# Next-Gen Immunotherapies: Analysis of Large TCR Repertoires to Create a TCR Cluster Benchmark Using the catELMO Embedding Technique

Muhammed Hunaid Topiwala, Computer Science Major
Mentor: Heewook Lee, Assistant Professor
School of Computing and Augmented Intelligence

## Objective & Research Question

Accurate clusters in large-scale TCR datasets provides candidate disease-specific receptors and is vital to repertoire classification for personalized immunotherapy. This study investigates why catElmo, a bidirectional LSTM model, outperforms GIANA, a handcrafted static embedder, through an examination of various metrics, and noise mitigation strategies using one density and one distribution-based algorithm for the Unsupervised Learning task: clustering of TCR-Epitope data.

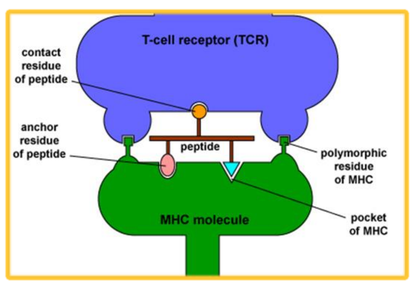
**Key words:** TCR- epitope interaction, benchmarking TCR CDR3 encoding, data

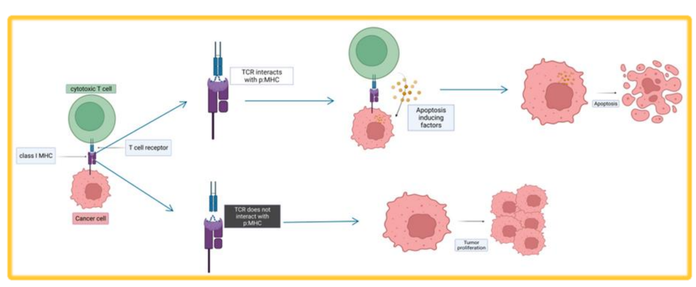
## Background

• T-cells initiate an immune response through binding to foreign antigens, where the T-cell receptor (TCR) recognizes the epitope of an antigen. This interaction exhibits many-to-many characteristics, as one TCR may bind with different epitopes, and one epitope can be targeted by various TCRs.

An epitope of an antigen may exist on other antigens through cross-reactivity too.

• Accurately predicting TCR-epitope clusters in large-scale TCR datasets provides candidate disease-specific receptors and could be used as the basis for a TCR-based non-invasive multi-disease diagnostic platform.



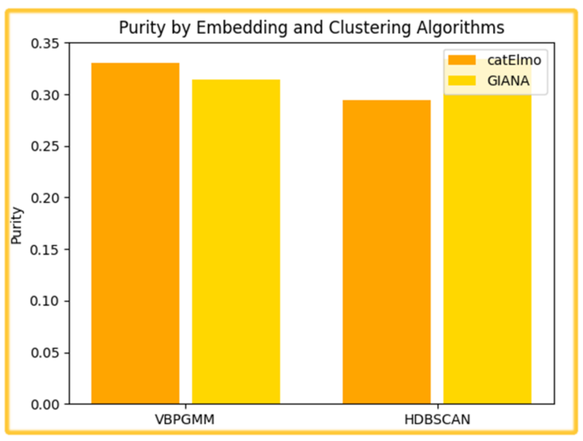



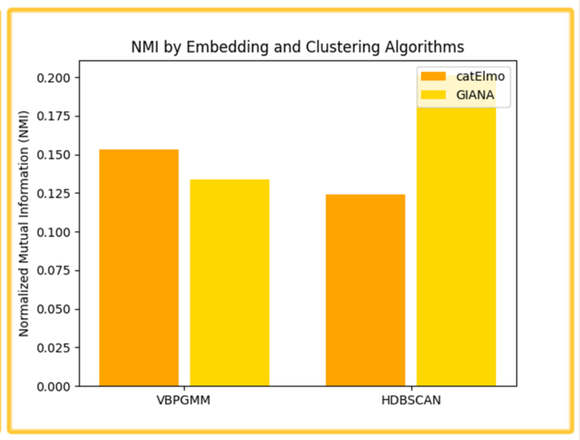

Main idea of TCR-Epitope interactions: If TCR binds to Epitope, tumor is locally stopped. If TCR fails to bind to Epitope, tumor is unchecked until another TCR recognizes the foreign pathogen.

