Report Title

line 1: Hongyu Qu  
line 2: *dept. name of organization   
(of Affiliation)*  
line 3: *name of organization   
(of Affiliation)*line 4: City, Country  
line 5: email address or ORCID

line 1: Yihui Ren  
line 2: *dept. name of organization*  
*(of Affiliation)*  
line 3: *name of organization   
(of Affiliation)*line 4: City, Country  
line 5: email address or ORCIDline 1: Zewei Yu  
line 2: *dept. name of organization   
(of Affiliation)*  
line 3: *name of organization   
(of Affiliation)*line 4: City, Country  
line 5: email address or ORCID

line 1: 4th Given Name Surname  
line 2: *dept. name of organization   
(of Affiliation)*  
line 3: *name of organization   
(of Affiliation)*line 4: City, Country  
line 5: email address or ORCID

*Abstract*—This electronic document is a “live” template and already defines the components of your paper [title, text, heads, etc.] in its style sheet. *\*CRITICAL: Do Not Use Symbols, Special Characters, Footnotes, or Math in Paper Title or Abstract*.

# 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, use tcrdist3 to calculate the distance matrices for the alpha and beta chains separately, and then use deepTCR to Calculate for the combined alpha and beta chains. Secondly 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.

# Literature Review

## Quantifiable predictive features define epitope-specific T cell receptor repertoires

As part of the study (Dash et al., 2017), quantifiable predictive features were defined for certain epitopes within the TCR library. TCR epitope specificity was predicted based on sequence features alone in a novel approach to sequencing and analysing TCRs. The analysis they produce is useful for grouping receptors that are related and selecting representatives to study specificity in further experiments. Using parameterised antigen-specific immune libraries based on different epitopes, they propose developing generalised TCR-pMHC recognition models, which can be applied to a variety of research areas, including cancer immunotherapy and infectious diseases diagnosis and treatment.

## T cell receptor sequence clustering and antigen specificity

The study (Vujovic et al., 2020) explored TCR sequence clustering based on antigen specificity, an important step towards understanding how TCRs detect different antigens and the potential for targeted therapies based on TCRs. Based on TCR sequences, this paper proposes a clustering algorithm for classifying antigen specificity. In addition to enhancing prediction specificity and accuracy, deep learning techniques are also highly effective. There is no single tool available to unambiguously classify TCR receptor specificity because of the TCR's cross-reactivity, its ability to bind multiple antigens with different affinities, and its inability to induce T-cell activation merely by binding to the TCR.

## DeepTCR

The DeepTCR framework was developed by (Sidhom et al., 2021). Deep learning algorithms are used for analysing TCR sequencing data. A structure based on TCR sequences is proposed for revealing the complex pattern of antigenic specificity found in TCR sequences. Analyzing TCR sequences from human and mouse datasets with DeepTCR uses both supervised and unsupervised deep learning methods. In order to efficiently learn these complex patterns, the framework extracts features from sequencing data and employs a variety of neural network architectures. Using deep learning from conventional T cells culture combined with TCR sequencing, they identify antigen-specific responses that identify both an immune response and its TCR sequence diversity. Their results demonstrate the power of detecting an immune response as well as its TCR sequence diversity. In their work, they demonstrate the value of this approach in generating previously unrecognized hypotheses.

## Multidimensional Scaling

The book Multidimensional Scaling (Kruskal & Wish, 1993) describes the statistical technique of multidimensional scaling (MDS), which transforms complex, multidimensional data into a more understandable two- or three-dimensional space. The MDS method can be applied to immunology, including the modeling of TCR specificity. The objective is to reduce dimensionality while maintaining similarity or dissimilarity in the distance between data points. The use of this technique is essential when analyzing TCRs. This is because receptors and their target antigens are both represented in multidimensional space, and their proximity indicates a likelihood of interaction or similarity in function. Using MDS, one can visualise and explore the high-dimensional space of TCR interactions for TCR specificity prediction. TCR sequences can be visually grouped into functional or specific groups by MDS's effective reduction of dimensionality.

# Methodology

## Tcrdist3

Tcrdist3 is an open source Python3 package designed for analyzing and visualizing T-cell receptor (TCR) libraries. This package is built around the TCRdist metric, a distance metric that quantifies the similarity between TCR sequences. (Dash et al., 2017) originally released it as a new API for calculating TCR distance measurements and for development of updated biomarkers. The scope of its application was expanded with an update by (Mayer-Blackwell et al., 2021). A high-performance on-the-fly compiler called NUMBA has been used to optimize the software for CPU efficiency. In this toolkit, a powerful comparison tool based on edit/Levenshtein distances is included for NUMBA encodings. Further, TCRdist3 can read AIRR, VDJDB export, MIXCR output, 10x Cell Ranger output, and AnnoesunoSeq data. Due to this, it is highly adaptable to the various data sources commonly used in clinical studies.

In tcrdist3, paired sequences are used to calculate distances fast and efficiently. In this program, distances between TCRs are measured using NUMBA enhanced TCRdist with a distance metric adjusted or any custom Python3 metric. With its ability to compute "rectangular" pairwise matrices, it is particularly helpful for computing distances between small sets of TCRs and larger sets, with about 70 million distances calculated per minute per CPU. As a secondary sampling technique, TCRsampler is designed for estimating TCR frequencies and neighbouring TCR frequencies in non-antigen-enriched background libraries. The database includes pre-compiled databases of human and mouse TCR sequences that represent the largest library of TCRs previously exposed to antigens. It allows stratified sampling according to an individual's V and/or J gene usage. By doing this, TCR neighbours can be compared to relevant background sets more efficiently. By using inverse probability weighting, the module also adjusts the biased sampling method. The purpose of this is to ensure that oversampled TCR frequencies can be estimated accurately. Tools for generating sequence identity maps are provided by the palmotif module. Maps like this can be used to visualize amino acid frequency and distribution in TCR sequences.

