Final Research Report

Course Code: BIOX7011  
Student Name: Yu Sun  
Student Number: 45105764  
Supervisor: [Jan Engelstaedter, Affiliation]  
Date: [Submission Date]

# Abstract

This section provides a concise summary of the entire project (200–300 words). It should clearly outline the research aims, the main methodologies used, the key findings, and the major conclusions derived from the study.

# Introduction

1. 抗菌素耐药性（AMR）的全球危机

1.1 全球流行与公共卫生威胁

简述AMR的定义与机制（突变、基因水平转移、选择压力）

Since the start of 21st century Antimicrobial Resistance (AMR) has become one of the most pressing threats to global public health. The World Health Organization estimates more than 10 million people will die annually by 20250 from AMR-related infections .AMR not only makes it more difficult to treat infections, but also significantly raises healthcare costs and length of hospitalization, placing greater pressure on public health systems in low-income countries . Among the many drug-resistant pathogens, drug-resistant tuberculosis (DR-TB) is of particular concern. approximately 450,000 people worldwide will have rifampicin-resistant tuberculosis (RR-TB) in 2022, with the majority of cases also showing resistance to isoniazid, thus constituting multidrug-resistant tuberculosis (MDR-TB) .

AMR 的成因与传播跨越 **人类、动物与环境**三重界面（One Health）：人群临床用药、畜牧与水产中的抗菌药使用、废水与环境耐药基因库之间彼此联动，要求卫生、农业与生态等多部门协同。其中，医院获得性感染中的 **ESKAPE 病原体**以**多重耐药**与**治疗失败**著称，是全球政策与研发的优先对象。虽然本文聚焦 **利福平（Rifampicin, RIF）** 的耐药机制（主要通过 **rpoB** 位点突变）及其跨物种分布与预测，但其研究逻辑与 ESKAPE 的抗性进化具有共通性：在持续的药物选择压力下，关键靶点突变、外排泵增强、酶介导失活、生物膜等机制可以**并行或叠加**出现，导致临床治疗窗口收窄。将 **RIF-耐药** 的分子谱系学证据与 **ESKAPE** 的临床与流行病学监测相结合，可为制定更具针对性的**经验用药策略**、**耐药预警指标**与**新药/伴随诊断**研发提供实证基础

在具体的抗药性细菌中，Louis B. Rice提出了“ESKAPE 细菌”即由 Enterococcus faecium, Staphylococcus aureus, Klebsiella pneumoniae, Acinetobacter baumannii, Pseudomonas aeruginosa, Enterobacter 等六大耐药病原的首字母组合而得。2008 年，Louis B. Rice 在其评论中指出，这些病原因为其高度的抗菌药物逃逸能力，成为医院获得性感染中最具挑战性的耐药威胁。

本研究针对的是另一类耐药机制——尤其关注于 Rifampicin 抗药性突变（rpoB 位点突变）在不同菌种间的分布与共演化。尽管 Rifampicin 耐药更多聚焦于结核分枝杆菌及相关病原，但从机制角度看，其耐药路径也与 ESKAPE 细菌中的多重耐药机制同样具备“选择压力下突变累积与传播”的特征。因此，探讨 Rifampicin 抗药突变与 ESKAPE 群体耐药模式之间的关联，有助于从更广义的抗菌耐药视角理解这些关键病原体如何“逃逸”治疗，并为监测、预防及治疗策略提供新的思路。

分枝杆菌属（Mycobacterium）以其独特的细胞壁结构和慢生特性，成为慢性病原感染的典型代表。其长期潜伏、细胞内存活及天然耐药性使其在感染生物学与药物耐受研究中占据核心地位。结核分枝杆菌（M. tuberculosis）作为该属最重要的病原，至今仍导致全球最严重的慢性感染疾病之一。

时至今日结核病（TB）仍是全球最具负担的慢性传染病之一。根据世界卫生组织（WHO, 2023）报告，2022 年全球约有 **1050 万** 新发 TB 病例和 **130 万** 相关死亡病例。其中，**耐药结核病（Drug-Resistant TB, DR-TB）** 的持续流行构成了抗击结核的最大障碍。  
**多重耐药结核（MDR-TB）** 指同时对至少异烟肼（Isoniazid）与利福平（Rifampicin）耐药的菌株；而 **广泛耐药结核（XDR-TB）** 则在 MDR 基础上进一步对氟喹诺酮类及至少一种二线注射药物（如阿米卡星或卷曲霉素）耐药。2022 年约有 **45 万例** 新发利福平耐药病例（其中多数为 MDR-TB），治愈率仅约 **60%**（WHO Global TB Report 2023）。

Rifampicin（RIF） 是目前针对分枝杆菌属（尤其是 M. tuberculosis）最有效的抗生素之一。

它属于 利福霉素类（Rifamycins），通过结合 RNA 聚合酶 β 亚基（由 rpoB 基因编码）来抑制细菌转录。

因此，RIF 可以有效杀灭活跃复制与部分静止状态的分枝杆菌，是打破“慢性感染壁垒”的关键药物。

在 WHO 推荐的“短程复合化疗方案”（HRZE 方案）中，RIF 与异烟肼（H）、吡嗪酰胺（Z）、乙胺丁醇（E）联合使用，是全球 TB 控制战略的基石。然而，RIF 的单药耐药（Rifampicin-Resistant TB, RR-TB）即被视作 MDR-TB 的预警指标，因为 rpoB 位点突变往往与其他药物耐受突变共存。近年来，基于 rpoB 的快速分子诊断（如 Xpert MTB/RIF 和 line-probe assay）已成为临床检测耐药结核的关键手段。

自 1960 年代问世以来，利福平（RIF）一直是结核病标准化治疗方案中的首选一线药物。其通过与 RNA 聚合酶 β 亚基（由 rpoB 基因编码）结合，抑制 RNA 合成，从而阻断细菌转录活动。由于其杀菌力强、组织渗透性好、对细胞内静止菌亦有效，RIF 被视为实现短程化疗（6 个月标准疗程）的关键药物。

在化学结构方面，利福平（Rifampicin, RIF）属于利福霉素类（Rifamycins）抗生素，为一种半合成的芳香族大环内酯化合物，分子式为 C₄₃H₅₈N₄O₁₂，分子量约 823.94 Da。其化学骨架由一个\*\*芳香萘醌环（aromatic naphthoquinone nucleus）与一个含氮吡嗪酮结构（piperazine-like or hydrazone moiety）\*\*通过脂肪族支链相连，形成高度共轭的刚性结构。这种结构赋予了 Rifampicin 优良的脂溶性，使其能够穿透结核分枝杆菌（Mycobacterium tuberculosis）富含分枝酸的细胞壁。

Rifampicin 的药理活性核心在于其能特异性结合细菌 RNA 聚合酶 β 亚基（由 rpoB 编码），通过与其“rifampicin binding pocket” 形成氢键及疏水相互作用，阻断 RNA 链的延伸，从而抑制转录。该作用机制对革兰阳性菌、部分革兰阴性菌及放线菌均有效，但对真核细胞 RNA 聚合酶无明显影响，因此选择性较高。

Rifampicin 口服后吸收迅速，在空腹状态下生物利用度约为 70–90%。其高脂溶性有助于广泛分布于机体组织，包括肺组织、脑脊液、肝脏、肾脏及吞噬细胞内，尤其在结核病灶组织中浓度较高。血浆蛋白结合率约 80%，半衰期为 2–5 小时（长期用药可因酶诱导而缩短）。

Rifampicin 主要在肝脏经细胞色素 P450（CYP3A4）及 CYP2C 系统代谢为去乙酰化活性产物（desacetyl-rifampicin），仍保留部分抗菌活性。RIF 同时是强效 诱导剂，可显著加快多种药物（如抗病毒药、激素、抗凝药）的代谢，可以作为促进药物相互作用的诱因。

药物及其代谢物主要通过胆汁排泄，部分经肠肝循环重新吸收，因此常见粪便呈橙红色。约 30% 的药物以原形或代谢物形式经尿排出。由于 Rifampicin 的代谢依赖肝脏功能，肝损伤或酶诱导状态（如长期饮酒或共用苯妥英、卡马西平）可显著影响其血药浓度。

在临床应用方面，RIF也被扩展至用于治疗多种细菌感染和预防。在麻风病（Leprosy, Hansen’s disease）的治疗中，世界卫生组织推荐将利福平与氯法齐明（Clofazimine）及氨苯砜（Dapsone）联合使用，构成多药联合疗法（MDT）核心方案，成人典型剂量为 600 mg 每月一次，疗程分别为多菌型 12 个月、少菌型 6 个月。

此外，利福平亦被用于某些假体相关感染（prosthetic joint infections, PJI）尤其由葡萄球菌（Staphylococcus spp.）引起的情况。研究指出，将利福平与其他抗菌药（如 β-内酰胺类或万古霉素）联合使用，可增强对生物膜内细菌的穿透与清除作用。例如一项涵盖 669 例患者的多中心观察研究显示，使用利福平联合治疗组的治疗失败率为 32.2%，而未使用组为 54.2%。

同时，利福平也被用于脑膜炎球菌（Neisseria meningitidis）携带者暴露后预防，旨在消除鼻咽部菌携带状态，从而减少爆发传播风险。

针对非结核分枝杆菌（NTM）感染，尽管不是其主要适应症，利福平常与大环内酯类、乙胺丁醇等联合应用，用于 Mycobacterium avium、M. kansasii 等病原的治疗

RIF的主要作用靶点是细菌的 DNA 依赖性 RNA 聚合酶（RNA polymerase, RNAP）。该酶复合体由多个亚基组成，包括 等，其中 亚基（RpoB） 是催化转录反应的核心成分，负责结合核苷酸并延伸 RNA 链。由于 RNAP 是所有细菌转录过程的必需酶，RpoB 在转录起始、DNA 解链、RNA 合成及转录泡稳定中发挥关键作用，因此成为多种广谱抗生素的重要靶点。

