npj | systems biology and applications

Article

Published in partnership with the Systems Biology Institute

https://doi.org/10.1038/s41540-025-00584-0

Minimal gene signatures enable high-accuracy prediction of antibiotic resistance in *Pseudomonas aeruginosa*

Nabia Shahreen[1,4], Syed Ahsan Shahid[2,4], Mahfuze Subhani[3,4], Adil Al-Siyabi[2] ✉ & Rajib Saha[1] ✉

Antimicrobial resistance (AMR) in *Pseudomonas aeruginosa* poses a critical global health challenge, with current diagnostics relying on slow, culture-based methods. Here, we present a ML framework leveraging transcriptomic data to predict antibiotic resistance with high accuracy. We applied a genetic algorithm to 414 clinical isolates to identify minimal, highly predictive gene sets (~35–40 genes) distinguishing resistant from susceptible strains for meropenem, ciprofloxacin, tobramycin, and ceftazidime. Automated ML classifiers trained on these sets achieved accuracies of 96–99% on test data (F1 scores: 0.93–0.99), surpassing clinical deployment thresholds. Multiple distinct, non-overlapping gene subsets exhibited comparable performance, suggesting that resistance acquisition is associated with changes in the expression of diverse regulatory and metabolic genes. Comparison with known resistance markers from CARD and operon annotations revealed a substantial number of previously unannotated clusters, highlighting significant knowledge gaps in current AMR understanding. Mapping these genes onto independently modulated gene sets (iModulons) revealed transcriptional adaptations across diverse genetic regions. Overall, this study presents a streamlined machine-learning workflow for transcriptomic data and offers a pathway toward rapid diagnostics and personalized treatment strategies against AMR.

Antimicrobial resistance has emerged as one of the most urgent threats to global public health, undermining the effectiveness of existing treatments and increasing the risk of untreatable infections. The World Health Organization (WHO) has identified AMR as one of the top ten global health threats, projecting that, without intervention, it could lead to 10 million deaths annually by 2050[1]. Among the most concerning pathogens is *P. aeruginosa*, a gram-negative opportunistic bacterium responsible for severe infections such as pneumonia, urinary tract infections, and bacteremia, particularly in immunocompromised patients[2–4]. The threat posed by this bacterium is exacerbated by its intrinsic resistance mechanisms, including efflux pumps and reduced outer membrane permeability, as well as its ability to rapidly acquire new resistance determinants, leading to the emergence of multidrug-resistant (MDR) and even pan-resistant strains[5,6].

Despite the escalating threat, clinical practice still relies primarily on culture-based antibiotic susceptibility testing, which, while reliable, can require 48–72 hours to yield results[7]. This delay necessitates empirical treatment with broad-spectrum antibiotics, which may be ineffective and can further drive resistance[8]. Moreover, culture-based assays provide limited insight into the genetic or molecular drivers of resistance, offering little guidance for targeted interventions[9].

In recent years, high-throughput sequencing, particularly transcriptomic profiling, has enabled a more fine-grained view of resistance by capturing the gene expression that underpins survival under antibiotic pressure[10–13]. Transcriptomic data provides a snapshot of the cellular state under antibiotic pressure, revealing pathways and regulatory networks that contribute to survival. This approach holds promise for identifying biomarkers of resistance, enabling earlier and more precise diagnostics[12,14,15]. However, leveraging transcriptomic data for AMR surveillance and prediction remains challenging because of the high dimensionality of the data, which complicates the identification of relevant features for accurate and interpretable predictions[16].

One promising avenue for overcoming these challenges is machine learning, which can handle large omics datasets and uncover complex patterns relevant to resistance phenotypes[17]. Prior studies have demonstrated the feasibility of ML-driven AMR prediction via genomic and transcriptomic data, highlighting the potential for faster and more precise

[1]Department of Chemical and Biomolecular Engineering, University of Nebraska-Lincoln, Lincoln, NE, USA. [2]Natural and Medical Sciences Research Center, University of Nizwa, Birkat Al-Mouz, Nizwa, Oman. [3]School of Computing, University of Nebraska-Lincoln, Lincoln, NE, USA. [4]These authors contributed equally: Nabia Shahreen, Syed Ahsan Shahid, Mahfuze Subhani. ✉e-mail: aalsiyabi@unizwa.edu.om; rsaha2@unl.edu

diagnostics[18]. However, a major challenge remains in identifying minimal yet robust gene subsets that preserve predictive accuracy while improving interpretability and reducing computational costs. In a notable study, Khaledi et al., 2020 demonstrated that integrating single nucleotide polymorphisms (SNPs), gene presence/absence (GPA), and transcriptomic expression data enabled predictive modeling of antibiotic resistance, with sensitivity and predictive values between 0.81 and 0.95 across four antibiotics. However, their reliance on a high-dimensional feature space (up to 93 markers per antibiotic) and mixed-data approaches limit clinical scalability owing to cost, interpretability, and generalizability constraints. Additionally, their pipeline required manual feature selection and hyperparameter tuning, lacking a fully optimized and automated feature selection mechanism.

To address these limitations, we implemented a hybrid methodology combining genetic algorithm (GA)-based feature selection with automated ML (AutoML), leveraging transcriptomic data from 414 *P. aeruginosa* clinical isolates (Fig. 1, Supplementary Fig. 1). By iteratively evolving gene subsets, the GA identified minimal, highly informative features, which were then used to train ML models with accuracies of 0.96–0.99 on held-out test data. Here, we define 'gene subsets' as the sets of genes selected based on their expression profiles across isolates. Despite minimal overlap among gene subsets, each yielded similarly high predictive power, suggesting a pervasive, multifactorial transcriptomic signature of resistance. To resolve whether these subsets shared underlying biological processes beyond direct gene overlap, we employed three complementary analyses. First, we compared GA-selected genes to known resistance determinants in the CARD, examining the extent to which recognized resistance markers drove the predictions[19]. Second, we explored operon-level organization to evaluate whether co-transcribed gene clusters were consistently selected, revealing potential regulatory "hotspots"[20,21]. Finally, we mapped the subsets to iModulons, which are coregulated gene modules derived from independent component analysis (ICA), to elucidate higher-order transcriptional control mechanisms associated with resistance[22–24]. Our results show that gene expression patterns predictive of AMR in *P. aeruginosa* extend beyond canonical resistance genes, suggesting that diverse transcriptional responses may characterize the resistant phenotype. While some GA-selected loci mapped to well-characterized efflux pumps or β-lactamase operons, many fell outside conventional AMR annotations, pointing to underexplored determinants. Operon-level analysis revealed recurrent co-transcribed clusters involved in osmotic stress, iron acquisition, and various metabolic pathways, whereas iModulon mapping revealed a convergence of transcriptional programs governing oxidative stress responses, DNA repair, efflux regulation, and ribosomal function. Collectively, these data highlight that AMR phenotypes correlate with transcriptomic patterns spanning diverse genetic loci, including both isolated resistance genes and genes implicated in broader cellular processes.

Overall, in this study, we present a multiscale framework that integrates transcriptomic profiling, evolutionary feature selection, and in-depth biological network analysis to enhance AMR diagnostics. By pinpointing compact yet functionally enriched gene sets linked to specific operons and regulatory modules, our approach offers a balance of predictive accuracy, interpretability, and clinical feasibility. These findings underscore the value of system-level perspectives on AMR and support the development of next-generation diagnostics and personalized therapeutic strategies.

Results

Automated ML and the genetic algorithm uncover minimal yet high-performing gene sets

Predicting antibiotic resistance from gene expression in *P. aeruginosa* requires a balance between accuracy and interpretability, as high-dimensional transcriptomic data present computational and clinical challenges. To address these constraints, we developed a hybrid GA-AutoML pipeline designed to systematically identify minimal, highly predictive gene subsets while optimizing classification performance. Initially, AutoML alone, which uses all available 6,026 genes, yielded strong baseline models with accuracies of up to 0.9 and F1 scores of up to 0.88 on a holdout set for each antibiotic: meropenem (MNM), ciprofloxacin (CIP), tobramycin (TOB), and ceftazidime (CAZ). Although these results demonstrate high predictive accuracy, reliance on the entire transcriptome poses substantial computational and sequencing challenges, limiting routine clinical adoption.

To address the challenge of high dimensionality in transcriptomic data, we employed GA[25] to systematically identify compact gene subsets capable of accurately predicting antibiotic resistance phenotypes. The process begins with a randomly initialized population of 40-gene subsets and iteratively refines them over 300 generations per run. In each generation, candidate subsets were evaluated via support vector machines (SVMs) and logistic regression (LR), with classification performance assessed through ROC-AUC and F1-score metrics (see Methods). High-performing subsets were preferentially retained and recombined via selection, crossover, and mutation operations, ensuring continued exploration of diverse gene combinations. This process was repeated independently for 1,000 runs per antibiotic, resulting in a broad array of high-performing but largely nonoverlapping feature sets.

Rather than converging on a single fixed subset, the GA produces thousands of distinct gene combinations, each achieving strong predictive performance. We observed that certain genes were consistently selected across many independent runs, suggesting their robust association with resistance phenotypes. To construct clinically practical and interpretable models, we generated consensus gene sets by ranking all genes on the basis of their frequency of selection across GA iterations. These top-ranked genes, representing the features most repeatedly selected by the GA, were then used to train final classifiers for each antibiotic.

