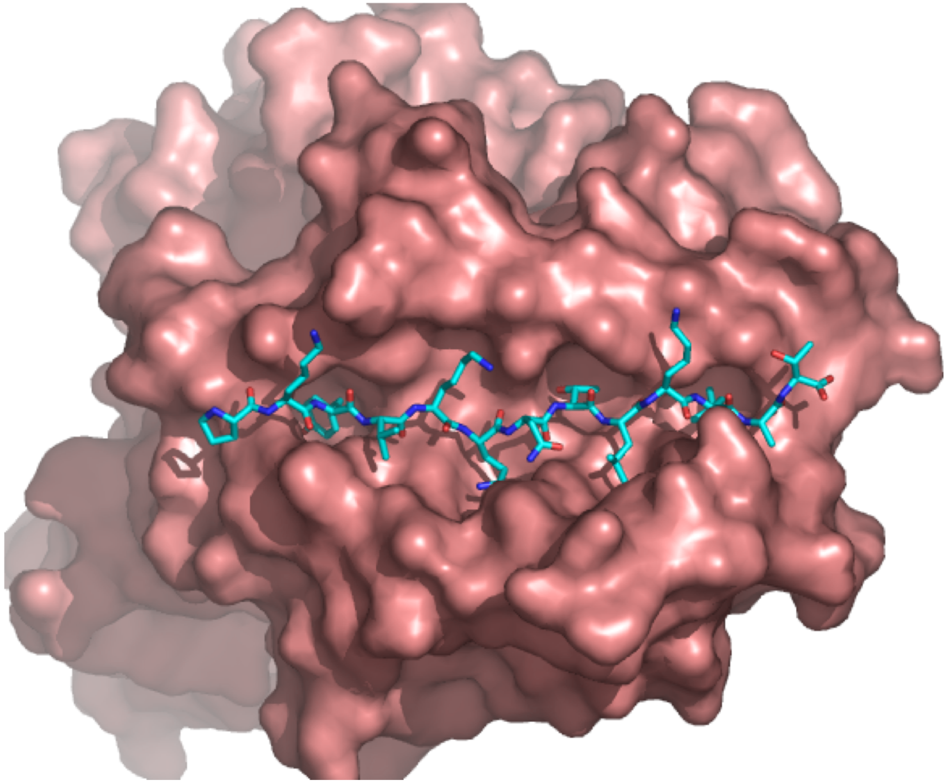
MHC class II binding predictions
NN-align
“Alignment using ANNs“

Class II MHC binding

- Binds peptides of length 9-18 (even whole proteins can bind!)
- Binding cleft is open
- Binding core is 9 aa
- Binding motif highly generate
- Amino acids flanking the binding core affect binding
- Peptide structure might determine binding



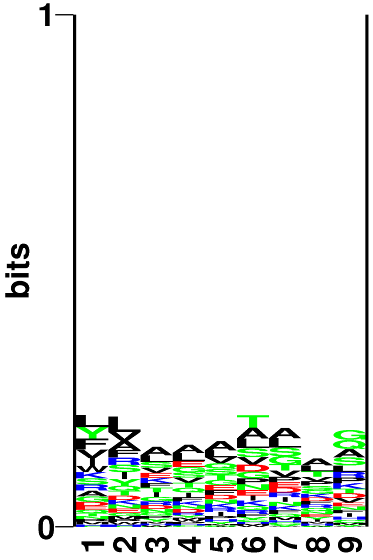
Gibbs sampler

www.cbs.dtu.dk/biotools/EasyGibbs

```
RFFGGDRGAPKRG
YLDPLIRGLLARPAKLQV
KPGQPPRLLIYDASNRATGIPA
GSLFVYNITTNKYKAFLDKQ
SALLSSDITASVNCAK
PKYVHQNTLKLAT
GFKGEQGPKGEP
DVFKELKVHHANENI
SRYWAIRTRSGGI
TYSTNEIDLQLSQEDGQTIE
```

100 10mer peptides

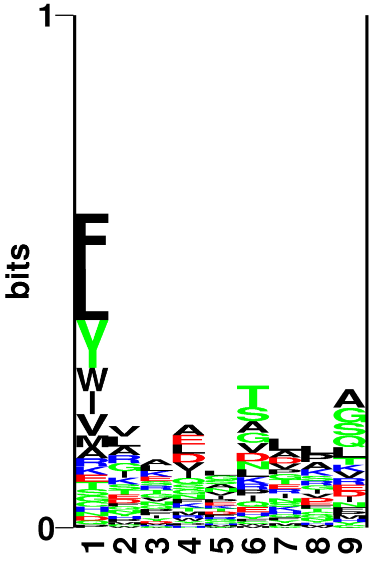
2^{100}~10^{30} combinations



```
      RFFGGDRGAPKRG
   YLDPLIRGLLARPAKLQV
KPGQPPRLLIYDASNRATGIPA
     GSLFVYNITTNKYKAFLDKQ
      SALLSSDITASVNCAK
      PKYVHQNTLKLAT
       GFKGEQGPKGEP
      DVFKELKVHHANENI
      SRYWAIRTRSGGI
       TYSTNEIDLQLSQEDGQTI
```

$$E = \sum_{p,aa} C_{pa} \log \frac{p_{pa}}{q_a}$$

Monte Carlo simulations can do it



NN-align

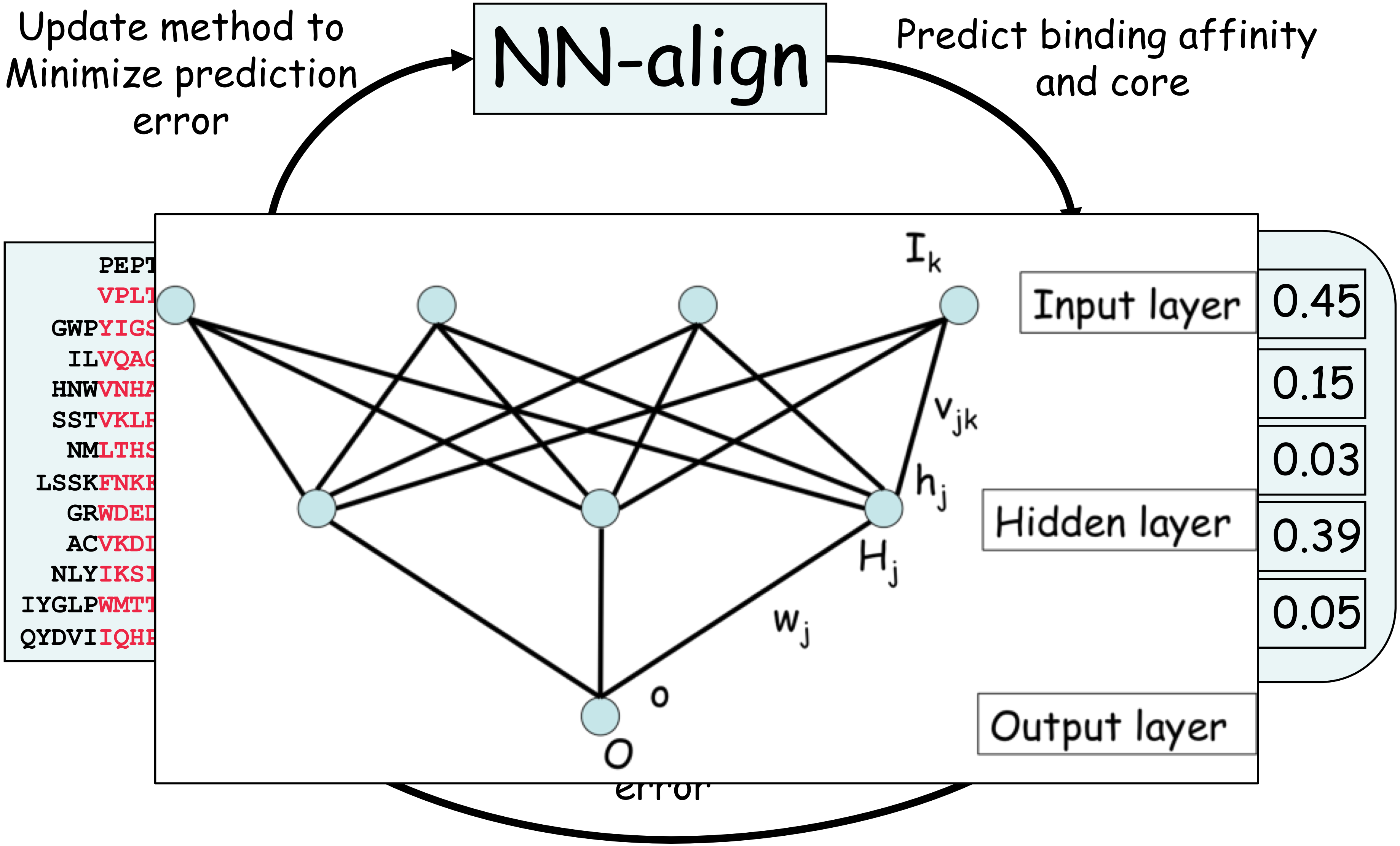
The problem. Where is the binding core?

