# Machine learning and phylogenetic analysis allow for predicting antibiotic resistance in *M. tuberculosis*

**Alper Yurtseven**[1,2], Sofia Buyanova[4], Amay A. Agrawal[1,2], Olga O. Bochkareva[4,5], Olga V. Kalinina[1,2,3]

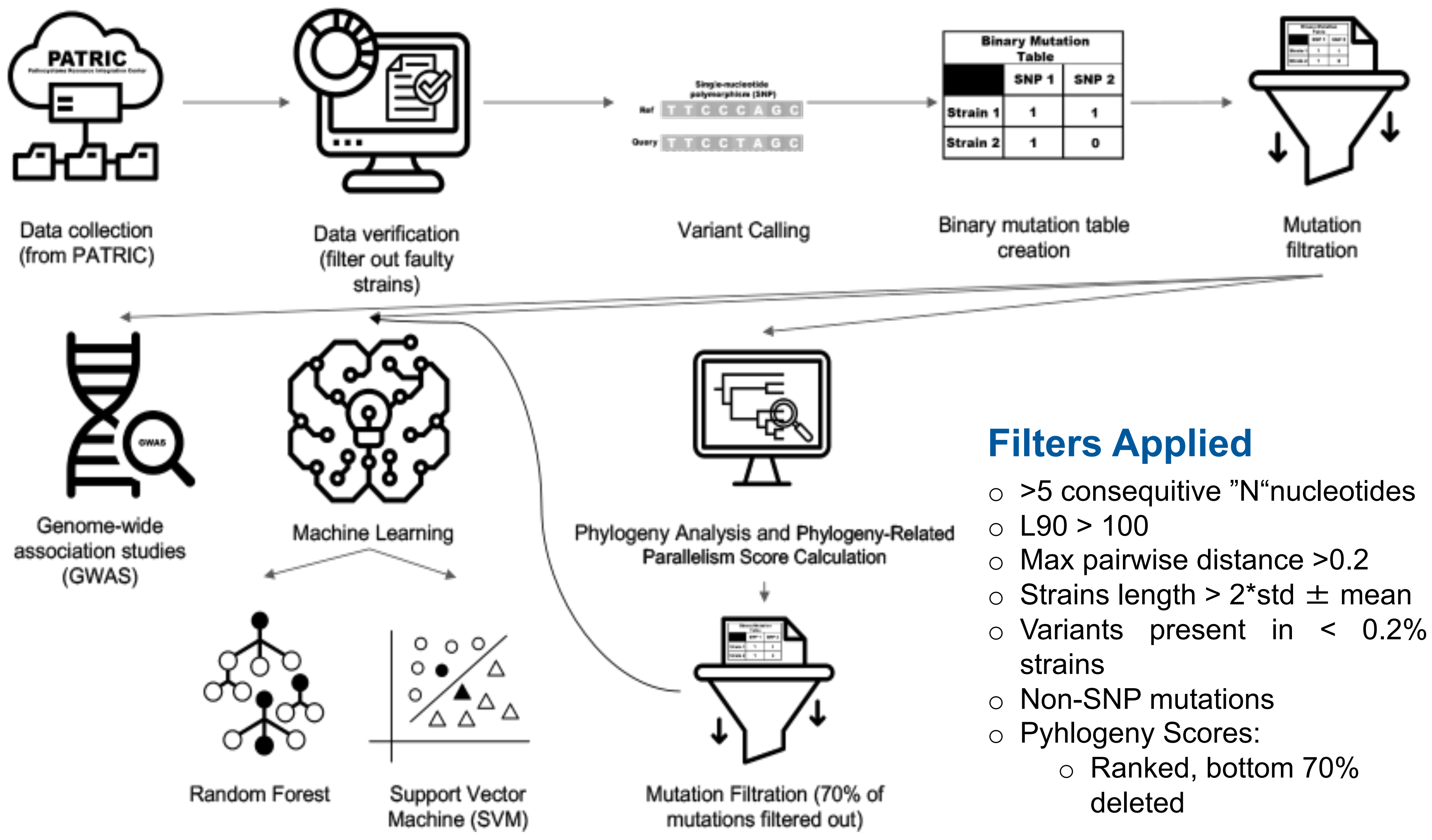
[1] Helmholtz Institute for Pharmaceutical Research Saarland (HIPS), Helmholtz Centre for Infection Research (HZI), Saarbruecken, Germany
[2] Center for Bioinformatics, Saarland University, Saarbruecken, Germany
[3] Faculty of Medicine, Saarland University, Homburg, Germany
[4] Institute of Science and Technology Austria (ISTA), Vienna, Austria
[5] CUBE, CMESS, University of Vienna, Austria