By using tcrdist3, we can efficiently calculate separate distance matrices for TCR alpha and beta.

## DeepTCR

The DeepTCR (Sidhom et al., 2021) framework is a deep learning framework that analyses TCR sequencing data. The method reveals complex patterns in TCR sequences and determines antigen specificity based on various features. For individual TCR sequences and entire T cell libraries, the suite offers both unsupervised and supervised deep learning methods. In order to work, it learns the joint representation used by CDR3 sequences and V/D/J genes of TCRs. By improving antigen-specific TCR classification and extraction from noisy datasets, the study aims to improve antigen-specific TCR detection.

DeepTCR extracts sequence-based features from CDR3 variable-length sequences by embedding them into a continuous number space, followed by a convolutional neural network (CNN). Also, the V/D/J genes are provided as categorical variables to the network, which are embedded in continuous continuous numerical space to convert them. The CDR3 sequences and the V/D/J genes are concatenated within the network to provide a complete representation of the TCR sequences. In DeepTCR, the unsupervised aspect uses a variational autoencoder (VAE) to learn the distribution of TCR sequences around a latent space for clustering similar antigenic specificities. As compared to traditional clustering and comparison methods, this method of TCR characterisation produces high quality clusters that correspond to true antigen-specific labels. In DeepTCR, supervised models use CNNs to classify TCR sequences based on antigen-specific labels. Compared to unsupervised learning methods and traditional machine learning models, these models provide better results. Furthermore, supervised learning models can be used to identify antigen sequences that are the most predictive. The deepTCR is especially relevant to TCR specificity prediction since it provides detailed information about how antigenicity sequences are determined. By combining high-throughput sequencing and deep learning models, researchers can better characterize and predict TCR responses. In order to advance personalised immunotherapy and learn more about T cells, it is crucial to use this approach.

We can easily use DeepTCR to calculate combined alpha and beta chain metrics.

## t-distributed Stochastic Neighbor Embedding

Aiming to analyze the relationship between two species (mouse and human) TCR sequences in the vdjdb database and their antigen epitopes binding specificity, t-distributed Stochastic Neighbor Embedding (t-SNE) is used to visualize the clusters. t-SNE is a technique for dimensionality reduction particularly suitable for clustering visualization of high-dimensional datasets. The foundational principle of t-SNE starts by calculating the pairwise Euclidean distances between all data points in the high-dimensional space. Then all high-dimensional Euclidean distances will be converted into conditional probability which represents similarity(Van der, 2008). Basically, the mathematic formula is given by:

图示, 示意图

描述已自动生成

Pj|i is the conditional probability from i point to j point, which is the similarity from i to j. The above process involves a centered on xi, σ i as the variance Gaussian distribution.图表, 散点图, 气泡图

描述已自动生成图表, 折线图

描述已自动生成

The objective of t-SNE is to reduce the dimension of the data, project data from high-dimensional space to a low-dimensional space while retaining as many features as possible. Ideally, it expects the conditional probability qj|i in the low dimension to be as close as possible to the pi|j in the high dimensions. This is achieved by minimizing the Kullback-Leibler divergences between the high-dimensional and low-dimensional probabilities. Specifically, t-SNE will employ gradient descent to optimize the loss function C. Formular of qj|i and loss function are given by:

手机屏幕截图

中度可信度描述已自动生成文本

中度可信度描述已自动生成

In particular, alpha and beta chain TCR distance matrices have been calculated by TCRdist previous, and t-SNE accepts the distance matrix as input. In other word, the distances matrix will be directly used to compute the probability of similarity between points (TCR sequence). Eventually, the high-dimensional TCR distance matrix is projected into a 2-dimensional space, and the low-dimensional clustering plot is more readable and beneficial for specificity analysis. Notably, t-SNE has a few significant parameters: n\_components, perplexity, init, metric, learning\_rate. Particularly, the init parameter indicates the starting point of the iterative algorithm. Another technique Multi-dimensional Scaling (MDS) is able to ensure the relative distances as similar are as possible in the low-dimensional space to those in the original high-dimensional space. If the init parameter is set to an output from MDS, t-SNE can converge more rapidly and cluster better.

## Gaussian Mixture Models

Gaussian mixture model (GMM) is a parametric probability density function characterized by a weighted sum of multiple Gaussian densities\cite{Gaussian}. The given equation could describe a Gaussian mixture model,

where is a D-dimensional vector, are the components Gaussian densities with their own mean vector and covariance matrix and each Gaussian density is assigned a weight .

The parameters of a GMM are estimated by the Expectation-Maximization(EM) algorithm. EM algorithm is an iterative algorithm consisting of two steps: the Expectation step (E-step) and the Maximization step (M-step). In the E-step, this algorithm calculates the expectation of the log-likelihood function of variables, and in the M-step, the algorithm will find the parameters that maximize the expectation computed in the E-step. This process continues until the algorithm converges. In this case,

E-step:

Calculate

M-step:

In this report, antigen epitopes are assumed to be in Gaussian distributions or in multimodal distributions where simple clusters like k-means are usually unable to capture varied and complicate features. Moreover, GMM provides high flexibility as it allows clusters with different shapes, sizes, and directions, which is important when dealing with biological data.

## UMAP

UMAP is an algorithm for dimension reduction based on Riemannian geometry and algebraic topology. This algorithm can retain the global structure of datasets while considering the local structure in low-dimensional space (McInnes, 2018).

