Machine learning models for antigen immunogenicity and T-cell recognition

Barbara Bravi

Department of Mathematics, Imperial College London (UK)

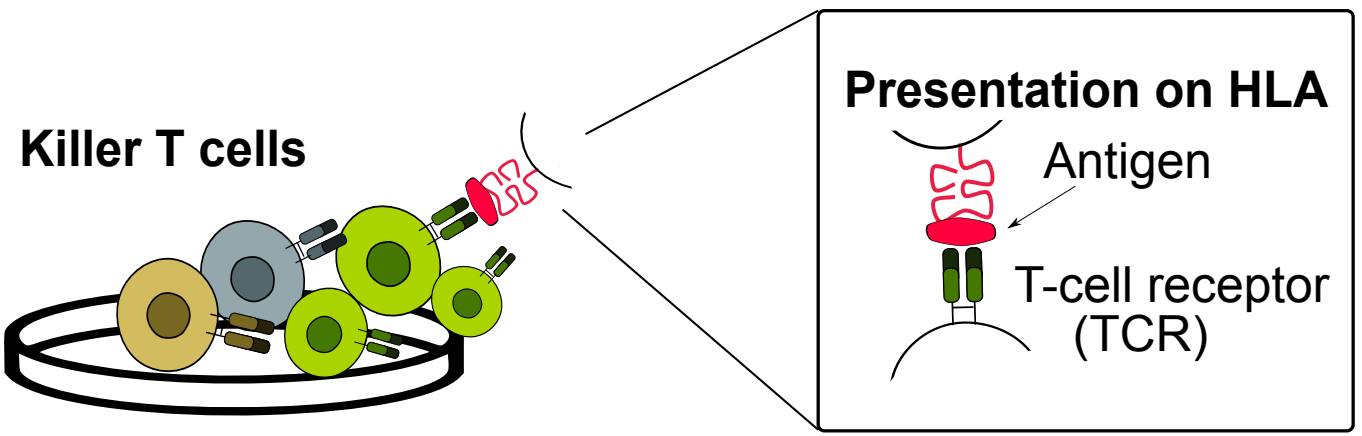
Imperial College London

Joint work with: A. Di Gioacchino, J. Fernandez-De-Cossio-Diaz, S. Cocco, R. Monasson, T. Mora, A.M. Walczak (ENS Paris)

BEvAS, EPFL 17/04/2023

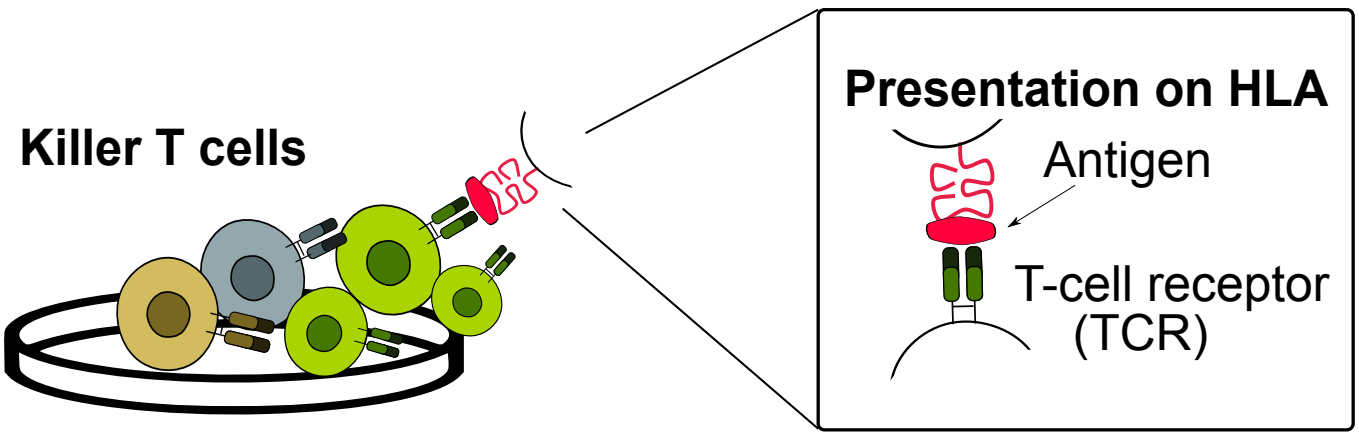
Objective:

Machine learning approach that is able to extract biologically interpretable features on antigen immunogenicity & T-cell epitope specificity



Objective:

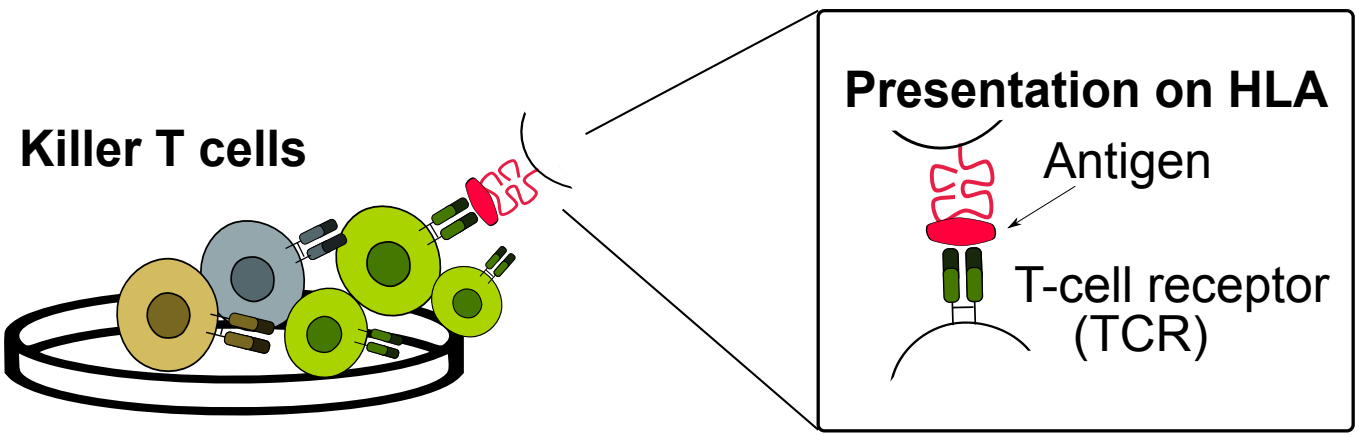
Machine learning approach that is able to extract biologically interpretable features on antigen immunogenicity & T-cell epitope specificity



Only a fraction of HLA-presented antigens are immunogenic (promote a T cell response).
Immunogenicity prediction: key in neoantigen discovery, low success rate (Wells et al. 2020)

Objective:

Machine learning approach that is able to extract biologically interpretable features on antigen immunogenicity & T-cell epitope specificity



Only a fraction of HLA-presented antigens are immunogenic (promote a T cell response). Immunogenicity prediction: key in neoantigen discovery, low success rate (Wells et al. 2020)

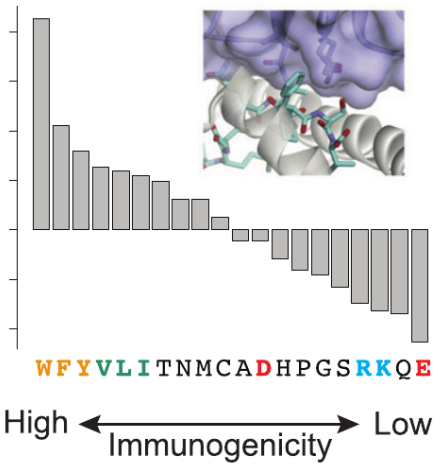
pMHC epitope elicits the response only of specific small subsets of TCRs, recent advances in prediction but insight into molecular properties is still challenging
(Gielis et al. 2019, Montemurro et al. 2021, Weber et al. 2021 and others)

Enrichment in distinctive patterns

Antigen immunogenicity and epitope-specificity of T cell receptors result from physico-chemical constraints on sequence composition

Enrichment in aromatic, hydrophobic residues in immunogenic peptides

From Schmidt et al. 2021



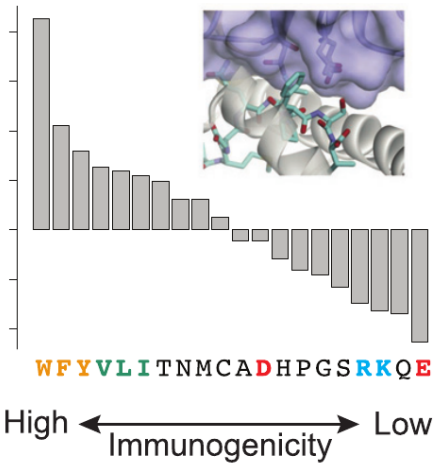
see also: Calis et al. 2013, Chowell et al. 2015

Enrichment in distinctive patterns

Antigen immunogenicity and epitope-specificity of T cell receptors result from physico-chemical constraints on sequence composition

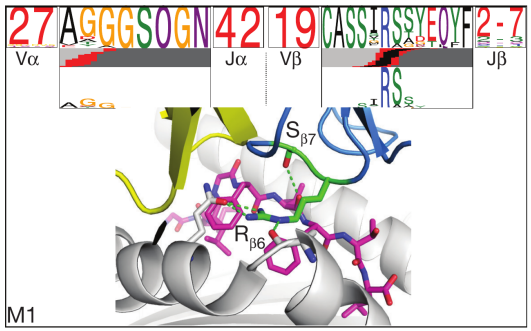
Enrichment in aromatic, hydrophobic residues in immunogenic peptides

From Schmidt et al. 2021



see also: Calis et al. 2013, Chowell et al. 2015

Convergence in sequence motifs in epitope-specific receptors



From Dash et al. 2017

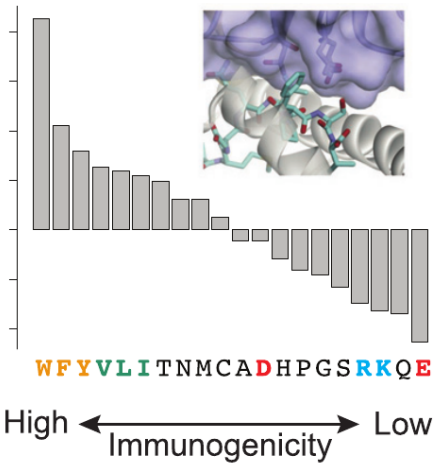
see also: Glanville et al. 2017, Meysman et al. 2019 Pogorelyy et al. 2019 , Mayer-Blackwell et al. 2021 etc

Enrichment in distinctive patterns

Antigen immunogenicity and epitope-specificity of T cell receptors result from physico-chemical constraints on sequence composition

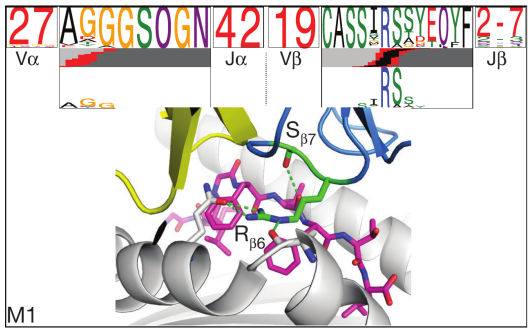
Enrichment in aromatic, hydrophobic residues in immunogenic peptides

From Schmidt et al. 2021



see also: Calis et al. 2013, Chowell et al. 2015

Convergence in sequence motifs in epitope-specific receptors



From Dash et al. 2017

see also: Glanville et al. 2017, Meysman et al. 2019
Pogorelyy et al. 2019 , Mayer-Blackwell et al. 2021 etc

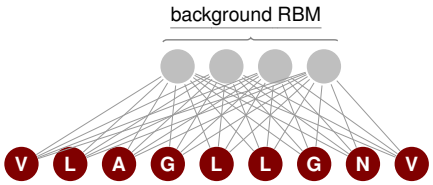
How to disentangle pattern enrichment from baseline constraints?
(ensuring e.g. in antigens high binding affinity to HLA)

Our approach:

Machine learning approach known as ‘transfer learning’ within the model known as Restricted Boltzmann Machines[1]

Based on the pre-print: B. Bravi, A. Di Gioacchino, J. Fernandez-de-Cossio-Diaz, A.M. Walczak, T. Mora, S. Cocco, R. Monasson, *Learning the differences: a transfer-learning approach to predict antigen immunogenicity and T-cell receptor specificity*, Biorxiv 2022.12.06.519259v1 (2022)

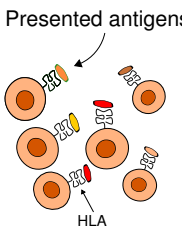
[1] (Smolensky 1986, Hinton 2002; Biophysical modelling: Tubiana, Cocco and Monasson 2019, Shimagaki and Weigt 2019)