The average performance of these consensus-based models was comparable to or exceeded that of full-transcriptome classifiers, with test set accuracies of ~99% for MNM and CIP and ~96% for TOB and CAZ (Fig. 2A, Supplementary Data 1). These compact classifiers, typically comprising 35–40 genes per antibiotic, offer an interpretable and computationally efficient alternative to transcriptome-wide approaches.

To evaluate model robustness, we examined the accuracy distributions across 1,000 GA iterations. The MNM and TOB classifiers showed narrower variability and higher median performance, whereas the CIP and CAZ classifiers demonstrated broader, yet still high-performing, distributions (see Fig. 2B). We also ranked genes by their selection frequency and plotted the top 10 genes that were most consistently selected across antibiotics (Fig. 2C).

We further assessed how model performance varied with the number of top-ranked genes used for classification. For each antibiotic, performance plateaued after the inclusion of ~35–40 genes, justifying the selection of the number of features (Fig. 2D–G). This plateau was observed for both accuracy and the macro average F1 score, underscoring that a limited number of frequently selected genes are sufficient to capture the transcriptomic signature of resistance.

To further enhance interpretability, we filtered iteration-specific subsets to retain ≥80% of the annotated genes. This ensured that the selected signatures remained both biologically informative and clinically actionable. The GA framework thus provided two complementary outputs: (1) iteration-specific subsets for assessing variability and classification performance and (2) consensus-ranked gene lists for biological interpretation and downstream regulatory analysis (see Methods).

Overall, our integrated GA–AutoML approach identified minimal yet highly predictive transcriptomic signatures across multiple antibiotics. The repeated selection of previously unannotated genes highlights underexplored regulatory or metabolic features of resistance, offering avenues for future biological investigation.

GA-derived subsets reveal limited CARD overlap and highlight uncharacterized genes

To determine whether our GA-AutoML-derived genes correspond to known antibiotic resistance markers or represent previously unrecognized

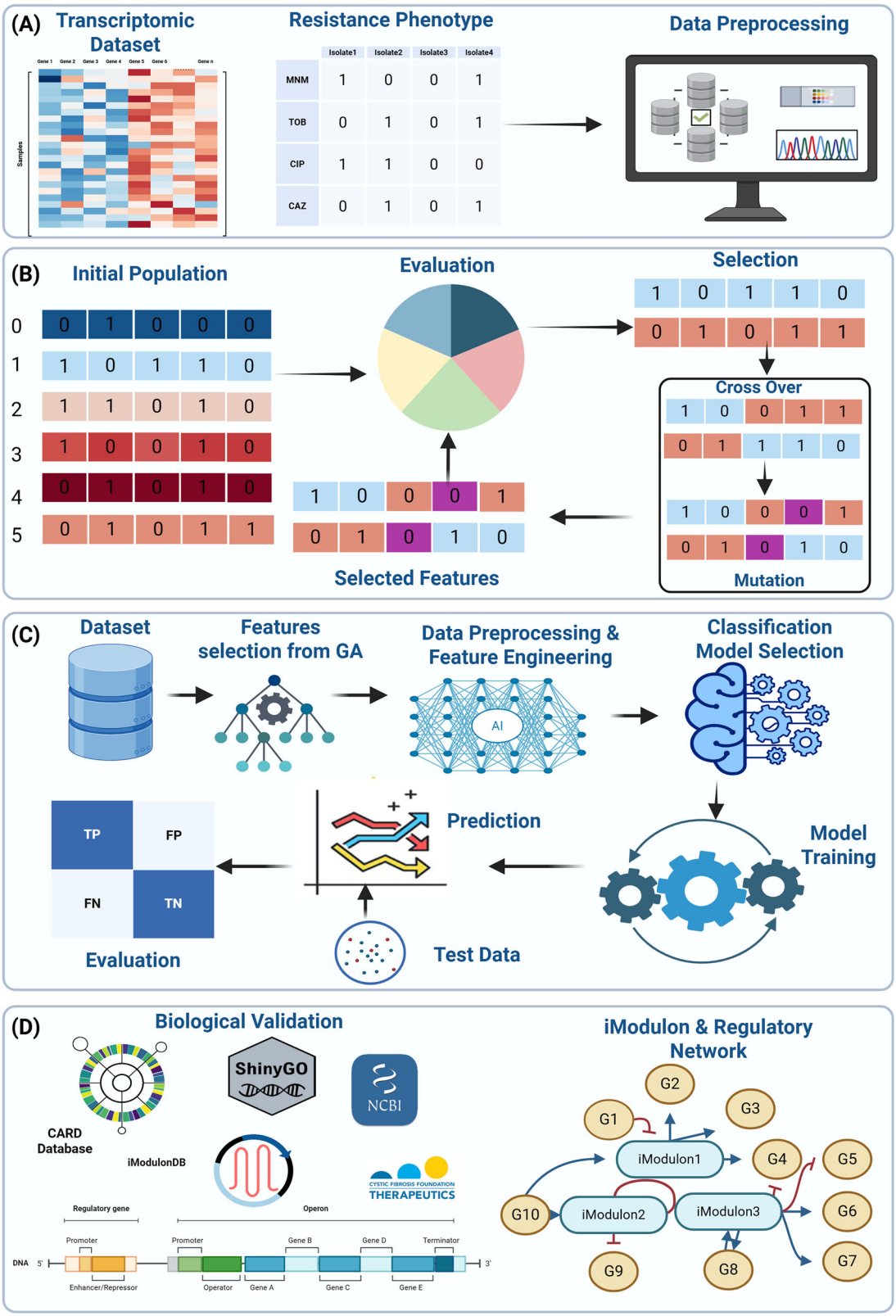

Fig. 1 | Overall pipeline for the GA–AutoML workflow and biological validation in *P. aeruginosa*. **A** Incorporation of RNA-seq data from 414 clinical isolates alongside antibiotic phenotypes and initial data processing. **B** Application of a genetic algorithm that iteratively identifies minimal, high-accuracy gene subsets through population initialization, selection, crossover, and mutation steps. **C** Integration of these GA-selected features into data preprocessing, classification, model selection, and final performance evaluation under an automated machine-learning pipeline. **D** Post hoc validation of selected genes through multiple external resources (CARD, Pseudomonas Genome DB, ShinyGO, NCBI, iModulonDB), with a focus on their operon assignments, potential regulatory networks, and broader functional context[19,20,60,63,64]. Created in BioRender. Saha, R. (2025) https://BioRender.com/0e3kj5x.

