Deep learning using protein structural modeling for prediction of T-cell receptor and peptide MHC binding

DTU xxxxxx xxxxxxx
Xxxxxxxx xxxxxxxxx xxxxxxxx

Master Thesis by Magnus Haraldson Høie

Summary: Binding of T-cell receptors (TCRs) to their cognate peptide-MHC target plays a key role in the activation of the adaptive immune system. Siamese like convolutional neural networks have shown state-of-the art performance on prediction of protein-protein interactions. Here a published CNN architecture is modfied and trained on 1464 generated TCR-p-MHC homology models from which the amino-sequence, FoldX energy terms and other structural features are extracted as training freatures. Training on features extracted from homology model structures significantly outperforms a model only trained on the complex amino-sequence.

Training set	MCC-score	AUC
Previous CNN model (baseline)	17 %	75 %
(1) Structure	16 %	63 %
(2) FoldX energy terms	34 %	75 %
(3) Amino-sequence	40 %	71 %
All features (1, 2 & 3)	**45 %**	**78 %**

Table X: The CNN model was trained independently on each feature set and one combining all features. Training on all features combined significantly increased model performance (MCC and AUC score) versus training on the complexes´ amino-sequence only.

Homology model generation:
Positive and negative TCR-p-MHC complexes from internal switching

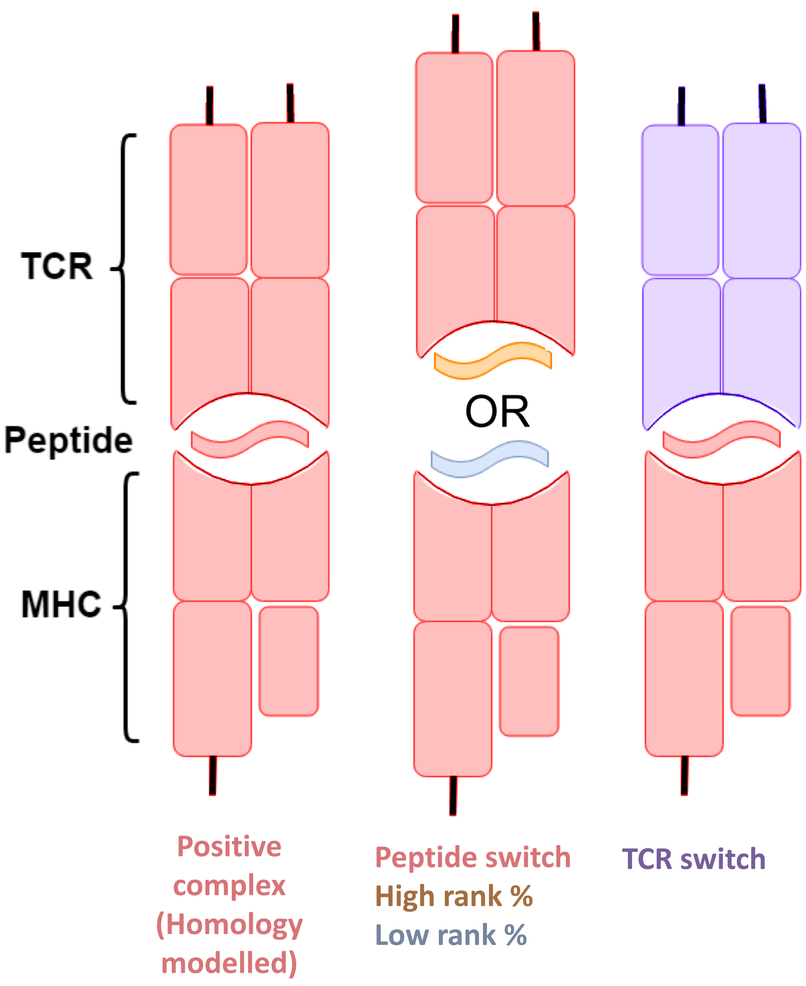

Figure X: From 61 available solved crystal structures of TCR-p-MHC complexes on IEDB (1), 4 sets of homology models were created, allowing at most a 95 % template identity. The positive set comprised 61 real complexes modelled on each other at 70, 80, 90 and 95 % identity. Negatives were created by switching out the complex peptide with one predicted to have either high or low binding activity as ranked by NetMHCpan, and by switching the TCR from within the dataset.

