PANGENSA: A GRAPH-CONSTRAINED MACHINE LEARNING FRAMEWORK FOR IDENTIFYING ANTIBIOTIC RESISTANCE DETERMINANTS IN BACTERIAL PANGENOMES

Tamsin James *
School of Computer Science
University of Birmingham, UK

Jordan Meadows
Independent Researcher, UK

Andreas Karwath
Department of Cancer and Genomic Sciences
University of Birmingham, UK

Nicole Wheeler & Peter Tino
School of Computer Science
University of Birmingham, UK

ABSTRACT

Antimicrobial resistance is a major global health threat, driving the need for genome-based diagnostics. While bacterial genome-wide association studies aim to identify causal resistance determinants, they are often impaired by the high multicollinearity and underdetermined regimes inherent in pangenomes. Standard machine learning approaches often fail to satisfy the requirement that fine-mapping should yield localised genomic loci rather than correlation-determined groupings.

We introduce PANGENSA, a graph-constrained machine learning framework that uses the topology of compacted de Bruijn graphs as a label-agnostic structural prior. PANGENSA partitions the pangenome graph into discrete communities and trains independent models on each, ensuring that subproblems are not overparameterized and that results are spatially localised by construction.

We demonstrate that these communities exhibit high internal connectivity and low inter-community leakage. Top-ranking community-level classifiers achieved high AUROC ($\geq$ 0.86) across 14 antibiotic phenotypes. Notably, PANGENSA localised known resistance signals for 11 of 14 antibiotics and recovered low-prevalence mechanisms that global baseline methods failed to detect. This demonstrates that encoding genomic locality as a structural prior can effectively amplify under-represented causal signals while controlling for population structure.

1 INTRODUCTION

Antimicrobial resistance (AMR) is recognised as a major global health threat, (on Antimicrobial Resistance, 2019), and culture-based antibiotic susceptibility testing (AST) delays treatment decisions (Irwin et al., 2021). Bacterial genome-wide association studies (bGWAS) seek to identify causal resistance determinants. Whereas earlier methods relied on univariate tests (Chen & Shapiro, 2015; Lees et al., 2016; Earle et al., 2016), contemporary approaches adopt multivariate genotype-to-phenotype (GP) mapping frameworks (Lees et al., 2020; Saber & Shapiro, 2020; Kim et al., 2022)

A convenient view is to decompose ML-bGWAS into the following stages: (1) representation, (2) lineage-aware evaluation splits, (3) genotype-to-phenotype (GP) prediction, and (4) candidate selection/fine-mapping.

Effective causal discovery via statistical learning methods requires three properties. First **(R1)**, the preprocessing stage (Stages 1–2) must preserve and encode the core genetic signals underlying phenotypic variation (Wu et al., 2025; Wang & Huang, 2022). Second **(R2)**, the modelling stage should

*Corresponding author: txj287@student.bham.ac.uk

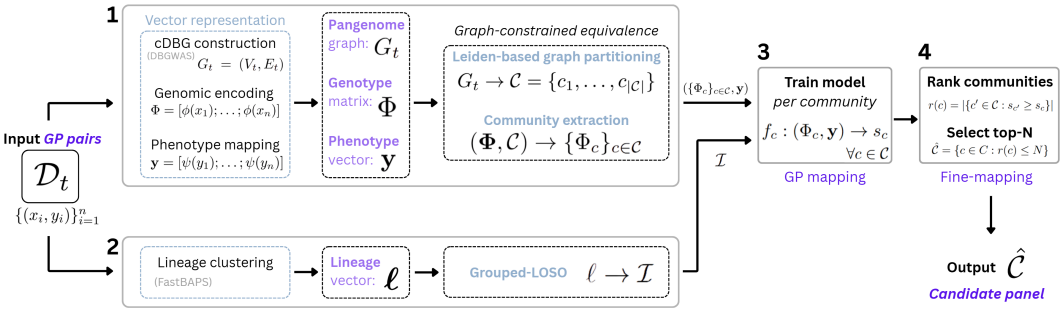

Figure 1: PANGENSA workflow. Stage 1 partitions the cDBG into communities; Stage 2 splits the data by lineage structures; Stage 3 fits independent classifiers per community under grouped, lineage-aware CV; Stage 4 selects candidates via sparsity-penalised ranking.

be well-posed, in alignment with the Hadamard criterion (Hadamard, 2014). Third **(R3)**, the candidate selection/fine-mapping stage must yield equivalence classes of genomic markers corresponding to unique genomic loci rather than correlation-determined groupings.

