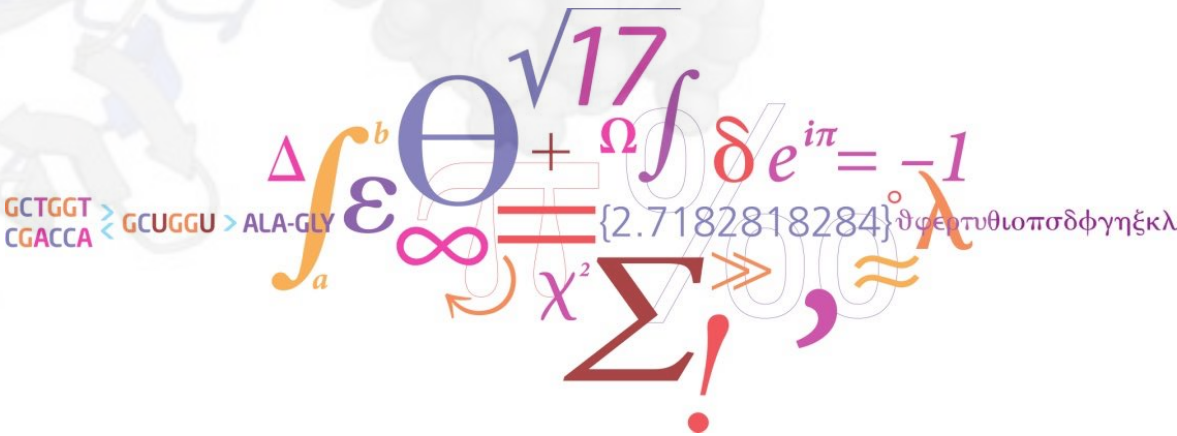
B- and T-cell Epitope prediction

Paolo Marcatili



DTU Biosys
Department of Systems Biology

About me

Assistant Professor at
Technical University of Denmark – Dept. of Systems Biology

Structural Bioinformatics
Immunoinformatics
Machine Learning

IMMUNE EPITOPE DATABASE AND ANALYSIS RESOURCE

La Jolla Institute FOR ALLERGY AND IMMUNOLOGY | Life Without Disease.

Tools for Immunological data

Immunological features

ArrayPitope
Residue-level epitope mapping of antigens based on peptide microarray data

BepiPred »
Linear B-cell epitopes

DiscoTope »
Discontinuous B-cell epitopes

HLArestrictor
Patient-specific HLA restriction elements and optimal epitopes within peptides

LYRA
Lymphocyte receptor automated modelling

MHCcluster
MHC class 1 clustring based on binding specififcity

NetChop »
Proteasomal cleavages (MHC ligands)

NetCTL »
Integrated class I antigen presentation

NetCTLpan »
Pan-specific integrated class I antigen presentation

NetMHC »
Binding of peptides to MHC class I alleles

NetMHCcons »
Binding of peptides to any known MHC class I molecule

NetMHCII »
Binding of peptides to MHC class II alleles

NetMHCIIpan »
Pan-specific binding of peptides to MHC class II alleles of known sequence

NetMHCpan »
Pan-specific binding of peptides to MHC class I alleles of known sequence

NetMHCstab
Stability of peptide:MHC-I complexes

NetTepi
T-cell epitopes restricted to prevalent HLA-A and HLA-B molecules

NNAlign
Identifying sequence motifs in quantitative peptide data

PickPocket »
Binding of peptides to any known MHC class I molecule using binding pocket matrix extrapolation

VDJsolver »
Analysis of human immunoglobulin VDJ recombination

http://www.cbs.dtu.dk/services/

Tabhu: Tool for AntiBody Humanization

Website: http://www.biocomputing.it/tabhu

Olimpieri P.P., Marcatili P. and Tramontano A. (2015) Tabhu: tools for antibody humanization. *In press.* do

proABC for the predicition of site-specific interactions in antibody-antigen comple

Website: http://www.biocomputing.it/proABC

Olimpieri P.P., Chailyan A., Tramontano A. and Marcatili P. (2013) Prediction of site-specific interactions i
server. Bioinformatics 29(18):2285-91. doi:10.1093/bioinformatics/btt369

DIGIT: a database of immunoglobulin variable domain sequences

Website: http://www.biocomputing.it/digit/

Chailyan A., Tramontano A., Marcatili P. (2012) A database of immunoglobulins with integrated tools: DIG
D1234. doi:10.1093/nar/gkr806

PIGS: Prediction of Immunoglobulin Structures

Website: www.biocomputing.it/pigs

1. Marcatili P., Rosi A. and Tramontano A. (2008) PIGS: Automatic prediction of antibody structures Bioinf
doi:10.1093/bioinformatics/btn341

2. Marcatili, P., Olimpieri, P. P., Chailyan, A., & Tramontano, A. (2014). Antibody modeling using the Predi
protocols, *9*(12), 2771-2783. doi:10.1038/nprot.2014.189

