KEGG Orthology-Based Machine Learning Reveals Functional Determinants of Antimicrobial Resistance in *Acinetobacter baumannii*

Zhihang Zheng, Bei Jiang, Abebe Mekuria Shenkutie, Jingyuan Bian, Yuyao Yan, Ruizhen Pi, Qing Xiong, Polly Hang-Mei Leung

Department of Health Technology and Informatics, The Hong Kong Polytechnic University, Kowloon, Hong Kong SAR, China

Microbiology Spectrum (2026) | doi:10.1128/spectrum.02592-25

1. BACKGROUND

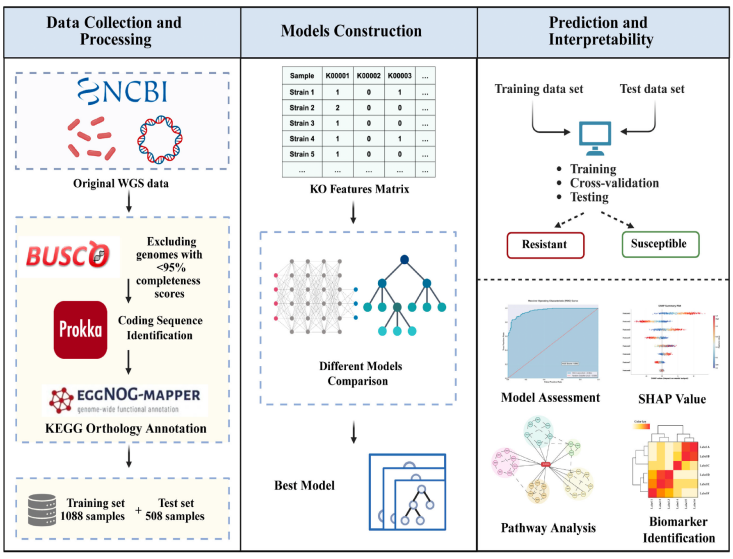
Antimicrobial resistance in *Acinetobacter baumannii* is a major global health threat.

Conventional antimicrobial susceptibility testing can take up to 48 hours, delaying appropriate treatment.

This study proposes a functional-genomics machine learning framework using KEGG orthology annotations from whole-genome sequencing to predict antimicrobial resistance phenotypes and uncover interpretable resistance mechanisms.

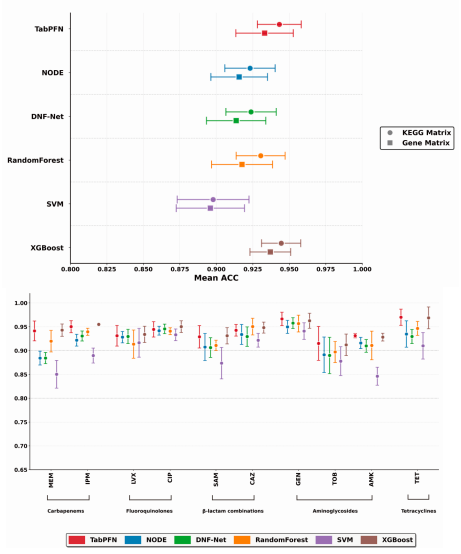
2. METHODS

- Training cohort: 1,088 *A. baumannii* isolates.
- Independent test cohort: 508 *A. baumannii* isolates.
- Whole-genome data were quality-controlled using BUSCO, retaining genomes with >95% completeness.
- Coding sequences were annotated with Prokka, followed by KEGG orthology assignment using eggNOG-mapper.
- A KO abundance matrix was constructed and used to train six machine learning models
- Model performance was evaluated using five-fold cross-validation and an independent test cohort.
- SHAP values and XGBoost feature importance were used to identify resistance-associated KO



