M.S. PROJECT PRESENTATION · DATA SCIENCE & AI

Predicting *E. coli* Drug Resistance

Using Traditional Machine Learning and Deep Learning

Abdoulfatah Mohamed Abdillahi

San Francisco State University · May 2026

COMMITTEE

Dr. Sara El Alaoui · Dr. Akila de Silva · Dr. Pleuni S. Pennings

OUTLINE

How this talk is organized.

01

Background & Motivation

Antimicrobial resistance in E. coli and why pangenome-based approaches matter for AMR prediction.

02

Methods

Pangenome construction from 588 isolates, gene categorization, and dual feature extraction (SNP & GPA matrices).

03

Results

Model comparison across RF, LR, XGB, and TabR; biological validation of gyrA and parC resistance signal.

04

Conclusion & Future Work

Key findings, limitations, and directions for unified SNP+GPA representations and deep learning extensions.

Antibiotic Resistance: A Global Health Crisis

>1M[1]

deaths annually

from antimicrobial resistance

#1

E. coli — most common

Gram-negative pathogen

Days

to get results from

culture-based testing

[1] Murray et al., *The Lancet*, 2022

Whole-genome sequencing combined with machine learning offers a faster and more scalable approach.

Why Ciprofloxacin in *E. coli*?

Clinical Relevance

- E. coli is the most common cause of urinary tract and bloodstream infections.
- Ciprofloxacin is widely used to treat severe E. coli infections.
- However, resistance to ciprofloxacin is increasing globally.

[1] Kot, *Polish J. Microbiology*, 2019

From Biology to Feature Design

Chromosomal Mutations → SNP-based Features

- Mutations in key genes (e.g., *gyrA*, *parC*) drive resistance
- Captured as SNPs in the model.

Horizontal Gene Acquisition → GPA Features

- Bacteria can acquire resistance genes from other organisms
- Represented as gene presence/absence features

→ *This dual mechanism provides a natural framework for comparing SNP-based and GPA feature representations.*

DATA SOURCES

Where Does the Data Come From?

All isolates from human bloodstream infections (bacteraemia) — clinically consistent across all three studies.

200 isolates[1]

Invasive E. coli — England

[1] Moradigaravand et al., *PLOS Comp. Biol.*, 2018

EUCAST breakpoints applied

200 isolates[2]

NORM Surveillance — Norway

[2] Gladstone et al., *The Lancet Microbe*, 2021

EUCAST breakpoints applied

200 isolates[3]

US Military Healthcare Network

[3] Mills et al., *Genome Medicine*, 2022

EUCAST breakpoints applied

600
initial isolates

→

18
intermediate
reclassified → susceptible (EUCAST)

→

−12
excluded
(no genomic data)

→

588
final matched isolates

Two ways to read a genome.

Two complementary genomic signals used for prediction.

SNPs · Single Nucleotide Polymorphisms

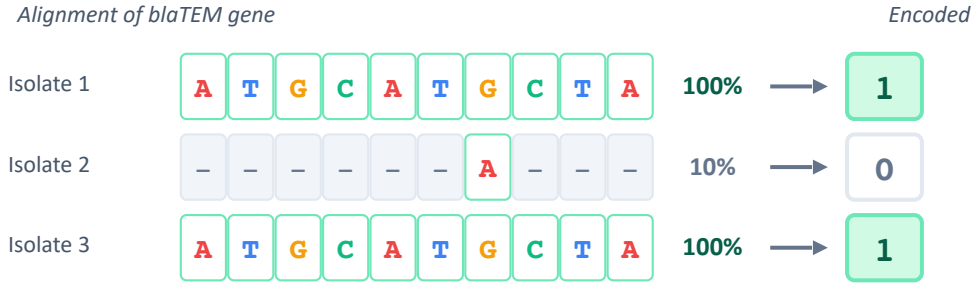
Letter-level differences in DNA at the same position.



- Captures point mutations (e.g., ***gyrA S83L***)
- Same gene across strains — sequence varies
- Output: **SNP matrix** — 600 × 166,415 features

GPA · Gene Presence / Absence

Map gene to isolate → encode by coverage.



Rule: coverage ≥ 50% → **1 (present)** · coverage < 50% → **0 (absent)**

- Captures acquired resistance genes (e.g., ***blaTEM***)
- Whole genes — present in some strains, absent in others
- Output: **GPA matrix** — 600 × 6092 features

SNP *= sequence variation* | ***GPA*** *= gene content variation*

Research Aims & Hypothesis

01

Construct Features

- SNP and gene presence/absence (GPA)

02

Benchmark Models

- 5 ML models + TabR (deep learning)
- Evaluated on both SNP and GPA datasets

03

Evaluate Impact

- Feature representation vs. model choice
- Predictive performance comparison

HYPOTHESIS

SNP features capture mutation-driven resistance (gyrA, parC) and are expected to perform strongly.

GPA features capture horizontally acquired resistance and provide complementary interpretability.

PART 01

Methods

From bacterial isolates to model-ready features

Class Distribution of Final Dataset

Label Processing & Sample Matching

01 Raw labels

282 R · 300 S · 18 I

02 Reclassify

I → S (EUCAST guideline)

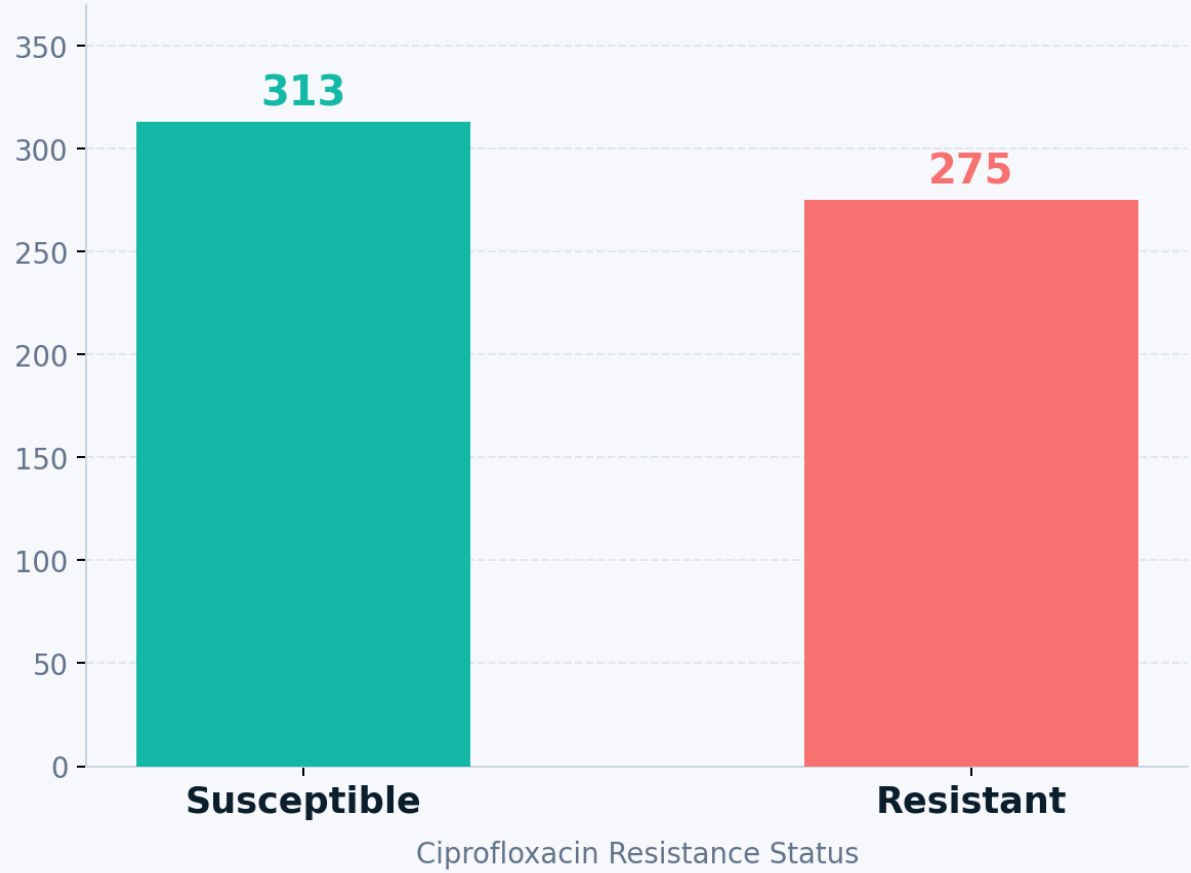
03 Filter

12 removed (no genomic data)

