TCR-epitope binding affinity prediction using multi-head self attention model

Michael Cai [1] **Seojin Bang** [2] **Heewook Lee** [1]

Abstract

TCR-epitope binding is the key mechanism for T cell regulation. Computational prediction of whether a given pair binds is of great interest for various clinical applications as well as understanding the underlying biological mechanisms of binding. Previously developed methods do not account for interrelationship between amino acids and suffer from poor out-of-sample performance. Our model uses the multi-head self attention mechanism to capture biological contextual information and to improve its generalization. We show that our model outperforms other models and we also demonstrate that the use of attention matrices can improve out-of-sample performance on recent SARS-CoV-2 data.

1. Introduction

The hallmark of the adaptive immune system is the T cells' ability to distinguish foreign invaders from host cells. T cells carry out this important task by utilizing their surface protein complex, called the T cell receptor (TCR) to bind to foreign peptides presented by major histocompatibility complex (MHC) molecules (also known as HLA molecules if the host is human) on host cell surface. The part of a peptide that a TCR binds to is called an epitope and the ability to computationally infer the binding affinity of TCR-epitope is important both for understating the underlying mechanism of the binding and for clinical applications in immunotherapy (Mösch et al., 2019).

Recently, immunotherapy has established itself as a promising treatment option for cancer patients. Since cancer is a disease caused by many random genetic mutations, tumor cells produce "neoantigens", peptides that are different from those produced by a patient's healthy cells (Schumacher et al., 2019). Assessing which TCRs will bind to the epitopes on these neoantigens becomes an important question for immunotherapy treatments. Also, with the current pandemic of SARS-CoV-2, the value of rapid screening for suitable candidate TCRs which may bind to foreign peptides produced by pathogens has become more clear. The ability to find candidate TCRs allows for the quick development of adaptive treatment strategies for diseases which threaten public health.

Immunology researchers need to determine the particular neoantigens present in a patient's body. Once they have collected the patient specific neoantigens cognate TCRs need to be found or engineered. Furthermore screening becomes necessary to determine the immunogenicity of a treatment. This task can be challenging, as there are over 10^{10} rearrangements of the VDJ genes on TCRs. Manually testing each TCR against its required epitope becomes an infeasible solution (Lythe et al., 2016b). A computational solution becomes necessary for the sake of time and labor.

Known bindings between TCRs and epitopes have been documented, however the biological reasons behind their specificity remains unknown. TCRs can bind to multiple epitopes while epitopes can also bind to multiple TCRs. This makes the search space for this problem much more difficult as a single concrete example may either be the only binding that either member of the pair has or possibly one of several bindings.

2. Background

Deep learning based methods have proven successful at extracting features and patterns from a sufficiently sized dataset and generalizing their findings to other examples. The advent of public databases containing epitope-specific T cell receptor (TCR) sequences, such as VDJdb (Shugay et al., 2018) and IEDB (Vita et al., 2019) has opened a door to computational methods for determining the immunogenicity of given epitope sequences. Solutions such as NetTCR (Jurtz et al., 2018), TCRGP (Jokinen et al., 2019), ERGO (Springer et al., 2020), and TCRex (Gielis et al., 2019) have been proposed. NetTCR utilizes convolutional networks

*Equal contribution [1]Biodesign Institute, Arizona State University, Tempe, Arizona [2]School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania. Correspondence to: Michael Cai <mrcai@asu.edu>, Heewook Lee <Heewook.Lee@asu.edu>.

The 2021 ICML Workshop on Computational Biology.

on the interactions between TCRs and peptides presented by the most common human allele HLA-A*02:01. ERGO utilizes an LSTM and autoencoder model to build a unified model using the CDR3's of TCRs. TCRGP focuses on utilizing the TCRα and TCRβ regions to determine which CDRs are important for epitope recognition. TCRex utilizes a random forest model to build a series of decision trees for each epitope.

These methods suffer severely from several major problems. The first of which is the loss of positional and contextual information of TCR and epitope sequences in the models. Additionally, TCRGP and TCRex both propose to build an epitope-specific model for each epitope to predict binding affinity of given TCR to the epitope, making only models with a sufficient number of known cognate TCRs applicable. Most importantly, they result in poor out-of-sample performance. In this paper, we present a new model using multi-head self attention mechanism. It helps to learn biological contextual representation of TCR and epitope, and explains how our model attends each amino acids for determining the binding affinity. We compare the prediction performance of our model with existing baseline methods. Furthermore, we demonstrate how to use attention matrix to improve out-of-sample performance.