Pangenomes make these requirements difficult to satisfy. Marker sets contain 10^5–10^7 features while sample sizes are typically 10^2–10^4 (Lees et al., 2020; Mosquera-Rendón et al., 2023), creating a severely underdetermined regime. Genome-wide linkage disequilibrium (LD) induces high multicollinearity, so multiple parameter configurations achieve equivalent predictive performance, rendering unconstrained GP mapping ill-posed (James et al., 2025; Saber & Shapiro, 2020; Collins & Didelot, 2018).

Structural priors can break this non-uniqueness by constraining the hypothesis space (Bach et al., 2012; Hazimeh et al., 2023). In bGWAS, features are often defined at gene/allele resolution (Hyun et al., 2020; Kavvas et al., 2018) or from de Bruijn graph (cDBG) substructures (Jaillard et al., 2018). However, these approaches pool features into a single global design matrix, discarding topological metadata before GP mapping. Consequently, fine-mapping returns correlation-determined equivalence classes spanning the entire genome (violating R3), irrespective of the input encoding used.

We address LD-induced non-identifiability via PANGENSA, a framework that partitions the pangenome cDBG into communities and reformulates the global bGWAS optimisation as independent, community-level subproblems. This guarantees that: (i) equivalence classes are contiguous genomic neighbourhoods by construction (satisfying R3); (ii) each subproblem operates on $|\mathcal{M}_c| < n$ markers (satisfying R2); and (iii) community-level predictive performance directly scores genomic loci rather than individual correlated markers. We demonstrate that this formulation recovers known resistance determinants, including low-prevalence mechanisms invisible to global methods, across 14 antibiotic resistance phenotypes in *Staphylococcus aureus*.

2 Problem Formulation

We formulate bGWAS as a structured inference task (Fig. 1), following the representation framework of James et al. (2025).

Inputs. We begin with an empirical dataset $\mathcal{D}_t = \{(x_i, y_i)\}_{i=1}^n$ of n observed genotype–phenotype pairs, where x_i denotes the whole-genome sequence and y_i denotes the raw phenotype measurement for isolate i. The subscript t indicates that samples are drawn from the distribution P_t at sampling time t, where samples from P_t are non-i.i.d. due to shared evolutionary history and population structure.

PANGENSA operates on four structures derived from $\mathcal{D}_t$. First, a cDBG $G_t = (V_t, E_t)$, constructed from the genomic sequences $\{x_i\}_{i=1}^n$ via DBGWAS (Jaillard et al., 2018). Nodes correspond to unitigs (maximal non-branching paths of overlapping k-mers in the DBG) and edges connect unitigs that are adjacent in at least one assembled genome. Each unitig induces a binary presence/absence marker, and we write $|\mathcal{M}| = |V_t|$ markers, establishing a bijection $V_t \cong \mathcal{M}$ between graph nodes

and markers. The second is a binary genotype matrix $\Phi \in \{0,1\}^{n\times|\mathcal{M}|}$, where entry $\Phi_{ij} = 1$ if unitig $m_j \in \mathcal{M}$ is present in genome x_i. The third is a binary phenotype vector $\mathbf{y} \in \{0,1\}^n$, obtained by binarising raw phenotype measurements according to clinical breakpoints. The fourth is a lineage assignment vector $\boldsymbol{\ell} = (\ell_1, \cdots, \ell_n)^\intercal$ assigning each isolate to a lineage cluster, derived via FastBAPS (Tonkin-Hill et al., 2019).

We now introduce notation for two operations that jointly determine the input to a GP mapping model.