HLA-A*02:01 presented antigens

VLAGLLGNV
GILGFVFTL
FLCLFLLPSL
SLQQELAHM
ALYGVWPLLL
ALAESIRPL

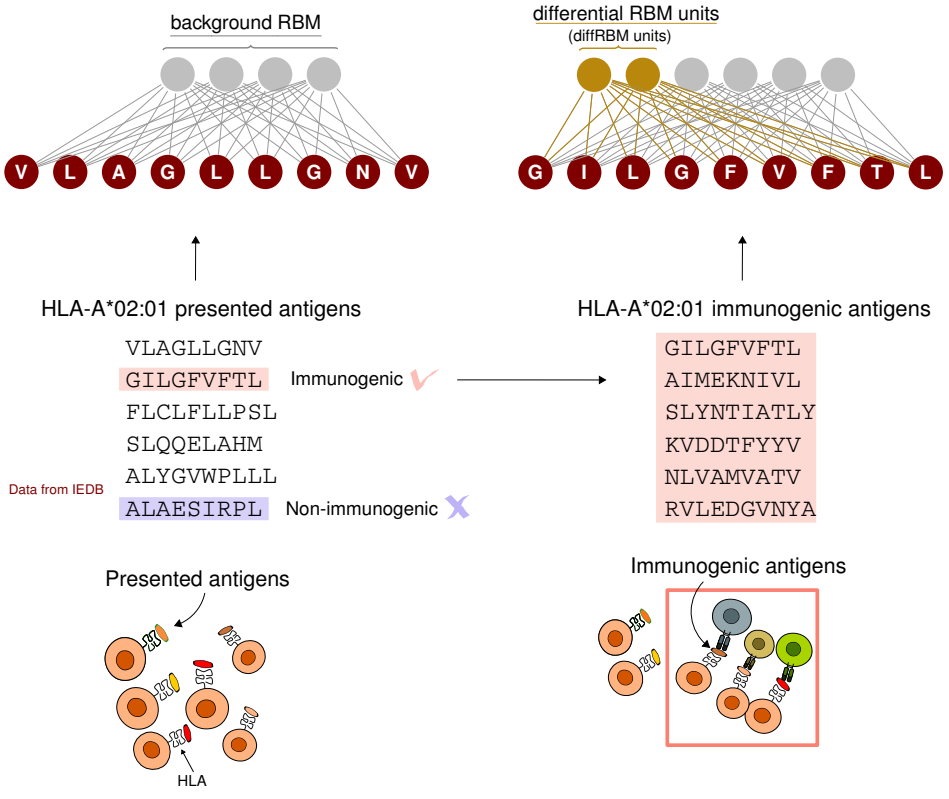
Data from IEDB

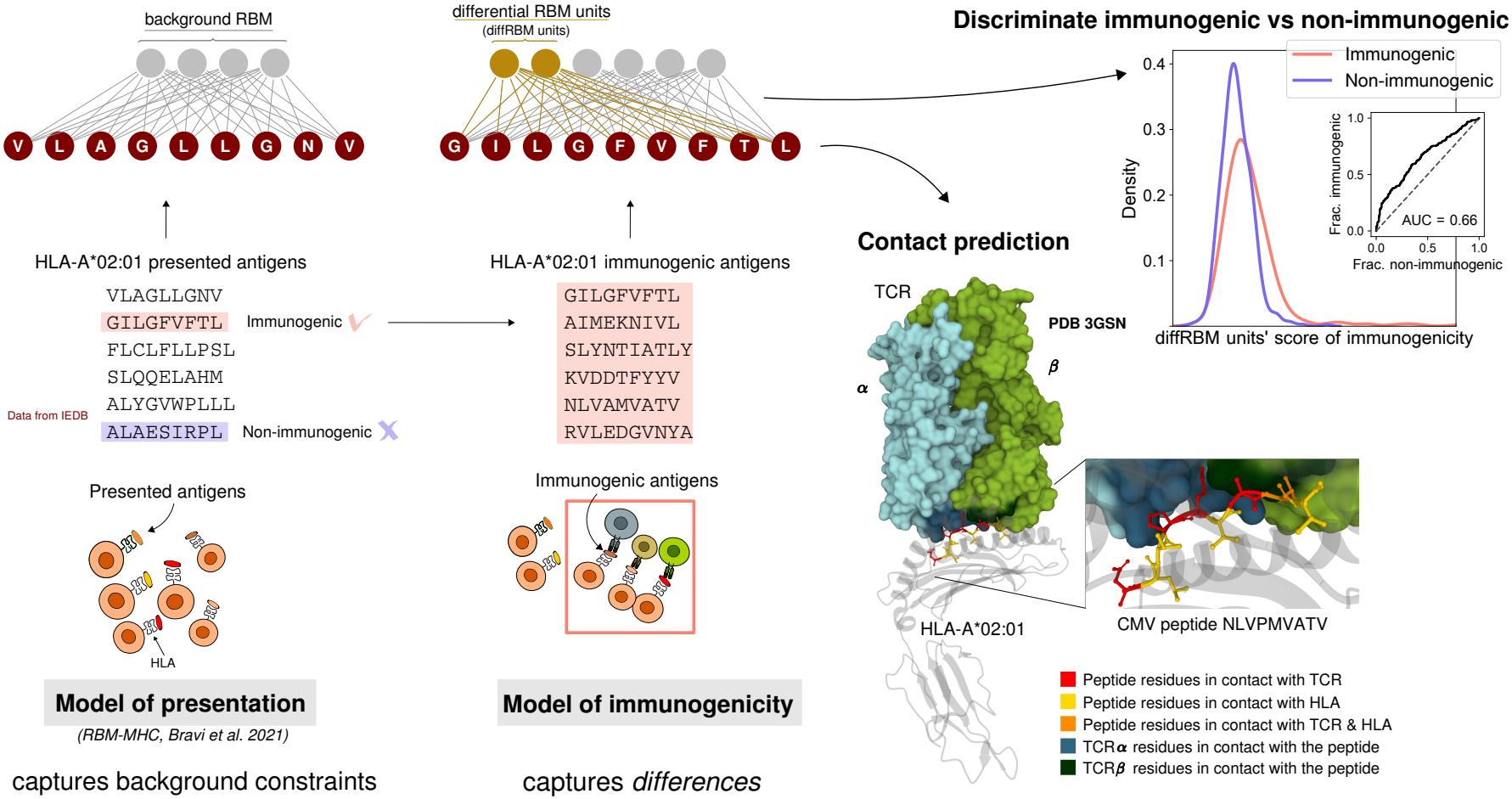


Model of presentation

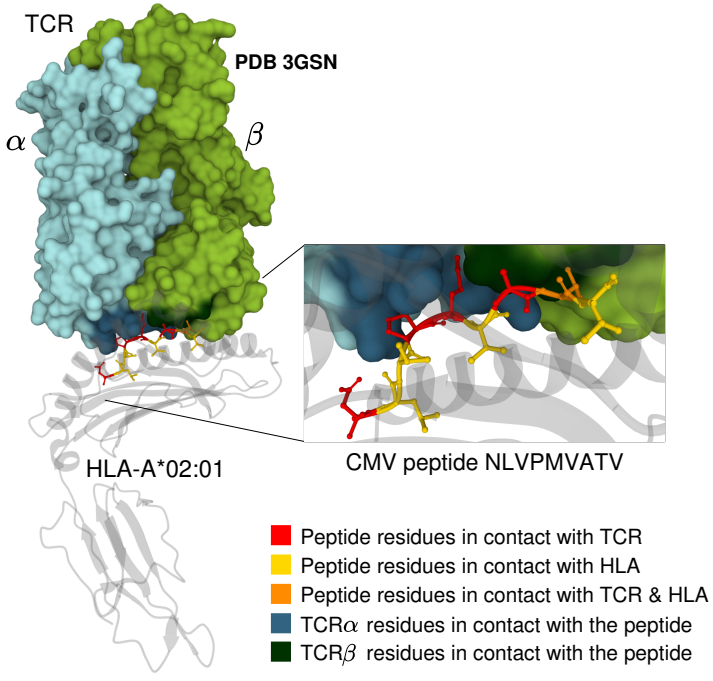
(RBM-MHC, Bravi et al. 2021)

captures background constraints
(binding affinity to HLA)

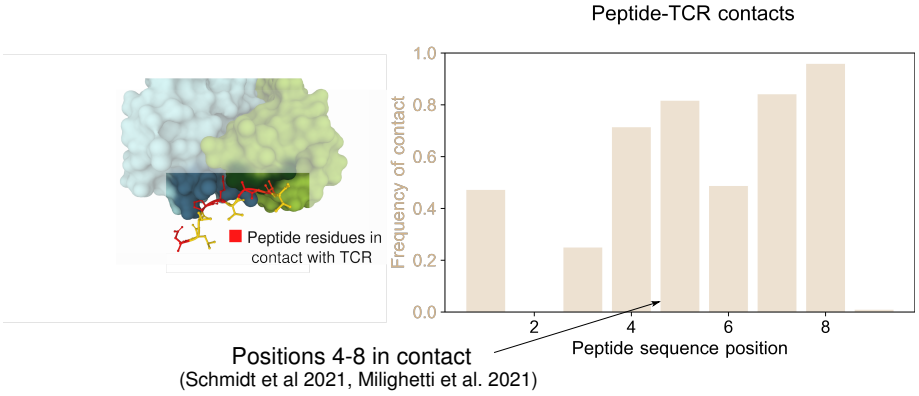

background RBM
differential RBM units
(diffRBM units)
V L A G L L G N V
G I L G F V F T L
HLA-A*02:01 presented antigens
HLA-A*02:01 immunogenic antigens
VLAGLLGNV
GILGFVFTL
Immunogenic
FLCLFLLPSL
SLQQELAHM
ALYGVWPLLL
ALAESIRPL
Non-immunogenic
Data from IEDB
GILGFVFTL
AIMEKNIVL
SLYNTIATLY
KVDDTFYYV
NLVAMVATV
RVLEDGVNYA
Presented antigens
HLA
Immunogenic antigens
Model of presentation
(RBM-MHC, Bravi et al. 2021)
Model of immunogenicity
captures background constraints
(binding affinity to HLA)
captures differences
(patterns of immunogenicity)

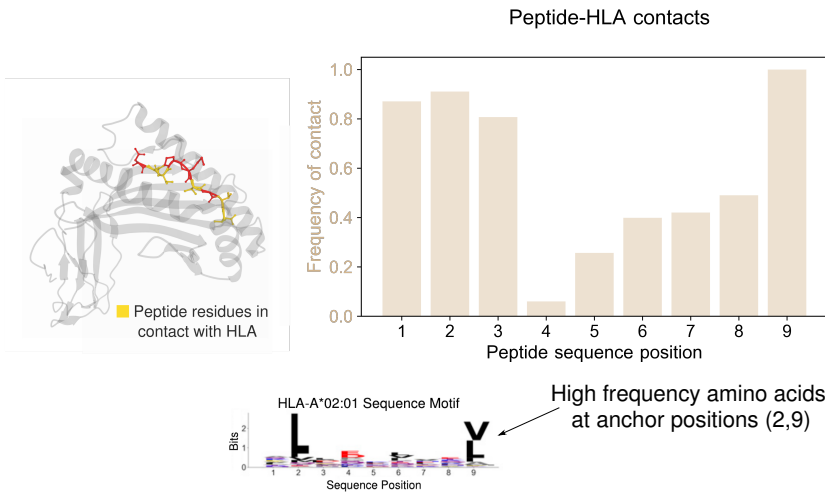

background RBM
V L A G L L G N V
differential RBM units
(diffRBM units)
G I L G F V F T L
HLA-A*02:01 presented antigens
VLAGLLGNV
GILGFVFTL
Immunogenic
FLCLFLLPSL
SLQQELAHM
ALYGVWPLLL
Data from IEDB
ALAESIRPL
Non-immunogenic
HLA-A*02:01 immunogenic antigens
GILGFVFTL
AIMEKNIVL
SLYNTIATLY
KVDDTFYYV
NLVAMVATV
RVLEDGVNYA
Presented antigens
HLA
Immunogenic antigens
Model of presentation
(RBM-MHC, Bravi et al. 2021)
Model of immunogenicity
captures background constraints
(binding affinity to HLA)
captures differences
(patterns of immunogenicity)
Discriminate immunogenic vs non-immunogenic
Immunogenic
Non-immunogenic
Density
0.4
0.3
0.2
0.1
Frac. immunogenic
1.0
0.5
0.0
AUC = 0.66
0.0
0.5
1.0
Frac. non-immunogenic
diffRBM units' score of immunogenicity
Contact prediction
TCR
PDB 3GSN
α
β
HLA-A*02:01
CMV peptide NLVPMVATV
Peptide residues in contact with TCR
Peptide residues in contact with HLA
Peptide residues in contact with TCR & HLA
TCRα residues in contact with the peptide
TCRβ residues in contact with the peptide

Contact positions in resolved structures:

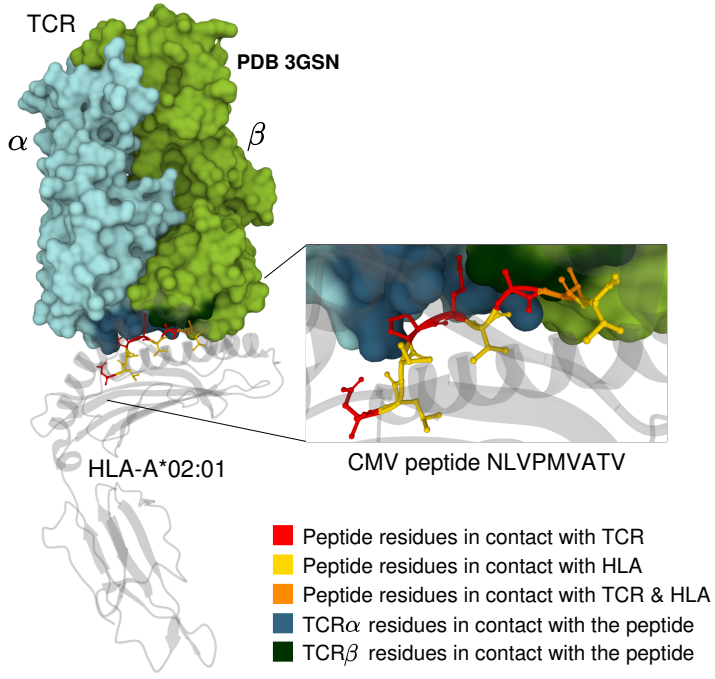


HLA-A*02:01-specific peptides

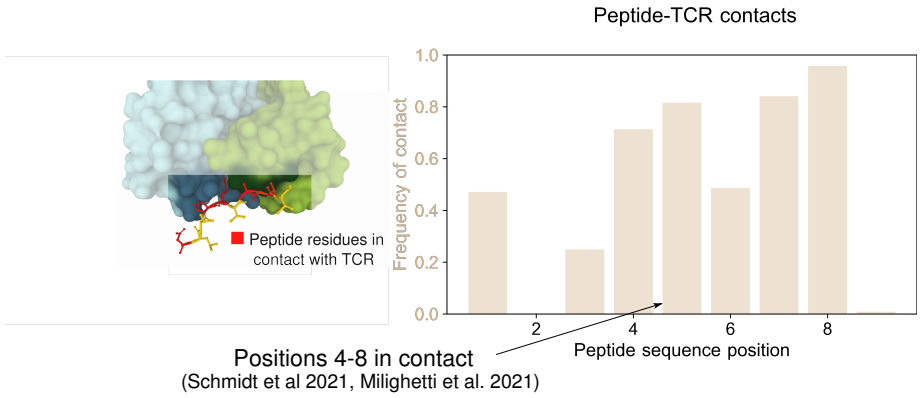




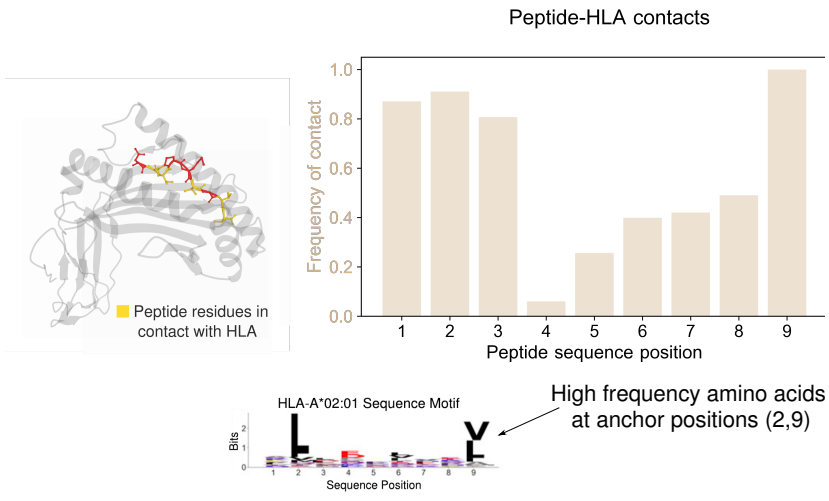
Contact positions in resolved structures:



HLA-A*02:01-specific peptides



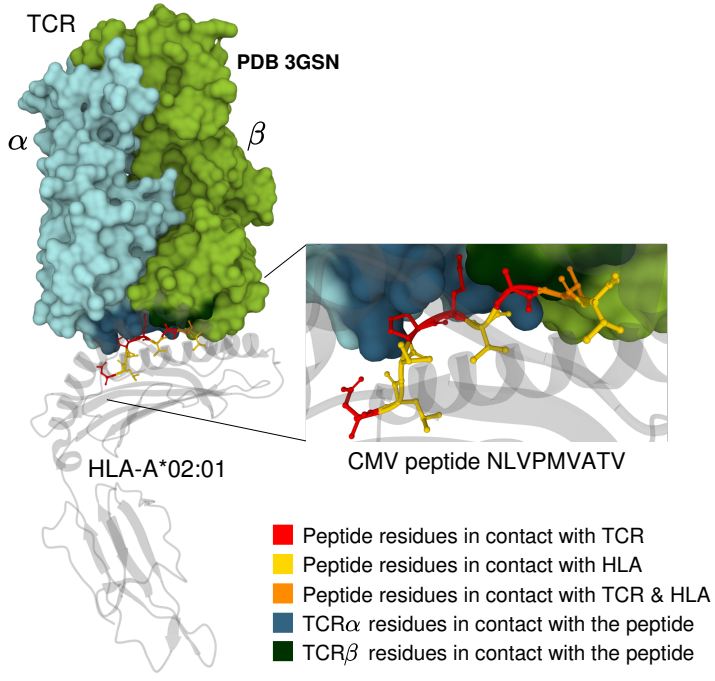
Positions 4-8 in contact
(Schmidt et al 2021, Milighetti et al. 2021)



High frequency amino acids at anchor positions (2,9)

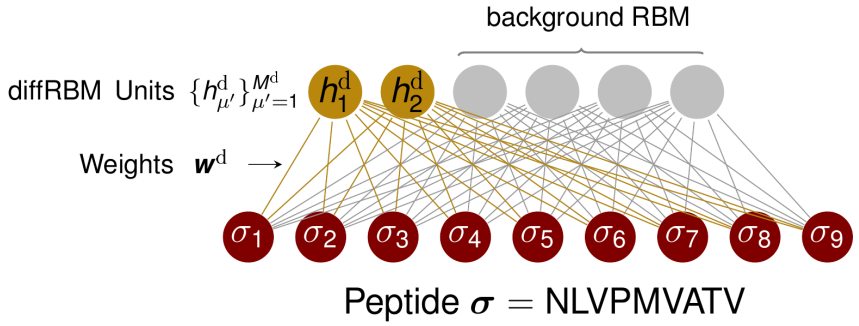
Differences in statistics of immunogenic peptides should reflect contacts

Contact positions in resolved structures:



Model prediction:

diffRBM architecture

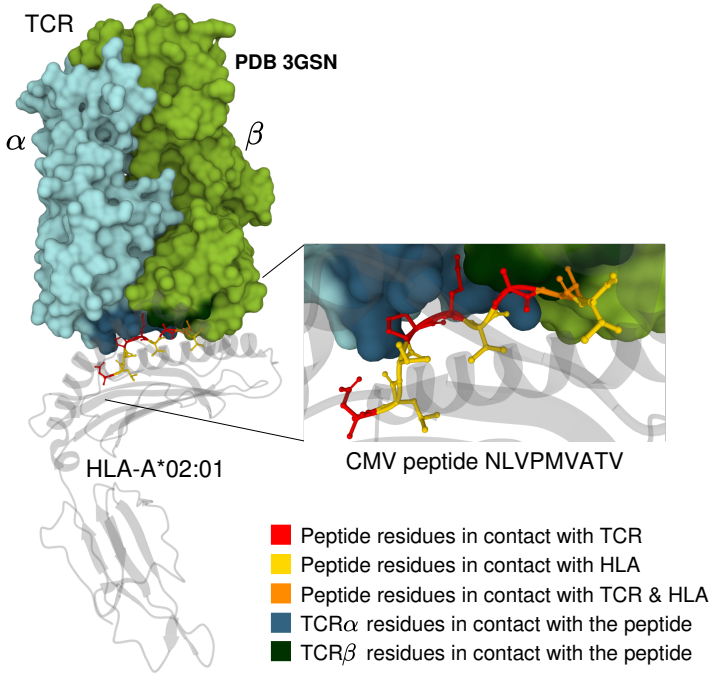


Single-site importance factors

$$T_i(\sigma_i) = \underbrace{g_i^{\mathrm{d}}(\sigma_i)}_{\text{related to amino acid frequency difference between immunogenic and presented}} + \sum_{\mu'=1}^{M^{\mathrm{d}}} \underbrace{w_{i\mu'}^{\mathrm{d}}(\sigma_i)\langle h_{\mu'}|\boldsymbol{\sigma}\rangle}_{\text{captures correlations between positions}}$$

