Understanding the Root Causes for Downstream Performance Improvements Using catELMo Embeddings Over Alternative Embeddings for the TCR-Epitope Binding Affinity Prediction Task

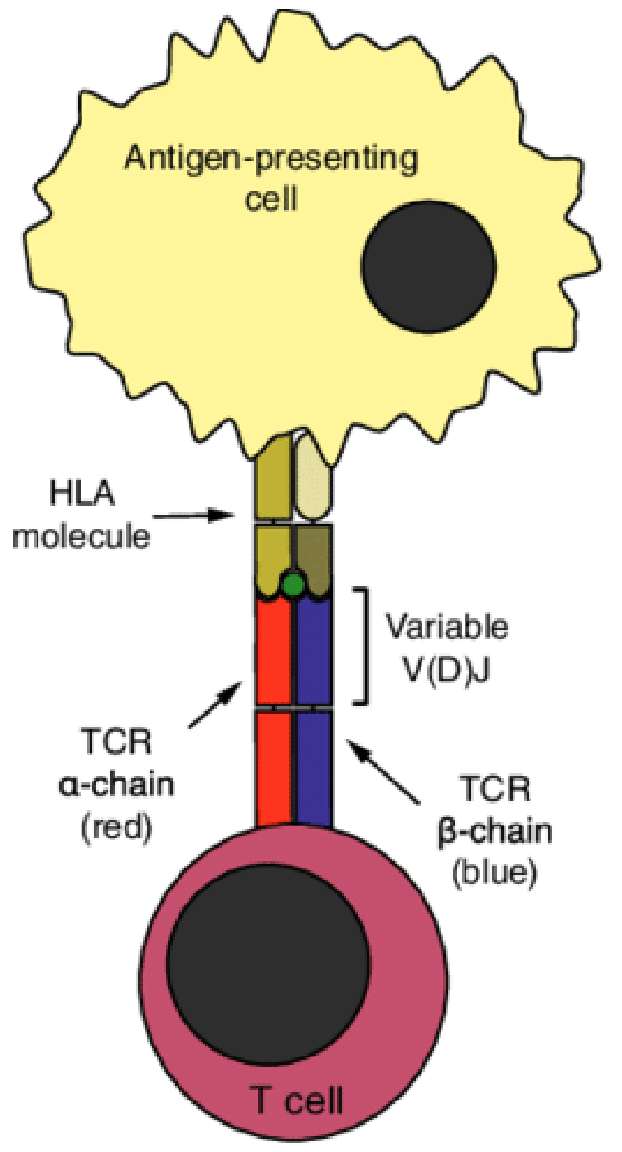
Ryan Connolly-Kelley, B.S. Computer Science
Mentor: Dr. Heewook Lee, Assistant Professor
School of Computing and Augmented Intelligence, and Biodesign Institute

Research Question

Why does using catELMo to embed T-Cell Receptors into a fixed-length vector representation yield better performance in downstream tasks than any other embedding method?

Background & Motivation

T Cells are responsible for binding to foreign antigens in the body and initiating an immune response. Specifically, the T-Cell Receptor (TCR) binds to the epitope of an antigen. This binding is many-to-many: one TCR may bind to several distinct epitopes, and one epitope may be bound by multiple distinct TCRs.