UMAP takes a comprehensive approach to dimension reduction. Before mapping data onto a low dimensional space, UMAP thoroughly learns the data's pattern in high dimension. It starts by using Nearest-Neighbour-Descent to find the data points' nearest neighbors. Then, for each point, it considers their local structures, creating a weighted graph where edges represent the distance between each point. UMAP also incorporates topological data analysis. Once the graph is created, UMAP uses manifold learning methods to capture the geometry and topological structure of the dataset, maintaining the distances between data points. While preserving the data structures, UMAP begins the task of mapping data points onto a low-dimensional space. It identifies a low dimension that best approximates the topological structure of the high-dimensional space. In this process, UMAP defines a cross entropy between the high dimensional and low dimensional spaces and then it minimizes this cross entropy using optimization methods such as stochastic gradient descent.

This is the cross entropy. In this equation, e represents the edges, are the known weights of edges from high-dimensional manifold approximation, and are the weights to be discovered for low-dimensional representation.

## K Neighbors Classifier

Knn is a simple but effective classification method that keeps all the training data for classification (Guo, 2003). The basic idea behind this algorithm is to learn based on the k nearest neighbors of each data point where k is an integer value decided by users (Pedregosa, 2011). After determining k's value, the algorithm computes the distances between the test points to be classified into those points in known categories. Usually, the distance is calculated using Euclidean distance or Manhattan distance. Based on the computed distance, the algorithm will find the k nearest neighbors to the point from the training dataset. Then, the category with the majority of votes will become the category of the test point.

# Data Description

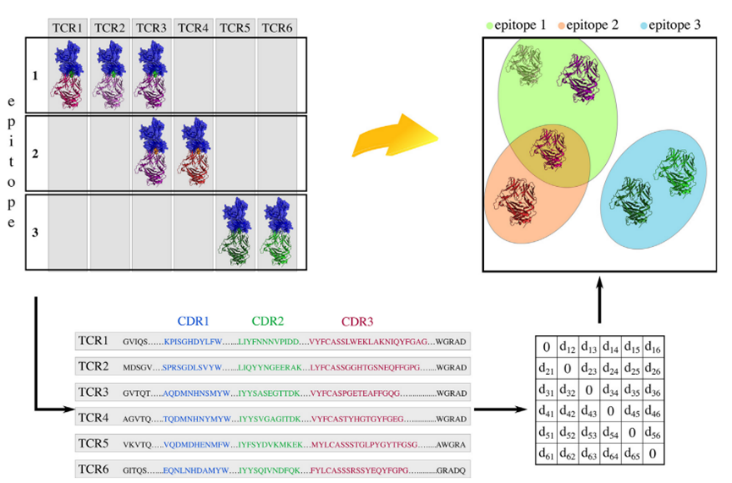
The dataset used for the project is vdjdb, which is a curated database of T-cell receptor (TCR) sequences with known antigen specificities. The database includes the TCR alpha and beta chains of HomoSapiens(human), MusMusculus(mouse), and MacacaMulatta(monkey) three species, the categories of antigen epitopes that bind specifically to them, and the corresponding diseases (antigen species). The total number of data is 92,772. This project aims to investigate the specific binding of TCR to antigen epitopes, so in the first step of data pre-processing, only the columns “complex.id”, “gene”, “cdr3”, “v.segm” , “ j,segm”, “species”, “antigen.species”, “antigen.epitope”, “vdjdb.score” are retained. There are some null values in the “v.segm”, “j,segm” columns that need to be removed firstly because the sequence in the v.segm, j,segm columns will be used in TCRdist to calculate the TCR distance matrix.

Subsequently, pre-processing the dataset differently depends on the tasks. For the task3 calculate the distance matrix of alpha, beta and combined chain, do a preprocessing on the “vdjdb.score” column by removing the data with vdjdb.score equal to 0 due to the data with 0 score does not have sufficient method details to draw any conclusion. Then only need to filter the chain based on genes and specific species based on species without extra pre-processing. For the task4 use t-SNE to reduce dimension and plot 2D clustering image, do the same pre-processing as task3, removing all data with score equal to 0. Besides, select only top 10 epitopes for clustering. On the one hand, this can greatly improve the quality of the data. On the other hand, the top 10 epitopes can also avoid too much data causing unclear clustering in 2D plots. For the task5, under the premise of removing the data with 0 score, select the top 8 antigen.species(disease) in the HomoSapiens dataset. Then, filter the number of antigen epitopes greater than 10 from these top 8 diseases and perform clustering on these epitopes. Considering the small number of mouse data, the score equal to 0 is preserved for the mouse clustering. Different from processing the human data, the mouse data are first selected for the top 5 antigen.species. For the task6:

Pre-processing does hierarchical sampling of the first 10 antigen epitopes and sampling 15000 data for each chain. Particularly, mouse data take all data from the first 10 antigen epitopes due to insufficient data.

# Results and Discussion

## Limitations of this approach in downstream analysis



As shown in the picture, this is a graphical representation of the epitope specificity of T-cell receptors (TCR) for different epitopes, as it has been described by Vujovic and colleagues. In the top part, we can see TCRs binding to different epitopes on different proteins. In the bottom part of the image, the CDR corresponding amino acid sequences for each TCR and also shows a matrix that probably represents some form of distance or similarity scoring between TCRs, which would consist of the amino acid sequences of the CDR and the matrix associated with it.

The limitations of the approach shown might include:

### *Variation of CDR3 length among TCRs*

If one-hot encoding assumes a fixed input length, it complicates the encoding process since the CDR3 region can vary significantly in length among a variety of TCRs. Because of this, it would be extremely difficult to compare them directly or input them into many machine learning models with fixed input lengths since the lengths are undefined.