## Introduction

**Antimicrobial resistance (AMR) is a global health concern requiring accurate prediction of bacterial resistance patterns. Machine learning (ML) methods often overlook evolutionary relationships, limiting their detection of resistance-associated features. We propose using PRPS (phylogeny-related parallelism score) with ML, which measures whether a certain feature is correlated with the population structure of a set of samples. When validated on *Mycobacterium tuberculosis* genomes screened against 6 antibiotics (Amikacin, Capreomycin, Ethionamide, Kanamycin, Ofloxacin, and Streptomycin), we re-discovered known mutations and uncovered new candidates. Therefore integrating PRPS with ML enhances AMR analysis, addressing limitations and improving treatment and control strategies.**

| Drug name | Classification by line | Pharmacological group | Number of strains | Number (fraction) of resistant strains | Number of mutations (features) |
|---|---|---|---|---|---|
| **Streptomycin** | First line | Aminoglycosides | 4726 | 1158 (24,5%) | 24425 |
| **Amikacin** | Second line | | 1149 | 208 (18,1%) | 18864 |
| **Capreomycin** | | | 1086 | 205 (18,9%) | 17045 |
| **Kanamycin** | | | 1362 | 297 (21,8%) | 17335 |
| **Ofloxacin** | | Fluoroquinolones | 795 | 307 (38,6%) | 14185 |
| **Ethionamide** | | Nicotinamide derivative | 571 | 210 (36,8%) | 12974 |

**Dataset**


PATRIC
Data collection (from PATRIC)
Data verification (filter out faulty strains)
Single-nucleotide polymorphism (SNP)
Ref T T C C C A G C
Query T T C C T A G C
Variant Calling
Binary Mutation Table
SNP 1
SNP 2
Strain 1
1
1
Strain 2
1
0
Binary mutation table creation
Mutation filtration
Genome-wide association studies (GWAS)
GWAS
Machine Learning
Random Forest
Support Vector Machine (SVM)
Phylogeny Analysis and Phylogeny-Related Parallelism Score Calculation
Mutation Filtration (70% of mutations filtered out)