04 Final dataset

588 isolates · 313 S · 275 R

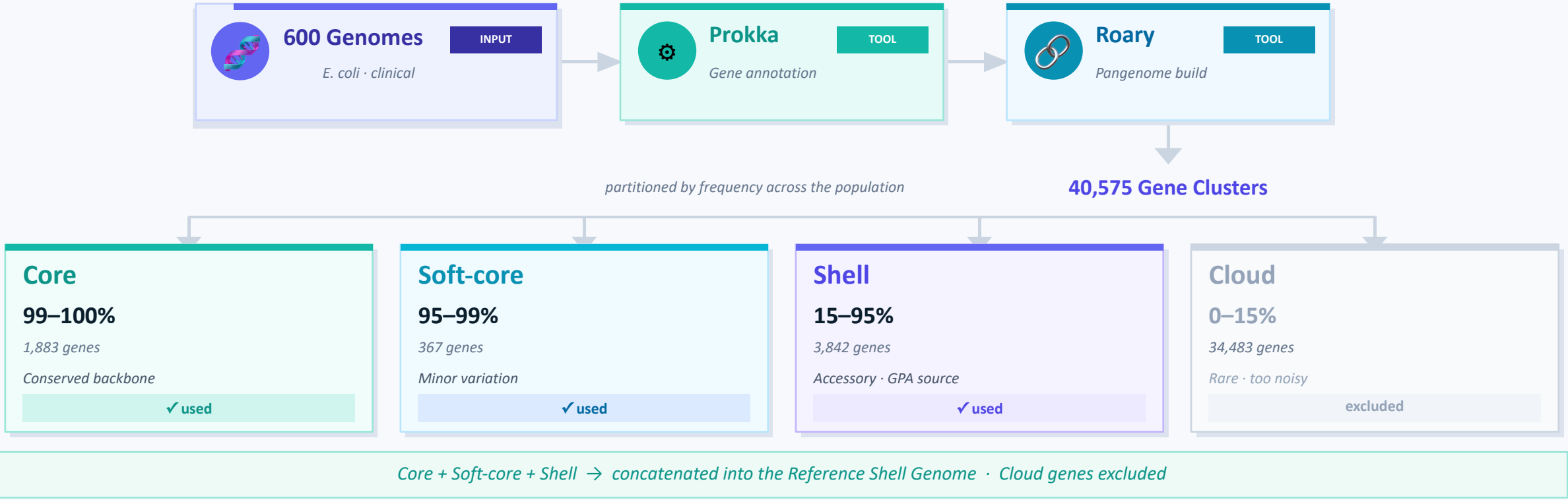
313 susceptible (53.4%) vs 275 resistant (46.6%)



Mild class imbalance (53.4% vs 46.6%) → motivates use of Balanced Accuracy as the primary evaluation metric.

How We Built the Pangenome.

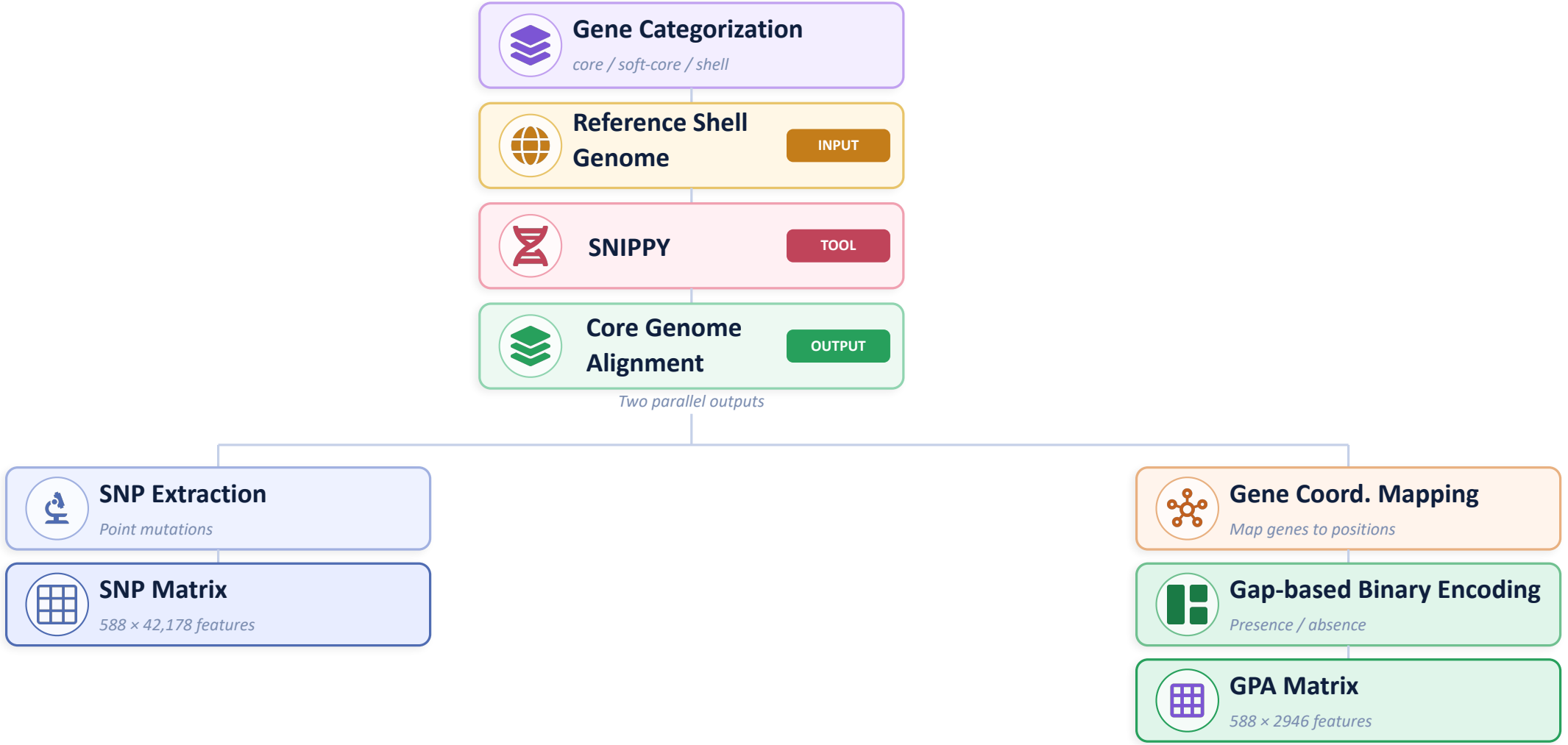
600 E. coli genomes annotated with Prokka · clustered with Roary · genes partitioned by frequency.



PIPELINE

From Reference Genome to Feature Matrices.

Continuing from pangenome output · SNIPPY aligns 600 isolates → two parallel ML feature sets extracted.



SNP matrix captures mutation-driven resistance · GPA matrix captures gene-level presence/absence — two complementary views of the genome.

SNP matrix: filtering & encoding.

High-dimensional SNP data requires filtering, encoding, and feature selection before ML.

166,415 raw variant sites from core genome alignment

↓

56,717 MAF ≥ 0.05 & gap ≤ 10% — remove rare variants & missing data

↓

67,259 one-hot encoding of multi-allelic loci — columns expand per allele

↓

42,178 MI feature selection (median threshold) — informative SNP features retained

⚠ *MultiSURF was computationally infeasible at 67,259 features — MI filter applied to SNPs instead.*

From Alignment to Gene Presence/Absence

Gene coordinate mapping

$$start_i = \sum_{j=1}^{i-1} length_j \qquad end_i = start_i + length_i$$

with $start_1 = 0$ · genes indexed $i = 1, 2, \ldots, n$

Example: cumulative positions in core alignment



$start(\text{atpE}) = length(\text{rpmG}) = 168$

Gap proportion

$$\text{gap_proportion}_{ij} = \frac{\text{number of gap characters in gene } i \text{ for isolate } j}{L_i}$$

Binary encoding

$$presence_{ij} = 1 \text{ if } \text{gap_proportion}_{ij} < 0.5, \quad \text{else } 0$$

Feature reduction

6,092

gene clusters

↓

3,625

after constants & missingness filter

↓

2,946

after MI + MultiSURF (median threshold)

MI captures marginal dependence · MultiSURF detects neighbourhood interactions

Model comparison across SNP and GPA features.

Same models applied independently to both feature sets — a fair, controlled comparison.

