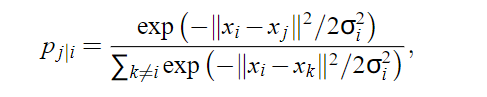
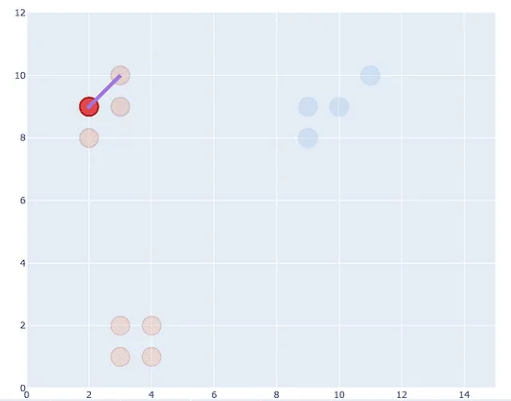
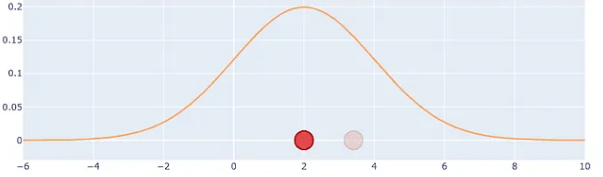
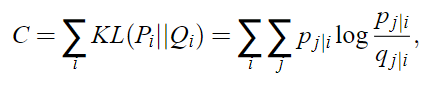
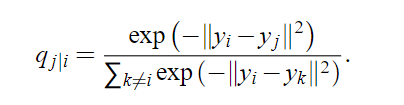
Aiming to analyze the relationship between two species (mouse and human) TCR sequences in the vdjdb database and their antigen epitopes binding specificity, t-distributed Stochastic Neighbor Embedding (t-SNE) is used to visualize the clusters. t-SNE is a technique for dimensionality reduction particularly 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 Maaten LHinton G2008Visualizing data using t-SNEJournal of machine learning research 9). Basically, the mathematic formula is given by:



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.  

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:

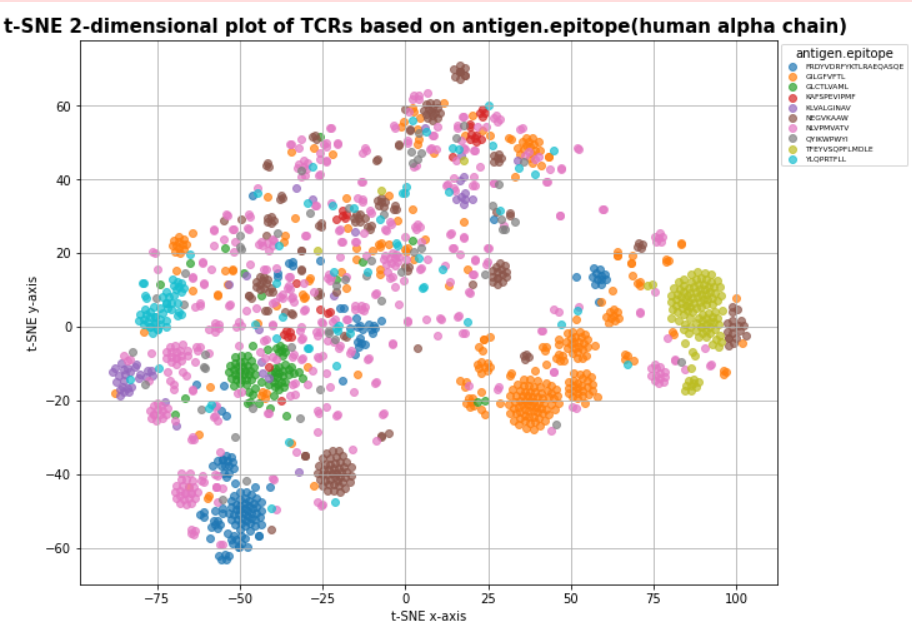
 

In particular, alpha and beta chain TCR distance matrices have been calculated by TCRdist previous, and t-SNE accepts the distance matrix as input. In other word, the distances matrix will be directly used to compute the probability of similarity between points (TCR sequence). Eventually, the high-dimensional 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.

# Results and Discussion

Reporting on the experiments with discussion on insights. Technical challenges are to be discussed here too.

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.