A feature selection step chooses a subset $\mathcal{S} \subseteq \mathcal{M}$ of markers to retain, yielding the submatrix $\Phi_{\mathcal{S}} \in \{0,1\}^{n\times|\mathcal{S}|}$ restricted to the columns corresponding to $\mathcal{S}$. A collapsing step then groups markers in $\mathcal{S}$ that share identical presence/absence patterns across all n isolates into equivalence classes $\mathcal{E}_{\mathcal{S}} = \{e_1, \ldots, e_{|\mathcal{E}_{\mathcal{S}}|}\}$, yielding a collapsed matrix $\tilde{\Phi}_{\mathcal{S}} \in \{0,1\}^{n\times|\mathcal{E}_{\mathcal{S}}|}$ with one representative column per class. The effective dimensionality of the input is $|\mathcal{E}_{\mathcal{S}}|$.

Causal inference objective. The goal of bGWAS is to identify causal markers $\mathcal{M}_\Theta \subseteq \mathcal{M}$ with direct mechanistic effect on the phenotype, where $|\mathcal{M}_\Theta| \ll \mathcal{M}$ (Lees et al., 2020). Non-causal markers $\mathcal{M} \setminus \mathcal{M}_\Theta$ may correlate with the phenotype through linkage-disequilibrium (LD) or population structure. As the input encoding induces equivalence classes $\mathcal{E}$ over $\mathcal{M}$ (sets of statistically indistinguishable markers), the practical target becomes to identify a set of equivalence classes $\hat{\mathcal{E}} \subseteq \mathcal{E}$ with candidate markers $\hat{\mathcal{M}} = \bigsqcup_{e\in\hat{\mathcal{E}}} \mathcal{M}_e$.

Conventional pipelines choose a global feature set $\mathcal{S}$, pool all retained markers into $\tilde{\Phi}_{\mathcal{S}}$, train a single model on $(\tilde{\Phi}_{\mathcal{S}}, \mathbf{y})$, and derive scores from model parameters. In bacterial populations, genome-wide LD correlates physically distant markers (Saber & Shapiro, 2020; Earle et al., 2016; San et al., 2020), causing $\mathcal{E}_{\mathcal{S}}$ to span the genome rather than correspond to localised loci; fine-mapping on such classes cannot resolve causal effects to specific genomic regions (Schaid et al., 2018; Broekema et al., 2020). This is compounded by the underdetermined regime: $|\mathcal{M}| \gg n$ (typically 10^5–10^7 markers vs. 10^2–10^4 isolates), so even after selection and collapsing $|\mathcal{E}_{\mathcal{S}}|$ often exceeds n, and no regularisation on a global model can distinguish causal from merely correlated markers without additional structural constraints (James et al., 2025).

Graph-constrained reformulation. PANGENSA addresses the localisation problem by constraining the feature space to the topology of G_t. We partition G_t into communities $\mathcal{C} = \{c_1, \ldots, c_{|\mathcal{C}|}\}$ using the Leiden algorithm (Traag et al., 2019), where each community $c \in \mathcal{C}$ is a connected subgraph containing markers $\mathcal{M}_c \subseteq \mathcal{M}$. Since communities are connected subgraphs of the cDBG, each corresponds to a contiguous genomic neighbourhood by construction. We write $\Phi_c \equiv \Phi_{\mathcal{M}_c} \in \{0,1\}^{n\times|\mathcal{M}_c|}$ for the submatrix restricted to community c. We set $\mathcal{E} = \mathcal{C}$ such that each community is itself an equivalence class for fine-mapping, fixing the output resolution at the community scale.

The global optimisation decomposes into independent community-level subproblems. From the lineage assignments $\boldsymbol{\ell}$, we construct a cross-validation (CV) index partition $\mathcal{I}$ via an adapted grouped leave-one-strain-out (grouped-LOSO) protocol (Lees et al., 2020) (see Appendices A and D for code and implementation details). For each community $c \in \mathcal{C}$, an independent classifier is trained on Φ_c and the shared phenotype vector $\mathbf{y}$:

$$F_c^* = \arg\min_{F_c\in\mathcal{F}} \sum_{i\in\mathcal{I}_{\text{train}}} \mathcal{L}\big(F_c(\Phi_c[i,:]),\, y_i\big), \tag{1}$$

where $\mathcal{F}$ is the hypothesis class and $\mathcal{L}$ is the loss function (see Sec. 3 for instantiation). Each community receives a predictive score averaged over folds:

$$s_c = \mathbb{E}_{\mathcal{I}}[\text{AUROC}(F_c^*,\, \Phi_c,\, \mathbf{y})]\,. \tag{2}$$

Because each model sees only Φ_c, genome-wide correlations between communities cannot confound local signals, and because fine-mapping operates on s_c rather than on marker-level parameters, the output resolution is set by graph topology rather than by genome-wide correlation structure. Provided $|\mathcal{M}_c| < n$ for all c, each subproblem is also better-conditioned than the global problem.

