全部的大纲：

Step 1. 基于 ASReview 的耐利福平突变文献筛选与模型优化

为系统收集与整理利福平（Rifampicin, RIF）耐药相关突变信息，本研究首先通过主动学习框架 ASReview 进行文献检索与半自动筛选。检索式设定为：

((rifampicin OR rifampin) AND (resistance OR resistant) AND (mutation OR polymorphism OR variant))

从主流数据库检索到的初始文献集合导入 ASReview 后，首先执行一次主动学习轮次（Round 0）：对系统初始排序靠前的 109 篇文献 进行人工标注，其中 46 篇 被确定为与 RIF 耐药突变直接相关，这部分人工标注结果作为初始训练集用于下游模型微调。

随后，本研究使用两种语言模型——Sentence-BERT (all-MiniLM-L6-v2) 与 SciBERT (allenai/scibert\_scivocab\_uncased) 对其余未标注文献进行自动分类预测。通过对比两模型在置信阈值 0.52 下的预测结果，共发现 1144 篇模型预测不一致的文献（即两模型分类结果不同）。从中随机抽取 113 篇文献 进行人工复核，新增 46 篇相关文献。

在第二轮模型不一致分析（Round 2）中，再次抽取 113 篇预测分歧文献 进行人工标注，新增 25 篇相关文献。

经三轮迭代，累计获得 335 篇标注文献（含相关与不相关两类），形成最终训练集。

随后，使用 SciBERT 模型 在完整标注集（334 条可用记录）上进行微调与 5 折分层交叉验证（Stratified 5-fold CV）。模型性能如下：

最佳阈值（Best Threshold）：0.1420 ± 0.0546

精确率（Precision）：0.4059 ± 0.0655

召回率（Recall）：0.8283 ± 0.0702

F1 分数：0.5385 ± 0.0434

微调完成后，将该模型应用于剩余 3139 篇未标注文献，预测得到 142 篇可能与 RIF 耐药突变相关的候选文献，用于后续数据提取与突变信息补充。

简言之，Step 1 通过“主动学习 + 模型不一致驱动的人工复核”策略，显著提高了相关文献筛选的覆盖率与效率，最终构建了高质量的突变来源文献集。

Step 2. 新突变数据的整合与标准化处理

在获得新的候选文献集合后，对其与原有突变数据库进行比对与去重。通过 DOI 匹配和人工校对，共识别出：

与既有文献重复（相同 DOI）：37 篇

为新增文献（不同 DOI）：170 篇

无 DOI 记录的文献：11 篇

其中，未被数据库覆盖的文献（尤其是 2015 年之后发表的研究）被纳入补充数据整合流程。

从这些新增文献中提取并规范化氨基酸突变信息（AA position 与 mutation type），统一参考物种为 Escherichia coli。对于来源于其他物种的突变记录，通过序列比对（alignment）将突变位置映射到 E. coli 编码体系下的等效位点，以便于跨物种分析。

所有新数据被整理并汇总于 Newrifmutdata/latestnewdata.xlsx 文件中，作为后续 unsupervised 与 supervised ML 分析的扩展输入。该文件包含标准化后的氨基酸替换形式（如 D516V, H526R, S531L 等）、来源物种及对应文献编号，从而保证突变注释的统一性与可追溯性。

Step 3.supervisedMl

Data and Representation

Object: Multi-species rpoB site-level mutation matrix (species × mutation; binary values/frequency are unified into binary values: presence = 1, absence = 0).

Filtering: Remove extremely low-frequency/single-sample mutation sites; retain a sufficiently informative site set (used for the "Full Mutation Map" and "Top-30 Mutation Map" visualizations).

Confounding Assessment and Stratification (confounder score → high / mid-high)

Calculate the confounder score for each species (based on comprehensive metrics such as known source bias, sequencing depth/sample size, and publication bias).

Stratification by Thresholding:

High group: High confounder score (confounder score >0.7 ,potentially more biased, independent stability assessment first).

Mid-high group: Second-highest confounder score (0.3<score< 0.7,serves as a control stratum for analysis parallely).

Clustering Strategy Grid (algorithms × distances)

Algorithms: K-means, GMM, hierarchical clustering (HDBSCAN is optional for robustness testing).

Distance/Similarity: Euclidean, Cosine, (optional) Jaccard/Manhattan.

2D Grid Combination: Run the full combination for both the high and mid-high groups.

Model Selection and Stability Assessment

Primary Score: Silhouette (cosine distance is preferred for evaluating "pattern similarity" for binary/sparse data).

Stability: Re-run concordance of subsampling/bootstrapping (median and IQR of NMI/ARI can be reported).