http://biocomputing.it/index.php/About-us/tools

Epitopes

Epitopes

T-cell

B-cell

Epitopes

T-cell

- Neural Networks
- Pan-allele
- Insertions
- Length Preference

B-cell

Sequence Based –

Structure based –

Epitopes

T-cell

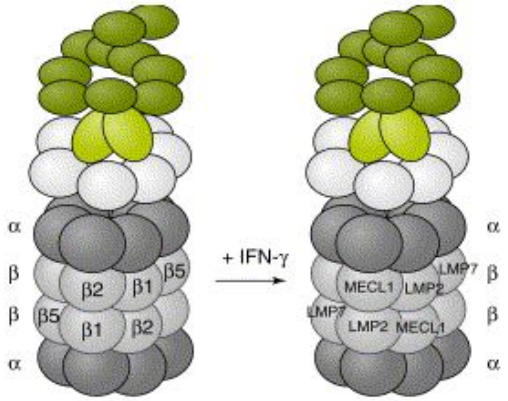
- NetMHC I & II
- NetMHCpan I & II

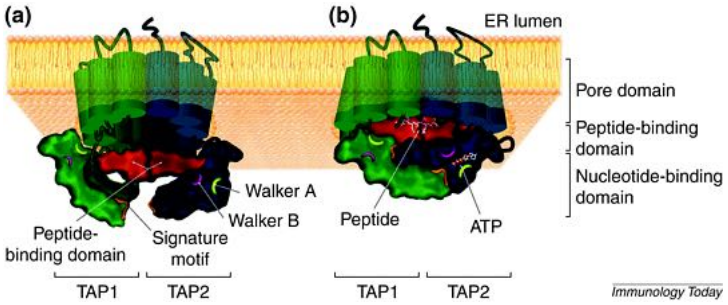
B-cell

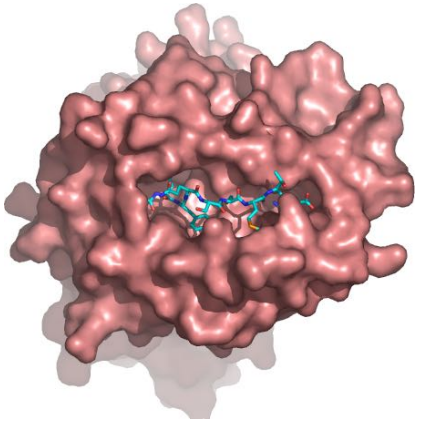
BepiPred –
Discotope –

T-cell epitope prediction

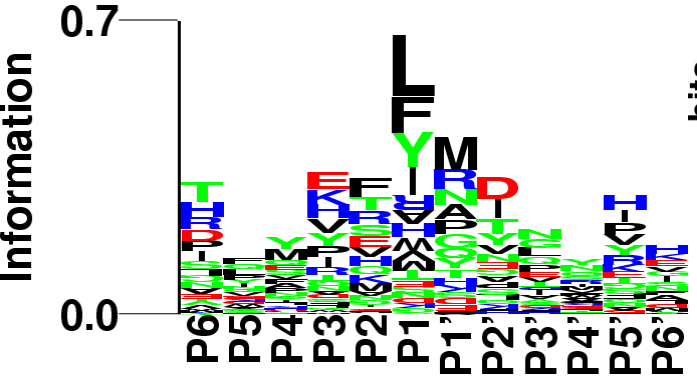
State of the art



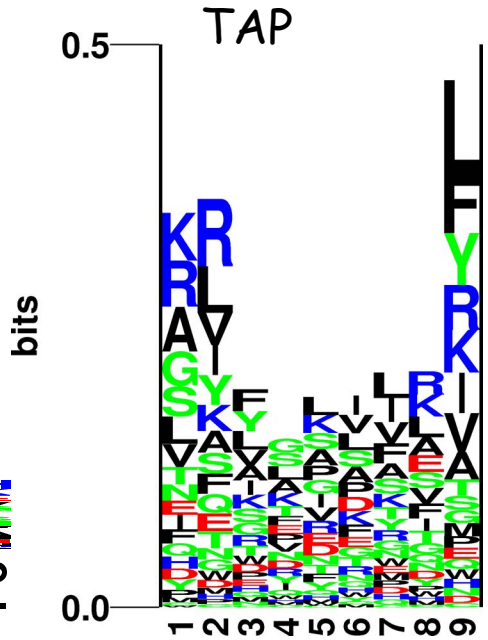




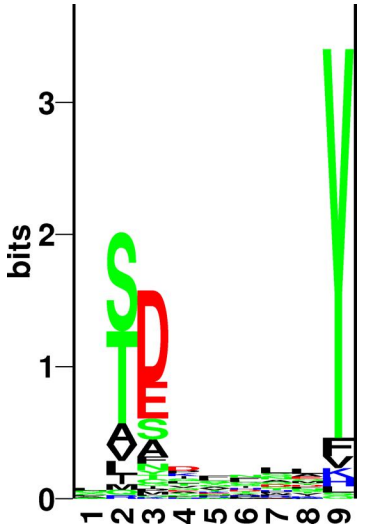
Immuno proteasome



TAP

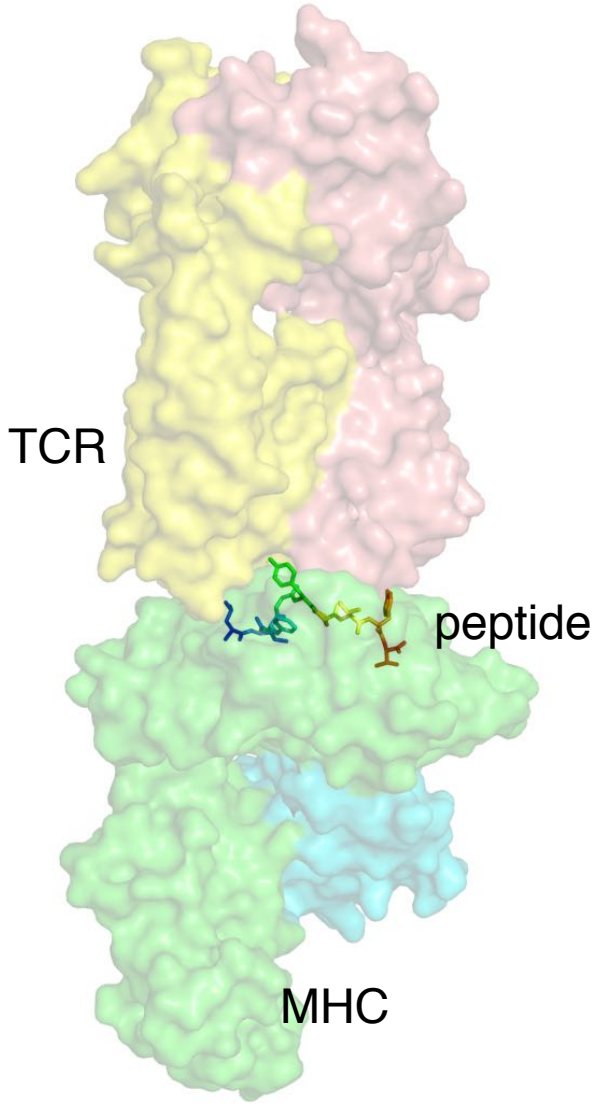


MHC



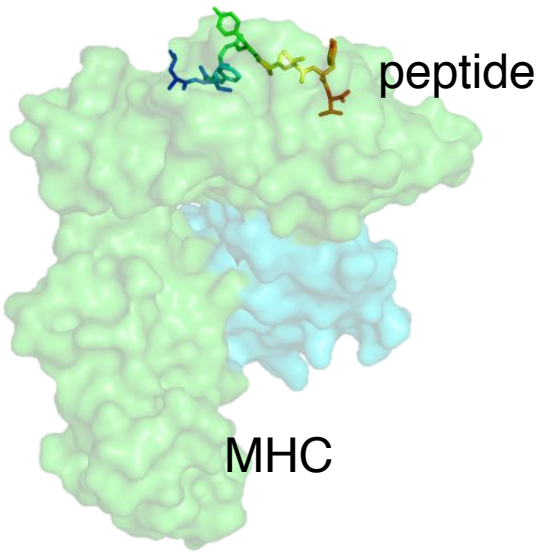
State of the art

MHC Class I	: 0.89 AUC
MHC Class II	: 0.81 AUC
Immunogenicity Class I	: ~ 0.7 AUC



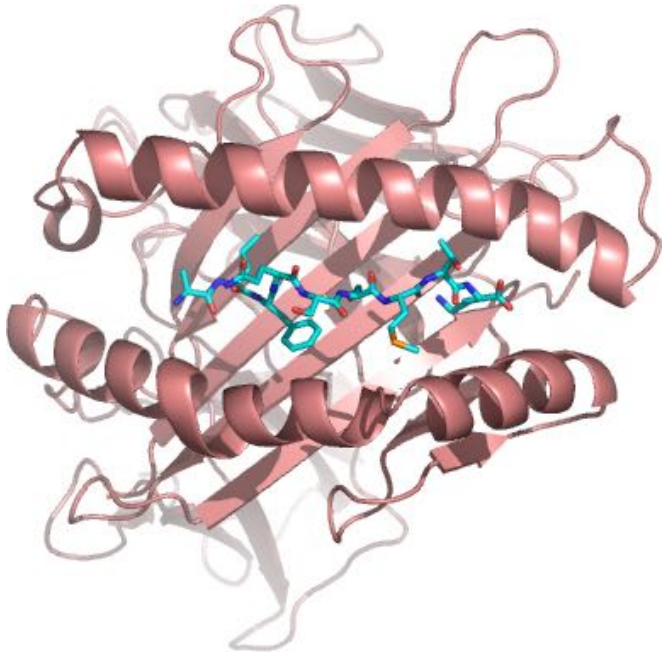
State of the art

MHC Class I	: 0.89 AUC
MHC Class II	: 0.81 AUC
Immunogenicity Class I	: ~ 0.7 AUC



State of the art

MHC Class I	: 0.89 AUC
MHC Class II	: 0.81 AUC
Immunogenicity Class I	: ~ 0.7 AUC

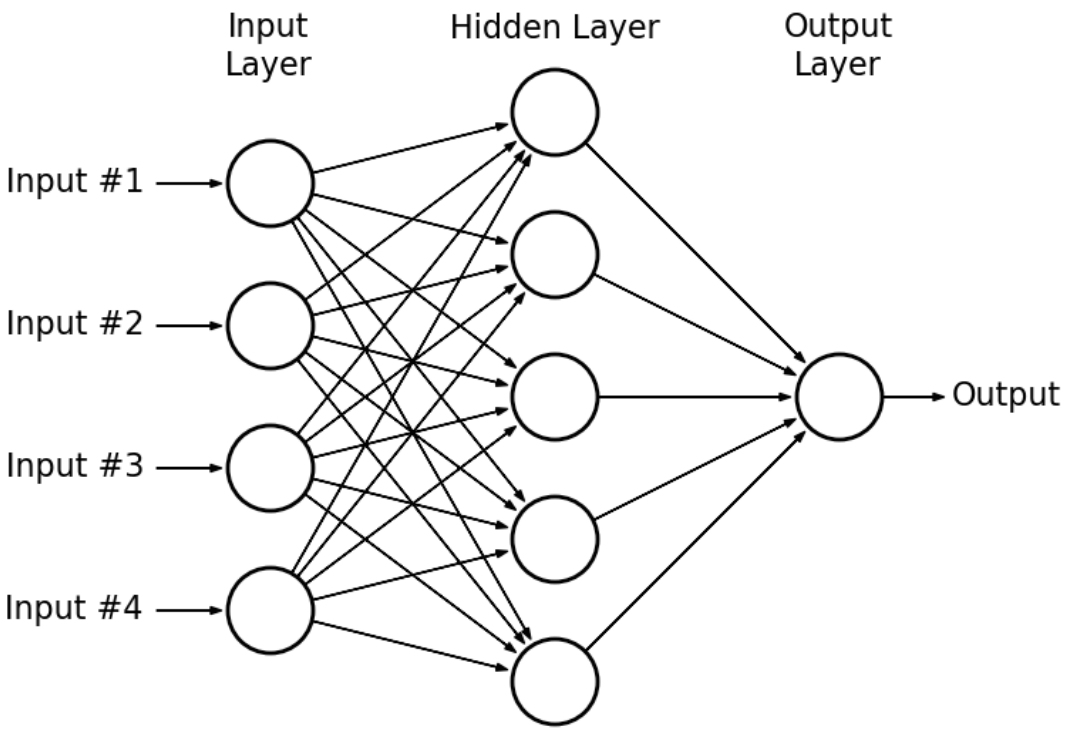


Neural Networks

SLLPAIVEL YLLPAIVHI TLWVDPYEV GLVPFLVSV KLLEPVLLL LLDVPTAAV LLDVPTAAV LLDVPTAAV
LLDVPTAAV VLFRGGPRG MVDGTLLLL YMNGTMSQV MLLSVPLLL SLLGLLVEV ALLPPINIL TLIKIQHTL
HLIDYLVTS ILAPPVVKL ALFPQLVIL GILGFVFTL STNRQSGRQ GLDVLTAKV RILGAVAKV QVCERIPTI
ILFGHENRV ILMEHIHKL ILDQKINEV SLAGGIIGV LLIENVASL FLLWATAEA SLPDFGISY KKREEAPSL
LERPGGNEI ALSNLEVKL ALNELLQHV DLERKVESL FLGENISNF ALSDHHIYL GLSEFTEYL STAPPAHGV
PLDGEYFTL GVLVGVALI RTLDKVLEV HLSTAFARV RLDSYVRSL 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 YIGEVLVSV CINGVCWTV VMNILLQYV
ILTVILGVL KVLEYVIKV FLWGPRALV GLSRYVARL FLLTRILTI HLGNVKYLV GIAGGLALL GLQDCTMLV
TGAPVTYST VIYQYMDDL VLPDVFIRC VLPDVFIRC AVGIGIAVV LVVLGLLAV ALGLGLLPV GIGIGVLAA
GAGIGVAVL IAGIGILAI LIVIGILIL LAGIGLIAA VDGIGILTI GAGIGVLTA AAGIGIIQI QAGIGILLA
KARDPHSGH KACDPHSGH ACDPHSGHF SLYNTVATL RGPGRAFVT NLVPMVATV GLHCYEQLV PLKQHFQIV
AVFDRKSDA LLDFVRFMG VLVKSPNHV GLAPPQHLI LLGRNSFEV PLTFGWCYK VLEWRFDSR TLNAWVKVV
GLCTLVAML FIDSYICQV IISAVVGIL VMAGVGSPY LLWTLVVLL SVRDRLARL LLMDCSGSI CLTSTVQLV
VLHDDLLEA LMWITQCFL SLLMWITQC QLSLLMWIT LLGATCMFV RLTRFLSRV YMDGTMSQV FLTPKKLQC
ISNDVCAQV VKTDGNPPE SVYDFFVWL FLYGALLLA VLFSSDFRI LMWAKIGPV SLLLELEEV SLSRFSWGA
YTAFTIPSI RLMKQDFSV RLPRIFCSC FLWGPRAYA RLLQETELV SLFEGIDFY SLDQSVVEL RLNMFTPYI
NMFTPYIGV LMIIPLINV TLFIGSHVV SLVIVTTFV VLQWASLAV ILAKFLHWL STAPPHVNV LLLLTVLTV
VVLGVVFGI ILHNGAYSL MIMVKCWMI MLGTHTMEV MLGTHTMEV SLADTNSLA LLWAARPRL GVALQTMKQ
GLYDGMEHL 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 SVRDRLARL SLREWLLRI LLSAWILTA AAGIGILTV AVPDEIPPL FAYDGKDYI
AAGIGILTV FLPSDFFPS AAGIGILTV FLPSDFFPS AAGIGILTV FLWGPRALV ETVSEQSNV ITLWQRPLV

