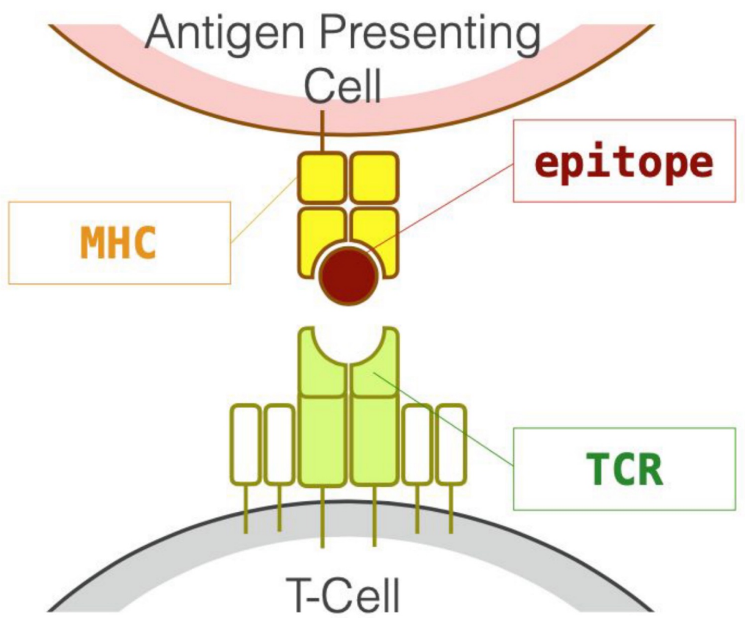
Towards Automated Selection of Embedding Models: Identifying the Optimal Parameters for the Baseline Model for TCR Embedding

Uttam Kumar, Computer Science
Mentor: Dr. Heewook Lee, Assistant Professor
School of Computing and Augmented Intelligence

Objective and Research Question



Analyzing T cell receptor (TCR)-epitope interactions is vital for identifying therapeutic targets, while TCR clustering reveals clonal expansion patterns, aiding intervention. Predicting TCR-epitope binding affinity helps screen TCRs against harmful antigens. Recent advances, like catELMo, enhance TCR tasks, yet its mechanisms are unclear.

This research is a large-scale study on TCR embeddings, focusing on optimizing catELMo parameters (e.g., learning rate, batch size, epochs). Despite transformer models, bidirectional Long Short-Term Memory (biLSTM)-based embeddings excel in prediction tasks. To grasp catELMo's success, a comparative study on TCR embeddings is proposed, focusing on optimizing baseline model parameters due to the study's scale.

