# III. 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.

## Reference:

1. 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.
2. 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.