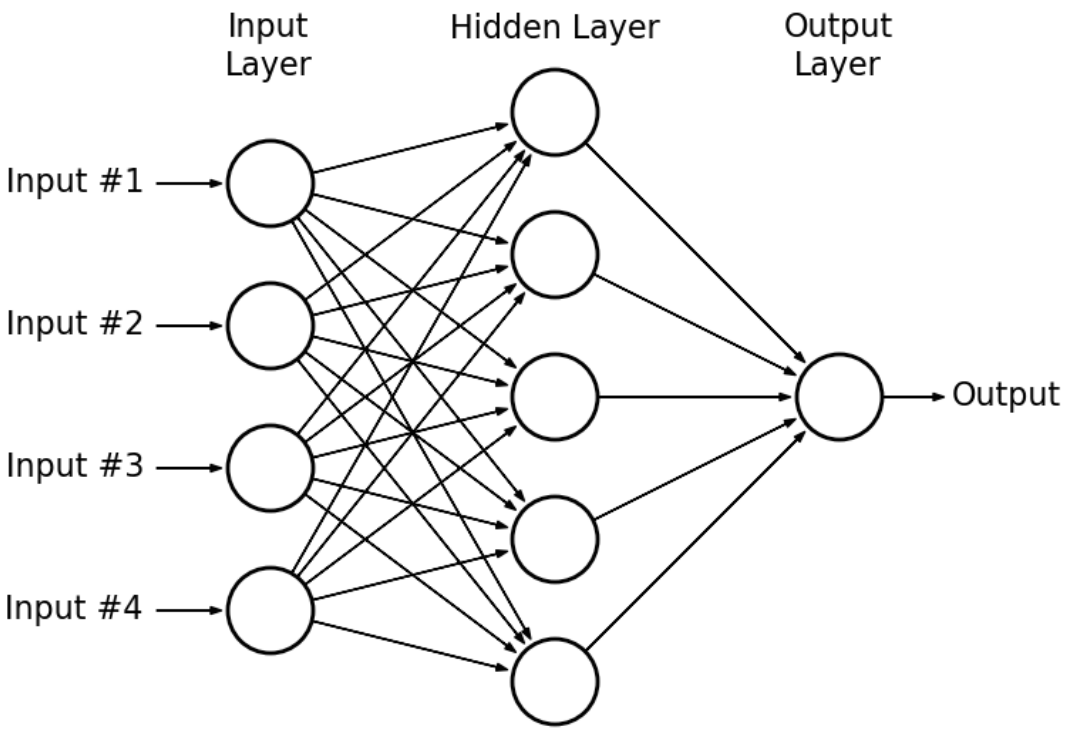
Neural Networks

SLLPAIVEL YLLPAIVHI TLWVDPYEV GLVP
LLDVPTAAV VLFRGGPRG MVDGTLLLL YMNG
HLIDYLVTS ILAPPVVKL ALFPQLVIL GILG
ILFGHENRV ILMEHIHKL ILDQKINEV SLAG
LERPGGNEI ALSNLEVKL ALNELLQHV DLER
PLDGEYFTL GVLVGVALI RTLDKVLEV HLST
ILGFVFTLT LLFGYPVYV GLSPTVWLS WLSL
KLVALGINA DLMGYIPLV RLVTLKDIV MLLA
KTWGQYWQV TITDQVPFS AFHHVAREL YLNK
SLLAPGAKQ KIFGSLAFL ELVSEFSRM KLTP
ILTVILGVL KVLEYVIKV FLWGPRALV GLSR
TGAPVTYST VIYQYMDDL VLPDVFIRC VLPD
GAGIGVAVL IAGIGILAI LIVIGILIL LAGI
KARDPHSGH KACDPHSGH ACDPHSGHF SLYN
AVFDRKSDA LLDFVRFMG VLVKSPNHV GLAP
GLCTLVAML FIDSYICQV IISAVVGIL VMAG
VLHDDLLEA LMWITQCFL SLLMWITQC QLSL
ISNDVCAQV VKTDGNPPE SVYDFFVWL FLYG
YTAFTIPSI RLMKQDFSV RLPRIFCSC FLWG
NMFTPYIGV LMIIPLINV TLFIGSHVV SLVI
VVLGVVFGI ILHNGAYSL MIMVKCWMI MLGT
GLYDGMEHL KMVELVHFL YLQLVFGIE MLMA
EAAGIGILT TLDSQVMSL STPPPGTRV KVAE
YLSTAFARV YLLEMLWRL SLDDYNHLV RTLD
FLDEFMEGV ALQPGTALL VLDGLDVLL SLYS
AAGIGILTV FLPSDFFPS SVRDRLARL SLRE
AAGIGILTV FLPSDFFPS AAGIGILTV FLPS



Neural Networks

SLLPAIVEL YLLPAIVHI TLWVDPYEV GLVPF
LLDVPTAAV VLFRGGPRG MVDGTLLLL YMNGT
HLIDYLVTS ILAPPVVKL ALFPQLVIL GILGF
ILFGHENRV ILMEHIHKL ILDQKINEV SLAGG
LERPGGNEI ALSNLEVKL ALNELLQHV DLERK
PLDGEYFTL GVLVGVALI RTLDKVLEV HLSTA
ILGFVFTLT LLFGYPVYV GLSPTVWLS WLSLI
KLVALGINA DLMGYIPLV RLVTLKDIV MLLAV
KTWGQYWQV TITDQVPFS AFHHVAREL YLNKI
SLLAPGAKQ KIFGSLAFL ELVSEFSRM KLTPI
ILTVILGVL KVLEYVIKV FLWGPRALV GLSRY
TGAPVTYST VIYQYMDDL VLPDVFIRC VLPDV
GAGIGVAVL IAGIGILAI LIVIGILIL LAGIG
KARDPHSGH KACDPHSGH ACDPHSGHF SLYNI
AVFDRKSDA LLDFVRFMG VLVKSPNHV GLAPI
GLCTLVAML FIDSYICQV IISAVVGIL VMAGV
VLHDDLLEA LMWITQCFL SLLMWITQC QLSLI
ISNDVCAQV VKTDGNPPE SVYDFFVWL FLYGA
YTAFTIPSI RLMKQDFSV RLPRIFCSC FLWGF
NMFTPYIGV LMIIPLINV TLFIGSHVV SLVIV
VVLGVVFGI ILHNGAYSL MIMVKCWMI MLGTH
GLYDGMEHL KMVELVHFL YLQLVFGIE MLMAQ
EAAGIGILT TLDSQVMSL STPPPGTRV KVAEI
YLSTAFARV YLLEMLWRL SLDDYNHLV RTLDK
FLDEFMEGV ALQPGTALL VLDGLDVLL SLYSF
AAGIGILTV FLPSDFFPS SVRDRLARL SLREM
AAGIGILTV FLPSDFFPS AAGIGILTV FLPSL



AUC >.9 for common alleles

HLA Variability

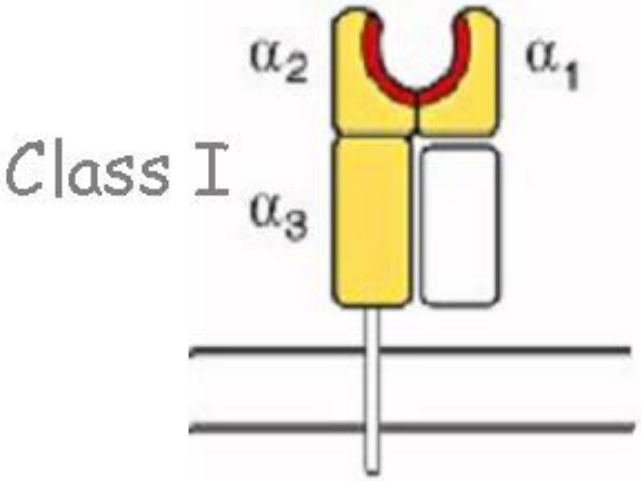
binding data for < 70 HLA alleles

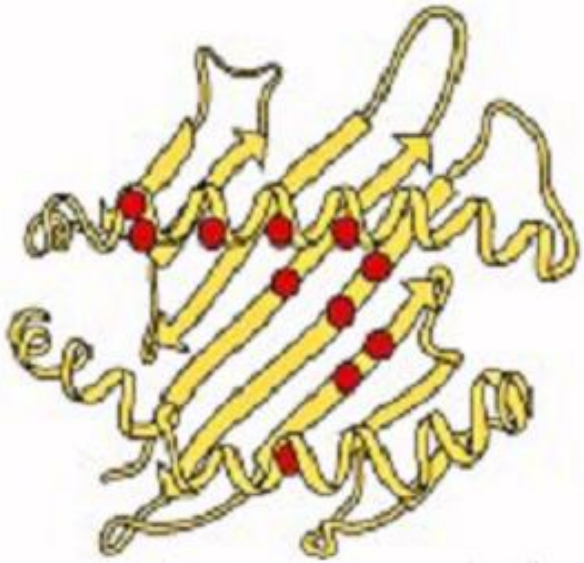
Reliable predictions (NetMHC-3.2)
for 57 HLA A and B molecules

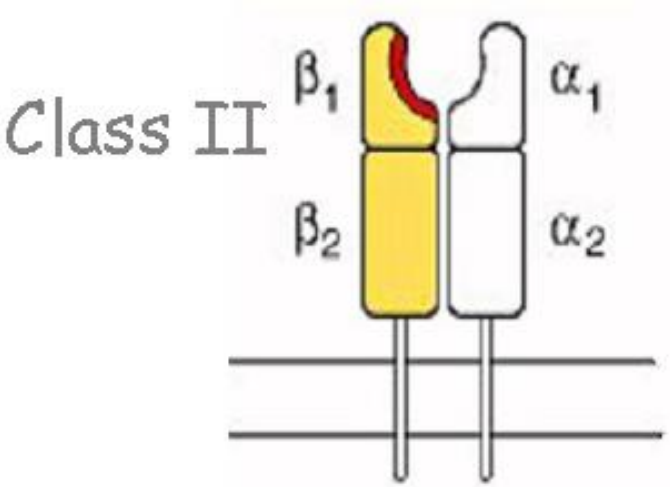
No methods for HLA-C, and HLA-E

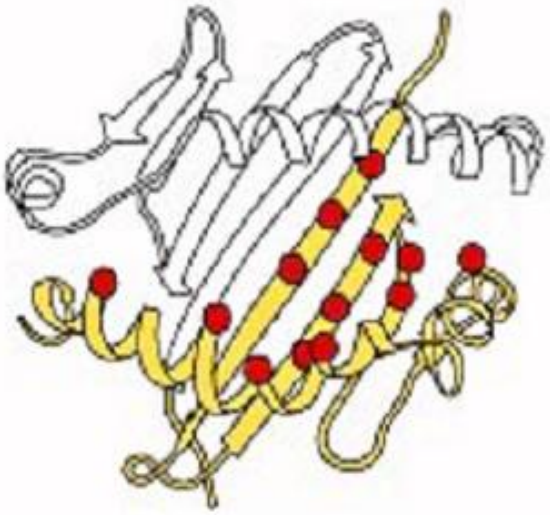
Long way to over 1500!

