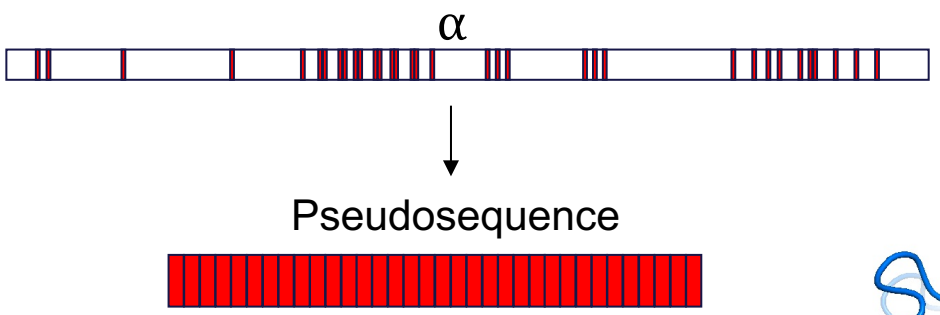
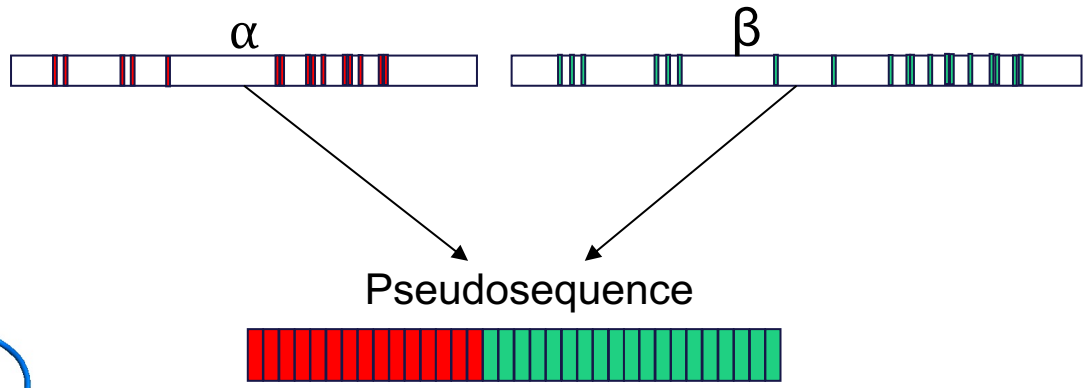
Stuff from my Master’s thesis

