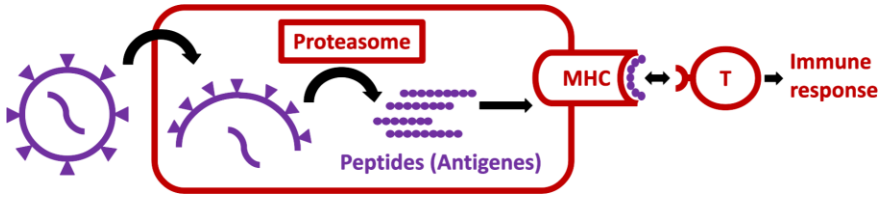
Predicting the Binding of SARS-CoV-2 Peptides to the Major Histocompatibility Complex with Recurrent Neural Networks

{johanna.vielhaben, markus.wenzel, eva.weicken, nils.strodthoff} @hhi.fraunhofer.de · AI Dept., Fraunhofer HHI, Berlin, Germany · arXiv:2104.08237 · github.com/nstrodt/USMPep

Motivation

- Virus peptides bind variably well to MHC, which presents peptides to T-cells & triggers immune response. Vaccines should contain (RNA that encodes) peptides that have a strong binding affinity to MHC.
- USMPep algorithm predicts peptide-MHC-binding-affinity. Potential application: Support development of more effective (multi-epitope) vaccines (that use remaining 7/8 viral proteome apart from spike protein) and rapid adaptation to virus (escape) variants.



Datasets and Targets

- Training: continuous binding affinity measurements from Immune Epitope Database (IEDB); binders identified by mass-spectrometry (IEDB), artificial neg. samples (binary)
- Test: binding stability between SARS-CoV-2 peptides and MHC [2]; more specific than binding affinity

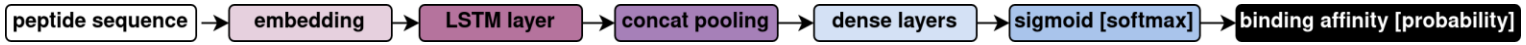
References

[1] Johanna Vielhaben et al. *USMPep: Universal Sequence Models for Major Histocompatibility Complex Binding Affinity Prediction.* BMC Bioinformatics 21, 279, 2020. https://doi.org/10.1186/s12859-020-03631-1

[2] Marek Prachar et al. *Identification and validation of 174 COVID-19 vaccine candidate epitopes reveals low performance of common epitope prediction tools.* Sci Rep 10, 20465, 2020. https://doi.org/10.1038/s41598-020-77466-4

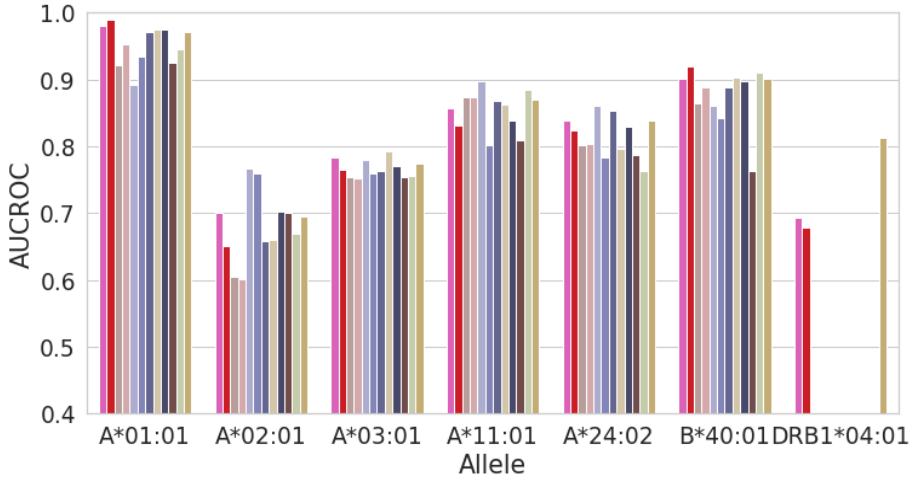
Model – USMPep [1]

- Language-model pretraining on unlabeled peptide data
- Finetuning of LSTM-model on peptide-MHC binding prediction task (for each MHC allele)
- Advantages: simple architecture/training procedure, arbitrary peptide length as input, no heuristics
- Novelty: Diversity through ensemble of binding affinity (BA) regressors and mass spectrometry (MS) classifiers



Results

- USMPep performance in benchmark [2]: *Spearman's ρ:* top 1 overall;
  AUCROC: top 4 overall; top 1 on selected alleles
- Ensemble of regressors (BA) and classifiers (MS) performs better than BA-only



Model	Spearman's ρ	AUCROC
USMPep_BAMS	**0.56085**	**0.84785**
NetMHCstab 1.0	**0.51745**	**0.86080**
NetMHCpan_BA 4.0	**0.51610**	**0.86080**
IEDB-AR Consensus	**0.51440**	**0.85470**
USMPep_BA	**0.50280**	0.82660
NetMHC 4.0	0.49545	0.83385
NetMHCpan_EL 4.0	0.49395	0.79235
NetMHCcons 1.1	0.49285	0.82865
MixMHCpred 2.0.2	0.48000	0.77485
SMMPMBEC 1.0	0.46845	0.83235
SMM 1.0	0.46540	**0.83760**
MHCFlurry 1.3.0	0.44265	0.82350

ICLR 2021 Workshop: Machine Learning for Preventing and Combating Pandemics

Fraunhofer HHI