Prediction of TCR-pMHC interactions using molecular modeling and recurrent networks

Ella Hedeboe s211253, Henrietta Holze s215945, Christian Johansen s202770, Paul Simon s202592

https://github.com/HenriettaHolze/TCR-pMHC-prediction

Introduction

- Recent advances within biological sequencing and deep learning methods have made it possible to investigate key interactions of the immune system computationally.
- The adaptive immune system is a key element for fighting diseases and the T-cells are responsible for cell-mediated immune response via their surface T-cell receptors (TCR).
- TCRs bind to peptide-Major Histocompatibility Complexes (pMHC) to form a complex that triggers an immune response.

Problem: Predict TCR-pMHC binding using molecular modeling and recurrent neural networks (RNN).

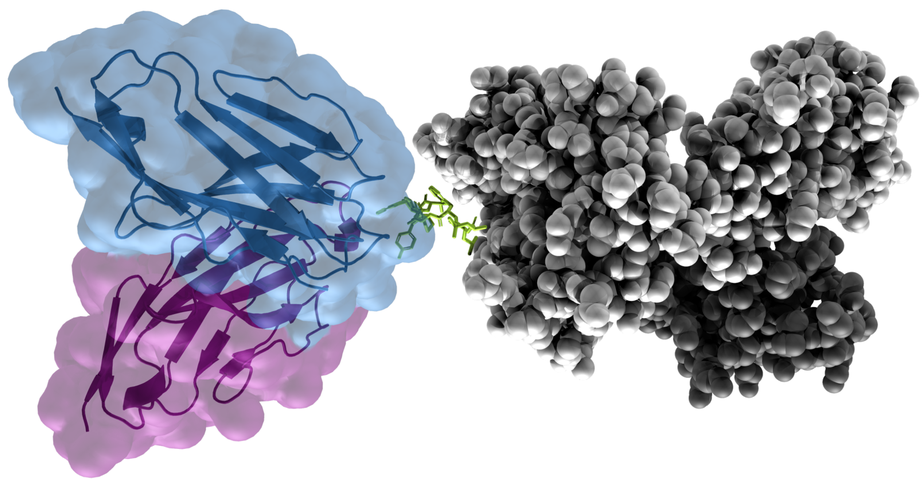

Figure 1. 3D-visualization of the TCR-pMHC complex. Blue and purple: TCR α and β chain, Green: Peptide, Grey: MHC.

Input data[1]

- Protein sequence of TCR, peptide, MHC (one-hot-encoding)
- Per-residue energy terms (one value per row)
- Global energy terms (constant, one value per column)

Input dimensions

- 6913 observations (4180 training, 1526 validation, 1207 test)
- 419 peptide positions (zero-padding where sequences are shorter)
- 54 channels (including sequence embedding and energy terms)

Methods

Pre-processing with protein embedding

- **BLOSUM (BLOcks SUbstitution Matrix):** Captures the biochemical properties of amino acids. It turns one-hot encoding into a non sparse vector.
- **ESM (Evolutionary Scale Modeling)[2]:** A transformer, i.e. a series of blocks that alternate self-attention with feed-forward connections. It has been trained beforehand with 250 million sequences and has 650 million weights. The output is a vector of size 1280 for each amino acid position.

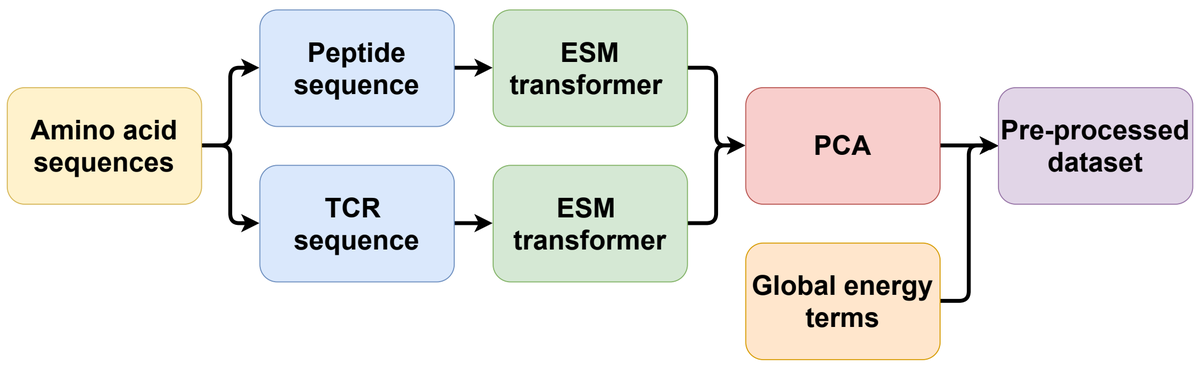

Figure 2. Pipeline for the pre-processing of the data.