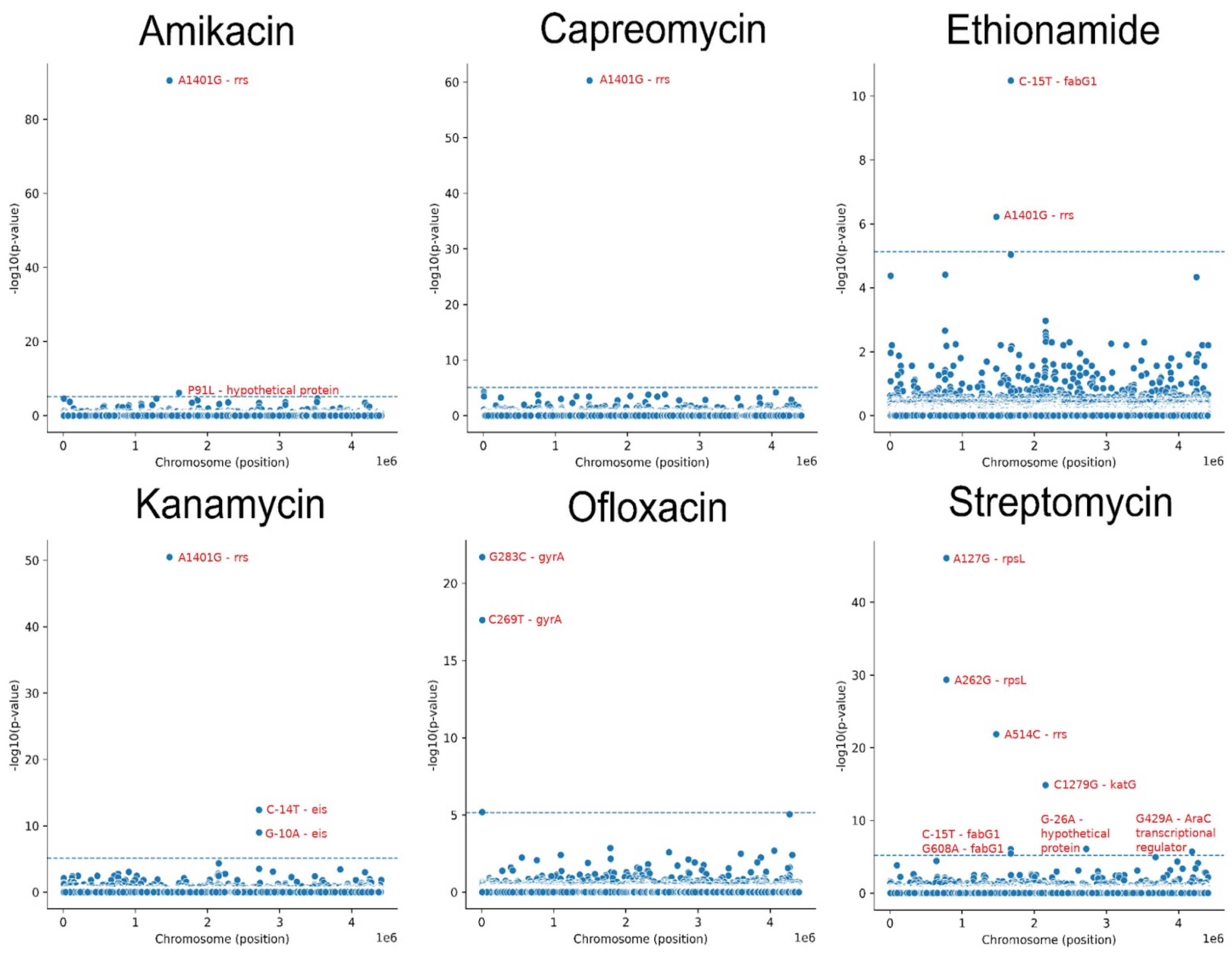
### Filters Applied

- >5 consequitive ”N“nucleotides
- L90 > 100
- Max pairwise distance >0.2
- Strains length > 2*std ± mean
- Variants present in < 0.2% strains
- Non-SNP mutations
- Pyhlogeny Scores:
  - Ranked, bottom 70% deleted