PEPTIDE	IC50(nM)
VPLTDLRIPS	48000
GWPYIGSRSQIIGRS	45000
ILVQAGEAETMTPSG	34000
HNWVNHAVPLAMKLI	120
SSTVKLRQNEFGPAR	8045
NMLTHSINSLISDNL	47560
LSSKFNKFVSPKSVS	4
GRWDEDGAKRIPVDV	49350
ACVKDLVSKYLADNE	86
NLYIKSIQSLISDTQ	67
IYGLPWMTTQTSALS	11
QYDVIIQHPADMSWC	15245

Effect of Peptide Flanking Residues

- PFR's can affect binding dramatically
  - **RFYKTLRAEQASQ 34 nM**
  - **YKTLRAEQA >10000 nM**

Alignment using ANN



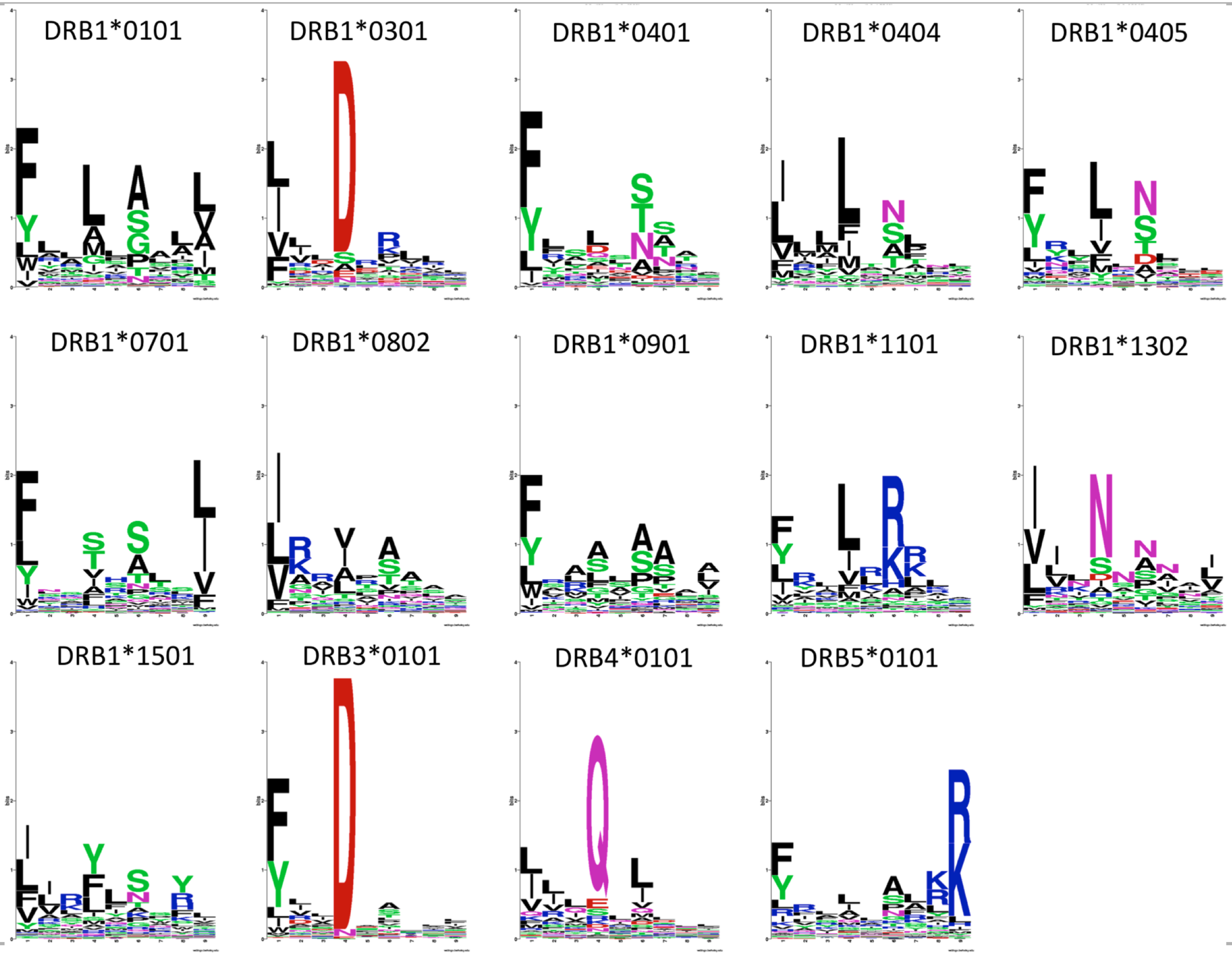
NN-align

Update method to Minimize prediction error

Predict binding affinity and core

Nielsen et al. BMC Bioinformatics 2009, 10:296

Binding motif of 14 HLA-DR molecules



DRB1*0101

DRB1*0301

DRB1*0401

DRB1*0404

DRB1*0405

DRB1*0701

DRB1*0802

DRB1*0901

DRB1*1101

DRB1*1302

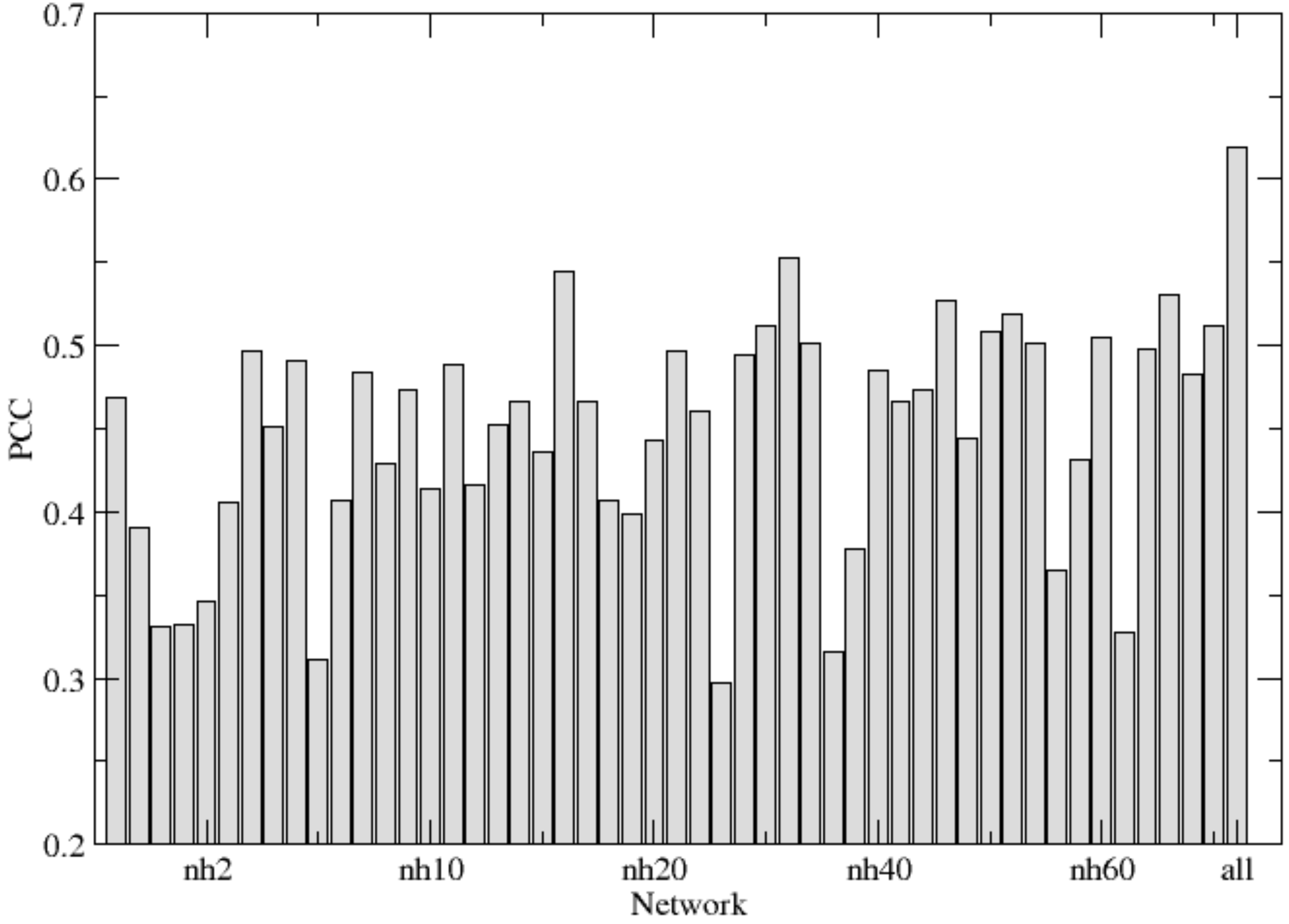
DRB1*1501

DRB3*0101

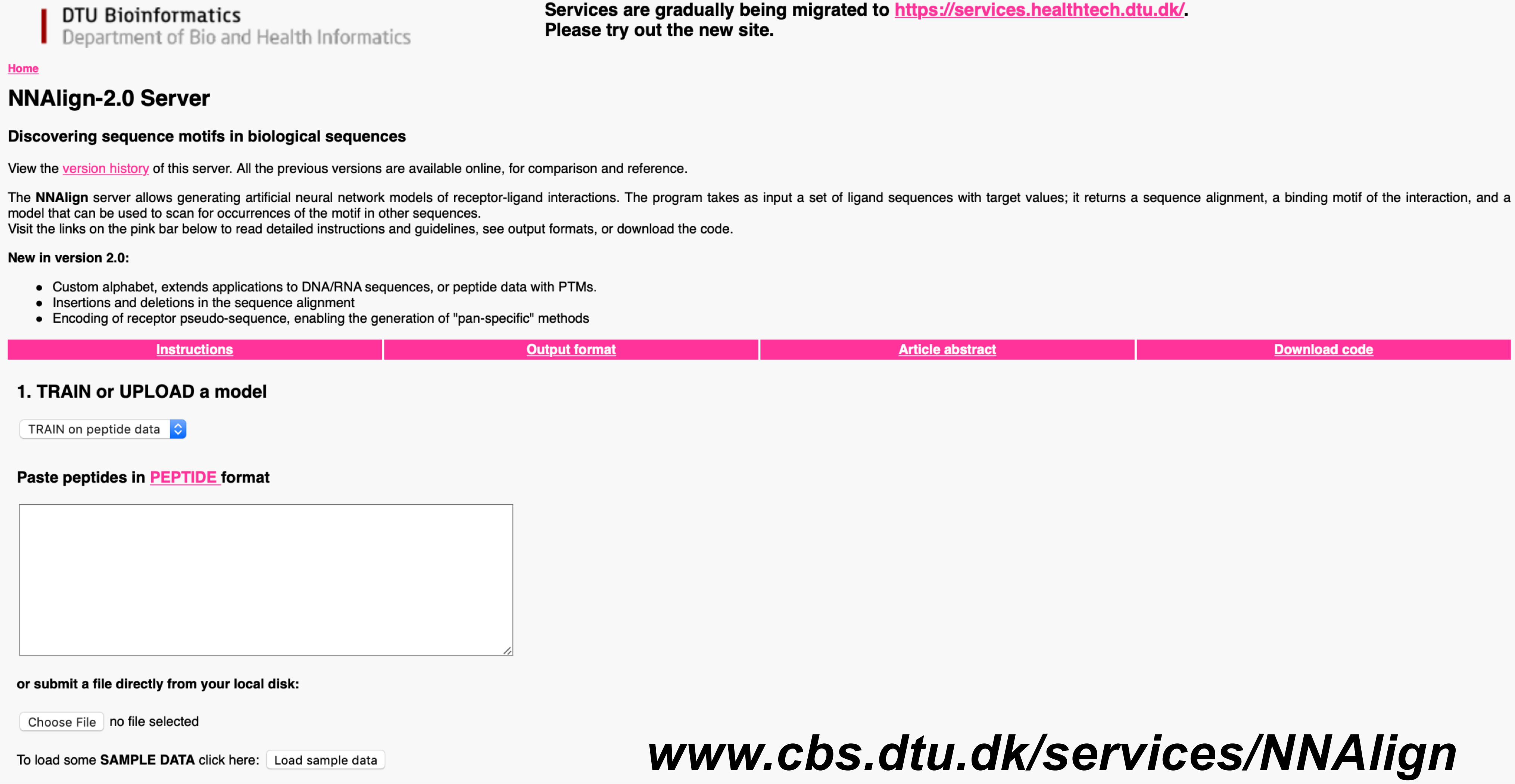
DRB4*0101

DRB5*0101

Network ensembles



NNAlign Server



DTU Bioinformatics
Department of Bio and Health Informatics

Services are gradually being migrated to https://services.healthtech.dtu.dk/.