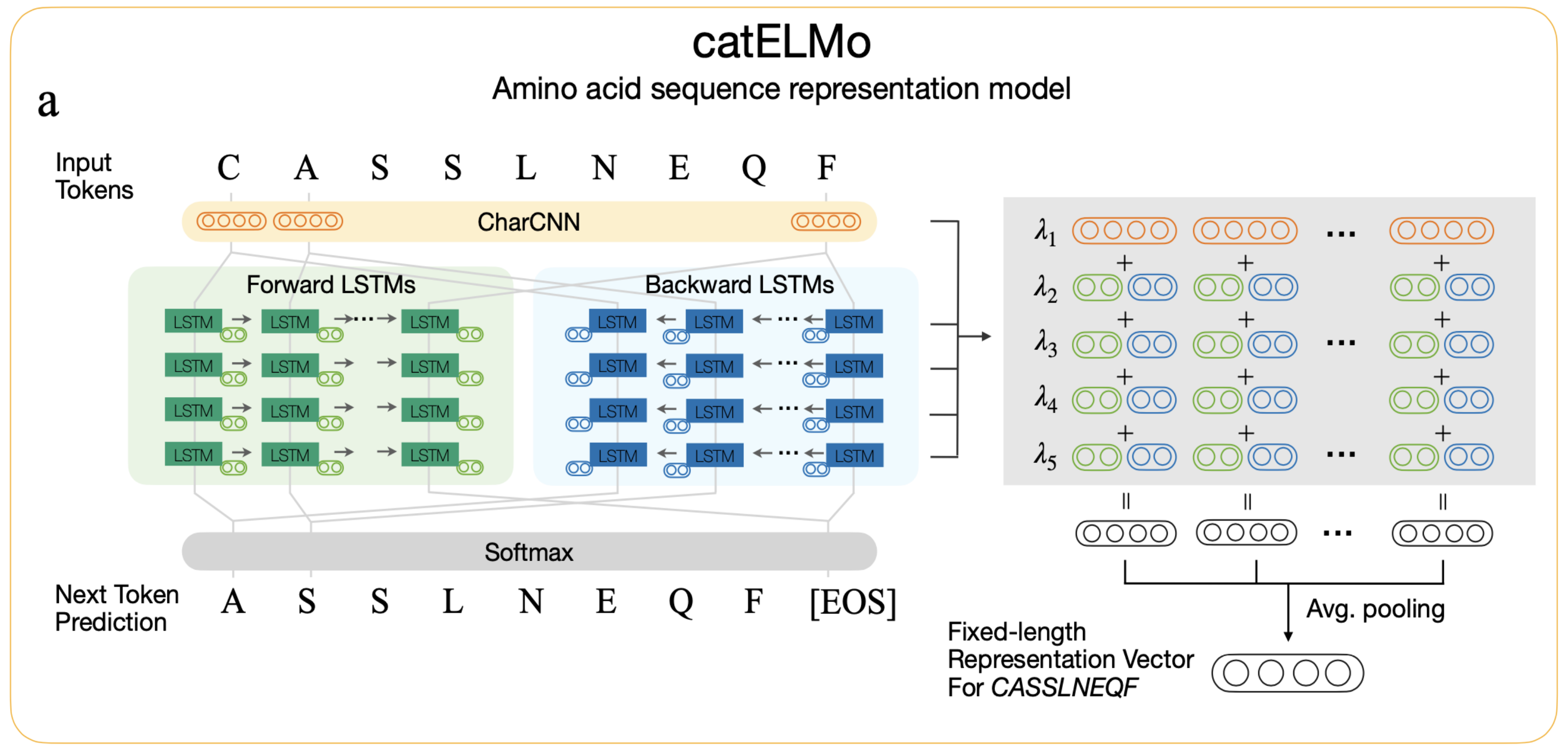
- We want to accurately predict whether any given TCR-Epitope pair will bind. We use a neural network trained on labeled TCR-Epitope pairs for this prediction task.
- In order to train the neural network, we first need to embed TCRs and epitopes into a fixed-length numeric vector representation. A good embedding has been shown to boost downstream performance in the prediction task.
- Using catELMo to generate embeddings yields better downstream performance than any other embedding method, *specifically for embedding TCRs and epitopes*.
- We investigate the root causes for this, with the goal of uncovering principles that can aid the design of other embedding techniques for specific types of biological data.

Methods

We want to start by determining the optimal hyperparameter settings for the baseline catELMo model. The following settings were tested:

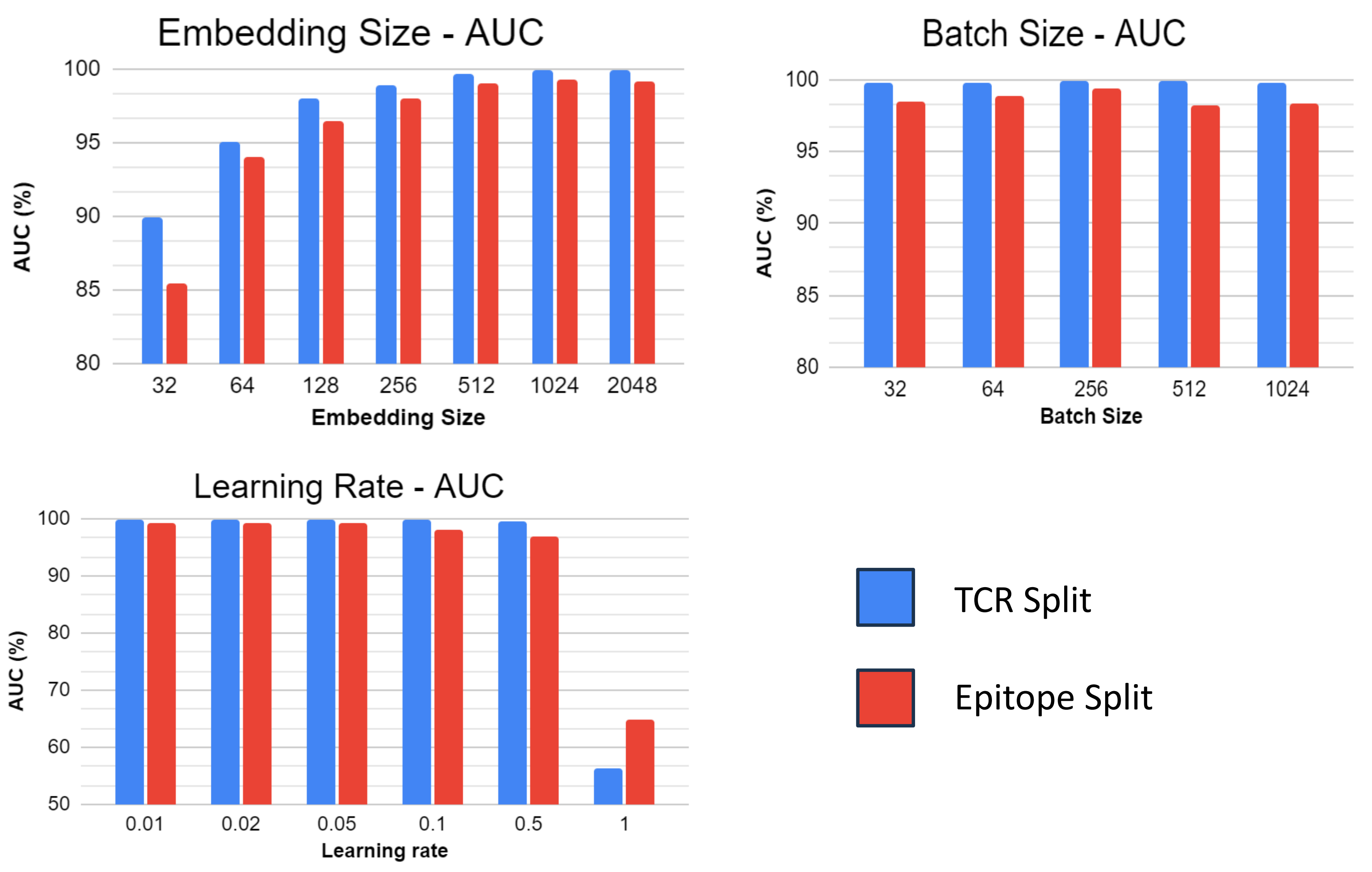
Embedding Size: *(baseline 1024)*	32	64	128	256	512	2048
Learning Rate: (*baseline 0.2)*	0.01	0.02	0.05	0.1	0.5	1.0
Batch Size: *(baseline 128)*	32	64	256	512	1024	

- catELMo was trained for 1 epoch on each setting, varying one hyperparameter at a time (while using baseline for others), and training perplexity of last batch was recorded.
- Dataset consists of 4 million TCR sequences – specifically CDR3 sequences of TCRβ chains.



- After training catELMo, weights were extracted and used to embed TCRs. (BLOSUM62, a static embedding matrix, was used to embed epitopes.)
- bap, a shallow NN, was then trained for the binding affinity prediction task 5 times each with TCR and epitope splits, and AUC/precision/recall/f1 scores were recorded.

Results



- Performance improves as embedding size increases, and when learning rate is less than or equal to 0.2
- Variations in batch size do not appear to affect performance

Future Work

We would like to test how varying the number of LSTM layers and their dimensions will impact downstream performance. Then we will investigate varying the model size. All future work is toward the goal of uncovering the underlying mechanisms for catELMo's superior performance as an embedder for TCRs.

References

- Zhang, P., Bang, S., Cai, M., & Lee, H. (2023). Context-aware amino acid embedding advances analysis of TCR-epitope interactions. bioRxiv, 2023-04.
- High-throughput sequencing of immune repertoires in multiple sclerosis - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/figure/Structure-function-and-diversification-of-antigen-receptors-A-The-T-cell-receptor_fig2_301233748

FURI

ASU Ira A. Fulton Schools of Engineering
Arizona State University