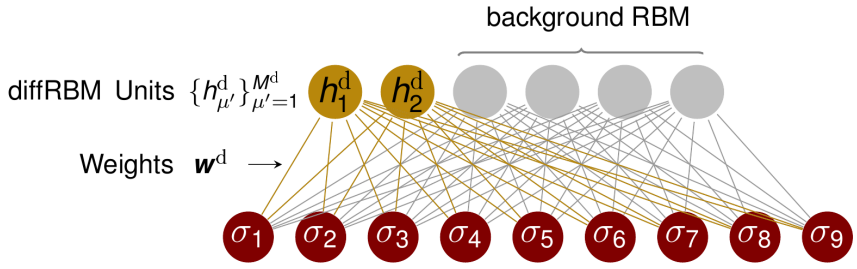
$\langle h_{\mu'}|\boldsymbol{\sigma}\rangle$: from $P(h_{\mu'}|I_{\mu'}(\boldsymbol{\sigma}))$, where $I_{\mu'}(\boldsymbol{\sigma}) = \sum_i w_{i\mu'}^{\mathrm{d}}(\sigma_i)$

Contact positions in resolved structures:



Model prediction:

diffRBM architecture



Peptide $\boldsymbol{\sigma}$ = NLVPMVATV

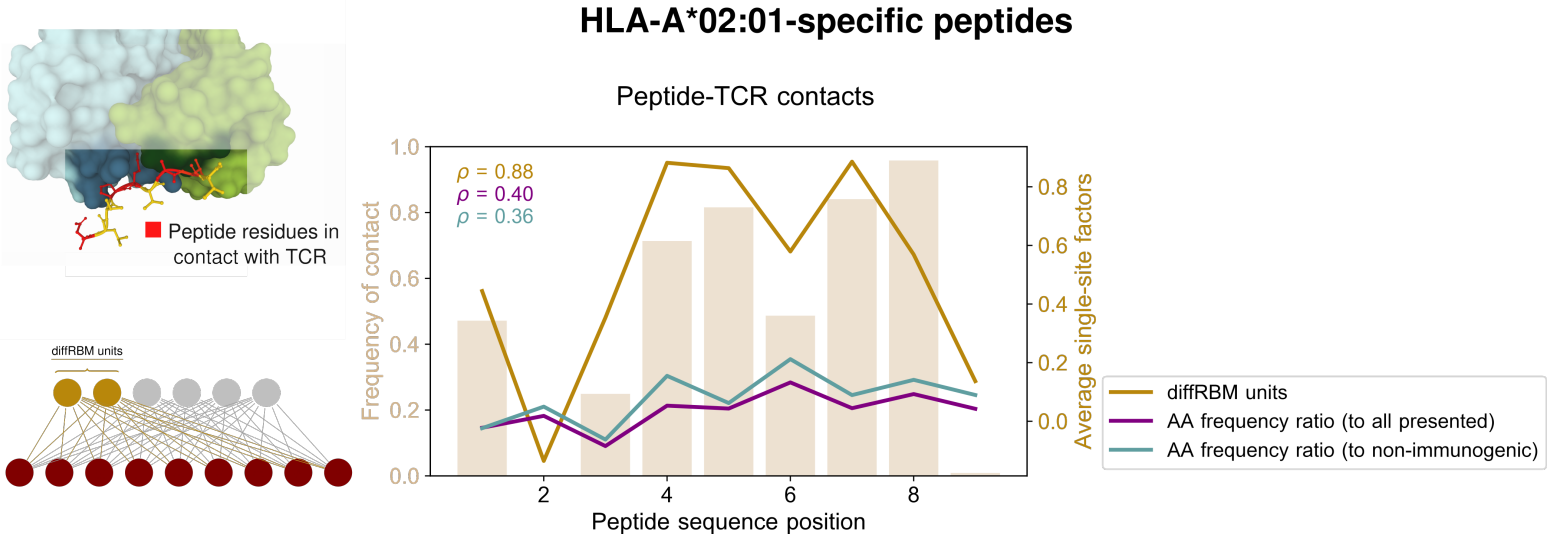
Single-site importance factors

$$T_i(\sigma_i) = \underbrace{g^{\mathrm{d}}_i(\sigma_i)}_{\text{related to amino acid frequency difference between immunogenic and presented}} + \sum_{\mu'=1}^{M^{\mathrm{d}}} \underbrace{w^{\mathrm{d}}_{i\mu'}(\sigma_i)\langle h_{\mu'}|\boldsymbol{\sigma}\rangle}_{\text{captures correlations between positions}}$$

$\langle h_{\mu'}|\boldsymbol{\sigma}\rangle$: from $P(h_{\mu'}|I_{\mu'}(\boldsymbol{\sigma}))$, where $I_{\mu'}(\boldsymbol{\sigma}) = \sum_i w^{\mathrm{d}}_{i\mu'}(\sigma_i)$

We hypothesize that sites at high $T_i(\sigma_i)$ are potential contacts

Structural interpretation

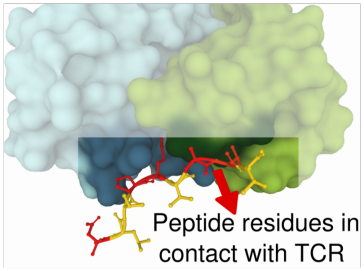


diffRBM identifies positions 4-8 as the most relevant to immunogenicity without restricting a priori the input sequences to a subset of positions

Comparison: independent-site models
based purely amino acid (AA) frequency (see e.g. IEDB tool)

Contact prediction (peptide-TCR)

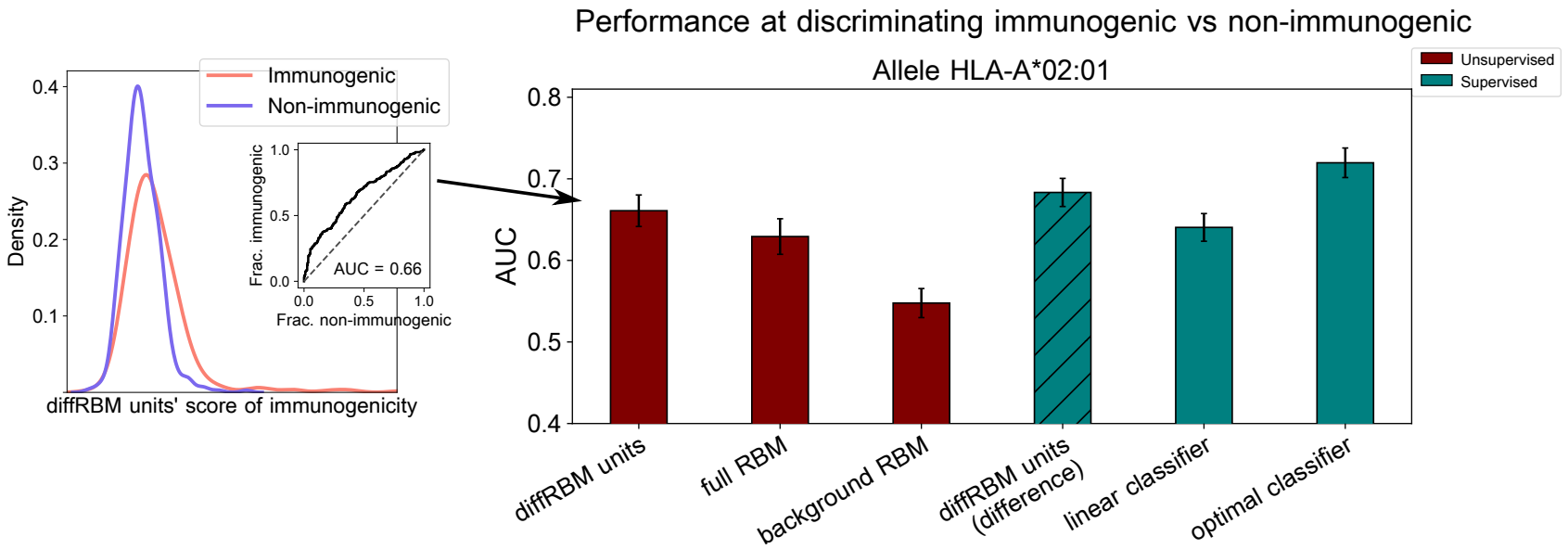
Top ranking positions by $T_i(\sigma_i) \rightarrow$ putative contacts



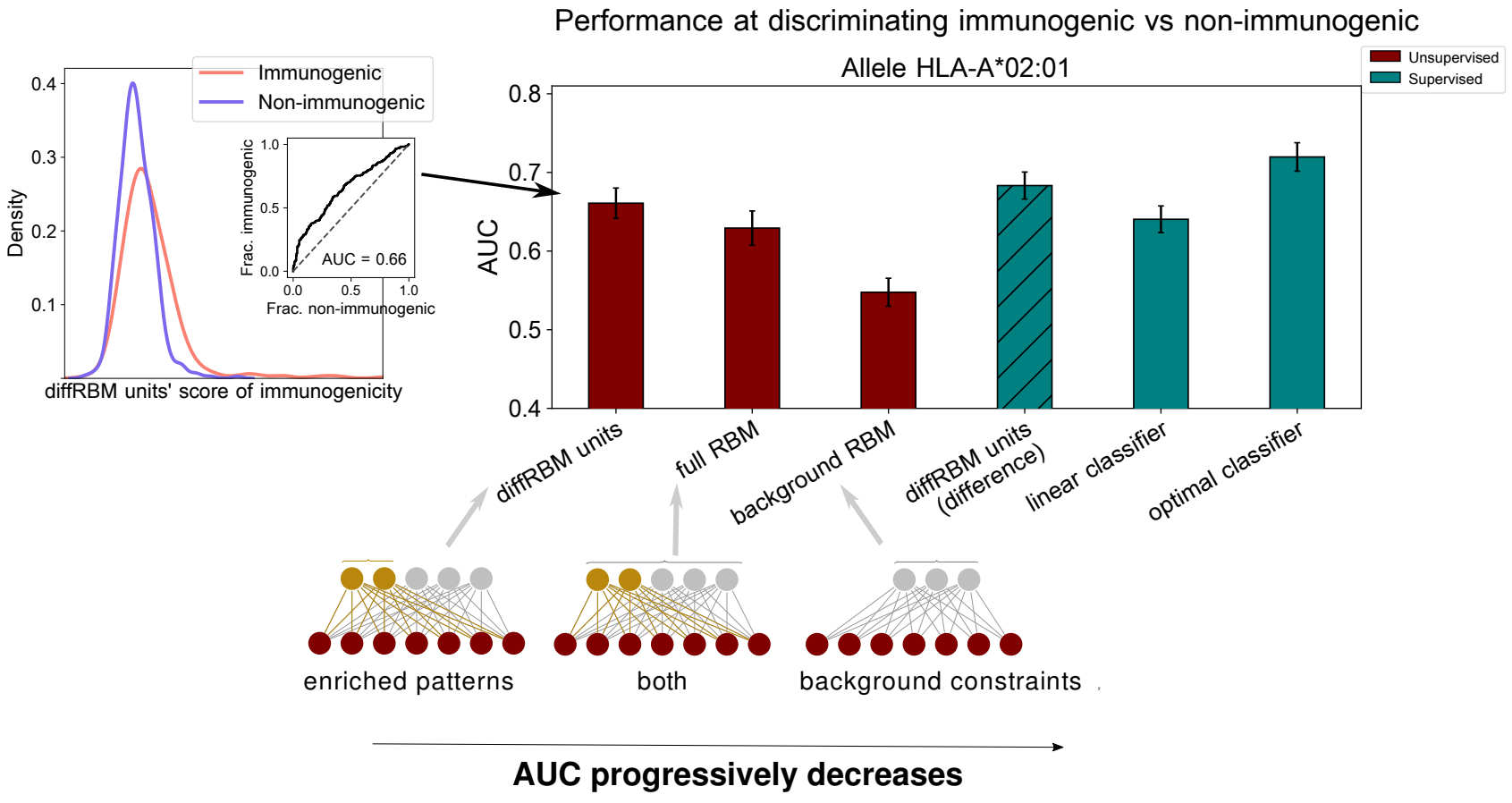
We test the prediction by the Positive Predictive Value (PPV):
fraction of ranked positions corresponding to true contacts

PPV averaged over structures for 3 HLA-I (HLA-A*02:01, HLA-B*35:01, HLA-B*07:02 - in total 46 structures)

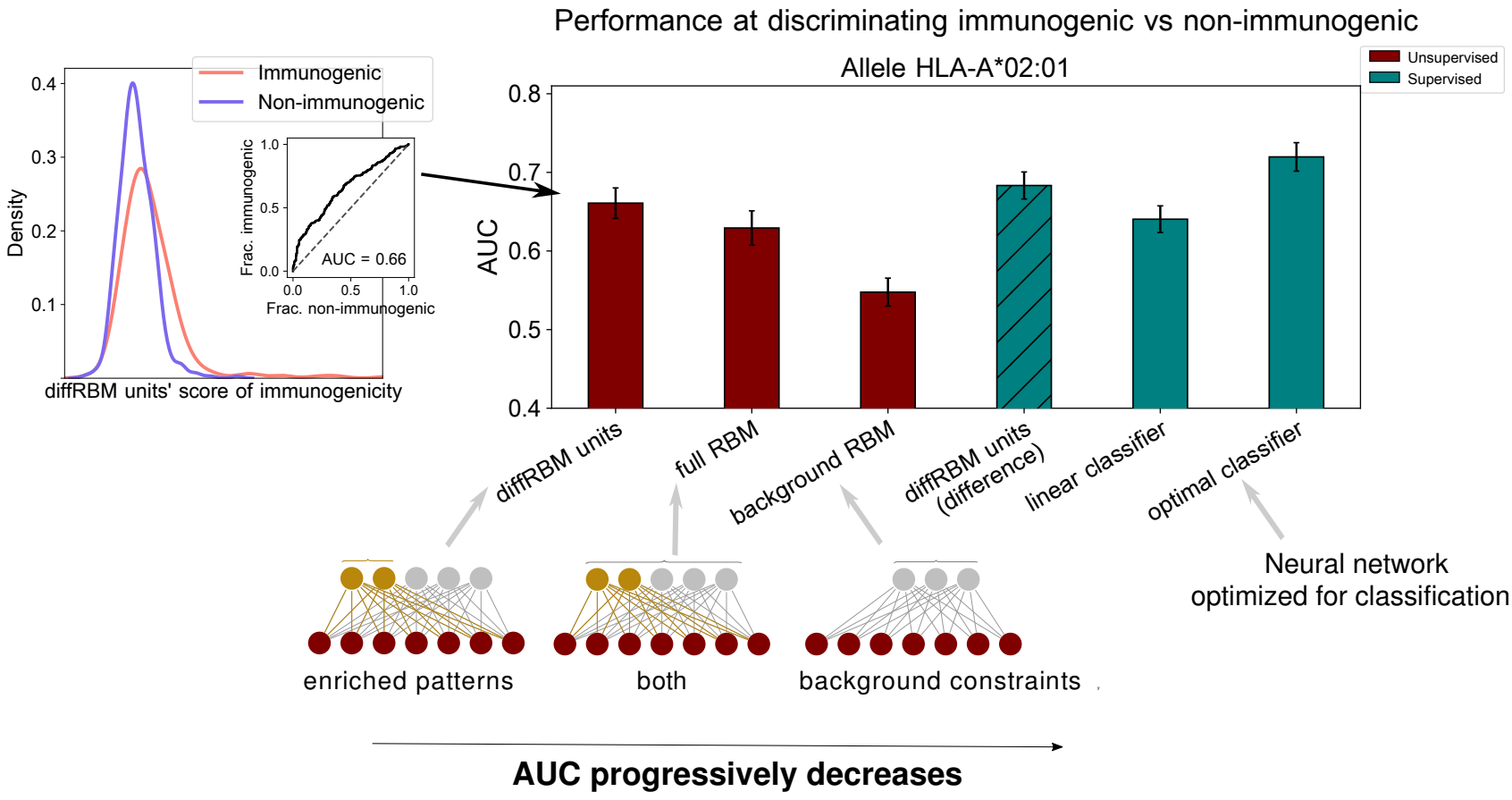
Classifying immunogenic vs non-immunogenic