HLA Variability



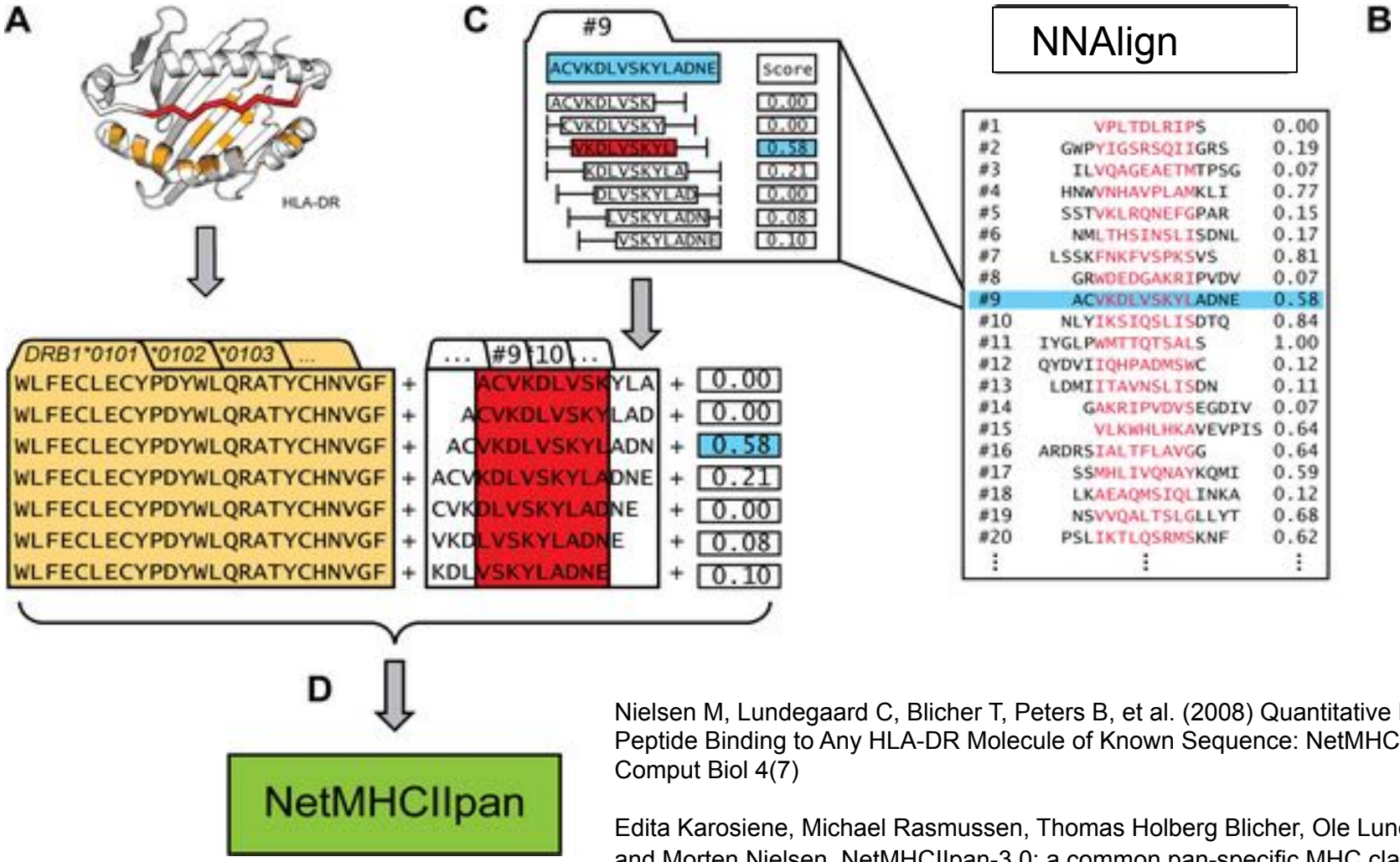






http://rheumb.bham.ac.uk/teaching/immunology/tutorials/mhc%20polymorphism.jpg

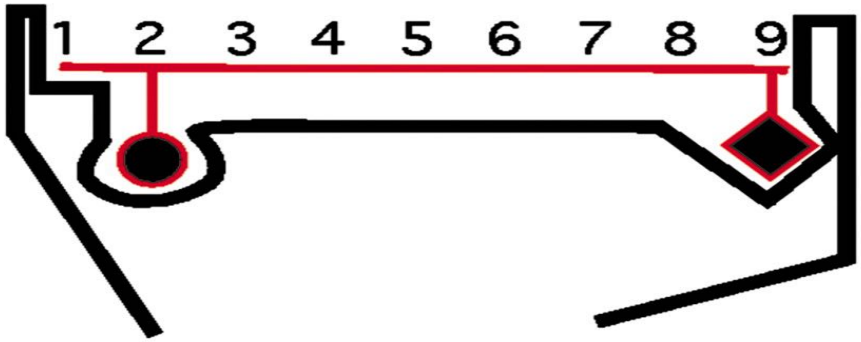
Pan specific Methods



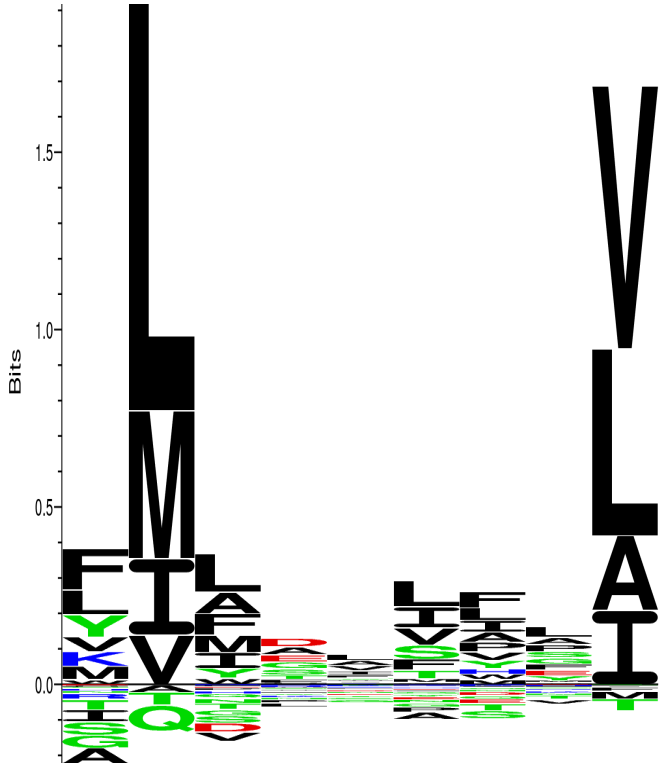
Nielsen M, Lundegaard C, Blicher T, Peters B, et al. (2008) Quantitative Predictions of Peptide Binding to Any HLA-DR Molecule of Known Sequence: NetMHCIIpan. PLoS Comput Biol 4(7)

Edita Karosiene, Michael Rasmussen, Thomas Holberg Blicher, Ole Lund, Søren Buus and Morten Nielsen. NetMHCIIpan-3.0; a common pan-specific MHC class II prediction method including all three human MHC class II isotypes, HLA-DR, -DP and –DQ. Immunogenetics. 2013 Jul 31.