Determining the Top Three: Sort by primary score. If the scores are close, weight the decision based on stability and interpretability (biological plausibility).

**Visualization and Comparison**

Heatmap 1 (Full Mutation): A binary matrix of species × all loci; rows are annotated with cluster groups, and columns are loci.

Heatmap 2 (Top-30): Top 30 sites filtered by overall frequency/information gain (easier to identify patterns).

**Upsetplot 3(top10)&all:**

Step4.

## ***1. Data Preparation***

## Known RIF-resistance mutations are used as positive samples (Positive).

## Unknown/unreported mutations are used as unlabeled samples (Unlabeled).

## A binary matrix (X\_dense) of species × mutations is formed.

## ***2. Model Selection and Training***

## Random Forest is used as the base classifier for supervised learning.

## Positive samples and unlabeled samples are combined using the PU-learning framework to build a model.

## During training, some known mutations are randomly masked to evaluate the model's resilience.

## ***3. Validation Strategy: Mask-then-Recover***

## Mask some known mutations.

## Use the model to predict candidate mutations.

## Calculate Recall@K to assess whether the masked mutations can be recovered in the top K predictions.

## ***4. Candidate Mutation Generation***

## **Top-K Strategy:** Select the top K mutations with the highest predicted probability for each species.

## Threshold Strategy: select mutations with p\_true ≥ τ(0.7) Generate a list of candidate mutations for subsequent analysis and experimental verification.

5. Novelty Filtering

Definition of novel. A mutation is “novel” if it does not appear in our compiled set of previously observed non-lab mutations after mapping all records to a unified E. coli rpoB amino-acid coordinate

Construction of the non-lab set. We parse the Google Sheet (Origin ≠ “Lab mutant”), map records to E. coli coordinates and build a blacklist S of observed mutations (two interchangeable modes):

**Global de-duplication**: S = {Mutation} across all species (default).

**Species-specific de-duplication** (optional): S = {(Species, Mutation)}.

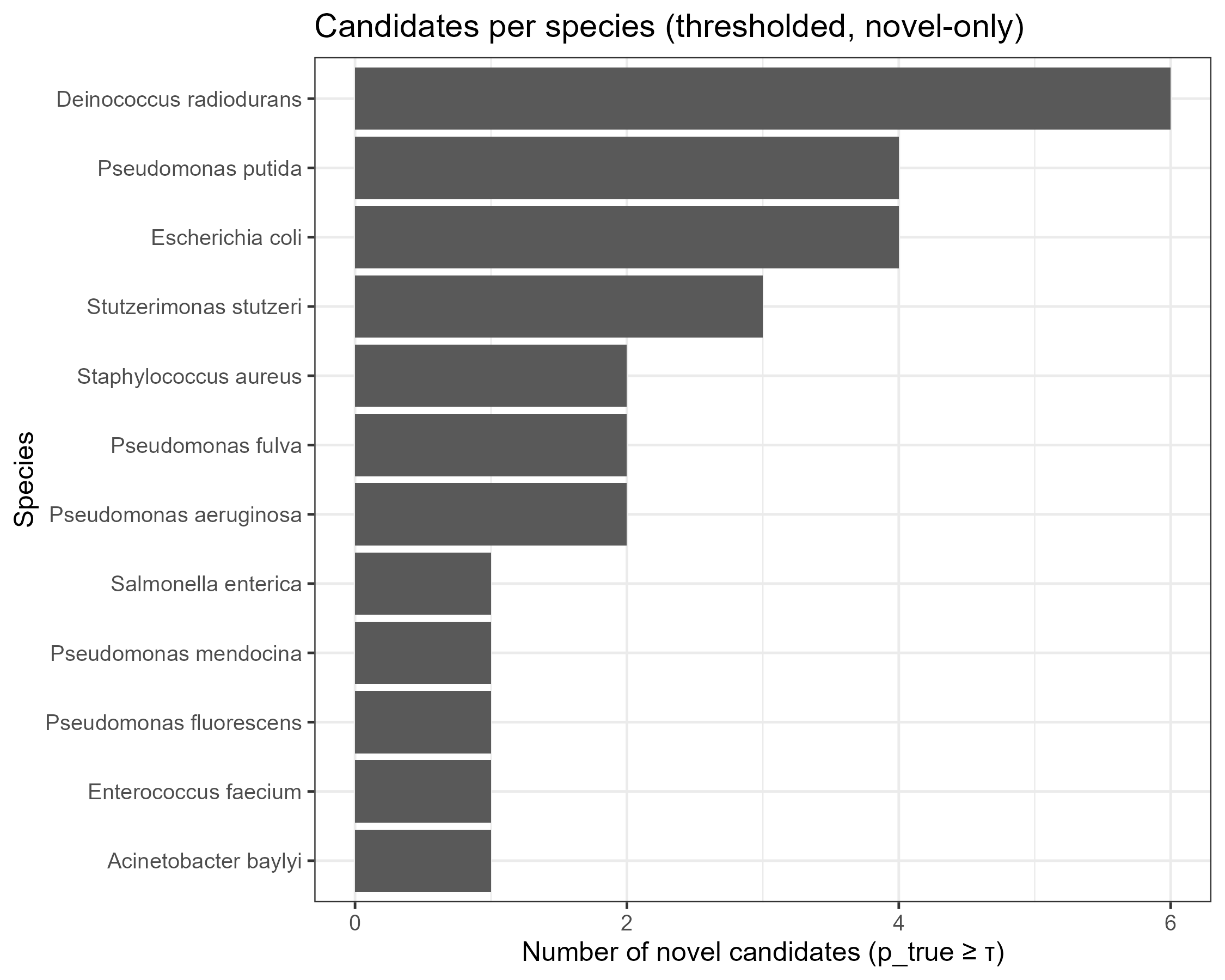
Filtering. We do not retrain the model. We filter the scored candidates by excluding anything in S.

