Predicting the diversity and specificity of T-cell receptors (TCRs) plays a crucial role in immunotherapy. Many studies employ deep learning methods for training, which are time-consuming and demand high computational performance. In this study, machine learning techniques were applied to analyze and predict the diversity and complexity of TCRs within bioinformatics. Through a series of research, filtering, and calculations, including Levenshtein distance, GIANA encoding, PCA, MDS, hierarchical clustering, and DBSCAN, TCR sequence data were processed and analyzed. The results from different algorithms were compared to elucidate the patterns and connections of TCRs.

Comparing the effectiveness of these algorithms revealed that GIANA encoding, combined with PCA and MDS techniques, excels in data dimension reduction and feature extraction, effectively capturing the subtle differences between TCR sequences. Additionally, the DBSCAN algorithm demonstrated high flexibility and robustness in handling TCR data with complex distributions, automatically identifying the number of clusters and dealing with outliers.

This study provides a novel approach to immunotherapy methods. Through this research, a better foundation is laid for future studies on the role of TCRs in immunological research.