Neural network architecture and training

- Early stopping, Adam optimizer with weight decay
- **Improvements:** More dense layers, division into local and global features, more drop out, additional batch normalization

Figure 3. Architecture of the neural network. 2 convolutional layers with max pooling followed by Long short-term memory (LSTM) and dense feed forward neural network.

Results

Network architecture	AUC	MCC	Precision	Recall	F1
Vanilla architecture	0.82	0.473	0.572	0.658	0.612
Improved architecture	0.86	0.469	0.5	0.794	0.614
Improved architecture with BLOSUM encoding	0.87	0.559	0.623	0.738	0.676
Improved architecture with ESM encoding	0.87	0.584	0.678	0.700	0.689

Table 1. Comparison of the performance on test set between different data preprocessing methods and network architectures.

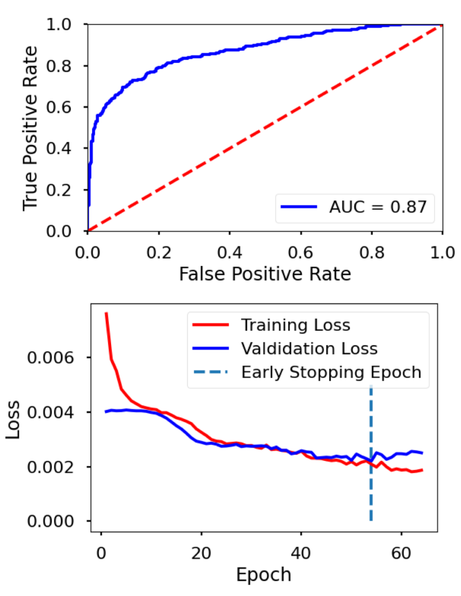

Figure 4. Evaluation of training and performance of improved architecture model with ESM encoding. Top: Receiver operating characteristic (ROC) curve on the test set. Bottom: Cross entropy loss of training and validation set and point of early stopping.

	Predicted: Negative	Predicted: Positive
Actual: Negative	806	100
Actual: Positive	90	211

Table 2. Confusion matrix on test set for model with improved architecture and ESM encoded

- Both changes in network architecture and protein sequence embeddings improve prediction of binding
- The highest MCC is obtained by setting the threshold to 0.7

Discussion

- (Challenge of imbalance of dataset) The dataset is really skewed towards negative (75%) which makes metrics like accuracy less reliable. Changing the threshold gives a better MCC because the model tends to predict more negatives. The best threshold is found to be 0.63 during validation and is applied for testing.
- The model overfits the dataset. If we were to predict binding for other peptides, the model would likely fail. This can be shown by leave-one-out cross-validation[3].
- TCR-BERT "outlook" embeddings work well → would be good to have TCR-specific embeddings (e.g. transfer learning) (like in preprint) but data availability is major challenge[4]. The ESM Transformer is trained using evolutionary sequences. To improve our model a TCR-specific transformer such as TCR-BERT could be used to generate TCR-specific embeddings[4]. These embeddings should be more specific for our problem than ESM.

References

[1]Magnus H. Høie. (2021). T-cell binding prediction challenge (TCR-pMHC). Github repository, https://github.com/CBH2021/tcr-pmhc
[2]Rives, A. et al. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences, 118(15).*
[3]Ida Kristine Sandford Meitil. (2021). Using deep learning for improving TCR homology modeling and its application to immunogenicity prediction [Master's Thesis, DTU]
[4]Wu, K. et al. (2021). TCR-BERT: learning the grammar of T-cell receptors for flexible antigen-xbinding analyses. *bioRxiv*.