3. Method

3.1. Data

Our dataset consists of TCR-epitope pairs known to bind collected from VDJDB, McPAS, and IEDB in December 2020. The collected data was processed into a unified format and then filtered down to only contain MHC I epitopes and TCRβ sequences. Quality control filters were applied to the dataset resulting in 6,388 pairs sourced from VDJDB, 11,936 pairs from McPAS, and 169,223 pairs from IEDB. The databases' information did overlap with each other a significant amount and hence duplicate pairs were removed. After applying the aforementioned modifications, the dataset consisted of 150,008 binding TCR-epitope pairs. Of which 982 unique epitopes and 140,675 unique TCRs were observed. Since TCR-epitope pairs known to not bind are not readily available, we generated negative data using random recombination. New unique pairs were created utilizing the same epitope distribution as the positive pairs. The final dataset consists of a 1:1 ratio of positive and negative data for training, validation, and testing.

An additional dataset of 332 TCR-epitope binding pairs was sourced from IEDB. This dataset contained two unique SARS-CoV-2 epitopes which were not present in either the training or testing set of data. The first epitope, YLQPRTFLL, had 304 cognate TCRs while the other epitope, RLQSLQTYV, had 28 cognate TCRs.

3.2. Training and Test Set Split

Epitope and TCR sequence spaces are extremely large. Novel epitope sequences continuously rise in nature due to new viral strains or mutations and a human body harbors about 10^{10} distinct TCR clonotypes (Lythe et al., 2016a) produced by random genetic recombination.

Therefore, it is of interest to accurately predict binding affinities for TCRs and epitopes that have never been seen before. In order to simulate this we designed two strategies for splitting the data into training and testing sets of TCR and epitope pairs. We use these strategies to evaluate our model's performance and ability to generalize on unseen TCRs and epitopes.

- **TCR Split**: The data was split such that any TCRs that are in the testing set are not found in the training set. This aims to evaluate prediction performance on out-of-sample TCRs.
- **Epitope Split**: The data was split such that any epitopes that are in the testing set are not found in the training set. This aims to evaluate prediction performance on out-of-sample epitopes.

A random sampling method was not used due to most TCRs in the data being unique. A randomized split would often produce splits similar to those of the TCR Split method.

3.3. Multihead-Attention Model

Our binding affinity prediction model consists of two encoders encoding TCR and epitope sequences separately and a linear decoder determining the affinity between the two sequences (Figure 1). The most distinctive characteristic of our model is that the two encoders use multi-head self attention mechanism (Vaswani et al., 2017). The multi-head self attention mechanism selectively attends amino acids based on how strongly amino acids in the sequence are correlated with each other. It helps to learn biological contextual representation of TCR and epitope sequences whose structure and function are determined by how amino acids are arranged and correlated with each other.

In the training process, each TCR and epitope sequence is fed into an initial embedding layer to obtain embeddings with the size p and the sequence lengths p_T and p_E. The sequence embeddings $\mathbf{T} \in \mathcal{R}^{p_T \times p}$ and $\mathbf{E} \in \mathcal{R}^{p_E \times p}$ are then fed into corresponding encoders. The attention mechanism in the two encoders f_t and f_e processes TCR and epitope sequences separately to learn a contextual representation of each. For this, each encoder quantifies the amount of attention by learning the strength of linear relationships of each amino acid with all others. The new representation is given by a linear average of input representations weighted