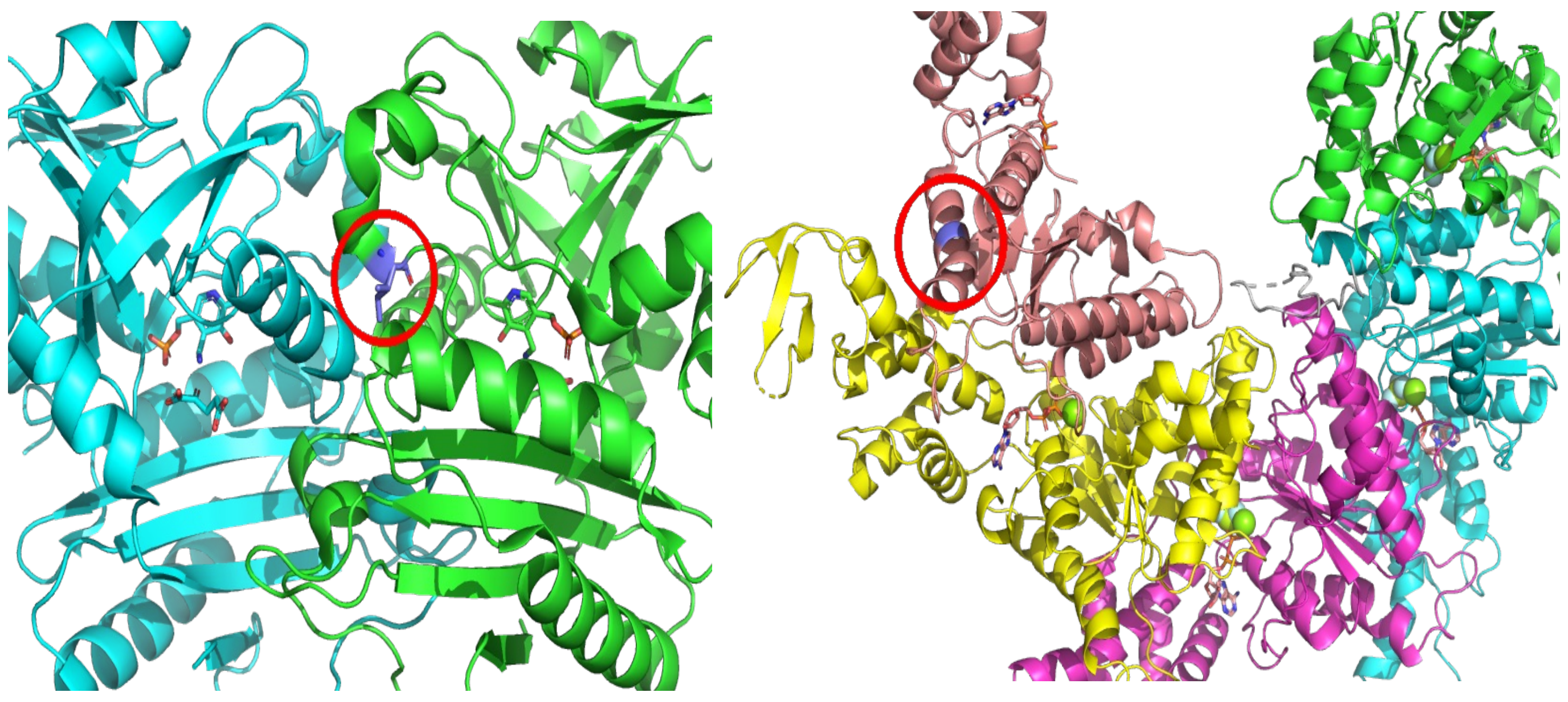
**Pipeline**

## GWAS

To identify genetic variations linked to antibiotic resistance, we utilized Pyseer[5], a leading tool for genome-wide association analysis. The analysis employed a linear mixed model (LMM) with random effects to control for population structure. To address multiple testing, we calculated Bonferroni p-value thresholds for each antibiotic, selecting significant variants associated with resistance phenotype.



