引言（Introduction）——抗药性、利福平机制与机器学习研究综述

一、抗菌素耐药性（AMR）的全球背景与挑战

自21世纪以来，抗菌素耐药性（Antimicrobial Resistance, AMR）已成为全球公共卫生面临的最严峻挑战之一。世界卫生组织（WHO）估计，到2050年，每年将有超过1000万人死于AMR相关感染。抗药性不仅导致感染治疗难度显著增加，还带来了更高的医疗成本与住院时间，对低收入国家的公共卫生系统造成沉重负担。在众多耐药病原体中，耐药结核病（Drug-resistant Tuberculosis, DR-TB）尤为令人担忧。2022年，全球约有45万人感染利福平耐药结核病（Rifampicin-resistant Tuberculosis, RR-TB），其中大多数同时对异烟肼耐药，构成多重耐药结核病（Multidrug-resistant Tuberculosis, MDR-TB）。

抗药性的形成与传播跨越“人类–动物–环境”三重界面（One Health）。临床滥用抗生素、畜牧与水产中的药物残留，以及废水中耐药基因的扩散形成相互作用的耐药网络，需要卫生、农业与生态部门的协作应对。

二、典型耐药病原体及耐药机制的演化原理

Louis B. Rice于2008年提出“ESKAPE病原体”这一概念，用于指代六类在医院获得性感染中最具代表性的耐药菌群：Enterococcus faecium、Staphylococcus aureus、Klebsiella pneumoniae、Acinetobacter baumannii、Pseudomonas aeruginosa与Enterobacter spp。这些病原体共同特征是可通过多种机制“逃避（escape）”抗菌药物的杀伤。

细菌耐药性的演化通常由以下机制驱动：

1. 外排泵增强：通过激活多药外排系统减少胞内药物浓度；
2. 酶介导失活：例如β-内酰胺酶水解抗生素结构；
3. 水平基因转移（HGT）：通过质粒、噬菌体和转座子传播耐药基因；
4. 生物膜形成：提供化学与免疫保护屏障。这些机制在持续的药物选择压力下可并行或叠加出现，推动多重耐药性（MDR）的进化。

；

1. 靶点突变：药物结合位点的氨基酸替换改变其亲和力，如rpoB突变导致RIF结合障碍

在不同抗菌药物类别中，作用于核心代谢或转录机制的药物尤为关键。  
例如，β-内酰胺类通过抑制细胞壁合成产生杀菌效应，氨基糖苷类干扰蛋白翻译，  
而利福平（Rifampicin, RIF）则直接结合细菌RNA聚合酶的β亚基（rpoB），  
阻断RNA链延伸，从而抑制转录与生长。  
由于这一靶点在原核生物中高度保守，rpoB突变的耐药机制不仅在结核分枝杆菌中普遍存在，也出现在多种临床与环境细菌中。因此，RIF耐药体系被认为是研究“靶点突变型耐药机制”的理想模型。这也反映了反映了在相似选择压力下的功能收敛（functional convergence）。

三、利福平（Rifampicin）及其药理作用机制

利福平（Rifampicin, RIF）是最重要的一线抗结核药物之一，属于利福霉素类抗生素。其药理作用机制是通过结合细菌RNA聚合酶的β亚基（由rpoB基因编码），阻断RNA链延伸，进而抑制转录过程。

RIF具有优异的脂溶性，能穿透富含分枝菌酸的细胞壁，对静止状态的细菌亦具活性。然而，RIF单药耐药常被视为多重耐药结核（MDR-TB）的早期信号，因为其耐药突变往往与其他药物的抗性共同出现。MDR-TB指同时对异烟肼（INH）与利福平（RIF）耐药的结核分枝杆菌株，是当前全球结核防控中的最严重挑战之一。

由于这两种药物分别作用于细胞壁合成与RNA转录，是短程联合化疗的核心药物，

MDR-TB的出现往往意味着治疗方案需更换为成本更高、副作用更大的二线药物。

若菌株进一步对氟喹诺酮类及注射用二线药物耐药，则被定义为广泛耐药结核（XDR-TB），

这类病例的治愈率通常不足60%（WHO, Global TB Report 2023）。”

四、RIF耐药机制与rpoB突变的跨物种演化特征

RIF通过与RNA聚合酶β亚基的“抗药性决定区”（Rifampicin Resistance-Determining Region, RRDR）结合发挥作用，Rifampicin resistance–determining region (RRDR), spanning codons **426–452 of rpoB** (equivalent to residues **507–533** in E. coli numbering。该区域在结核分枝杆菌中对应RpoB蛋白的426–452号氨基酸残基。典型突变如S450L、H445Y与D435V等，其中S450L显著降低RIF结合力，H445Y显著降低其结合亲和力，而D435V会干扰药物进入其通道从而导致耐药性。约90–95%的RIF耐药菌株的突变位于RRDR内，因此该区域被视为RIF耐药性的分子“指纹”。

在其他细菌中（如E. coli、Pseudomonas aeruginosa、Staphylococcus aureus），也观察到功能相似的rpoB突变，提示RIF耐药的演化具有“趋同进化（convergent evolution）”特征。对这些跨物种突变模式进行系统分析，有助于揭示抗药性形成的普适规律。

