

Head to head of feature sets over maximum of 102 problems

t-tests between accuracy distributions for each feature set and problem combination were calculated and p-values corrected using Holm–Bonferroni method for each pairwise feature set comparison over all problems. Tile labels follow a wins/ties/losses format.