A candidate panel $\hat{\mathcal{C}}$ is selected to maximise cumulative signal subject to a sparsity penalty:

$$\hat{\mathcal{C}} = \arg\max_{\mathcal{S}\subseteq\mathcal{C}} \sum_{c\in\mathcal{S}} s_c \; - \; \lambda\,|\mathcal{S}|, \tag{3}$$

where $\lambda > 0$ controls the trade-off between informativeness and compactness. Because scores are computed independently, Eq. 3 is separable: the optimal panel is $\hat{\mathcal{C}} = \{c \in \mathcal{C} : s_c > \lambda\}$. Rather than setting λ directly, we determine the threshold from the score distribution via a rank-gap procedure.

Panel selection. To operationalise Eq. 3, we rank communities by score $s_{(1)} \geq s_{(2)} \geq \cdots \geq s_{(|\mathcal{C}|)}$ and identify where signal transitions to noise. Let $\Delta_r = s_{(r)} - s_{(r+1)}$ denote the score gap at rank r. For each gap we compute a leave-one-out z-score:

$$z_r = \frac{\Delta_r - \bar{\Delta}_{-r}}{\sigma_{-r}}\,, \tag{4}$$

where $\bar{\Delta}_{-r}$ and σ_{-r} are the mean and standard deviation of all gaps excluding Δ_r. The panel size is:

$$N = \arg\max_{r}\; \frac{z_r}{\log(r+1)}\,. \tag{5}$$

The logarithmic denominator penalises larger panels, encoding the expectation that resistance mechanisms are sparse (Lees et al., 2020; Wheeler et al., 2019). If no gap has a positive penalised z-score, N defaults to 1. The candidate panel and markers are:

$$\hat{\mathcal{C}} = \{c \in \mathcal{C} : r(c) \leq N\}, \qquad \hat{\mathcal{M}} = \bigsqcup_{c\in\hat{\mathcal{C}}} \mathcal{M}_c\,. \tag{6}$$

3 Experimental setup

3.1 Dataset and graph construction

We use a published cohort of *Staphylococcus aureus* from Wheeler et al. (Wheeler et al., 2019), comprising $n = 4{,}139$ isolates drawn from national bacteraemia surveillance, a tertiary-care blood culture collection, and a regional MRSA collection, each with laboratory AST data and Illumina whole-genome sequencing. Phenotypes were encoded as binary susceptible/resistant labels for a panel of 14 antibiotics: CIP, CLI, ERY, FUS, GEN, LIN, MET, MIN, MUP, PEN, RIF, TEI, TET, and TMP (full names in Table 2). Label availability ranged from $n = 908$ to $n = 4{,}139$ per drug, with resistance prevalence varying from 0.14% (LIN) to 97.54% (PEN); class balances and label coverage are reported in Appendix B (Fig. 2).

Population structure was inferred using FastBAPS (Tonkin-Hill et al., 2019) (level-2 fine clustering), yielding 45 lineage clusters (33 clonal complexes and 12 singleton sequence types), dominated by CC22/EMRSA-15 ($n = 2{,}802$; Fig. 3 in Appendix B). A cDBG was constructed from all assemblies using DBGWAS (Jaillard et al., 2018), producing $|\mathcal{M}| = 1{,}238{,}044$ unitigs. A single giant connected component contained 1,237,935 nodes, and five minor components (sizes 2–84) were excluded. Collapsing identical presence/absence patterns reduced the effective feature count to 733,415 unique equivalence classes. Unitigs were annotated with known resistance genes by searching both strands against a curated library of gene sequences from NCBI (Brown et al., 2015).