### Alignment of the sequences

Since the CDR3 sequences differ in the lengths and compositions of their CDR regions in different organisms, aligning these sequences in order to compare them can be a challenge. There is a possibility that inaccuracies in alignments may affect distance or similarity measures.

### Loss of Sequential Information

In contrast to one-hot encoding, one-hot encoding does not capture the sequential nature of amino acid chains. In other words, it treats the sequence as a set of independent features, which means that it misses the patterns that can be discerned from the order in which amino acids are arranged.

### Sparse Representation:

Some models can have a difficult time processing large-scale data set because of the high-dimensionality of the data during encoding, so encoding one-hot can lead to sparse matrices that are inefficient and might lead to high-dimensional data that some models cannot handle.

## Method to overcome the limitations of this approach in downstream analysis

To overcome these limitations, the following steps can be taken:

### Variable Length Sequence Handling:

The handling of variable length sequences should be done using methods such as zero padding to ensure that all CDR3 sequences are the same length when the matrix is constructed. To achieve this, all sequences could be padded so that they match the length of the longest sequence in the dataset.

### Using more sophisticated encoding techniques:

Rather than using one-hot encoding methods, it is a good idea to use more complex encoding techniques, including embedding layers that can learn the amino acids in an efficient manner and capture the sequential nature of sequence data, as opposed to one-hot encoding.

### Dimensionality Reduction:

To mitigate the curse of dimensionality, use techniques such as PCA, t-SNE, or UMAP before training models to reduce the dimensionality of one-hot encoded data as much as possible, which can help reveal patterns in the data that may otherwise be hard to discover.

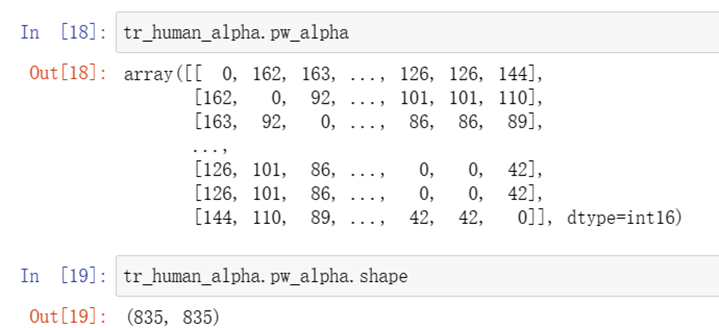
### 4. Alignment Algorithms:

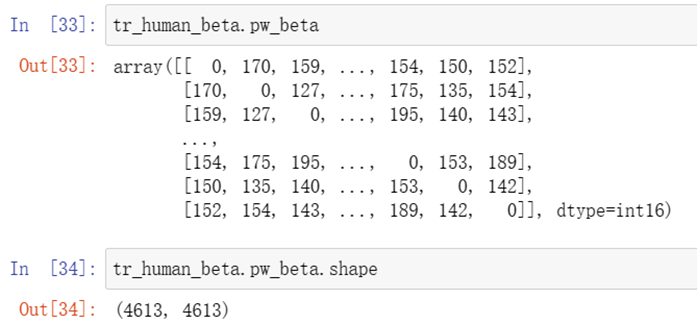
Utilize alignment algorithms that can effectively handle different lengths and compositions of sequences and determine associated similarity scores in a manner that ensures that the similarity scores are determined by accurately aligned sequences.

These methods can be integrated into the preprocessing and model development stages to improve TCR-specific prediction of sequence data, and some of them will be used in the following tasks.

## Calculate metrics for the alpha and the beta chains separately

Various distance metrics were represented by using the default parameter pwseqdist >= 0.2.0 of tcrdist3 (Dash et al., 2017), which typically weights CDR3 more heavily as it is in direct contact with antigenic peptides. We can calculate the distance matrix for alpha and beta chain metrics separately as shown in the screenshot below:

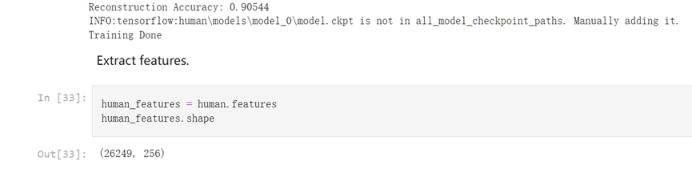


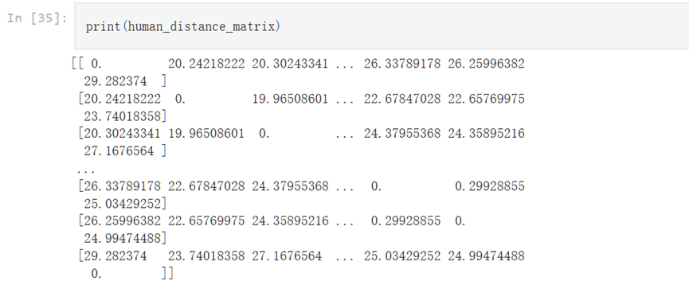


These two graphs show the distance matrices of the alpha and beta chains for the human TCR alone, the results are similar for other species, only the human results are shown here. We can find that the distance matrices of the TCR for humans as well as other species have their diagonal lines at 0, and each value in each row represents the distance between the TCR and all other rows, respectively, and the distance matrices of the alpha and beta chains are not one-to-one correspondences to be able to compute their binding chains directly. In order to further analyse and calculate their alpha and beta binding chains, we need to use deeptcr's method.

