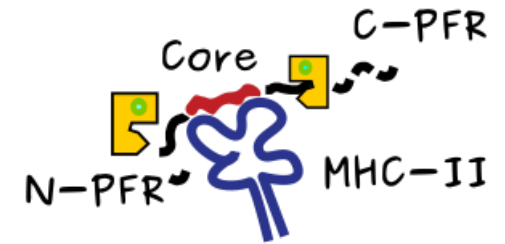
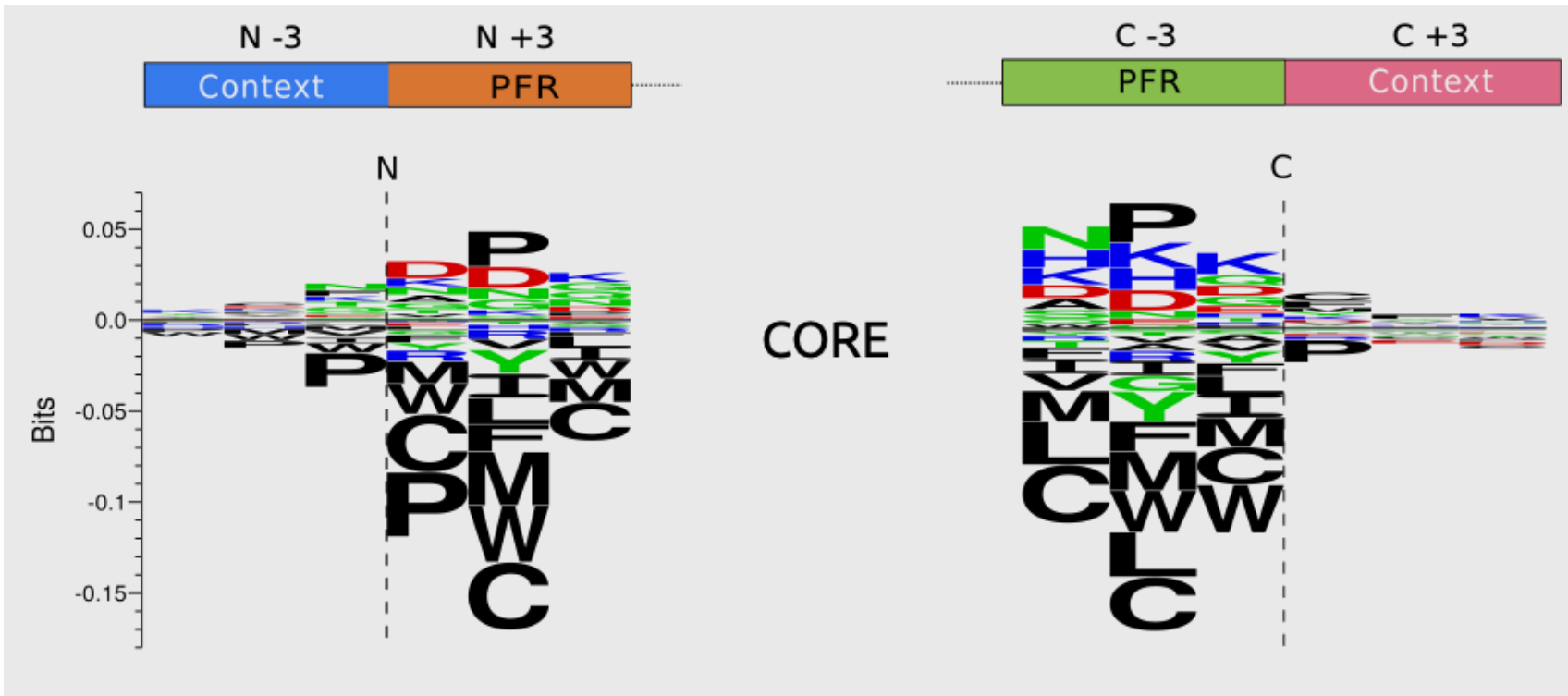
Peptides from mass spectrometry eluted ligands