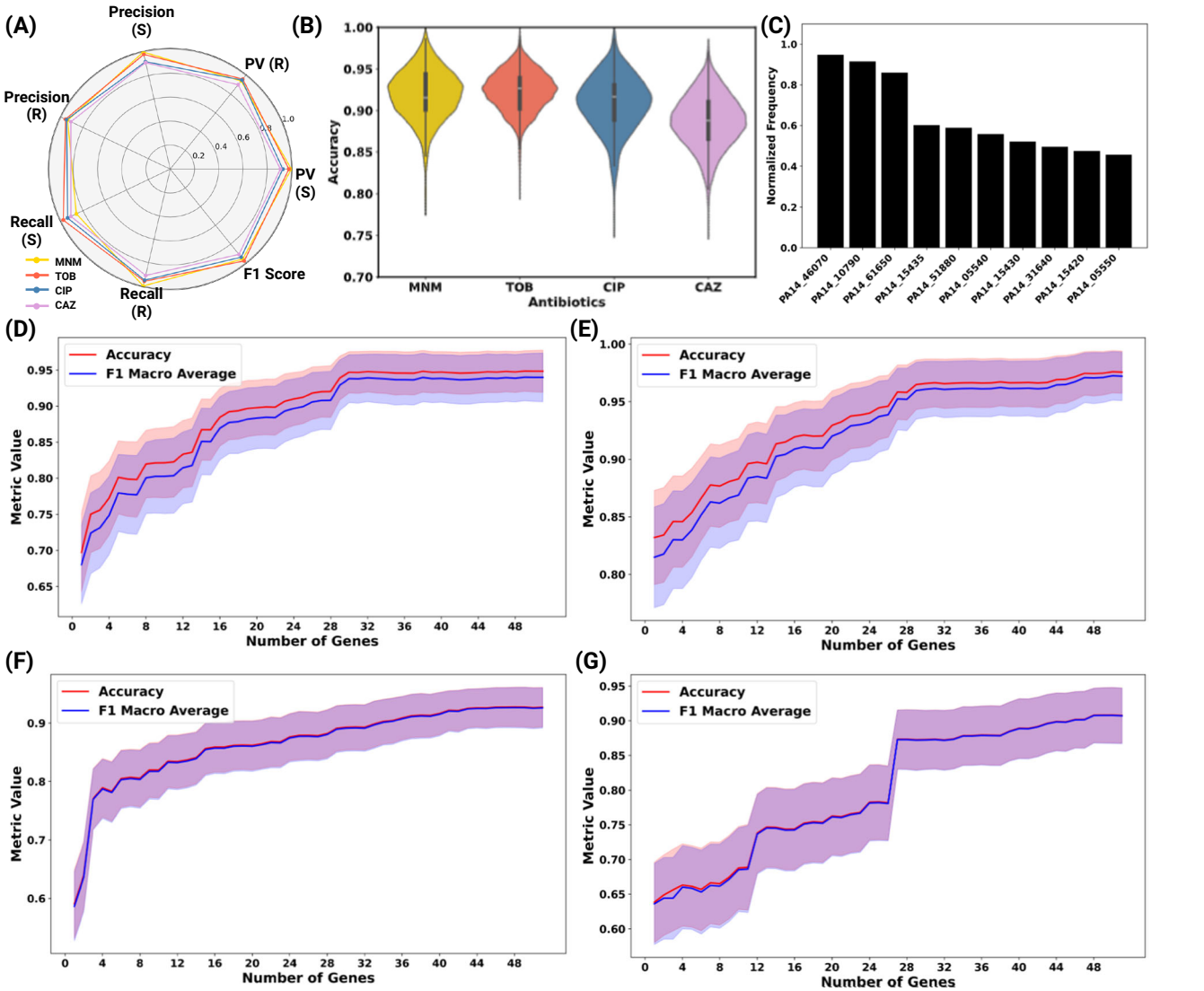

Fig. 2 | Classification performance and feature set optimization across four antibiotics. A Radar chart displaying precision, recall, predictive value (PV), and F1 score (0–1 range) for susceptible (S) and resistant (R) classes in the top-performing feature sets for each antibiotic. **B** Violin plots illustrating the accuracy distribution across 1000 genetic algorithm iterations, with median accuracy (dotted line), interquartile range (shaded area), and density. MNM and CAZ exhibit slightly broader variability, whereas TOB and CIP demonstrate narrower distributions around high accuracies. **C** Representation of the top 10 genes that were most enriched in GA across all antibiotics. **D**–**G** Classifier performance (accuracy and F1 macro average) as a function of the number of most frequently selected genes used for model training, shown separately for MNM (**D**), TOB (**E**), CIP (**F**), and CAZ (**G**). Lines represent the mean performance across cross-validation runs; shaded areas denote the standard deviation. The performance improvements plateau at ~35–40 genes, supporting the use of compact consensus feature sets for accurate and stable prediction of resistance phenotypes.

determinants, we compared the top-performing antibiotic-specific gene subsets to genes annotated in the Comprehensive Antibiotic Resistance Database (CARD)[19]. Across all antibiotics, this analysis revealed that only a minor fraction (2–10%) of our predictive gene subsets overlapped with the established AMR genes listed in the CARD (Supplementary Data 2), indicating substantial novelty within our gene signatures (Fig. 3).

For MNM, the overlap with CARD-annotated genes was approximately 3–5%, involving the efflux pump-related loci *mexA* and *mexB*, both of which were selected in >50% of GA iterations. These genes are well documented contributors to resistance via antibiotic efflux[26,27]. Interestingly, the most consistently selected gene by the GA, *gbuA*, chosen in approximately 95% of the iterations, is absent from the CARD. *GbuA* encodes guanidinobutyrase, an enzyme involved in arginine catabolism[28]. The upregulation of *gbuA* in meropenem-resistant isolates may serve as a compensatory biomarker for impaired arginine uptake linked to loss-of-function mutations in *OprD*, a porin gene critical for carbapenem entry[29]. While *gbuA* itself is not directly involved in resistance mechanisms, its strong association with resistant phenotypes highlights its utility as a diagnostic marker.

In predicting TOB resistance, only a limited overlap (2–5%) was found between GA-derived genes and CARD annotations. The gene PA14_15435, encoding a cyclic di-GMP-specific phosphodiesterase, was prominently selected (~80% iterations) but was not annotated in the CARD. Its activity may influence biofilm formation, motility, and antibiotic penetration, indirectly affecting aminoglycoside susceptibility[30]. Other highly selected genes (PA14_15430, PA14_15420, and PA14_15410) correspond to transposon-associated elements linked to a genomic island carrying mercury resistance genes[31]. Although these events do not directly confer resistance, their repeated selection indicates that the genomic context or co-selection events might mark isolates adapted to aminoglycoside stress. Known aminoglycoside resistance genes, such as *amrB* (PA14_60860), were less frequently selected (~30%), suggesting that despite their direct

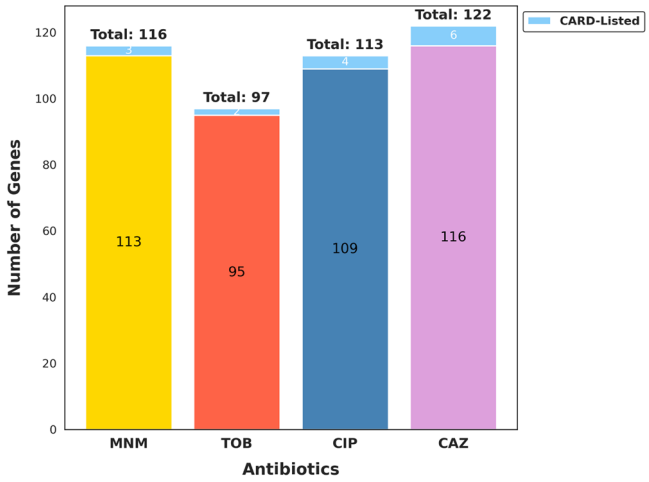

Fig. 3 | GA-derived subsets reveal minimal overlap and predominantly novel AMR genes. Relative to the CARD. Stacked bar plot comparing the number of GA-selected genes overlapping with the CARD database versus those absent from the CARD database across the four antibiotics: MNM, TOB, CIP, and CAZ. For each antibiotic, the total number of unique genes identified from high-performing GA-derived subsets is partitioned into CARD-listed and novel categories. The small size of the known segment in most cases underscores the substantial novelty captured by our approach.

resistance roles, they present weaker transcriptomic signatures distinguishing resistant strains.

For CIP, the genes most frequently selected by the GA were predominantly absent from CARD annotations as well. PA14_61650 (*pagL*, >70% iterations), encoding lipid A 3-O-deacylase, modifies the LPS structure, potentially influencing antibiotic permeability[32]. PA14_31640 (glyoxalase I homolog, ~40% iterations) may mitigate the oxidative stress induced by ciprofloxacin treatment. PA14_36990 encodes another cyclic di-GMP phosphodiesterase that influences biofilm dynamics and possibly efflux mechanisms, indirectly affecting fluoroquinolone susceptibility[30]. Moreover, genes known for quinolone resistance (e.g., *mexA, mexB, and gyrA*), identified in the CARD, were infrequently selected (10–20%), likely due to subtler transcriptomic differences.

In CAZ resistance prediction, significant overlap with CARD was observed due to the frequent selection (>90%) of ampC (PA14_10790), a known β-lactamase[33]. However, PA14_33680 (*fpvA*, >40% iterations), encoding the ferripyoverdine receptor, was repeatedly selected despite not being in the CARD. Elevated *fpvA* expression has been correlated with ceftazidime resistance, possibly linking iron acquisition responses with antibiotic resistance phenotypes[34].

Overall, our analysis revealed that GA-selected genes frequently fall outside conventional AMR annotations, likely because they present pronounced transcriptomic differences between resistant and susceptible isolates rather than directly conferring resistance. Many of these genes play indirect roles, such as metabolic compensation, membrane modification, stress response, or biofilm regulation. These results suggest that many resistance-associated transcriptomic signatures fall outside current AMR annotations, suggesting that conventional databases may not capture the full landscape of expression-based adaptations observed in resistant strains. While not direct resistance determinants, such genes may serve as robust biomarkers or proxies for resistant phenotypes, meriting further experimental characterization.

Operon-level analysis reveals limited functional overlap in predictive gene sets

In addition to genes lacking prior associations with resistance mechanisms, our GA-AutoML pipeline generated multiple gene subsets per antibiotic, each containing approximately 35-40 genes, as predictive models. These subsets achieved high predictive accuracy (~96–99%) despite minimal gene overlap (~5–8%) within each antibiotic category (Fig. 4 A-D). This unexpected observation prompted us to examine whether these subsets share common biological functions at the operon level. An operon is a group of genes in prokaryotes (bacteria) that are transcribed together as a single mRNA molecule[35]. By analyzing which operons the selected genes belonged to, we asked whether different subsets might still capture similar biological responses, even if the specific genes varied.

Several antibiotic-specific operons were indeed consistently represented across multiple predictive gene subsets (Supplementary Data 3). For MNM, two operons consistently stand out: the well-characterized mexAB-oprM operon, encoding a multidrug efflux pump system directly involved in antibiotic efflux[27], and the gbuA operon, encoding guanidinobutyrase and related enzymes (Fig. 5A). The frequent identification of mexAB-oprM aligns with its established role in direct antibiotic resistance, particularly through reducing intracellular meropenem concentrations[36]. The recurrent selection of the gbuA operon suggests an association with broader metabolic adaptations, potentially compensating for nutrient uptake disruptions such as those resulting from impaired porin function (e.g., OprD deficiency), which is a known determinant of carbapenem resistance[37,38].