Please try out the new site.

Home

NNAlign-2.0 Server

Discovering sequence motifs in biological sequences

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

The **NNAlign** server allows generating artificial neural network models of receptor-ligand interactions. The program takes as input a set of ligand sequences with target values; it returns a sequence alignment, a binding motif of the interaction, and a model that can be used to scan for occurrences of the motif in other sequences.
Visit the links on the pink bar below to read detailed instructions and guidelines, see output formats, or download the code.

New in version 2.0:

- Custom alphabet, extends applications to DNA/RNA sequences, or peptide data with PTMs.
- Insertions and deletions in the sequence alignment
- Encoding of receptor pseudo-sequence, enabling the generation of "pan-specific" methods

1. TRAIN or UPLOAD a model

TRAIN on peptide data

Paste peptides in PEPTIDE format

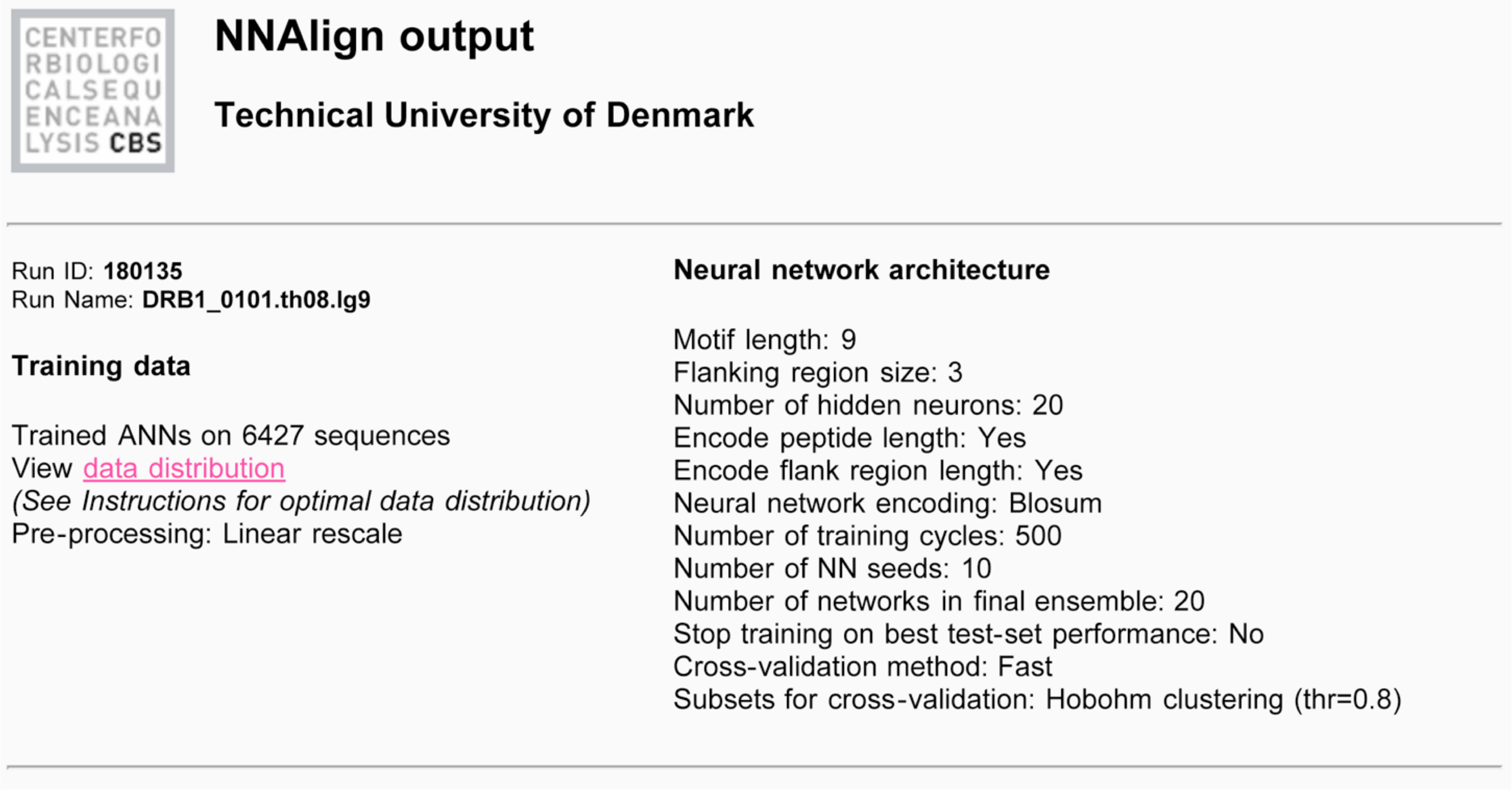
or submit a file directly from your local disk:

no file selected

To load some **SAMPLE DATA** click here:

www.cbs.dtu.dk/services/NNAlign

NNAlign Server - Output (1)



NNAlign output

Technical University of Denmark

Run ID: **180135**
Run Name: **DRB1_0101.th08.lg9**

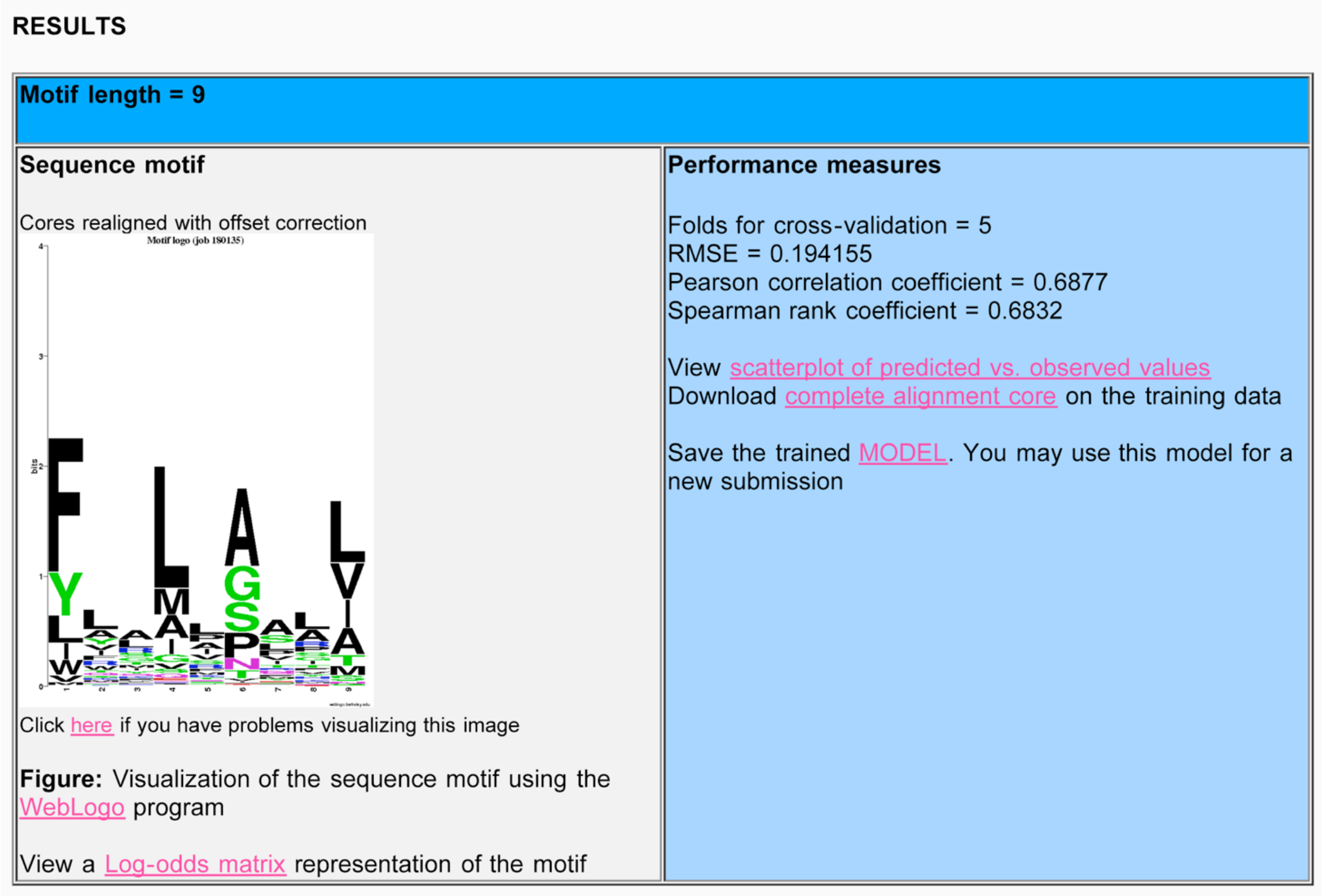
Training data

Trained ANNs on 6427 sequences
View data distribution
(See Instructions for optimal data distribution)
Pre-processing: Linear rescale

Neural network architecture

Motif length: 9
Flanking region size: 3
Number of hidden neurons: 20
Encode peptide length: Yes
Encode flank region length: Yes
Neural network encoding: Blosum
Number of training cycles: 500
Number of NN seeds: 10
Number of networks in final ensemble: 20
Stop training on best test-set performance: No
Cross-validation method: Fast
Subsets for cross-validation: Hobohm clustering (thr=0.8)

NNAlign Server - Output (2)



RESULTS

Motif length = 9

Sequence motif

Cores realigned with offset correction

Click here if you have problems visualizing this image

Figure: Visualization of the sequence motif using the WebLogo program

View a Log-odds matrix representation of the motif

Performance measures

Folds for cross-validation = 5
RMSE = 0.194155
Pearson correlation coefficient = 0.6877
Spearman rank coefficient = 0.6832

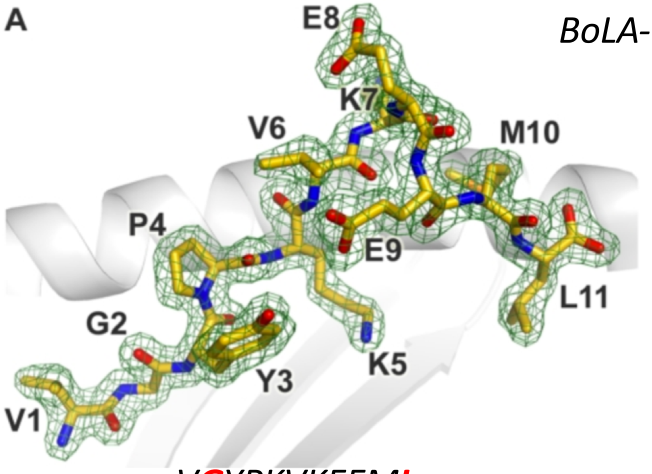
View scatterplot of predicted vs. observed values
Download complete alignment core on the training data

Save the trained MODEL. You may use this model for a new submission

The first challenge - Moving beyond 9mers

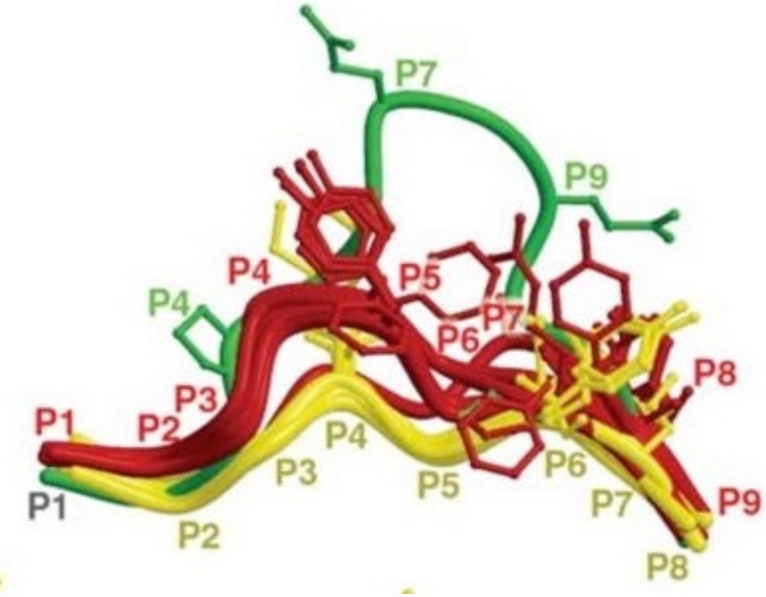
- Most MHC class I binding methods are trained on 9mer peptide binding data only
- Close to 30% of binding data available have length <> 9