For Top-K, we refill from the remaining pool (highest p\_true) until reaching K or the pool is exhausted.

For Threshold, we simply keep items with p\_true ≥ τ and Mutation ∉ S.

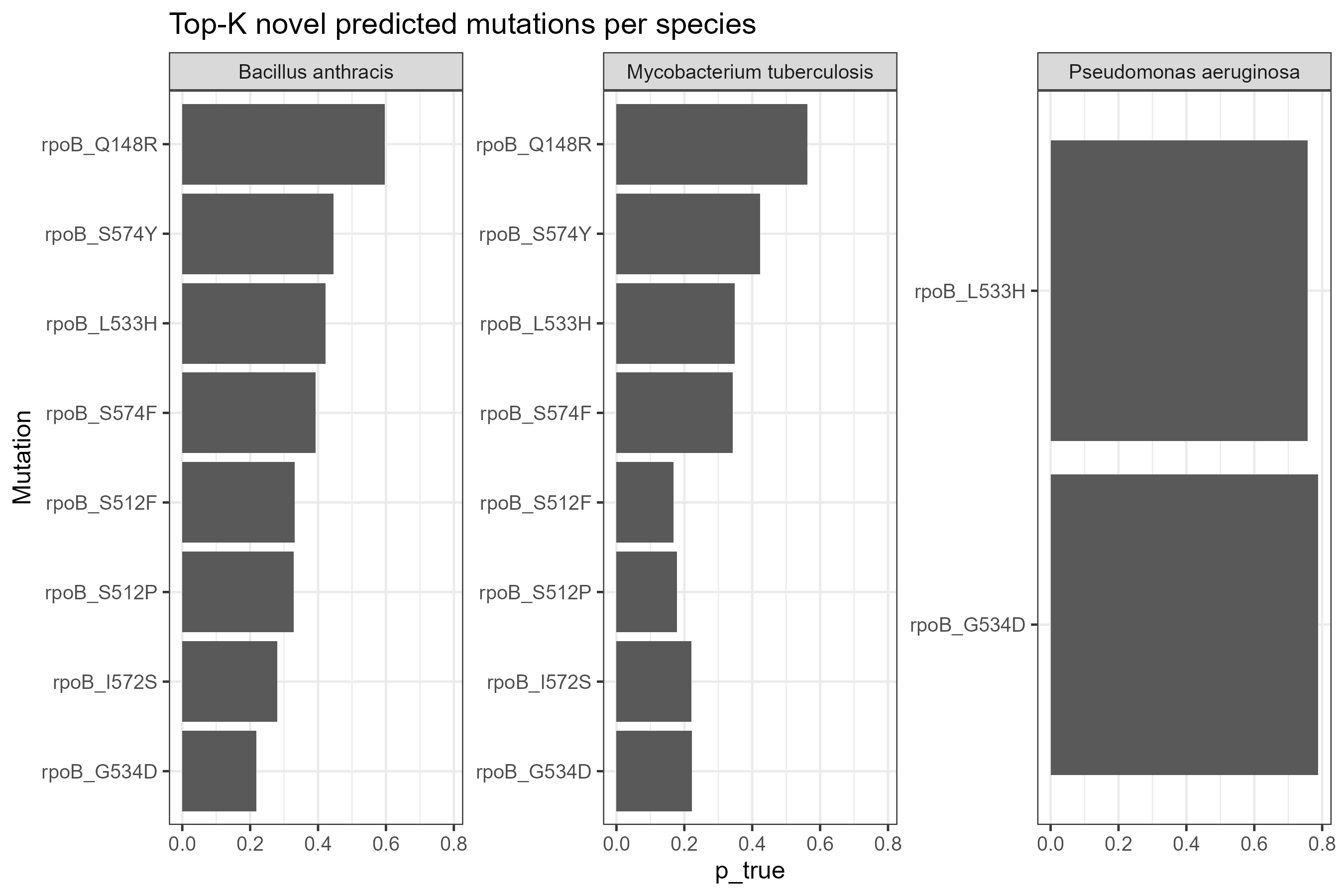
**Analysis of Figures：**

Fig1 – Novel threshold method: the number of candidate species