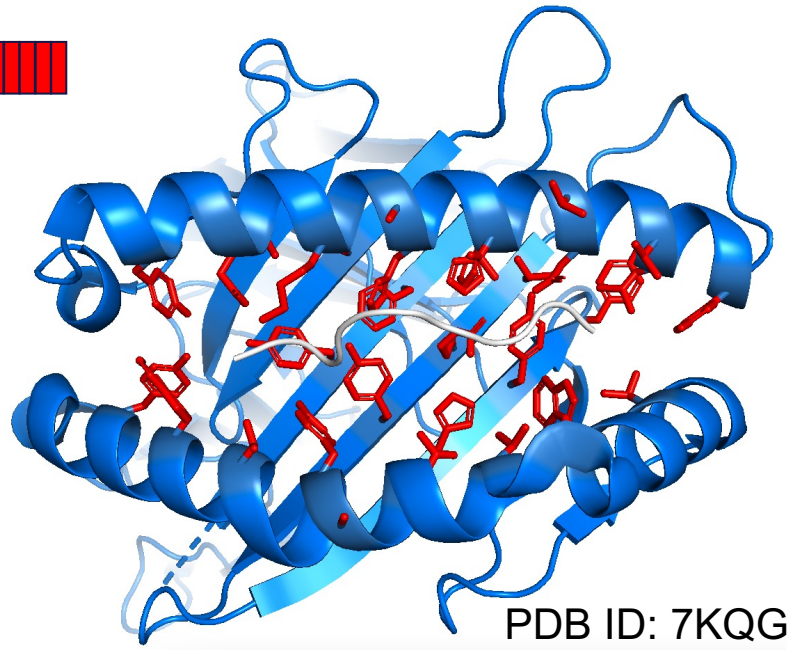
MHC sequences

Class I: Alpha chain



Class II: alpha + beta chain





NNAlign-2.1

Update network weights

Predict binding cores and affinities

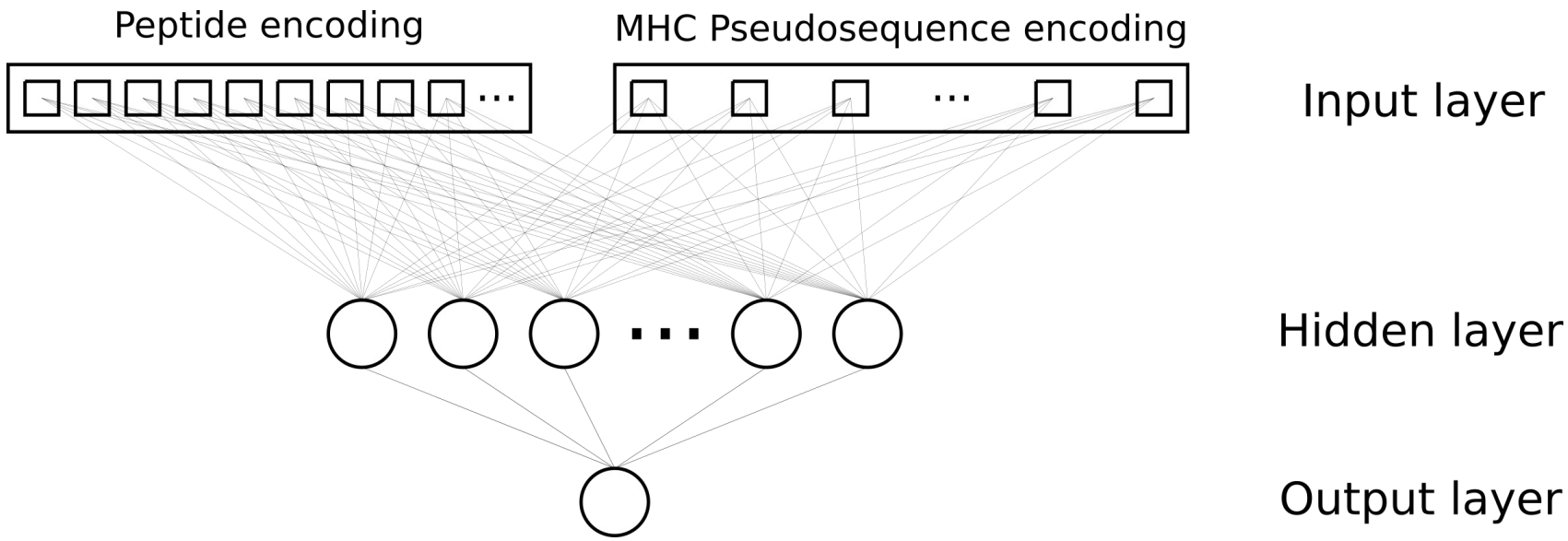
Peptide	Measurement	Prediction	Error
IPQSLWTSL (DSY)	0.679	0.619	0.060
-ALPPVAPV	0.503	0.507	0.002
FSLPFPFLYKFLL	0.590	0.482	0.108
EVVMAYVGIK	0.577	0.448	0.129
QLLGKLPHLRM	0.010	0.020	0.010
-RGYVFQGL	0.198	0.206	0.008
RLRDLLLIVTR	0.059	0.023	0.036
⋮			

Peptide	Measurement
ALPPVAPV	0.503
Binding core	**Prediction**
-ALPPVAPV	0.362
A-LPPVAPV	0.473
AL-PPVAPV	0.507
ALP-PVAPV	0.416
⋮	⋮