Reconciling multiple binding models - Peptide binding to MHC class I

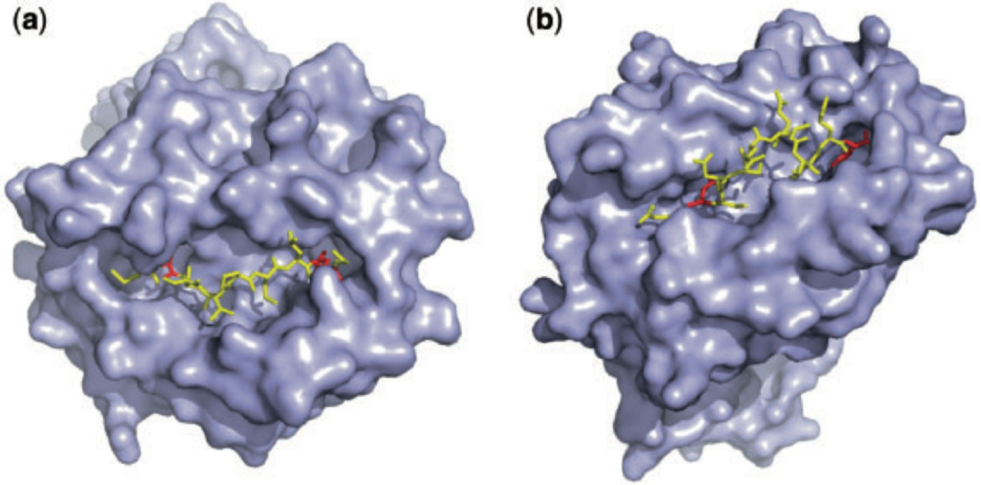


BoLA-HD6

VGYPKVKEEML



HLA-A*03:01

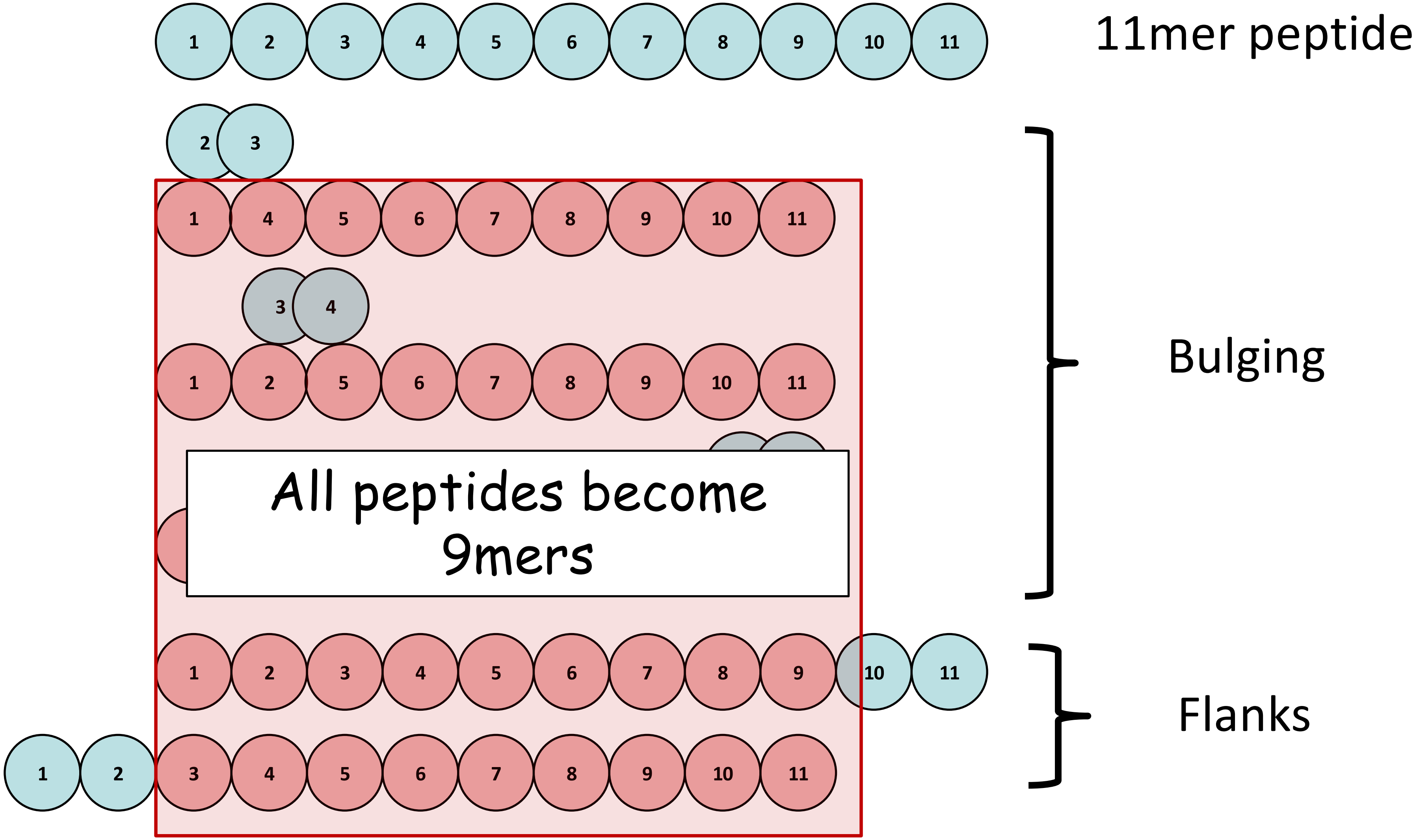


MLLSVPLLLG

EECDSELEIKRY

MLQGRGPLK	*1.0nM*	*MLQGRGPLK*
CAHHFWTK	*169nM*	*CAHHFWT-K*
TMRIYCSLFK	*1.2nM*	*TMRYCSLFK*
RMRGAHTNDVK	*1.0nM*	*RMRGAHTNK*

Different possible binding modes



All peptides become 9mers

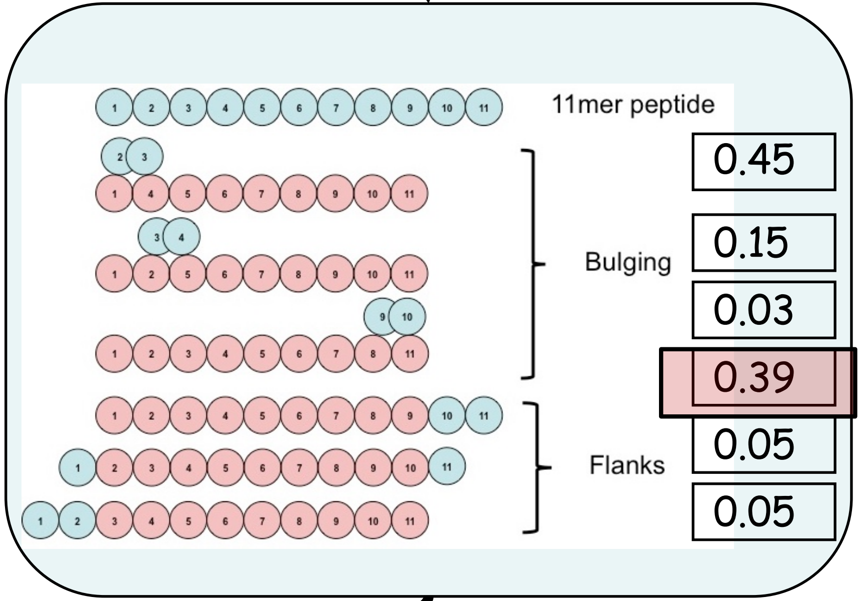
A "CNN like" model before the era of CNN's