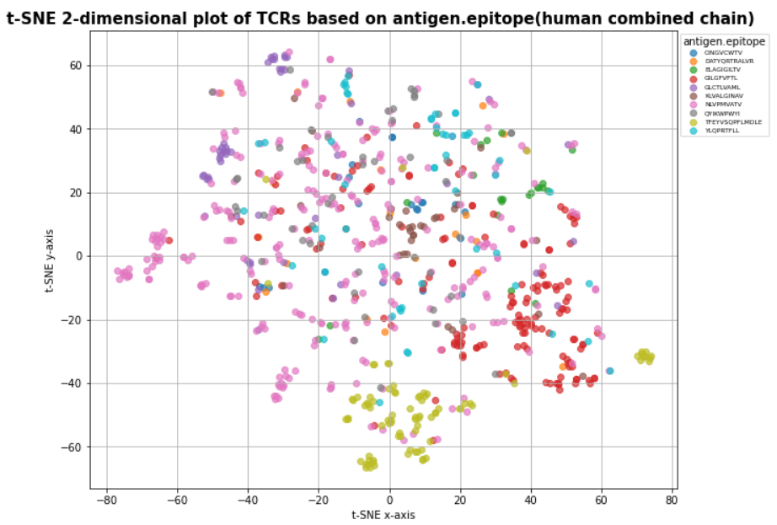
In the 2-dimensional plot of human alpha chain TCR based on antigen epitope specificity, a diverse clustering pattern can be observed apparently. Each clear color represents an antigen epitope respectively, and each point represents a TCR from a high-dimensional project into 2-dimensional space. Such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

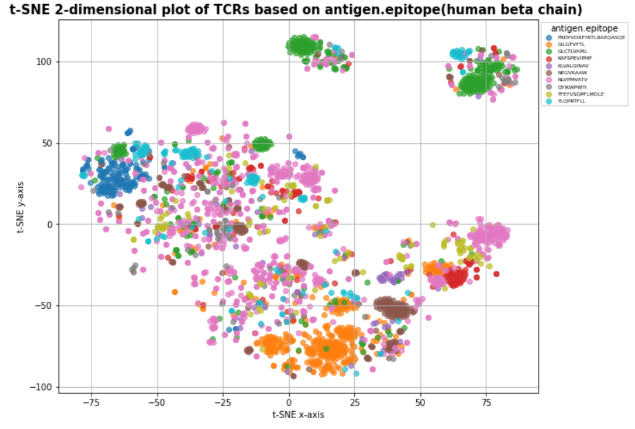
the clustering of certain epitopes is more concentrated and forms large clusters such as green, yellow, orange, and blue clusters. That indicates the TCR sequences corresponding to those epitopes have higher similarities. On the contrary, some epitope clusters are more dispersed and structurally forming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