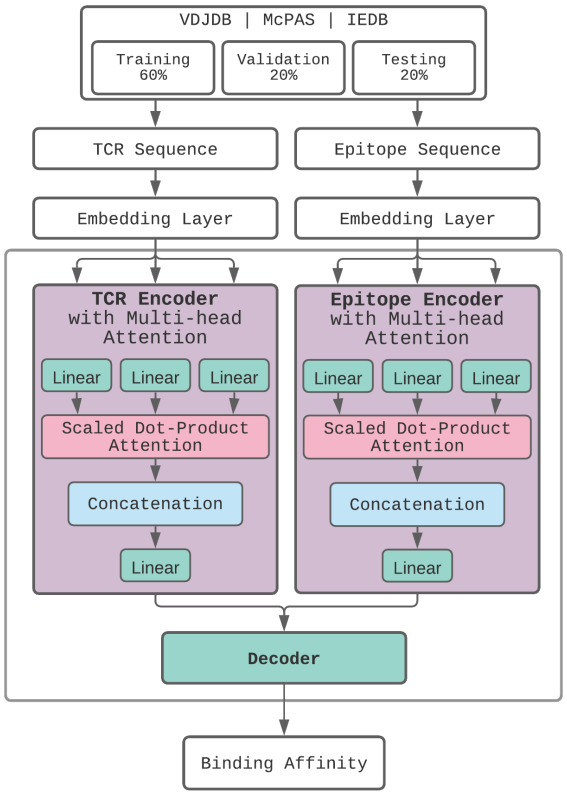

Figure 1. A depiction of the TCR-epitope binding affinity prediction model using the multi-head attention mechanism.

by the strength of relationship. In detail, a TCR sequence $\mathbf{T}$ is fed into three linear layers returning (linear-transformed) *key* ($\mathbf{K}$), *query* ($\mathbf{Q}$), and *value* ($\mathbf{V}$) matrices as follows:

$$\mathbf{K} = \mathbf{T}\mathbf{W}^K,\ \mathbf{Q} = \mathbf{T}\mathbf{W}^Q,\ \mathbf{V} = \mathbf{T}\mathbf{W}^V$$

The relationship strength between i-th amino acid and the others, denoted as $\mathbf{w}_i$, is determined by the scaled dot-product of the i-th row of $\mathbf{Q}$ with all rows of $\mathbf{K}$ as follows:

$$\mathbf{w}_i = Softmax\left(\frac{\mathbf{q}_i\mathbf{K}^T}{\sqrt{p_T}}\right)$$

where $\mathbf{q}_i$ is the i-th row of $\mathbf{Q}$. The contextual representation of i-th amino acid is then defined as a linear sum of all amino acid vectors weighted by $\mathbf{w}_i$.

$$\mathbf{t}_i^* = \mathbf{w}_i\mathbf{V} = \mathbf{w}_{i1}\mathbf{v}_1 + \cdots + \mathbf{w}_{ip_t}\mathbf{v}_{p_t},$$
$$\mathbf{T}^* = \mathbf{W}\mathbf{V}$$

where $\mathbf{w}_i$ is the i-th row of $\mathbf{W}$. Each element of $\mathbf{w}_i$ can be interpreted as an importance score of each amino acid in the sequence for determining the new representation of the i-th amino acid. Similarly, an epitope sequence is processed through corresponding encoder to obtain a new representation matrix. Such attention mechanism is called self-attention. We spread out cognitive load of attention mechanism by concatenating and passing multiple self-attention outputs through a single dense layer. Finally, output of the multi-head attention layer forms expected dimensions of TCR and epitope representation ($\mathbf{T}^*$ and $\mathbf{E}^*$).

Then, the decoder f_d determines binding affinity between the outputs of which, where the two encoded sequences are concatenated and fed into a dense layer of several linear transformations. The output of decoder is fed into a Sigmoid activation function to receive a binding affinity score between 0 and 1 as follows:

$$Score(\mathbf{T}, \mathbf{E}) = \frac{1}{1 + exp(-f_d(\mathbf{T}^*, \mathbf{E}^*))}$$

The binding affinity score is rounded to the closest integer value to receive either 0 or 1 indicating a positive or negative predicted binding respectively. We optimized the model with the following hyperparameter search space (bold indicate the choice for our final model): embedding matrix initialization – {Blosum45, Blosum50, Blosum62, **None**}, TCR and epitope sequence padding type – {Left, **Middle**, Right}, dropout rate for dense layer – $\{\mathbf{30}\%, 50\%\}$, the maximum length for TCR – $\{10, 15, \mathbf{20}\}$, the maximum length of epitope – $\{10, \mathbf{15}, 20\}$, and the dimensions of the dense layer – $\{128, 256, 512, \mathbf{1024}\}$. They are tuned via grid search and learned via Adam algorithm (Kingma & Ba, 2014). The best performing hyperparameter set was determined using 5-fold nested cross validation. The other models were trained and tested using their best performing hyperparameters as reported by their corresponding literature.

Table 1. Prediction performance of TCR-epitope binding affinity prediction models. Average and standard deviation of AUC over the 5-fold test sets has been reported.