Introduction: A proof-of-concept for increased prediction performance from integrating energy terms into a TCR-p-MHC binding CNN prediction model was demonstrated by Olsen et al 2018 (baseline model) (x). Additionally, a sequence only model by Jurtz et al 2018 (x) demonstrated high accuracy for prediction of binding for known TCR and peptide parts, but model predictions failed when generalizing to unseen sequences. Here, increased model performance and recall of binding (positive) complexes is shown by taking advantage of additional training features extracted from generated homology models.

Methods: 1464 homology models of known binding (positive) and predicted non-binding TCR-p-MHC structures were generated using an in-house pipeline using MODELLER and LYRA (x). Protein free energies were extracted from the models using FoldX, and structural terms using PDB_Tool (x). Training data was split into 5 separate partitions with at most 70 % shared sequence identity, weighted 1/3 for TCR, peptide and MHC parts respectively. The Siamese like CNN architecture used exploits two parallel pipelines for processing the TCR and p-MHC separately. A protein-protein interaction is computationally mimicked using a custom random projection module. Each vector protein representation is multipled over a mirrored gaussian noise-matrix. Finally each protein representation is element-wise multiplied to generate a final vector from which binding or non-binding is predicted upon. A sytematic comparison was made of effect of different training features´ on model performance.

Increased model recall from training on all features vs amino-sequence only

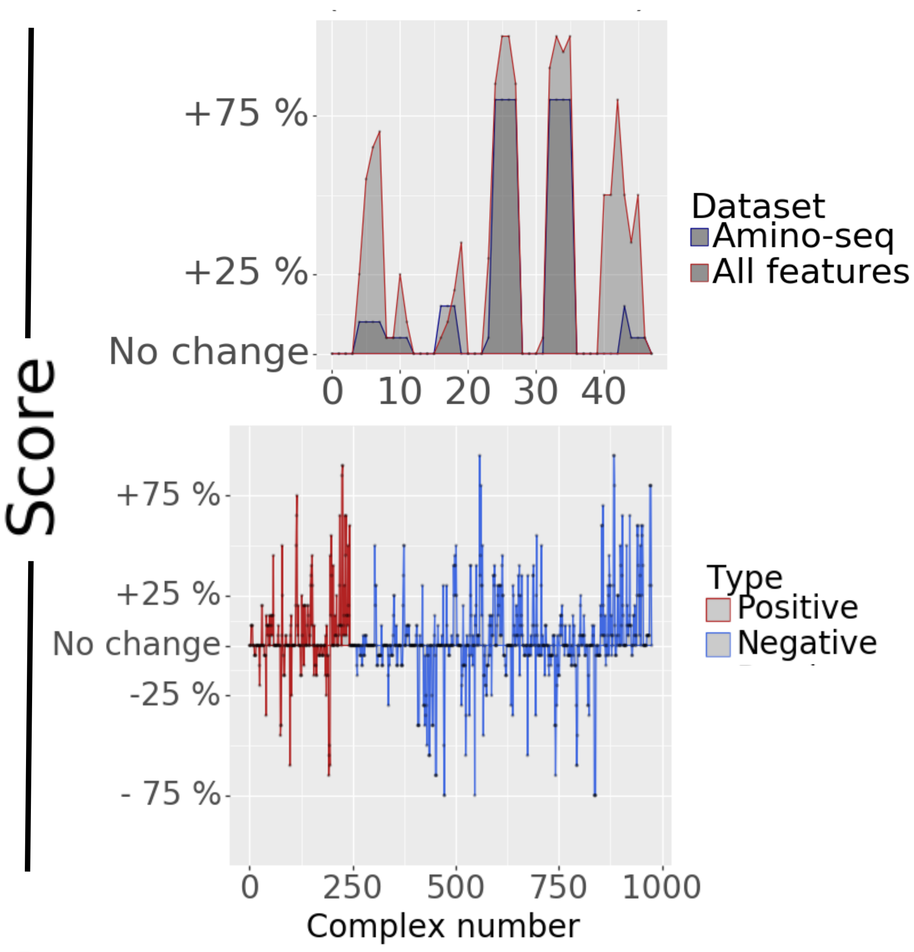

Figure X: In order to comprehensively test effect of model predictions based on input training features, 40 separate CNN models were trained. 20 training on only the amino-sequence and 20 training on all features combined. Training on all features is seen to significantly increase the chance of recall by independently trained models, both for previously unrecognized positive complexes, and increase recall. Top figure shows increased recall for a single validation set. and bottom for all complexes across all partitions also including negative complexes. Low rank % peptide and TCR switched data not shown.