结论：Candidates were highly heterogeneous (Deinococcus radiodurans≈6, Pseudomonas putida / E. coli≈4,rest of species

Fig2 – novel per-species Top-K column figs

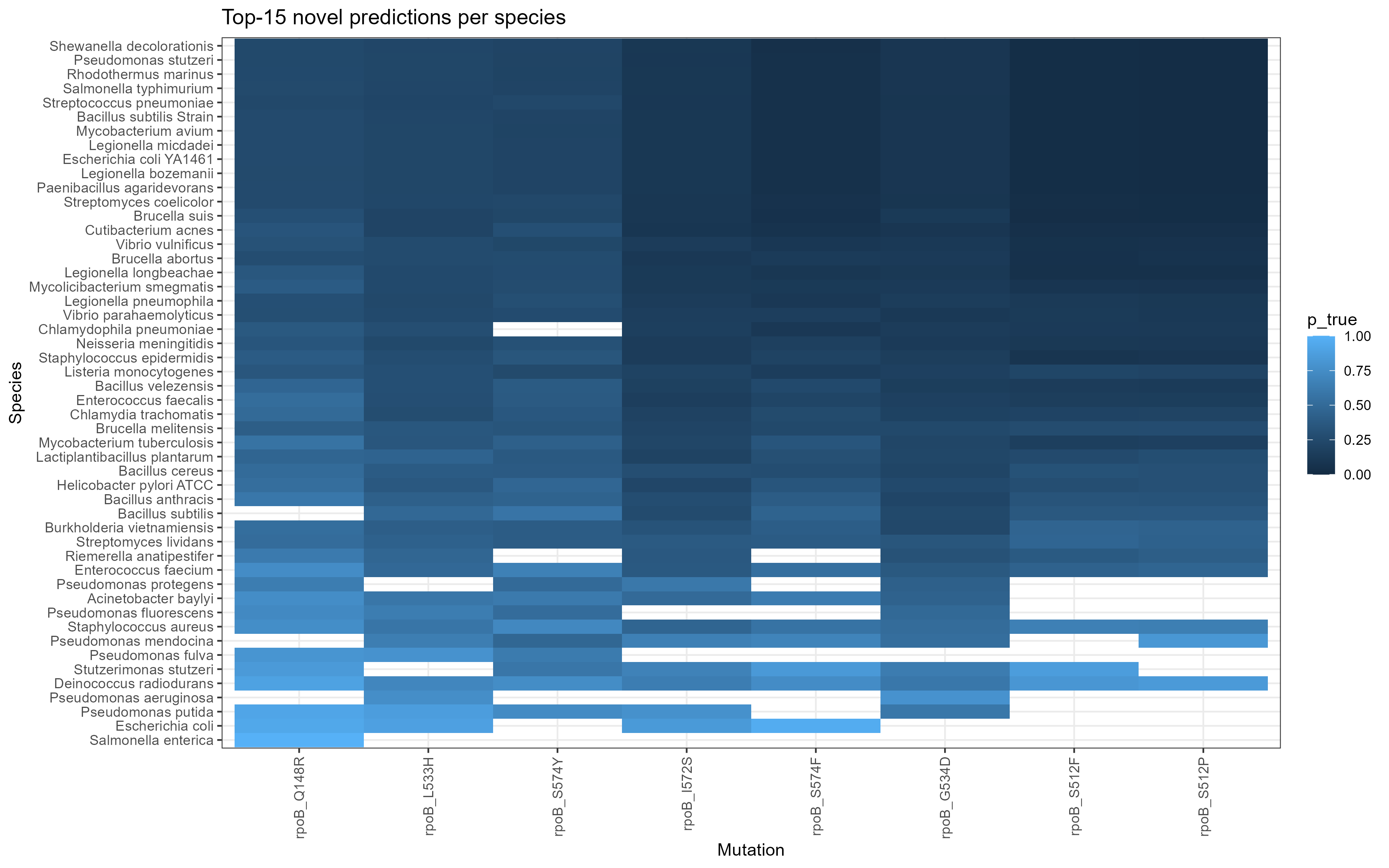


Both B. anthracis and M. tuberculosis ranked rpoB\_Q148R very highly (p\_true ~0.7–0.8), along with the S574 series and L533H.