SNP Matrix

588 × 47,178

SNP features

Models

CLASSICAL ML

Decision Tree	DT
Random Forest [1]	RF
XGBoost [2]	XGB
Logistic Regression	LR
K-Nearest Neighbours	KNN
TabR [3]	DL

GPA Matrix

588 × 2,946

GPA features

Prediction

Ciprofloxacin resistance

Evaluation

Bal. Accuracy · F1 · ROC-AUC

Prediction

Ciprofloxacin resistance

Evaluation

Bal. Accuracy · F1 · ROC-AUC

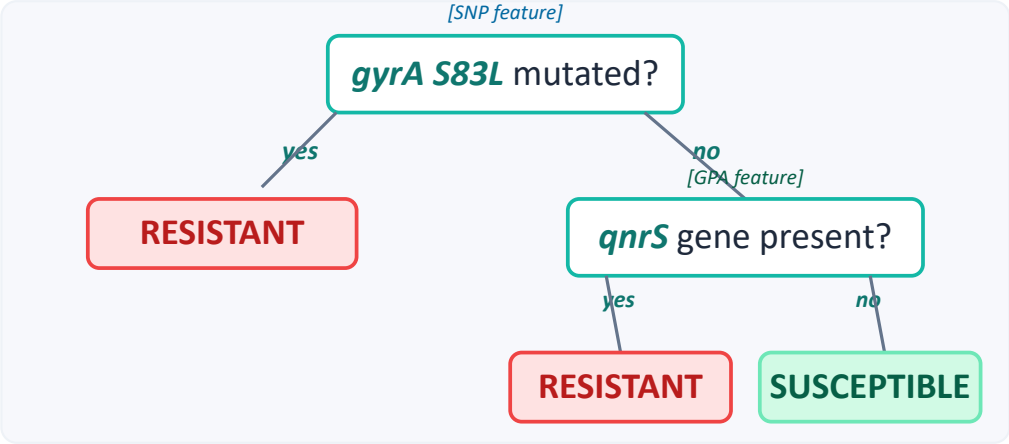
Same metrics applied to both pipelines — no data leakage between SNP and GPA.

[1] Breiman, *Machine Learning*, 2001 [2] Chen & Guestrin, *KDD*, 2016 [3] Gorishniy et al., *ICLR*, 2023

How the models learn.

Classical ML · learns rules

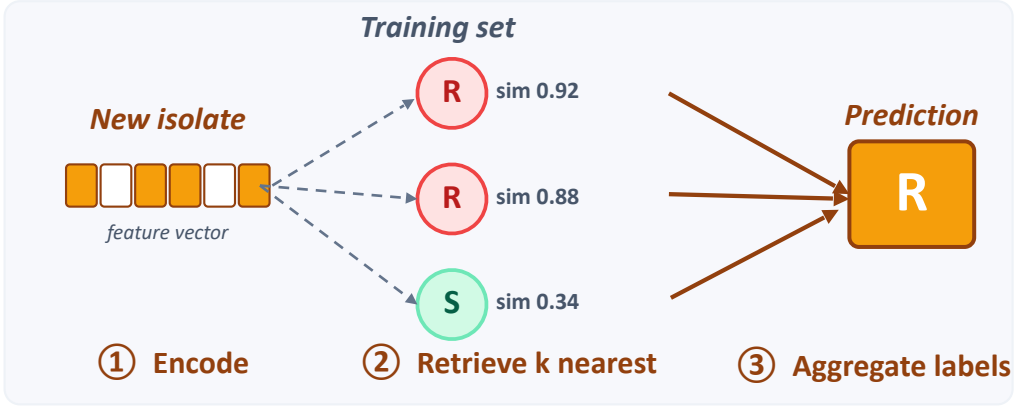
"If gyrA mutated → resistant. Otherwise check if qnrS gene is present."



- **Single concatenated matrix:** SNPs + GPA features per isolate.
- **RF / XGBoost** — many trees vote together.
- **Logistic Regression / kNN** — different decision boundaries, same input.

TabR · learns by comparison

"This isolate looks like these — predict accordingly."



- **Retrieval + aggregation** — weights neighbours by learned similarity.
- No hand-crafted rules — learns its own feature embeddings.

So the question is: *at our dataset size, which approach works better?*

EVALUATION

Experimental setup.

Stratified split

80 / 20

470 train · 118 held-out test

Cross-validation

3-fold

- Stratified 3-fold CV for classical ML
- Fixed train/val split for TabR

Hyperparameters

Optuna

TPE sampler
100 trials per model per fold

Primary metric

Balanced Accuracy.

+ F1, ROC-AUC, Precision, Recall

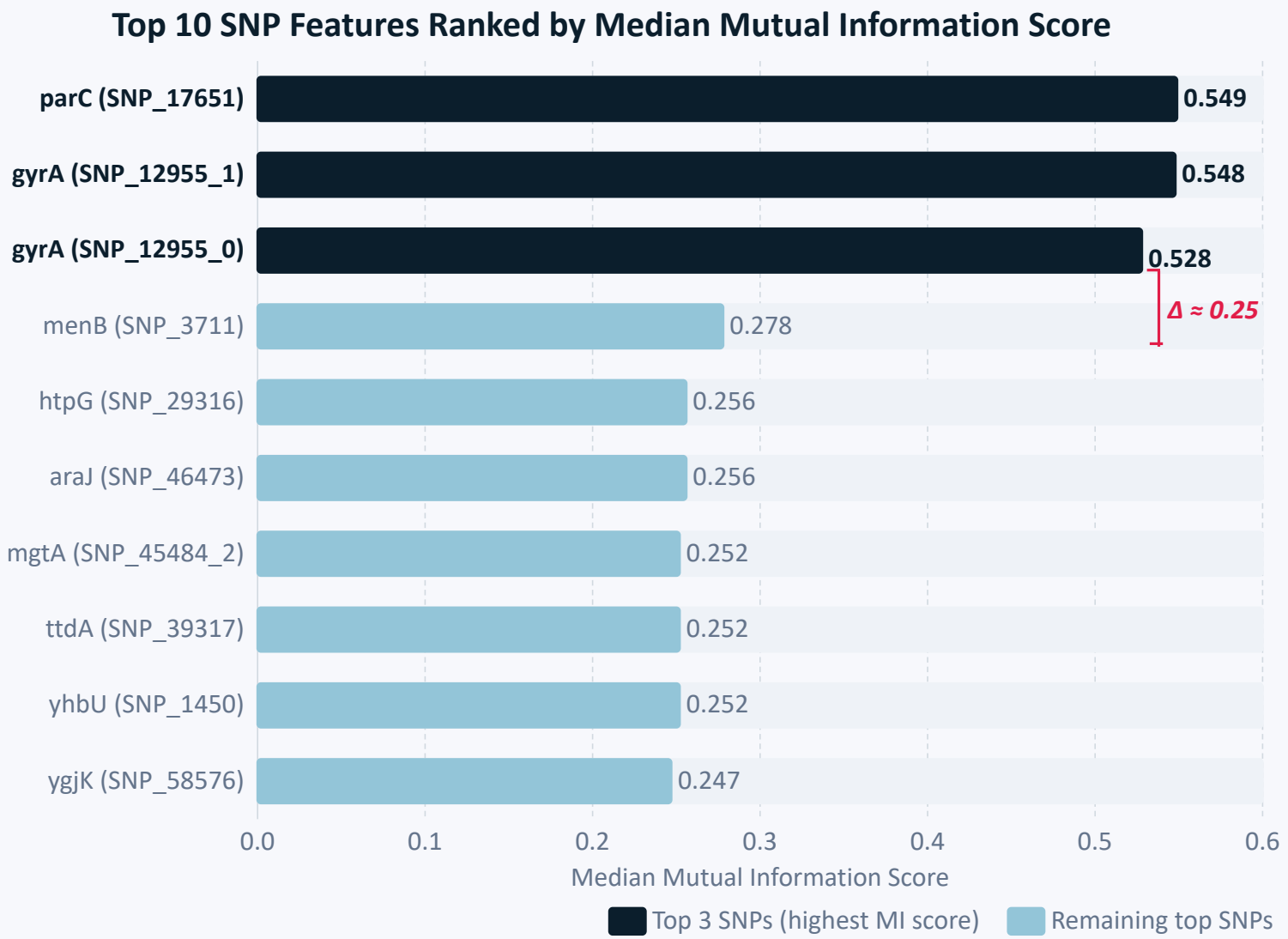
Test set never used during training or hyperparameter tuning · Framework: STREAMLINE + scikit-learn + XGBoost + Optuna

PART 02

Results

Feature importance, model performance, and the GPA vs SNP

Top SNP signals are driven by gyrA and parC mutations.



