Supplementary Materials for `ActiveTCR`

Algorithm 1: ActiveTCR

Inputs : $D = (X_{tcr}, X_{epi}, y)$: Initial training set
$P_u = (X_{tcr}, X_{epi})$: Unlabeled TCR-epitope pool `// only for use case a`
$P_l = (X_{tcr}, X_{epi}, y)$: Labeled TCR-epitope pool `// only for use case b`
Q: Query strategy
k: Query size per iteration

Outputs: M_t: Optimized prediction model at iteration t
L_t: Queried TCR-epitope pairs at iteration t

Initialize $L_0 \leftarrow \varnothing$, $t \leftarrow 0$

while *stopping criterion is not met* **do**
  Train TCR-epitope binding affinity prediction model M_t on D;
  if *use case* `a`*: reduce annotation cost* **then**
    $L_{t+1} \leftarrow Q(M_t, P_u, k)$ `// select` k `most informative samples from` P_u
    Obtain wet-lab annotation of L_{t+1};
    $P_u \leftarrow P_u \setminus L_{t+1}$;
  else if *use case* `b`*: reduce redundancy* **then**
    $L_{t+1} \leftarrow Q(M_t, P_l, k)$ `// select` k `most informative samples from` P_l
    $P_l \leftarrow P_l \setminus L_{t+1}$;
  $D \leftarrow D \cup L_{t+1}$
  $t \leftarrow t + 1$
end

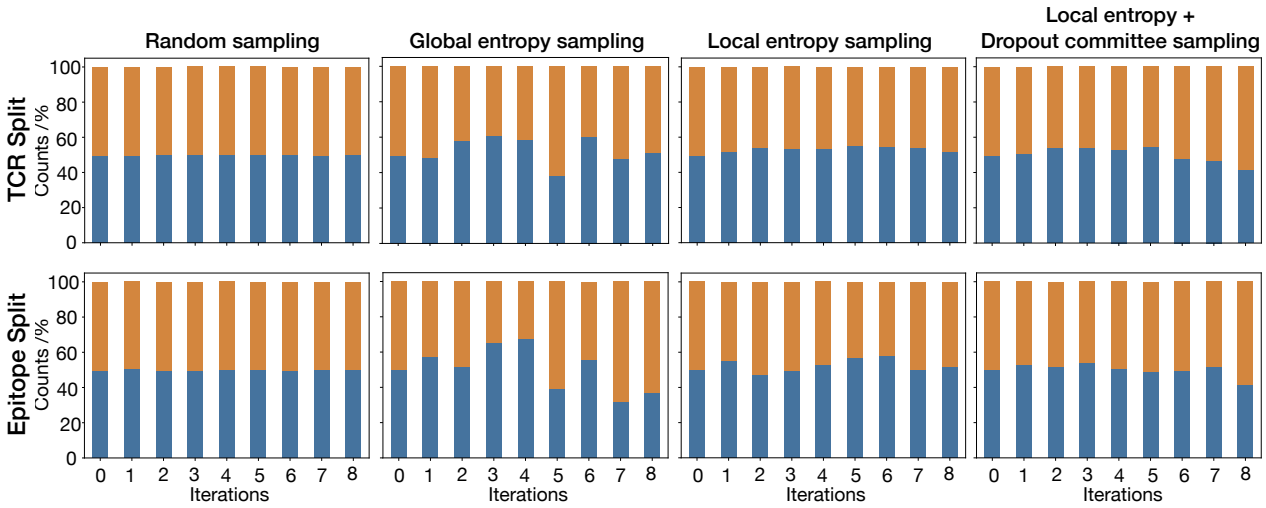

Fig. S1. Ratio of the queried positive (colored in orange) and negative (colored in blue) pairs, selected by different query strategies. The top row is For TCR split and the bottom row is for epitope split. Each column represents a query strategy.