Rifampicin 通过与 RpoB 的特定位点结合来抑制转录，该位点位于 亚基内一个长度约 81 bp 的保守区域，称为 Rifampicin Resistance-Determining Region (RRDR)。该区域在结核分枝杆菌 (Mycobacterium tuberculosis) 中对应 RpoB 的第 426–452 位氨基酸残基。结构研究表明，Rifampicin 的芳香萘环结构能够嵌入 RNAP 的通道中，与多个关键残基（如 Ser450、His445、Asp435 等）形成氢键与疏水相互作用。结合后，Rifampicin 会阻止 RNA 链从起始阶段进入延伸阶段，使新生 RNA 链无法超过 2–3 个核苷酸长度，从而终止转录过程并导致细菌死亡。

然而，当 rpoB 基因发生突变时，特别是在 RRDR 区域内的氨基酸替换（如 S450L、H445Y、D435V 等），会引起 RpoB 结合口袋的构象和电荷环境改变，破坏 Rifampicin 结合所需的关键氢键网络与疏水相互作用。这些突变显著降低 Rifampicin 与 RNAP 的亲和力或阻止药物进入结合口袋，从而导致抗药性产生。约 90–95% 的 Rifampicin耐药结核菌株可在 RRDR 区检测到突变（Zhao et al., 2020）。此外，部分突变还可在一定程度上影响 RNA 聚合酶的催化效率，形成代谢成本与适应性之间的权衡，为耐药菌株的长期存活提供进化基础（Helmann et al., 2023；Das et al., 2020）。

因此，RpoB 不仅是 Rifampicin 的直接靶点，也是结核分枝杆菌耐药性最重要的分子标志之一。对 rpoB 突变的结构和功能研究，不仅为理解 Rifampicin 的作用机制提供了精确分子基础，也为开发新一代 RNA 聚合酶抑制剂和快速耐药诊断工具奠定了理论依据。

3. 利福平抗药性（Rif-resistance）的分子基础

3.1 rpoB突变与耐药机制

经典突变位点（S531L, H526Y/D, D516V）

突变如何改变Rif结合亲和力

不同突变导致不同水平的耐药性与适应代价（fitness cost）

3.2 其他辅助耐药机制

外排泵系统（efflux pump）上调

细胞膜通透性改变

代谢或氧化应激反应增强

联合用药时的交互耐药性（例如Isoniazid与Rif的共同压力）

3.3 耐药突变的可预测性与进化模式

同源位点的进化趋同（convergent evolution）

结构保守区域突变的适应性边界

不同物种rpoB序列差异下的结构敏感性

4. 已有的利福平耐药性研究进展

4.1 临床与实验研究的主要模型物种

结核分枝杆菌（Mycobacterium tuberculosis）

麻风分枝杆菌（M. leprae）

大肠杆菌（E. coli）、金黄色葡萄球菌（S. aureus）、铜绿假单胞菌（P. aeruginosa）

环境菌与土壤放线菌（例如 Streptomyces）

4.2 利福平抗药性数据库与突变数据积累

TB-Profiler、WHO RDB、CARD、ResFinder等数据库

重点说明这些数据库局限于临床病原体，非模式物种覆盖不足

4.3 跨物种耐药性比较与结构预测研究

比较不同物种rpoB突变的结构效应

分子动力学模拟（MD）或计算机辅助药物设计（CADD）研究

展示耐药性预测在分子演化、生物信息学、药理学上的意义

5. 研究空白与科学问题定位

5.1 跨物种耐药预测的必要性

环境细菌作为耐药基因储库（resistome）的潜在威胁

利福平使用后的环境残留导致选择压力

当前研究集中于少数病原体 → 缺乏对非临床细菌的预测框架

5.2 利福平抗药性突变的可迁移性问题

同源rpoB位点突变在不同物种间是否具有相同效应？

结构与序列差异对突变适应性的影响

预测新物种耐药突变的潜在意义：

指导未来病原风险评估

辅助药物再设计与靶点改进

6. 研究目标与本项目定位

6.1 项目总体目标

构建跨物种rpoB序列比较框架

利用机器学习预测不同细菌物种中可能导致利福平抗药的突变位点

6.2 科学意义与潜在应用

揭示抗药突变的保守与可塑性边界

为新出现或非模式细菌提供耐药风险预警

为后续抗生素设计、耐药传播监测提供基础

# Although Mycobacterium tuberculosis is the most intensively studied model for rifampicin resistance, rifampicin is exposed to a wide range of bacteria, both in clinical use and through environmental contamination. Therefore, this study aims to predict cross-species mutation patterns, which may provide insights into the evolutionary convergence of antibiotic resistance.

As humans truly explore the Earth, they encounter more than just E. coli or M. tb. Therefore, if humans rely on a limited set of RIF drugs to combat a wide variety of bacteria in nature, they are likely to exhibit varying adaptability or resistance. Therefore, this project aims to predict as many potential RIF-resistant mutations as possible in unconventional or less widely studied bacterial species.

# Materials and Methods

Materials：

1

你如何系统性地检索文献；

使用了哪款软件（ASReview）、版本号、核心功能；

使用了哪些关键词；

主动学习的流程（机器模型与人工复核如何结合）；

输出结果数量。

2

你具体使用了哪些语言模型（Sentence-BERT、SciBERT）；

微调方式（HuggingFace Transformers，PyTorch 环境）；

性能评估方法（5-fold CV + Precision/Recall/F1）；

软件库。

3

数据整合步骤（DOI 匹配、去重、跨物种突变映射）；

提及使用的辅助工具（Excel、Python、Biopython、Clustal Omega 等）；

输出文件（newrifmutdata.xlsx）。

4

输入数据矩阵（X\_dense\_high / X\_dense\_midhigh）；

使用的降维算法与聚类算法；

软件与库；

筛选 top methods 的依据（平均轮廓系数）。

5

说明：

目的（验证模型能否学习已知突变规律）；

方法（PU-learning, Random Forest）；

软件与实现（Scikit-learn）；

评估策略（Mask–then–Recover + Recall@K）。

Methods:

2. Methods

2.1 Data Sources and Literature Screening

2.1.1 Literature Search and Active Learning (ASReview)

To systematically collect literature reporting on rifampicin (RIF) resistance mutations, this study first tested multiple keyword combinations on the Web of Science website, including "RIF," "resistance," and "mutant screen," to search for as many articles as possible containing experimentally generated RIF resistance mutation data.

The search query was: ((rifampicin OR rifampin) AND (resistance OR resistant) AND (mutation OR polymorphism OR variant)).

Then, at the recommendation of my supervisor, I used the open-source active learning platform ASReview (version 1.3.1) for semi-automated screening.

After initially importing the literature, ASReview continuously updated the ranking using an active learning algorithm to maximize the model's efficiency in identifying relevant literature. The researchers manually annotated the first 109 articles in the system, obtaining 46 relevant samples as the initial training set. This was then combined with three rounds of model disagreement sampling, ultimately yielding a total of 335 labeled articles.

2.2 Language Model Fine-tuning and Document Classification

To further automatically identify articles with potential drug-resistant mutations, two models based on deep language representation were used: Sentence-BERT (all-MiniLM-L6-v2) and SciBERT (allenai/scibert\_scivocab\_uncased).

Model fine-tuning was performed in Python 3.10, relying on HuggingFace Transformers (v4.30) and PyTorch (v2.0).

Stratified 5-fold cross-validation was performed on each model on the 334 annotated samples, and the average precision, recall, and F1 score were calculated. The SciBERT model performed best (Precision = 0.41 ± 0.07, Recall = 0.83 ± 0.07, F1 = 0.54 ± 0.04) and was therefore used to predict the remaining unannotated articles.

2.3 Integration and Standardization of New Mutation Data

After comparing the candidate articles predicted by the model with the existing mutation database, a total of 37 duplicate articles, 170 newly added articles, and 11 articles without DOI records were identified.

For newly added articles, the mutation site, amino acid substitution pattern, and species of origin were extracted.

If the mutation originated from a species other than E. coli, Clustal Omega (v1.2.4) was used for sequence alignment, and the mutation position was mapped to the corresponding site in E. coli.

The resulting standardized mutation table (latestnewdata.xlsx) includes a uniformly formatted amino acid position (AA\_pos), mutation pattern (AA\_change), species name, and reference number.

2.4 Unsupervised ML

To identify the clustering structure of mutation spectra across species, the authors first constructed a binary mutation matrix X\_dense (rows = species, columns = mutations) using the Lab mutant dataset generated in the previous step.****To make the results more readable and intuitive, the authors filtered the species × mutation matrix (X\_dense) for species with low confounder scores (i.e., <0.3).****

Besids,after multiple attempts at plotting, the authors removed data from several species because their mutation data had too low overlap with other species. This was likely due to a misalignment in the coordinate system constructed using E. coli as the standard, or other issues beyond the scope of this project, leading to manual filtering.

UMAP (uwot v0.1.15) was used for nonlinear dimensionality reduction (n\_neighbors=15, min\_dist=0.3, seed=123). Four clustering algorithms were then applied to the embedding space: HDBSCAN (dbscan v1.1.11), DBSCAN, k-means, and GMM (mclust v6.0).

Three distance metrics (Euclidean, Manhattan, and Cosine) were combined to generate 12 schemes. The average silhouette score is calculated for each combination. The top three methods are selected based on the scores: Cosine–HDBSCAN, Euclidean–GMM, and Euclidean–HDBSCAN. The clustering results are visualized on their UMAP embeddings.

2.5 Supervised ML

Using the clustering results and confounder score from unsupervised clustering, the authors narrowed the species and mutations used for training and prediction. They then trained a supervised learning model to test its ability to learn the distribution patterns of drug-resistant mutations.