P. aeruginosa only had two remaining entries (L533H and G534D), both of which were highly ranked (~0.75–0.8).

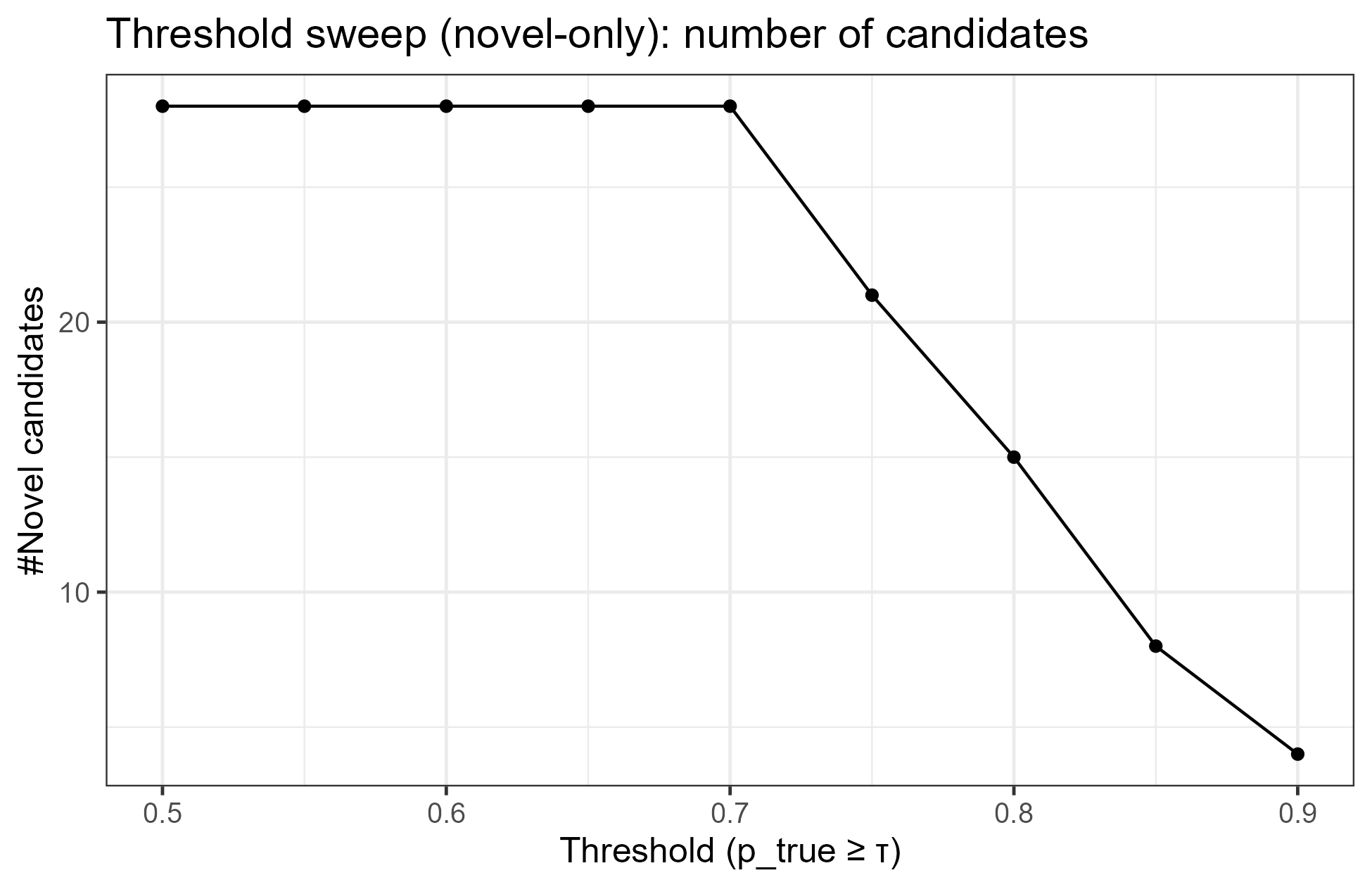
Implications: Cross-species co-occurring sites (148, 533, 574, 512, 572, etc.) appear repeatedly, and the model considers these sites "unreported but likely to occur."