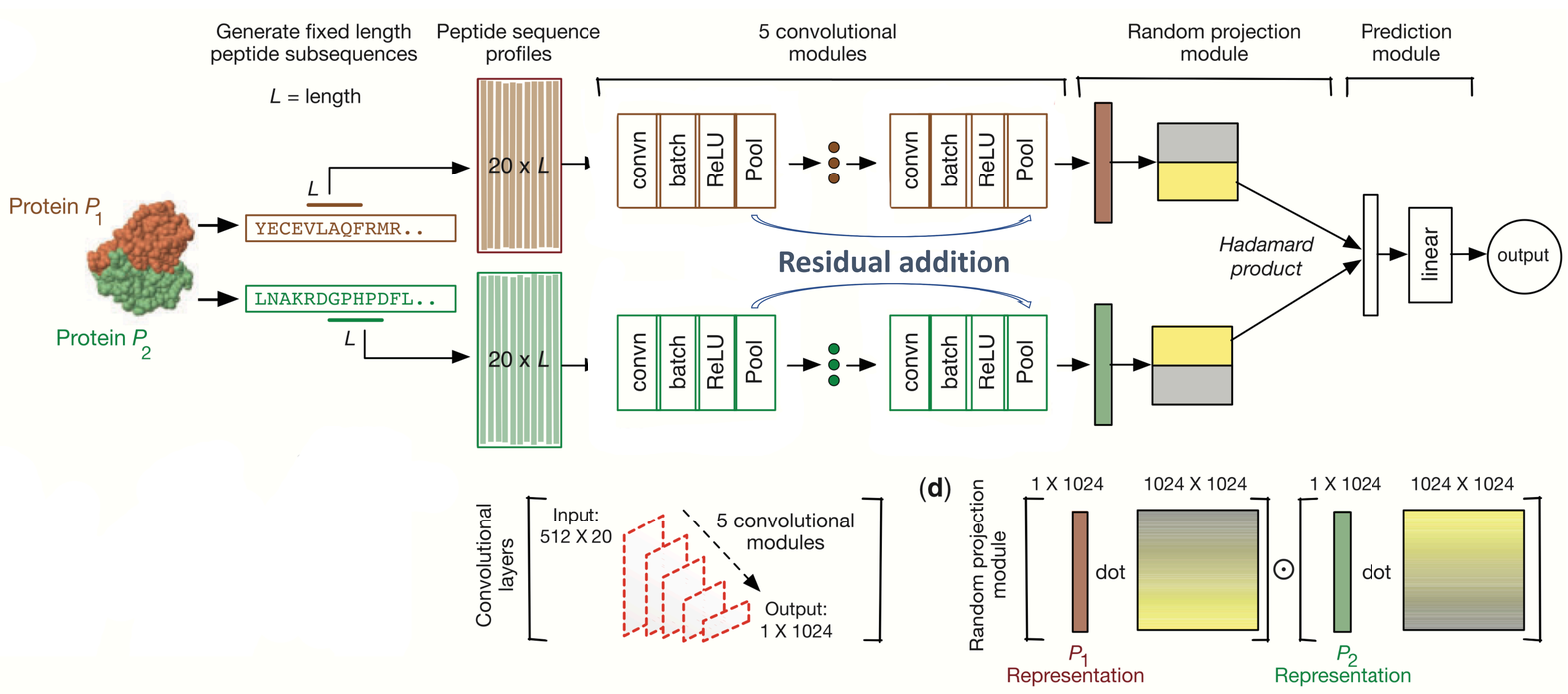

Figure X: A modified Siamese-like CNN architecture using for prediciting TCR-p-MHC binding, adapted from Hashemifar et al 2018. The TCR and p-MHC protein parts are divided into 2 separate tensors which are padded to size 1x122x512. The amino-acid sequence position is represented along the last dimension, while the 122 feature channels represent the one-hot encoded amino-acids (21), structural features such as conformational letter sequence, solvent accessibility and secondary structure (35) and FoldX energy terms (66). Each protein part goes through separate but parallell set of 5 convolutional modules. Before exiting each module, the input vector is halved in size through average-pooling, before entering the next module,. Residual skip-connections (see ResNet) are made possible by residual output of one module directly to the end average-pool layer of the next module and significantly improves performance.

Conclusion

Integrating energy and structural features from generated homology models increased model performance by 5 % MCC and 7 % AUC score. Notably, increased recall was seen for both otherwise missed and previously recognized complexes ...

...

...

References:
0) Olsen
1) Jurtz
1) IEDB
2) NetMHCpan
3) LYRA
4) Hashemifar et al 2018
5) PDB_Tool
6) Resnet