For TOB, operon-level recurrence was more limited, highlighting diverse, isolate-specific strategies potentially reflecting antibiotic exposure history or genomic context. The aminoglycoside efflux operon amrAB was identified in two subsets, which is consistent with its direct role in antibiotic expulsion. Interestingly, the mercury resistance operon (mer) and alginate biosynthesis (alg) operon each appeared in single subsets (Fig. 5B). Although mercury resistance (mer operon, e.g., gene PA14_15435) is not directly linked to antibiotic resistance, its selection could indicate a broader cellular stress response or biofilm adaptation influencing aminoglycoside penetration or susceptibility. Similarly, the selection of the alg operon may reflect the known association between biofilm formation and reduced antibiotic susceptibility[39].

In the prediction of CIP resistance, the efflux operon mexAB-oprM appeared frequently, underscoring its well-documented role in fluoroquinolone efflux[40] (Fig. 5C). Additionally, the pel operon (biofilm polysaccharide production), cupA operon (fimbrial adhesion), and ribosomal operons emerged recurrently, pointing to additional transcriptomic correlates of antibiotic response, such as biofilm formation and stress-induced translational adaptation, potentially mitigating ciprofloxacin-induced DNA damage[41]. The diverse representation of operons involved in biofilm formation (pel) and adhesion (cupA) highlights broader physiological adaptations affecting antibiotic permeability and tolerance[42,43].

Moreover, CAZ exhibited a distinct operon signature heavily dominated by the β-lactamase-encoding ampC-ampR operon[44], which was selected across all subsets (Fig. 5D). This was consistent with its direct enzymatic role in β-lactam hydrolysis. Secondary selection included iron scavenging (pvd operon) and anaerobic respiration (nar operon) operations[45], reflecting metabolic and nutritional adaptations potentially associated with resistance phenotypes, particularly under nutrient-limited infection conditions where antibiotic tolerance may increase.

Critically, however, despite the consistent selection of certain operons, most predictive genes (approximately 70–85% across antibiotics) did not map to these recurrent operons. Instead, these genes were broadly dispersed throughout the genome, spanning various unrelated functional groups and metabolic pathways. This extensive dispersion suggests that antibiotic resistance-associated transcriptomic signatures capture generalized transcriptional responses rather than being strictly limited to specific operon-based regulatory adaptations.

The broad genomic distribution of predictive genes further emphasizes that antibiotic resistance involves not only direct antibiotic neutralization (e.g., efflux pumps, antibiotic-modifying enzymes) but also indirect responses, including metabolic shifts, stress responses, biofilm dynamics, and cell-envelope modifications. These diverse transcriptomic adaptations reflect the multilayered complexity underlying antibiotic resistance in *P. aeruginosa*.

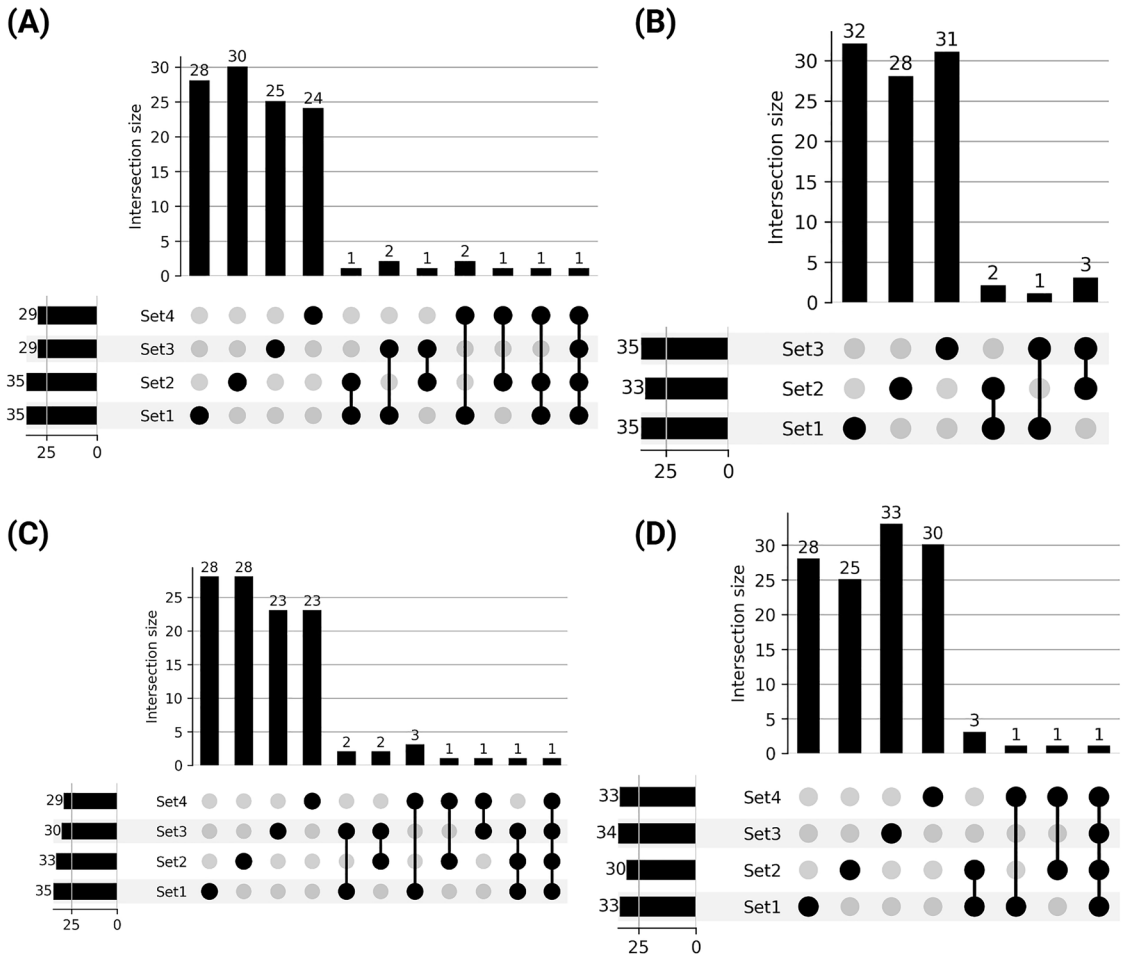

Fig. 4 | Minimal overlap among multiple high-accuracy GA-selected gene subsets. UpSet plots illustrating the pairwise and higher-order intersections among the top GA-selected gene subsets for MNM (**A**), TOB (**B**), CIP (**C**), and CAZ (**D**). Each UpSet plot underscores the limited overlap across these equally high-accuracy subsets, highlighting distinct routes to resistance.

iModulon mapping reveals regulatory convergence and strain-specific adaptations

To further explore the regulatory basis of these resistance signatures, we examined whether these diverse operons may be orchestrated by shared regulatory modules, as identified through iModulon mapping. First, we mapped the feature sets to published iModulons of *P. aeruginosa* PAO1, which represent coregulated gene groups identified through ICA[24]. Approximately 40–60% of the feature set genes consistently mapped to known iModulons, forming a clear regulatory framework. Conversely, 30–50% remained unmatched, with 12% of these corresponding to PA14-specific genes absent in PAO1. These results underscore the regulatory complexity and strain-specific adaptations underlying AMR mechanisms.

Mapping the feature sets to PAO1 iModulons revealed that distinct gene subsets converged on shared regulatory modules associated with critical cellular processes (Supplementary Fig 2, Supplementary Data 4). Several feature sets mapped to iModulons involved in oxidative stress adaptation, a common bacterial response to antibiotic-induced damage. These iModulons include genes encoding catalases, superoxide dismutases, and other antioxidants, which help mitigate the effects of the reactive oxygen species generated by antibiotics[46]. Similarly, multiple feature sets, such as the MexAB-OprM and MexCD-OprJ operons, were enriched in iModulons associated with efflux pump systems. These systems are known to confer resistance by expelling antibiotics from the cell, reducing intracellular drug concentrations to sublethal levels (Lorusso et al., 2022).

The feature sets for TOB and other antibiotics frequently map to iModulons regulating ribosomal proteins and the translation machinery, reflecting the role of protein synthesis inhibition in antibiotic action[47]. For CIP, the feature sets consistently mapped to iModulons involved in DNA repair and recombination, aligning with the mechanism of action of fluoroquinolones, which target DNA gyrase and topoisomerase IV. These iModulons include genes encoding RecA and other DNA repair proteins, which are critical for surviving CIP-induced DNA damage.

To investigate strain-specific regulation, we performed a complementary ICA using 414 transcriptomes of clinical PA14 isolates, yielding a PA14-specific iModulon set (Supplementary Data 5). Mapping the GA-selected gene subsets onto this PA14 compendium resulted in high coverage for each antibiotic. For MNM, the four predictive subsets exhibited 83–88% alignment with PA14 iModulons (Fig. 6A), predominantly mapping to modules related to metabolism and nutrient adaptation, the general stress response, efflux mechanisms, and cell envelope integrity. The CIP-associated subsets demonstrated 81–83% mapping (Fig. 6B), with prominent allocation to metabolic adaptation, DNA repair, and general stress response iModulons. In the CAZ subset, 80–82% of the genes mapped primarily to modules associated with metabolic pathways, cell envelope remodeling, and stress responses (Fig. 6C). Finally, the TOB subsets exhibited 79–81% mapping (Fig. 6D) and were most strongly associated with metabolic regulation, ribosomal functions, and stress adaptation. Across all antibiotics, the "Metabolism and Nutrient Adaptation" category consistently accounted for the largest fraction of mapped genes (Fig. 6A–D), reflecting the broad impact of antibiotic exposure on global metabolic rewiring beyond direct drug-target interactions.