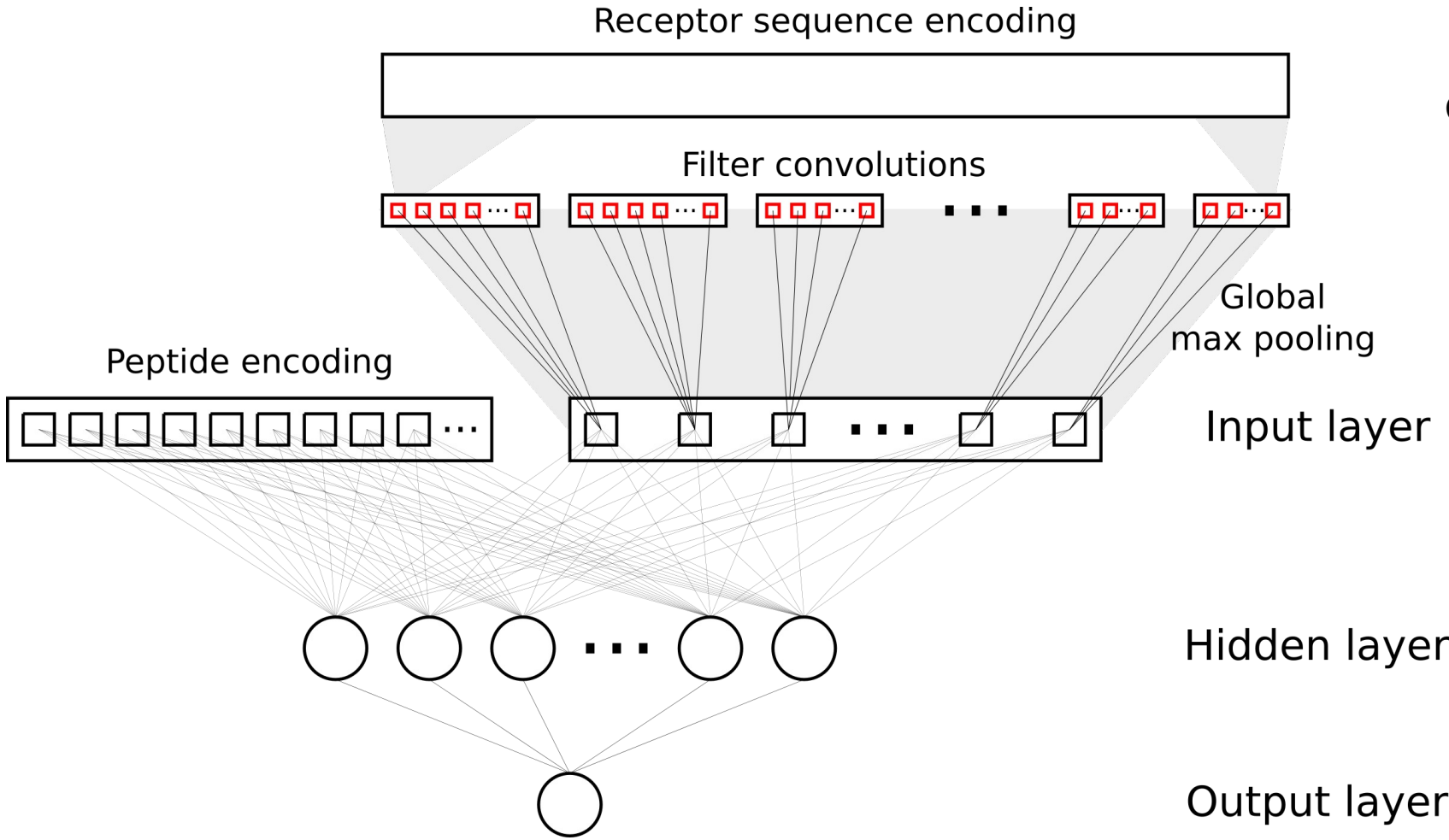
Calculate prediction errors

Peptide	Measurement
IPQSLDSYWTSL	0.679
Alignment	**Binding core**	**Prediction**
IPQSLDSYWTSL	IPQSLDSYW	0.543
IPQSLDSYWTSL	PQSLDSYWT	0.327
IPQSLDSYWTSL	QSLDSYWTS	0.430
⋮	⋮	⋮
IPQSLDSYWTSL	IQSLDSYWT	0.523
IPQSLDSYWTSL	IPSLDSYWT	0.495
IPQSLDSYWTSL	IPQLDSYWT	0.347
⋮	⋮	⋮
IPQSLDSYWTSL	IPQSYWTSL	0.567
IPQSLDSYWTSL	IPQSLWTSL	0.619
IPQSLDSYWTSL	IPQSLDTSL	0.583
⋮	⋮	⋮