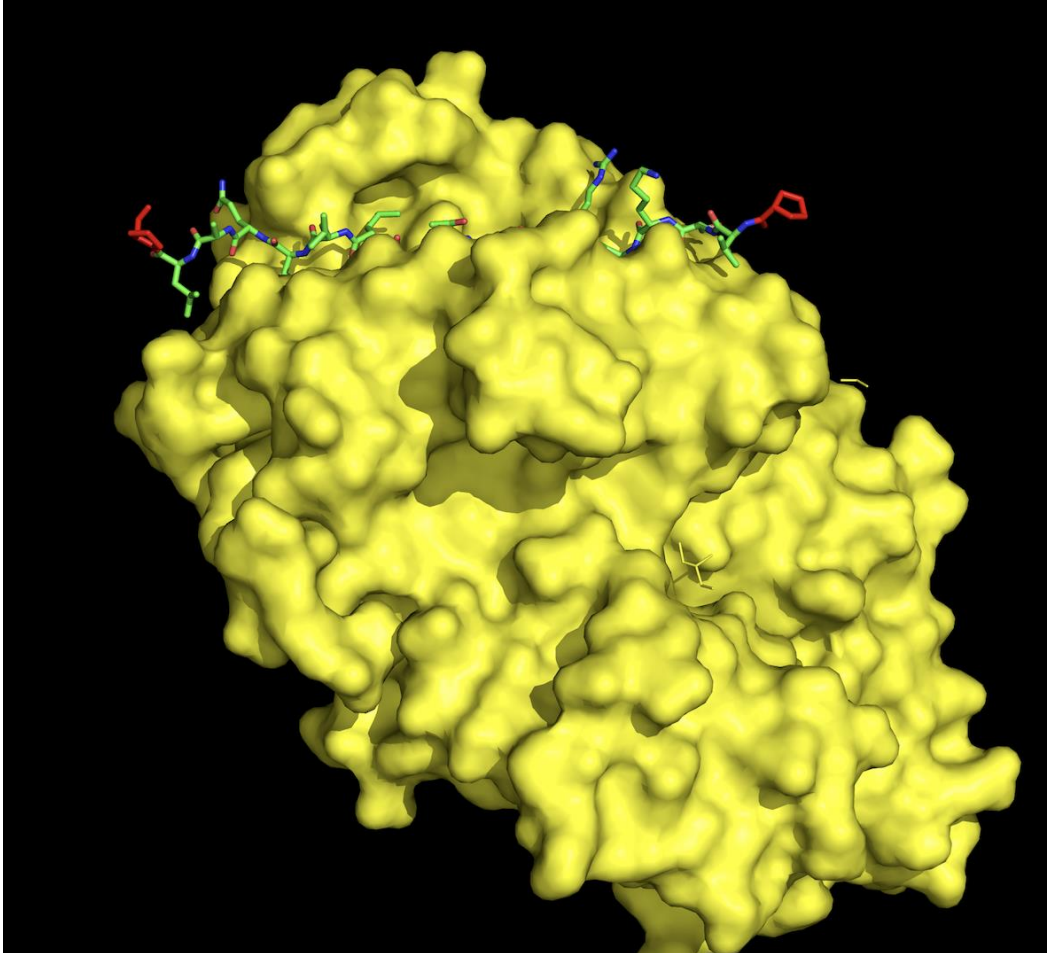
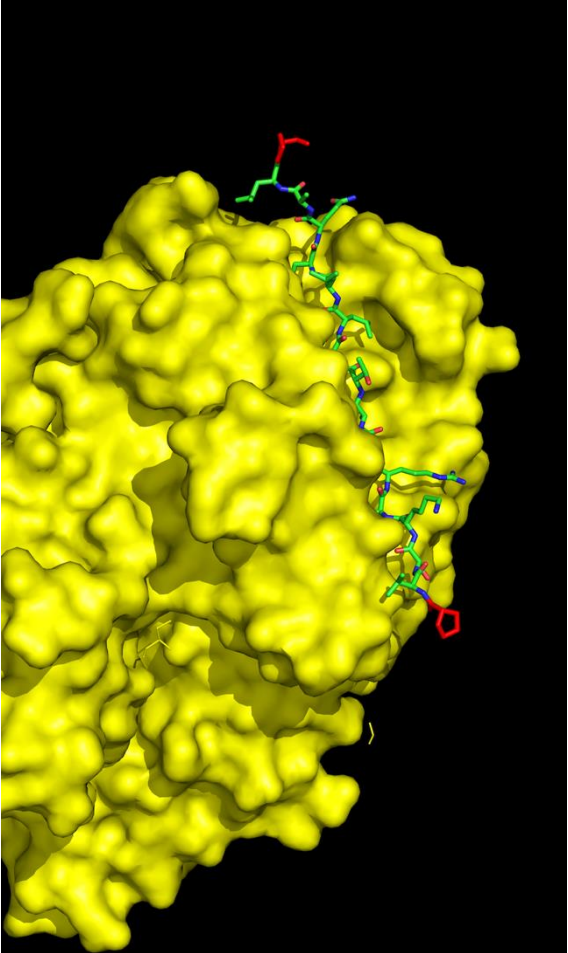
Eluted ligands
Evaluation dataset