Using the Positive–Unlabeled Learning (PU-learning) framework, a binary classification model based on Random Forest (scikit-learn v1.3) was trained using known drug-resistant mutations as positive examples and unlabeled mutations as background samples. Furthermore, since the ALJE team had already collected a considerable number of "non-laboratory" mutations, the authors excluded these mutations from their predictions to ensure the "novelty" of the final results. A "novel mutation" is defined as one that 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 system.

To evaluate the generalization ability of the model, a "mask-then-recover" strategy was employed: some positive mutations were randomly masked and then observed to see if the model could recover these mutations in the top K predictions.

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.

Model performance was measured using Recall@K, Precision, and ROC-AUC metrics, and was analyzed using the ROI-based validation dataset. The results were averaged from 100 repeated experiments.

2.6 Visualization and statistical analysis

所有可视化分析均在 R (version 4.3.1) 环境中完成，主要使用 ggplot2 (v3.4.4)、ComplexHeatmap (v2.16) 与 UpSetR (v1.4.0)。

热图用于展示突变分布与聚类一致性；

UpSet 图用于展示突变交集结构；

统计检验包括 Fisher 精确检验与 Benjamini–Hochberg FDR 校正。

# Results

****Clustering results are systematically compared all combinations of three distance metrics (Euclidean, Manhattan, and Cosine) and four clustering algorithms (HDBSCAN, k-means, DBSCAN, and GMM). Each combination was first embedded in two dimensions using UMAP with uniform parameters (n\_neighbors=15, min\_dist=0.3, and seed=123). Clustering was then performed on the embedded space (minPts and eps were used for HDBSCAN/DBSCAN, k=4 for k-means, and optimality was automatically chosen for GMM). To mitigate noise, the average silhouette coefficient was calculated only on valid samples (with positive cluster labels and ≥2 clusters) as a quality metric. The top\_k methods were selected from the highest to lowest silhouette coefficients. If restrict\_metric is set, the optimization is performed within the specified metric; otherwise, the optimization is performed across all combinations. Scores and labels for all solutions were exported and archived to ensure reproducibility.****

****For the top-selected methods, we present their UMAP projections and corresponding cluster heatmaps (all mutations and Top-30 versions) to compare the consistency and differences in clustering achieved by different algorithms in the mutation spectrum space.****

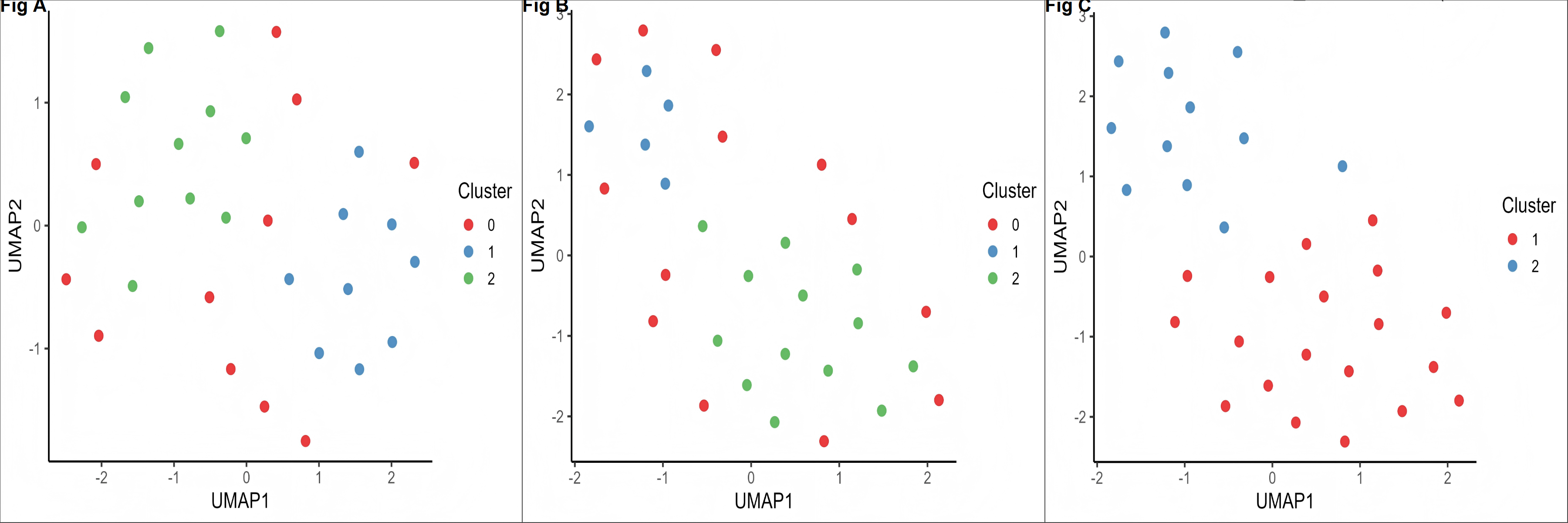
****Based on the parameters of the X\_dense\_midhigh matrix, this study performed multi-algorithm clustering and visualization analysis of mutation distribution patterns across species. To verify the robustness and consistency of the clustering results, we used three top-performing clustering methods: Cosine–HDBSCAN, Euclidean–GMM, and Euclidean–HDBSCAN.****

****The HDBSCAN algorithm automatically identifies clusters and excludes noise points in non-spherical data, while the GMM (Gaussian Mixture Model) captures the underlying continuous distribution of the data through probability density modeling. All three methods take a standardized mutation presence matrix as input. The dimensionality reduction and visualization uses the UMAP (Uniform Manifold Approximation and Projection) algorithm to preserve local topological relationships between samples. The UMAP scatter plot ( ) shows the distribution and cluster boundaries of each species in the two-dimensional latent space using the three methods. Different colors represent different cluster labels, which can intuitively reflect the similarity of mutational profiles between species. For example, Cosine–HDBSCAN effectively groups species with highly similar mutational profiles into the same cluster; Euclidean–GMM shows a relatively regular distribution with clear boundaries; and Euclidean–HDBSCAN demonstrates higher resolution when dealing with intermediate or noisy species. Overall, the cluster structures obtained by the three methods are highly consistent, indicating that mutational patterns are stable and reproducible across different metric spaces.****

****The heatmap ( ) further illustrates the shared mutations among species within each cluster. Rows represent species, columns represent mutation sites, and colors indicate the presence or absence of mutations (1/0). The color bars on the sides correspond to the cluster numbers in the UMAP clustering results. It can be seen that different clusters exhibit distinct complementarity or specificity in mutation distribution. For example, Cluster 0 is primarily concentrated in RRDR core mutations (such as D516, H526, and S531), while Cluster 2 is enriched in marginal or low-frequency mutations (such as Q148R and L533R). These clustering characteristics indicate that the mutation spectrum has certain phylogenetic and evolutionary differentiation characteristics.****

Clustering Framework Overview

Based on the X\_dense\_midhigh matrix, the authors used three top-performing clustering methods: Cosine–HDBSCAN, Euclidean–GMM, and Euclidean–HDBSCAN. All three methods take a normalized mutation presence matrix as input and employ the Unified Mapping (UMAP) algorithm for dimensionality reduction and visualization, preserving local topological relationships between samples.

The UMAP scatter plots show the distribution and cluster boundaries of each species in the two-dimensional latent space using the three methods, with different colors representing different cluster labels.  


**Mutation Distribution and Intra-Cluster Characteristics**

The heatmap further illustrates the shared mutations among species within each cluster. Rows represent species, columns represent mutation sites, and colors indicate the presence or absence of mutations.

The color bars on the sides correspond to the cluster numbers in the UMAP clustering results. Different clusters exhibit distinct complementarities or specificities in their mutation distributions.

**Analysis of the best clustering methods reveals that the best clustering method for mid-high-level data is cosine–HDBSCAN, followed by euclidean–GMM.**