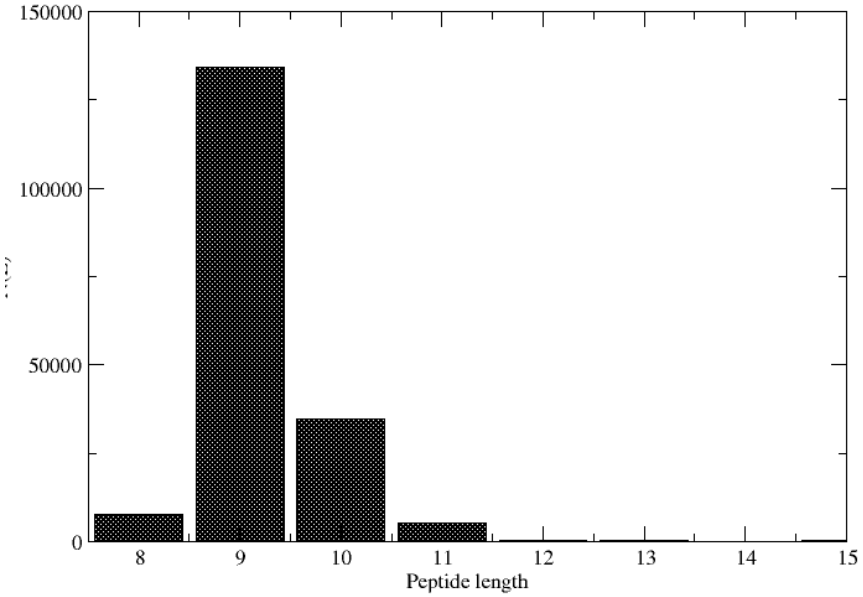
Canonical binding

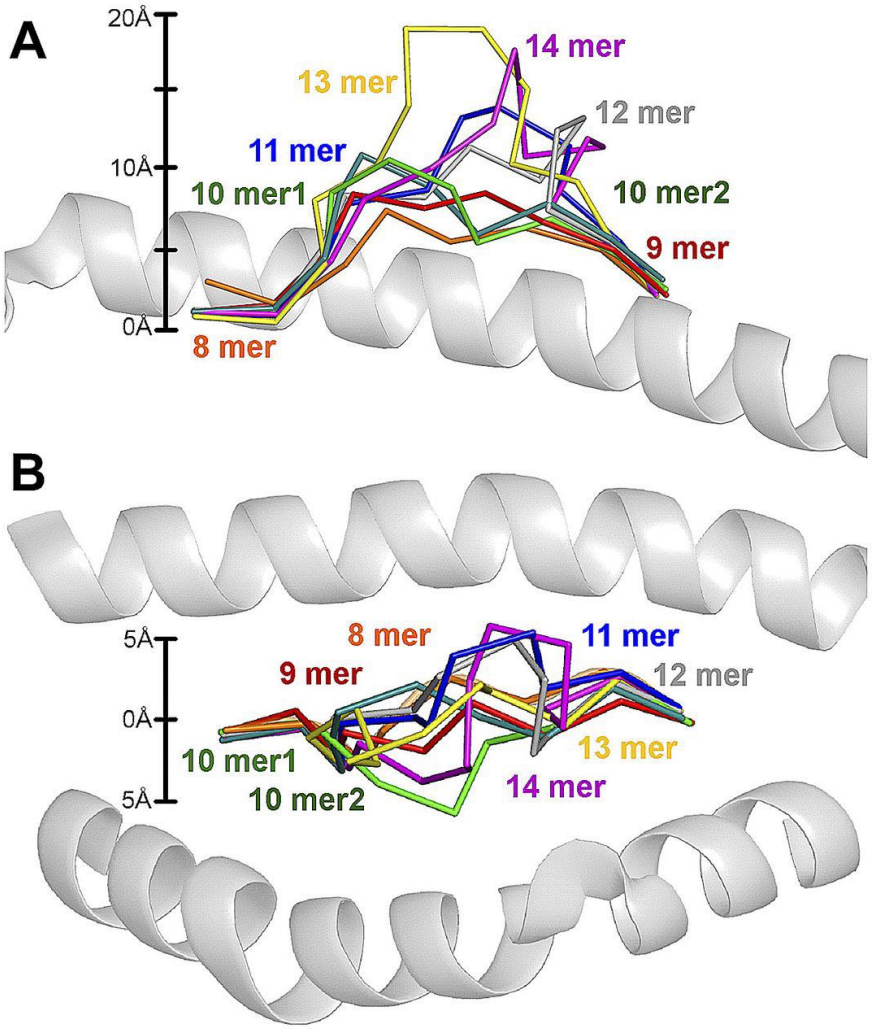


HLA-A*02:01



30% of epitopes are not 9mers





Non-canonical binders

Non-canonical binders

- Ignore non 9mers
  - NetMHCpan

Non-canonical binders

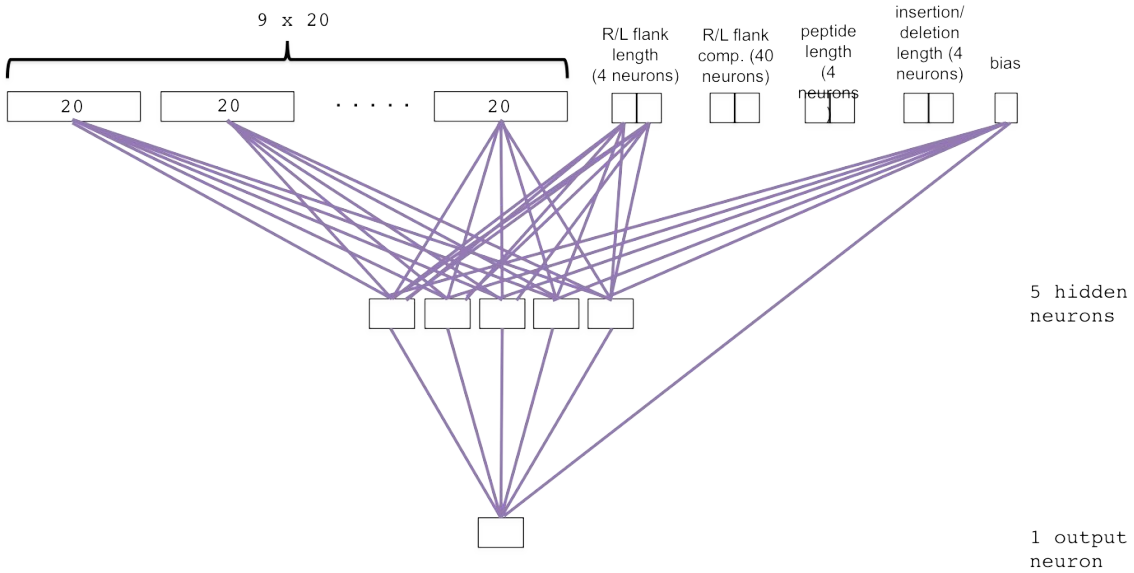
- Ignore non 9mers
  - NetMHCpan
- Train a method for each length
  - SMM
  - NetMHC

Non-canonical binders

- Ignore non 9mers
  - NetMHCpan
- Train a method for each length
  - SMM
  - NetMHC
- Few data for non-9mers

NetMHC 4.0 – gapped alignments

SMMVILPDKI 0.83



SMVILPDKI 0.42
SMVILPDKI 0.42
SMMILPDKI 0.42
SMMVLPDKI 0.43
SMMVIPDKI 0.38
SMMVILDKI 0.37
SMMVILPKI 0.75
SMMVILPDI 0.72
SMMVILPDK 0.62
MMVILPDKI 0.02

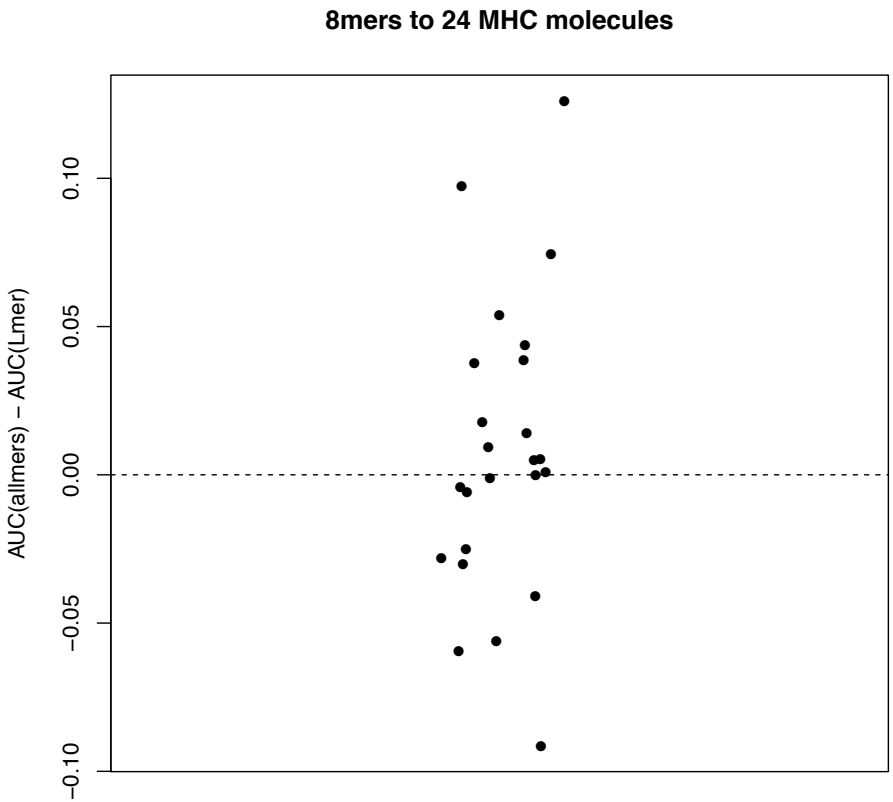
Bulging

Flanks

231*5 + 5 + 1 = 1161 total weights to optimize

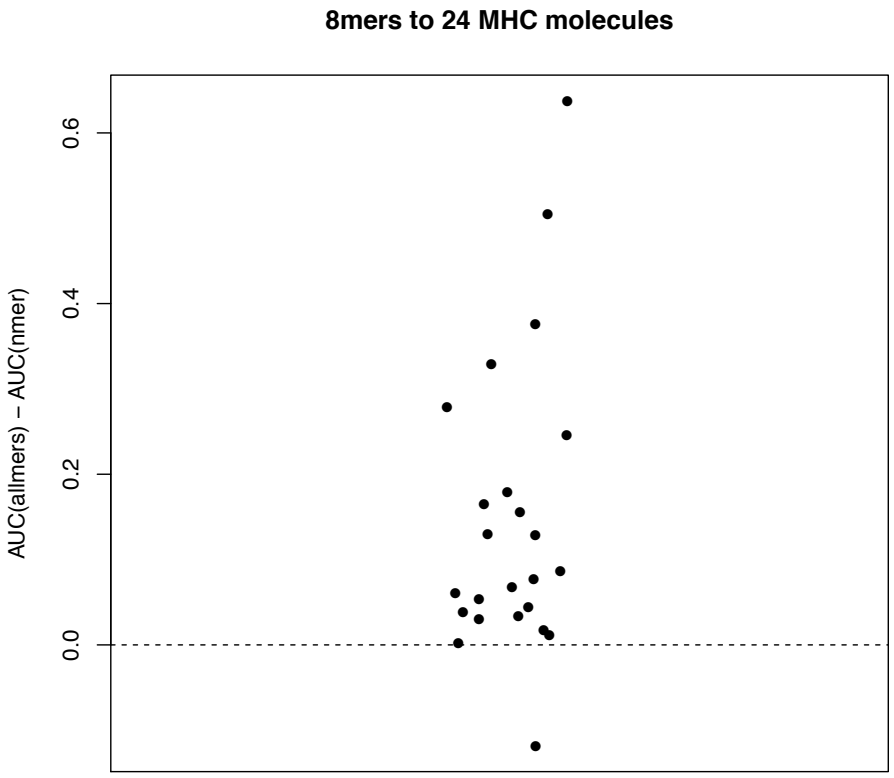
Enrichment of dataset with peptides of different length (8-mers)

all-mers networks vs. networks trained on 9mers only (L-mer approximation)



Networks trained on all-mers have higher AUC in 13/24 cases (p=0.4)

all-mers networks vs. networks trained on 8mers only

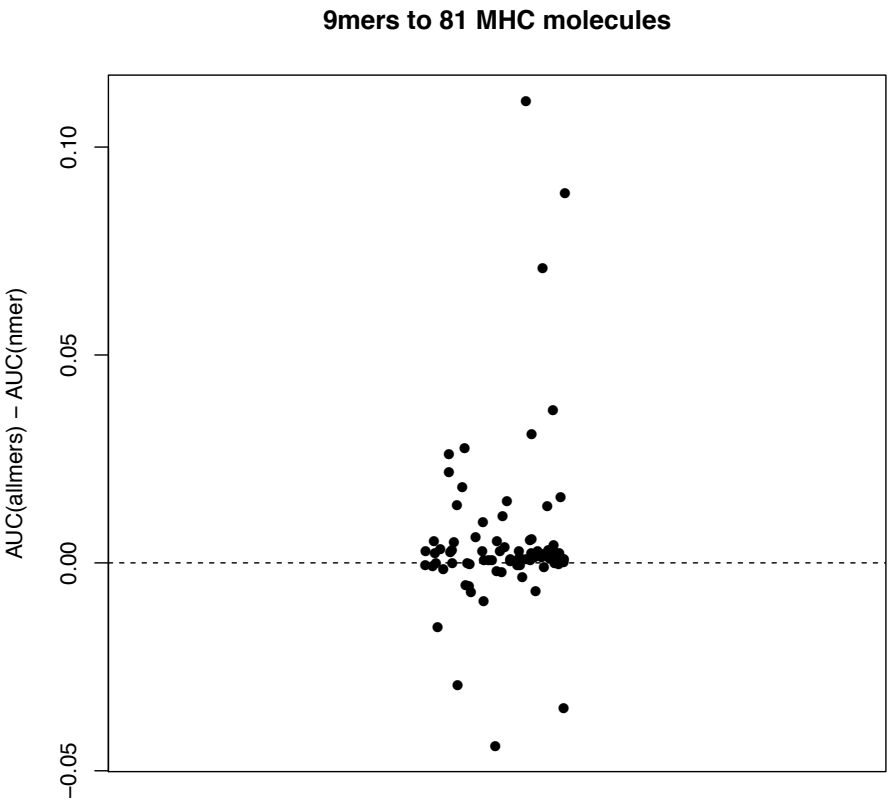


Networks trained on all-mers have higher AUC in 23/24 cases ($p=2*10^{-6}$)

(molecules measured with at > 20 peptides of which at least 4 are binders)