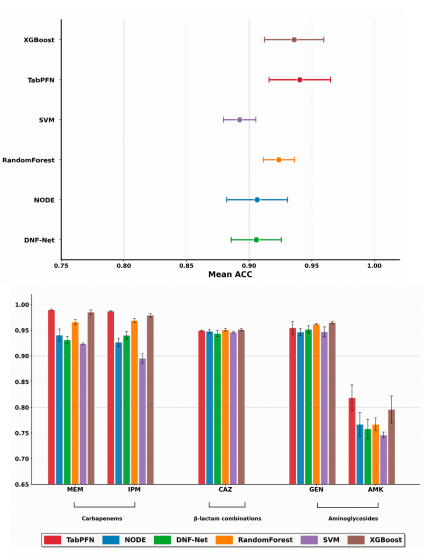
3. RESULTS

A. Model Performance Comparison



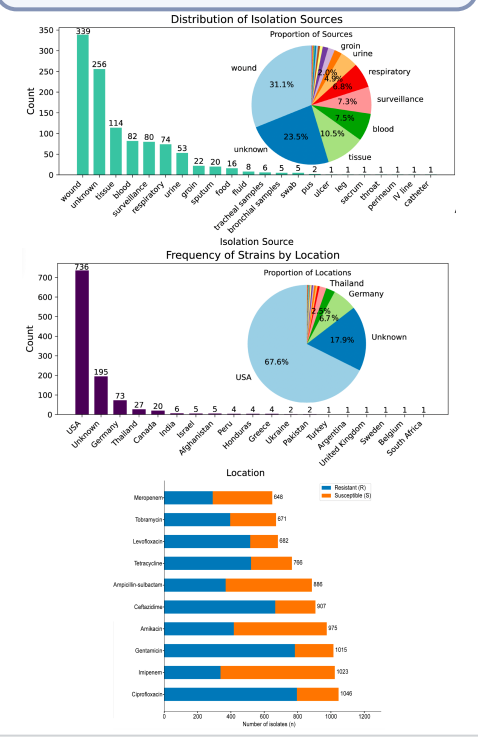
1. KEGG orthology-based functional representation consistently outperformed gene presence–absence encoding across all machine learning models.
2. In five-fold cross-validation, XGBoost achieved the highest weighted accuracy of 94.44%, followed closely by TabPFN at 94.32%.

B. Test Performance



1. XGBoost performed strongly for carbapenem resistance prediction, while TabPFN maintained robust performance across multiple antibiotics.
2. Amikacin showed a notable generalization gap in the independent cohort

C. Epidemiology&Resistance Overview



4. KEY RESISTANCE DETERMINANTS

XGBoost interpretability identified biologically meaningful KO terms.

Carbapenems: K19213 (*bla*OXA-12) was a top predictor.

Multidrug marker: K18974 (*sul1*) emerged as a highly prevalent shared determinant.

Non-conventional mechanism: K21801 (*iaaH*) suggested a role for indole-mediated signaling.

Aminoglycosides: K19299 and K19300 aminoglycoside phosphotransferases.

Tetracycline: K08151 (*tetA*) and K18476 (*tetR*).

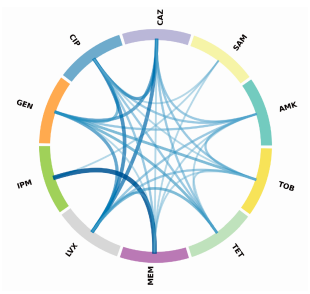
Pathway enrichment highlighted β-lactam resistance pathways.

Top Predictive KO Terms (XGBoost Feature Importance)

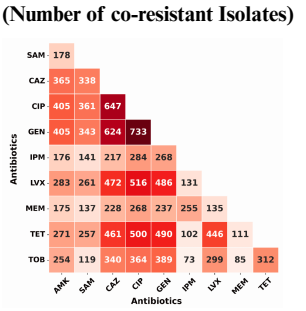
KO Term	Gene / Function
K19213	*bla*$_{OXA-12}$ (carbapenemase)
K18974	*sul1* (multidrug efflux)
K21801	*iaaH* (indole-3-acetamide hydrolase)
K19299	*aphA6* (aminoglycoside phosphotransferase)
K19300	*aph(3')-Ia* (aminoglycoside phosphotransferase)
K08151	*tetA* (tetracycline efflux)
K18476	*tetR* (tetracycline repressor)
K02022	*ampC* (β-lactamase)
K07410	*oprD* (porin D)
K03321	*gyrA* (DNA gyrase subunit A)

5. MULTIDRUG RESISTANCE

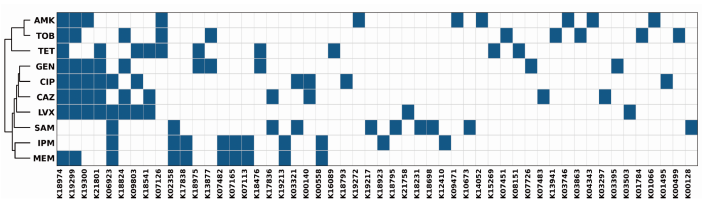
A. Shared Resistance Network



B. Co-resistance Matrix (Number of co-resistant Isolates)



C. Molecular Connectivity Heatmap (Top 10 Resistance-related KOs)



5. CONCLUSION & SIGNIFICANCE

KEGG orthology provides a more informative functional representation than simple gene presence–absence.

Results reveal both canonical resistance genes and broader metabolic or signaling contributors to resistance.

The framework combines strong predictive performance with biological interpretability.

This provides a foundation for rapid genome-based AMR prediction and future precision antimicrobial therapy.