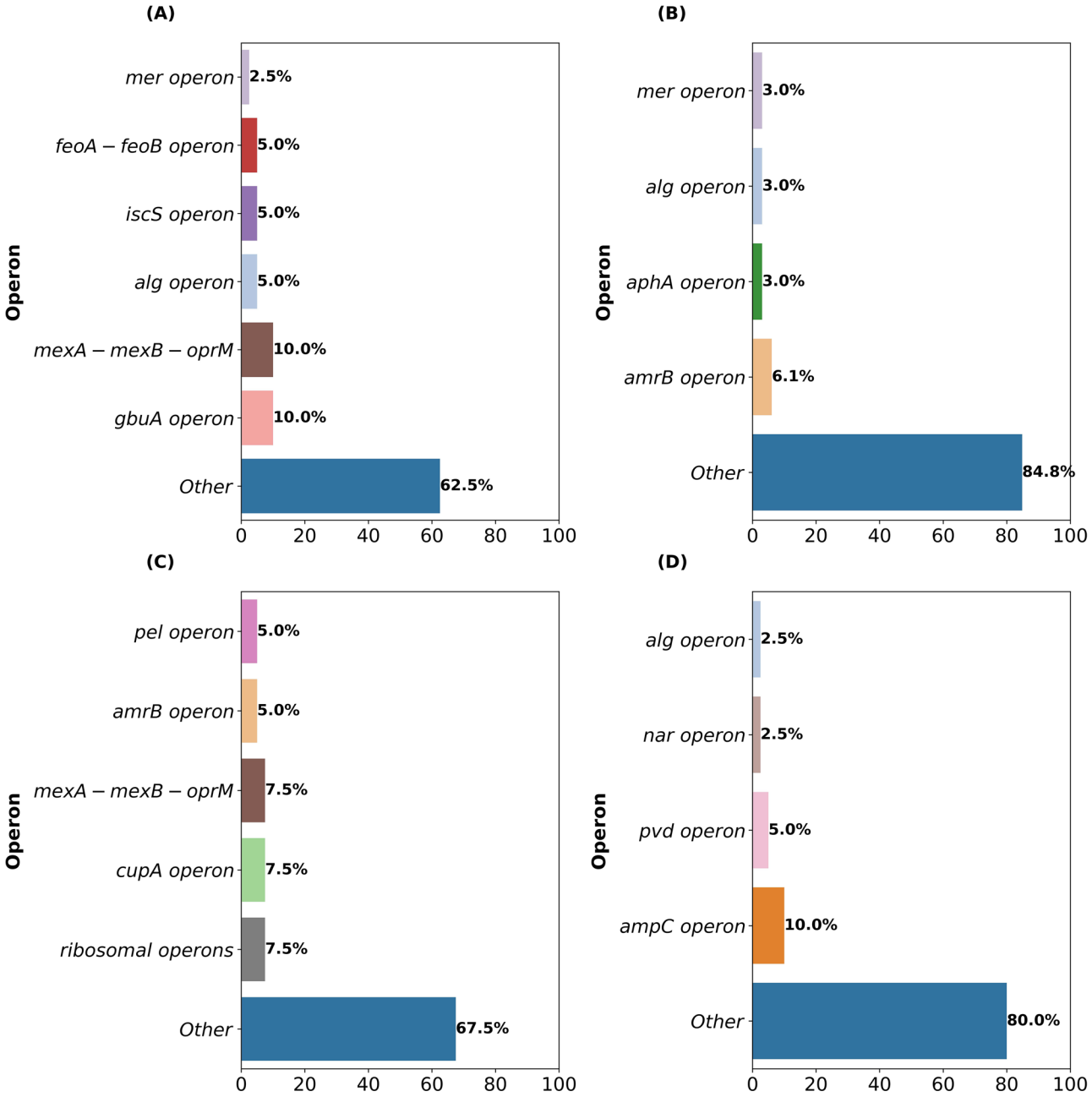

Fig. 5 | Recurrent and dispersed operon involvement across GA-predictive gene sets. Bar plots showing how often genes from specific operons were selected across multiple high-performing subsets for each antibiotic: MNM (**A**) TOB (**B**) CIP (**C**) and CAZ **D** While known resistance operons such as *mexAB-oprM* and *ampC* were repeatedly selected, most predictive genes were derived from diverse operons, highlighting that resistance-associated transcriptomic signatures involve diverse genetic loci beyond classical resistance operons.

These PA14-based findings mirror the regulatory themes observed with PAO1 analysis, reinforcing the recurrent involvement of SOS-mediated repair, efflux systems, oxidative stress responses, ribosomal regulation, and envelope remodeling in antibiotic resistance across strains. Notably, several PA14-specific iModulons lacked direct counterparts in PAO1, including modules linked to heavy-metal detoxification and type VI secretion systems, highlighting strain-specific regulatory architectures potentially associated with multifactorial resistance strategies.

Discussion

In this study, we demonstrated a successful automated approach for antibiotic susceptibility profiling in *P. aeruginosa* using only transcriptomic data. By applying a genetic algorithm to select informative features and an AutoML framework to train classifiers, we identified minimal gene expression signatures that achieved accuracies approaching 99% on a holdout set. Notably, this level of performance is comparable to that of existing genome-based approaches but relies on a markedly smaller set of transcriptomic features[12,13,48,49]. It is worth noting that genotype-based AMR predictors often achieve high accuracy with very few genomic markers. For instance, Drouin et al. (2019)[50] developed interpretable classifiers achieving ≥90% accuracy using a median of only two or three k-mer presence rules per dataset. Conversely, our transcriptomic models required 35–40 genes per antibiotic to achieve comparable or higher accuracy. This difference reflects the broader biological information captured by gene expression data, encompassing condition-dependent regulatory and metabolic responses to antibiotic exposure beyond static gene presence or absence. Indeed, Khaledi et al. (2020)[48] reported significantly lower predictive performance for GPA-only models trained on the same dataset (sensitivities 0.53–0.88; PPVs