MODEL	AUC	
	TCR SPLIT	EPITOPE SPLIT
OUR MODEL	**76.38**% (± 0.14%)	53.92% (± 5.23%)
NETTCR	73.43% (± 0.57%)	54.65% (± 4.03%)
ERGO - LSTM	74.70% (± 0.24%)	51.78% (± 4.98%)

4. Results

4.1. Prediction Performance

We evaluated our performance for the TCR-epitope binding affinity prediction task on both the TCR split and epitope split. Table 1 shows that our method significantly outperforms the baseline models in the TCR split. However, there is no significant difference in the epitope split. The model

which was trained on the TCR split data was additionally evaluated against the previously mentioned SARS-CoV-2 dataset. The model was able to predict the binding affinities with an individual epitope recall of 65.13% for YLQPRTFLL and 57.14% for RLQSLQTYV. The overall recall on the dataset was 64.26%.

Notably, all models perform better on the task using the TCR split compared to the same task on the epitope split. Although most TCRs within the dataset only appear once, they share many features with other TCRs present. It is common for an epitope in the dataset to bind to two TCRs which differ by only one or two changes in its amino acid sequence. The epitope sequences do not have this same overlap of features and motifs, most sequences in the dataset are fairly unique and so it is difficult to extrapolate the information they provide to another epitope. The epitope split ensures that an epitope in the testing set has never been seen by the model before, and so it struggles to make an accurate prediction. The near 50% AUC on the epitope split with high variance for all models indicates that a key feature is still missing from existing models.

Three hyperparameters had the greatest effect on model performance: (1) the size of the linear transformations in the dense layer, (2) number of heads in the attention layer, and (3) the maximum sizes of the sequences. Our attention based model benefited from having length bounds on sequences. Reducing either the TCR or epitope sequence size below 15 amino acids, however was very detrimental to model performance. We also observed that initializing the embedding with a blosum matrix often had little to no effect compared to utilizing a random initialization. No single blosum matrix consistently out-performed the others. Even in cases where a model with a blosum matrix performed better than a random initialization the increase in performance was negligible.

4.2. Attention matrix

Our model, as well as the baseline models, showed limited performance on the epitope split and out-of-sample data (SARS-CoV-2), indicating there is still an unknown key feature that determines binding affinity. In this section, we illustrate that attention matrix $\mathbf{W}$ can help us further improve the prediction performance using the SARS-CoV-2 epitope YLQPRTFLL as an example.

The SARS-CoV-2 epitope YLQPRTFLL and YYVGYLQPRTFLL, a training set epitope, both share a subsequence and bind to similar TCR sequences (Figure 2). Our idea is based on the assumption that TCRs binding to similar epitopes would also share similar inter-relationships between amino acids, and hence the model's attention mechanisms for those TCRs would be similar. In order to validate this assumption, we generated reference attention matrices

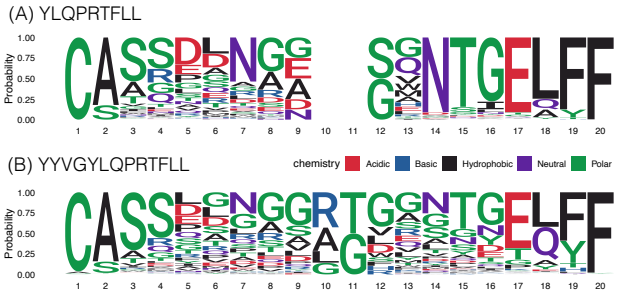

Figure 2. Sequence logos of TCRs binding to epitope sequences (A) YLQPRTFLL and (B) YYVGYLQPRTFLL

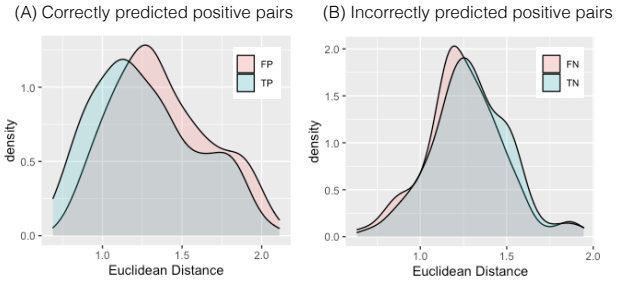

Figure 3. Distribution of the distance between (A) the correctly predicted TCRs and the FP/TP reference matrices, and (B) the incorrectly predicted TCRs and the FN/TN reference matrices.

