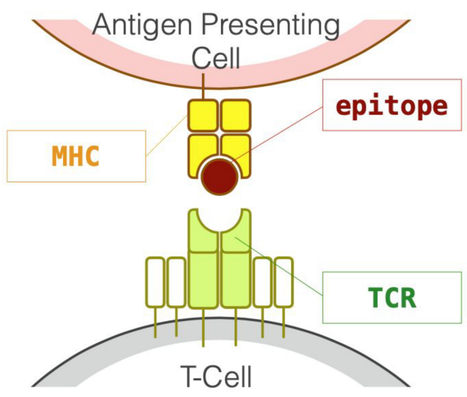
Identifying the Optimal Orientation for Selecting Embedding Models for TCR-Epitope Binding Affinity Prediction

Aiko Muraishi, Computer Science Major
Mentor: Heewook Lee, Assistant Professor
School of Computing and Augmented Intelligence

Objective & Research Question

Accurate TCR-epitope binding prediction is vital for personalized healthcare and immunotherapy. This study investigates why catELMo, a bidirectional LSTM model, outperforms TCRBert, a Transformer-based model, by exploring two potential reasons: structural suitability and variations in learning objectives. GPT and unidirectional LSTM structures were trained to identify the optimal choices of a model for TCR-epitope affinity prediction.



Background

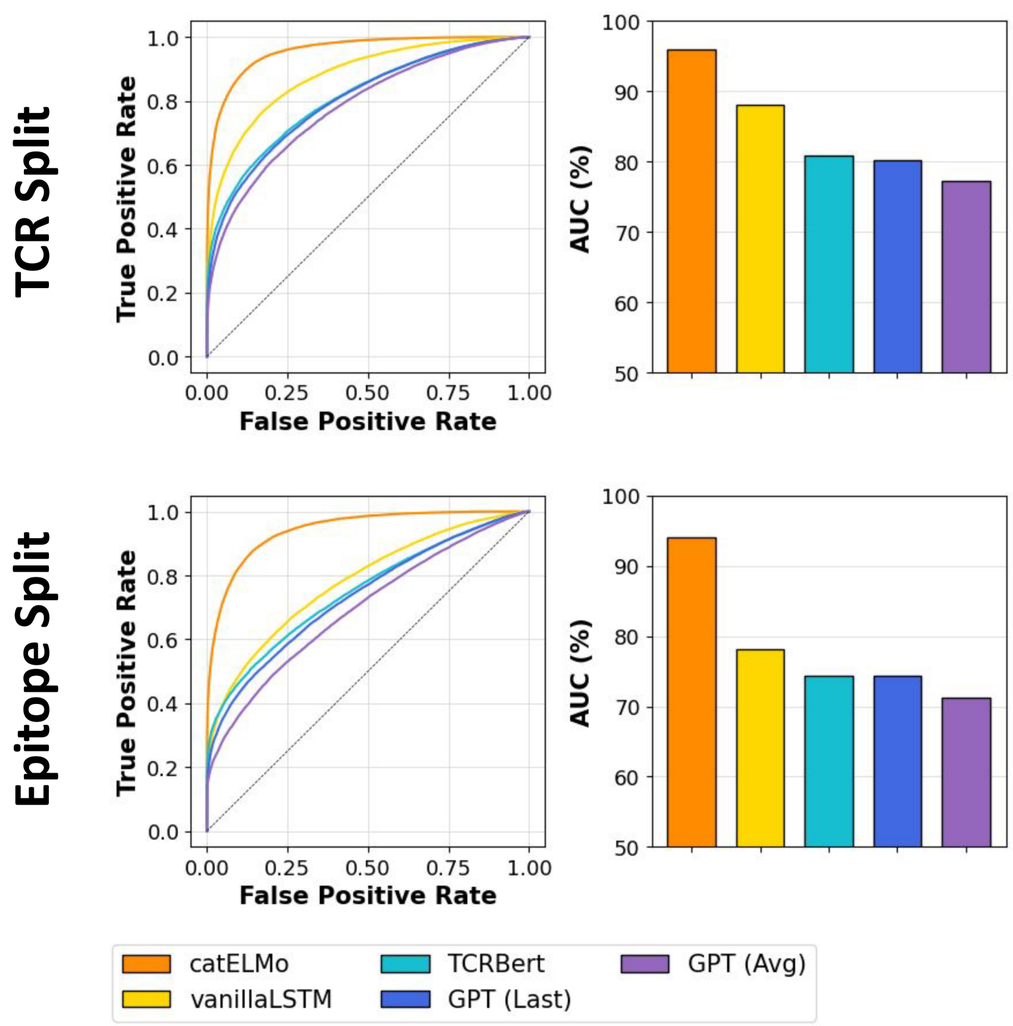
- The crucial role of T cell receptors (TCRs) in the adaptive immune system lies in their ability to facilitate killer T cells in distinguishing between abnormal cells and healthy cells.
- Using computational methods to predict their binding can significantly decrease both the cost and time required to refine a set of potential TCR targets, thereby expediting the advancement of personalized immunotherapy.
- While Transformer models, like TCRBert, have gained traction in Natural Language Processing, recent research highlights catELMo's superior accuracy in predicting TCR-epitope binding.