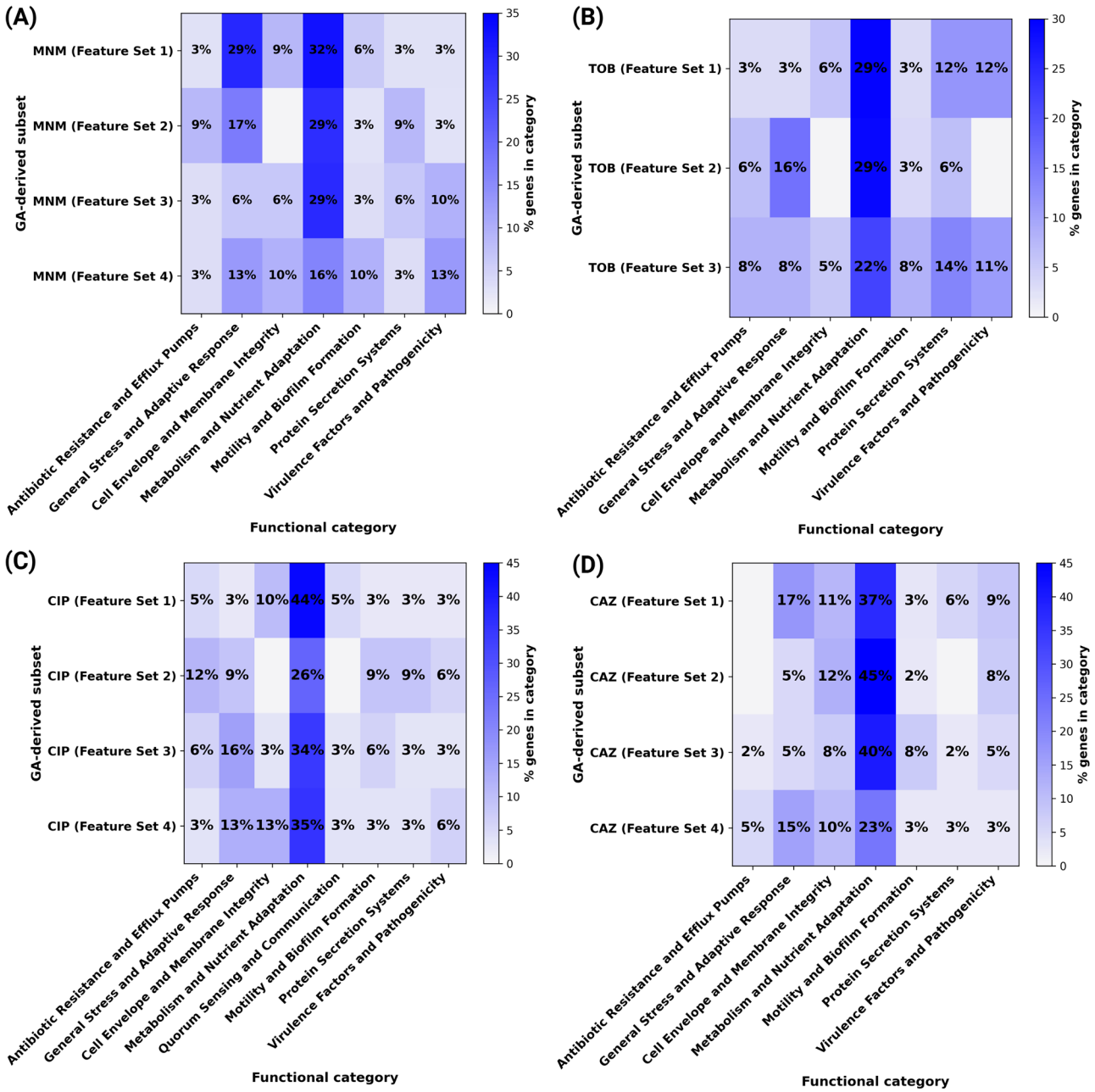

Fig. 6 | Functional allocation of GA-selected genes within PA14 iModulons. Heatmaps display the percentage of genes in each high-accuracy feature set (rows) that fall into specific PA14 iModulon functional classes (columns). for MNM (**A**) TOB (**B**) CIP (**C**) and CAZ (**D**). The color scale indicates the percentage of genes assigned to individual functional categories, showing that seemingly distinct gene sets converge on shared regulatory modules such as efflux, general stress, and metabolic adaptation.

0.56–0.92), while our transcript-based models improved these metrics by 7–17 percentage. Moreover, a major limitation of sparse genotype-based rules is their reliance on known resistance loci, potentially missing newly emerging strains with mutations in novel genes. Transcriptomic signatures, however, capture the functional state of cells and thus remain informative even as resistance evolves via new or unexpected genetic changes. Consequently, transcriptomic models provide greater discriminatory power and robustness over time.

To better understand the biological underpinnings of our transcriptomic classifiers, we examined the specific genes selected by the genetic algorithm across different antibiotics. Many of the top-ranked genes have previously been implicated in resistance, lending biological credibility to our approach. For example, *gyrA*, the primary target of fluoroquinolones[51], was recurrently selected in ciprofloxacin resistance signatures. Likewise, *ampC* and *oprD*, both of which are well-established β-lactam resistance determinants[52], are frequently selected among the ceftazidime and meropenem classifiers. Multidrug efflux pump components such as *mexA*, *mexB*, and *oprM* are also commonly identified across multiple antibiotics. The recurrence of these canonical resistance markers serves as internal validation and confirms that our feature selection strategy recovers biologically meaningful predictors.

Importantly, several highly ranked genes have not been previously annotated as resistance determinants. These include loci associated with stress adaptation, cell envelope remodeling, alginate biosynthesis, and metabolic regulation. Similar to earlier findings[53,54], our results suggest that resistant strains adopt broader physiological and regulatory adaptations beyond direct drug-target modifications. These genes may not confer resistance but may act as transcriptional proxies or participate in

compensatory mechanisms associated with enhanced bacterial fitness under antibiotic stress. The repeated selection of these genes across independent GA runs highlights their potential utility as biomarkers, even if they are not mechanistically causative.

Interestingly, the gene subsets identified for each antibiotic showed minimal overlap, indicating that multiple, equally predictive transcriptomic signatures can distinguish resistant strains from susceptible strains. Each subset included a small number of genes associated with known regulatory modules, such as the oxidative stress response, SOS-mediated DNA repair, and efflux regulation modules, which represented only a fraction of each predictive set. For example, among the ciprofloxacin classifiers, a subset of genes belongs to the SOS regulon, which is consistent with the known induction of DNA damage repair pathways under fluoroquinolone treatment[55]. Similarly, components of the SoxR regulon, which regulates redox-responsive genes such as *mexGHI-opmD*[56], were observed in classifiers for multiple antibiotics. These findings align with reports that antibiotic exposure triggers global stress response programs that facilitate bacterial survival[57–59].

However, the majority of selected genes did not map onto annotated operons or iModulons, suggesting that resistance-associated expression signatures extend beyond established regulons. This broad distribution highlights the multifactorial nature of resistance and the potential involvement of under characterized or strain-specific regulatory programs. Some markers may reflect cellular states shaped by prior antibiotic exposure, mutations in upstream regulators, or metabolic rewiring in response to treatment.

Given the transcriptomic basis of our classifiers, distinguishing between genes that directly confer resistance and those that serve as indirect markers reflecting cellular adaptations or physiological responses is important. While this distinction does not diminish their diagnostic utility, it does emphasize the need for functional characterization through follow-up experiments such as gene disruption or controlled overexpression. Furthermore, while the high accuracy and minimal feature set provide an attractive clinical diagnostic framework, broader validation across diverse clinical populations and independent cohorts remains necessary to confirm the robustness and generalizability of these transcriptomic signatures.

Overall, our approach highlights the power of integrating advanced machine learning with systems biology methodologies to uncover predictive biomarkers and elucidate the broader adaptive landscapes associated with antibiotic resistance. By providing interpretable, minimal transcriptomic signatures, this strategy offers promising avenues for rapid diagnostics, personalized treatment regimens, and targeted antimicrobial stewardship.

Methods

Data retrieval and preprocessing

Gene expression and antibiotic resistance data for this study were obtained from the publicly available dataset published by Khaledi et al.[48]. In this dataset, RNA-seq reads were aligned to the *Pseudomonas aeruginosa* PA14 reference genome using Stampy (v1.0.23), and gene-level expression counts were calculated following a previously described method[12], in which uniquely mapped reads were assigned to coding sequences and scaled by both gene length and library size. To handle zero counts, a pseudocount of 1 was added prior to log transformation. The resulting expression matrix included 6026 genes across 414 clinical isolates. We used this processed matrix directly as input to our machine learning pipeline. Isolates labeled as intermediately resistant were excluded from classification tasks, and samples missing resistance labels for a given antibiotic were removed for that analysis. The expression dataset was used to construct the feature matrix (X), while the corresponding resistance classifications for the four antibiotics formed the target vector (y). These datasets enabled the training of machine learning models to predict antibiotic resistance based on transcriptomic signatures.

Gene expression data were preprocessed to align with resistance labels, ensuring consistency in sample representation. The expression values were normalized, and genes with excessive missing values were excluded. Samples with incomplete resistance phenotype annotations were removed from the analysis.

Threshold for resistance determination

Resistance and susceptibility classifications were assigned on the basis of the minimal inhibitory concentrations determined in the original study[48]. MIC values were binarized according to Clinical & Laboratory Standards Institute (CLSI) breakpoint criteria, with samples classified as resistant or susceptible to each antibiotic.

Genetic algorithm optimization for feature selection

To reduce the number of transcriptomic features systematically while maintaining predictive accuracy, a genetic algorithm-based feature selection approach was implemented. The GA iteratively refines feature subsets by mimicking the principles of natural selection, favoring gene sets that maximize predictive performance.

The GA pipeline was initialized with 1,000 randomly generated feature sets, each consisting of 40 genes. The selection of 40 genes was determined through an empirical evaluation process, where feature set sizes ranging from 16-40 genes were systematically tested via LR and SVM, with the ROC-AUC and F1 score as primary evaluation metrics. The feature sets containing 40 genes consistently provided the best trade-off between predictive accuracy and feature set interpretability, making them the optimal configuration (Fig. 2D-G). Each GA iteration refines these feature sets over 300 generations, systematically converging on the most informative subsets.

Two types of outputs were generated from the GA selection process: iteration-specific feature sets, representing the highest-performing gene subset identified in each iteration, and a ranked gene list, which quantified how frequently individual genes appeared in the top-performing subsets across all GA iterations. These outputs provided complementary insights, allowing both the selection of stable predictive markers and an assessment of the broader resistance-associated transcriptomic landscape.

To assess feature set stability, the top-ranked gene lists from two independent GA runs were compared. The overlap between the highest-scoring gene sets in early runs was limited (~30%), suggesting substantial variability in feature selection. To address this, the number of GA iterations was increased to 1000, and a ranked gene list was generated, quantifying the frequency with which specific genes appeared in the top-performing subsets across all iterations. The ranked gene list was further segmented into subsets of 20-50 genes and systematically evaluated to pinpoint the highest-performing gene sets.

A major limitation of GA-derived ranked lists is the presence of hypothetical genes, which reduces interpretability. To address this, iteration-specific feature sets were filtered to retain at least 80% of the well-characterized genes on the basis of functional annotation databases. Although this filtering was also applied to the ranked gene list, the evaluation shifted toward iteration-specific gene sets because of their superior classification performance. The final GA-selected gene subsets were validated externally and consistently maintained high precision and recall across all antibiotics (Fig. 2 A-G).

Machine learning model training and evaluation

Automated machine learning was employed to optimize model training and hyperparameter selection. The AutoML pipeline was implemented via the auto-sklearn Python library, which systematically explores a range of models and hyperparameter configurations to identify the best-performing classifier for each antibiotic-specific dataset.