We partitioned G_t using the Leiden algorithm(Traag et al., 2019) at a data-driven resolution $\gamma^* \approx 28.02$, selected by identifying a gap in the $\log_{10}$ community size distribution across a scan of $\gamma \in [0.1, 100]$. The initial 3,947 communities were reduced to $|\mathcal{C}| = 705$ by greedy size-constrained merging (Algorithm 1, Appendix C), with maximum community size $S_{\max} = 2{,}352$, ensuring $|\mathcal{M}_c| < n$ for all communities.

3.2 Evaluation Protocol

Cross-validation. To prevent lineage-based confounding, we evaluate all models under a grouping adaptation of the leave-one-strain-out (LOSO) protocol (Lees et al., 2020). The 45 FastBAPS (Tonkin-Hill et al., 2019) clusters are collapsed into K folds such that no fold shares a lineage

with its corresponding training set. This rewards GP relationships that generalise across genetic backgrounds, biasing toward higher precision. Grouped LOSO relaxes strict invariance in favour of average-case generalisation, a deliberate choice given lineage-varying penetrance and small within-lineage sample sizes(Pfister et al., 2021).

Community-level classifiers. For each community $c \in \mathcal{C}$, a random forest classifier is trained on Φ_c to predict $\mathbf{y}$ under the LOSO splits (hyperparameters in Appendix D; source code in Appendix A). The community score s_c (Eq. 2) is the mean AUROC across folds, and the panel $\hat{\mathcal{C}}$ is selected via the z-score (Sec. 2).

Ground truth and metrics. We evaluate against 24 known resistance genes spanning 29 recoverable gene–drug mechanism pairs across 14 phenotypes (Wheeler et al., 2019) (Appendix B). A community c recovers a known mechanism if $\mathcal{M}_c$ contains at least one unitig annotated to that gene. We report: AUROC@1 (mean AUROC of the top-ranked community over K folds); Precision@1 and Precision@N (fraction of top-1 or top-N communities containing annotated resistance unitigs); Recall@N (fraction of known gene–drug pairs recovered); and nDCG@N (whether known mechanisms appear early in the ranking). A cross-drug nDCG@N variant credits communities containing resistance-associated unitigs for any phenotype.

3.3 Baselines and ablations

We compare PANGENSA against three baselines, each representing an established paradigm in bGWAS, and two ablation conditions. All methods use the same unitig presence/absence features, grouped-LOSO splits, and phenotype labels.

Baselines. The first baseline (χ^2-top$k$ pooled) ranks all markers by χ^2 association with the phenotype, selects the top k ($k \in \{5\text{k}, 10\text{k}, 20\text{k}\}$), and trains a single global XGBoost classifier on the resulting pooled design matrix. This mirrors widely used significance-based filtering pipelines (Lees et al., 2016; Earle et al., 2016). The second baseline (PANGENSA-top$k$ pooled) uses PANGENSA's community-level AUROC ranking to select the top k features via ranking communities by score, but then discards locality by pooling their unitigs into a single global XGBoost model ($k \in \{5\text{k}, 10\text{k}, 20\text{k}\}$ unitigs from top-ranked communities). This ablation tests whether graph-informed feature selection alone is sufficient, or whether maintaining locality during GP mapping is necessary. The approach is analogous to methods that extract features from cDBG substructures but model them globally (Jaillard et al., 2018). The third baseline (Causal@γ^* pooled) restricts the feature space to unitigs residing in communities that contain at least one known resistance unitig. This oracle condition presupposes access to ground-truth causal knowledge and serves as a proxy for catalogue-based approaches (e.g. AMRFinder, ResFinder) that rely on curated resistance gene databases (Hyun et al., 2020; Kavvas et al., 2018). Any competitive performance achieved by PANGENSA relative to this oracle demonstrates effective causal localisation in the absence of prior annotation.

Ablations. To isolate the contribution of graph-constrained locality from the choice of learner, we performed two sensitivity analyses with cross-validation splits held fixed. First, at the data-driven resolution γ^*, we replaced the default RF with XGBoost and L_1-regularised logistic regression (L_1-LR), keeping all other settings identical. This tests model-class sensitivity at fixed locality.