**Both algorithms tend to capture density differences and multimodal distributions.**

**This "density difference" can be seen in the changes in the UpSet graph:**

**Top 10 species → one or two density peaks (highly shared mutations);**

**Mid-high → multiple flat peaks (local commonality but overall sparseness).**

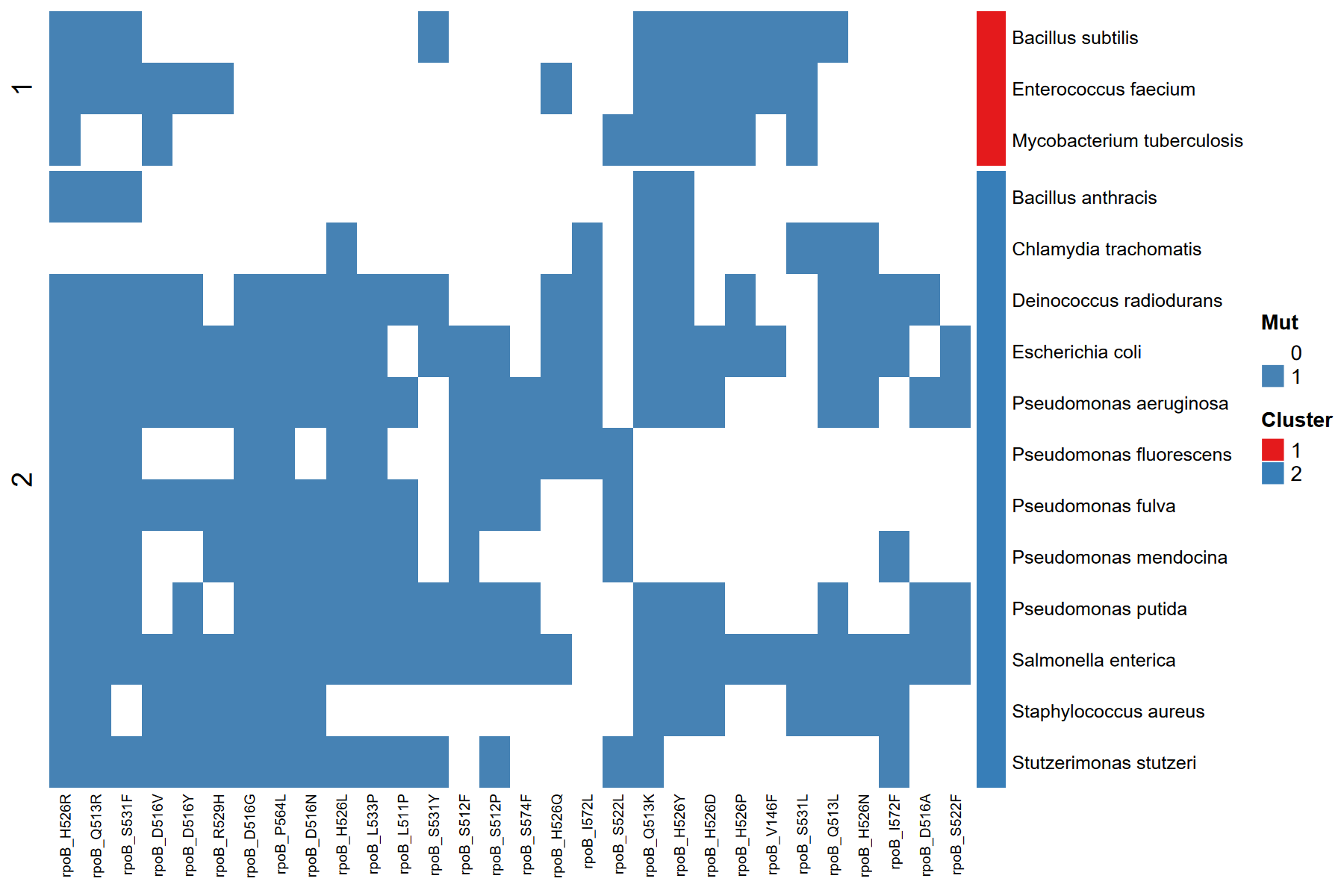
**This suggests that the broadened species dimension of the UpSet graph already reflects the tendency of your cluster structure to become sparser at the mutation level.**

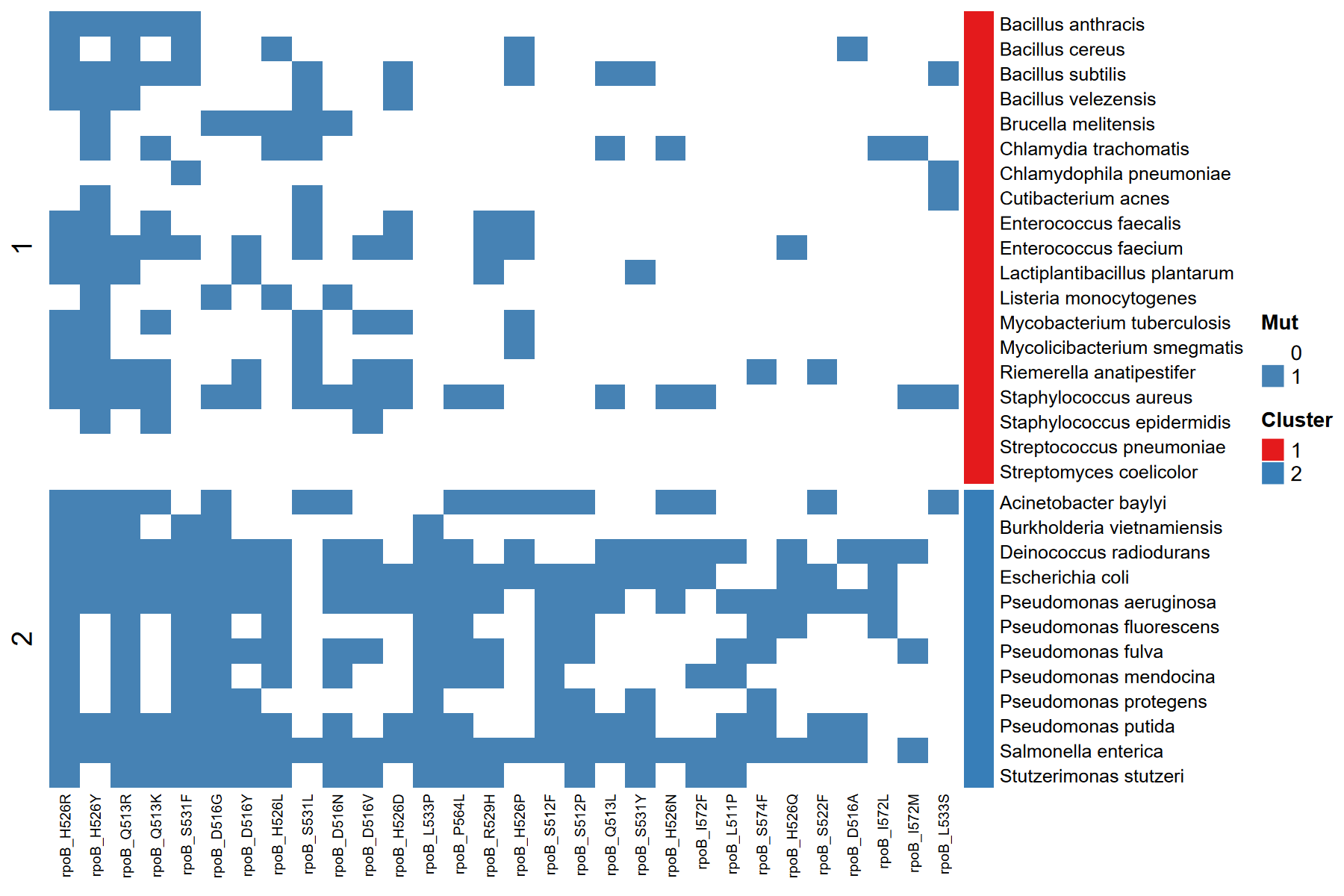
********

****Comparison of Clustering Patterns at Different Confounder Levels****

**To ensure that clustering results are not significantly biased towards highly studied species versus less studied ones, the authors specifically compared clustering results for data with high confounder and mid-high confounder levels.**

**Despite varying noise levels, the clustering structures obtained by the three methods are generally consistent, demonstrating that mutation patterns are robust and reproducible across different metric spaces.**

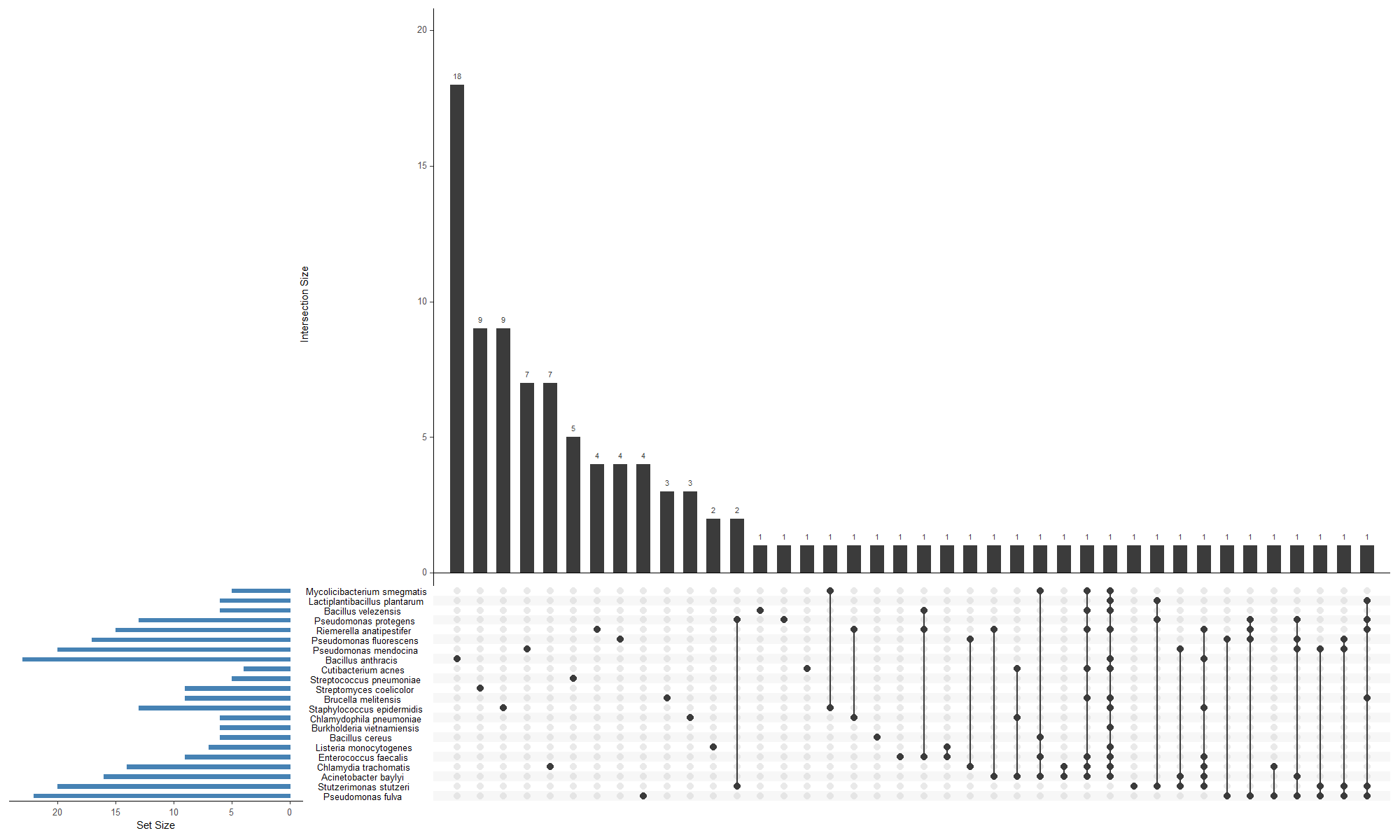
****

****

****Mutation Intersection Relationships (UpSet Analysis)****

**UpSet plots demonstrate the shared mutation relationships between species. Species are set elements, and the bar graphs represent the size of the intersections.**

**In the mid-high dataset, some bacterial genera (e.g., Pseudomonas and Bacillus) share multiple high-frequency rpoB mutation sites, suggesting convergence in resistance mechanisms across lineages.**

********

The blue horizontal bar on the left (Set Size) represents the total number of mutations in each species (set), that is, the number of mutations possessed by that species.