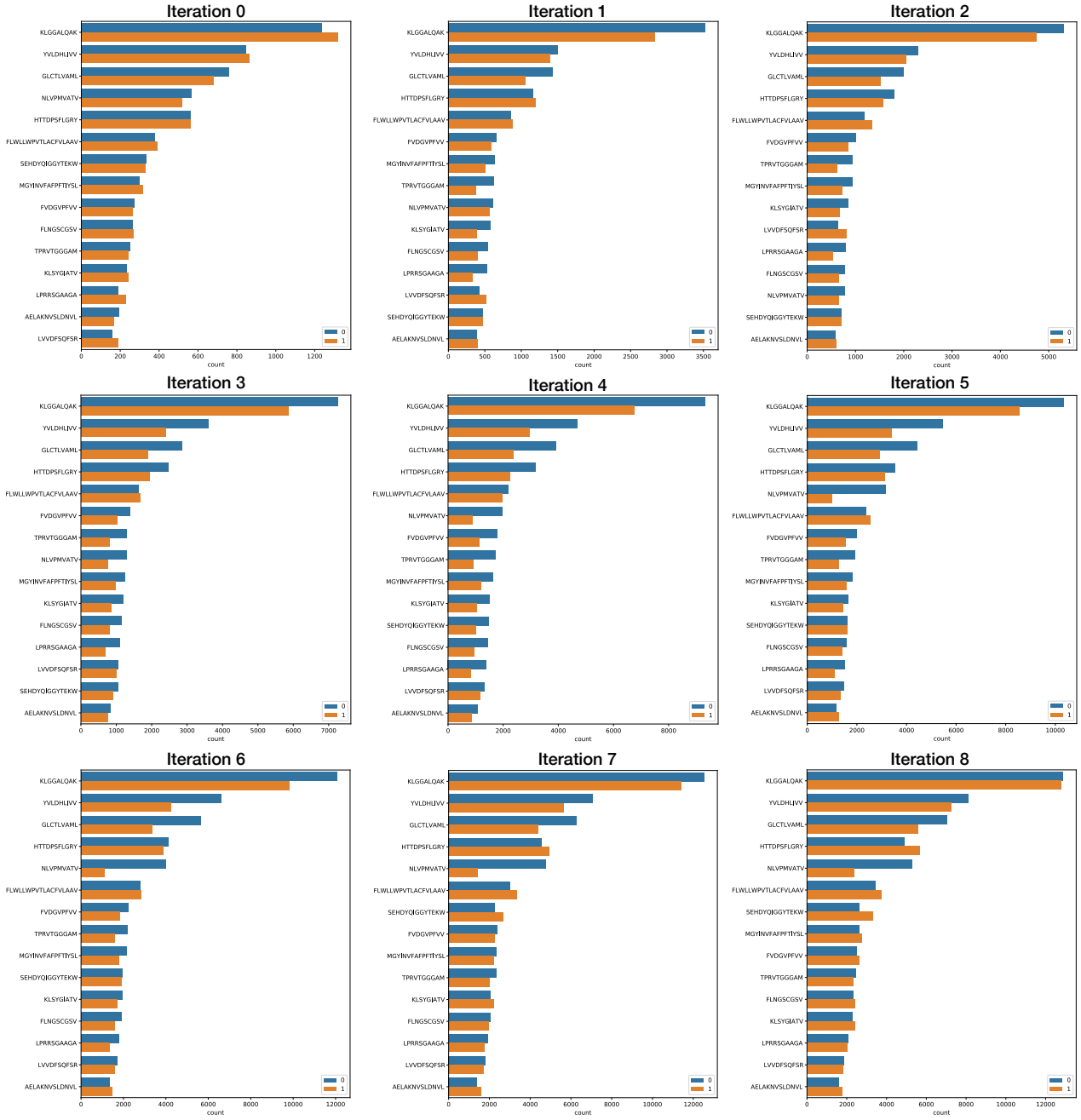

Fig. S2. Number of the queried positive (colored in orange) and negative (colored in blue) pairs, selected by *global entropy sampling*. Epitope-specific numbers are measured for the top 15 frequent epitopes, and on epitope split. Values at each iteration are the cumulative number of the queried pairs.