The conventional NNAlign method

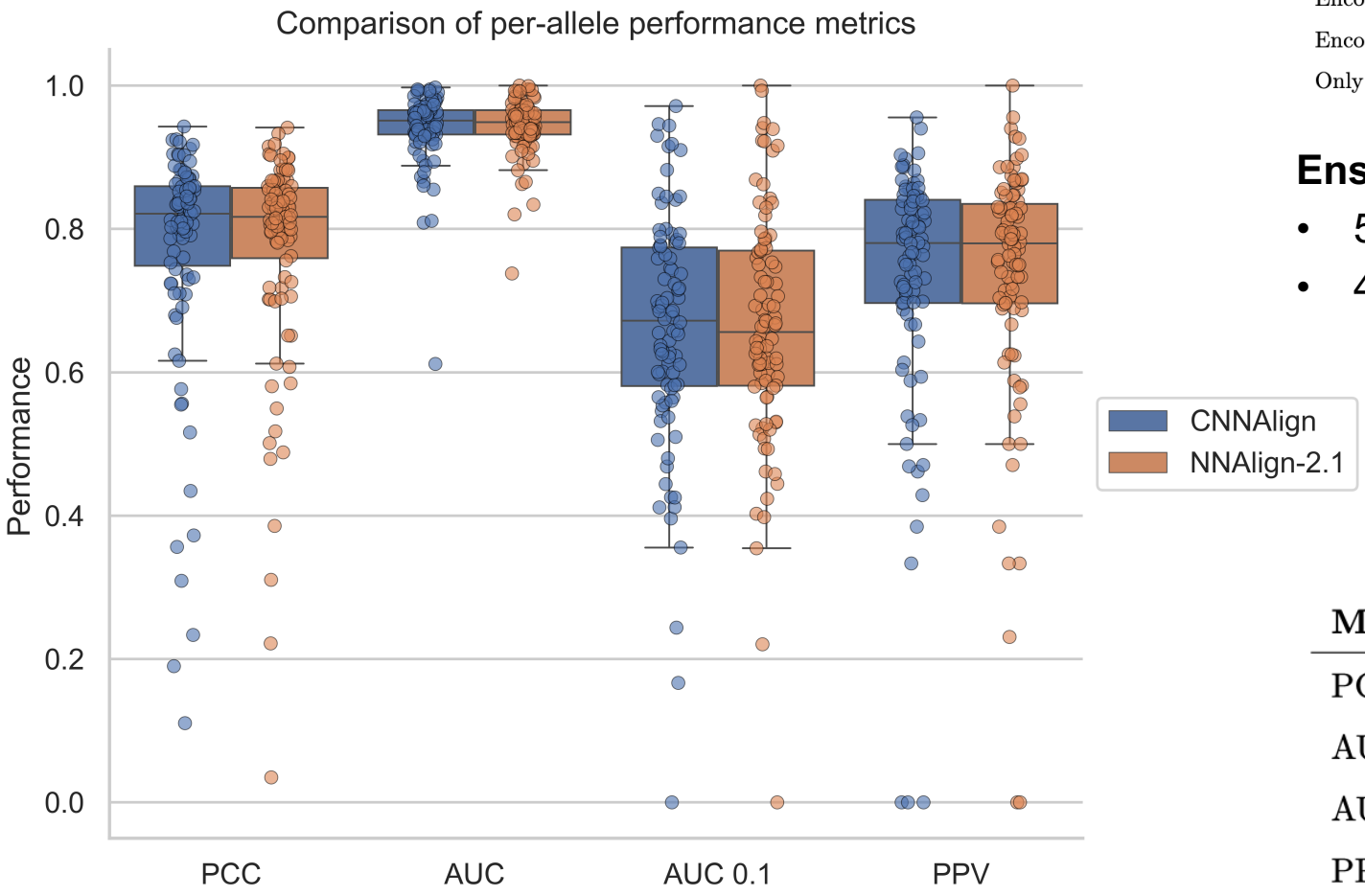


CNNAlign



Code implemented using the C language

Results: MHC class I



Parameter	Value	Number of input values
Peptide encoding	Blosum	-
Motif (binding core) length	9	180 (9×20)
PFR length	0	-
PFR length encoding	False	-
Max length of insertions	1	-
Max length of deletions	10	-
Individual peptide lengths encoded	8,9,10,11,12,13,14	7
Encode insertion length	True	2
Encode deletion length	True	2
Only make insertions in peptides shorter than the motif length	True	-