Classifying immunogenic vs non-immunogenic

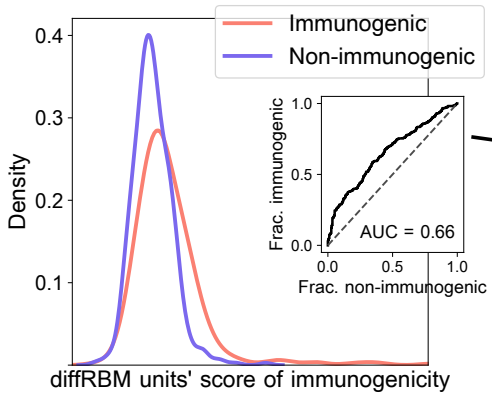


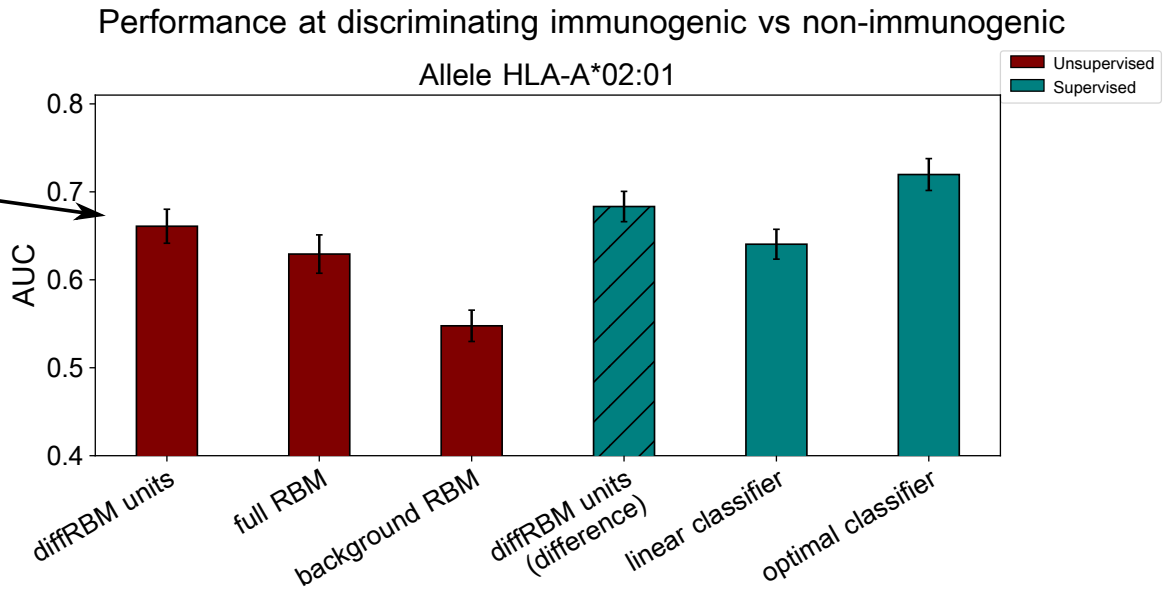
Classifying immunogenic vs non-immunogenic



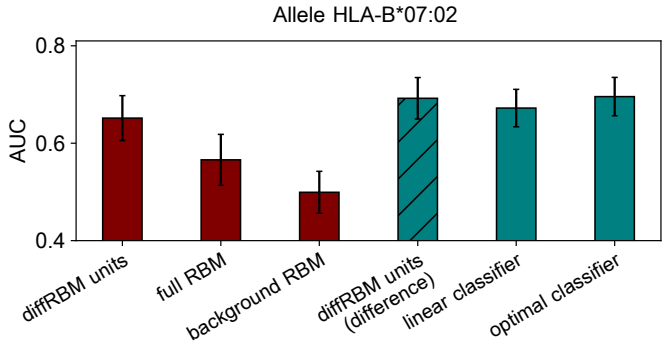
For efficient classification: score of diffRBM (immunogenic) - diffRBM (non-immunogenic)

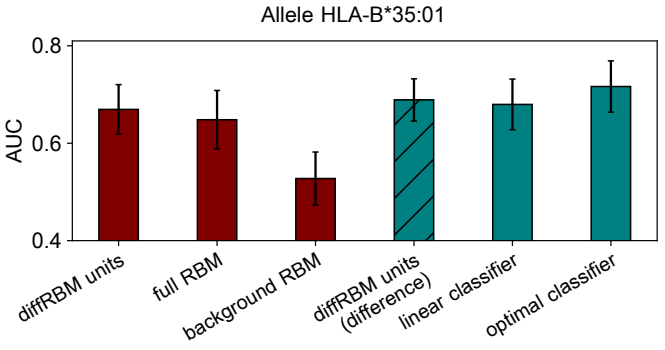
Classifying immunogenic vs non-immunogenic





Consistent trend across HLA-specific models

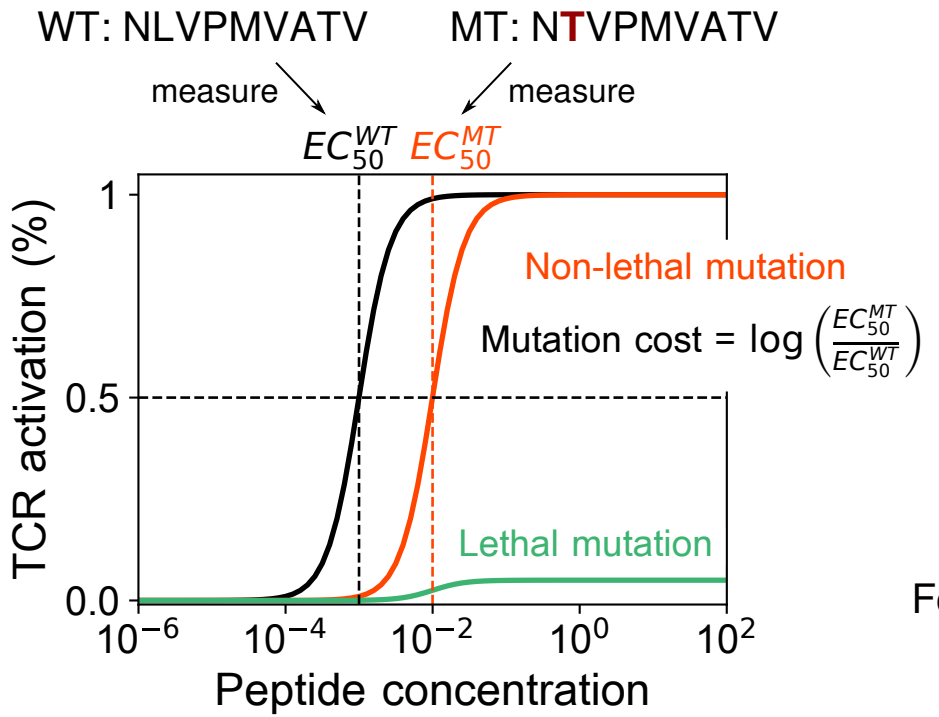




Comparison to TCR avidity assays

Data on TCR cross-reactivity to NLVPMVATV mutants

From Łuksza et al. Nature 2022

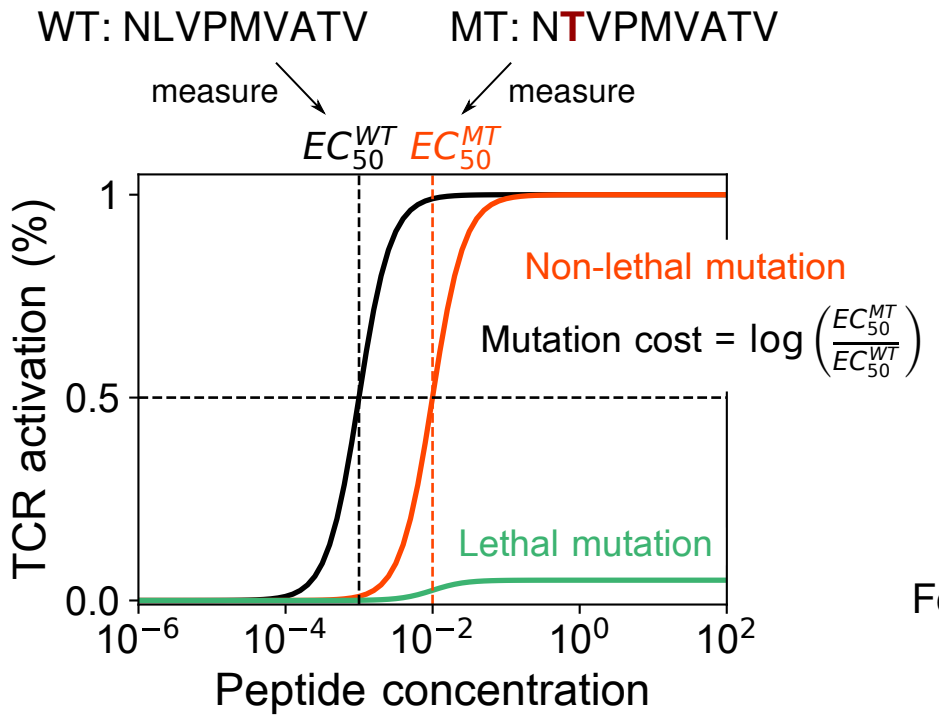


For 3 TCRs

Comparison to TCR avidity assays

Data on TCR cross-reactivity to NLVPMVATV mutants

From Łuksza et al. Nature 2022



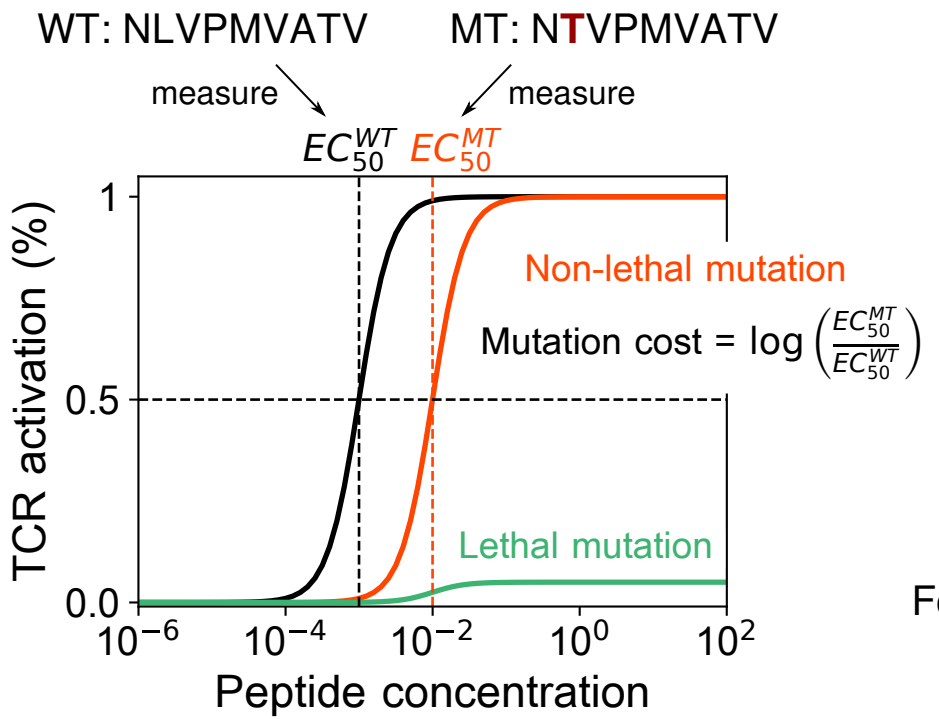
For 3 TCRs

Positive mutation costs $\rightarrow$ loss in TCR response & decrease in antigen immunogenicity

Comparison to TCR avidity assays

Data on TCR cross-reactivity to NLVPMVATV mutants

From Łuksza et al. Nature 2022



For 3 TCRs

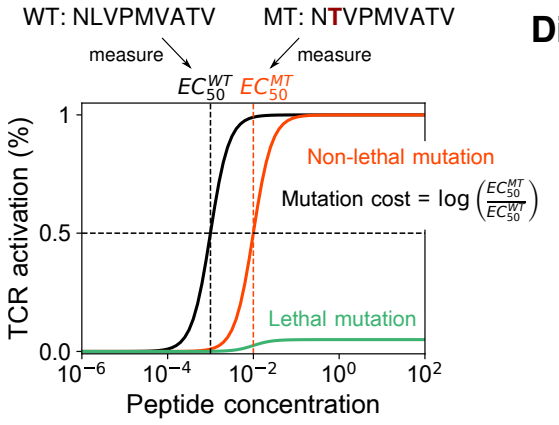
Positive mutation costs $\rightarrow$ loss in TCR response & decrease in antigen immunogenicity

We distinguish 2 groups of mutations: lethal and non-lethal (for TCR response)

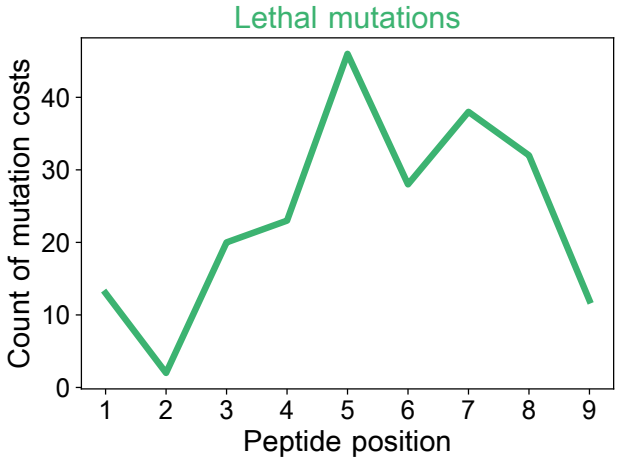
Comparison to TCR avidity assays

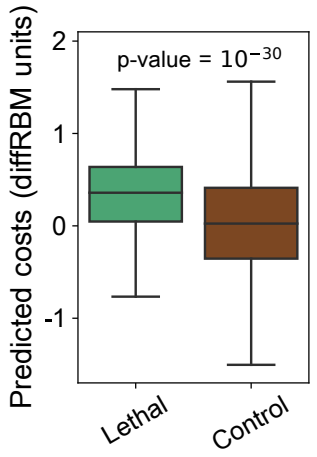
Data on TCR cross-reactivity to NLVPMVATV mutants

From Łuksza et al. Nature 2022



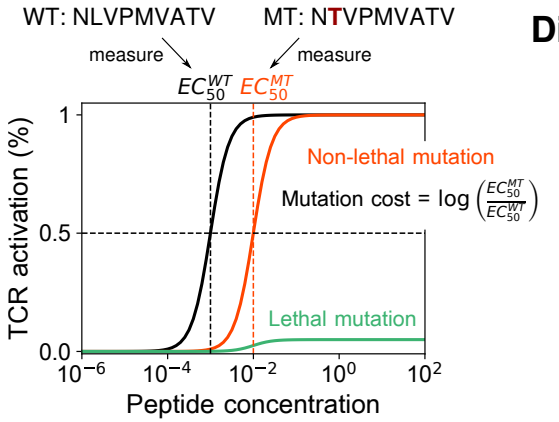
DiffRBM predicts the loss in immunogenicity upon lethal mutations



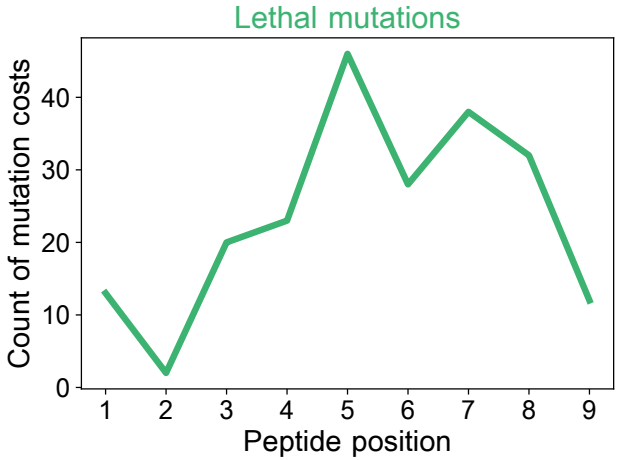


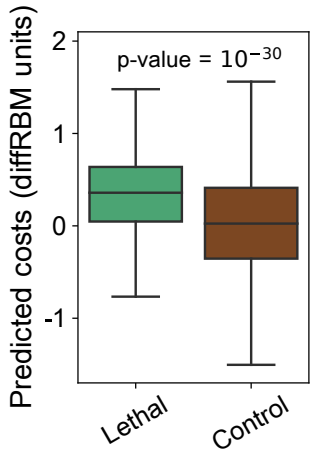
Comparison to TCR avidity assays

Data on TCR cross-reactivity to NLVPMVATV mutants
From Łuksza et al. Nature 2022

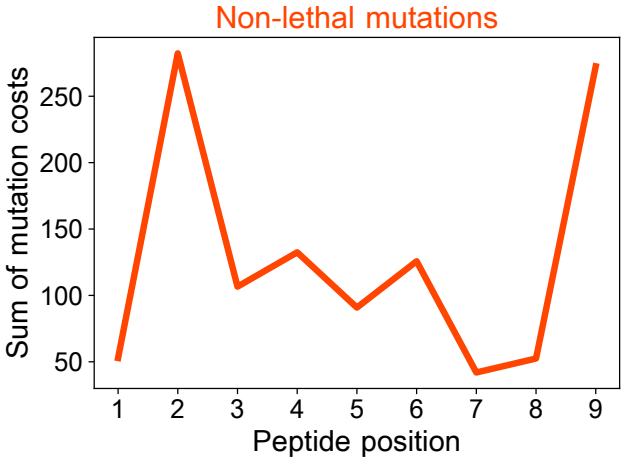


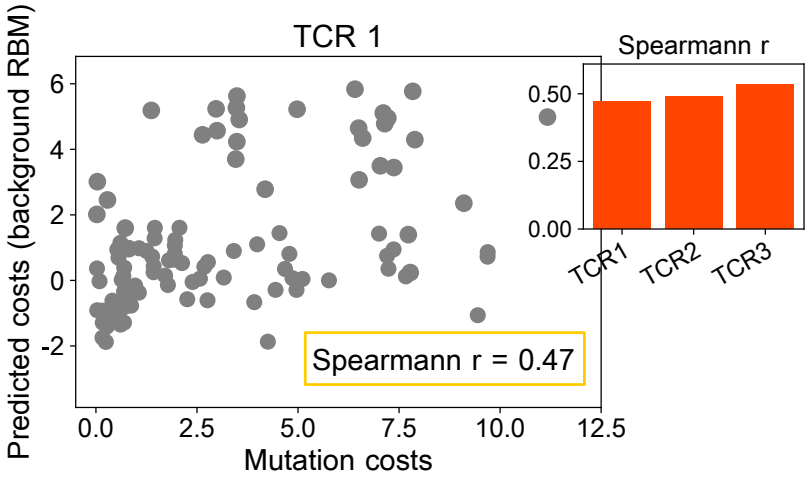
DiffRBM predicts the loss in immunogenicity upon lethal mutations