by averaging TCR attention matrices of (1) true positive (TP), (2) false positive (FP), (3) true negative (TN), and (4) false negative (FN) predictions for the training epitope YYVGYLQPRTFLL. We then compared Euclidean distances between each TCR attention matrix of YLQPRTFLL and the reference attention matrices. We observed that the positively (i.e. correctly) predicted TCR-YLQPRTFLL pairs have attentions significantly closer to the true positive than the false positive reference attention (p-value $< 2.2 \times 10^{-16}$ from paired t-test). Similarly, the negatively (i.e. incorrectly) predicted TCR-YLQPRTFLL pairs have attentions significantly closer to the false negative than the true negative reference attention (p-value $< 2.2 \times 10^{-16}$ from paired t-test). We also observed none of the distances is significantly correlated with the model scores (correlation between 0.05–0.23), indicating that the distance to each reference matrix can be a distinct measure for evaluating the confidence of model predictions.

5. Conclusion

We have developed TCR-epitope binding affinity prediction model using multi-head self attention mechanism. This mechanism helps us to learn biological contextual representation of TCR and epitope sequences by correlating amino acids within each sequence. We showed that our model outperform the baseline models in the TCR split, and demonstrated that attention matrix can be used as a confidence score for a prediction.

References

Gielis, S., Moris, P., Bittremieux, W., De Neuter, N., Ogunjimi, B., Laukens, K., and Meysman, P. Detection of enriched T cell epitope specificity in full T cell receptor sequence repertoires. *Frontiers in Immunology*, 10:2820, 2019.

Jokinen, E., Huuhtanen, J., Mustjoki, S., Heinonen, M., and Lähdesmäki, H. Determining epitope specificity of T cell receptors with TCRGP. *bioRxiv*, pp. 542332, 2019.

Jurtz, V. I., Jessen, L. E., Bentzen, A. K., Jespersen, M. C., Mahajan, S., Vita, R., Jensen, K. K., Marcatili, P., Hadrup, S. R., Peters, B., et al. NetTCR: sequence-based prediction of TCR binding to peptide-MHC complexes using convolutional neural networks. *bioRxiv*, pp. 433706, 2018.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Lythe, G., Callard, R. E., Hoare, R. L., and Molina-París, C. How many TCR clonotypes does a body maintain? *Journal of Theoretical Biology*, 389:214–224, 2016a.

Lythe, G., Callard, R. E., Hoare, R. L., and Molina-París, C. How many tcr clonotypes does a body maintain? *Journal of Theoretical Biology*, 389:214–224, 2016b. doi: 10.1016/j.jtbi.2015.10.016.

Mösch, A., Raffegerst, S., Weis, M., Schendel, D. J., and Frishman, D. Machine learning for cancer immunotherapies based on epitope recognition by T cell receptors. *Frontiers in Genetics*, 10, 2019.

Schumacher, T. N., Scheper, W., and Kvistborg, P. Cancer neoantigens. *Annual Review of Immunology*, 37:173–200, 2019.

Shugay, M., Bagaev, D. V., Zvyagin, I. V., Vroomans, R. M., Crawford, J. C., Dolton, G., Komech, E. A., Sycheva, A. L., Koneva, A. E., Egorov, E. S., et al. VDJdb: a curated database of T-cell receptor sequences with known antigen specificity. *Nucleic Acids Research*, 46(D1): D419–D427, 2018.

Springer, I., Besser, H., Tickotsky-Moskovitz, N., Dvorkin, S., and Louzoun, Y. Prediction of specific tcr-peptide binding from large dictionaries of tcr-peptide pairs. *Frontiers in Immunology*, 11:1803, 2020. ISSN 1664-3224. doi: 10.3389/fimmu.2020.01803. URL https://www.frontiersin.org/article/10.3389/fimmu.2020.01803.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. *CoRR*, abs/1706.03762, 2017. URL http://arxiv.org/abs/1706.03762.

Vita, R., Mahajan, S., Overton, J. A., Dhanda, S. K., Martini, S., Cantrell, J. R., Wheeler, D. K., Sette, A., and Peters, B. The immune epitope database (IEDB): 2018 update. *Nucleic Acids Research*, 47(D1):D339–D343, 2019.