# Data Description / Preparation

Includes description of data sources, samples and steps for pre-processing if any.

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.