THE GAP

Δ ≈ 0.25

between rank 3 and rank 4

A small number of SNPs (gyrA, parC) dominate the signal—consistent with known ciprofloxacin resistance mechanisms.

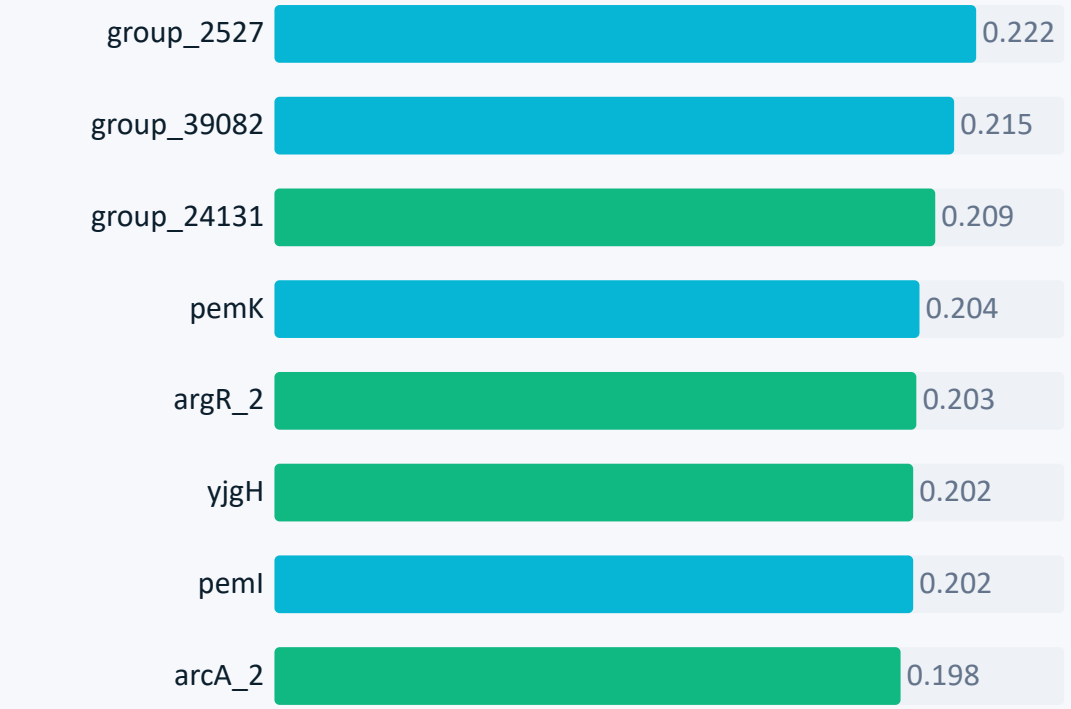
GPA: distributed signal across many genes.

Unlike SNPs, where three mutations dominate, GPA importance is spread across many genes.

✓ **13 of 20** top genes are shared across MI and MultiSURF -> consistent signal across methods

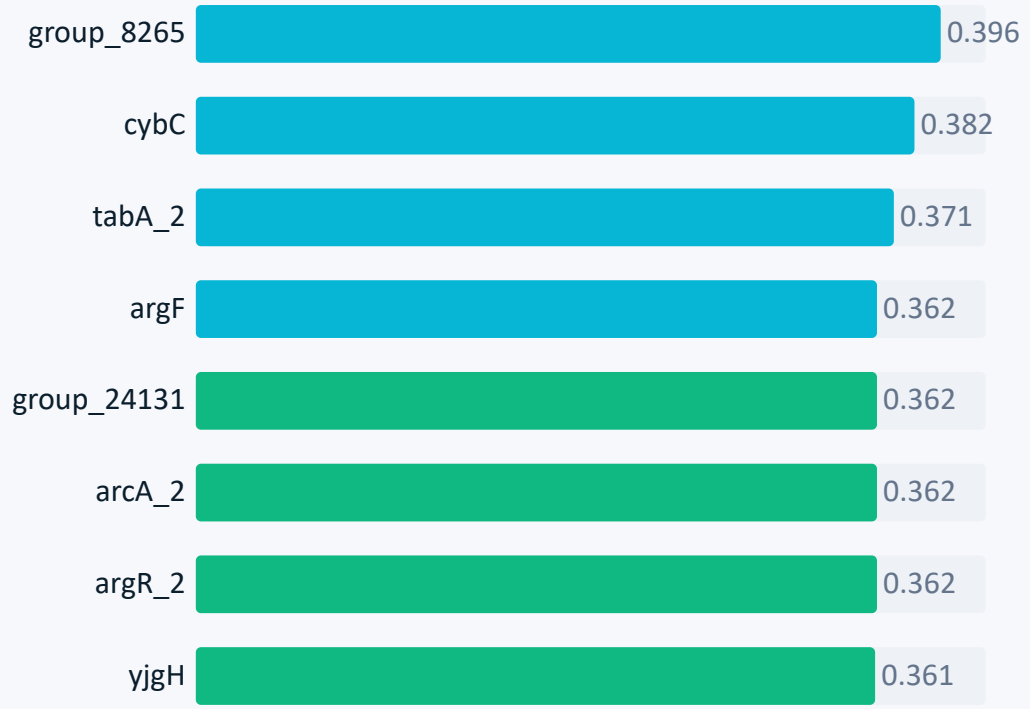
Mutual Information

marginal feature relevance



MultiSURF

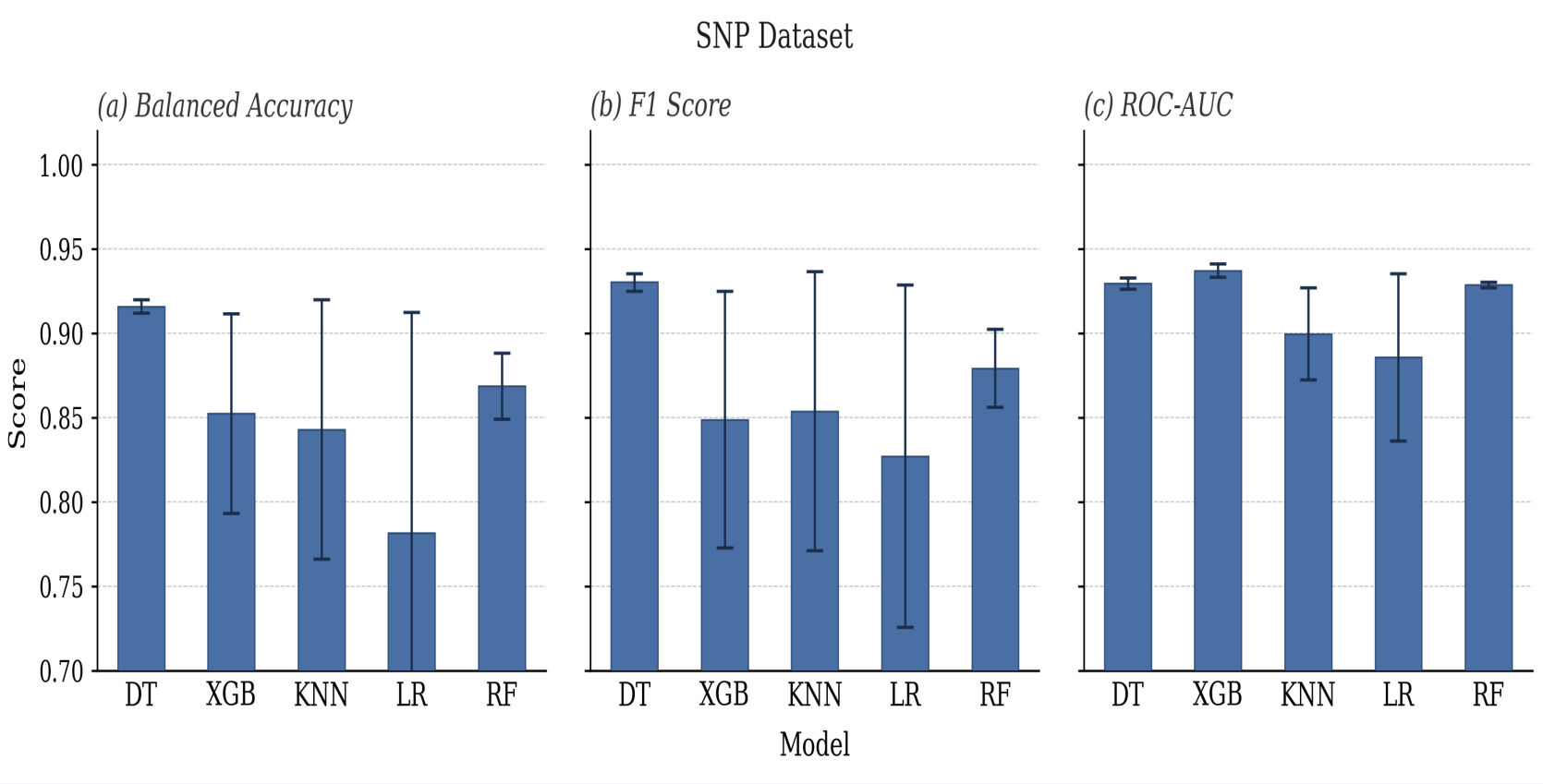
neighbourhood + interaction signal



Shared by both methods Method-specific

SNP test set : result

Primary metric: Balanced Accuracy — accounts for class imbalance in the resistance label.