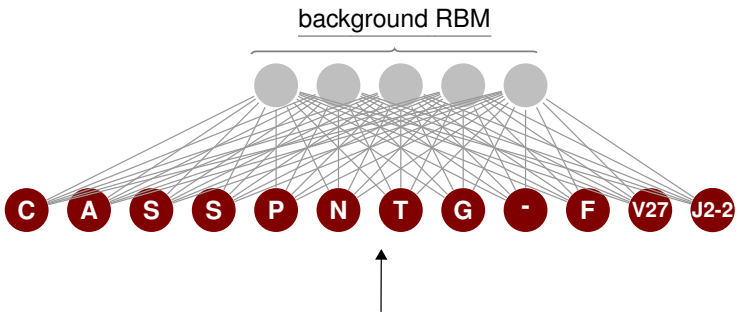


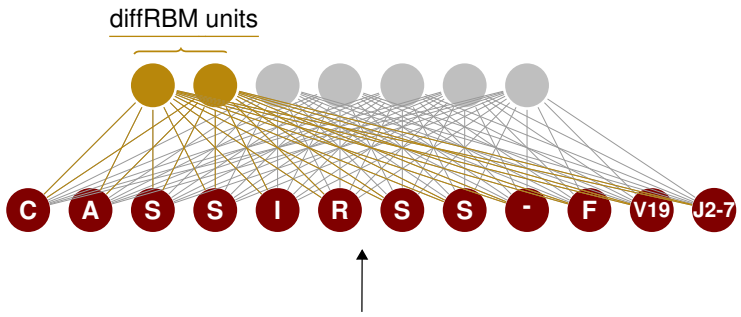
Non-lethal mutation costs are predicted by the presentation model





Model of T-cell epitope specificity





CDR3β sequences from bulk repertoire

CDR3β of TCRs specific to epitope GILGFVFTL

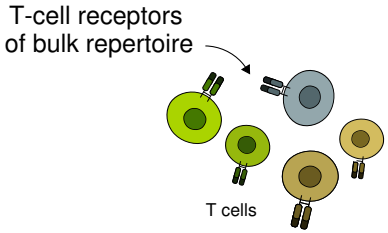
```
CASSPNTGELFF V27 J2-2
CASSIRSSYEQYF V19 J2-7
CASSRDRATDTQYF V28 J2-3
CASSPRRHGDTEAFF V14 J1-1
CASKQGSSYEQYF V6-6 J2-7
CASSPWDGNTEAFF V18 J1-1
```

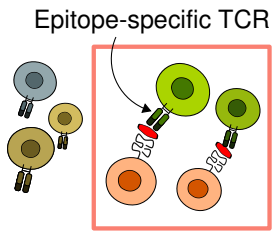
Specific to GILGFVFTL

```
CASSIRSSYEQYF V19 J2-7
CASSQMTGLNTEAFF V28 J1-1
CASSSRSAGELFF V19 J2-2
CASSGGRNQPQHF V5-4 J1-5
CASSVRSSYEQYF V19 J2-7
CASSSRSAYEQYF V19 J2-7
```

Data from Emerson et al. 2017

Data from VDJdb



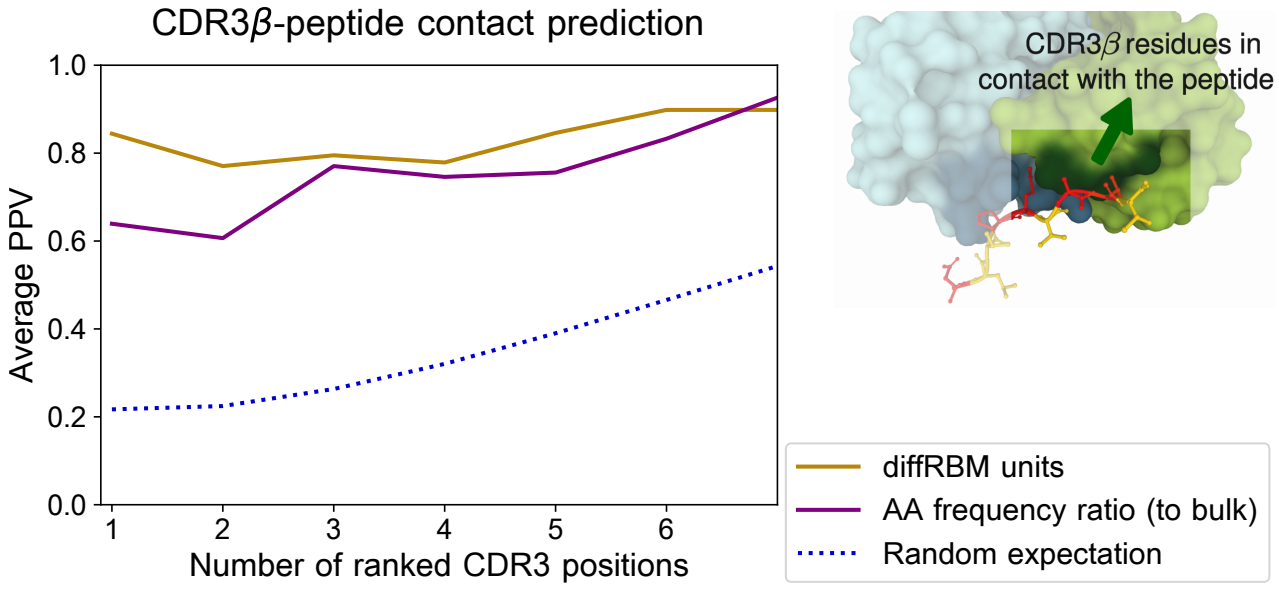


Model of bulk repertoire

Model of epitope specificity

diffRBM units: capture antigen-driven convergent features

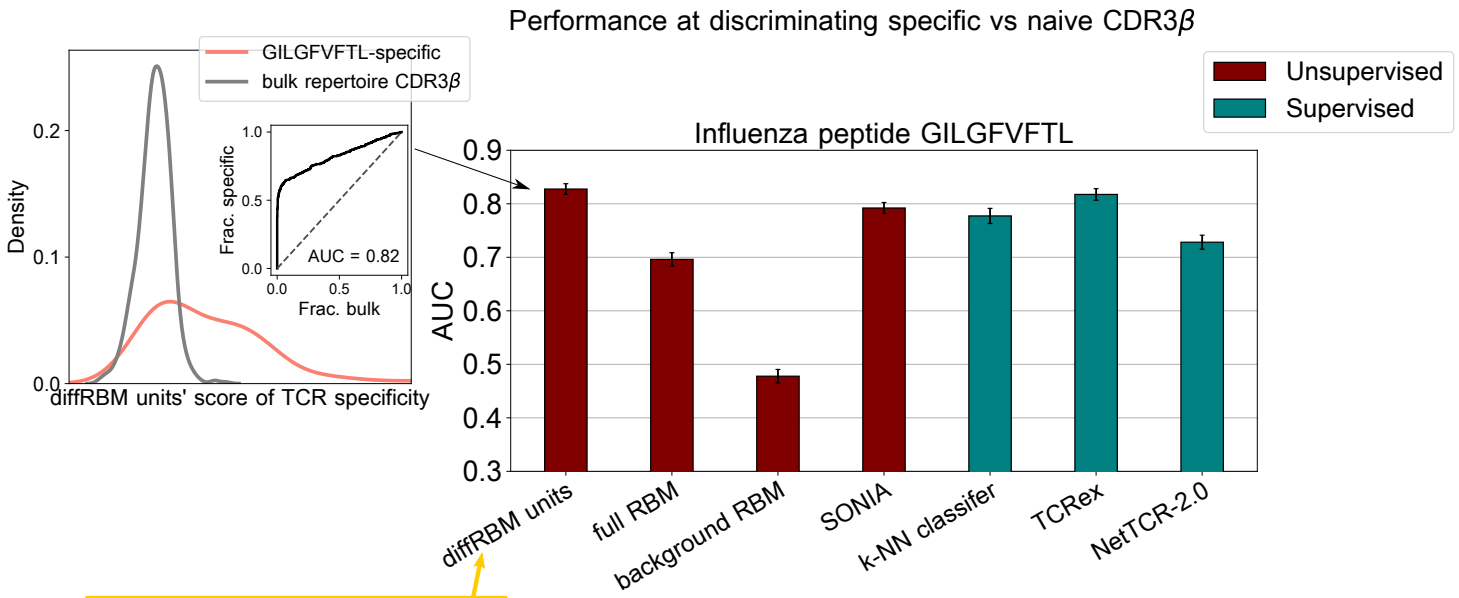
Contact prediction (CDR3β-peptide)



PPV averaged over structures for 4 epitopes (CMV, Influenza, EBV, Sars-Cov-2)

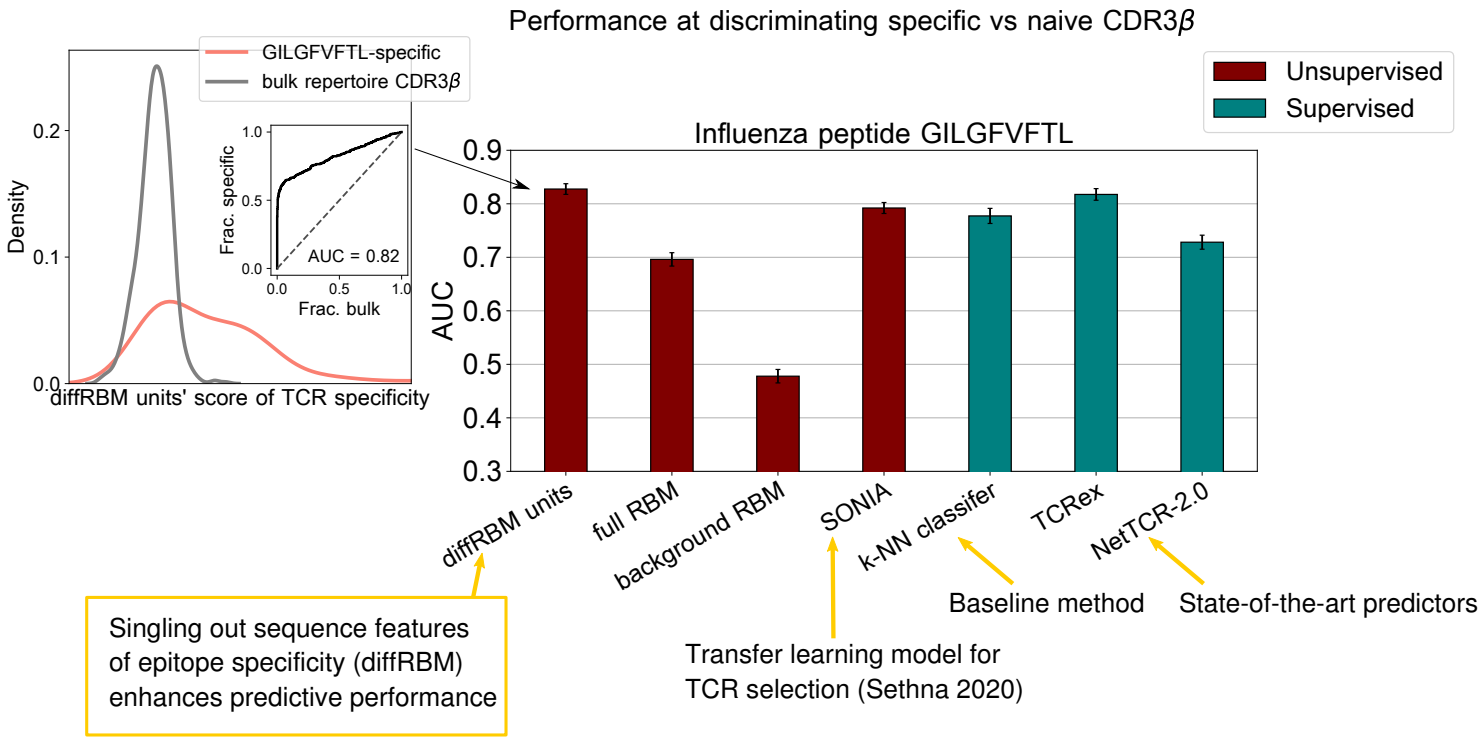
diffRBM performs better than independent models, both higher than random baseline

Classifying epitope-specific receptors

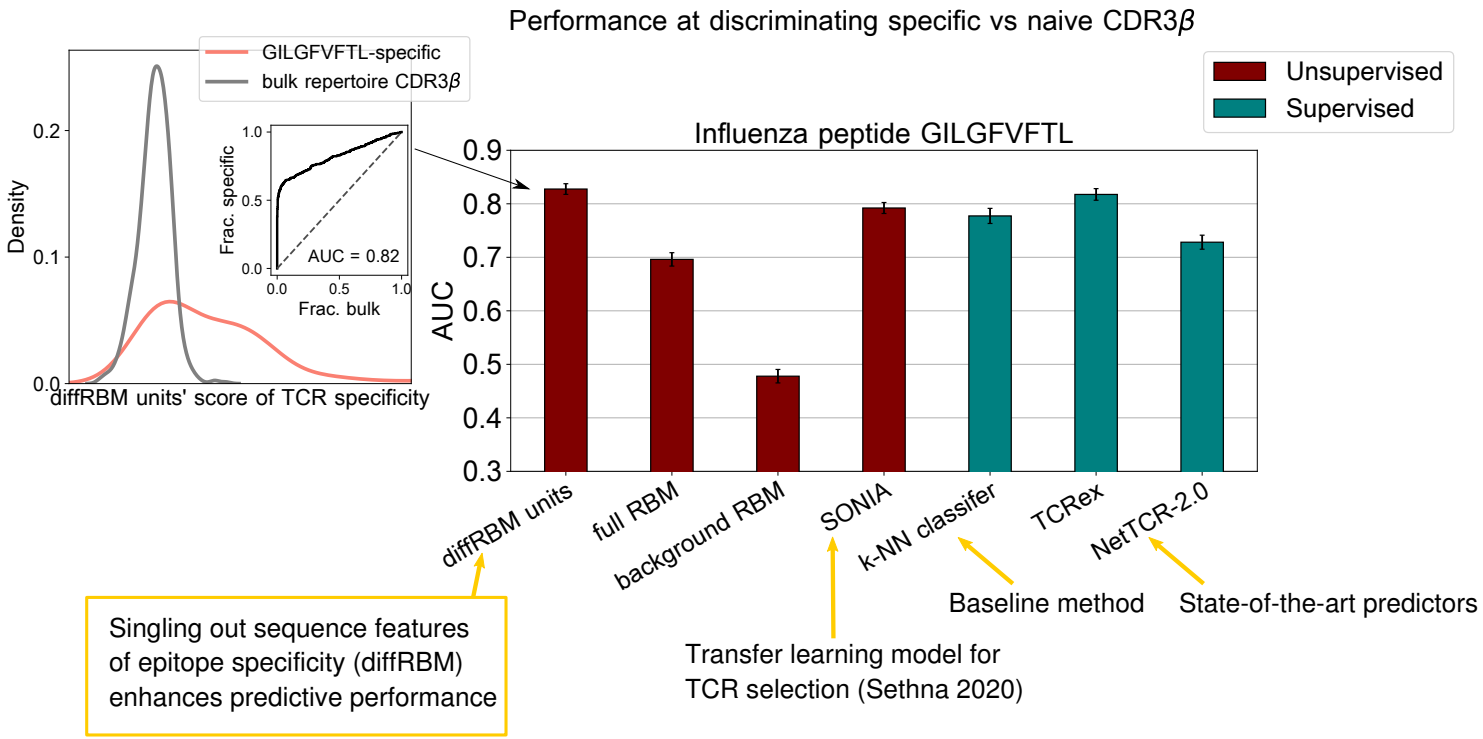


Singling out sequence features of epitope specificity (diffRBM) enhances predictive performance