## Calculate for the combined alpha and beta chains.

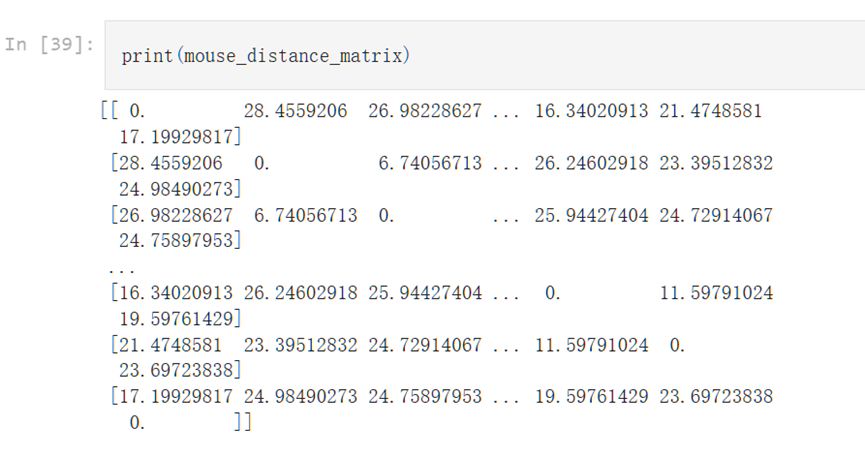
Because apes (MacacaMulatta) do not have paired α and β chains, they were not considered in the preparation of the data. We first categorised the paired α and β chains according to the antigenic species, the files were named after the antigenic species, and the paired α and β chains in each file were saved as a tsv file. According to the approach 'DeepTCR is a deep learning framework for revealing sequence concepts within T-cell repertoires' (Sidhom et al., 2021), we use a variational autoencoder (VAE) to take the CDR3 sequences from both the α- and β-chains along with their corresponding V, D, and J gene usage and learn a joint representation of these inputs.





As shown above, the final feature matrix shows that there's only 26249 rows left. So there are around 1000 rows missing. The reasons can be a little bit complex. I think the main reasons are that some features share same cdr3 representations and v, j segments information or these tcrs may also be found having similar information when the neutral network processed them. Overall, these reasons are just our assumptions and the ture causes need further development.

For other species, the procedures are same with human.

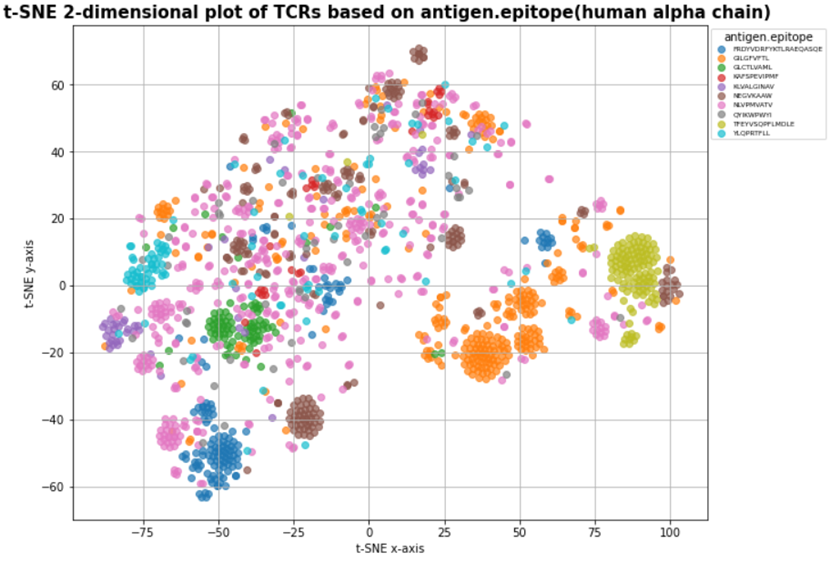


For further analysis, we next perform clustering such as Multidimensional Scaling (MDS) to perform cluster analysis to find similar groups of TCR sequences to visualise the distance relationship between sequences.

## Plot the TCRs in 2 dimensions and colour them based on specificity. Compare the plots for the alpha, the beta and the combined alpha-beta chains.

For task 4 results, exclude the MacacaMulatta specie from the dataset, focus on HomoSapiens and MusMusculus. In order to study and analyze the specificity relationship between TCR sequences and antigenic epitopes. The clustering plot is a significant evaluation of the specificity.

Simply apply t-SNE technique to reduce the dimensional of alpha, beta, and combined chains of human and mouse to two dimensions respectively. Specifically, clustering plots for human are shown below.



For better capturing valuable information from the clustering plot, the analysis focuses on the data with the top 10 number of antigenic epitopes.

In the 2-dimensional plot of human alpha chain TCR based on antigen epitope specificity, a diverse clustering pattern can be observed apparently. Each clear color represents an antigen epitope respectively, and each point represents a TCR from a high-dimensional project into 2-dimensional space. Such as the orange color cluster represents a specific antigen epitope: GILGFVFTL. Overall,

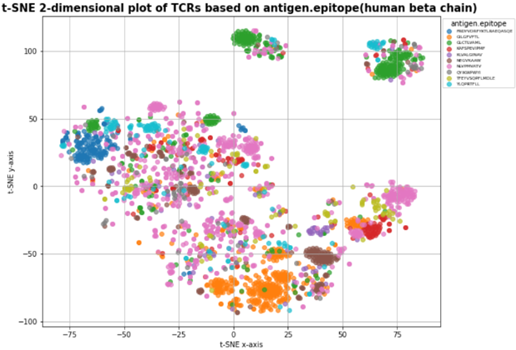
the clustering of certain epitopes is more concentrated and forms large clusters such as green, yellow, orange, and blue clusters. That indicates the TCR sequences corresponding to those epitopes have higher similarities. On the contrary, some epitope clusters are more dispersed and structurally forming many smaller clusters. Such as the brown epitope. In particular, the pink epitopes (NLVPMVATV) are very scattered and only form a few small clusters.