Second, we perturbed the graph partition resolution to $\gamma = {}^1\!/_2\gamma^*$ and $\gamma = 2\gamma^*$, testing robustness to the granularity of the locality prior. As a locality ablation, we additionally constructed a shuffled partition at γ^*, denoted $\gamma^\dagger$: unitigs were randomly reassigned to size-matched communities, preserving the community size distribution but destroying all cDBG topology. This isolates the effect of graph context from the effect of community size alone.

4 RESULTS

4.1 PARTITION QUALITY

We assess partition quality in Appendix C. At $\gamma^* \approx 28.02$, the merged partition of $|\mathcal{C}| = 705$ communities exhibited strong internal connectivity: degree-corrected density enrichment R_{dens} was orders of magnitude above the Chung–Lu null across all community sizes (Chung & Lu, 2002), indicating edge concentration far beyond random chance. Conductance improved after merging ($\Delta\varphi = -0.0038$), and partitions at neighbouring resolutions ($1/2\gamma^*$, $2\gamma^*$) showed comparably low conductance, confirming robustness to moderate resolution changes. The most localised resistance gene signals (e.g. *ermA*, *ileS*, *tetK*) each confined to a single community, while the most distributed (*vanZ*; TEI) appeared in seven.

4.2 BENCHMARK COMPARISON

Table 1 compares PANGENSA against the chosen baselines (Sec. 3). The χ^2-top$k$ baselines achieved moderate AUROC@1 (0.81–0.82) but produced candidate panels spanning hundreds of communities ($|\mathcal{C}_{\mathrm{out}}| = 165$–$211$), with precision and recall near zero. Increasing k improved neither performance nor localisation. The PANGENSA-top$k$ pooled ablation achieved comparable AUROC@1 (0.75–0.82) but even lower precision, demonstrating that graph-informed selection alone is insufficient without maintaining locality during GP mapping. The oracle baseline achieved AUROC@1 of 0.86 ± 0.11 but still spread signal across 5 ± 3 communities with recall of only 0.12. PANGENSA achieved the highest AUROC@1 (0.93 ± 0.05), substantially higher precision (0.45 vs. ≤ 0.02) and recall (0.33 vs. ≤ 0.12), while confining output to 1–17 communities, outperforming even the oracle condition without prior annotation.

Table 1: Comparison of GP mapping and fine-mapping performance across methods. All methods use unitig presence/absence features. Metrics averaged over 14 phenotypes (mean ± std). Genomic spread measured against the $|\mathcal{C}| = 705$ merged communities at γ^*. Precision and recall evaluated against the full set of known gene–drug mechanism pairs.

	Input		GP mapping (Stage. 3)	Fine mapping (Stage. 4)			**Genomic spread** (#/705)	
Method	$\lvert\mathcal{M}_{\mathrm{GP}}\rvert$	$\lvert\mathcal{E}_{\mathrm{GP}}\rvert$	**AUROC@1**	$\lvert\hat{\mathcal{E}}\rvert$	**Prec.@**$\lvert\hat{\mathcal{E}}\rvert$	**Rec.@**$\lvert\hat{\mathcal{E}}\rvert$	$\lvert\mathcal{C}_{\mathrm{in}}\rvert$	$\lvert\mathcal{C}_{\mathrm{out}}\rvert$
Global pooled baselines ($\mathcal{E} = \mathcal{E}_{\tilde{\Phi}}$; global model)								
χ^2	6942 ± 1064	5 k	0.81 ± 0.17	492 ± 285	0.0215 ± 0.0733	0.0707 ± 0.1397	518 ± 113	165 ± 90
χ^2	13874 ± 3358	10 k	0.82 ± 0.15	592 ± 342	0.0118 ± 0.0394	0.0670 ± 0.1351	600 ± 86	191 ± 99
χ^2	27667 ± 5614	20 k	0.82 ± 0.16	695 ± 422	0.0050 ± 0.0150	0.0707 ± 0.1397	657 ± 29	211 ± 108
Ablation ($\mathcal{E} = \mathcal{E}_{\tilde{\Phi}}$; graph- and prediction-informed selection, global model)								
Pooled–$\mathcal{C}$	235242 ± 20635	5689 ± 457	0.82 ± 0.13	563 ± 339	0.0012 ± 0.0022	0.0466 ± 0.0978	16 ± 2	7 ± 4
Pooled–$\mathcal{C}$	294357 ± 12165	10726 ± 307	0.80 ± 0.10	794 ± 481	0.0001 ± 0.0002	0.0013 ± 0.0048	31 ± 3	18 ± 8
Pooled–$\mathcal{C}$	357087 ± 3553	20562 ± 413	0.75 ± 0.16	1002 ± 618	0.0002 ± 0.0005	0.0060 ± 0.0151	61 ± 3	31 ± 14
Oracle ($\mathcal{E} = \mathcal{E}_{\tilde{\Phi}}$; ground-truth causal labels, global model)								
Pooled–$\mathcal{C}_\Theta$	10008 ± 5693	6808 ± 3251	0.86 ± 0.11	595 ± 385	0.01 ± 0.02	0.12 ± 0.16	5 ± 3	5 ± 3
Proposed method ($\mathcal{E} = \mathcal{E}_{\mathcal{C}}$; local models per community)								
PANGENSA	$\sim \mathbf{1.2}\,\mathrm{M}$	**705**	$\mathbf{0.93 \pm 0.05}$	$\mathbf{1-17}$	$\mathbf{0.45 \pm 0.36}$	$\mathbf{0.33 \pm 0.34}$	**705**	$\mathbf{1-17}$

