**Biasing of the RF bins with the verified negative features –** The specificity of the LLL antigen associated T cell receptors (TCRs) presented a unique problem in generating an accurate RF model. Because the LLL-specific TCRs all contained a specific alpha V-gene (TRAV12-2), the established model would often and correctly pick up on this feature within the CDR3 region and assign those containing this feature a score associated with high binding to LLL. This led to problems with accurately reading and separating between known non-binding TCRs that contained TRAV12-2. To remediate this issue, we have prioritized testing the individual RF models with these known, non-binding, TRAV12-2 containing TCRs to enforce the successful confirmation of the associated features of these TCRs. Through constant exposure and biasing these sequences, the RF model is better able to select specific features separating out these known binding LLL TCRs and TRAV12-2 containing but non-binding TCRs. This resulted in a marked improvement to previously unpresented TCRs expressing TRAV12-2 with known binding affinity, increasing the accuracy of determination from 66.05% to 82.55% with only a small decrease to overall accuracy of determination of new TCRs of 92.64% to 92.23%. Below is the discussion of the addendum modifications done to bias the model.

TRAV12-2 Without Biased Binning TRAV12-2 With Biased Binning

(AUC=66.05%) (AUC=82.55%)

Graphical user interface, chart

Description automatically generated Graphical user interface

Description automatically generated

Inputs:

**Largeneg:** The large array containing all the negative TCRs from the initial sampling pool, minus those containing TRAV12-2. Must contain the Alpha V-gene, Alpha J-Gene, CDR3 Alpha, Beta V-Gene, Beta J-Gene and CDR3 Beta sequences as described within the attached negative test file.

**Smallneg:** A smaller CSV file in the same format as previously specified of just the TCRs that are required to be biased and included in every random forest model. In our case this was the TRAV12-2 expressing TCRs that were known to be non-binding to the LLL antigen.

**Bincount:** The number of models that was previously generated without biasing the presentation of these TRAV12-2 expressing TCRs.

**Output name:** The attached label for each of the resulting TCR files designating the new collection of TCRs that need to be used to generate new individual RF models.

Outputs:

**New TCR files:** Individual csv that match the format of attached test files. These files are effectively the bins specified within step 3 of our established pipeline and given an existing dictionary generated prior, can be encoded and used to generate a single RF model that will be collected into a grouping of RF models like the process done by Step 3. This collection can be treated as a normal collection of RF models and be used in Step 4 of our established pipeline like normal.