Table 1. Summary of Model Structures and Learning Objectives

	Model Structure	Objective Function	Performance (AUC %)
catELMo	Bidirectional LSTM	Next Word Prediction	94
TCRBert	Transformer Encoder	Masked Word Prediction	74
GPT	Transformer Decoder	Next Word Prediction	**We want to know this!**
Vanilla LSTM	Unidirectional LSTM	Next Word Prediction	**We want to know this!**

Results

- catELMo continues to outperform GPT.
- Unidirectional LSTM also shows superior performance compared to the Transformer models.
- The results indicate that the structure of LSTM seems to be more suitable for the TCR-epitope binding problem.
- A comprehensive investigation is needed that aims to unravel the reasons behind the effective performance of catELMo. Future work will include exploring different model structures, data characteristics and volumes, and model size (number of parameters).

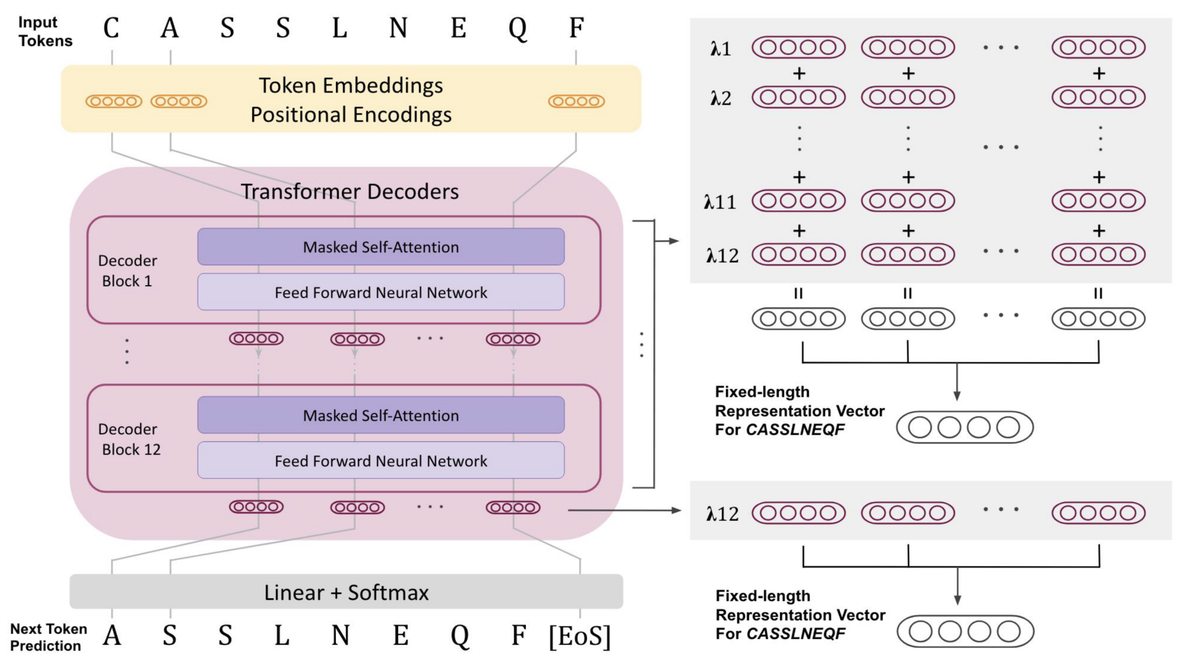


Methods

- GPT and vanilla LSTM were trained for 10 epochs with batch sizes of 256 and 128, respectively. The length of representation vectors is 768 (GPT) and 1024 (LSTM).
- The sequences of the third complementarity-determining region (CDR3) within the TCR-β chain were used for training due to the utmost significance in establishing its binding specificity to an epitope.
- To check the performance of the embedders, classification on binding/ non-binding was conducted using a simple three-layer Feedforward Network, inputting the obtained representation vectors of TCRs and epitopes.