Ensemble of 20 networks

- 5-fold CV
- 4 initializations per fold

CNN filters used (180 in total)

- 50 of length 10
- 50 of length 8
- 40 of length 5
- 40 of length 3

Metric	#CNNAlign wins	#Ties	P-value (without ties)
PCC	71	0	**0.00001**
AUC	60	2	**0.00103**
AUC 0.1	62	0	**0.00056**
PPV	38	21	0.31765

Positional importance

Rescaled filter weights (5 x 20)

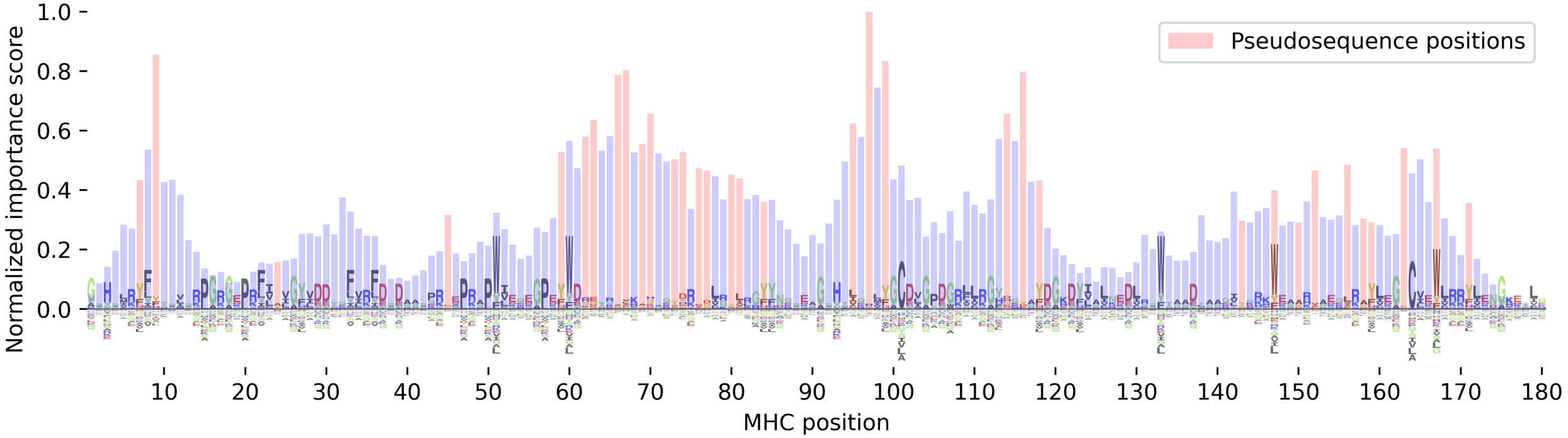
	i=1	i=2	i=3	i=4	i=5
A	-0.55	0.51	0.23	-0.19	0.39
R	-0.10	0.02	2.11	-0.66	0.23
N	-0.59	-0.30	0.67	-0.91	1.02
D	-0.50	-0.14	0.67	-0.82	1.24
C	-0.58	0.70	0.34	0.35	0.45
Q	-0.49	-0.06	0.83	-0.77	0.46
E	-0.33	0.01	0.83	-0.73	0.92
G	-0.65	-0.01	0.71	-0.49	0.46
H	0.11	-0.39	0.77	-0.10	0.38
I	0.67	2.11	0.30	0.95	-0.05
L	0.81	1.78	0.59	1.18	-0.28
K	-0.41	-0.19	1.58	-0.97	0.53
M	0.44	1.48	0.33	0.80	-0.20
F	1.62	1.02	0.44	1.90	-0.57
P	-0.29	0.21	0.50	-0.25	0.78
S	-0.68	-0.08	0.50	-0.52	0.55
T	-0.38	0.58	0.33	-0.10	0.22
W	1.56	0.51	1.05	1.71	-0.12
Y	1.25	0.42	0.30	1.29	-0.47
V	0.41	1.97	0.27	0.74	0.02