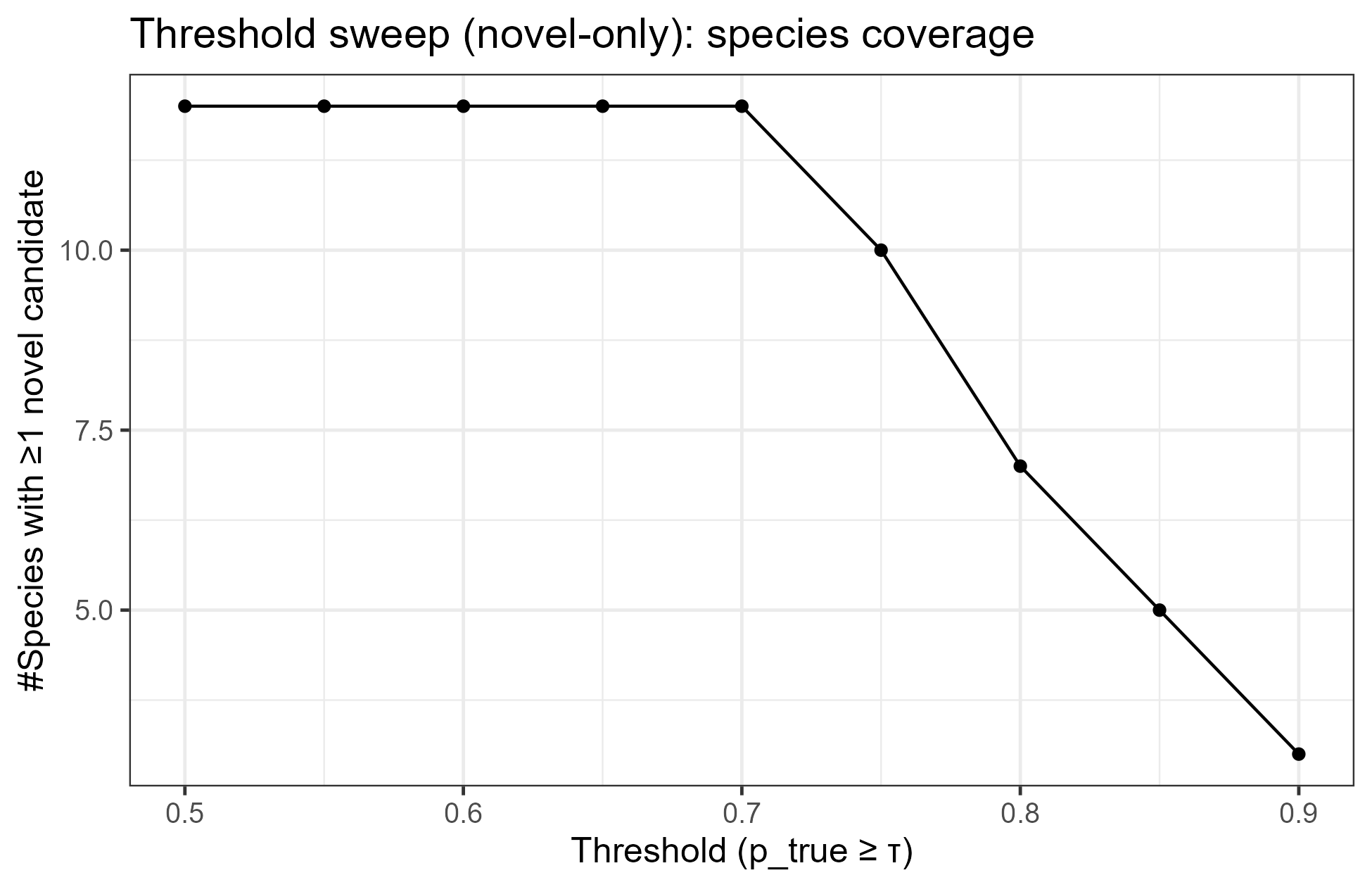
Fig3 – novel Top-15 in species heatmap



The columns are basically the hotspot sites mentioned above. Many species have dark blocks (high scores) in these columns, and white blocks indicate that the species does not list the mutation in the top 15.