## Results

• catELMo's context-aware, unsupervised deep learning embedding model outperforms the static BLOSUM-based GIANA embedding in hierarchical clustering tasks, by higher NMI scores and clustering homogeneity, showcasing its capacity to capture nuanced sequence variability and epitope-specific interactions.

• GIANA's clusters perform better in HDBSCAN which highlights the potential advantages of embeddings optimized for density separations




## Future Works

• catELMo could benefit from exploring geometric isometric embedding techniques or hybrid models to enhance performance in density-based methods.

• Investigate the incorporation of dimensionality reduction techniques or simplified vector representations to maintain high context-awareness while boosting the speed and efficiency of clustering tasks.

## Methods

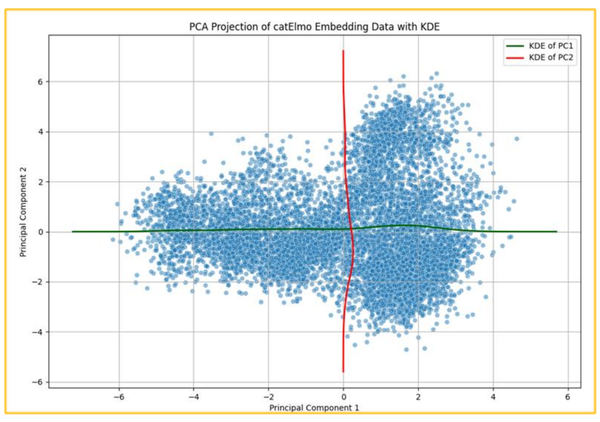
• catElmo was trained for 10 epochs, and GIANA uses a hand-crafted embedder requiring no training. The length of the representation vectors is 1024 (LSTM) and 96 (GIANA).

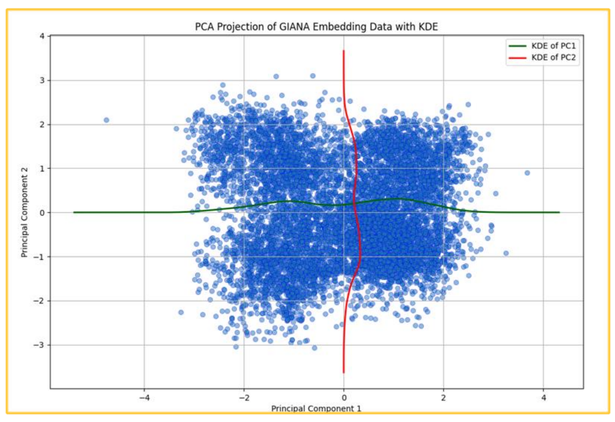
• The training dataset was initially sourced from the catELMo project, which provides a comprehensive collection of human repertoires from seven distinct projects.

• The testing dataset of 9033 rows of "TCR (CDR3) sequences with antigen-epitope labels" is sourced from the GIANA project. The ground truth number of clusters are 25.

• The choice of a clustering algorithm depends on the dimensionality, density and distribution of the embedded real-valued representation. While dimensionality is known for the embeddings-density and distribution is inferred using PCA, KDE and Epsilon point from k-distance graph.