Preprocessed gene expression data were split into training (80%) and test (20%) sets while maintaining a stratified distribution of resistance phenotypes. To prevent overfitting, a stratified k-fold cross-validation approach ($k = 3$) was applied during training. The AutoML classifier was configured with a runtime limit of five hours per model to ensure exhaustive search while maintaining computational feasibility. The feature sets derived from the GA were used as inputs, with model selection based on the optimization of the ROC-AUC and F1 score.

Model performance was assessed via multiple evaluation metrics to ensure robustness. The ROC-AUC was used to measure the ability of the model to distinguish between resistant and susceptible strains, whereas the F1 score provided a balance between precision and recall, ensuring that both classes were correctly classified. Precision quantified the proportion of correctly predicted resistant strains, whereas recall measured the fraction of actual resistant isolates correctly identified. Confusion matrix analysis was conducted to evaluate true positives, true negatives, false positives, and false negatives, providing further insights into misclassification patterns.

The final models were selected based on test set performance, and the best-performing classifiers were used for further downstream analysis.

Independent component analysis and iModulon mapping

To elucidate the regulatory architecture underlying antibiotic resistance in *Pseudomonas aeruginosa*, we applied ICA via the iModulonMiner framework[60]. ICA decomposes transcriptomic data into independent components, termed iModulons, which represent coregulated gene groups under shared regulatory control[61,62]. This approach enables the identification of transcriptional regulatory networks that drive resistance phenotypes.

We analyzed the normalized transcriptomic profiles of 414 *P. aeruginosa* PA14 clinical isolates via the FastICA algorithm implemented in the iModulonMiner pipeline (https://github.com/SBRG/iModulonMiner). The input data consisted of a centered and scaled gene expression matrix (genes × samples). Stability analysis revealed that the optimal number of components was 203. Each iModulon was characterized by two outputs: a gene weight matrix, which quantifies gene contributions, and a sample activity matrix, which represents iModulon activity across samples. Genes with absolute weights > 2.5 were considered significant contributors. The functional annotations of iModulons were assigned based on Gene Ontology (GO) terms and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways, enabling insights into their biological roles.

The gene subsets identified via the GA were mapped onto the derived iModulons to explore their regulatory basis. Distinct feature sets for different antibiotics were analyzed for their convergence on iModulons, reflecting potential regulatory mechanisms underlying resistance. Strain-specific iModulons unique to PA14 were also identified, highlighting novel regulatory adaptations. These analyses facilitated a deeper understanding of transcriptional regulation in the context of antibiotic resistance.

Data availability

The transcriptomic and phenotypic data used in this study were obtained from Khaledi et al., 2020[48] and are available via the NCBI Gene Expression Omnibus (GEO) under accession number GSE123544. Processed expression matrices and phenotype labels used in the GA-AutoML pipeline are provided in the Supplementary Data. These materials are sufficient to replicate the machine learning and validation steps described in this study. Code and scripts for data analysis, machine learning, and visualization are publicly available on GitHub at https://github.com/AAlsiyabi-Research-Group/Predicting-P.-aeruginosa-Resistance-with-Minimal-Gene-Signatures.

Code availability

Code and scripts for data analysis, machine learning, and visualization are publicly available on GitHub at https://github.com/AAlsiyabi-Research-Group/Predicting-P.-aeruginosa-Resistance-with-Minimal-Gene-Signatures.

Received: 20 May 2025; Accepted: 24 August 2025;
Published online: 02 October 2025

References

1. World Health Organization. *No Time to Wait: Securing the Future from Drug-Resistant Infections. Report to the Secretary-General of the United Nations*. https://www.who.int/publications/i/item/no-time-to-wait-securing-the-future-from-drug-resistant-infections (2019).
2. Horcajada, J. P. et al. Epidemiology and Treatment of Multidrug-Resistant and Extensively Drug-Resistant Pseudomonas aeruginosa Infections. *Clin. Microbiol Rev.* **32**, e00031–19 (2019).
3. Qin, S. et al. Pseudomonas aeruginosa: pathogenesis, virulence factors, antibiotic resistance, interaction with host, technology advances and emerging therapeutics. *Signal Transduct. Target Ther.* **7**, 199 (2022).
4. Pang, Z., Raudonis, R., Glick, B. R., Lin, T.-J. & Cheng, Z. Antibiotic resistance in Pseudomonas aeruginosa: mechanisms and alternative therapeutic strategies. *Biotechnol. Adv.* **37**, 177–192 (2019).
5. Breidenstein, E. B. M., de la Fuente-Núñez, C. & Hancock, R. E. W. Pseudomonas aeruginosa: all roads lead to resistance. *Trends Microbiol* **19**, 419–426 (2011).
6. Lister, P. D., Wolter, D. J. & Hanson, N. D. Antibacterial-resistant *Pseudomonas aeruginosa*: clinical impact and complex regulation of chromosomally encoded resistance mechanisms. *Clin. Microbiol Rev.* **22**, 582–610 (2009).
7. van Belkum, A. et al. Developmental roadmap for antimicrobial susceptibility testing systems. *Nat. Rev. Microbiol* **17**, 51–62 (2019).
8. Tacconelli, E. et al. Discovery, research, and development of new antibiotics: the WHO priority list of antibiotic-resistant bacteria and tuberculosis. *Lancet Infect. Dis.* **18**, 318–327 (2018).
9. Dyar, O. J., Huttner, B., Schouten, J. & Pulcini, C. What is antimicrobial stewardship? *Clin. Microbiol. Infect.* **23**, 793–798 (2017).
10. Lázár, V. et al. Antibiotic-resistant bacteria show widespread collateral sensitivity to antimicrobial peptides. *Nat. Microbiol* **3**, 718–731 (2018).
11. Alsiyabi, A., Shahid, S. A. & Al-Harrasi, A. An Automated Machine Learning Framework for Antimicrobial Resistance Prediction Through Transcriptomics. Preprint at https://doi.org/10.1101/2024.06.22.600223 (2024).
12. Khaledi, A. et al. Transcriptome Profiling of Antimicrobial Resistance in Pseudomonas aeruginosa. *Antimicrob. Agents Chemother.* **60**, 4722–4733 (2016).
13. Suzuki, S., Horinouchi, T. & Furusawa, C. Prediction of antibiotic resistance by gene expression profiles. *Nat. Commun.* **5**, 5792 (2014).
14. Bhattacharya, S. et al. Transcriptomic Biomarkers to Discriminate Bacterial from Nonbacterial Infection in Adults Hospitalized with Respiratory Illness. *Sci. Rep.* **7**, 6548 (2017).
15. Wang, H. et al. Paving the way for precise diagnostics of antimicrobial resistant bacteria. *Front Mol Biosci.* **9**, (2022).
16. Topçuoğlu, B. D., Lesniak, N. A., Ruffin, M. T., Wiens, J. & Schloss, P. D. A Framework for Effective Application of Machine Learning to Microbiome-Based Classification Problems. *mBio* **11**, e00434–20 (2020).
17. Libbrecht, M. W. & Noble, W. S. Machine learning applications in genetics and genomics. *Nat. Rev. Genet.* **16**, 321–332 (2015).
18. Macesic, N. et al. Predicting Phenotypic Polymyxin Resistance in Klebsiella pneumoniae through Machine Learning Analysis of Genomic Data. *mSystems* **5**, e00656–19 (2020).
19. Alcock, B. P. et al. CARD 2023: expanded curation, support for machine learning, and resistome prediction at the Comprehensive Antibiotic Resistance Database. *Nucleic Acids Res.* **51**, D690–D699 (2023).
20. Winsor, G. L. et al. Enhanced annotations and features for comparing thousands of *Pseudomonas* genomes in the Pseudomonas genome database. *Nucleic Acids Res.* **44**, D646–D653 (2016).
21. Cao, H., Ma, Q., Chen, X. & Xu, Y. DOOR: a prokaryotic operon database for genome analyses and functional inference. *Brief. Bioinform* **20**, 1568–1577 (2019).
22. Lee, T.-W. Independent Component Analysis. in *Independent Component Analysis* 27–66 (Springer US, Boston, MA, (1998). https://doi.org/10.1007/978-1-4757-2851-4_2.
23. Sastry, A. V. et al. Independent component analysis recovers consistent regulatory signals from disparate datasets. *PLoS Comput Biol.* **17**, e1008647 (2021).