The longer the blue bar, the greater the total number of mutations in that species.

The dot matrix at the bottom (Sets) represents the intersection of different species combinations.

Each vertical column corresponds to a species combination; a black dot indicates that the species is included in the set, and a gray dot indicates that it is not.

A line connecting multiple black dots in a column indicates that those species share the same set of mutations.

The black bar above (Intersection Size) represents the number of mutations shared by that species combination.

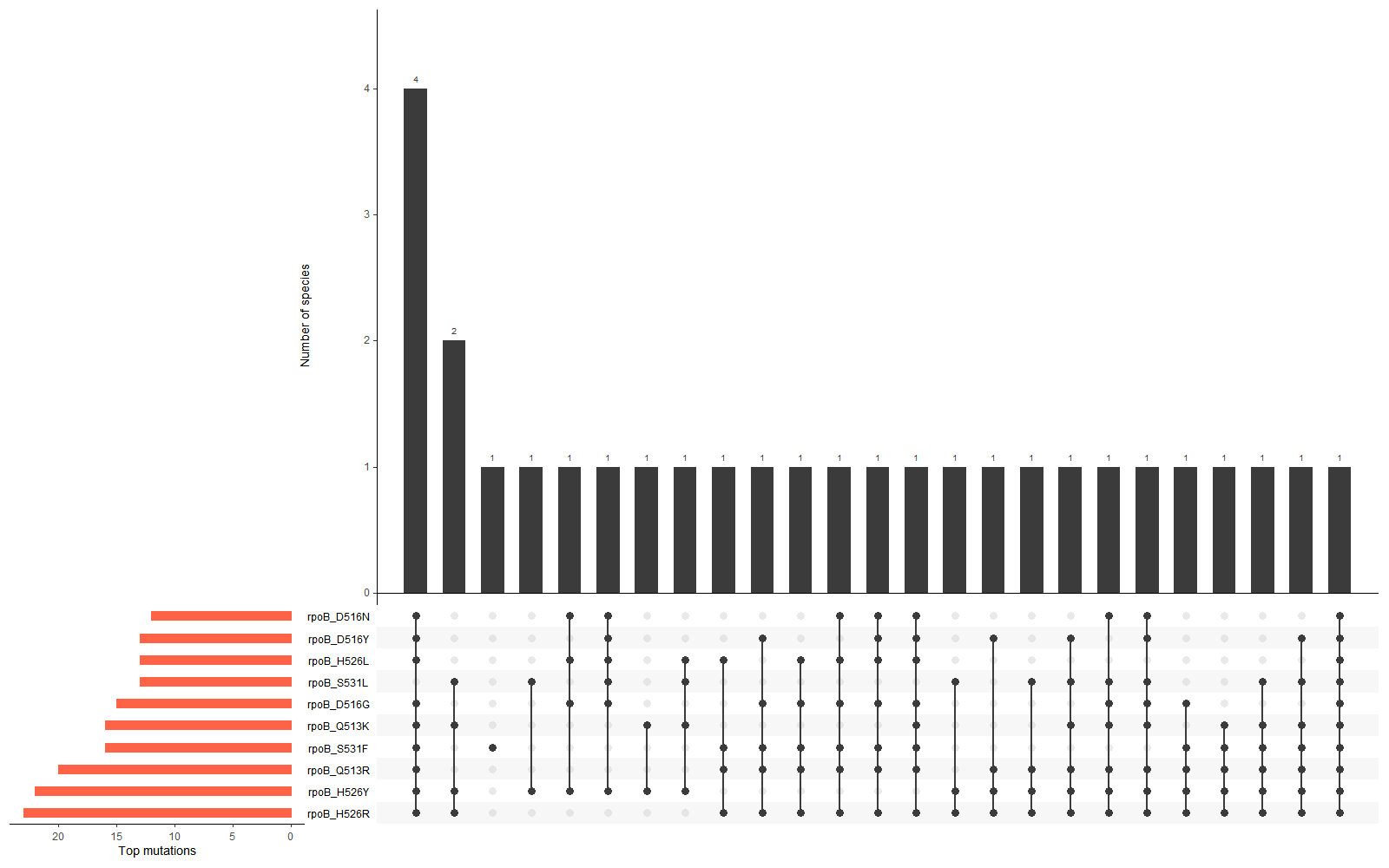
The taller the bar, the more mutations are shared among the species in that group.

Some mutations are shared by a very large number of species (for example, 20 species share certain mutations).

**Only a few mutations are widely distributed across multiple species, while most mutations occur in only a few species.**

The distribution of mutations across species is significantly uneven.

Certain core mutation combinations are common across multiple species, potentially reflecting conserved drug resistance mechanisms.

********

**The red horizontal bar on the left (Selected mutation frequency) indicates the number of species in which each mutation occurs. Longer red bars indicate more common mutations.**

**The bottom matrix (Sets): Each column represents a mutation combination (i.e., which mutations co-occur in the same species).**

**A black dot indicates that the mutation is included in the set, and a line connecting the dots indicates that these mutations co-occur.**

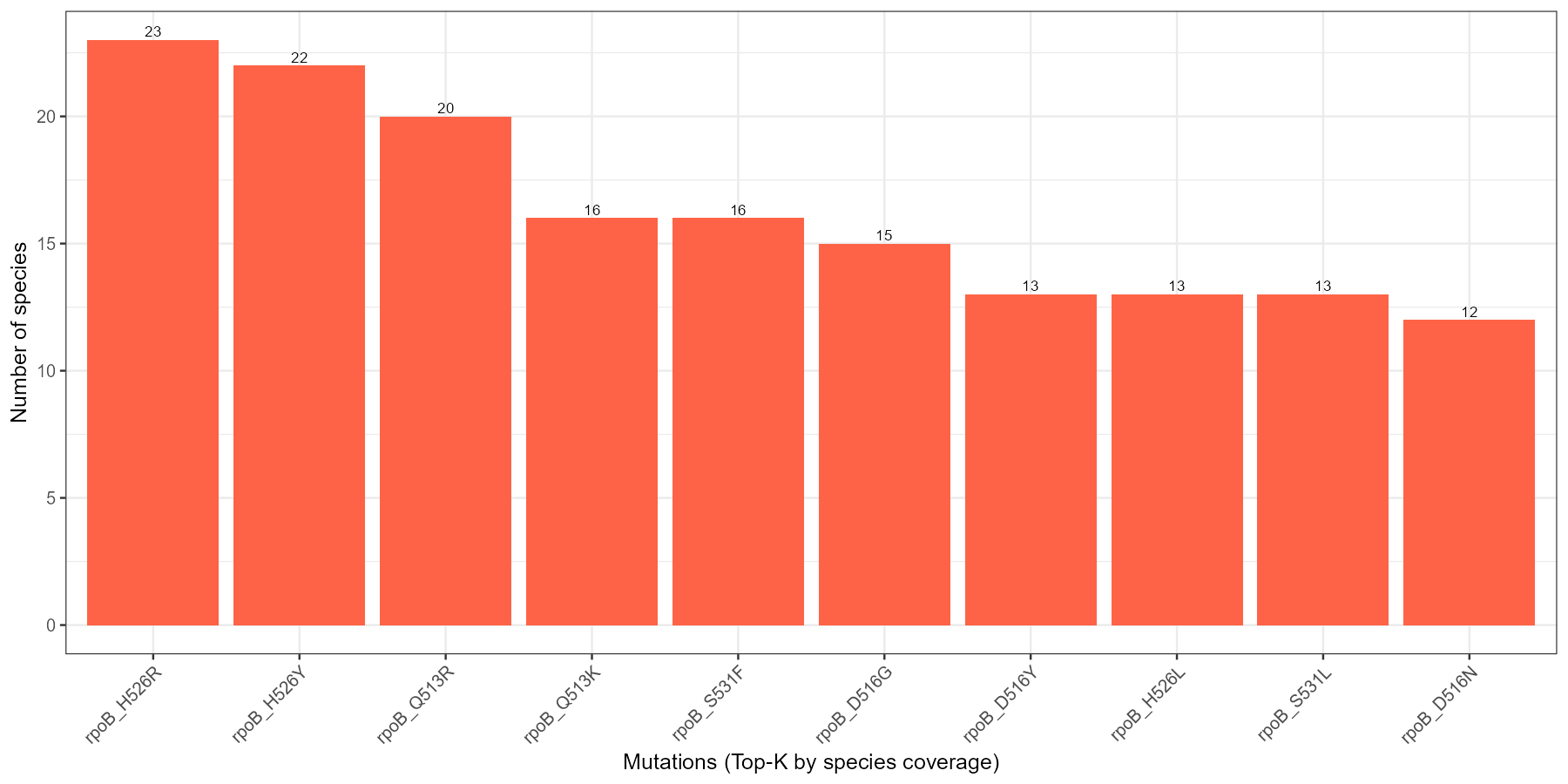
**The upper bars (Number of species): Show how many species share that set of mutations.**

