# Literature Review

## Introduction

As part of adaptive immunity, T cells detect and eliminate infected or cancerous cells in the body. T cell receptors (TCR) which expressed on the membrane of T cells are able to recognize the target. Basically, T cell is composed of a beta and an alpha chain. Because of the biological complexity, the genetic encoding, VDJ recombination for each chain providing an abundant diversity of TCR species. As a result, a wide range of TCRs are produced. Since TCR repertoires respond very strongly to health status, TCR specificity can be very useful for therapeutic approaches. More specifically, the study of TCR sequence similarity features for antigenic epitopes (key of T cell and peptide-MHC complex binding) has great value and significance in implementing marker tracking, treatment, and control of disease.

In this project, tasks are concentrated on understanding and revealing the relationship between TCR sequence represented by cdr3 sequence and TCR binding specificity. Firstly, using t-SNE method to perform clustering analysis based on distance matrix composed of known TCR sequences to investigate the specificity between cdr3 sequences and antigenic epitopes. In the two-dimensional clustering plot, the similarity between TCR sequences and antigenic epitopes can be better demonstrated, which brings new insights into TCR disease treatment.

Due to the high-dimensional complexity of TCR sequence data, directly analyzing these data is impractical. As a result, this report turns to Multidimensional Scaling (MDS) to reduce the dimensionality while preserving data’s features as much as possible. Moreover, to explore TCR sequence data’s structures and features, Gaussian Mixture Models (GMM) are employed to cluster these dimensions reduced data. This method is able to perform refined clustering of TCT sequences based on probabilistic models. Moreover, considering TCR specificities in each cluster and in order to display the relationships between cluster results and TCR specificities, UMAP is employed to perform data visualization. The combination of GMM, MDS and UMAP provides a useful analysis structure which is able to enhance the understanding of TCR specificities.

Furthermore, develop machine learning models that can predict the specificity of TCRs to certain epitopes based on its cdr3 sequence. This report uses K nearest neighbours (KNN) to classify TCR sequences. Knn as a simple but effective machine learning method, is widely used because of its efficiency and intuitiveness when dealing with classification tasks. This report employs KNN aiming to explore its potentials in classifying different TCRs’ antigen specificities and manage to make some conclusions based on classification results.