Decision Tree — Best on SNP

Balanced Accuracy
0.916

F1 Score
0.930

ROC-AUC
0.929

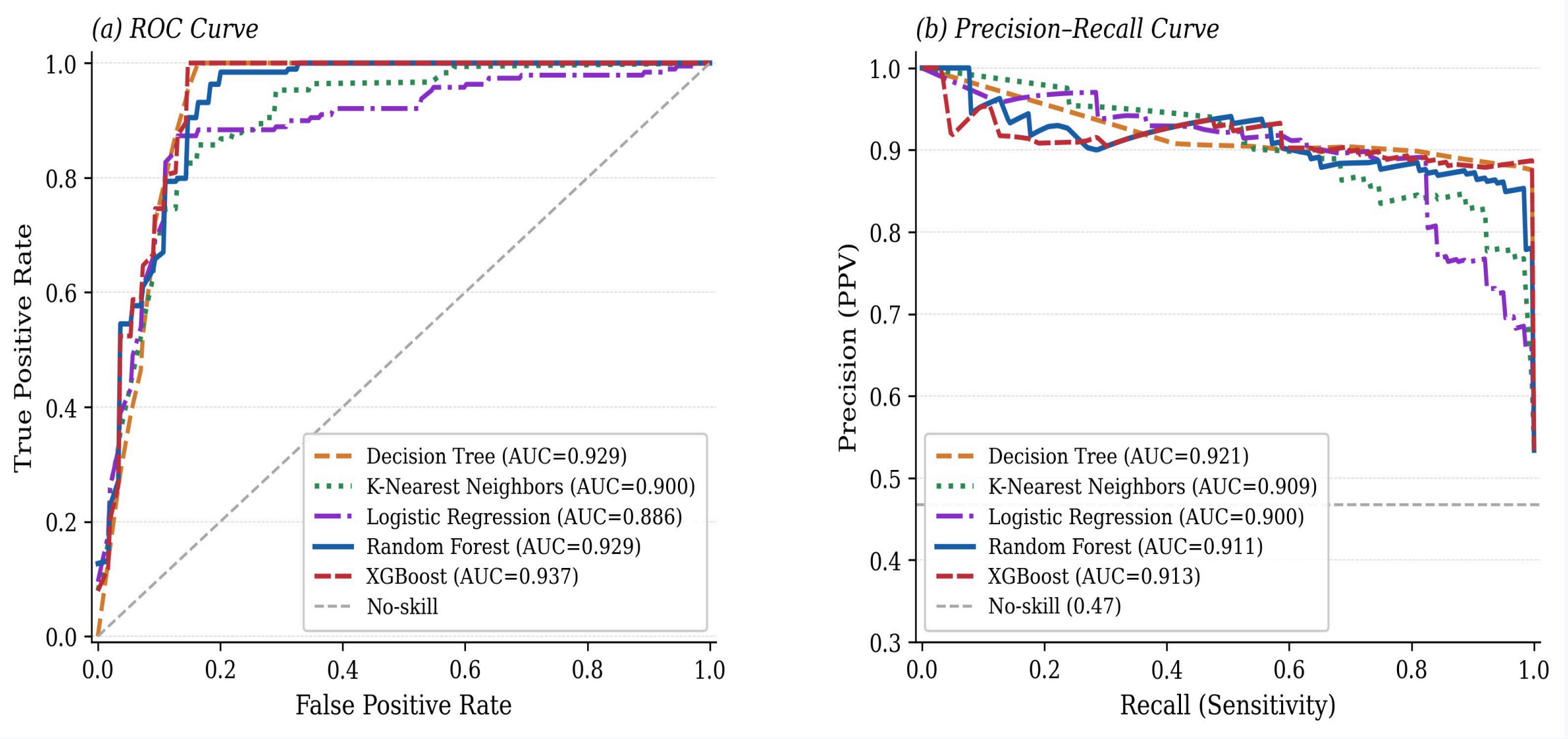
- XGBoost edges ROC-AUC at **0.937** — but trails on Bal. Acc.
- LR lowest (**0.782**) — linear boundaries fail in 47k-dim SNP space.

ROC & PR curves confirm the story — full curves in the next slide.

Tree-based non-linear structure captures gyrA / parC signal most effectively.

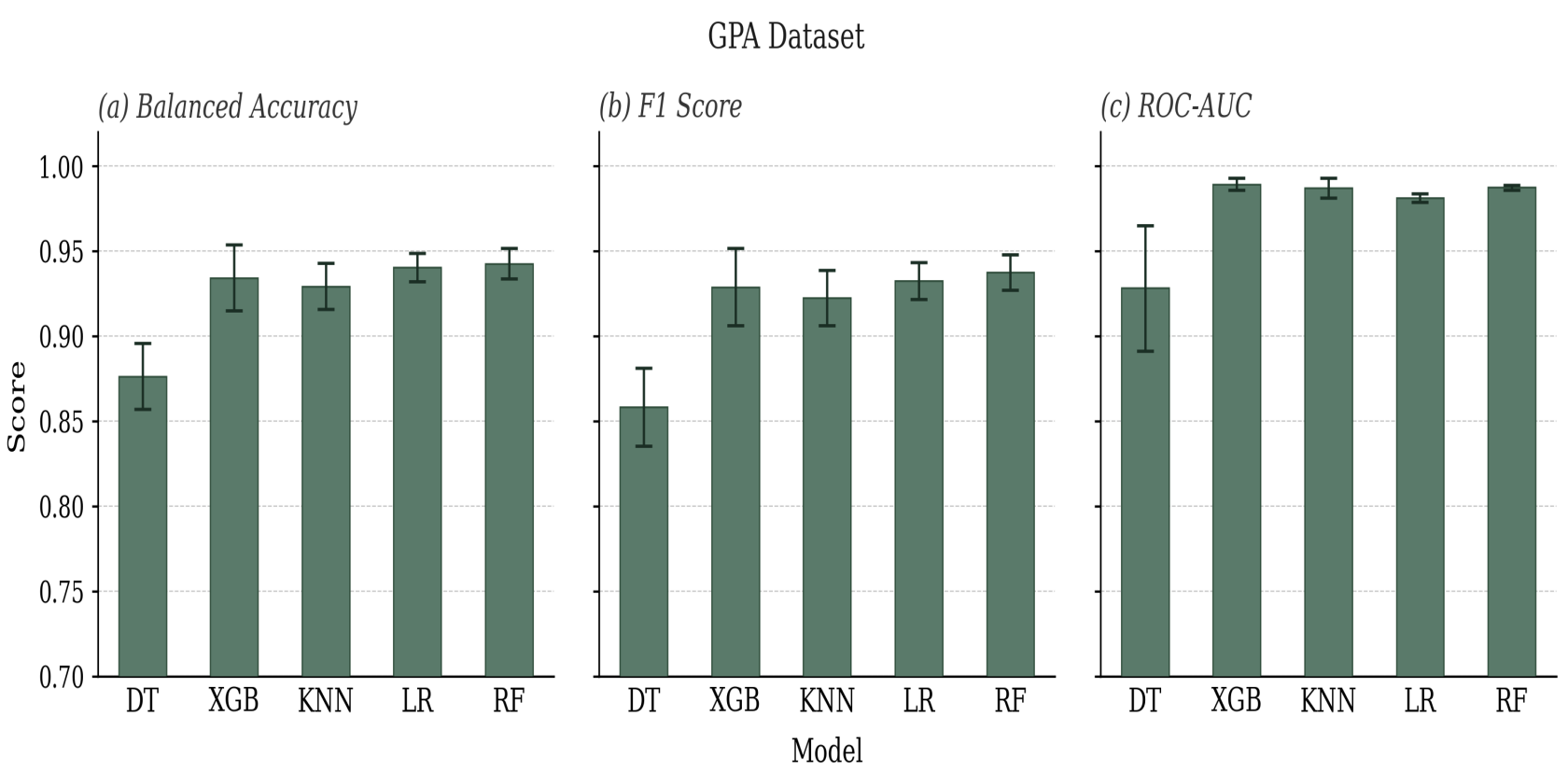
SNP: ROC and Precision-Recall curves.

