

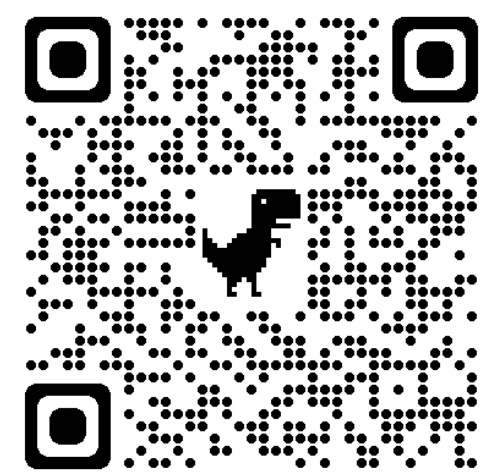
# PiTE: TCR-epitope Binding Affinity Prediction Pipeline using Transformer-based Sequence Encoder



Pengfei Zhang<sup>1,2</sup>, Seojin Bang<sup>2</sup> and Heewook Lee<sup>1,2,\*</sup>

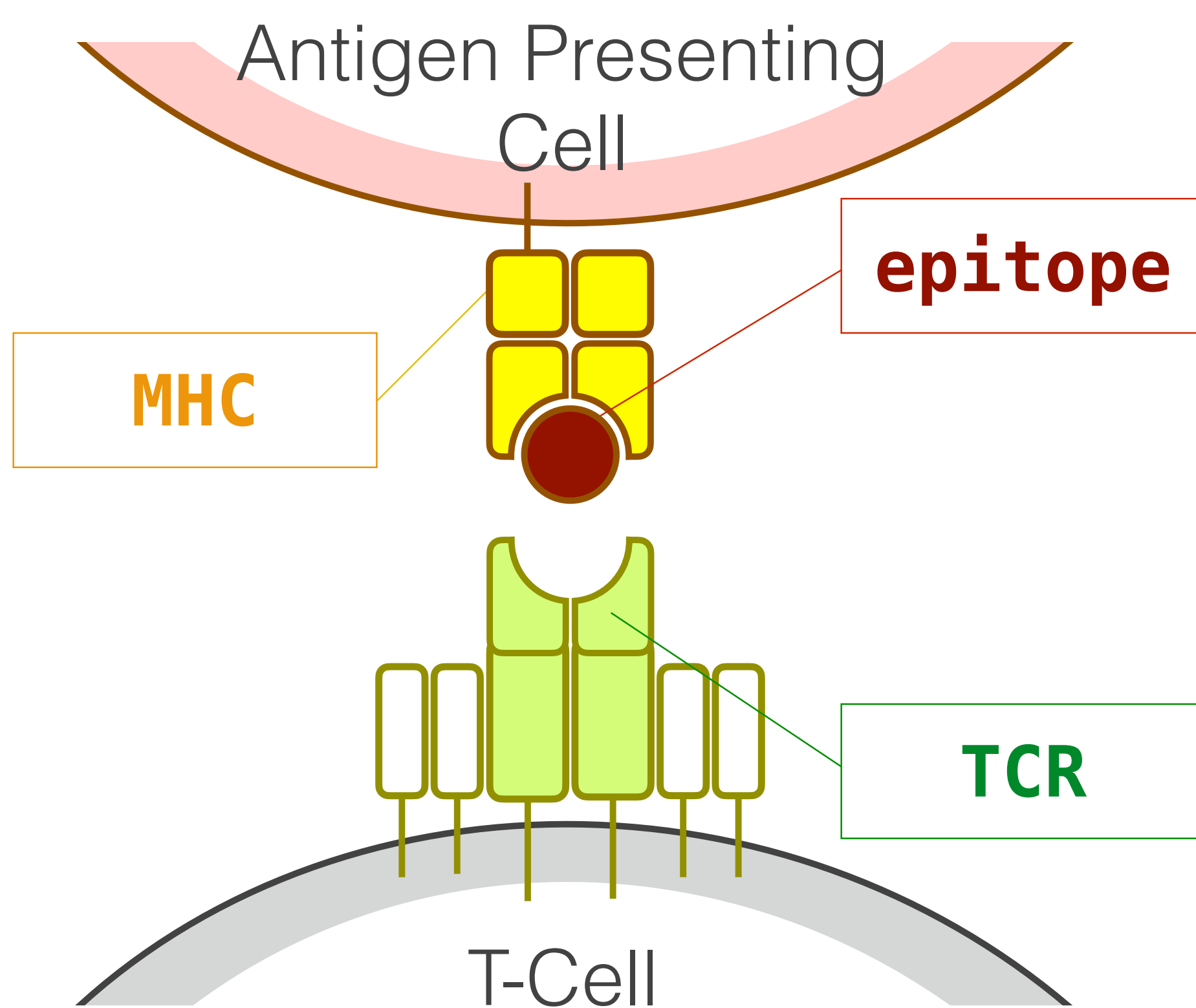
<sup>1</sup>School of Computing and Augmented Intelligence, Arizona State University

<sup>2</sup>Biodesign Institute, Arizona State University



## INTRODUCTION

T cell receptor (TCR) is critical in adaptive immune systems. It can bind to MHC complex and determine whether the presented antigen is foreign or not.



**The Computational Problem:**

To improve the binding affinity prediction between TCR and epitope sequences.