Fig4a/4b – novel threshold scanning

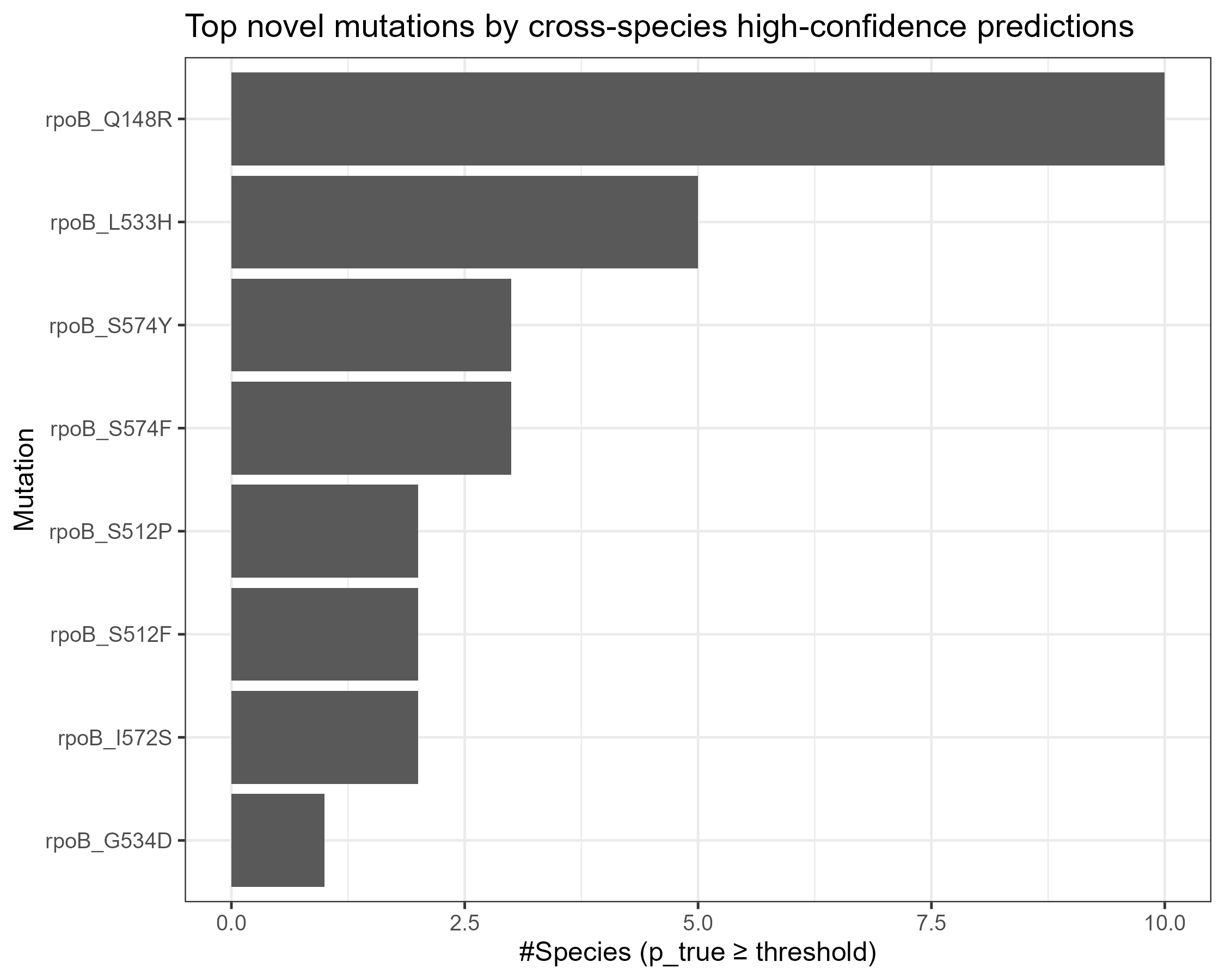




Conclusion: The number of candidates and coverage remain high at 0.7 (but decrease significantly after 0.75).

Thus, in the paper, I can say that τ≈0.7 is a good balance; >0.8 significantly sacrifices coverage.

Fig5



rpoB\_Q148R entered high scores in the most species (~10), followed by L533H, S574Y/F, S512P/F, I572S, and G534D.

**Masked vs Unmasked\_cf description**

**Masked**: Mask some known true positives as unlabeled (0), then train and retrieve them in the unlabeled pool, and calculate Recall@K. → Closer to the difficulty of "real discovery", should be more conservative.