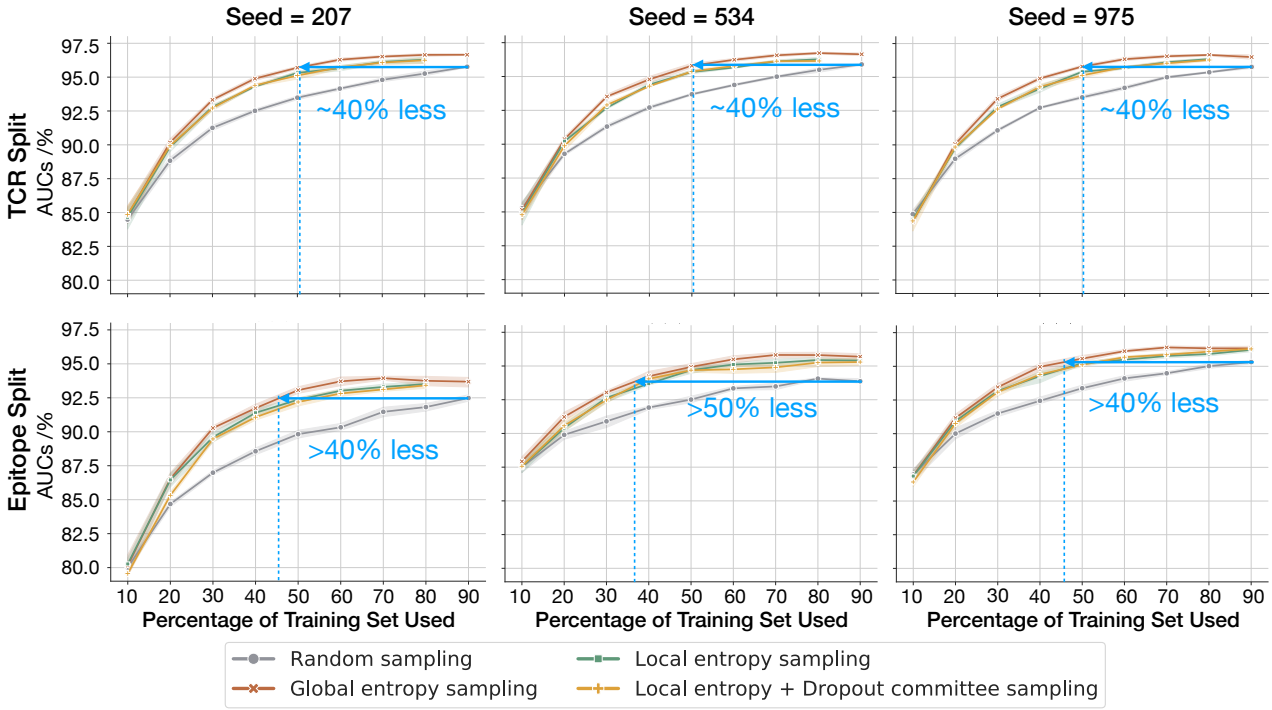

Fig. S3. Performance of `ActiveTCR`in reducing annotation costs for different initial training sets. Each row represents TCR split (top) and epitope split (bottom). Each column represents different random seeds to choose the initial training set. Average (solid line) and standard error (band) AUC of 5 independent runs for each query strategy are reported. The amount of annotation cost saved using `ActiveTCR`is represented by blue arrows and numbers.

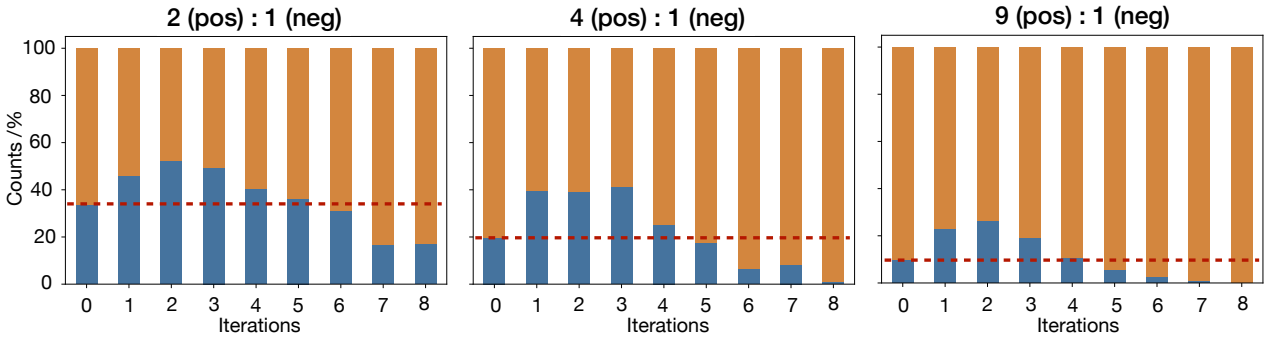

Fig. S4. Ratio of positive (colored in orange) and negative (colored in blue) pairs queried by *global entropy sampling*. The initial training set had different positive-negative ratios, which are shown as red dashed lines. The left panel has a 2:1 ratio, the middle panel has a 4:1 ratio, and the right panel has a 9:1 ratio.