Update method to Minimize prediction error

NNAlign-2.0

Predict binding affinity and core

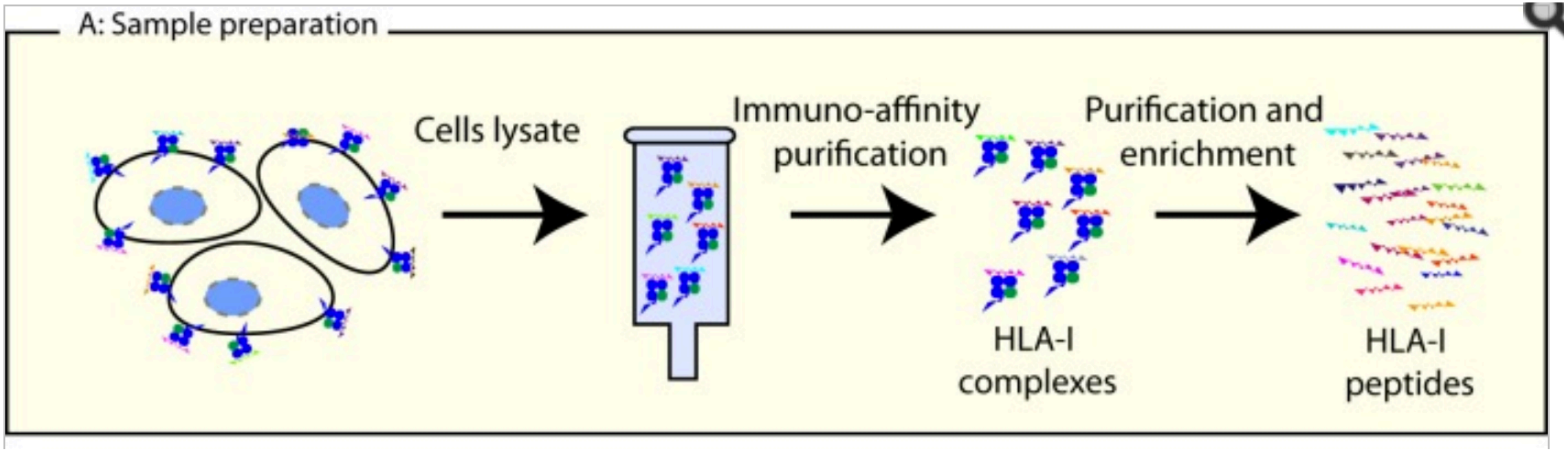
PEPTIDE	Pred	Meas	9mer_Core
AEMKTDAA	0.022	0.000	AEMK-TDAA
HHIWQNLL	0.029	0.000	HHI-WQNLL
APLAHRLGM	0.065	0.085	APLAHRLGM
GGFTFKRTK	0.385	0.547	GGFTFKRTK
LPTWLGAAI	0.029	0.085	LPTWLGAAI
DILGVLTIK	0.376	0.351	DILGVLTIK
DIVNNFITK	0.430	0.361	DIVNNFITK
RRRKGWIPL	0.058	0.213	RRRKGWIPL
SLSEPWRDF	0.078	0.085	SLSEPWRDF
RELVRKTRF	0.028	0.085	RELVRKTRF
IISDMYDPR	0.412	0.556	IISDMYDPR
LQAGFFLLTR	0.443	0.394	LQAGFFLLR



Calculate prediction error

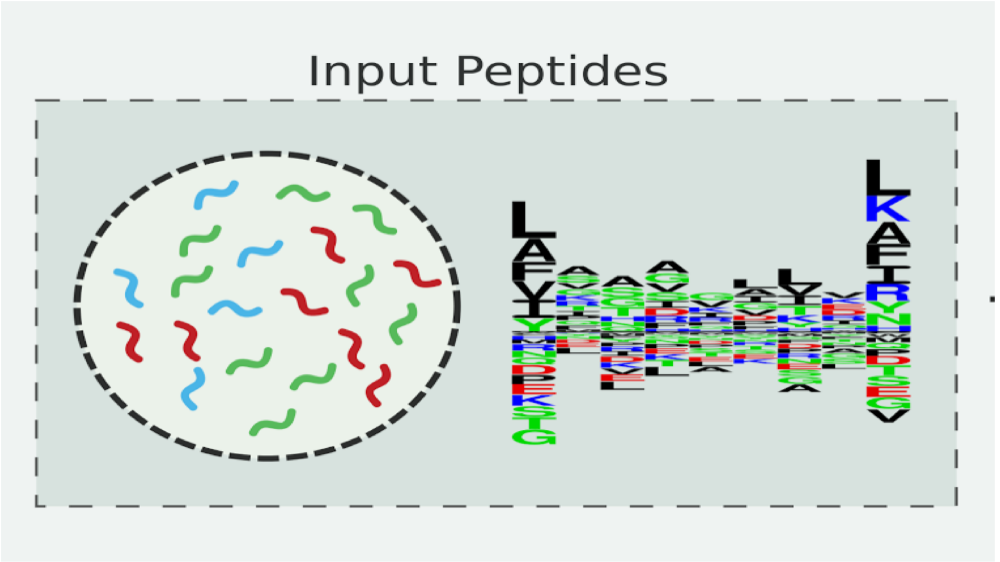
Andreatta, Nielsen, Bioinformatics 2016
Nielsen, Andreatta, Genome Medicine, 2016
Nielsen M, Andreatta M., NAR 2017

Interpreting (and benefitting from) MS eluted ligand data sets

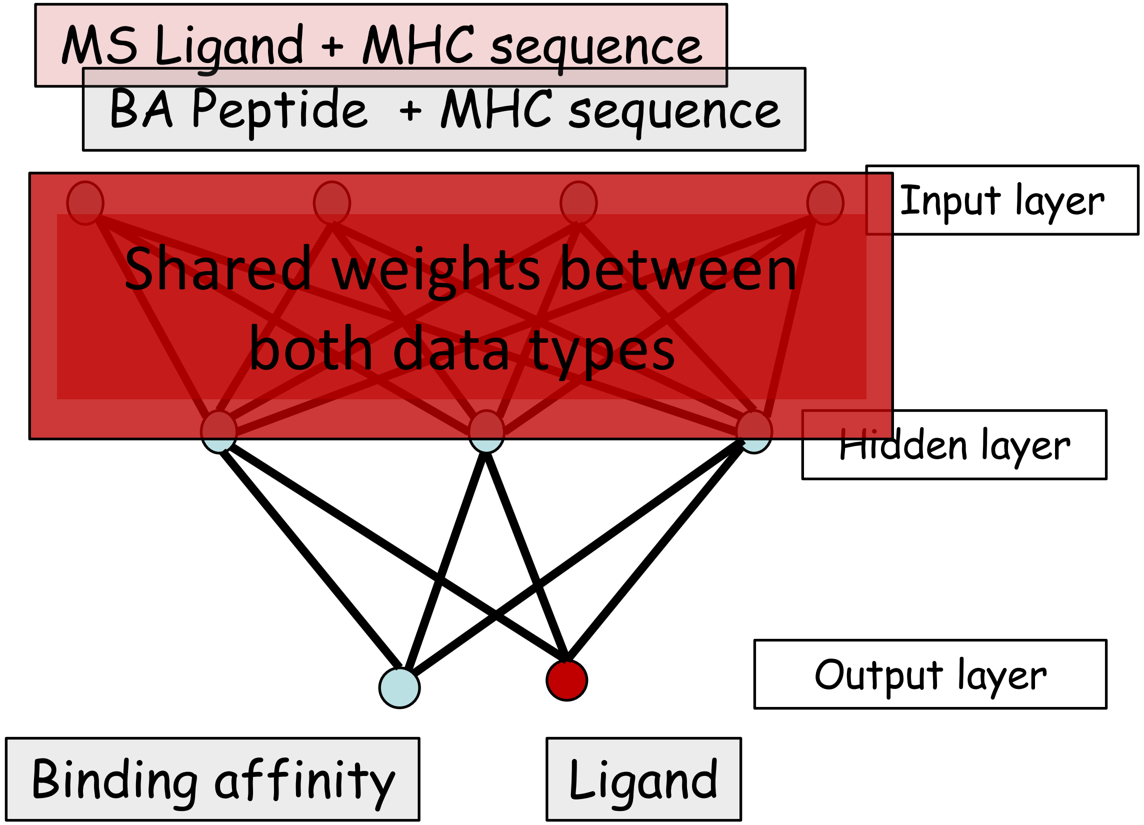


Michal Bassani-Sternberg et. al, MCP, 2015

In silico analysis



How to train on mixed data types (benefiting from MS ligand data)?