Enrichment of dataset with peptides of different length (9-mers)

all-mers networks vs. networks trained on 9mers only

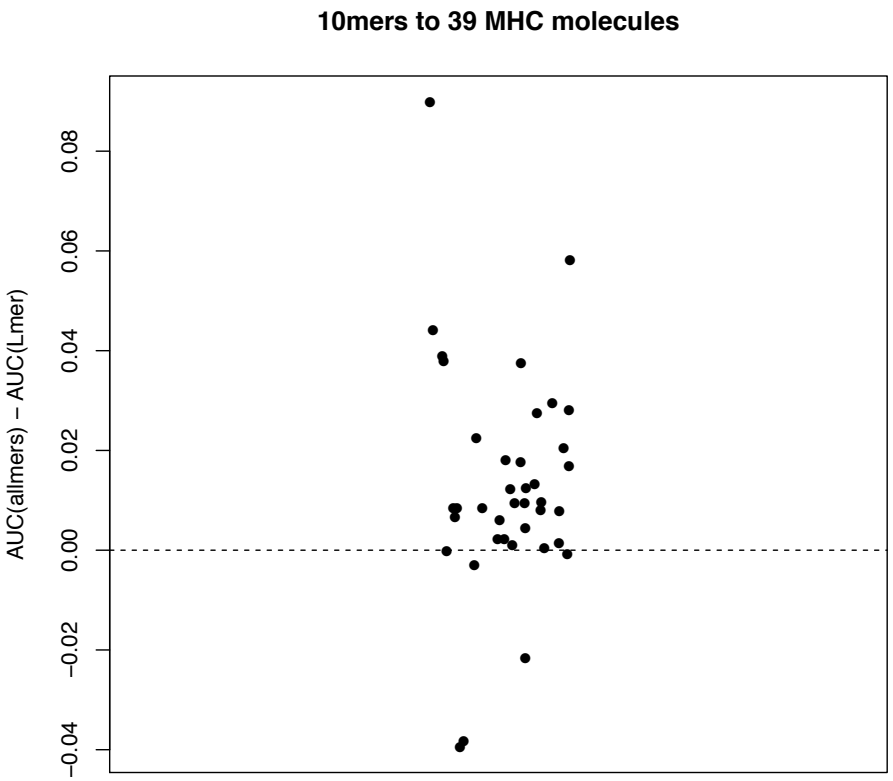


Networks trained on all-mers have higher AUC in 56/81 cases (p=0.0004)

(molecules measured with at > 20 peptides of which at least 4 are binders)

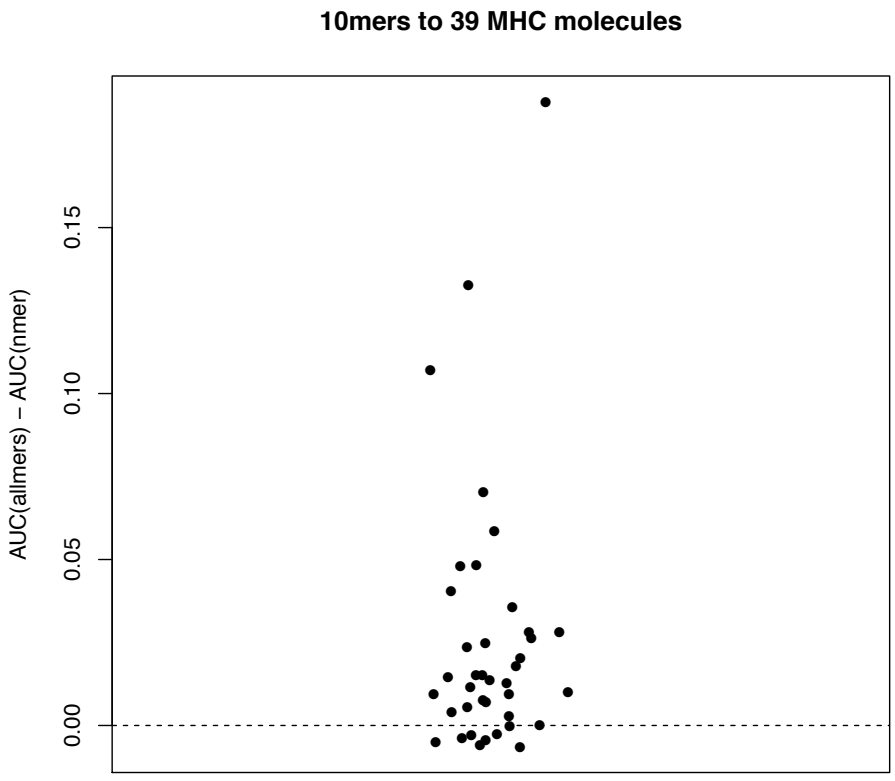
Enrichment of dataset with peptides of different length (10-mers)

all-mers networks vs. networks trained on 9mers only (L-mer approximation)



Networks trained on all-mers have higher AUC in 33/39 cases ($p=10^{-5}$)

all-mers networks vs. networks trained on 10mers only



Networks trained on all-mers have higher AUC in 31/39 cases (p=0.0002)

(molecules measured with at > 20 peptides of which at least 4 are binders)

Conclusions I

- Methods are continuously evaluated at IEDB

http://tools.iedb.org/auto_bench/mhci/

- ~ .95 AUC for class I
- ~ .85 AUC for class II

- Use rank, not affinity!

- Beware of length preference

B-cell epitope prediction

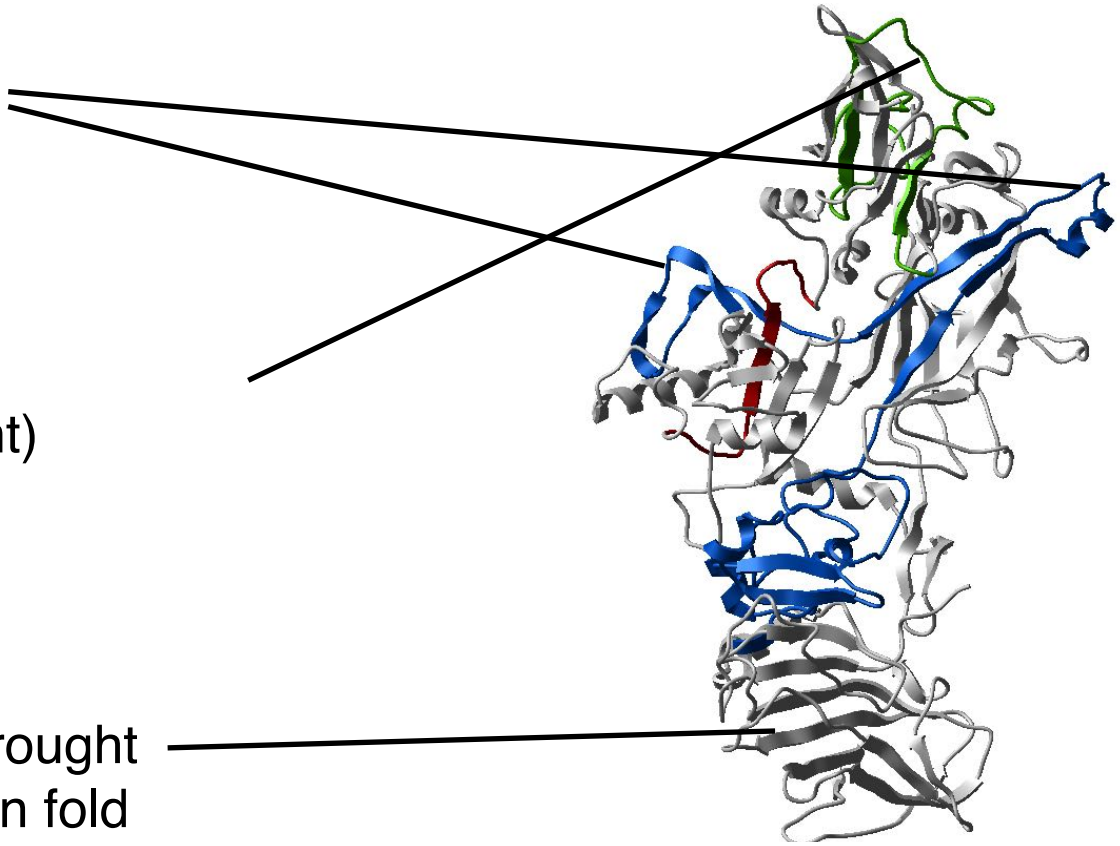
B-cell epitope classification

B-cell epitope: structural feature of a molecule or pathogen, accessible and recognizable by B-cell receptors and antibodies

Linear epitopes
One segment of the amino acid chain

Discontinuous epitope
(with linear determinant)

Discontinuous epitope
Several small segments brought into proximity by the protein fold



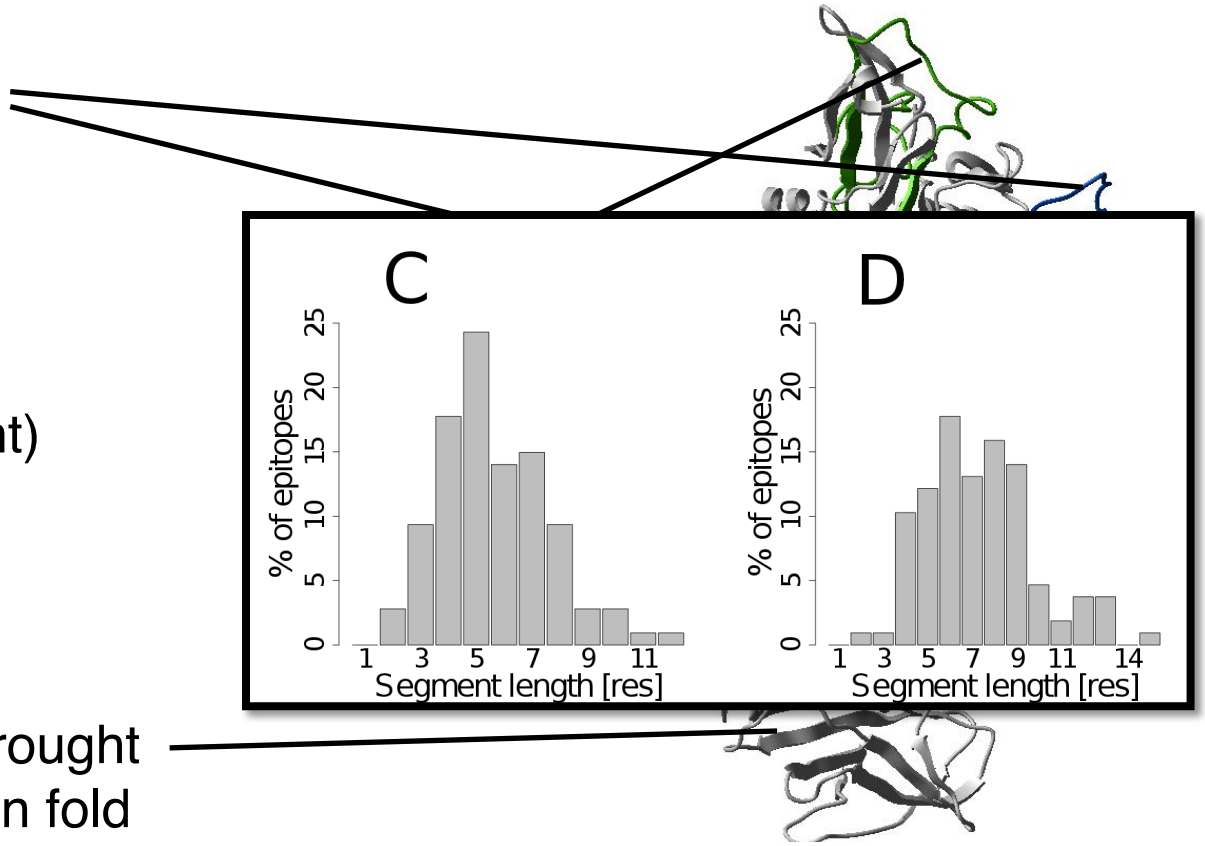
B-cell epitope classification

B-cell epitope: structural feature of a molecule or pathogen, accessible and recognizable by B-cell receptors and antibodies