**Applications:**

- Vaccine development
- Personalized immunotherapy
- Cancer treatment

## DATA & SPLIT

**Binding Pairs:** 150,008 pairs collected from publicly available databases. Used as positive data

**Non-binding Pairs:** 150,008 pairs generated by pairing healthy TCRs to epitopes. Used as negative data.

**TCR Split:** designed to measure out-of-sample TCR performance.

**Epitope Split:** designed to measure out-of-sample epitopes performance.

## CONCLUSIONS

PiTE with Transformer-like sequence encoder establishes a new state-of-the-art performance for TCR-epitope binding affinity prediction.

## References

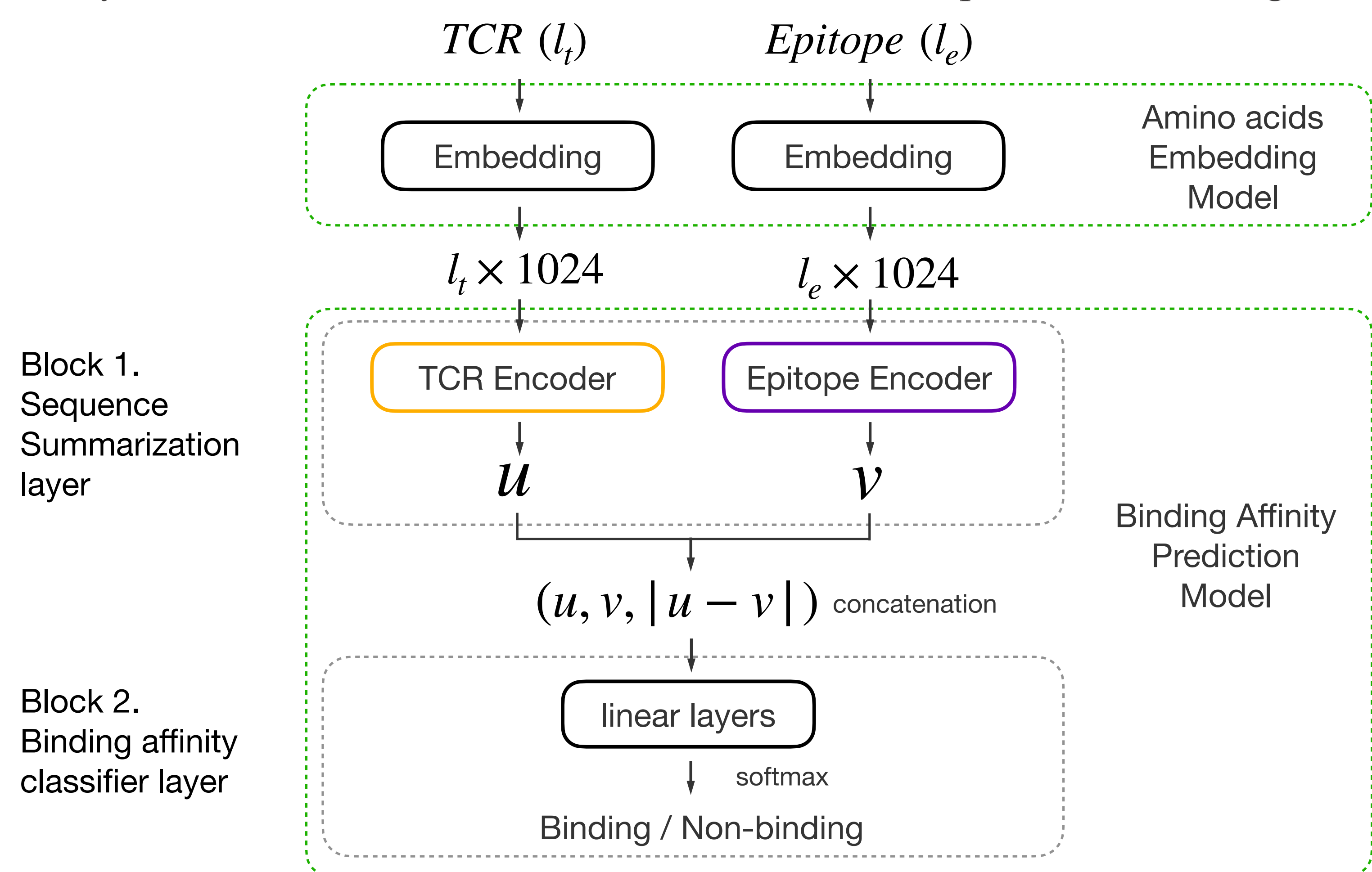
1. Cai, Michael, et al. ATM-TCR. In Front Immunol.
2. Zhang, Pengfei, et al. catELMo. In submission.
3. Tickotsky, Nili, et al. McPAS-TCR. In Bioinformatics.
4. Fleri, W., et al. IEDB. In Front Immunol.
5. Shugay, Mikhail, et al. VDJdb. In Nucleic acids research.

## OVERVIEW OF OUR METHOD

**Motivation:** Develop sequence encoders for TCR and epitope sequences to **better summarize amino acid level embedding to sequence level embedding** than average pooling method.

**Method:**

- 1) Amino acid level embedding is obtained from pre-trained catELMo embedding model, which has been shown as the current state-of-the-art embedding method in TCR analysis.
- 2) Sequence summarization layer encodes amino acid level embeddings into a sequence level embedding. The Transformer-based encoder can capture context-aware information because of its self-attention mechanism.
- 3) Linear layers with a softmax activation function is used to predict the binding affinity result.



## PiTE PERFORMANCE

- The Transformer-based sequence encoder significantly outperforms the rest three sequence encoders in TCR-epitope binding affinity prediction.
- The choice of model architecture can be more important than the number of model parameters.