$\sigma_i = \mathbf{std.dev}(s_{i,A}, s_{i,R}, \ldots, s_{i,V})$

σ_1	σ_2	σ_3	σ_4	σ_5
0.74	0.76	0.46	0.90	0.48

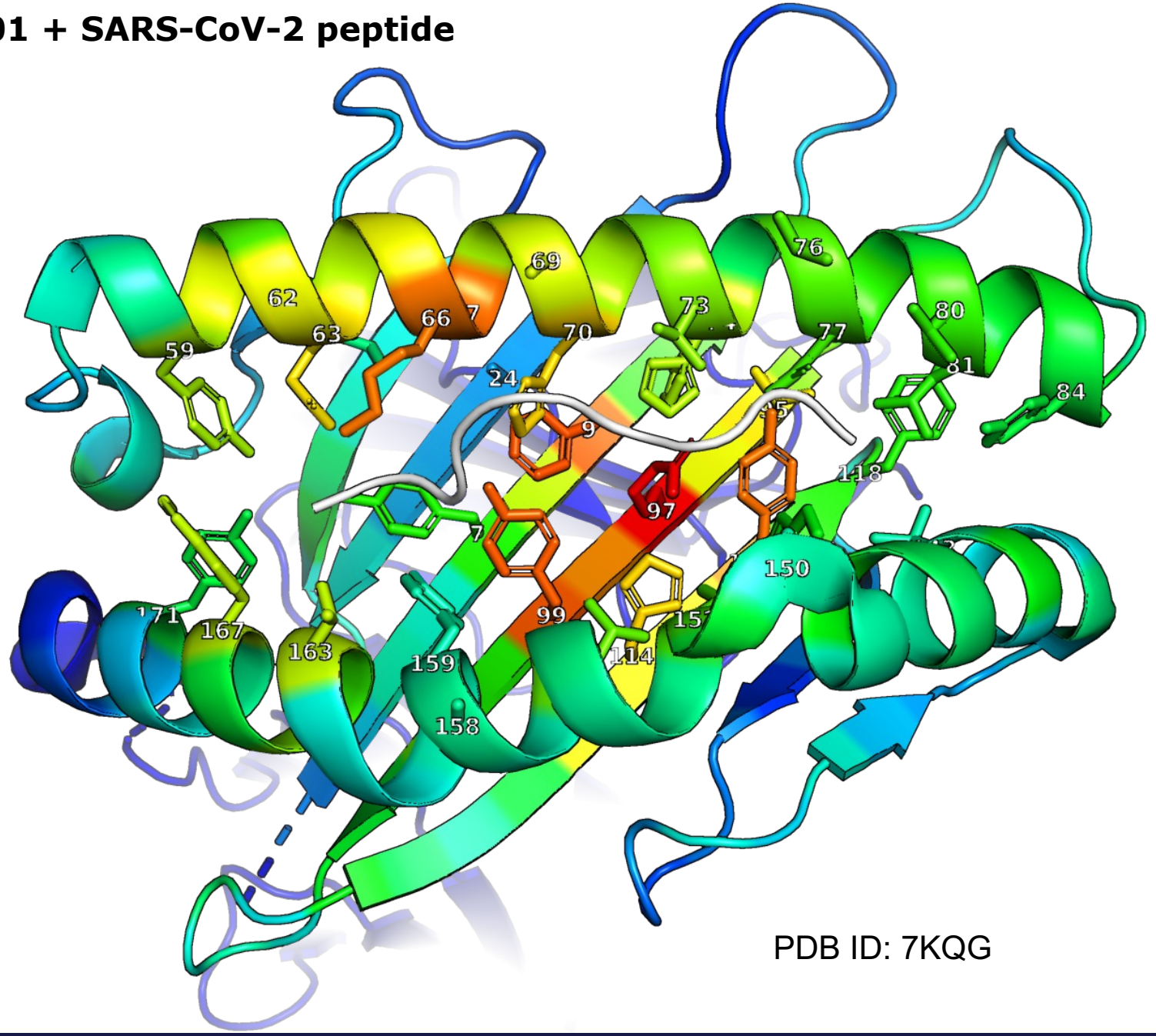
- Weight each position in the MHC based on how many filters had their max score around that position
- The weight depends on how much each filter has ‘learned’

	33	34	35	36	37	
...	0.00	0.00	0.00	0.00	0.00	...
			+			
	0.19	0.19	0.12	0.23	0.12	
			=			
	33	34	35	36	37	
...	0.19	0.19	0.12	0.23	0.12	...