Background

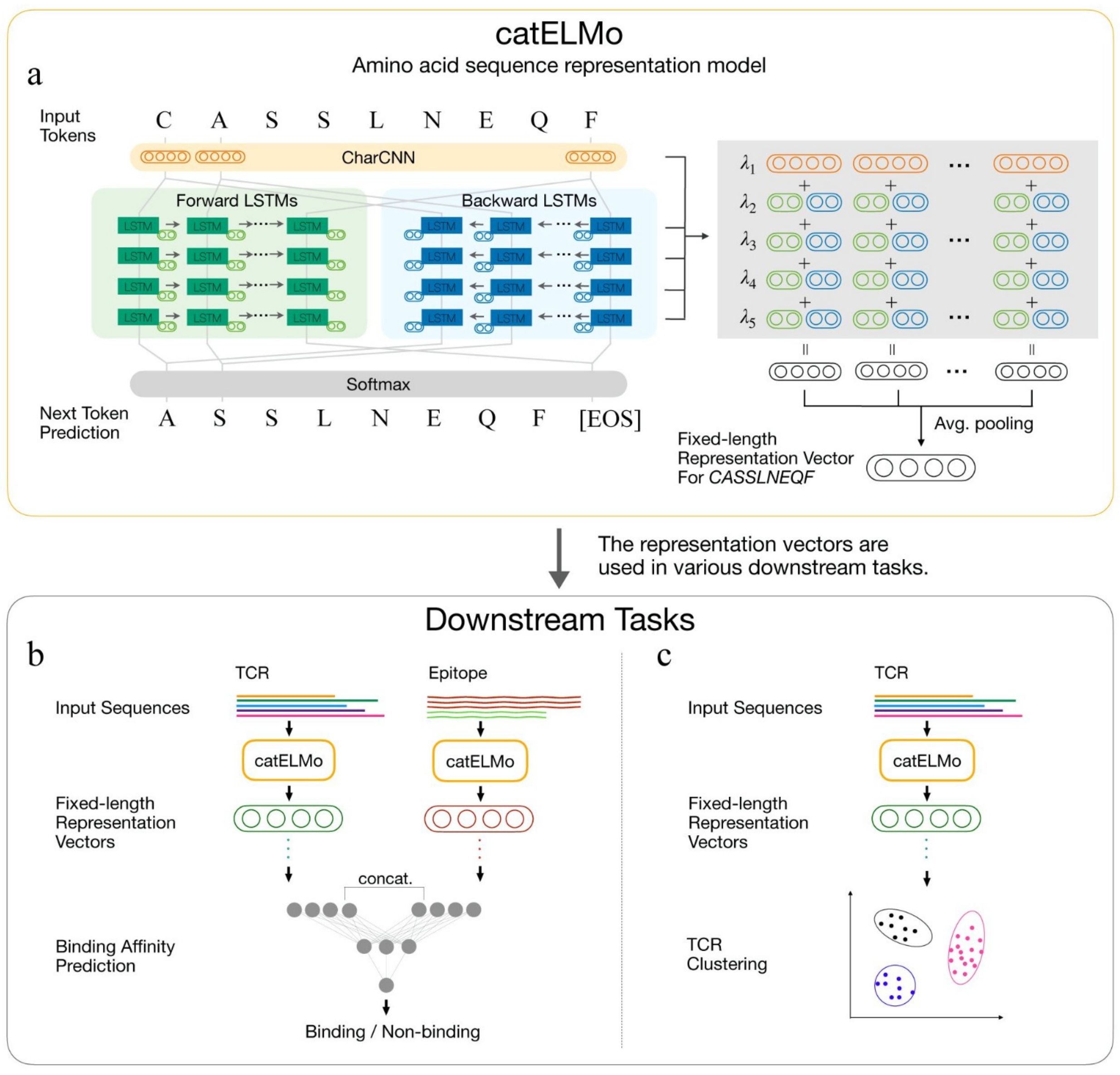
- The crucial role of the T cell receptors (TCRs) in the adaptive immune system lies in their ability to facilitate killer T cells in distinguishing between abnormal cells and normal cells.
- Using computational methods to predict their binding can significantly decrease both the cost and time required 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.

	AUC (%)	Precision (%)	Recall (%)	F1 (%)
BLOSUM62	$\underline{82.03 \pm 0.25}$	67.16 ± 1.01	$\underline{82.04 \pm 1.01}$	70.57 ± 0.73
Yang et al.	75.03 ± 0.20	62.54 ± 0.78	79.71 ± 1.45	65.22 ± 0.69
ProtBert	77.86 ± 0.29	70.01 ± 1.47	69.90 ± 2.65	69.85 ± 0.41
SeqVec	81.61 ± 0.21	69.30 ± 1.33	79.02 ± 2.02	71.75 ± 0.66
TCRBert	80.79 ± 0.17	$\underline{74.19 \pm 1.17}$	70.48 ± 1.60	$\underline{72.89 \pm 0.23}$
catELMo (ours)	**96.04 ± 0.12**	**86.88 ± 0.92**	**91.83 ± 0.98**	**88.94 ± 0.21**
p-value	6.28×10^{-23}	1.94×10^{-15}	1.82×10^{-14}	1.29×10^{-29}

TCR-epitope binding affinity prediction performance of TCR split.