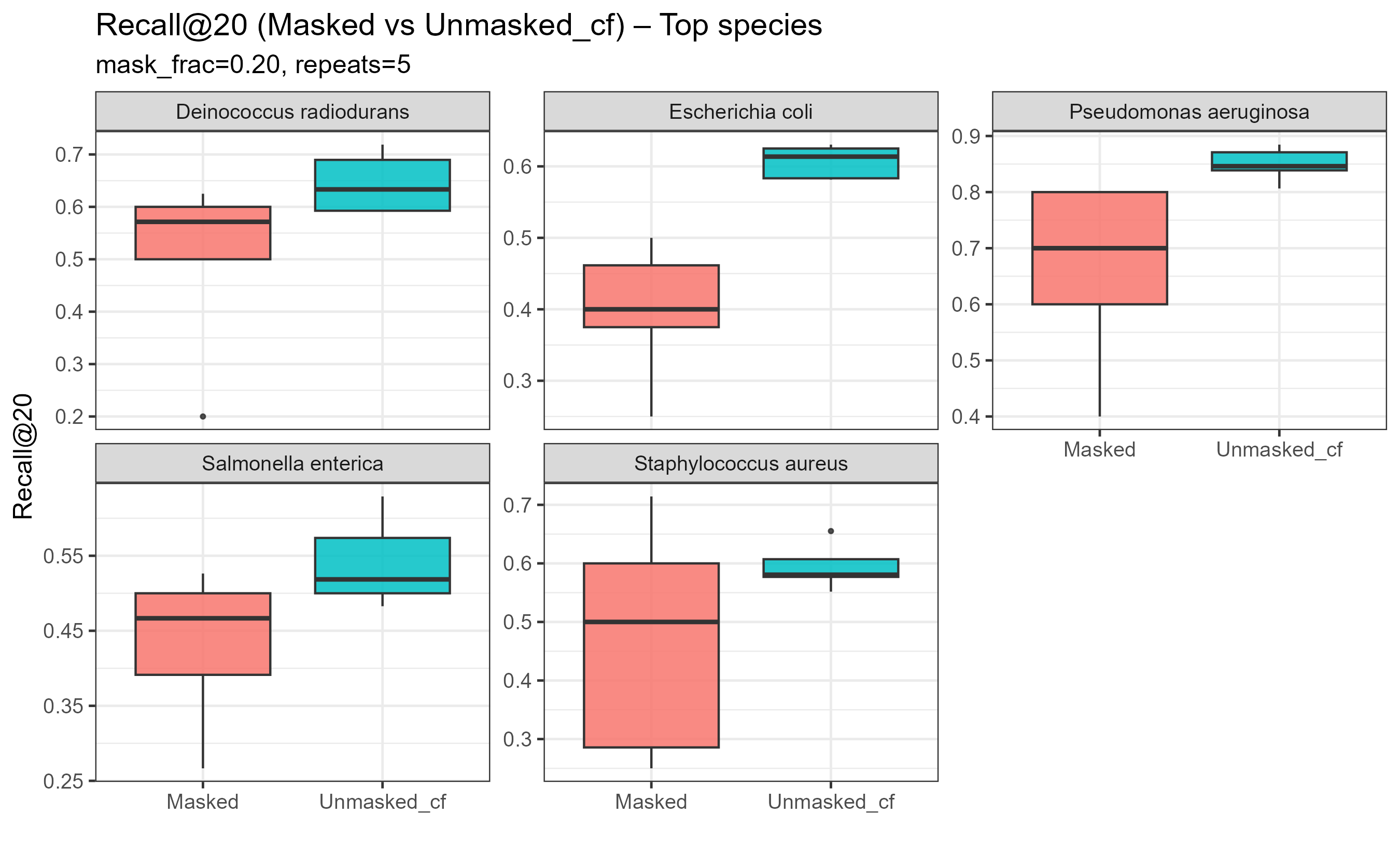
**Unmasked\_cf (counterfactual)**: These true positives are not masked (still participate in training as positive examples), but we use the scores of the same model to put them into the unlabeled pool, and compare them with the unlabeled scores to calculate "how high they would be ranked if they were also unlabeled." This is an "upper bound/optimistic" reference, usually higher than Masked.

Box plot (gray/pink/blue rectangle): The interquartile range of the group (Masked or Unmasked\_cf) under multiple repeats.

Therefore, the "height" of the box reflects the dispersion of the distribution (IQR = Q3 − Q1); the "higher" means the greater the fluctuation.

"Whiskers" (thin lines outside the box): The minimum/maximum values ​​extending from the box to non-outliers (usually the extreme values ​​in the range Q1 − 1.5 × IQR to Q3 + 1.5 × IQR).

Black dots: Outliers, i.e., individual replicates that fall outside the whiskers (significantly lower or higher under certain folds/samples).·



· 五个物种里，Unmasked\_cf 全都高于 Masked（预期现象）。Unmasked\_cf 是“未被遮蔽的真阳性”的反事实上界：它们仍作为正例参与训练，所以得分更乐观；而 Masked 是把一部分真阳性塞进未标注池再去“找回”，更接近真实难度。

· · P. aeruginosa：表现最好，Masked 的中位也在 ~0.7；Unmasked\_cf 接近 0.85–0.9，gap 小，说明模型对该物种的排序很稳。

· · E. coli / Salmonella / S. aureus：Masked 中位在 0.4–0.55，且 S. aureus 方差较大，说明重复采样下稳定性稍弱（多半因为该物种的正例分布更稀/不平衡）。

· · D. radiodurans：两组都高（~0.6–0.7），整体稳

·