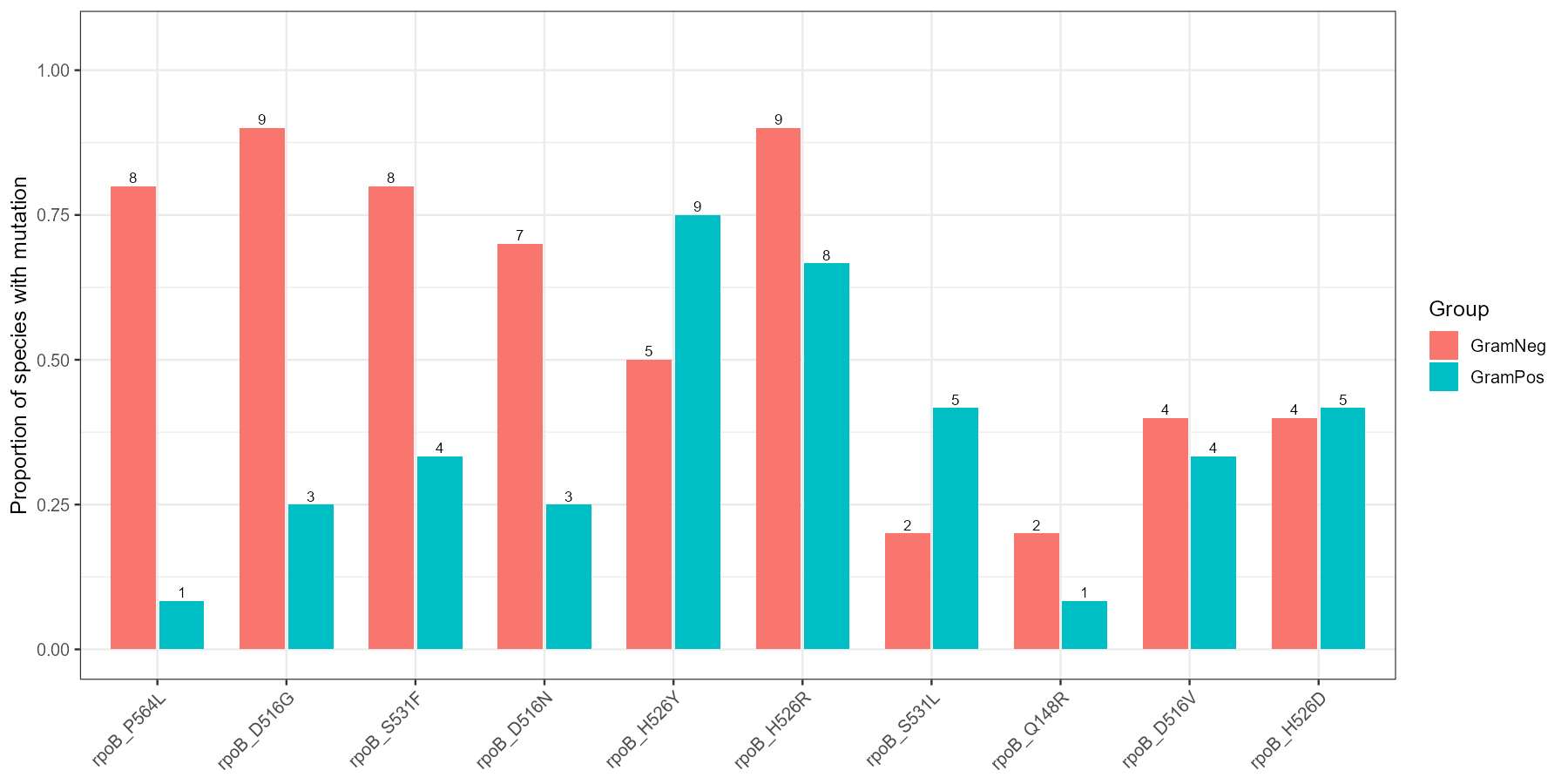
**The taller the bar, the higher the bar on the far left, is 4, indicating that four species share the mutation combination shown in that column (typically including the classic combination of S531F and H526Y).**

**The remaining bars are mostly 1 or 2 in height, indicating that most mutation combinations co-occur in only a very small number of species.**

**rpoB\_S531F and rpoB\_H526Y are the most frequently occurring mutations, each appearing in over 20 species. They are typical rifampicin resistance loci.**

****Core sites of resistance mutations are highly conserved, but multi-site co-occurrence patterns are species-specificGram-positive and -negative differentiation analysis.****

****

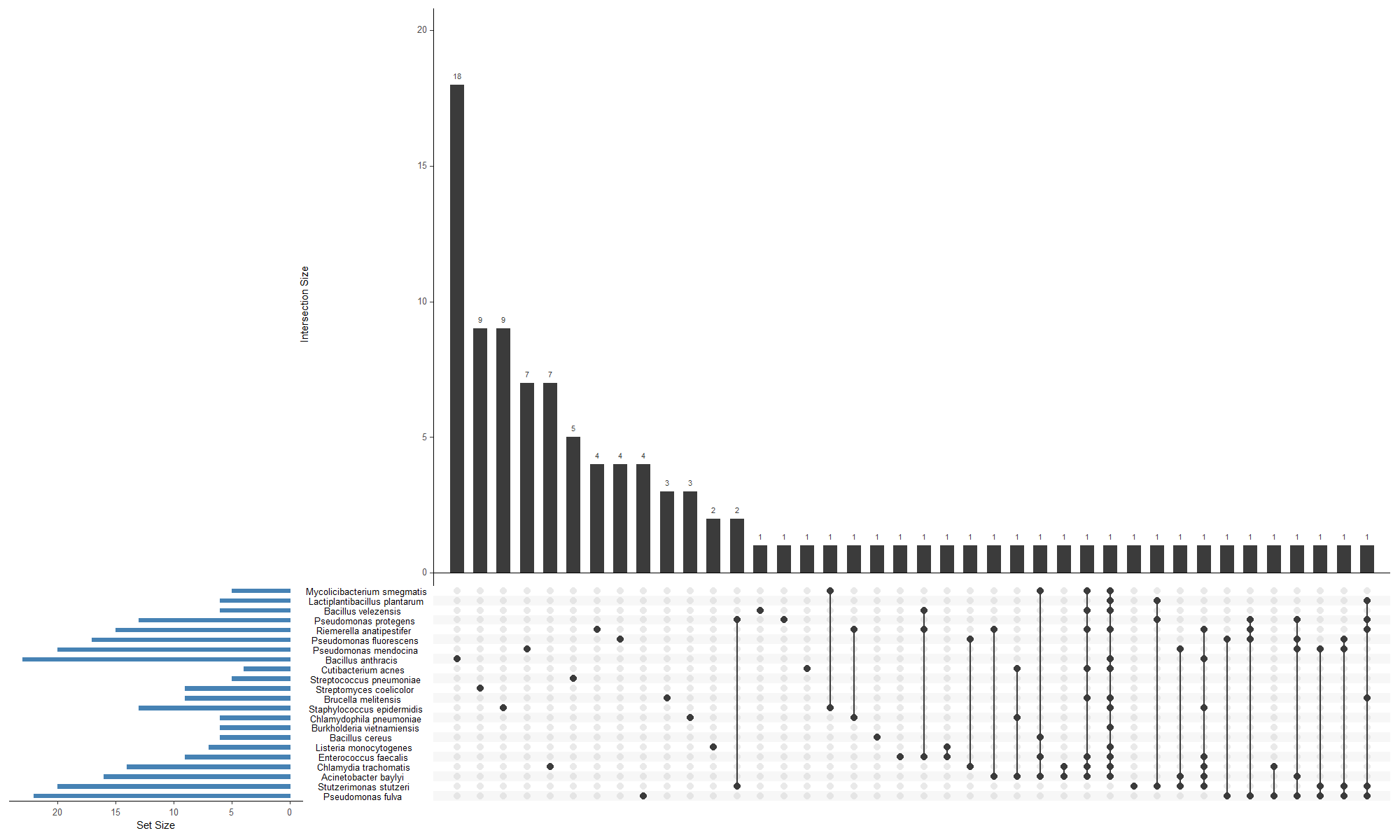
****

**Leveraging the highly recognized Gram-positive/-negative differentiation method, the authors further compared mutation patterns in the mid-high data by Gram staining group (Gram+/Gram–). They found that the two bacterial groups differed in their preferences at typical rpoB loci.**

Most classical RRDR mutations (e.g., P564L, D516G, S531F) were enriched in Gram-negative species, whereas a few peripheral variants (S531L, Q148R) showed relative enrichment in Gram-positive taxa.

This suggests that Gram-negative species rely predominantly on canonical RRDR substitutions conferring strong resistance, whereas Gram-positive taxa accumulate peripheral or compensatory variants that may fine-tune rifampicin susceptibility with reduced fitness costs.