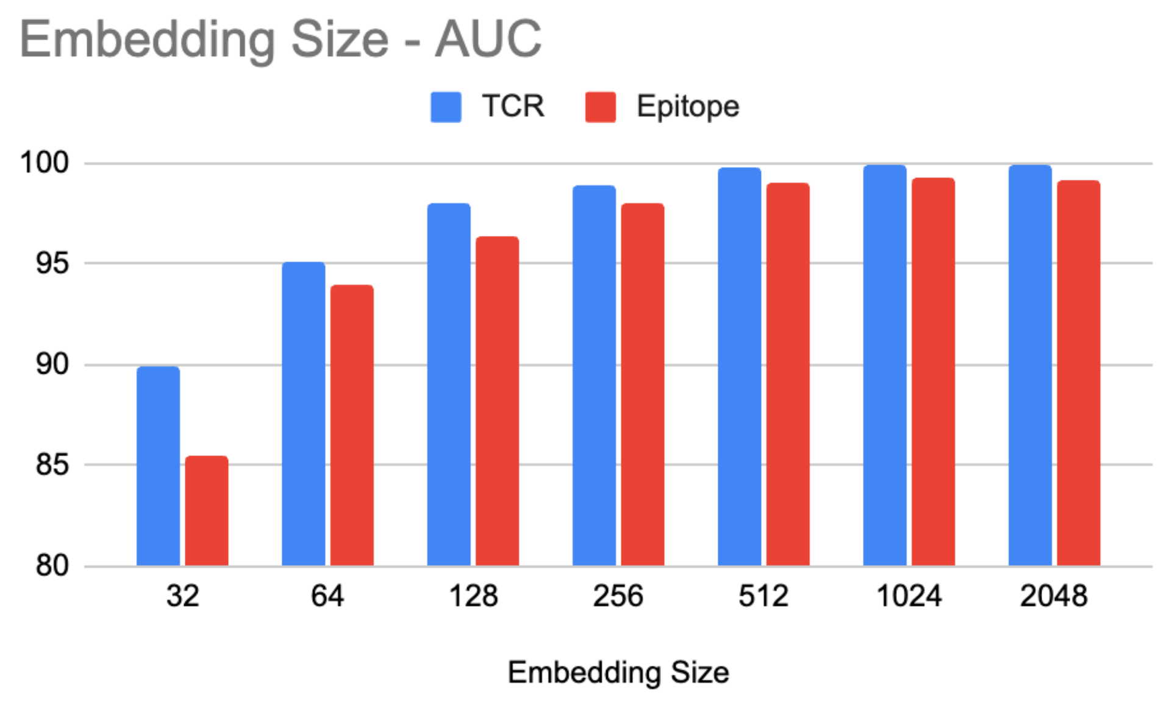
Methods

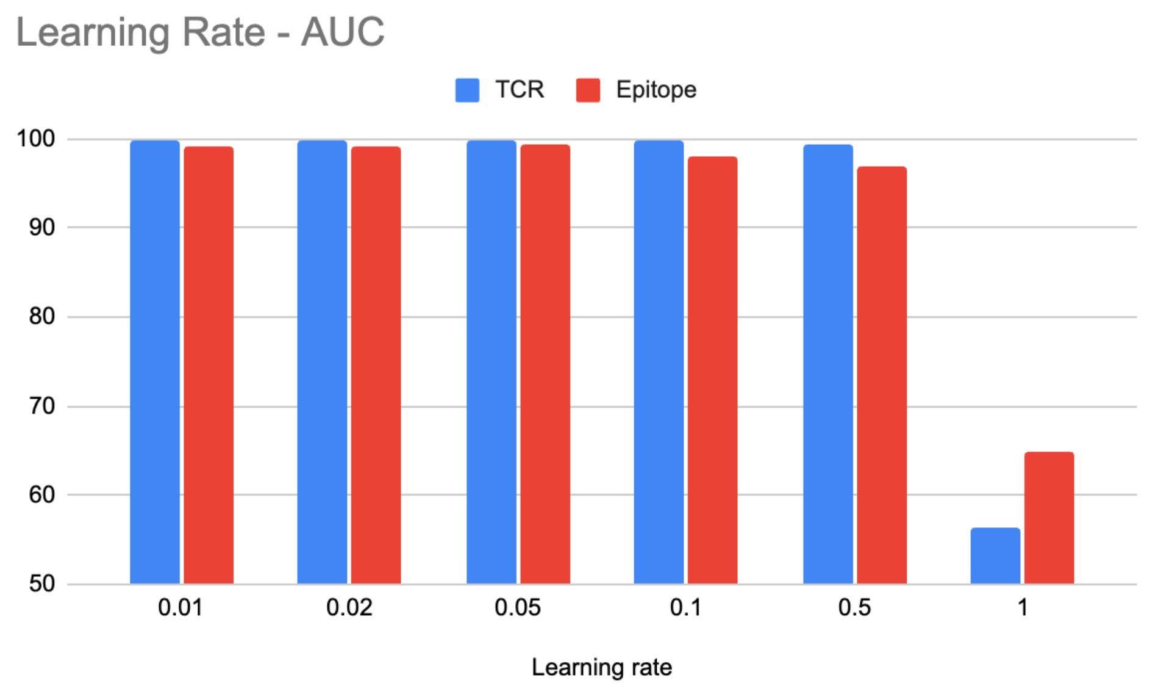
- Baseline model parameters were selected with Learning rate, Batch Size, Embedding Size, LSTM Layers and LSTM Dimensions.
- It has been trained on 4,173,895 TCR*β* CDR3 sequences (52 million of amino acid tokens) from ImmunoSEQ.
- Then from trained models , TCR-Epitope embeddings were extracted.
- We investigated and recorded the downstream performance of TCR-epitope binding affinity prediction models trained using these catELMo embeddings.