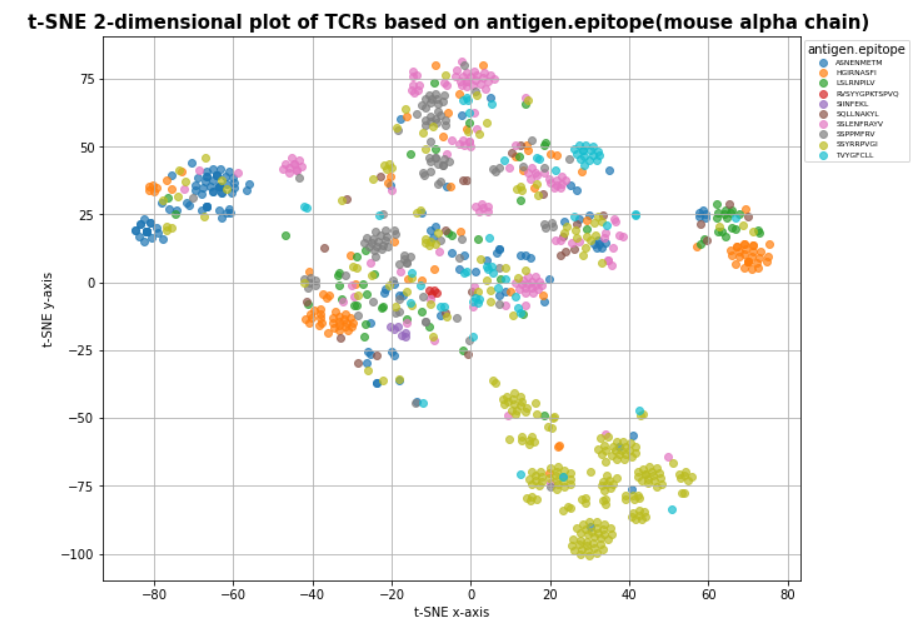
## Specificity

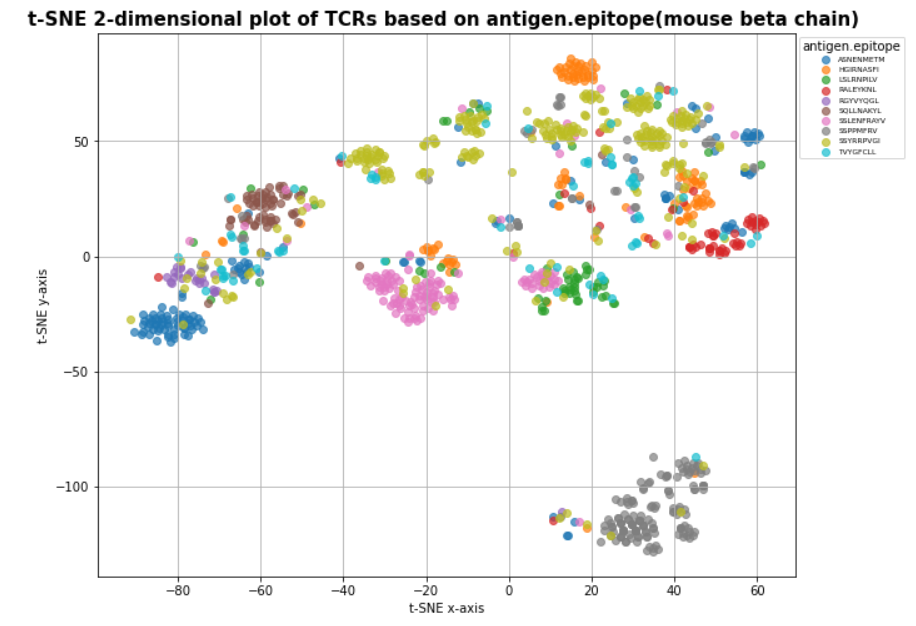
The phenomenon of cluster formation illustrates that similarity exists in TCR sequences corresponding to the same antigenic epitope. This is critical for TCR disease prediction, tracking, and treatment. Notably, same color antigen epitopes form few scattered clusters which reflect that different TCR 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 (T cell receptor sequence clustering and antigen specificity). 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.

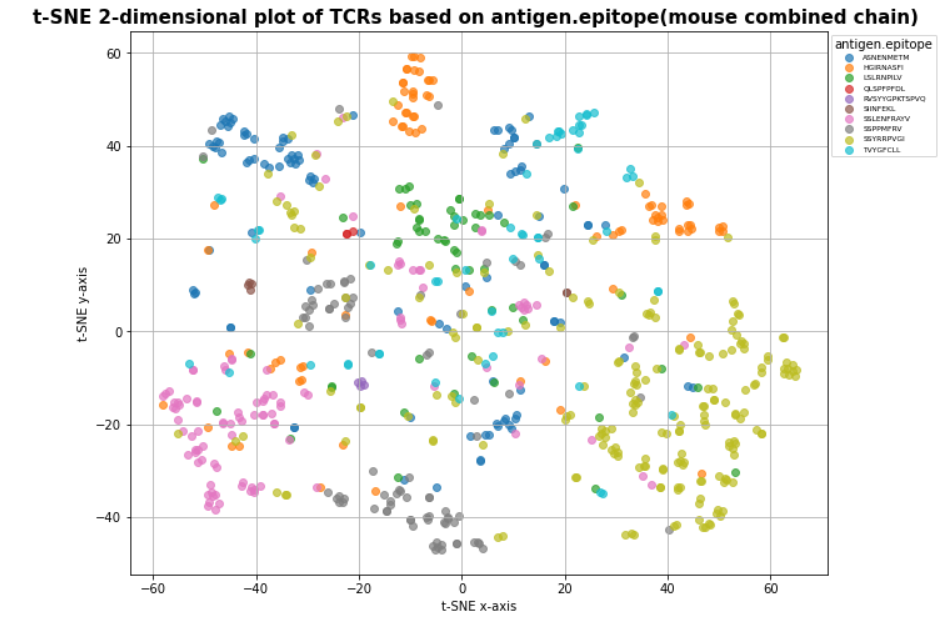
## Cross-reactivity

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.







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.