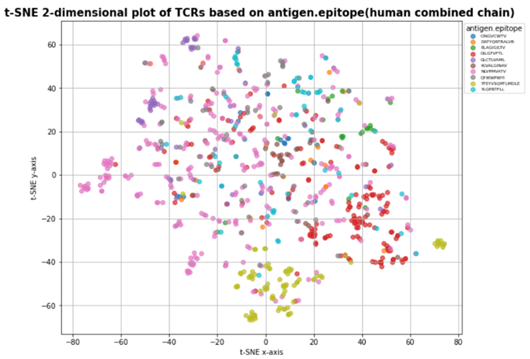
### Specificity

The phenomenon of cluster formation illustrates that similarity exists in TCR sequences corresponding to the same antigenic epitope. This is critical for TCR disease prediction, tracking, and treatment. Notably, same color antigen epitopes form few scattered clusters which reflect that different TCR sequences can recognize the same antigenic epitope in some circumstances. there is a large variety of TCRs in human due to the recombination of VDJ gene segments. Moreover, different individuals have a unique set of TCR specificity responses (Vujovic et al., 2020). To ensure that each individual is able to recognizes specific antigen epitope in the immune system, many TCRs can recognize the same epitope. Generally, it reflects the specificity recognition of TCRs.

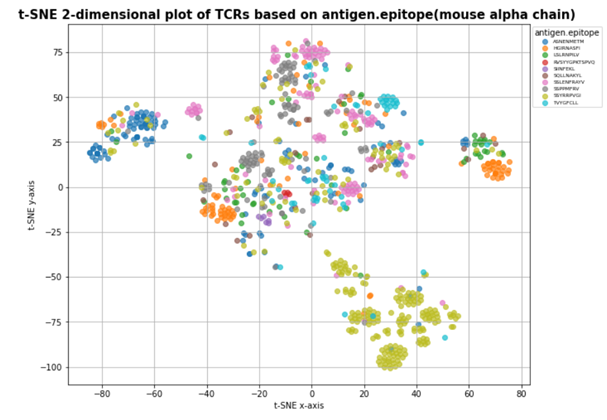
### Cross-reactivity

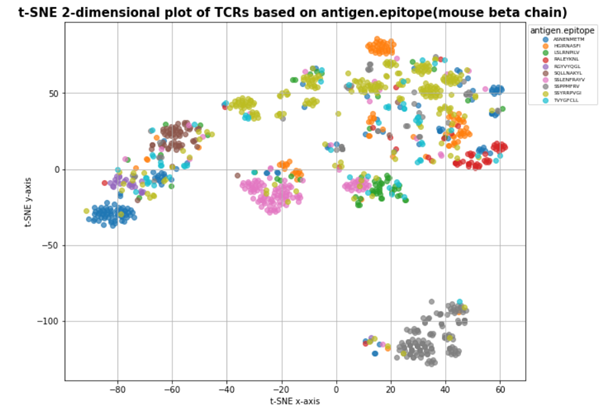
On the other hand, it was observed that different color clusters represent TCR clusters corresponding to different antigen epitopes, are close to each other or even overlap. For instance, clusters at (90, 0) and (-45, -10). It is possible that this phenomenon shows a TCR property called cross-reactivity. Human has limited number of t cells in the body. So limited TCR has to respond specificity to a huge number of antigen epitopes. As an important consequence is that unique TCR may be able to recognize more than one antigen epitope. Nevertheless, t-SNE technique focuses more on preserving local rather than global data structure. When the t-SNE projects high-dimensional TCR features into a low-dimensional 2D plot, it is possible that TCR with distant spatial distance will be projected together. Therefore, they appear to be close or map to each other in 2D image. That is to say, it does not accurately reflect the cross-reactivity of TCRs. More biological experiments are needed to verify the phenomenon.





Furthermore, in the 2D plot of TCRs of the human beta chain, most of the clusters are closer to each other compared to alpha chains’ cluster distribution e.g. cluster at (70, -35). That potentially indicates a higher cross-reactivity of beta chain TCR. At the same time, it may also indicate that alpha chain TCRs are more specific in recognizing different epitopes. Additionally, alpha and beta chains reflect different clustering patterns for the same antigen epitope. In the 2D plot of TCRs of human combined chain, points of the same epitope are more scattered in a unique cluster which potentially reflecting TCR sequence or structure of the combined chain can be more complex than single alpha or beta chains. TCRs have a more integrative property for antigen recognition. Certainly, normally alpha and beta chains are combined inside human body and TCRs will have more comprehensive features of the antigen epitope specificity recognition.





图表, 散点图

描述已自动生成

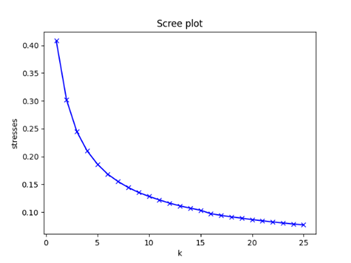
The data of mouse species are implemented the same t-SNE dimensionality reduction clustering.

Although the mouse data is much less than human and its antigen epitopes are totally different from humans, but it is easy to find out that mouse TCRs clusters of epitopes are formed clearly as well from the observation of alpha chain beta chain, and combined chain. It can be inferred that the binding of TCR to antigen epitopes also has specificity and cross-reactivity in mouse species.