MS Ligand + MHC sequence

BA Peptide + MHC sequence

Shared weights between both data types

Input layer

Hidden layer

Output layer

Binding affinity

Ligand

185,985 data points covering 153 MHC-I molecules

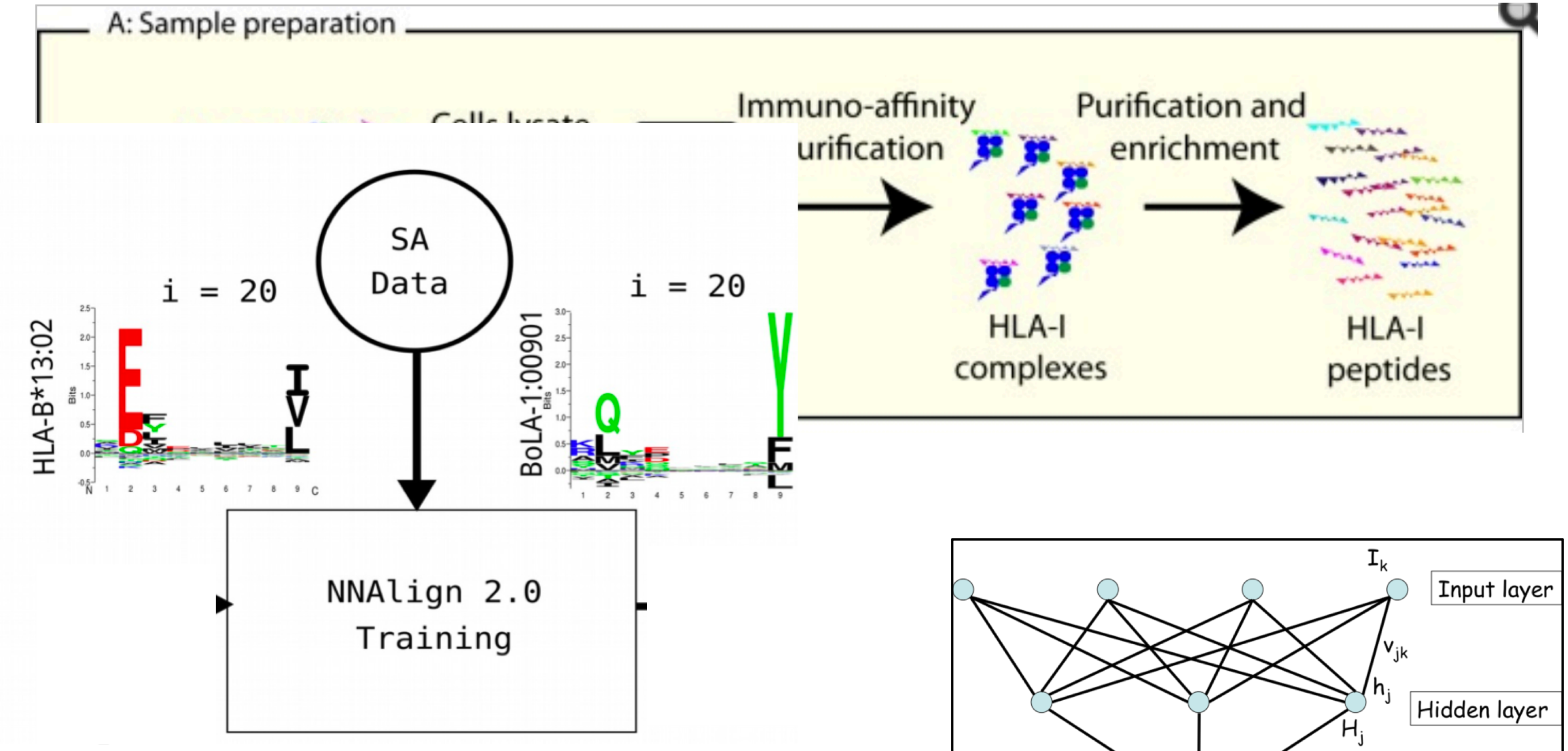
84,717 data points covering 55 HLA-I molecules

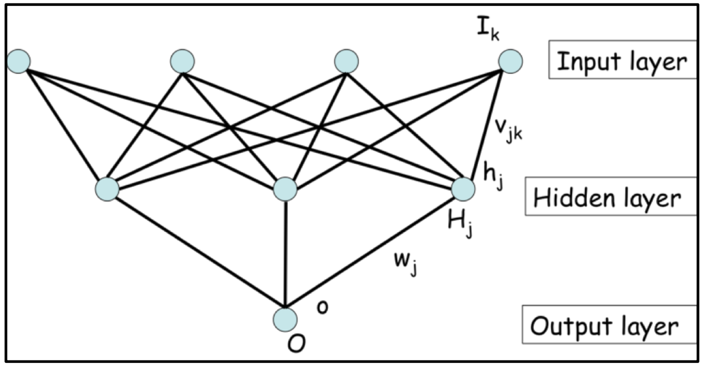
Neural network model

- We expand the NNalign approach by adding a second output neuron
- Training is performed on both data simultaneously
- Resulting model is able to predict binding affinity value and probability of peptide being an eluted ligand