Classifying epitope-specific receptors

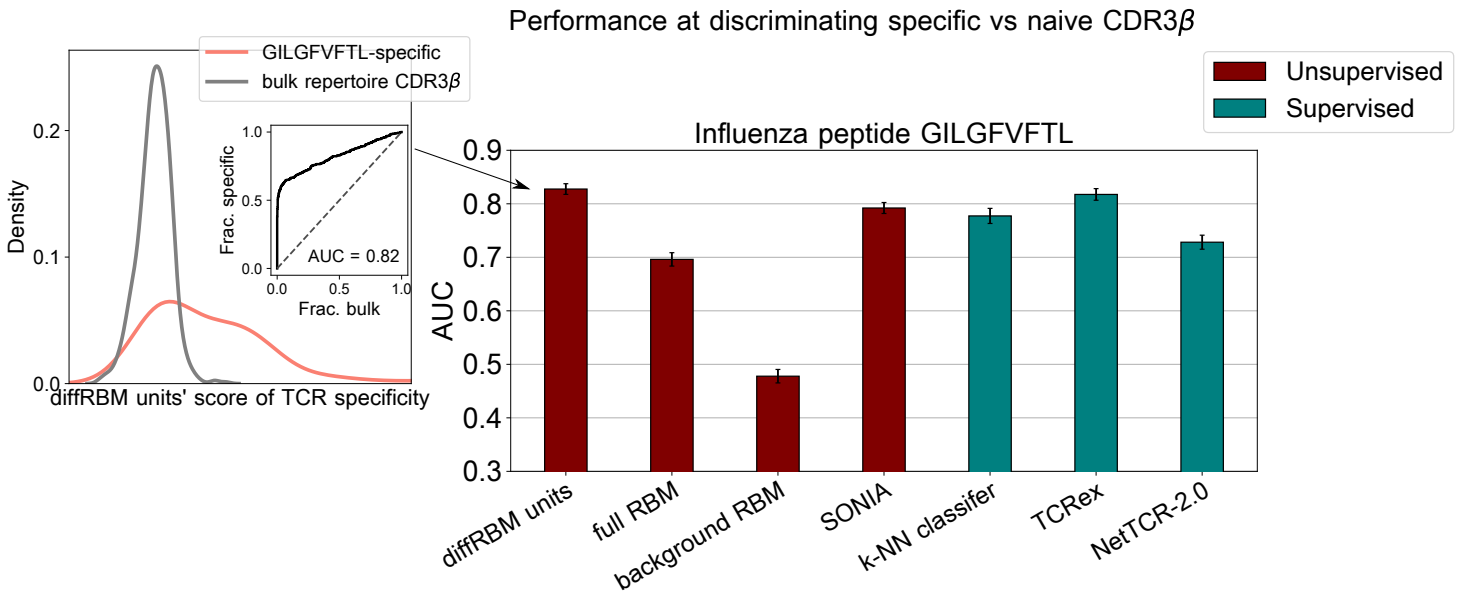


Classifying epitope-specific receptors

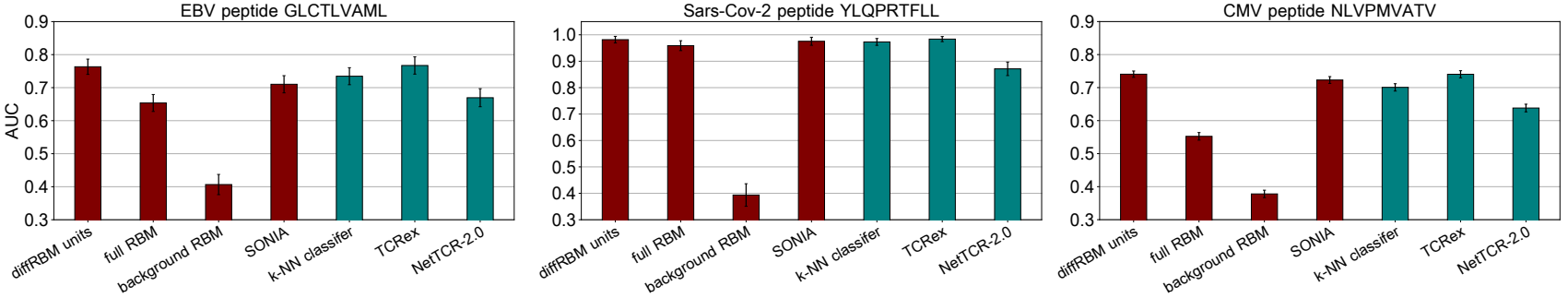


Classifying epitope-specific receptors vs generic non-binders (bulk):
diffRBM reaches the performance of state-of-the-art TCR specificity tools

Classifying epitope-specific receptors



Consistent trend across epitopes



Summary

- Transfer learning: diffRBM parameters capture characteristic differences of immunogenic peptides and epitope-specific receptors

Summary

- Transfer learning: diffRBM parameters capture characteristic differences of immunogenic peptides and epitope-specific receptors
- They allow us to estimate peptide-CDR3β contacts

Summary

- Transfer learning: diffRBM parameters capture characteristic differences of immunogenic peptides and epitope-specific receptors
- They allow us to estimate peptide-CDR3β contacts
- Probabilistic scores distinguish immunogenic vs non-immunogenic peptides, epitope-specific vs generic receptors, with performance comparable to classifiers

Summary

- Transfer learning: diffRBM parameters capture characteristic differences of immunogenic peptides and epitope-specific receptors
- They allow us to estimate peptide-CDR3β contacts
- Probabilistic scores distinguish immunogenic vs non-immunogenic peptides, epitope-specific vs generic receptors, with performance comparable to classifiers
- Applications in vaccine design, TCR engineering, cancer neoantigen discovery, study of viral evolution and immunoediting in cancer

Summary

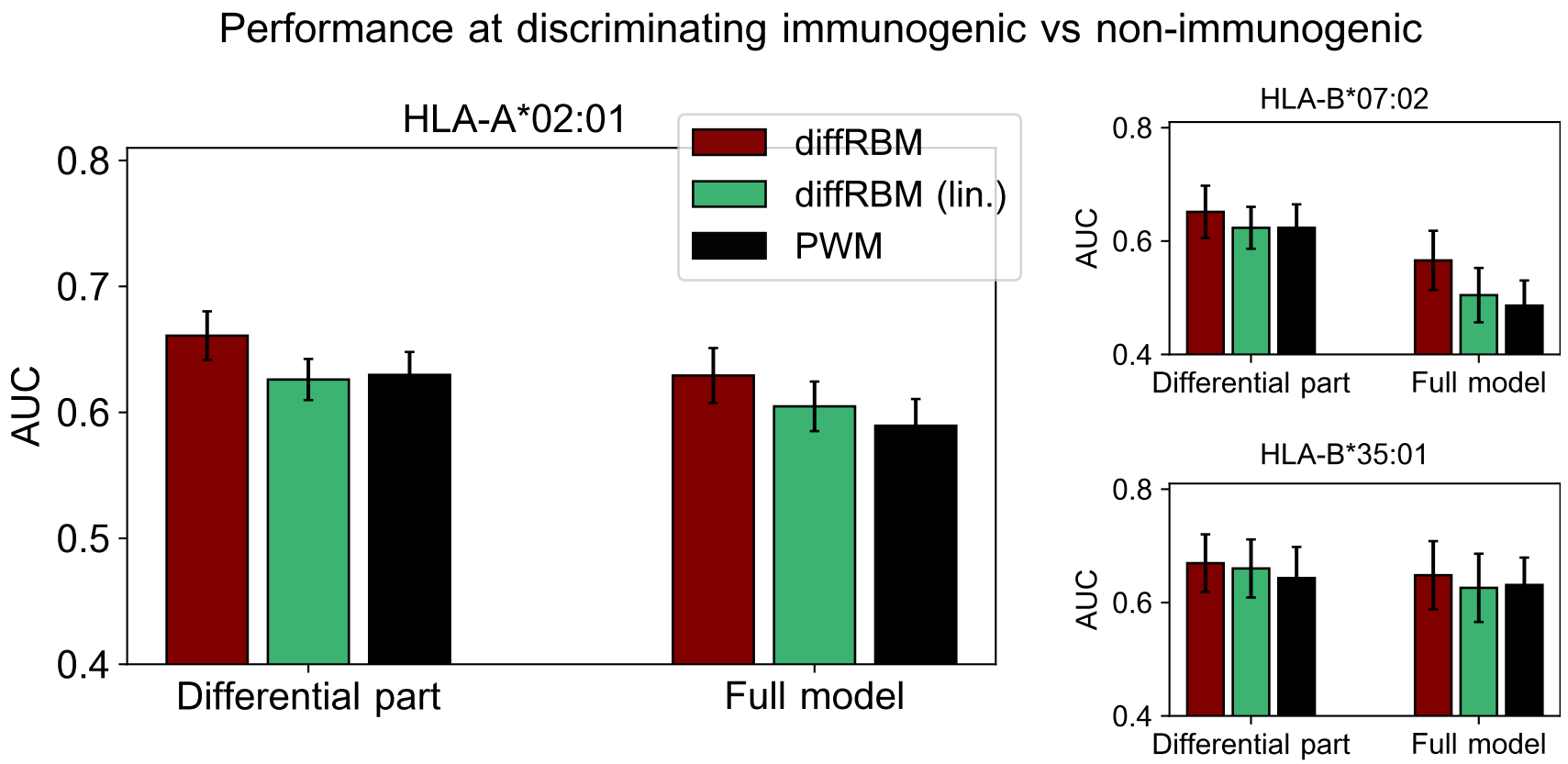
- Transfer learning: diffRBM parameters capture characteristic differences of immunogenic peptides and epitope-specific receptors
- They allow us to estimate peptide-CDR3β contacts
- Probabilistic scores distinguish immunogenic vs non-immunogenic peptides, epitope-specific vs generic receptors, with performance comparable to classifiers
- Applications in vaccine design, TCR engineering, cancer neoantigen discovery, study of viral evolution and immunoediting in cancer
- Broader domain of application: distinctive sequence features that are selected upon (directed evolution, etc.)

ACKNOWLEDGEMENTS

Statistical Physics and Inference for Biology Laboratory, ENS Paris
A. Di Gioacchino, J. Fernandez-De-Cossio-Diaz, S. Cocco, R. Monasson, T. Mora, A.M. Walczak

Thank you for your attention!

Comparison of differential models (Immunogenicity)



Comparison to PRIME and IEDB tool

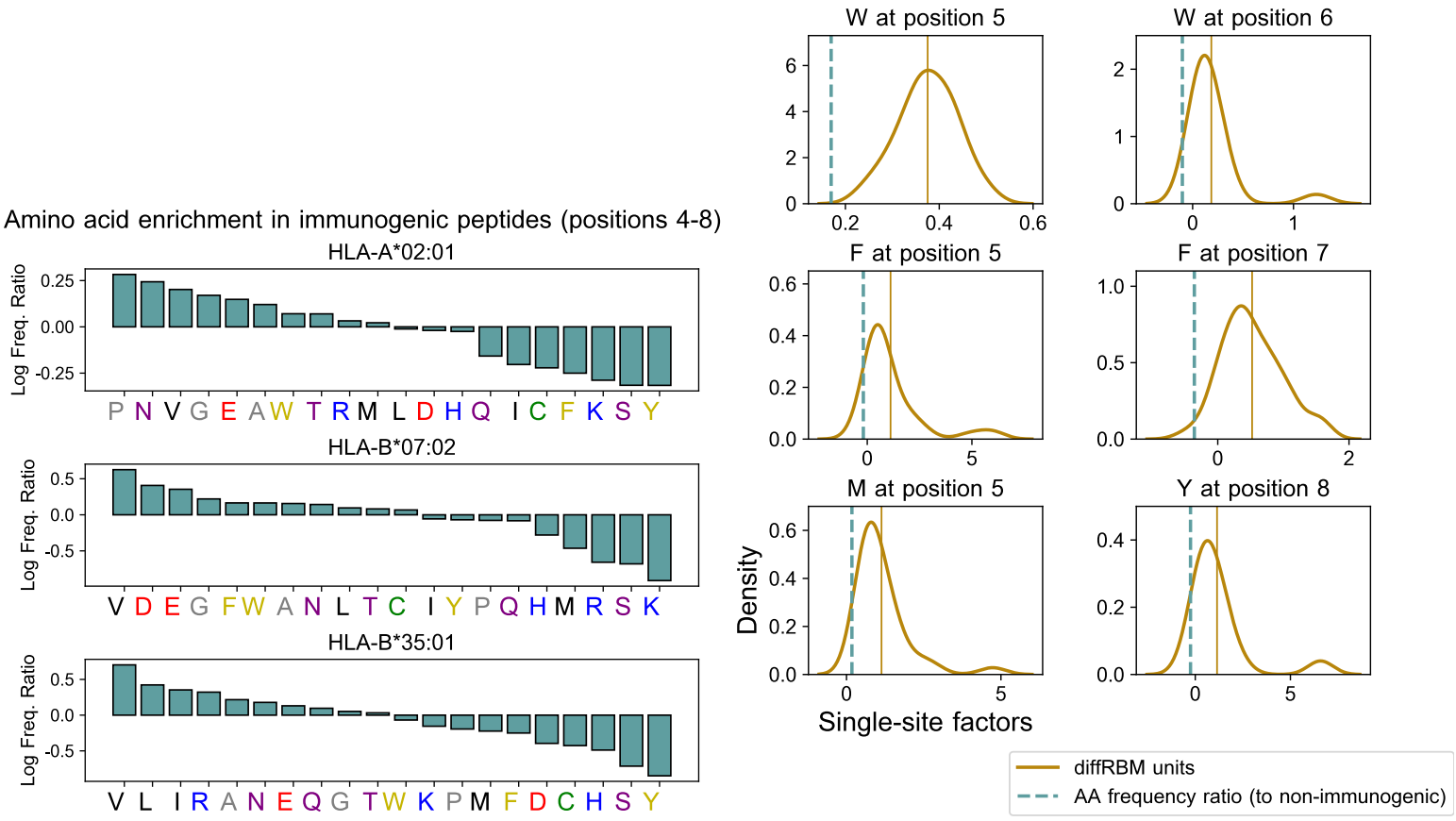
top RBM: AUC = 0.66 (HLA-A*02:01), AUC = 0.65 (HLA-B*07:02), AUC = 0.67 (HLA-B*35:01)

PRIME (Schmidt et al. 2021): AUC = 0.56 (HLA-A*02:01), AUC = 0.52 (HLA-B*07:02), AUC = 0.58 (HLA-B*35:01)

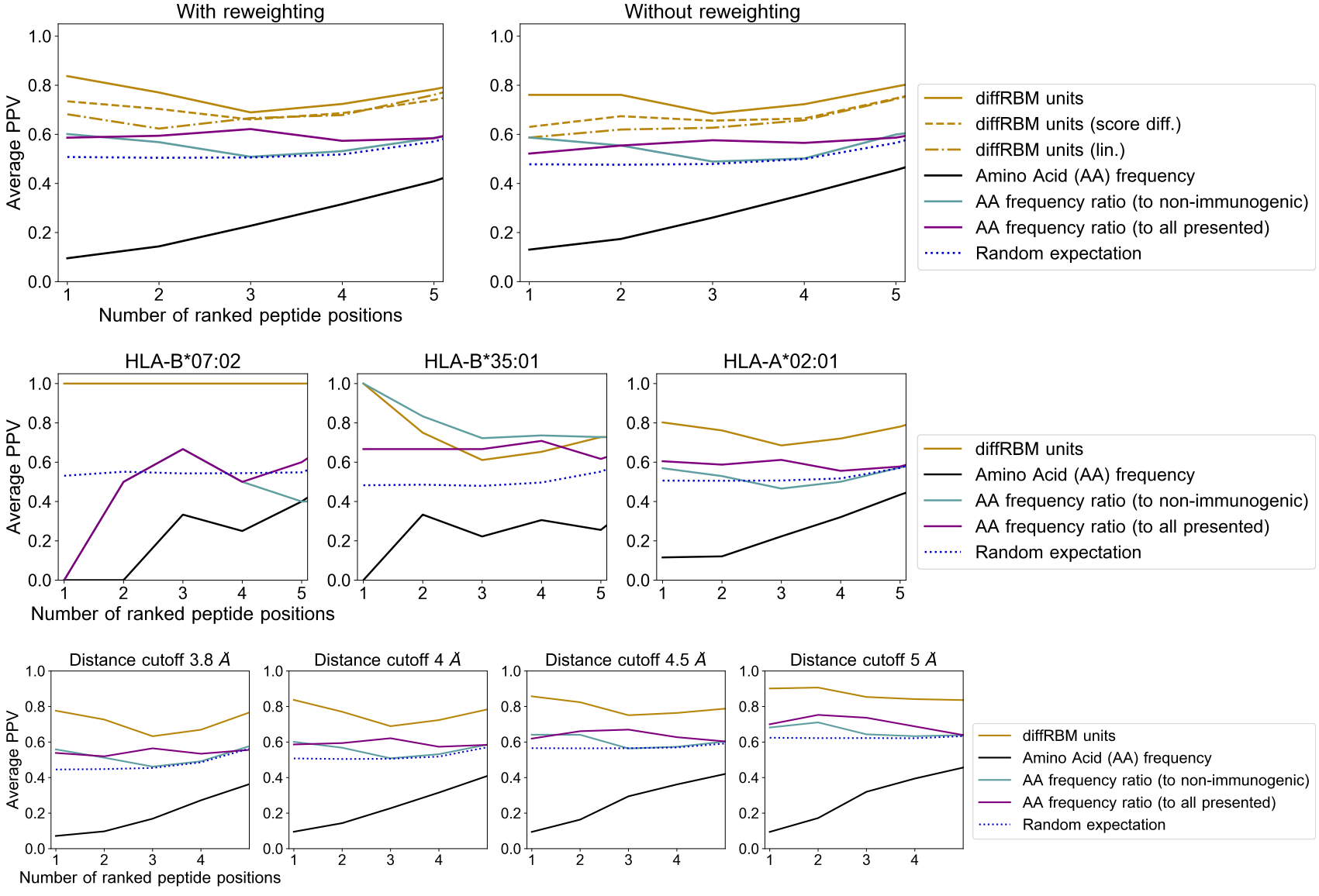
IEDB tool (Calis et al. 2013): AUC = 0.53 (HLA-A*02:01), AUC = 0.60 (HLA-B*07:02), AUC = 0.57 (HLA-B*35:01)