XGBoost leads ROC-AUC (0.937) — all models well above no-skill baseline.



GPA test set: consistent performance across models.

Same metric, same structure — direct comparison with SNP results.



Random Forest — Best on GPA

Balanced Accuracy
0.943

F1 Score	ROC-AUC
0.938	**0.987**

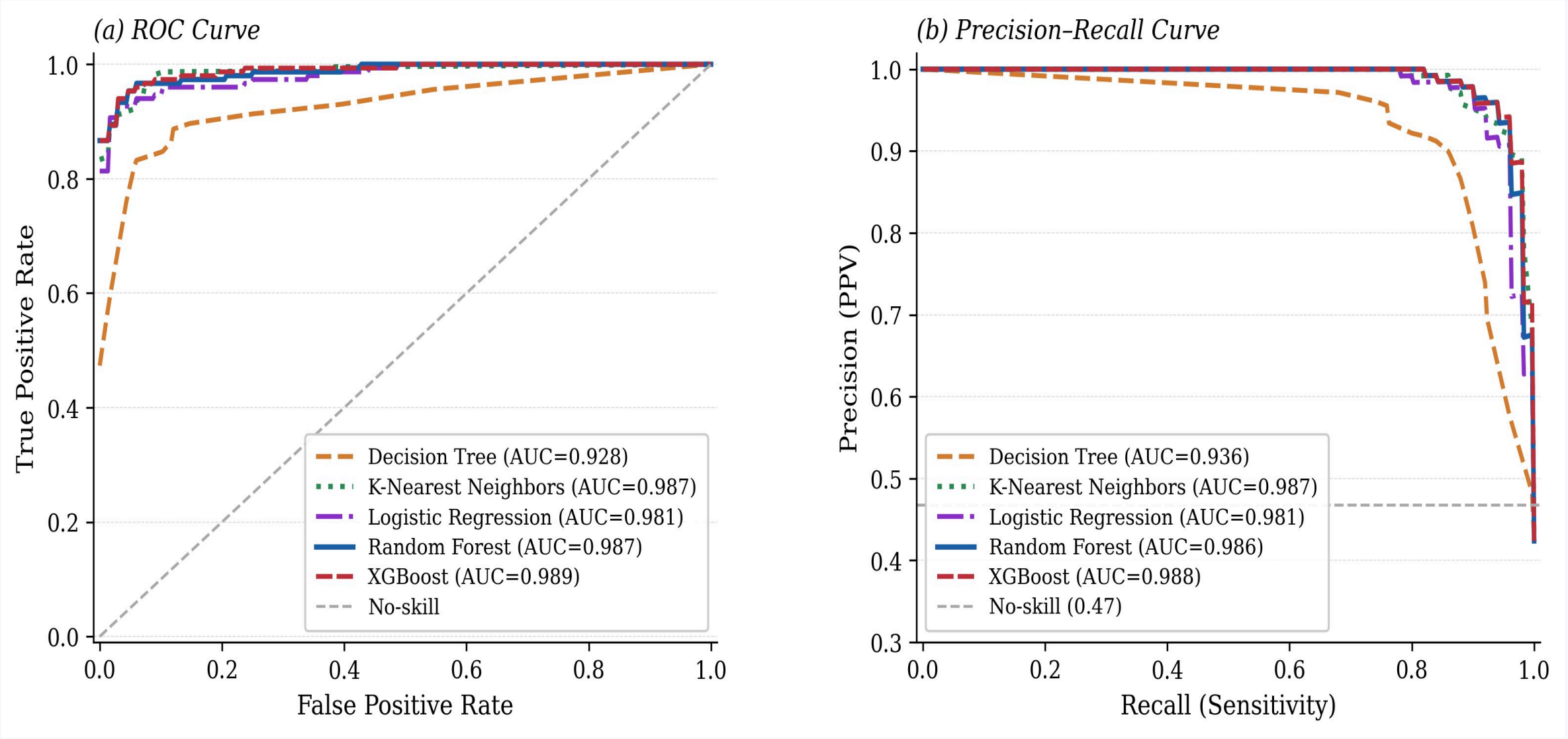
- Four models above **0.930** — far more consistent than SNP.
- Decision Tree lags (**0.876**) — shallow splits miss distributed gene signals.

ROC & PR curves confirm near-perfect discrimination — full curves in appendix.

Structured gene features reduce model sensitivity — GPA generalises better.

GPA: near-perfect ROC and Precision-Recall curves.

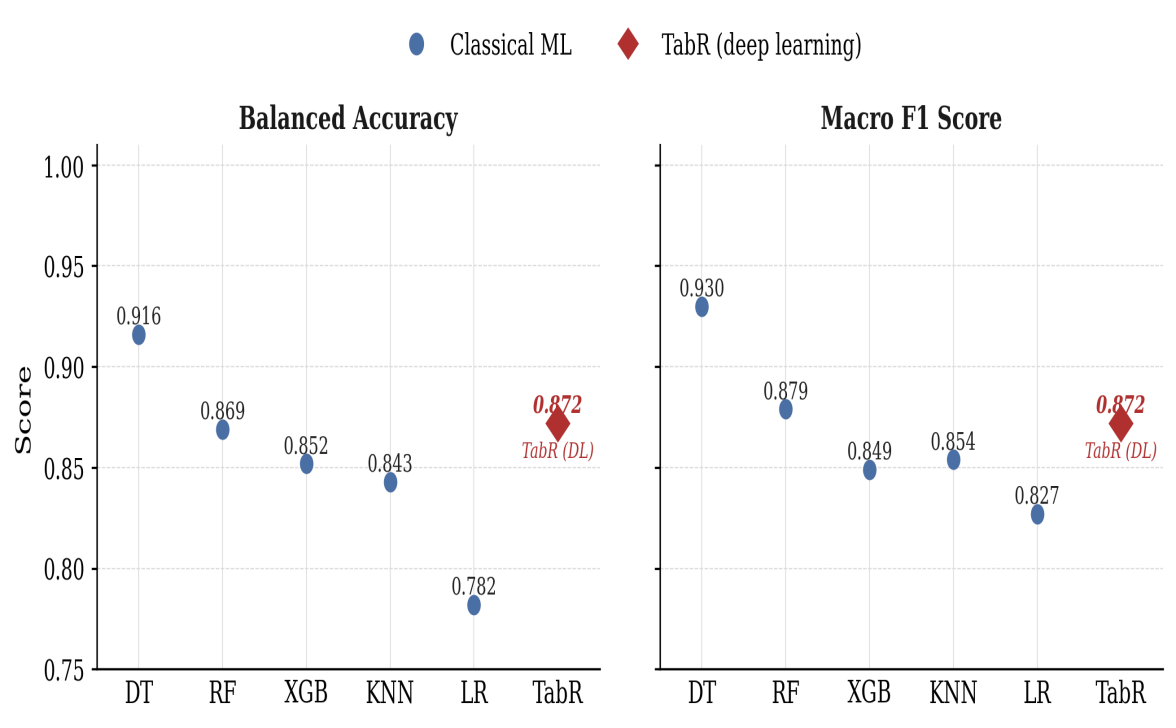
Four models above AUC 0.981 — GPA curves hug the top-left corner, far stronger than SNP.



Classical ML outperforms TabR on both datasets.

TabR — retrieval-augmented deep learning — trails classical models on both SNP and GPA feature sets.

SNP Dataset



GPA Dataset



SNP
Best classical (DT) 0.916
TabR 0.872 Δ = −0.044

GPA
Best classical (RF) 0.943
TabR 0.925 Δ = −0.018

TabR underperforms on both datasets — retrieval-augmented learning likely requires larger training sets than 588 isolates.

GPA outperforms SNP across every model except DT



✓ *Same trend holds for F1 Score and ROC-AUC — GPA leads on both metrics across all models.*

Best model per dataset

SNP · Decision Tree
0.916

GPA · Random Forest
0.943

Key takeaway

GPA outperforms SNP across all models

GPA gene presence/absence captures a cleaner, more structured signal for ciprofloxacin resistance.