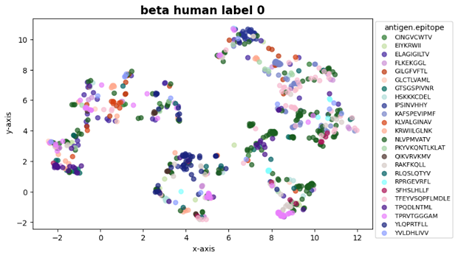
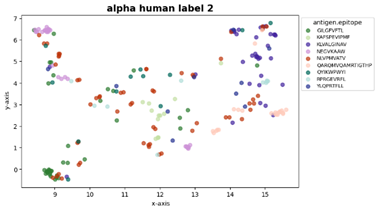
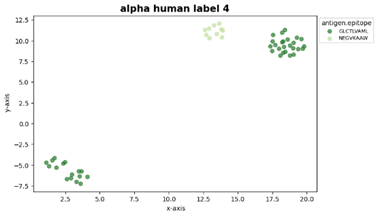
## Cluster TCRs

This report uses Gaussian Mixture Models to cluster TCR sequences based on their specificity. This strategy is applied to human and mouse data separately across alpha and beta chains.

Considering the only known thing is the distance matrix, MDS is applied to this distance matrix to map it to a specific space. MDS can reveal data points' relative positions and yield valuable insights(). Moreover, MDS allows users to use a precomputed distance matrix as input, which is suitable for this case. The output of MDS represents the relative positions of each TCR, which is equivalent to mapping the distance matrix onto a feature space about TCR. However, before applying MDS, like PCA, the number of principal components must be determined to represent critical structures of the dataset while avoiding overfitting and unnecessary complexity. From (Kruskal & Wish, 1993), a method called Scree plot is introduced where in each step 'stress' which is the square root of a normalized "residual sum of squares" is recorded. Like the elbow plot, the turning point is usually chosen as the number of components that best balance the information and complexity.



In this case, the number of components is all chosen as eight from the plots for both human and mouse. After using MDS to map the distance matrix onto a feature space, GMM is employed to do clustering. From the dimension reduction results discussed in the last chapter, it is clear that some antigen epitopes cluster together. Still, many antigen epitopes are scattered in the space, which means that some TCR sequences may not have a specific boundary, and they are likely to overlap in feature space. Therefore, in contrast to other cluster methods like K-means, GMM allows a data point belonging to multiple clusters, so it is suitable for dealing with fuzzy classification boundaries and overlapping distribution, which is possible for TCR sequences. When implementing GMM, the initial mean state is set to be the mean of each kind of antigen species as their distributions are assumed to be in Gaussian. Moreover, consistent with the number of antigen species decided to do analysis, the numbers of components for humans and mice in this experiment are set to be 8 and 5 individually.

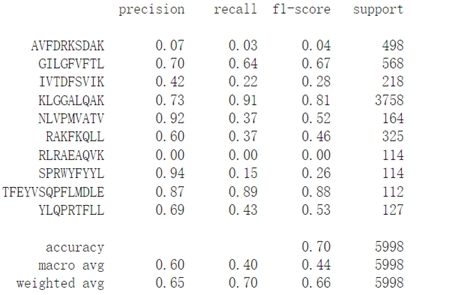


The cluster results show that for both chains and species, antigen epitopes in some clusters show strong clustering patterns, but for some clusters, the data points seem to scatter everywhere. The reasons for such a phenomenon are complicated. Firstly, from a technique perspective, to do visualization, data have to be mapped to a 2-dimensional space, so in this process, a lot of information is lost, which may lead to the condition that data is clustered in high dimensions but scattered in low dimensions. From a biological aspect, the complexity of TCRs may become a reason. Some epitopes may have a diverse array of TCR responses, which could result in a dispersed pattern in each cluster. Then, conservative antigen epitopes usually behave more densely in space, which could lead to good cluster results. In contrast, those with polymorphism or frequent mutations may tend to scatter in the space. Moreover, TCRs exhibit cross-reactivity and can bind to multiple antigens, each with different degrees of affinity, which may also contribute to some scattering phenomena.

## K neighbors classifier

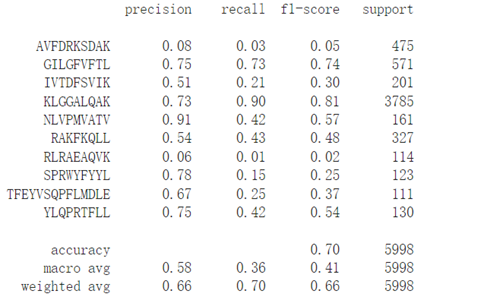
In task6, for the human dataset, K neighbors classifier is used to classify the 15000 data with the top 10 epitopes sampled hierarchically.

### Prediction of human epitopes (alpha chain)



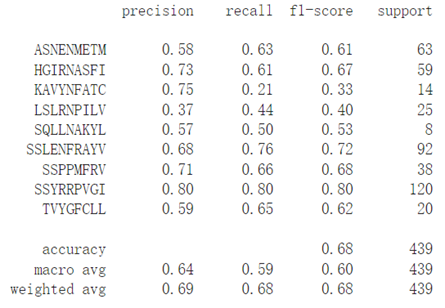
The overall accuracy of knn model is 0.7 and the macro average F1 score is 0.44. Notably, the F1 score shows significant difference in predicting different epitopes. For instance, F1 score of KLGGALQAK is the highest (0.81) and F1 score of RLRAEAQVK is 0. Obviously, a part of the lower f1 value reduces the average f1 value. The model performs well on some epitope categories and poorly on others. One potentially reason is related to the diversity of the TCR sequence corresponding to the epitopes. If the model can predict epitopes well, the TCRs sequence corresponding to them may have a higher similarity. This is reflected in the clustering plot by the more concentrated clustering of the same epitope and the smaller number of small clusters. On the contrary, if there are TCRs with a lot of categories and complex structure sequences that can specifically bind with a single epitope, the model prediction performance will be bad. In other word, TCRs’ specificity and cross-reactivity have a significant impact on model’s performance.