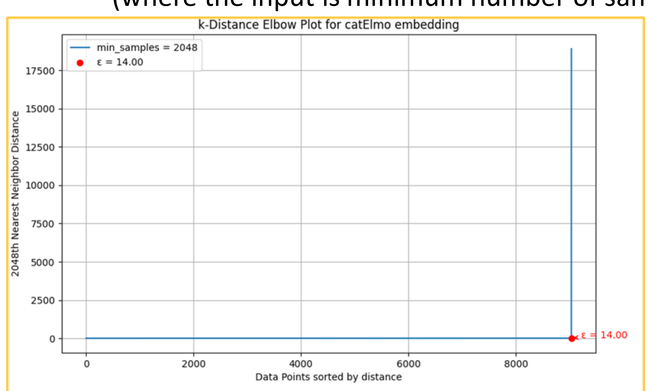
**Distribution** is inferred through Principal Component Analysis (PCA), and Kernel Density Estimation (KDE) on each representation

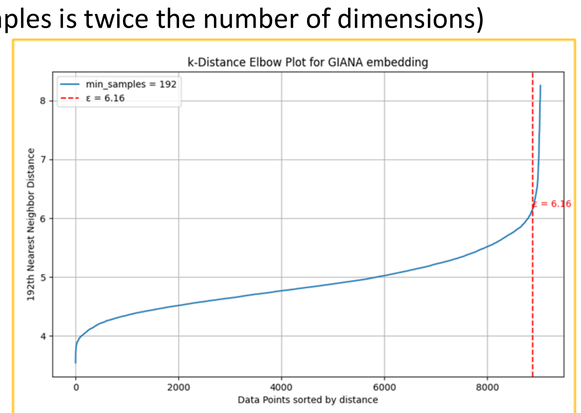



Skewed distribution and multimodality suggest underlying subpopulations, justifying clustering with a Variational Bayesian Gaussian Mixture (VBGMM) (1/2). VBGMM clustering algorithm fits multiple Gaussian distributions to capture the underlying structure.

**Density** is inferred using Elbow point through k-distance graph on each representation (where the input is minimum number of samples is twice the number of dimensions)




The gradual slope and clear elbow at Eps = 14, 6.16 in the k-distance plot indicate varying densities, justifying clustering with a Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) (2/2).

• Evaluating TCR-epitope clustering results needs consideration of important factors like the coverage of the clustering method, the extent to which epitope-specific sequences are clustered together and how consistent this process is. The following metrics for the evaluation of clustering quality were set out to evaluate these criteria. Although each metric has its limitations, collectively they offer a solid framework:

$$\mathrm{NMI}(U,V) = \frac{2 \cdot MI(U,V)}{H(U) + H(V)}$$

$$\text{where} \quad MI(U,V) = \sum_{u \in U} \sum_{v \in V} P(u,v) \log \frac{P(u,v)}{P(u)P(v)},$$

$$H(U) = -\sum_{u \in U} P(u) \log P(u)$$

$$\text{purity} = \frac{\sum |\gamma(s \in c) = \gamma_{max(c)}|}{\sum |s \in c|}$$

1. **Normalized Mutual Information** (NMI) measures the agreement between predicted and true labels within a single cluster, quantifying how much information is shared between the two label sets. It ranges from 0 to 1, where 0 indicates no shared information between the predicted and true labels (indicating they are completely independent), and 1 indicates perfect agreement (where the predicted labels perfectly match the true labels)

2. **Purity** is calculated as the fraction of sequences within one cluster targeting the same epitope. For each cluster c, we count the number of sequences s specific for the most common epitope γ, sum the values and divide them by the total number of sequences in any cluster.

## References

• Zhang, P., Bang, S., Cai, M., & Lee, H. (2023). Context-aware amino acid embedding advances analysis of TCR-epitope interactions. *eLife*. https://doi.org/10.7554/elife.88837.1

• Zhang, H., Zhan, X., & Li, B. (2021). Giana allows computationally-efficient TCR clustering and multi-disease repertoire classification by isometric transformation. Nature Communications, 12(1). https://doi.org/10.1038/s41467-021-25006-7

• Valkiers, S., Van Houcke, M., Laukens, K., & Meysman, P. (2021). ClusTCR: A python interface for rapid clustering of large sets of CDR3 sequences with unknown antigen specificity. Bioinformatics, 37(24), 4865–4867. https://doi.org/10.1093/bioinformatics/btab446

FURI

Ira A. Fulton Schools of Engineering
Arizona State University