Normalized importance score
1.0
0.8
0.6
0.4
0.2
0.0
Pseudosequence positions
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
170
180
MHC position

HLA-A*02:01 + SARS-CoV-2 peptide

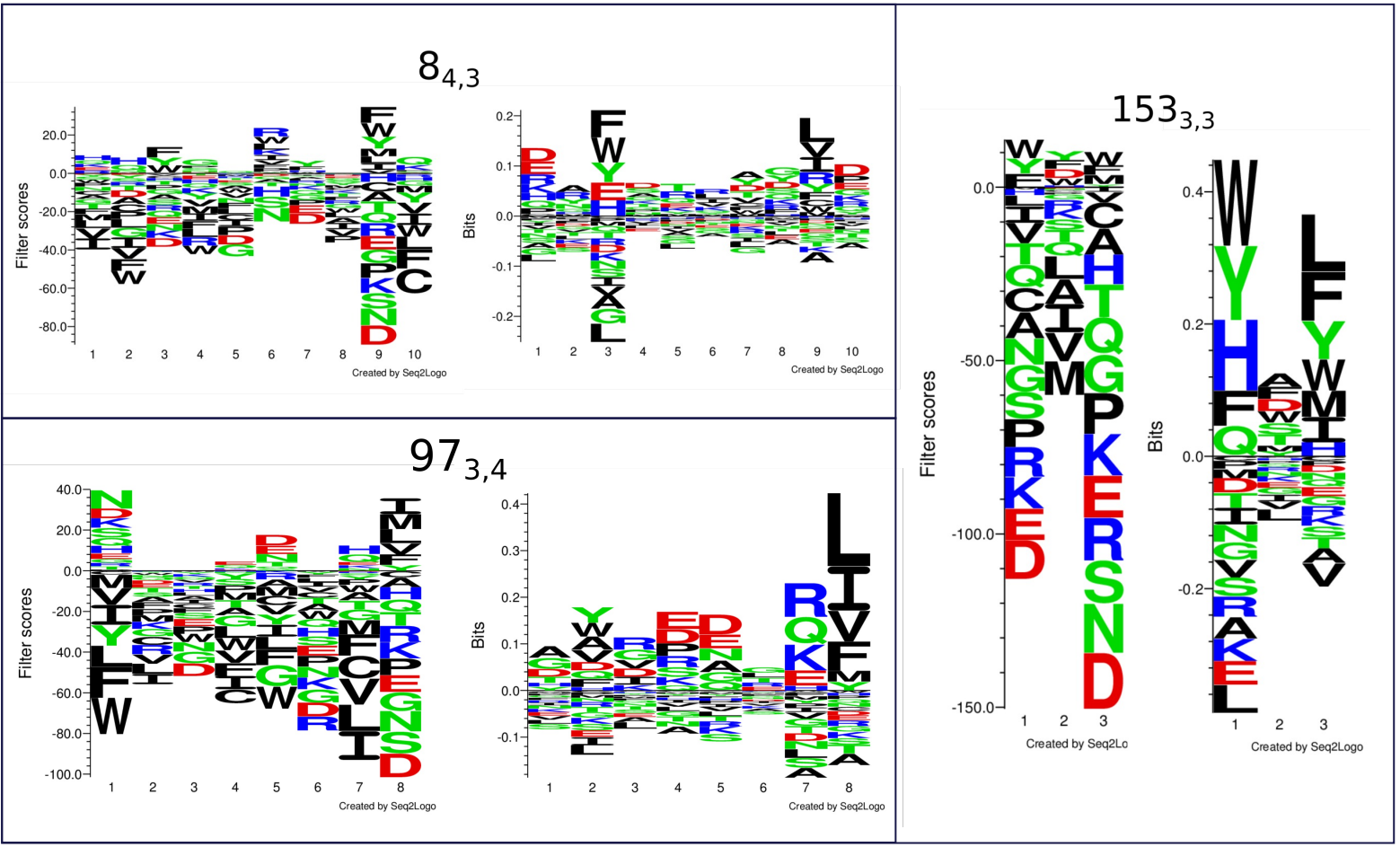
Sticks:
pseudosequence
positions



PDB ID: 7KQG

Filter sequence logos

Filters with high ‘knowledge’



Filters with low ‘knowledge’