24. Rajput, A. et al. Advanced transcriptomic analysis reveals the role of efflux pumps and media composition in antibiotic responses of *Pseudomonas aeruginosa*. *Nucleic Acids Res.* **50**, 9675–9688 (2022).
25. Mirjalili, S. Genetic Algorithm. in 43–55 (2019). https://doi.org/10.1007/978-3-319-93025-1_4.
26. Lorusso, A. B., Carrara, J. A., Barroso, C. D. N., Tuon, F. F. & Faoro, H. Role of Efflux Pumps on Antimicrobial Resistance in Pseudomonas aeruginosa. *Int J. Mol. Sci.* **23**, 15779 (2022).
27. Poole, K. Efflux-mediated antimicrobial resistance. *J. Antimicrobial Chemother.* **56**, 20–51 (2005).
28. Nakada, Y. & Itoh, Y. Characterization and Regulation of the *gbuA* Gene, Encoding Guanidinobutyrase in the Arginine Dehydrogenase Pathway of *Pseudomonas aeruginosa* PAO1. *J. Bacteriol.* **184**, 3377–3384 (2002).
29. Li, H., Luo, Y.-F., Williams, B. J., Blackwell, T. S. & Xie, C.-M. Structure and function of OprD protein in Pseudomonas aeruginosa: From antibiotic resistance to novel therapies. *Int. J. Med. Microbiol.* **302**, 63–68 (2012).
30. Valentini, M. & Filloux, A. Biofilms and Cyclic di-GMP (c-di-GMP) Signaling: Lessons from Pseudomonas aeruginosa and Other Bacteria. *J. Biol. Chem.* **291**, 12547–12555 (2016).
31. Partridge, S. R., Kwong, S. M., Firth, N. & Jensen, S. O. Mobile Genetic Elements Associated with Antimicrobial Resistance. *Clin Microbiol Rev* **31**, (2018).
32. Rutten, L. et al. Crystal structure and catalytic mechanism of the LPS 3- *O* -deacylase PagL from *Pseudomonas aeruginosa*. *Proc. Natl Acad. Sci.* **103**, 7071–7076 (2006).
33. Livermore, D. M. beta-Lactamases in laboratory and clinical resistance. *Clin. Microbiol Rev.* **8**, 557–584 (1995).
34. Ramsay, K. A. et al. Ceftazidime resistance in Pseudomonas aeruginosa is multigenic and complex. *PLoS One* **18**, e0285856 (2023).
35. Salgado, H., Moreno-Hagelsieb, G., Smith, T. F. & Collado-Vides, J. Operons in *Escherichia coli*: Genomic analyses and predictions. *Proc. Natl Acad. Sci.* **97**, 6652–6657 (2000).
36. Avakh, A., Grant, G. D., Cheesman, M. J., Kalkundri, T. & Hall, S. The Art of War with Pseudomonas aeruginosa: Targeting Mex Efflux Pumps Directly to Strategically Enhance Antipseudomonal Drug Efficacy. *Antibiotics (Basel)* **12**, (2023).
37. Wang, M., Zhang, Y., Pei, F., Liu, Y. & Zheng, Y. Loss of OprD function is sufficient for carbapenem-resistance-only but insufficient for multidrug resistance in Pseudomonas aeruginosa. *BMC Microbiol* **25**, 218 (2025).
38. Jagmann, N., Bleicher, V., Busche, T., Kalinowski, J. & Philipp, B. The guanidinobutyrase GbuA is essential for the alkylquinolone-regulated pyocyanin production during parasitic growth of *Pseudomonas aeruginosa* in co-culture with *Aeromonas hydrophila*. *Environ. Microbiol* **18**, 3550–3564 (2016).
39. Uruén, C., Chopo-Escuin, G., Tommassen, J., Mainar-Jaime, R. C. & Arenas, J. Biofilms as Promoters of Bacterial Antibiotic Resistance and Tolerance. *Antibiotics (Basel)* **10**, (2020).
40. Piddock, L. J. V. Multidrug-resistance efflux pumps? not just for resistance. *Nat. Rev. Microbiol* **4**, 629–636 (2006).
41. Friedman, L. & Kolter, R. Genes involved in matrix formation in *Pseudomonas aeruginosa* PA14 biofilms. *Mol. Microbiol* **51**, 675–690 (2004).
42. Trastoy, R. et al. Mechanisms of Bacterial Tolerance and Persistence in the Gastrointestinal and Respiratory Environments. *Clin Microbiol Rev.* **31**, (2018).
43. Meissner, A. et al. *Pseudomonas aeruginosa cupA* -encoded fimbriae expression is regulated by a GGDEF and EAL domain-dependent modulation of the intracellular level of cyclic diguanylate. *Environ. Microbiol* **9**, 2475–2485 (2007).
44. Glen, K. A. & Lamont, I. L. β-lactam Resistance in Pseudomonas aeruginosa: Current Status, Future Prospects. *Pathogens* **10**, 1638 (2021).
45. Punchi Hewage, A. N. D. et al. Mobilization of Iron Stored in Bacterioferritin Is Required for Metabolic Homeostasis in Pseudomonas aeruginosa. *Pathogens* **9**, 980 (2020).
46. Higazy, D., Ahmed, M. N. & Ciofu, O. The impact of antioxidant-ciprofloxacin combinations on the evolution of antibiotic resistance in Pseudomonas aeruginosa biofilms. *NPJ Biofilms Microbiomes* **10**, 156 (2024).
47. Arenz, S. & Wilson, D. N. Bacterial Protein Synthesis as a Target for Antibiotic Inhibition. *Cold Spring Harb Perspect Med.* **6**, (2016).
48. Khaledi, A. et al. Predicting antimicrobial resistance in *Pseudomonas aeruginosa* with machine learning-enabled molecular diagnostics. *EMBO Mol. Med.* **12**, e10264 (2020).
49. Su, M., Satola, S. W. & Read, T. D. Genome-Based Prediction of Bacterial Antibiotic Resistance. *J Clin Microbiol* **57**, (2019).
50. Drouin, A. et al. Interpretable genotype-to-phenotype classifiers with performance guarantees. *Sci. Rep.* **9**, 4071 (2019).
51. Akasaka, T., Tanaka, M., Yamaguchi, A. & Sato, K. Type II Topoisomerase Mutations in Fluoroquinolone-Resistant Clinical Strains of *Pseudomonas aeruginosa* Isolated in 1998 and 1999: Role of Target Enzyme in Mechanism of Fluoroquinolone Resistance. *Antimicrob. Agents Chemother.* **45**, 2263–2268 (2001).
52. Freed, S. & Hanson, N. D. AmpC induction by imipenem in Pseudomonas aeruginosa occurs in the absence of OprD and impacts imipenem/relebactam susceptibility. *Microbiol Spectr.* **12**, e0014224 (2024).
53. Torrens, G. et al. Profiling the susceptibility of Pseudomonas aeruginosa strains from acute and chronic infections to cell-wall-targeting immune proteins. *Sci. Rep.* **9**, 3575 (2019).
54. Kuper, T. J., Islam, M. M., Peirce-Cottler, S. M., Papin, J. A. & Ford, R. M. Spatial transcriptome-guided multi-scale framework connects P. aeruginosa metabolic states to oxidative stress biofilm microenvironment. *PLoS Comput Biol.* **20**, e1012031 (2024).
55. Dörr, T., Lewis, K. & Vulić, M. SOS response induces persistence to fluoroquinolones in Escherichia coli. *PLoS Genet.* **5**, e1000760 (2009).
56. Sakhtah, H. et al. The *Pseudomonas aeruginosa* efflux pump MexGHI-OpmD transports a natural phenazine that controls gene expression and biofilm development. *Proceedings of the National Academy of Sciences* **113**, (2016).
57. Dalbanjan, N. P., Kadapure, A. J. & Praveen Kumar, S. K. A comprehensive review on latent role of stress proteins in antibiotic resistance. *Microbe* **4**, 100151 (2024).
58. Dawan, J. & Ahn, J. Bacterial Stress Responses as Potential Targets in Overcoming Antibiotic Resistance. *Microorganisms* **10**, (2022).
59. Niu, H., Gu, J. & Zhang, Y. Bacterial persisters: molecular mechanisms and therapeutic development. *Signal Transduct. Target Ther.* **9**, 174 (2024).
60. Sastry, A. V. et al. iModulonMiner and PyModulon: Software for unsupervised mining of gene expression compendia. *PLoS Comput Biol.* **20**, e1012546 (2024).
61. Hyvärinen, A. & Oja, E. Independent component analysis: algorithms and applications. *Neural Netw.* **13**, 411–430 (2000).
62. Comon, P. Independent component analysis, A new concept? *Signal Process.* **36**, 287–314 (1994).
63. Ge, S. X., Jung, D., Jung, D. & Yao, R. ShinyGO: A graphical gene-set enrichment tool for animals and plants. *Bioinformatics* **36**, 2628–2629 (2020).
64. O'Leary, N. A. et al. Exploring and retrieving sequence and metadata for species across the tree of life with NCBI Datasets. *Sci. Data* **11**, 732 (2024).

Acknowledgements

This research was supported by the National Institutes of Health (NIH) R35 MIRA grant (5R35GM143009) and the Nebraska Ethanol Board Award (26-1106-0157-001) awarded to Rajib Saha. Syed Ahsan Shahid acknowledges funding support from the University of Nizwa.

Author contributions

N.S., S.A.S., and M.S. designed the study, preprocessed transcriptomic data, developed and implemented the GA–AutoML pipeline, conducted feature selection and classifier evaluation, performed biological interpretation, and wrote the manuscript. A.A.S. and R.S. supervised the study, provided conceptual guidance on methodology and biological validation, secured funding, and critically revised the manuscript. All authors reviewed and approved the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41540-025-00584-0.

Correspondence and requests for materials should be addressed to Adil Al-Siyabi or Rajib Saha.