Linear epitopes
One segment of the amino acid chain

Discontinuous epitope
(with linear determinant)

Discontinuous epitope
Several small segments brought into proximity by the protein fold



BepiPred

Input

```
TSQDLSVFPLASCCKDNIASTSVTLGCLVTGYLP
MSTTVTWDTGSLNKNVTTFPTTFHETYGLHSIVS
QVTASGKWAKQRFTCSVAHAESTAINKTFSACAL
NFIPPTVKLFHSSCNPVGDTHTTIQLLCLISGYV
PGDMEVIWLVDGQKATNIFPYTAPGTKEGNVTST
HSELNITQGEWVSQKTYTCQVTYQGFTFKDEARK
CSESDPRGVTSYLSPPSPL
```


Output

```
TSQDLSVFPLASCCKDNIASTSVTLGCLVTGYLP
MSTTVTWDTGSLNKNVTTFPTTFHETYGLHSIVS
QVTASGKWAKQRFTCSVAHAESTAINKTFSACAL
NFIPPTVKLFHSSCNPVGDTHTTIQLLCLISGYV
PGDMEVIWLVDGQKATNIFPYTAPGTKEGNVTST
HSELNITQGEWVSQKTYTCQVTYQGFTFKDEARK
CSESDPRGVTSYLSPPSPL
```

Parker hydrophilicity scale

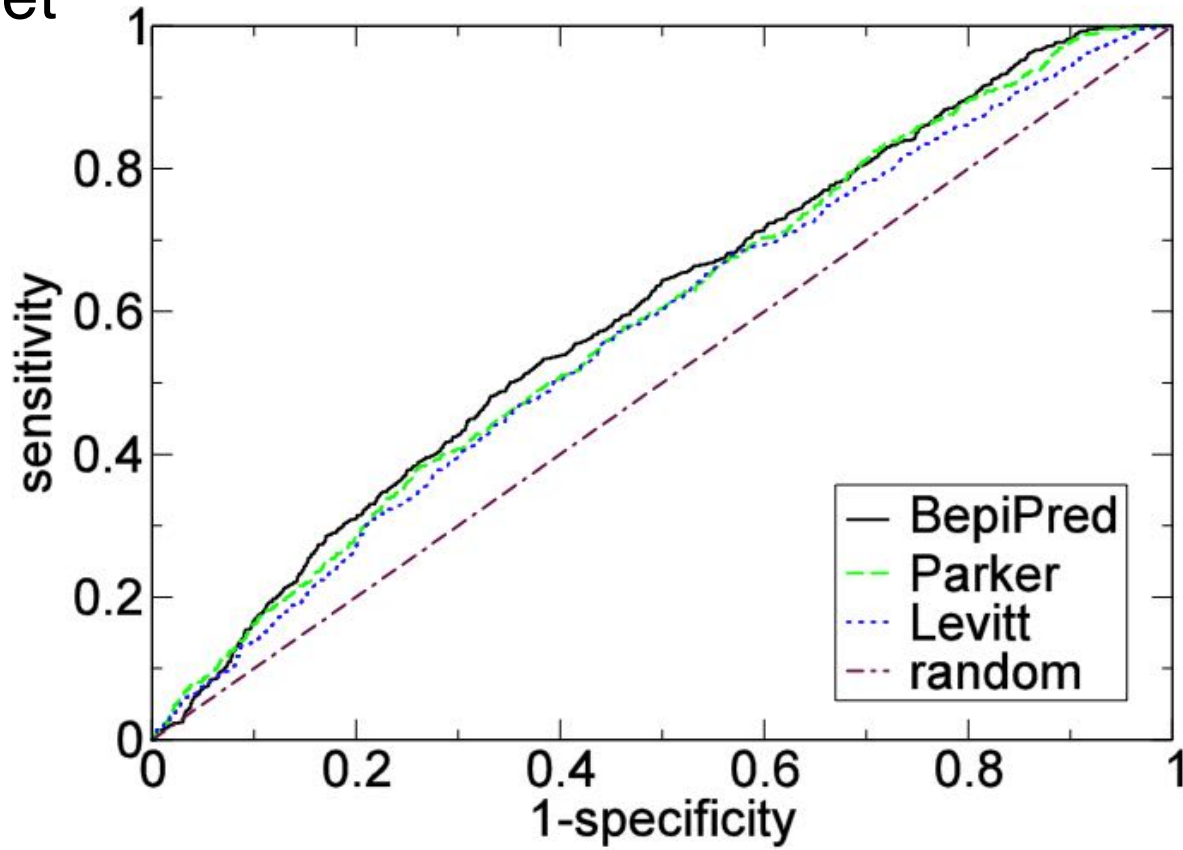
PSSM based on linear epitopes extracted from the AntiJen database

Combination of the Parker prediction scores and PSSM leads to prediction score

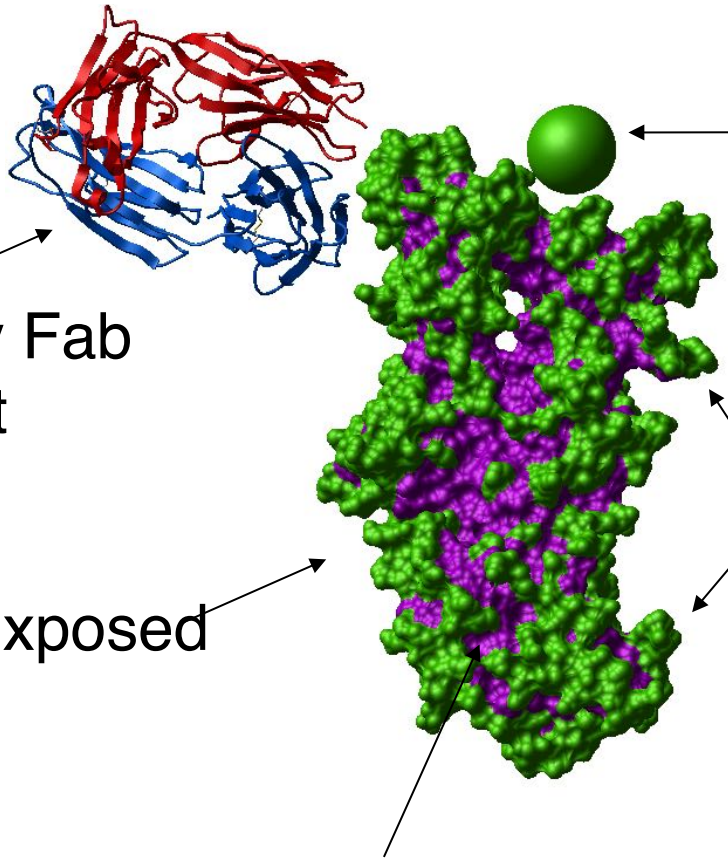
Tested on the Pellequer dataset and epitopes in the HIV Los Alamos database

BepiPred Accuracy

- Pellequer data set:
  - Levitt AUC = 0.66
  - Parker AUC = 0.65
  - BepiPred AUC = 0.68
- HIV Los Alamos data set
  - Levitt AUC = 0.57
  - Parker AUC = 0.59
  - BepiPred AUC = 0.60



What does antibodies recognize in a protein?



