**The main bottleneck is "too few trainable columns" - 77.6% of the mutation columns are invalid columns (pos\_count <= MIN\_POS or c\_est fails), resulting in many species having almost no entries in the "unreported and with valid predictions" set; coupled with the strict threshold of 0.7, the threshold candidate table has very few species. The topk file actually covers all species (you verified it as character(0)), indicating that the lack of candidates is mainly caused by the threshold and invalid columns.**

**Possible Improvment**

**per\_species: For most species, n\_unlabeled is large, but n\_valid\_pred (unlabeled and with valid p\_true) is significantly smaller, and n\_above\_thr is even smaller → indicating that there are too few available prediction columns + the threshold is too high.**

****

****

**mut\_summary: The percentage of columns with any\_valid=FALSE is 0.7755, and the pos\_count of the first 20 columns you printed is all 1 → This conflicts with the current MIN\_POS=3, so these columns are considered invalid.**

****

****

**missing\_in\_topk: Empty set → The topk file already contains all species; the problem lies primarily with the threshold file.**

****