**This grouping trend is consistent with the phylogenetic clustering reported by Bolourchi et al. (2025), indicating that mutational spectrum structure exhibits reproducible evolutionary clustering across species.**



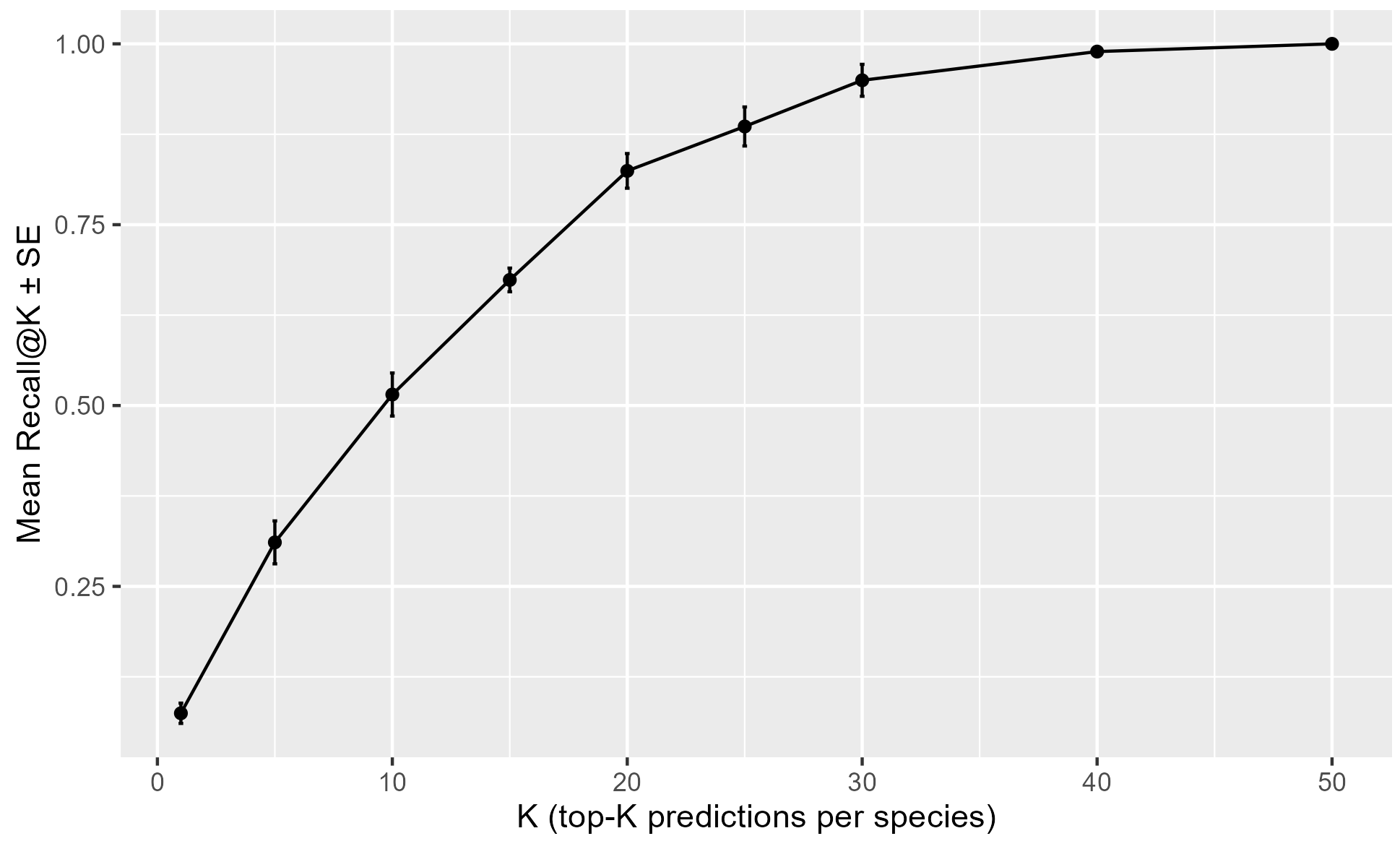
4. Supervised Learning Results

4.1 Overview of PU-learning framework

- Briefly explain model, data, and Mask–then–Recover design.

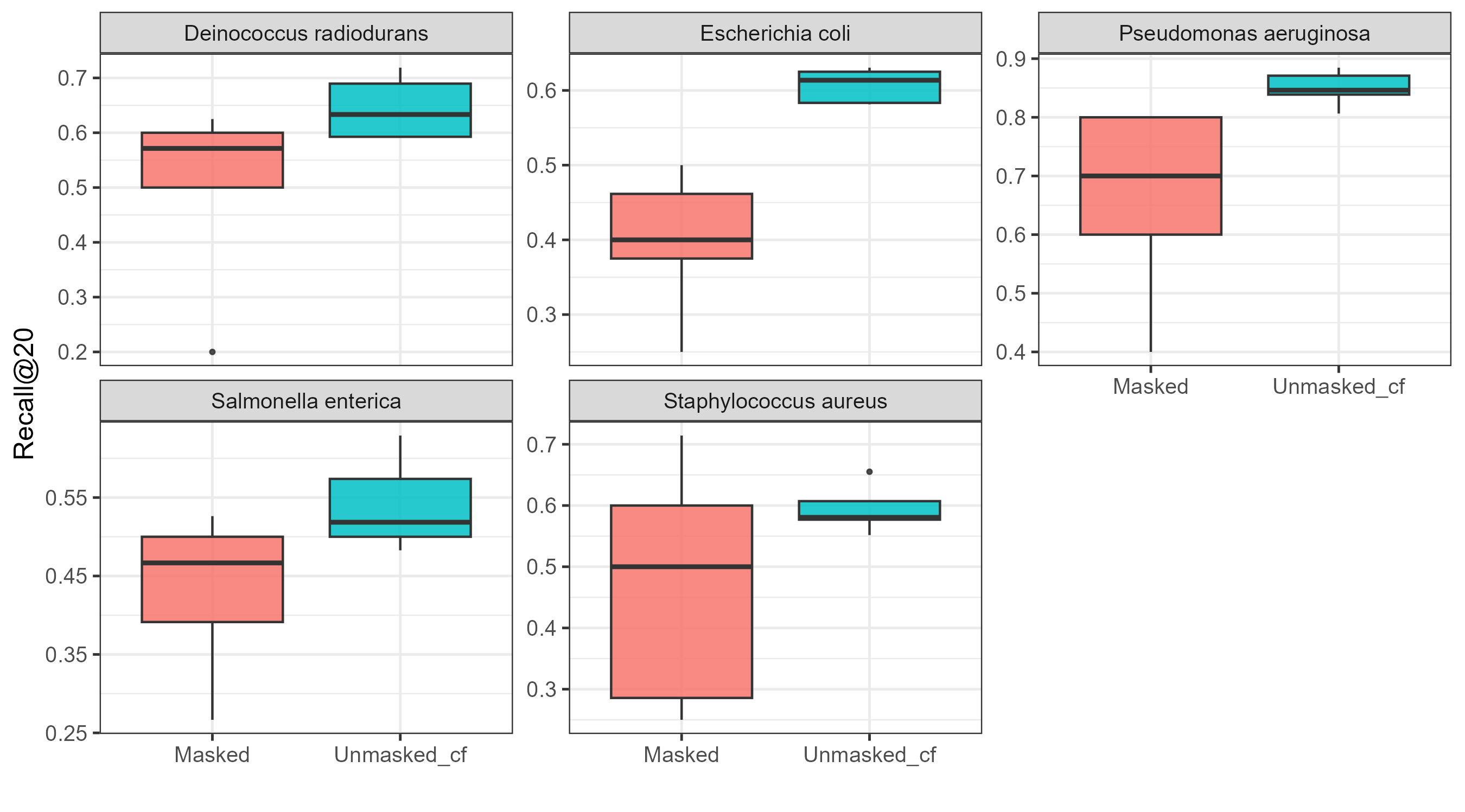
4.2 Model performance and recovery ability

- Recall@K curves (Fig. X)



|  |  |  |  |
| --- | --- | --- | --- |
| Species | Masked | Unmasked\_cf | Delta |
| Brucella abortus | 0.16666666666666666 | 1 | 0.8333333333333334 |
| Listeria monocytogenes | 0.1 | 0.6733333333333333 | 0.5733333333333334 |
| Brucella melitensis | 0.48 | 0.935 | 0.45500000000000007 |
| Helicobacter pylori ATCC | 0.43333333333333335 | 0.8742063492063492 | 0.44087301587301586 |
| Neisseria meningitidis | 0.5 | 0.9 | 0.4 |

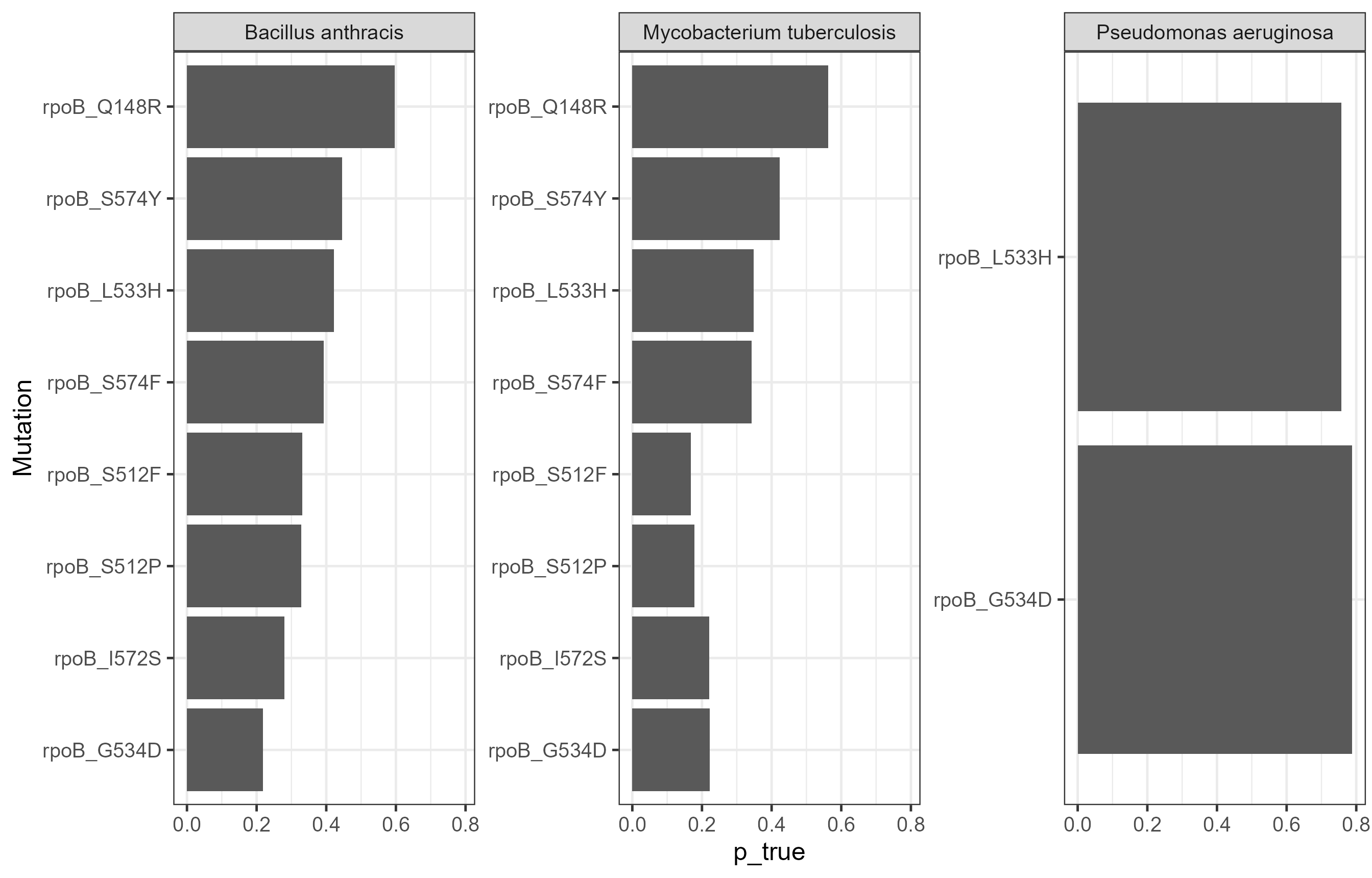
- Interpretation: model generalizes well to resistance-associated mutations.