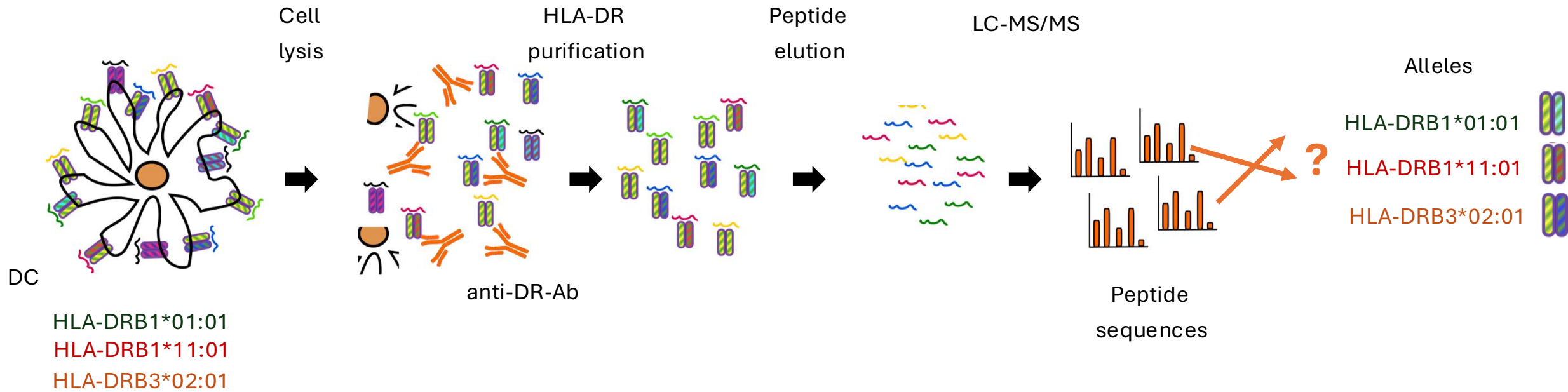
Footprints of antigen processing in MHC class II MS data



Footprints of antigen processing in MHC class II MS data



Challenges on immunopeptidomics



What is the IEDB?



- Database of experimentally-derived epitope assay results
- Maintained and curated by experts
- Data from literature and submissions
 - ~ 1 Million Epitopes
 - ~ 4 Million Assay Results
 - ~ 22K References
- Graphical User Interface
- Analysis Resource - Prediction Tools

IEDB Home Page

Intro+Stats

Home Page Search

Analysis Resource



Welcome

The Immune Epitope Database (IEDB) is a freely available resource funded by NIAID. It catalogs experimental data on antibody and T cell epitopes studied in humans, non-human primates, and other animal species in the context of infectious disease, allergy, autoimmunity and transplantation. The IEDB also hosts tools to assist in the prediction and analysis of epitopes.

[Learn More](#)

Upcoming Events

Antibody Society Booth	Dec 9-13
AAAAI 2020 Booth	Mar 13-16
AAI 2020 Booth	May 8-12
FOCIS 2020 Booth	June 23-26

Summary Metrics

Peptidic Epitopes	618,745
Non-Peptidic Epitopes	2,886
T Cell Assays	367,360
B Cell Assays	492,022
MHC Ligand Assays	1,299,413
Epitope Source Organisms	3,788
Restricting MHC Alleles	787
References	20,791

START YOUR SEARCH HERE

Epitope

Any Epitopes
 Linear Epitope
 Discontinuous Epitopes
 Non-peptidic Epitopes

Exact Match: Ex: SIINFEKL

Assay

Positive Assays Only
 T Cell Assays
 B Cell Assays
 MHC Ligand Assays

Ex: neutralization [Find](#)