V. Jurtz et al. J Immunol. 2017

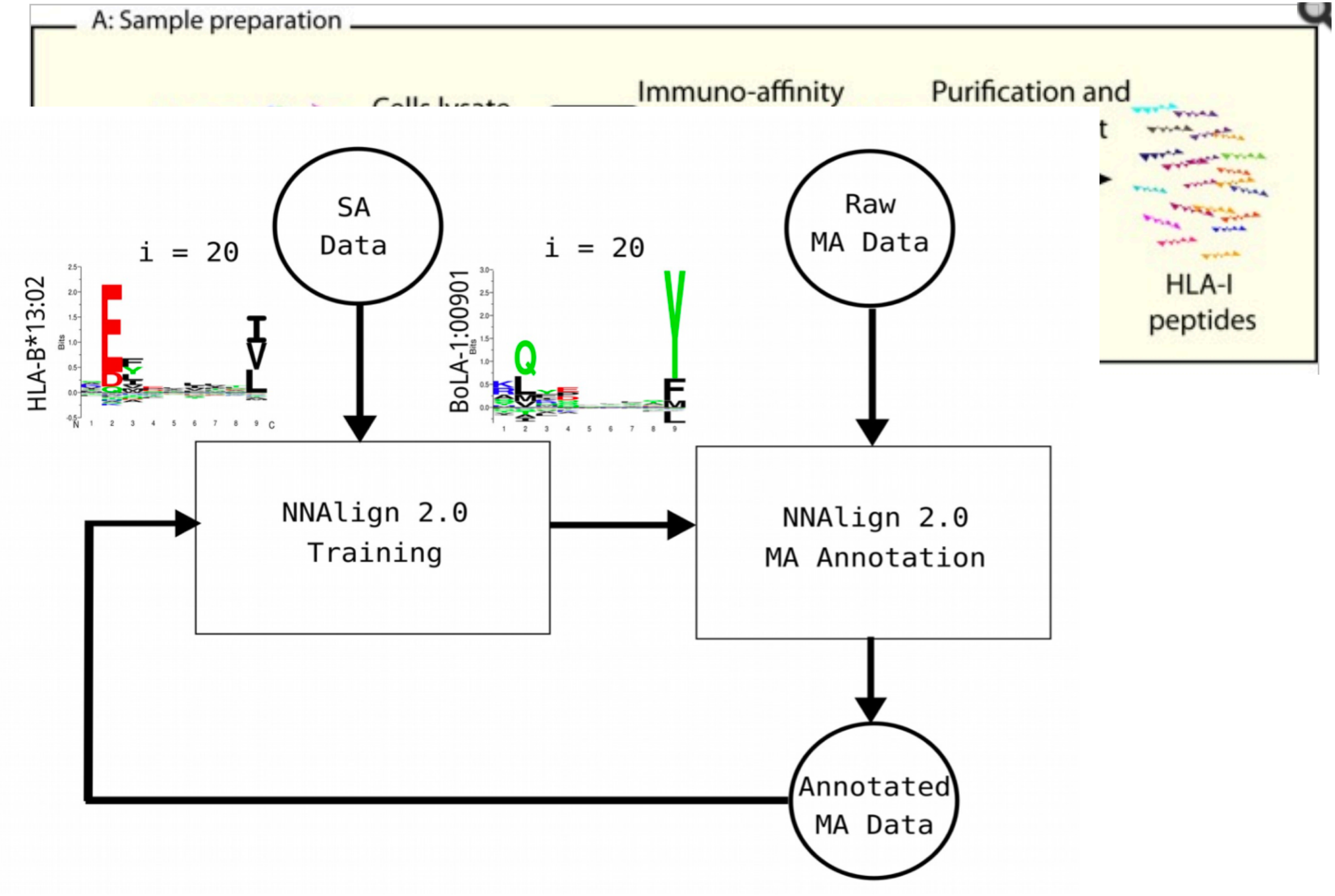
Learning from raw MS data - NNAlign_MA





Alverez et al., 2020

Learning from raw MS data - NNAlign_MA



SA Data

Raw MA Data

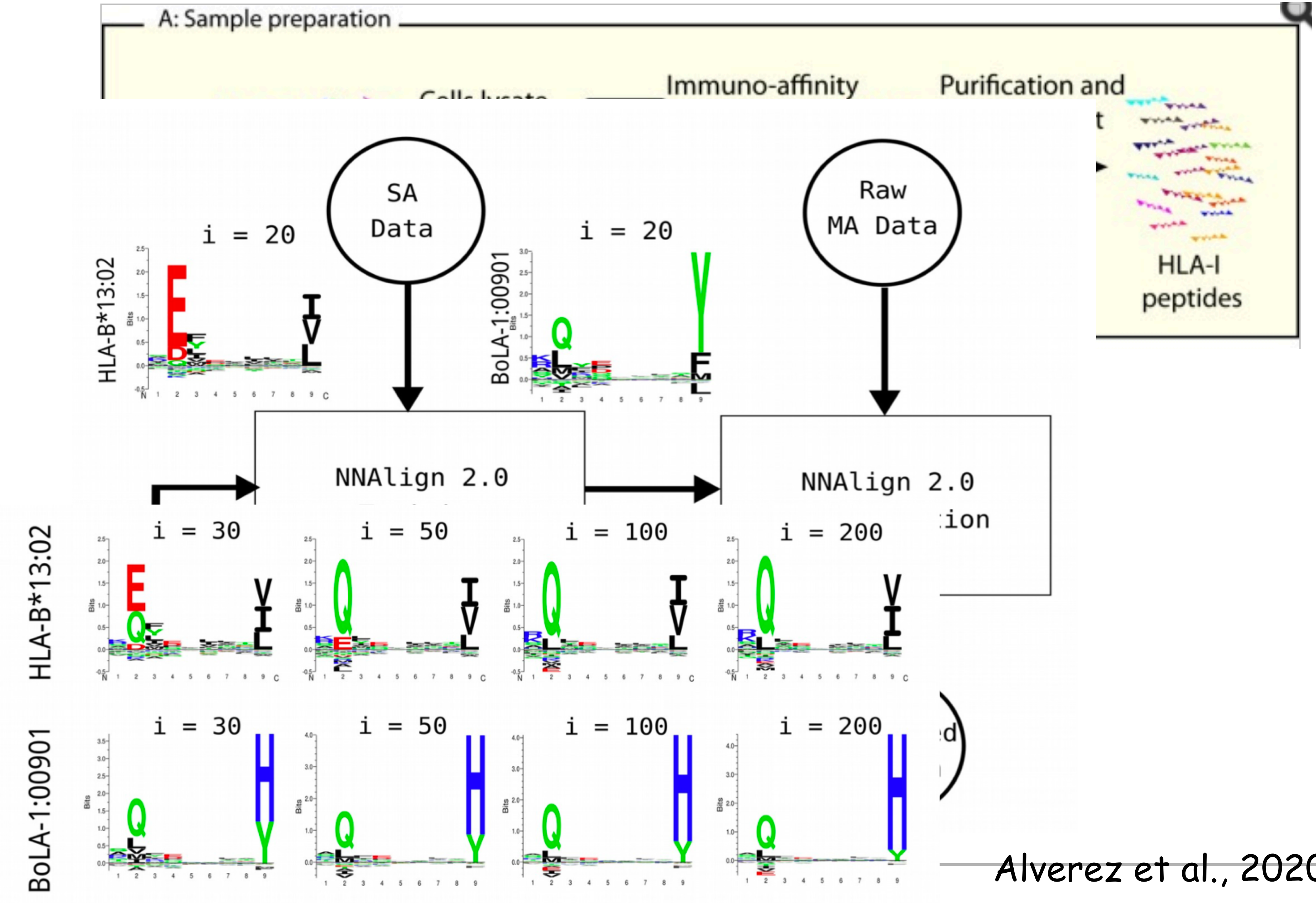
NNAlign 2.0 Training

NNAlign 2.0 MA Annotation

Annotated MA Data

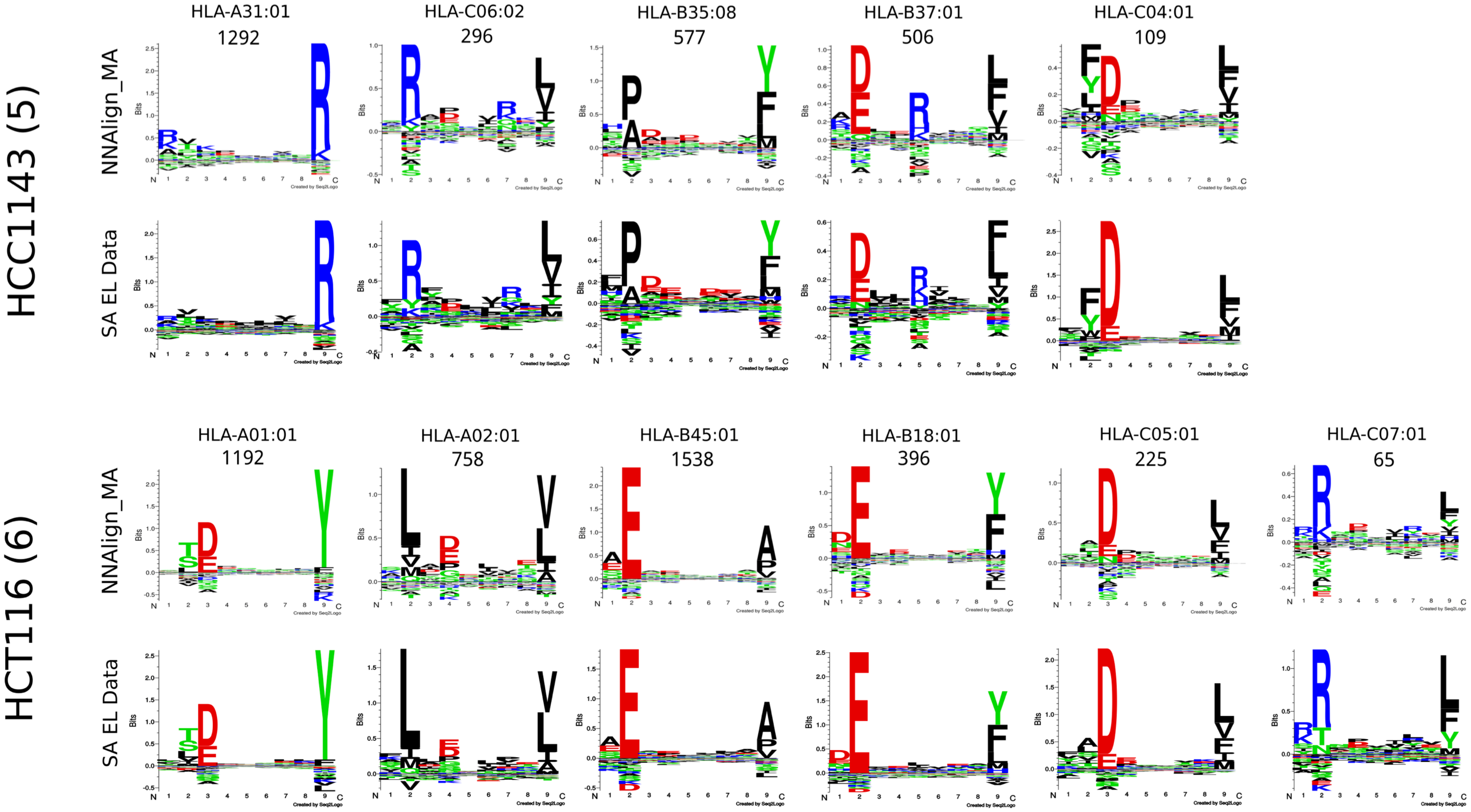
Alverez et al., 2020

Learning from raw MS data - NNAlign_MA



Alverez et al., 2020

NNAlign_MA



Alverez et al., 2020

Conclusions and perfectives

- Receptor ligand systems are effectively characterized using swallow ANN combined with biological intuition
- Knowing how to program a simple FFNN allows you to modify the implementation to integrate mixed data types and to reconcile different peptide binding modes
- You might not always have to go Deep :=)