4.3 Candidate prediction and novelty analysis

- Probability distribution and novelty filtering (Fig. Y)

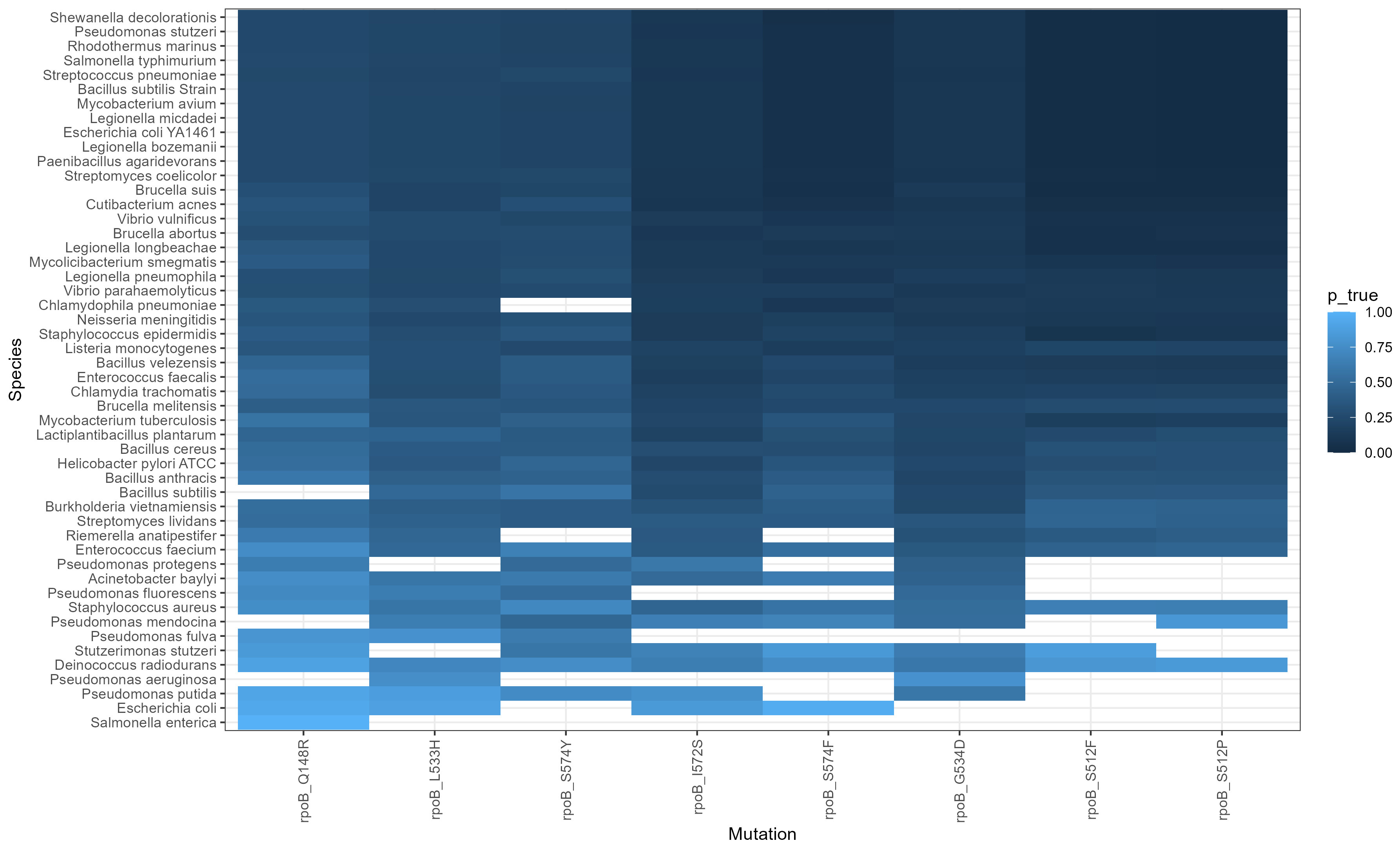
- Highlight representative novel sites (e.g., L533R, Q148R).



|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Species | Novel\_Count | Mean\_Prob | Candidate1\_Mutation | Candidate2\_Mutation | Candidate3\_Mutation |
| Deinococcus radiodurans | 8 | 0.748578812656373 | rpoB\_Q148R | rpoB\_S512P | rpoB\_S512F |
| Staphylococcus aureus | 8 | 0.6132635026502922 | rpoB\_Q148R | rpoB\_S574Y | rpoB\_S512F |
| Enterococcus faecium | 8 | 0.5098281033263364 | rpoB\_Q148R | rpoB\_S574Y | rpoB\_S574F |
| Streptomyces lividans | 8 | 0.4192693982589603 | rpoB\_Q148R | rpoB\_S512F | rpoB\_L533H |
| Burkholderia vietnamiensis | 8 | 0.3989763259749059 | rpoB\_Q148R | rpoB\_S512F | rpoB\_S512P |

4.4 Species-level mutation prediction pattern

- Heatmap showing predicted mutation profiles per species (Fig. Z)



4.5 Biological interpretation

- Summarize Gram+/– divergence and functional implications.

- Connect unsupervised and supervised findings.

# Discussion

**Can ASreview be optimized?**

Currently, ASreview does not offer batch automatic labeling, relying entirely on users to manually assign labels (relevant/irrelevant). Its AI algorithm is only used to rank recommended articles. Users must rely on ASreview's recall regression curve for the current project to determine whether to stop reading subsequent articles deemed "irrelevant." The author believes this process is still cumbersome and should be combined with a language understanding model to automatically perform all subsequent screening after a small amount of manual labeling, or to leave only a few "suspicious" articles for users to judge.

**Why are RIF mutation sites highly disproportionate in some species compared to others?**

Vibrio parahaemolyticus, V. vulnificus, Streptomyces lividans, Brucella suis & Brucella melitensis

In the PU-learning and mask and recall processes, how does masking affect the final prediction model?

**Is random masking the optimal approach?**

Can ASreview's recommendation prioritization or other language vocabulary analysis models be combined to perform high-value, targeted masking?

**Are the final predictions too credible?**

# 5. Conclusions

Summarize the most significant findings and their implications. State whether the research aims were achieved and highlight potential applications or follow-up studies.

# 6. References

All cited literature should appear here in a consistent referencing style (APA, Harvard, or a scientific journal format). Example:  
Foster, L., Mouse, M., & Christ, J. (1972). The effect of hypoxia on free divers. J. Irrep. Res., 23, 490–512.

Federal Funding for the Study of Antimicrobial Resistance in Nosocomial Pathogens: No ESKAPE — Louis B. Rice，The Journal of Infectious Diseases, 2008, 197(8): 1079–1081

WHO. Guidelines for the diagnosis, treatment and prevention of leprosy (2018).

Sivakumaran P, Barros Bd, Antonio Dias VL, Lockwood DN, Walker SL. “A retrospective cohort study of monthly rifampicin, ofloxacin and minocycline in the management of leprosy…” PLoS Neglected Tropical Diseases (2024).

Pupaibool J. “The Role of Rifampin in Prosthetic Joint Infections: Efficacy, Challenges, and Clinical Evidence.” Antibiotics 13(12):1223 (2024).

Karlsen Ø et al. “Rifampin combination therapy in staphylococcal prosthetic joint infections: a randomized controlled trial.” J Orthopaedic Surgery and Research 15:365 (2020).

Zhao L-L, Wan K-L. rpoB mutations and effects on rifampin resistance. Infect Drug Resist. 2020;13:2599-2610. DOI: 10.2147/IDR.S283855.

Das A, et al. The Structural Basis of Mycobacterium tuberculosis RpoB Drug-Resistant Clinical Mutations on Rifampicin Drug Binding. Molecules. 2020;27(3):885. DOI: 10.3390/molecules27030885.

Helmann JD, et al. Mutations in rpoB That Confer Rifampicin Resistance Can Alter Levels of Peptidoglycan Precursors and Affect β-Lactam Susceptibility. mBio. 2023;14(1):e03168-22. DOI: 10.1128/mbio.03168-22.

“Consensus numbering system for the rifampicin resistance-associated rpoB gene mutations in pathogenic mycobacteria.” Clin Microbiol Infect. 2016;22(11):981-986. DOI: (article)

# 7. Figures and Tables

All figures and tables should be labeled and inserted close to where they are first mentioned in the text. Each figure should include a descriptive caption below, while tables should have a title above.

# 8. Appendices

Include any supplementary materials, extended data tables, code snippets, or detailed methodology not essential to the main text.