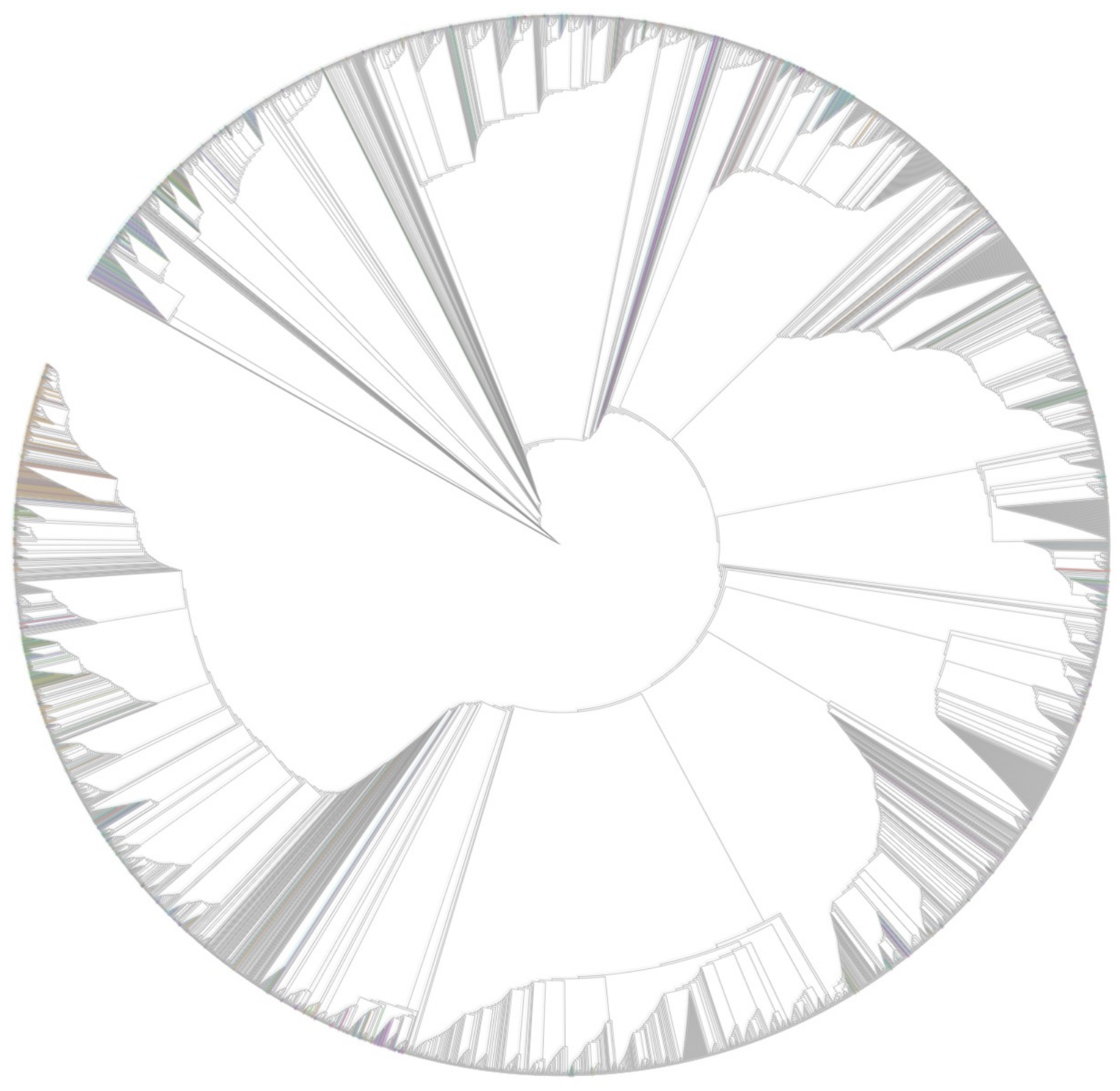
**GWAS Results**



**Left:** Mutation Ile145Met in the probable amino acid aminotransferase PabC
**Right:** Mutation Lys179Gln in the zinc metalloprotease FtsH

## Phylogenetic Analysis

Phylogeny was built using the PanACoTA pipeline[2]. Orthologous groups were formed with an 80% protein identity threshold. A maximum-likelihood phylogenetic tree was constructed using fasttree[3] from a concatenated codon alignment of 161 common genes. The tree, along with resistance profiles, was visualized using the iTOL online tool[4].



## PRPS

To calculate PRPS for each single-nucleotide polymorphism (SNP), we generate a pairwise distance matrix, collapse clades with SNPs, and calculate the logarithm of the sum of pairwise distances between nodes with SNPs. PRPS reflects SNP frequency and mutation distances.