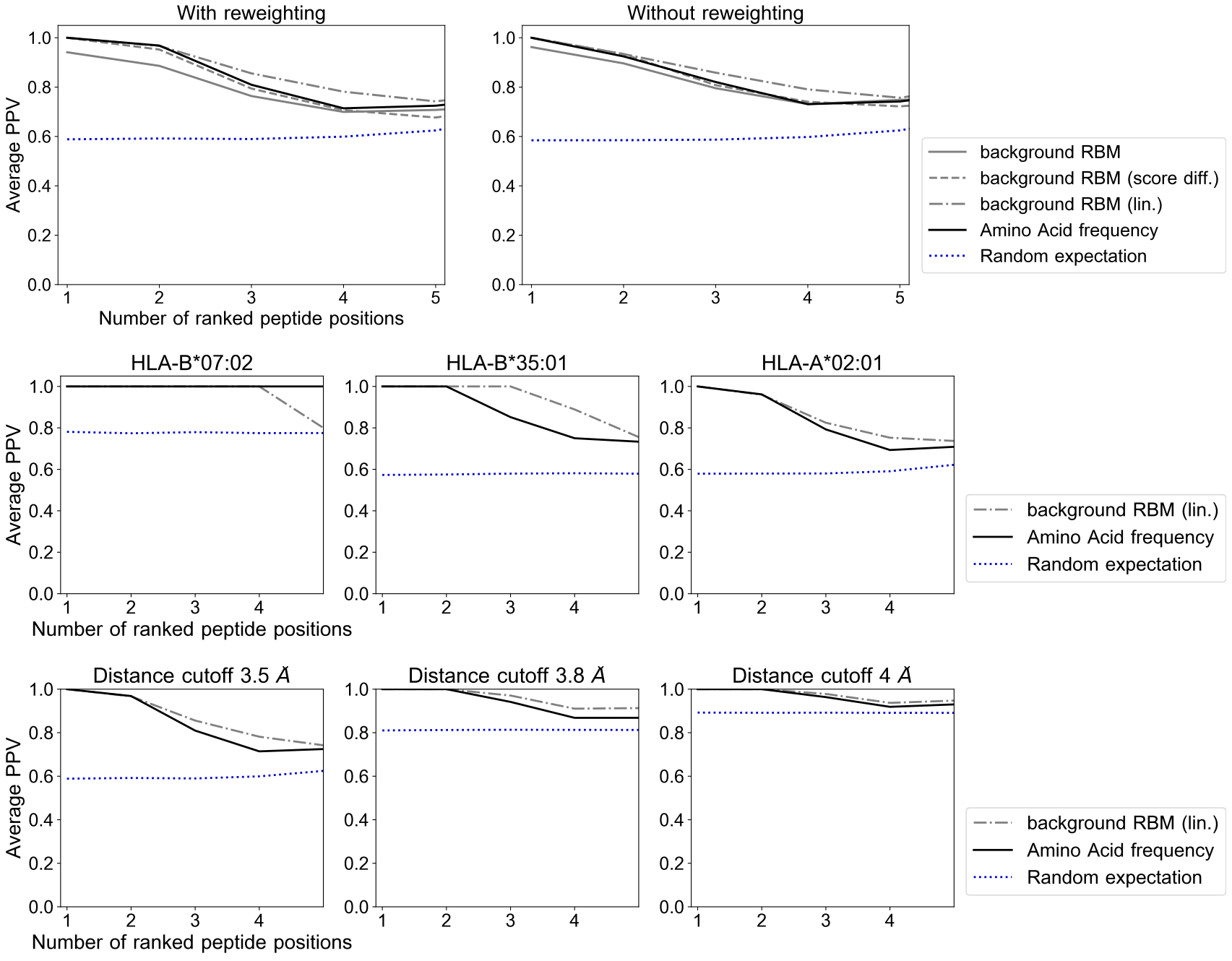
(Note: different training set)

Residues' contribution to immunogenicity

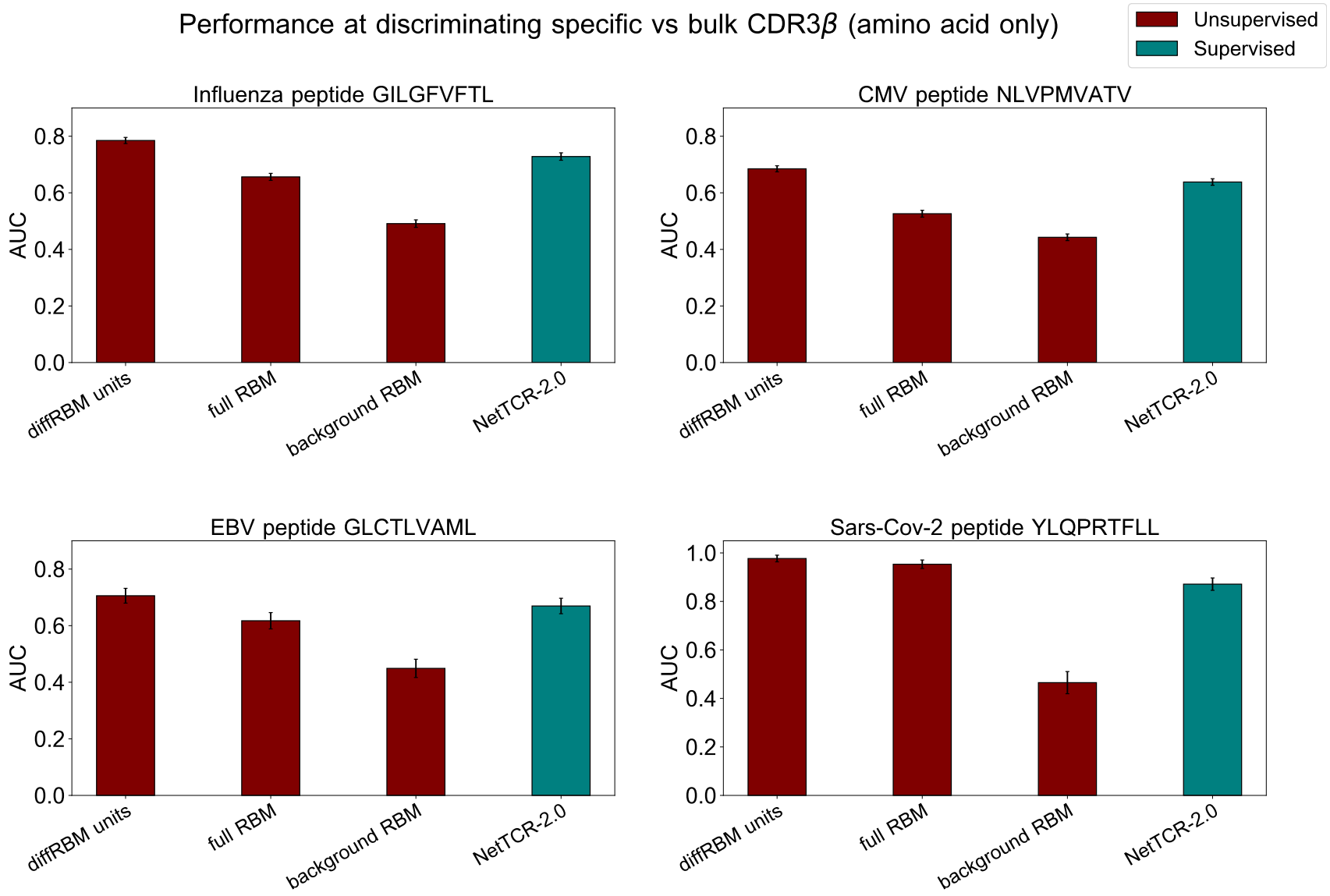


W at position 5 and 6, F at position 5 and 7 (Schmidt et al. 2021), M at position 5 (Luksza et al. 2022), Y at position 8 (Piepenbrink et al. 2013)

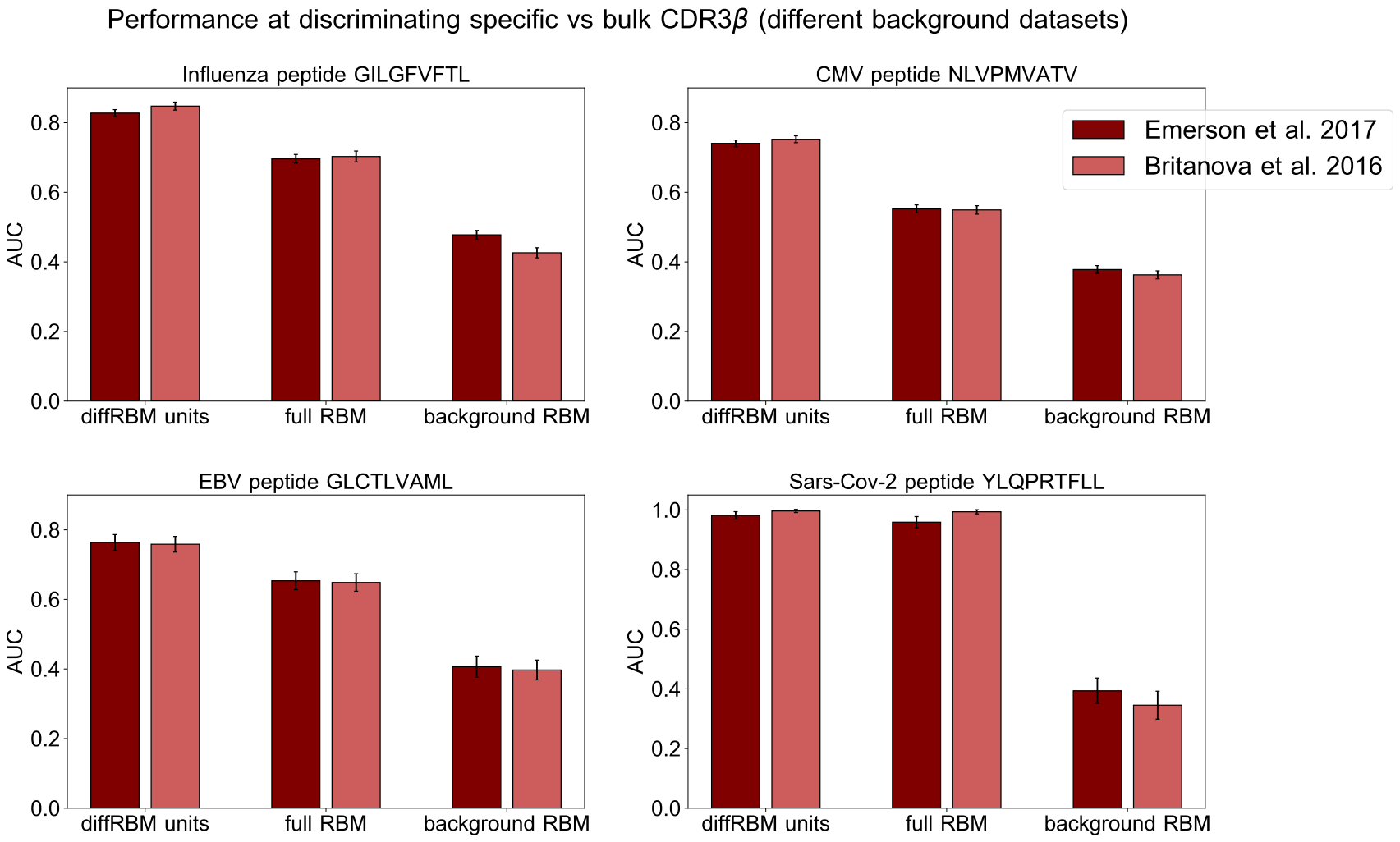

With reweighting
Without reweighting
Average PPV
Number of ranked peptide positions
diffRBM units
diffRBM units (score diff.)
diffRBM units (lin.)
Amino Acid (AA) frequency
AA frequency ratio (to non-immunogenic)
AA frequency ratio (to all presented)
Random expectation
HLA-B*07:02
HLA-B*35:01
HLA-A*02:01
Distance cutoff 3.8 Å
Distance cutoff 4 Å
Distance cutoff 4.5 Å
Distance cutoff 5 Å

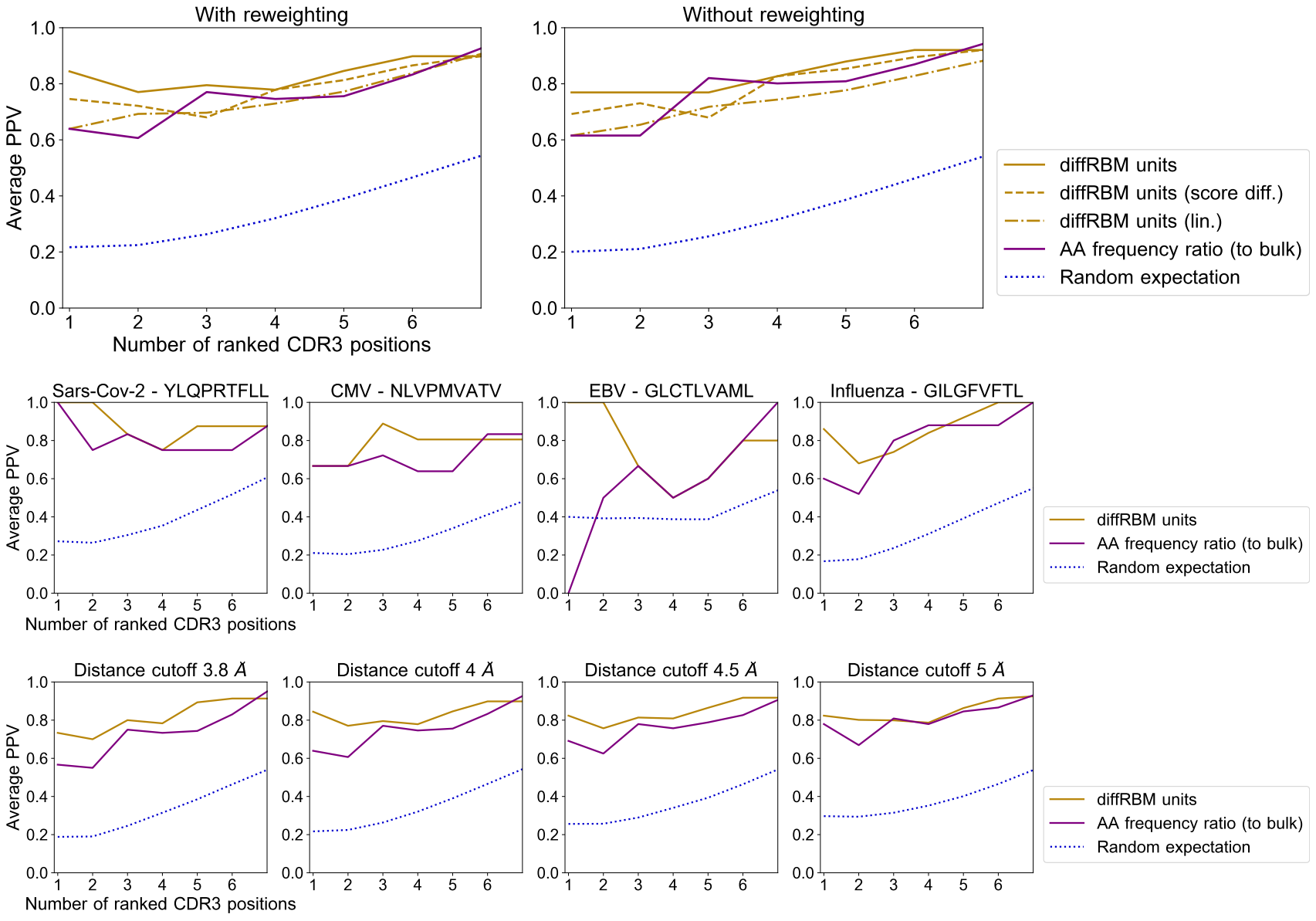

With reweighting
Without reweighting
Average PPV
Number of ranked peptide positions
background RBM
background RBM (score diff.)
background RBM (lin.)
Amino Acid frequency
Random expectation
HLA-B*07:02
HLA-B*35:01
HLA-A*02:01
Distance cutoff 3.5 Å
Distance cutoff 3.8 Å
Distance cutoff 4 Å