Antigen

Organism
Ex: influenza, peanut

Antigen Name
Ex: core, capsid, myosin [Find](#)

MHC Restriction

Any MHC Restriction
 MHC Class I
 MHC Class II
 MHC Nonclassical

Ex: HLA-A*02:01 [Find](#)

Host

Any Host
 Humans
 Mice
 Non-human Primates

Ex: dog, camel [Find](#)

Disease

Any Disease
 Infectious Disease
 Allergic Disease
 Autoimmune Disease

Ex: asthma, diabetes [Find](#)

[Reset](#) [Search](#)

Epitope Analysis Resource

T Cell Epitope Prediction

Scan an antigen sequence for amino acid patterns indicative of:

- MHC I Binding
- MHC II Binding
- MHC I Processing (Proteasome, TAP)
- MHC I Immunogenicity

B Cell Epitope Prediction

Predict linear B cell epitopes using:

- Antigen Sequence Properties

Predict discontinuous B cell epitopes using antigen structure via:

- Discotope
- ElliPro

Epitope Analysis Tools

Analyze epitope sets of:

- Population Coverage
- Conservation Across Antigens
- Clusters with Similar Sequences

Home Page Search

- Epitope Sequence
- Antigen
- Host
- Assay
- MHC
- Disease

Positive epitopes by default

START YOUR SEARCH HERE ?

Epitope ?



Any Epitopes
 Linear Epitope
Exact Ma Ex: SIINFEKL
 Discontinuous Epitopes
 Non-peptidic Epitopes

Assay ?



Positive Assays Only
 T Cell Assays
 B Cell Assays
 MHC Ligand Assays
Ex: neutralization

Antigen ?




Organism
Ex: influenza, peanut
Antigen Name
Ex: core, capsid, myosin

MHC Restriction ?



Any MHC Restriction
 MHC Class I
 MHC Class II
 MHC Nonclassical
Ex: HLA-A*02:01

Host ?




Any Host
 Humans
 Mice
 Non-human Primates
Ex: dog, camel

Disease ?



Any Disease
 Infectious Disease
 Allergic Disease
 Autoimmune Disease
Ex: asthma, diabetes

Results Summary



IMMUNE EPITOPE DATABASE
AND ANALYSIS RESOURCE

[Home](#) | [Specialized Searches](#) | [Analysis Resource](#)

Query Results

Tab Separated Summary

Each row is a record/epitope

Pending Filters

Reset Search

Epitope

Any Epitopes

Linear Epitope

Discontinuous Epitopes

Non-peptidic Epitopes

3D structure available

Amino Acid Modification

Antigen

Organism

Ex: influenza, peanut

Antigen Name

Ex: core, capsid, myosin

Receptor

Has receptor sequence

Type: Any Type

Chain: Any Type

Sequence

Exact Matches

Assay

Positive Assays Only

Current Filters: ✖ Positive Assays Only

Epitopes (536986)
Antigens (42541)
Assays (1233456)
Receptors (24786)
References (19791)