五、机器学习在抗药性研究中的崛起与代表性应用

传统抗药性研究依赖表型药敏试验或有限基因检测，难以捕捉多基因、多通路的复杂耐药机制。随着大规模基因组测序与多组学数据的积累，机器学习（Machine Learning, ML）成为刻画高维基因型–表型关系的关键工具。

多项研究已验证其有效性。例如：

Yang等人（2024, PLOS Computational Biology）利用XGBoost回归，在超过10,000株结核分枝杆菌的基因组上预测药物MIC值；

Walker等人（2022, Nature Communications）采用随机森林预测13种抗结核药物的耐药性并分析突变特征；

Davis等人（2019, Nature Communications）使用梯度提升模型实现E. coli与Klebsiella间的耐药泛化；

Arango-Argoty等人（2018, Nucleic Acids Research）提出DeepARG模型识别宏基因组中的潜在耐药基因；

Moradigaravand等人（2018, mBio）通过Lasso逻辑回归预测金黄色葡萄球菌的β-内酰胺与MRSA耐药性。

这些研究证明，ML模型不仅能高效预测已知耐药性，还能识别潜在的新型突变特征。然而，现有模型多局限于单一物种或固定药物类别，缺乏跨物种的系统预测框架。

六、本研究的创新点与目标

本研究旨在构建一个跨物种的RIF抗药突变预测框架，整合多层次机器学习方法：

（1）使用Sentence-BERT与SciBERT模型，对文献数据进行自动语义筛选与突变证据提取；

（2）采用无监督聚类（UMAP + HDBSCAN/GMM）探索RIF耐药突变在不同菌种间的分布模式；

（3）通过半监督正负不平衡学习（PU Learning）结合随机森林（Random Forest），预测潜在RIF耐药突变。

在实际的生物信息与医学研究中，研究者往往只能获取部分已确认的“正例”，而缺乏系统性的“负例”信息。

传统的监督学习依赖明确的正负样本划分，因此难以直接适用于此类问题。

正例–未标注学习（Positive–Unlabeled Learning, PU Learning）提供了一种自然的解决方案，

它通过同时利用“已知正例”和“大量未标注样本”进行建模，估计其中潜在的负例或未知正例分布，

从而在缺乏负样本标注的条件下实现可靠的分类或排序。

PU Learning 最初广泛应用于文本检索与网页分类任务（Elkan & Noto, 2008），

在这些任务中，系统只记录“被点击的网页”作为正例，而未点击的网页并不等于不相关。

随后，该框架被拓展至推荐系统的隐式反馈学习（Hu et al., 2008），

仅凭“用户看过或购买过的物品”建立偏正例学习模型。

在生态学的物种分布建模中（Phillips et al., 2006），

研究者也仅掌握“物种被观测到的存在点”，并使用类似PU逻辑的MaxEnt方法进行预测。

近年来，PU Learning 在生物信息学与医学领域的应用尤为广泛，

包括疾病基因优先级排序（Yang et al., 2020）、基因功能预测、蛋白相互作用发现等任务，

其核心思想是：已知致病基因为正例，其余未标注基因不等同于无关或阴性样本。

在本研究中，这一思路被自然地引入至抗药性突变预测任务中。

已报道的耐药突变可视为“正例”，而其余未研究突变则属于“未标注样本”。

由于真实的“非耐药突变”信息极度稀缺，

采用PU Learning结合随机森林（Random Forest）可在不依赖人工负例的情况下，

评估各突变产生抗药性的可能性并进行候选优先级排序。

这一方法为探索未验证突变的潜在功能效应提供了一种可推广的新思路。

该研究的创新点在于将自然语言处理与基因组聚类、机器学习预测相结合，实现从文献知识到基因变异再到新突变预测的多层信息融合。研究结果有望揭示RIF耐药突变的跨物种演化规律，并为新型抗药性监测、耐药诊断及药物设计提供数据支撑。

参考文献（部分）

Yang J. et al. (2024). Quantitative drug susceptibility testing for Mycobacterium tuberculosis using unassembled sequencing data and machine learning. PLOS Computational Biology, 20(6), e1012260.

Walker T. M. et al. (2022). Prediction of Mycobacterium tuberculosis antibiotic resistance from whole genome sequencing: an interpretable machine learning approach. Nature Communications, 13, 3078.

Davis J. J. et al. (2019). Antimicrobial resistance prediction in PATRIC and RAST using curated genome annotation and machine learning. Nature Communications, 10, 5182.

Arango-Argoty G. A. et al. (2018). DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. Nucleic Acids Research, 46(9), 5549–5560.

Moradigaravand D. et al. (2018). Prediction of antibiotic resistance in Escherichia coli from large-scale pan-genome data using machine learning. mBio, 9(3), e02118-17.

Elkan, C., & Noto, K. (2008). Learning classifiers from only positive and unlabeled data. SIGKDD.

Hu, Y., Koren, Y., & Volinsky, C. (2008). Collaborative filtering for implicit feedback datasets. ICDM.

Phillips, S. J., Anderson, R. P., & Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. Ecological Modelling.

Yang, P. et al. (2020). Positive-unlabeled learning for disease gene identification. Bioinformatics.