Three findings, taken together.

01

GPA > SNP for prediction

★ Key finding

GPA features predict better than SNP features, even though SNP mutations in gyrA/parC are the known biological cause. Counter to our hypothesis

02

Classical ML > Deep Learning

Tree ensembles beat TabR on 588 isolates — deep learning needs more data.

03

Biology and ML agree on the signal

SNPs point to gyrA/parC mutations. GPA points to horizontally acquired resistance genes — ML captures both mechanisms.

These three findings together answer the central research question.

LIMITATIONS

Honest about the boundaries.

Dataset size

588 isolates — sufficient for classical ML, insufficient for deep learning.

Shell genome only

Only shell genes used. Full species-wide pangenome not computed — some E. coli diversity may be missed.

No feature integration

SNP and GPA evaluated separately. Combined representations not explored.

Selection thresholds

Median MI threshold used for feature selection. MultiSURF not feasible at SNP dimensionality.

Despite these limitations, results are consistent and biologically meaningful.

What we found — and where to go next.

Conclusion

GPA features consistently outperform SNP features across all models — structured gene representations generalise better.

Classical ML (RF, LR, XGB) outperforms TabR on 588 isolates — deep learning requires more data.

gyrA and parC SNPs carry the dominant ciprofloxacin resistance signal — biology and ML agree.

A pangenome-based pipeline is a viable, scalable framework for AMR prediction in E. coli.

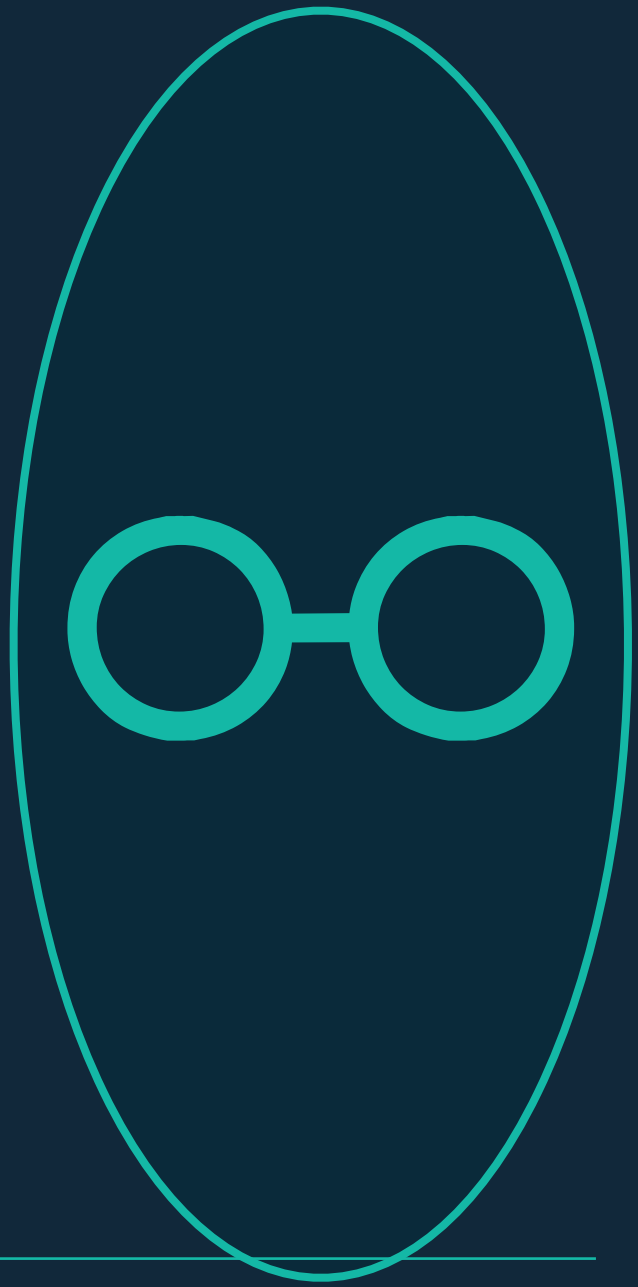
Future Work

- Integrate SNP + GPA into a unified feature representation to capture complementary biological signals.
- Apply transformer-based or graph neural network models designed for high-dimensional genomic data.
- Expand dataset with additional public E. coli bloodstream isolates to power deep learning.
- Extend framework to multi-drug resistance prediction beyond ciprofloxacin.

GPA features provide a more robust and learnable representation for predicting antibiotic resistance.

THANK YOU

Questions?



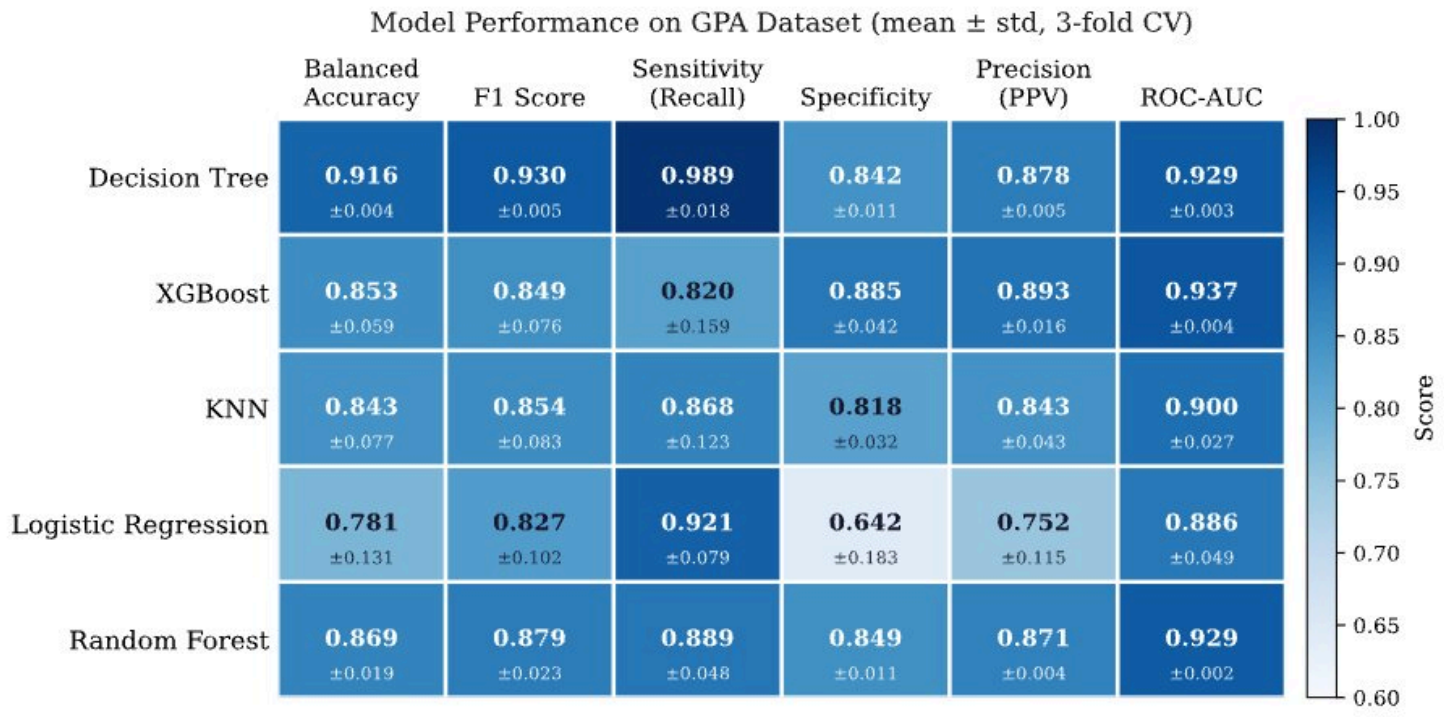
ACKNOWLEDGMENTS

Committee: Dr. Sara El Alaoui · Dr. Akila de Silva · Dr. Pleuni S. Pennings

CodeLab team: Estefanos Kebebew, Myco Torest, Gian Carlo, Juve Barajas

Funding: NIH–UCSF/SFSU Bridges to the Doctorate Fellowship (2023–2025) · SFSU SEO program

SNP Dataset — Model Performance



Mean ± std across 3-fold stratified CV. Note: in-figure title reads "GPA" — this panel reports the SNP dataset.