Go To Records Starting At 1200 GO Export Results

536986 Records Found 25 Per Page

Details	Epitope	Antigen	Organism	# References	# Assays
123885	cardiolipin			320	1028
44920	NLVPMTATV	65 kDa phosphoprotein	Human herpesvirus 5 (Human cytomegalovirus)	281	693
20354	GILGFVFTL	Matrix protein 1	Influenza A virus	207	563
113645	MEVGWYRSPFSRVVHLYRNGK	Myelin-oligodendrocyte glycoprotein	Mus musculus (mouse)	187	975
58560	SIINFELK	Gal d 2	Gallus gallus (chicken)	171	444
4602	ASNENMETM	Nucleoprotein	Influenza A virus	146	402
112741	2,4-dinitrophenyl group			140	477
20788	GLCTLVAML	mRNA export factor ICP27 homolog	Human herpesvirus 4 (Epstein Barr virus)	128	271
130694	1-O-(alpha-D-galactosyl)-N-hexacosanoylphyto sphingosine			118	579
24786	HSLGKWLGHDPKF	Myelin proteolipid protein	Mus musculus (mouse)	115	697
48237	PKYVKQNTLKLAT	Hemagglutinin	Influenza A virus	109	388
130649	alpha-D-Galp-(1->3)-beta-D-Galp-(1->4)-D-GlcNAc-yl group	Envelope glycoprotein	Murine leukemia virus	108	425
6435	CINGVCWTV	Genome polyprotein	Hepatitis C virus	106	302
112742	2,4,6-trinitrophenyl group			106	309
32208	KLVALGINAV	Genome polyprotein	Hepatitis C virus	96	283
53112	RAHYNIVTF	Protein E7	Alphapapillomavirus 9	96	239
61086	SSIEFARL	Envelope glycoprotein B	Human herpesvirus 1	93	305
61151	SSLENFRAYV	Polymerase acidic protein	Influenza A virus	88	264
16833	FLPSDFFPV	Capsid protein	Hepatitis B virus	86	246
30001	KAVYNFATC	Pre-glycoprotein polyprotein GP complex	Lymphocytic choriomeningitis mammarenavirus	85	250
65748	TPRVTGGGAM	65 kDa phosphoprotein	Human herpesvirus 5 (Human cytomegalovirus)	85	173
6568	CLGLLTMV	Latent membrane protein 2	Human herpesvirus 4 (Epstein Barr virus)	84	216
7493	DAEFRHDSGYEVHHQKLVFFAEDVGSNKGAIIGLMVGGVVIA	Amyloid beta A4 protein	Homo sapiens (human)	81	254
16878	FLPDRYGL	Epstein Barr nuclear antigen 2	Human herpesvirus 4 (Epstein Barr virus)	81	249
			Lymphocytic choriomeningitis mammarenavirus	80	193

of 21480 25 Per Page


ing At 1200 GO Export Results

Search Refinement

Add filters to your initial results

Even more search fields than in home page

Results Summary - Refined - Immunome Browser



IMMUNE EPITOPE DATABASE
AND ANALYSIS RESOURCE

Help | More IEDB

Home | Specialized Searches | Analysis Resource

✕
Pending Filters
✕ Positive Assays Only
✕ Organism: Influenza A virus (ID:11320, influenza A)
✕ No B cell assays

✕ No MHC ligand assays
✕ MHC Restriction Type: Class II
✕ Host: Homo sapiens (human)

Pending Filters

Reset

Search

Epitope ?

Any Epitopes

Linear Epitope

Discontinuous Epitopes

Non-peptidic Epitopes

3D structure available

Amino Acid Modification



Antigen ?

Organism

Influenza A virus (ID:11320) 1

Antigen Name

Ex: core, capsid, myosin

Epitopes
(810)

Antigens
(11)

Assays
(1465)

Receptors
(197)

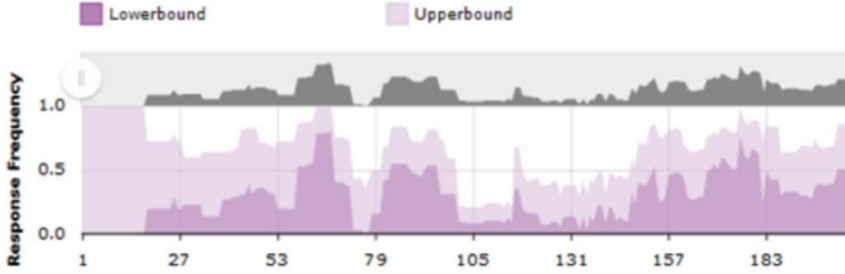
References
(113)

Go To Records Starting At GO Export Results ↗

11 Records Found ⏪ ⏩ Page 1 of 1 ⏪ ⏩ 25 Per Page

Antigen	Organism	# Epitopes	# Assays	# References
Hemagglutinin	Influenza A virus (ID:11320)	1	1	1
Matrix protein 1	Influenza A virus (ID:11320)	1	1	1
Nucleoprotein	Influenza A virus (ID:11320)	1	1	1
RNA-directed RNA polymerase catalytic subunit	Influenza A virus (ID:11320)	1	1	1
Neuraminidase	Influenza A virus (ID:11320)	1	1	1
Non-structural protein 1	Influenza A virus (ID:11320)	1	1	1
Polymerase acidic protein	Influenza A virus (ID:11320)	1	1	1
Polymerase basic protein 2	Influenza A virus (ID:11320)	1	1	1
Nuclear export protein	Influenza A virus (ID:11320)	1	1	1