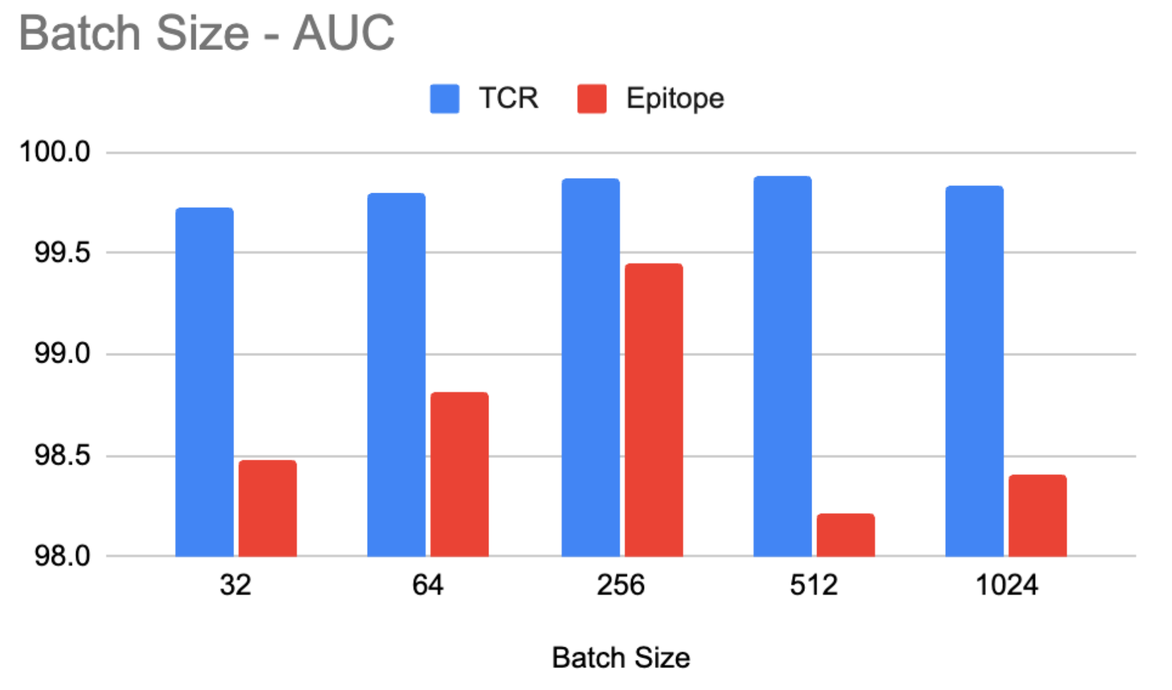


Results

- catELMo continues to outperform transformer models with further parameter tuning.
- Increasing the embedding size (1024, 2048) for the catELMo model improves performance.
- The batch size of 256 outperforms lower batch and higher batch sizes for the Epitope split.
- A learning rate close to 0.1 will have much better results compared to smaller or larger learning rates.







Future Works

- We look forward to further exploring other hyperparameters and how they affect the performance of the model.
- Further exploration of the model will also shed light on how the catELMo model performs much better than transformer models.

Acknowledgment

I extend heartfelt gratitude to Pengfei Zhang for his invaluable support and guidance throughout the research process. His expertise shaped the work's direction significantly, and I deeply appreciate his contributions. I also want to thank Aiko Muraishi for her dedication in resolving technical challenges and providing patient guidance daily. Her expertise ensured smooth progress, and I am truly grateful for her unwavering support.

References

- Zhang Pengfei, Bang Seojin, Cai Michael, Lee Heewook (2023) Context-Aware Amino Acid Embedding Advances Analysis of TCR-Epitope Interactions eLife 12:RP88837 https://doi.org/10.7554/eLife.88837.1
- Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). Scaling Laws for Neural Language Models. *ArXiv*. /abs/2001.08361
- Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. D., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., Driessche, G. V., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., . . . Sifre, L. (2022). Training Compute-Optimal Large Language Models. *https://doi.org/10.48550/arXiv.2203.15556*
- Cai, M., Bang, S., Zhang, P., & Lee, H. (2022). ATM-TCR: TCR-epitope binding affinity prediction using a multi-head self-attention model. Frontiers in Immunology, 13. 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., & Nielsen, M. (2018). NetTCR: sequence-based prediction of TCR binding to peptide-MHC complexes using convolutional neural networks https://doi.org/10.1101/433706
- Wu, K., Yost, K.E., Daniel, B., Belk, J.A., Xia, Y., Egawa, T., Satpathy, A., Chang, H.Y. and Zou, J., 2021. TCR-BERT: learning the grammar of T-cell receptors for flexible antigen-xbinding analyses. bioRxiv, pp.2021-11.