TCR specificity

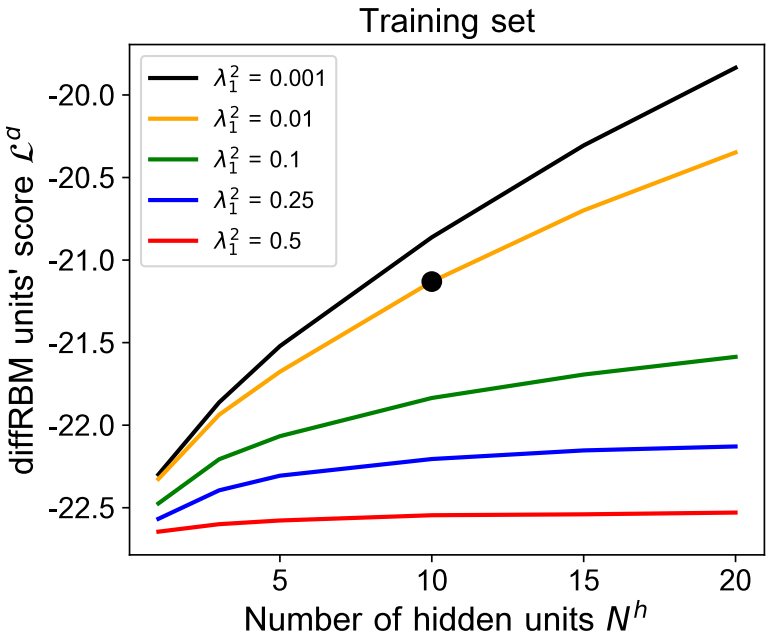


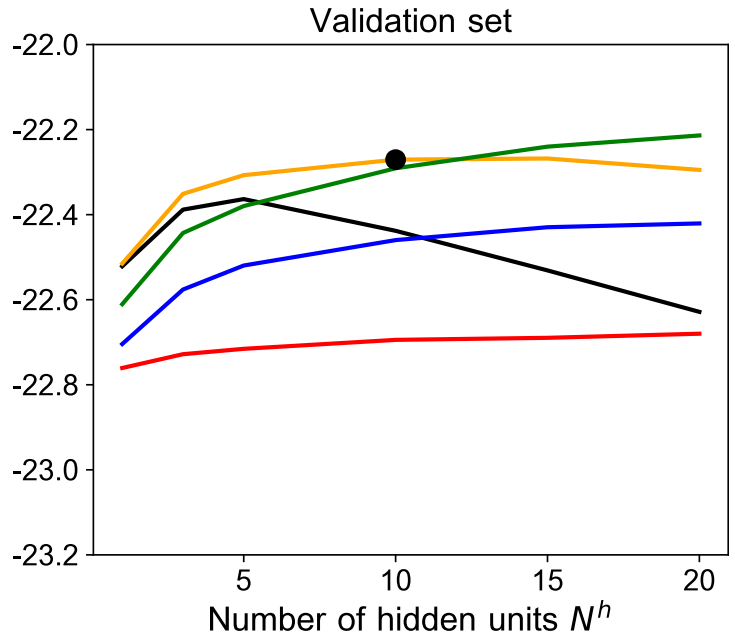
Robustness to choice of naive TCR repertoire

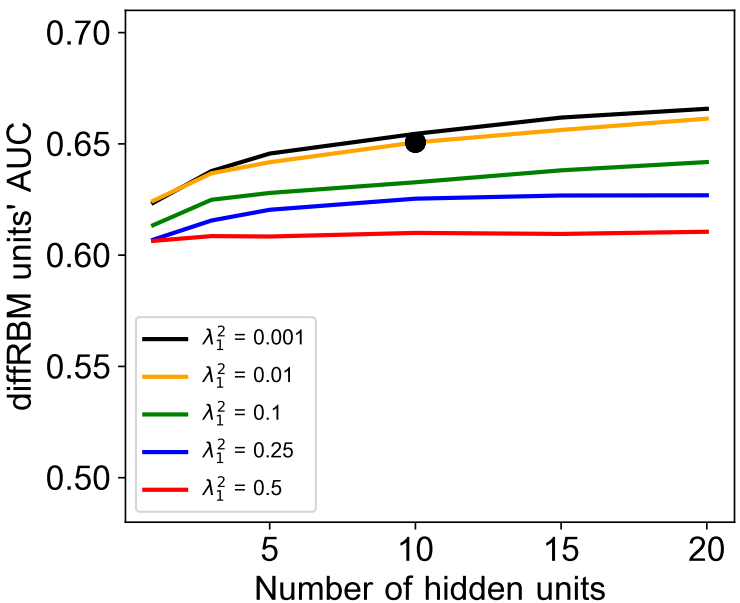


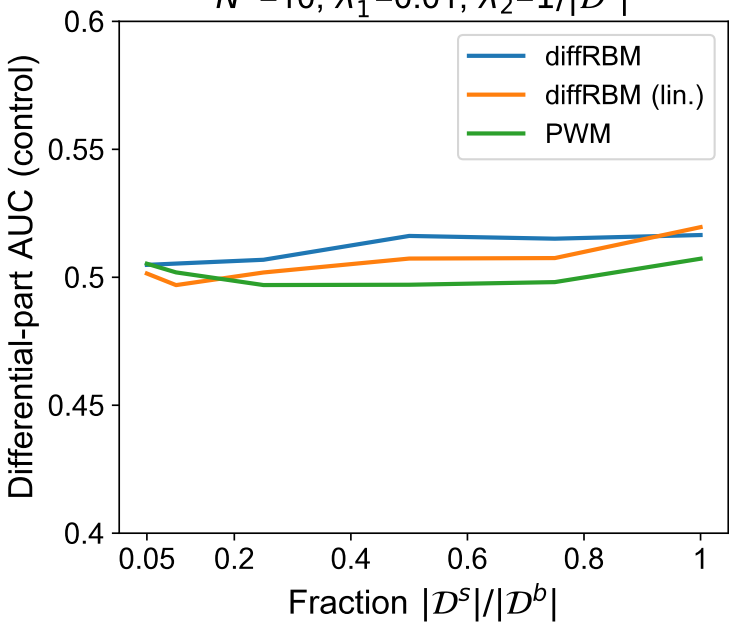

With reweighting
Without reweighting
Average PPV
Number of ranked CDR3 positions
diffRBM units
diffRBM units (score diff.)
diffRBM units (lin.)
AA frequency ratio (to bulk)
Random expectation
Sars-Cov-2 - YLQPRTFLL
CMV - NLVPMVATV
EBV - GLCTLVAML
Influenza - GILGFVFTL
Distance cutoff 3.8 Å
Distance cutoff 4 Å
Distance cutoff 4.5 Å
Distance cutoff 5 Å

Hyper-parametric search

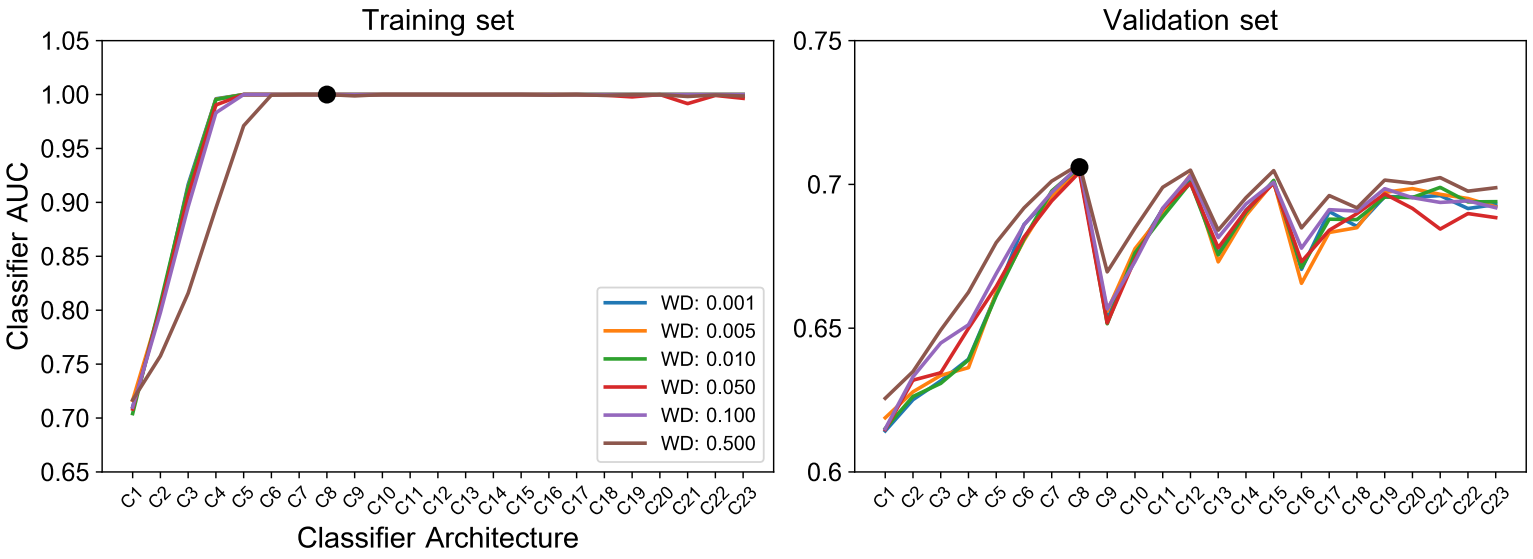




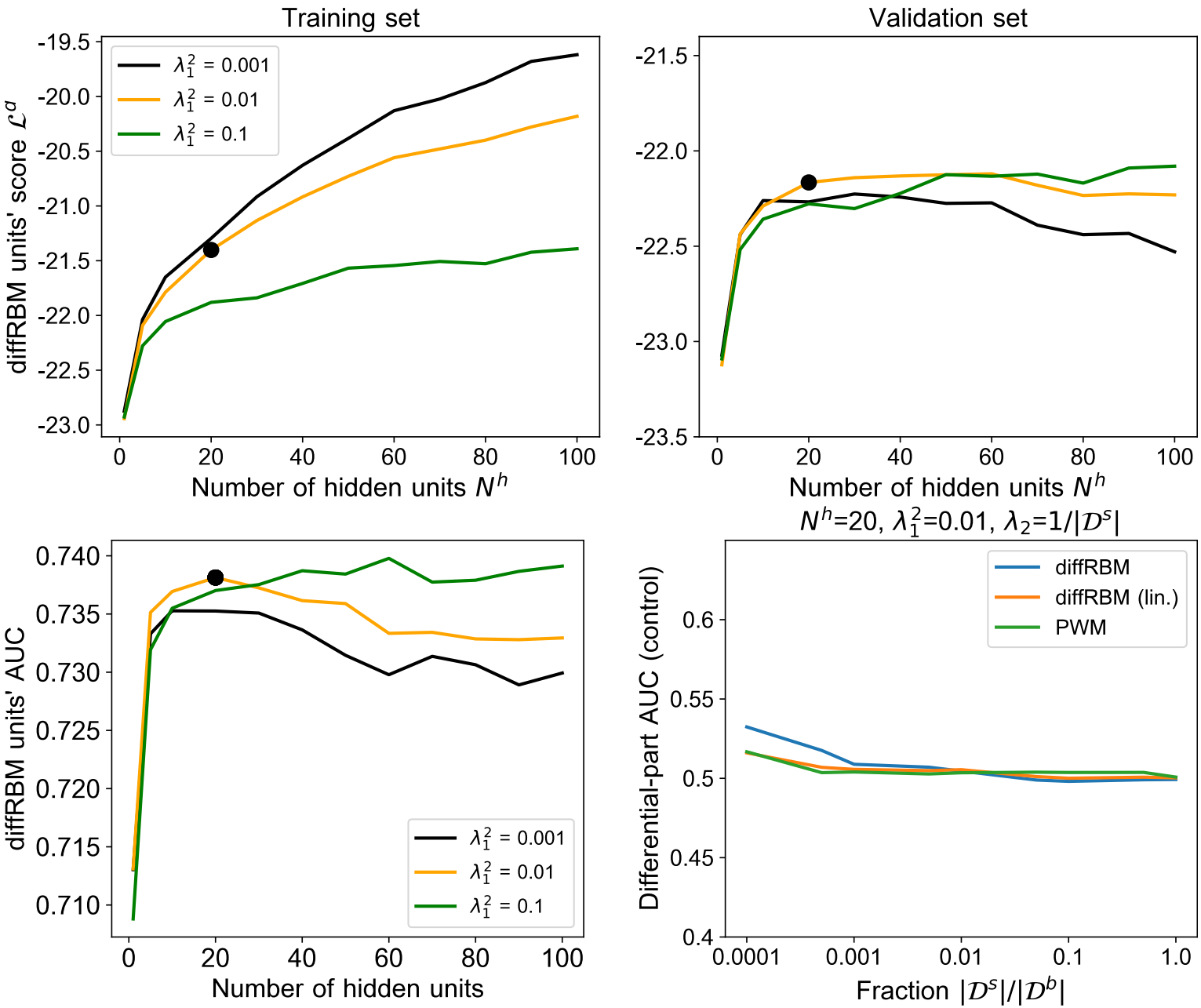




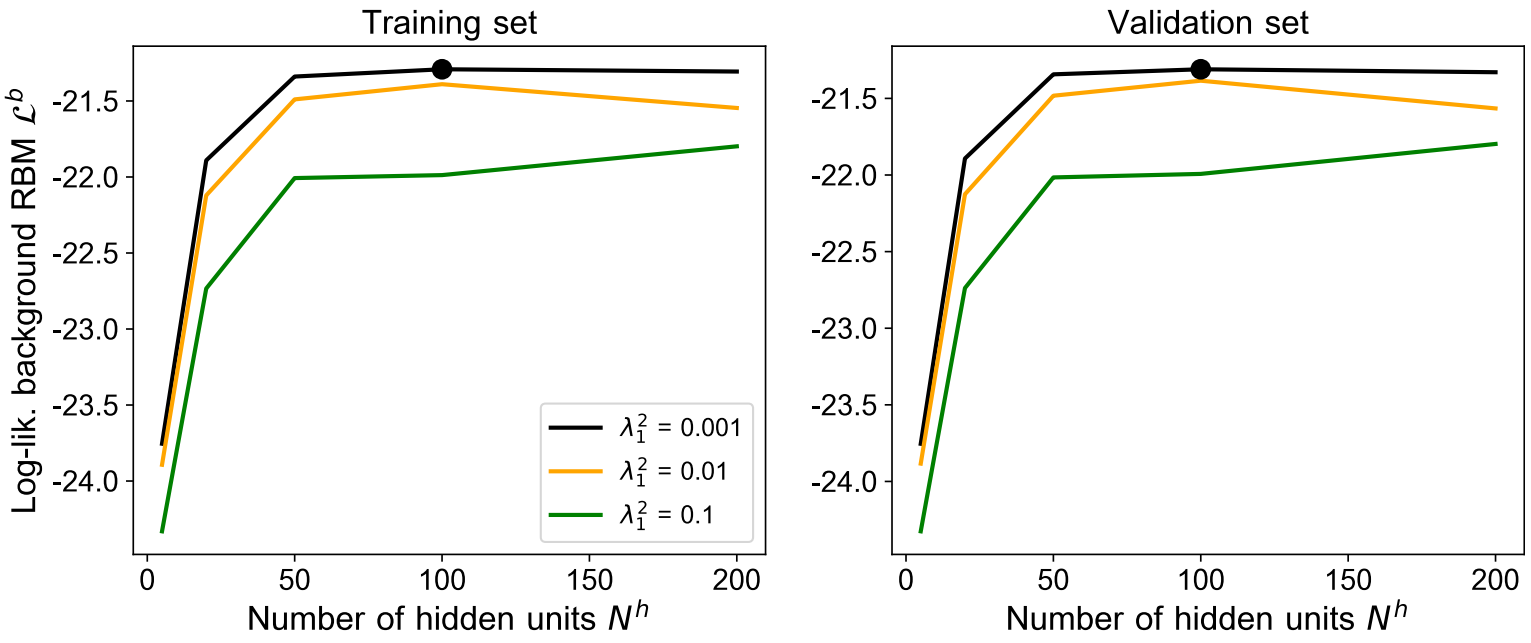
Hyper-parametric search



Hyper-parametric search



Hyper-parametric search



Hyper-parametric search