### Prediction of human epitopes (beta chain)

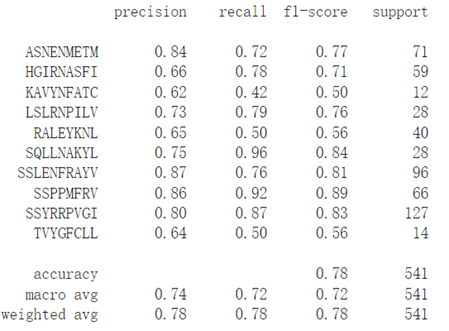


Comparing the prediction results of human alpha chains, the beta chain has an overall accuracy 0.70 and macro average F1 score 0.41 which is similar with the prediction evaluation of alpha chain. Similarly, the model performs poorly in predicting some specific epitopes such as AVFDRKSDAK and RLRAEAQVK. Generally, TCR specificity dictates that more than one TCR category can recognize the same epitope. That will reduce the predictive performance of some normal machine learning models.

### Prediction of mouse epitopes (alpha chain)



### Prediction of mouse epitopes (beta chain)



From the prediction of mouse epitopes, alpha chain and beta chain have overall accuracy 0.68, 0.78 and F1 score 0.6, 0.72 respectively. Noticed that the overall performance of mouse epitope prediction is greatly better than human. This may be related to the simpler structure of the mouse.

On the one hand, biological TCRs have a high complexity in feature space initially. On the other hand, the knn model is sensitive to the unbalanced data. Moreover, it has a high requirement for the feature space distance, which will have an impact on the final prediction results. Optimizing to a more complex and advanced model may be more helpful.

# Further Work and Improvement

For the future work of t-SNE technique application on TCRs clustering, the first one is t-SNE parameters optimization. Some parameters in t-SNE have significant impact on the visualization. For instance, the local neighborhood of each point is small and may miss important features of the global data when perplexity is set too low. On the contrary, if the perplexity is set too large, the clustering plot may miss some local data structure. Besides, the learning rate decides the stability of t-SNE’s optimization process. Adjusting the finest combination of those parameters is one of the future work.

There are a number of TCRs sequences which correspond to the same epitope scattered in the 2-dimensional image. Its reason needs to be further investigated. Additionally, t-SNE dimensional reduction is a process of projection and realistic epitope recognition occurs in three-dimensional space. Therefore, it does not exactly reflect the specificity and cross-reactivity of TCRs. It is necessary to combine multiple dimensional reduction method or add novel experiments to improve clustering effect in the future.

About the model, this report uses Gaussian mixture models and K neighbors to do the clustering and classification. However, the experiment results show that these models could only display some TCR features related to TCRs' differences. TCR is complicated, so it is hard to describe its features using a simple model. Therefore, in the future, to explore TCR with more information, some complicated models can be considered to capture TCR structures. Deep learning is a good direction, and it has already created some excellent works, such as DeepTCR. Moreover, (Madi et al., 2017) use network analysis, showing that CDR3 sequences with a similar annotation tend to be linked with the same cluster, and their experiments show that this method has good performance, So Network analysis can also become a future study direction.

# Conclusions

A brief summary of the key insights in your report.

##### References

1. Dash, P. et al. (2017) ‘Quantifiable predictive features define epitope-specific T cell receptor repertoires’, Nature, 547(7661), pp. 89–93. doi:10.1038/nature22383.
2. Guo, G. et al. (2003) ‘KNN model-based approach in classification’, On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE, pp. 986–996. doi:10.1007/978-3-540-39964-3\_62.
3. Kruskal, J.B. and Wish, M. (1993) Multidimensional scaling. Newbury Park, CA: SAGE publ.
4. Madi, A. et al. (2017) ‘T cell receptor repertoires of mice and humans are clustered in similarity networks around conserved public CDR3 sequences’, eLife, 6. doi:10.7554/elife.22057.
5. Mayer-Blackwell, K. et al. (2021) ‘TCR meta-clonotypes for biomarker discovery with TCRDIST3 enabled identification of public, HLA-restricted clusters of SARS-COV-2 tcrs’, eLife, 10. doi:10.7554/elife.68605.
6. McInnes, L. et al. (2018) ‘UMAP: Uniform manifold approximation and projection’, Journal of Open Source Software, 3(29), p. 861. doi:10.21105/joss.00861.
7. Reynolds, D. (2009) ‘Gaussian mixture models’, Encyclopedia of Biometrics, pp. 659–663. doi:10.1007/978-0-387-73003-5\_196.
8. Sidhom, J.-W. et al. (2021) ‘DeepTCR is a deep learning framework for revealing sequence concepts within T-cell repertoires’, Nature Communications, 12(1). doi:10.1038/s41467-021-21879-w.
9. Vujovic, M. et al. (2020) ‘T cell receptor sequence clustering and antigen specificity’, Computational and Structural Biotechnology Journal, 18, pp. 2166–2173. doi:10.1016/j.csbj.2020.06.041.
10. Pedregosa, F. et al. (2011) ‘Scikit-learn: Machine Learning in Python’, Journal of Machine Learning Research, 12, pp.2825-2830.

[11] van der Maaten, L. et al. (2008) ‘Visualizing Data using t-SNE’, Journal of Machine Learning Research, 9(Nov), pp.2579-2605.

##### Appendix

The document up to this section should be no more than 8 pages. The appendix section is optional. You can include additional material here, but it will not be marked.