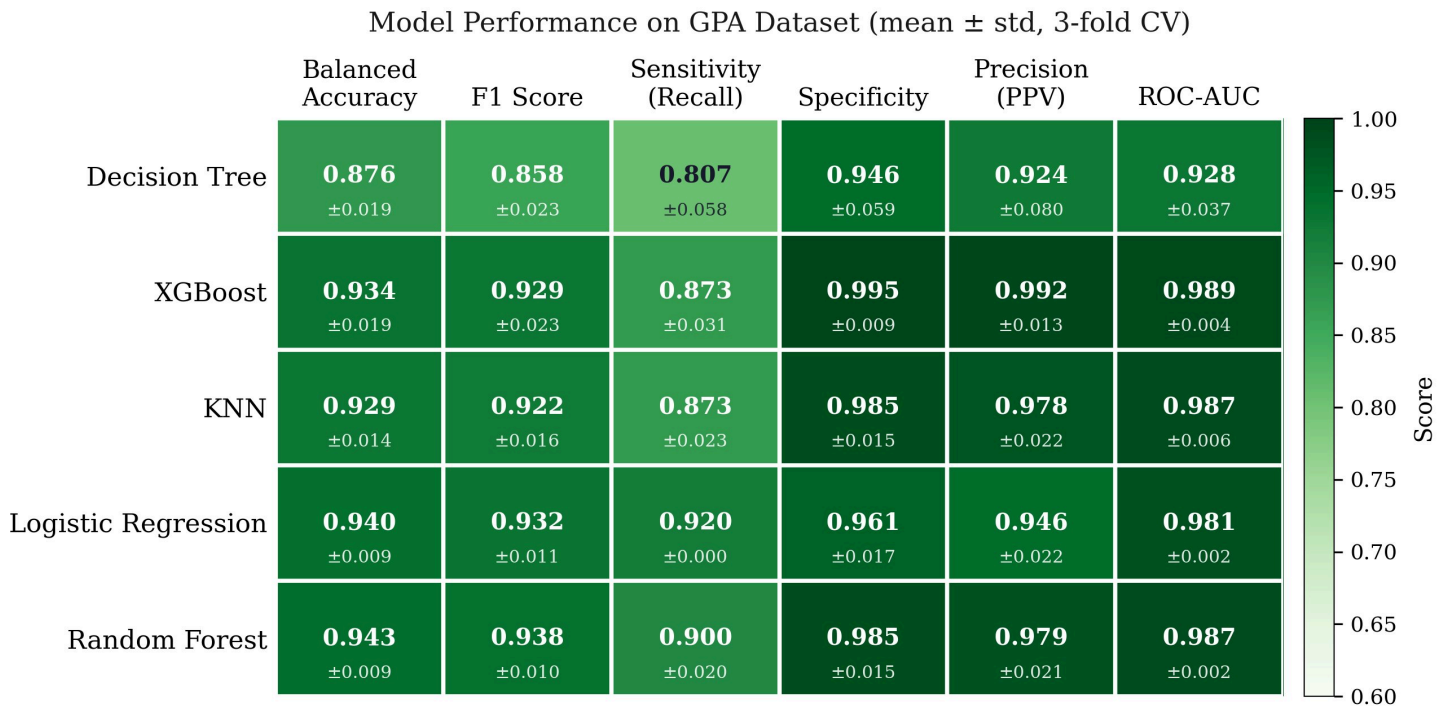
KEY OBSERVATIONS

Best: Decision Tree

Balanced Accuracy: 0.916 ± 0.004

- Decision Tree leads on BA, F1, Recall (0.989) — strongest sensitivity overall.
- XGBoost: highest ROC-AUC (0.937), best calibrated ranker.
- Logistic Regression: high recall (0.921) but low specificity (0.642) — high false-positive rate.
- Wide variance across models suggests SNP space is harder to learn than GPA.
- Feature space: 47,178 selected per fold.

GPA Dataset — Model Performance



Mean ± std across 3-fold stratified CV. Gene Presence/Absence binary feature matrix.

KEY OBSERVATIONS

Best: Random Forest

Balanced Accuracy: 0.943 ± 0.009

- All models ≥ 0.876 BA — GPA is consistently easier to learn than SNP.
- RF, LR, XGBoost cluster within 0.009 BA — robust signal across model families.
- XGBoost: highest specificity (0.995) and precision (0.992) — fewest false positives.
- ROC-AUC ≥ 0.981 across top 4 models — strong class separability.
- Feature space: 2,946 binary features after selection.

Selected Hyperparameters (Optuna, TPE)

100 Optuna trials per algorithm per fold · TPE sampler · objective: Balanced Accuracy · seed = 42

Classical ML — Best Hyperparameters

Model	Hyperparameter	SNP	GPA
DT	criterion	gini	gini / entropy
	max_depth	29	12 — 21
	min_samples_leaf	2 — 30	3 — 24
RF	n_estimators	381	354 — 496
	max_depth	18	23 — 27
	criterion	gini	gini / entropy
XGB	n_estimators	868	107 — 913
	max_depth	5	19 — 27
	eta (lr)	0.0072	0.0003 — 0.20
	subsample	0.510	0.524 — 0.776
KNN	n_neighbors	4 — 42	2 — 11
	weights	distance	distance / uniform
LR	solver	newton-cg / lbfgs / liblinear	newton-cg / sag / saga
	C	0.20 — 9.8e4	0.07 — 393.8

SNP folds (F0–F2) generally agreed on values; GPA shows wider per-fold variation. Range shown where folds differ.

TabR — Final Hyperparameters

Group	Parameter	Value
Training	batch_size	256
	n_epochs	100
	context_size	96
Model	d_main	256
	d_multiplier	2.0
	predictor_n_blocks	1
	context_dropout	0.2
	normalization	LayerNorm
Optim.	type	AdamW
	lr	3e-4
	weight_decay	1e-4

Identical hyperparameters across SNP and GPA datasets — suggests architectural robustness to feature representation.

References

References cited in presentation

1 **Murray, C.J.L., et al. (2022).** Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *The Lancet*, 399(10325), 629–655.

2 **Kot, B. (2019).** Antibiotic resistance among hospital pathogens. *Polish Journal of Microbiology*, 68(4), 403–415.

3 **Versporten, A., et al. (2019).** Quantifying the contribution of resistance mechanisms to ciprofloxacin MIC in E. coli. *Journal of Antimicrobial Chemotherapy*, 74(2), 298–310.

4 **Moradigaravand, D., et al. (2018).** Prediction of antibiotic resistance in Escherichia coli from large-scale pan-genome data. *PLOS Computational Biology*, 14(12), e1006258.

5 **Gladstone, R.A., et al. (2021).** Emergence and dissemination of antimicrobial resistance in Escherichia coli. *The Lancet Microbe*, 2(12), e688–e699.

6 **Mills, M.C., et al. (2022).** Antimicrobial resistance patterns in E. coli from the US Military Healthcare Network. *Genome Medicine*, 14, 32.

7 **Breiman, L. (2001).** Random Forests. *Machine Learning*, 45, 5–32.

8 **Chen, T., & Guestrin, C. (2016).** XGBoost: A scalable tree boosting system. *Proceedings of KDD*, 785–794.

9 **Gorishniy, Y., et al. (2023).** TabR: Tabular deep learning meets nearest neighbours. *ICLR*, 2024.

References

Additional thesis references

10 **Page, A.J., et al. (2015).** Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics*, 31(22), 3691–3693.

11 **Seemann, T. (2014).** Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, 30(14), 2068–2069.

12 **Seemann, T. (2015).** Snippy: fast haploid variant calling from NGS reads. *GitHub*, github.com/tseemann/snippy.

13 **Akiba, T., et al. (2019).** Optuna: A next-generation hyperparameter optimization framework. *Proceedings of KDD*, 2623–2631.

14 **Urbanowicz, R.J., et al. (2023).** STREAMLINE: A simple, transparent, end-to-end automated machine learning pipeline. *Journal of Biomedical Informatics*, 145, 104424.

15 **Tettelin, H., et al. (2005).** Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial pan-genome. *PNAS*, 102(39), 13950–13955.

16 **Orcales, P.B., et al. (2024).** Machine learning approaches for predicting antimicrobial resistance. *PLOS Computational Biology*, 20(1), e1011937.

17 **Kamel, A.M., et al. (2021).** Ciprofloxacin resistance among uropathogenic Escherichia coli isolates. *BMC Infectious Diseases*, 21, 996.

18 **Rottier, W.C., et al. (2021).** Escherichia coli resistance to fluoroquinolones in community-acquired urinary tract infection. *Antimicrobial Agents and Chemotherapy*, 65(1).