## Machine Learning

| | Support Vector Machine (SVM) | | | Random Forest | | |
|---|---|---|---|---|---|---|
| **MCC** (Matthew's correlation coefficient) | All Features | TOP 30% PRPS | 30% Random | All Features | TOP 30% PRPS | 30% Random |
| **Amikacin** | 0,720 | 0,752 | 0,302 | 0,881 | 0,883 | 0,481 |
| **Capreomycin** | 0,539 | 0,620 | 0,334 | 0,779 | 0,780 | 0,569 |
| **Ethionamide** | 0,325 | 0,370 | 0,269 | 0,550 | 0,605 | 0,488 |
| **Kanamycin** | 0,766 | 0,685 | 0,415 | 0,812 | 0,856 | 0,546 |
| **Ofloxacin** | 0,508 | 0,549 | 0,294 | 0,778 | 0,778 | 0,452 |
| **Streptomycin** | 0,602 | 0,613 | 0,477 | 0,782 | 0,801 | 0,650 |

**Machine Learning Results:** MCC values of ML models (67% Training, 33% Test)

| Variant ID | Gene | Mutation | Uniprot ID | RIN-based simple classification | Observed associated with resistance to | Reported to be associated with resistance to |
|---|---|---|---|---|---|---|
| **Known resistance-associated mutations** | | | | | | |
| **(7570, 'C,T', 'snp')** | GyrA | A90V | P9WG47 | Ligand interaction | Ofloxacin, Ethionamide | Fluoroquinolone |
| **(7362, 'G,C', 'snp')** | GyrA | E21Q | P9WG47 | Protein interaction | Ofloxacin | Fluoroquinolone |
| **(7585, 'G,C', 'snp')** | GyrA | S95T | P9WG47 | DNA interaction | Ofloxacin | Fluoroquinolone |
| **(781687, 'A,G', 'snp')** | RpsL | K43R | P9WH63 | RNA interaction | Streptomycin | Streptomycin |
| **Previously unreported variants** | | | | | | |
| **(906857, 'A,G', 'snp')** | PabC | I145M | Q79FW0 | Protein interaction | Ethionamide, Ofloxacin | - |
| **(4052349, 'T,G', 'snp')** | FtsH | K179Q | P9WQN3 | Protein interaction | Ethionamide | - |
| **Previously unreported variants with no structural templates (AlphaFold models used)** | | | | | | |
| **(4120926, 'A,G', 'snp')** | anion transporter ATPase | N378D | I6Y498 | Surface | Ethionamide | - |
| **(1896581, 'T,C', 'snp')** | membrane protein | M36T | O53918 | Surface | Ethionamide | - |
| **(835611, 'C,T', 'snp')** | hypothetical protein | T153M | I6X9N8 | Surface | Ofloxacin | - |

**Structural Classification Results:** Structural classification of known resistance-associated and novel predictive variants with StructMAn[6]

## Conclusion

ML has demonstrated its value in clinical treatment as a predictive tool for medical staff[1]. We utilized traditional ML to detect the significance of unreported mutations. To enhance our ML approach, we devised a procedure that analyzes phyletic patterns of features. This procedure effectively eliminates numerous false variants and enhances performance metrics. While retaining most known resistance markers, we also discovered several new mutations.

Therefore, we suggest that applying phylogeny analysis before ML model training is a good in-between step to increase model scores and explainability of the models as well as reduce computation time and resources.

## References

(1) Moran et al. (2020) *J. Antimicrobial Chemotherapy*
(2) Perrin, A. et al. (2021) *NAR genomics and bioinformatics*
(3) Price, M. N. et al. (2010) *PLoS ONE*
(4) Letunic, I. et al. (2006) *Bioinformatics*
(5) Lees, John A., et al. (2018) *Bioinformatics*
(6) Gress et al. (2016) *Nucleic Acids Research*