Task 2 Exploration:

**Limitations of using one-hot for amino acid representation:**

Loss of information: The single-point representation only provides information about the amino acid type and ignores information about the spatial structure and interactions between amino acids. This means that two TCRs with identical amino acid sequences, but different 3D structures may be incorrectly considered identical.

Length diversity: Machine learning models typically require fixed-length inputs, and CDR3s are likely to be different for different TCRs.

Loss of complexity: Single-point representations cannot capture the chemical (e.g., hydrophilic, hydrophobic) and biological functional complexity of amino acids, properties that are critical for understanding the interactions between TCRs and epitopes.

**Some ways to overcomes this problem:**

Embedding Vector Representation: some limitations can be overcome by using amino acid embedding vectors instead of a single point representation. With this approach, each amino acid is represented as a point in a high-dimensional space, and this point captures more properties of the amino acid (e.g., chemical properties, common contexts in the protein structure). This representation can be obtained by pre-trained embedding models, such as those similar using word embedding techniques in natural language processing.

**Potential Replacement**

**TCRdist**

TCRdist is a method for quantifying differences between T-cell receptor (TCR) sequences. Its advantages include:

Considering sequence similarity: Unlike One-hot coding, TCRdist captures biological similarity between sequences by calculating the "distance" between them.

Application to immunobank analysis: TCRdist enables researchers to identify functionally relevant clusters in high-dimensional TCR sequence data, which is critical for understanding immune response and disease-related immune cell diversity.

Flexibility and scalability: TCRdist provides a flexible framework to study TCR diversity and specificity, which can be applied to a wide range of different immunological studies and applications.

**GLIPH**

GLIPH (Global Indexed PCR High Throughput Sequence) is a tool for characterising TCR sequence commonality and specificity. Its advantages include:

Identification of functionally related TCR clusters: GLIPH is able to identify clusters of TCRs with common functionality, which is useful in areas such as vaccine design, disease surveillance and immunotherapy.

High throughput and accuracy: it can handle large-scale datasets and efficiently identify TCR sequences with potential common antigenic specificity.

Consideration of TCR sequence diversity: GLIPH takes into account the structural and functional diversity of TCR sequences, which contributes to a deeper understanding of the complexity of the immune response.