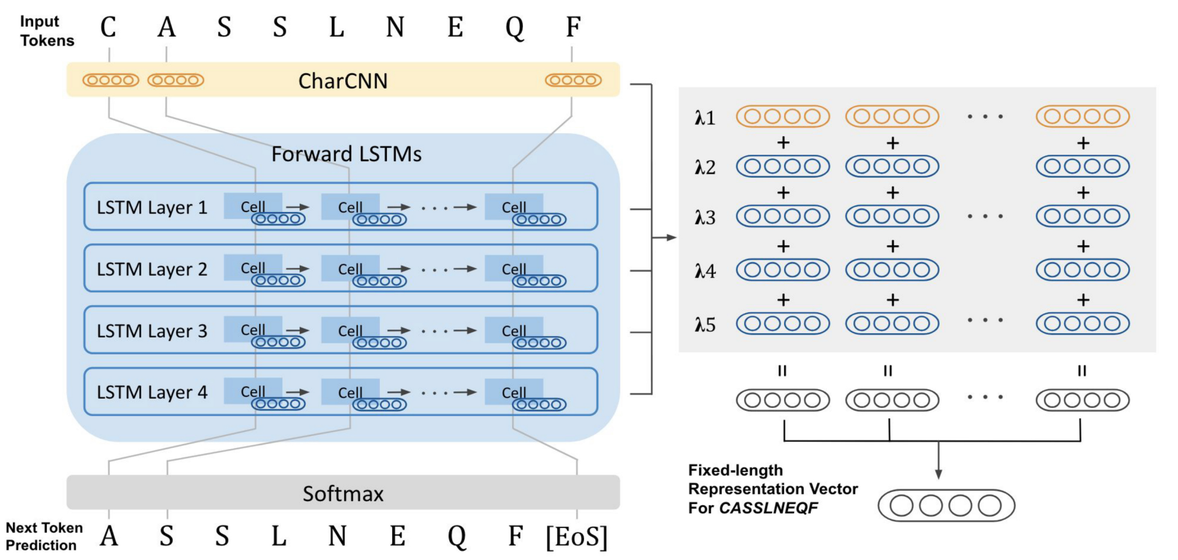
GPT Embedder

Amino acid sequence representation model using GPT structure



Vanilla LSTM Embedder

Amino acid sequence representation model using unidirectional LSTM



The representation vectors are used in the binary classification

Classification

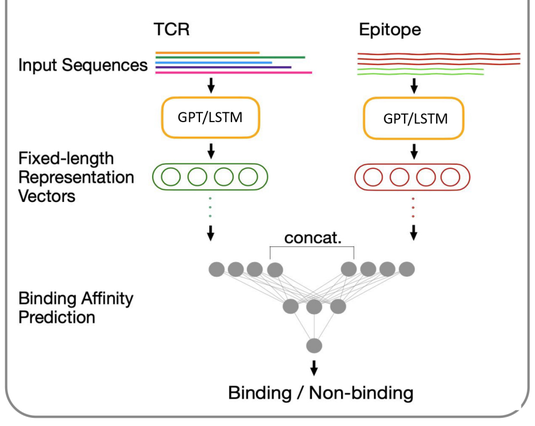


Table 2. Data Summary. The number of unique epitopes, TCRs, and TCR-epitope pairs used for the embedder training and binding affinity prediction.

Usage	Source	Unique Epitopes	Unique TCRs	Unique TCR-epitope Pairs	Amino Acid Tokens
Embedder Traning	ImmuneSEQ	X	4,173,895	X	52,546,029
Binding Prediction	VDJdb	187	3,915	4,047	X
	McPAS	301	9.822	10,156	X
	IEDB	1189	136,492	145,678	X
	Total (Duplicates removed)	982	140,675	150,008	X

References

- M. Attaf, M. Legut, D. K. Cole, and A. K. Sewell, "The T cell antigen receptor: the Swiss army knife of the immune system," *Clinical and Experimental Immunology*, vol. 181, no. 1, pp. 1–18, Jul. 2015.
- P. Moris *et al.*, "Current challenges for unseen-epitope TCR interaction prediction and a new perspective derived from image classification," *Briefings in Bioinformatics*, vol. 22, no. 4, Dec. 2020.
- K. Wu *et al.*, "TCR-BERT: learning the grammar of T-cell receptors for flexible antigen-binding analyses," *bioRxiv (Cold Spring Harbor Laboratory)*, Nov. 2021.
- Zhang Pengfei, Bang Seojin, Cai Michael, Lee Heewook, "Context-Aware Amino Acid Embedding Advances Analysis of TCR-Epitope Interactions", *eLife* 12:RP88837, 2023.
- M. M. Davis and P. J. Bjorkman, "T-cell antigen receptor genes and T-cell recognition," *Nature*, vol. 335, no. 6192, pp. 744–744, Oct. 1988.
- M. Krogsgaard and M. M. Davis, "How T cells 'see' antigen," *Nature Immunology*, vol. 6, no. 3, pp. 239–245, Feb. 2005
- M. Cai, S. Bang, P. Zhang, and H. Lee, "ATM-TCR: TCR-Epitope Binding Affinity Prediction Using a Multi-Head Self-Attention Model," *Frontiers in Immunology*, vol. 13, Jul. 2022.

FURI

Ira A. Fulton Schools of Engineering
Arizona State University