probe

Antibody Fab fragment

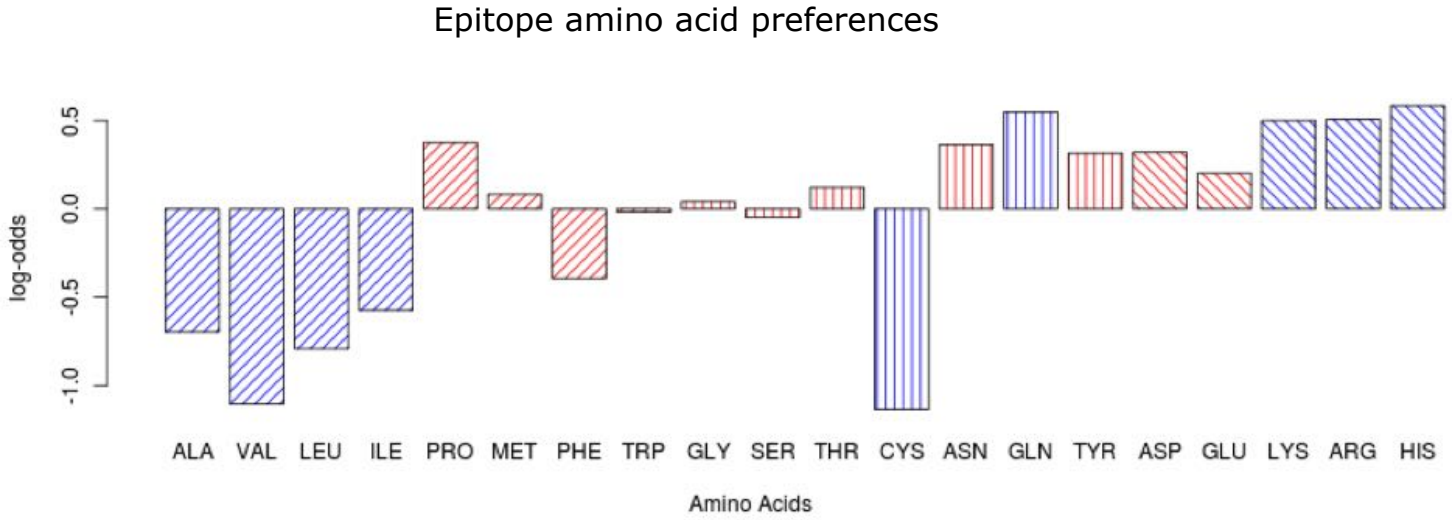
Protrusion index

Surface Exposed

Hydrophobic region

Discotope -

- Some amino acids are preferred and disliked in the epitope



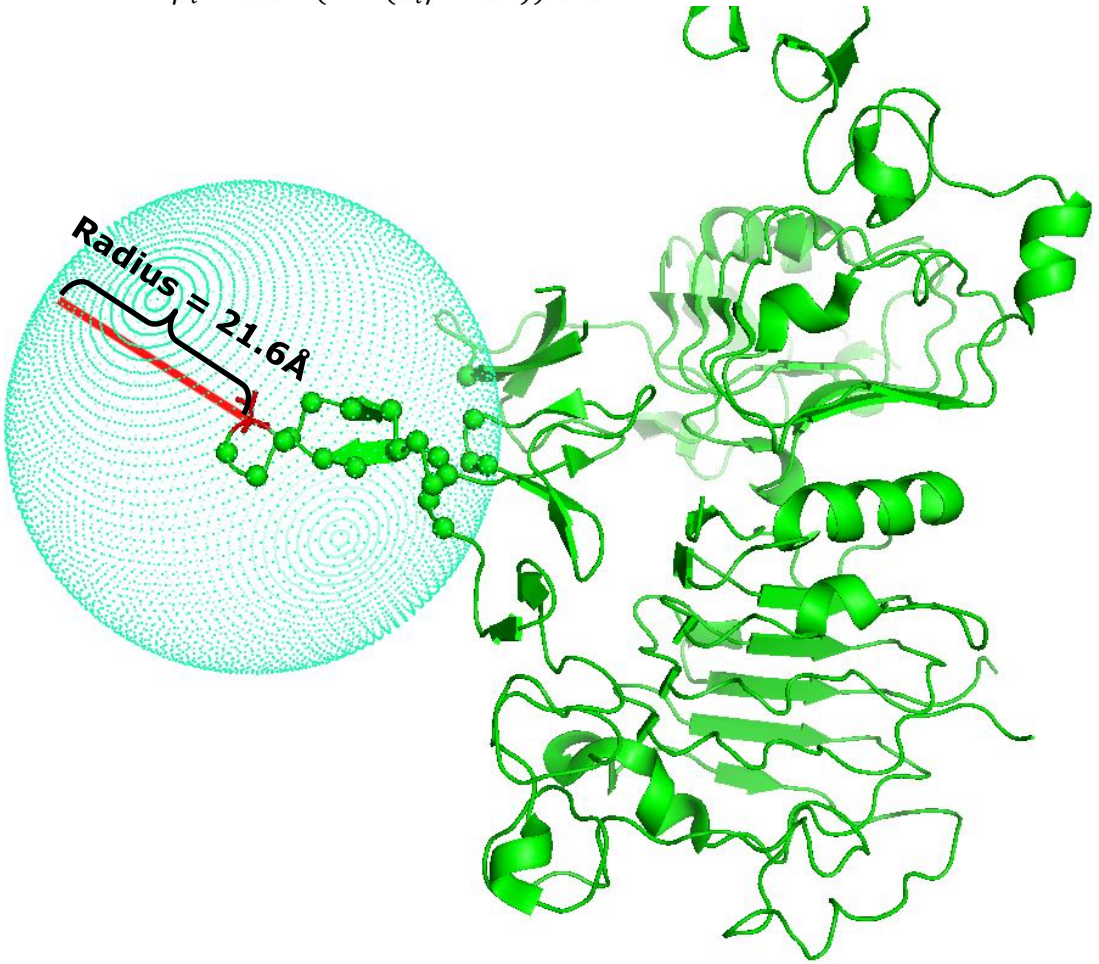
Kringelum et al, 2011

Propensity Score

$$PS(r) = \sum_i \beta_i \cdot ls(r_i)$$

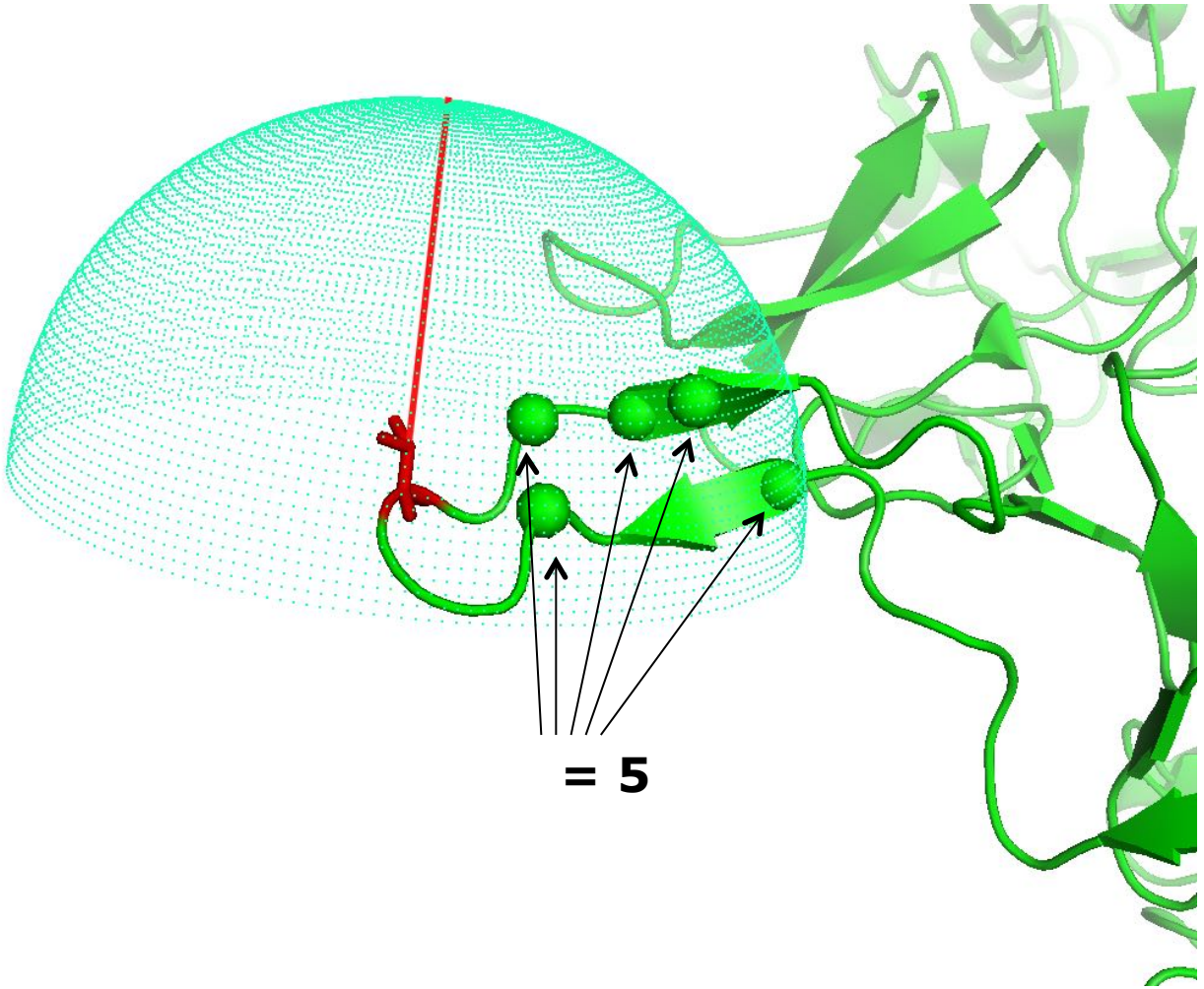
$$\beta_i = 0.8 * (1 - (d_i/21.6\text{Å})) + 0.2$$

Distance-weighted propensity score



Half-sphere exposure

Accessibility and
Protrusion



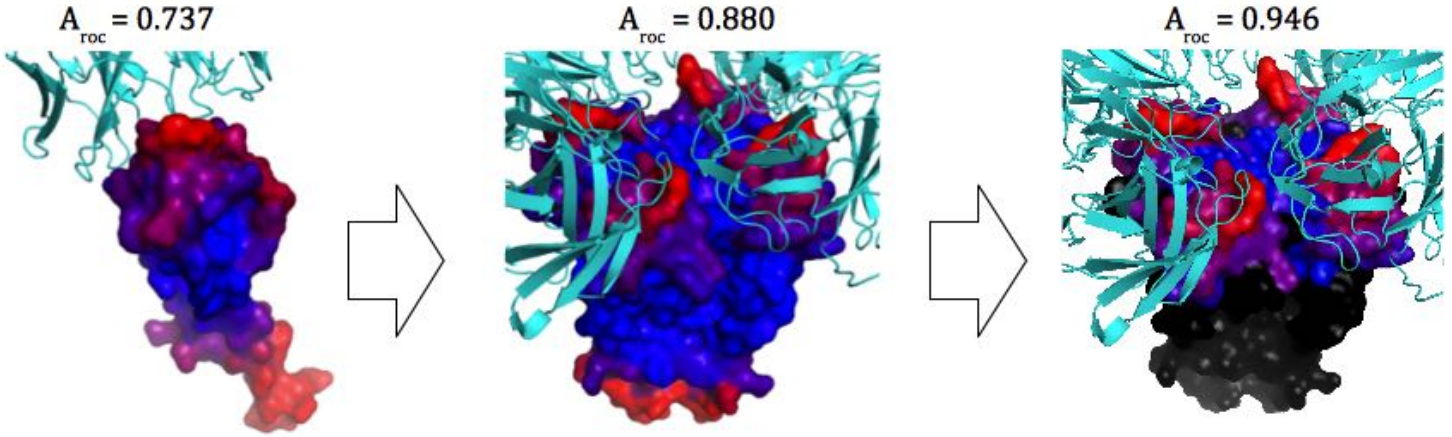
Performance and Limitations

- External Benchmark Dataset
    - 52 antigen:antibody structures
    - 33 homology groups
- Performance: 0.73 AUC

Performance and Limitations

- External Benchmark Dataset
  - 52 antigen:antibody structures
  - 33 homology groups
- Performance: 0.73 AUC

- Inclusion of biological units enhance performance



Potassium Channel

Conclusions II

- DiscoTope V2.0 outperforms similar methods
- Inclusion of surface measures does only slightly enhance predictions
- Use the entire biological unit, when possible
  - Small fragments (< 120 residues) have lower performance
- PTMs are not considered!

Current Projects

TCR Immunogenicity

Improved B-cell epitope prediction

Antibody information in epitope prediction

Combine Machine Learning with Docking

Acknowledgments

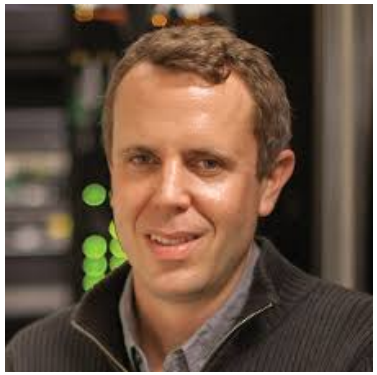
Morten Nielsen
Ole Lund

T-cells
Vanessa Jurtz
Thomas Trolle
Edita Karosiene

B-cells
Martin Jespersen
Kamilla Jensen

Bjørn Peters

Massimo Andreatta



Thanks!