Response Frequency



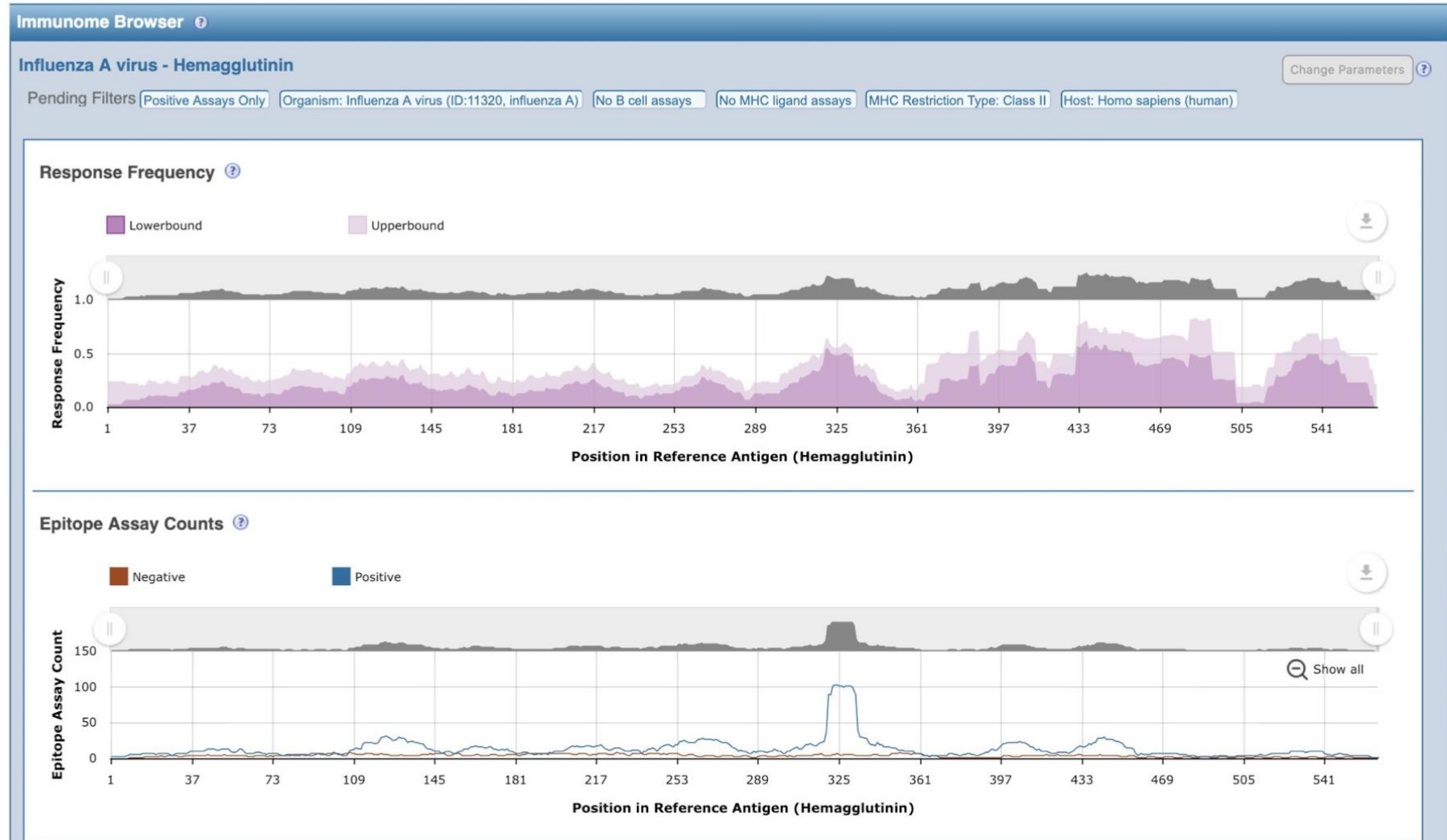
✕

Click icon to view Immunome Browser

Influenza A Hemagglutinin
Host: Homo sapiens
Assay: B cell assays

Immunome Browser

- Visual Representation of epitopes, mapped onto antigen sequence



Now, explore it yourself!

Links

[Detailed Course Program](#)