Columns. $|\mathcal{M}_{\mathrm{GP}}|$: total markers considered. $|\mathcal{E}_{\mathrm{GP}}|$: number of equivalence classes in the input (effective dimensionality; $\mathcal{E}_{\mathcal{C}}$ = community-induced, $\mathcal{E}_{\tilde{\Phi}}$ = correlation-induced over the filtered genotype matrix; see text). AUROC@1: AUROC of the top-ranked equivalence class (baselines: single global model). $|\hat{\mathcal{E}}|$: candidate panel size in equivalence classes. Prec./Rec.@$|\hat{\mathcal{E}}|$: precision and recall of the candidate panel evaluated against all known gene–drug mechanism pairs. $|\mathcal{C}_{\mathrm{in}}|$/$|\mathcal{C}_{\mathrm{out}}|$: number of communities containing $\geq$1 marker in the input feature set / candidate panel $\mathcal{M}^*$, respectively; lower $|\mathcal{C}_{\mathrm{out}}|$ indicates greater genomic localisation.

4.3 COMMUNITY-LEVEL CLASSIFICATION

Table 2 reports per-drug results. Across all 14 phenotypes, top-ranked communities achieved AUROC@1 ≥ 0.86, even at extreme prevalence. PANGENSA localised resistance-associated signal for 11 of 14 antibiotics, recovering 19 of 29 known gene–drug mechanism pairs within the selected panels. Of these, 14 were resolved at rank 1, with 2 at rank 2 and 1 at rank 3.

Table 2: Panel performance and diagnostic metrics per antibiotic. **Gap (Δ)** indicates the marginal loss in predictive power at rank N, while **z (σ)** quantifies the statistical significance of this drop relative to the background noise.

		Model	Selection Diagnostics			Panel Utility				Mechanism Recovery
Drug	**Known Mechanisms**	**AUROC@1**	**N**	**Gap (Δ)**	**z (σ)**	**Prec@1**	**Rec@N**	**Prec@N**	**nDCG@N**	**Recovered**
Known mechanism recovered within selected panel										
MIN	*tetM*†	1.00 ± 0.00	4	0.09	27.25	1.0	0.50	0.25	0.61 (+0.39)	*tetM*
GEN	*aac6Ie*†	0.99 ± 0.01	4	0.11	34.34	1.0	0.33	0.25	0.47 (-0.08)	*aac6Ie*
RIF	*rpoB**†	0.97 ± 0.04	1	0.25	286.65	1.0	1.00	1.00	1.00 (+0.00)	*rpoB**
CIP	*parC**†, *parE**, *gyrA**†, *gyrB**	0.94 ± 0.09	1	0.14	59.45	1.0	1.00	1.00	1.00 (+0.00)	*parC**, *parE**, *gyrA**
CLI	*lnuA*, *ermA*, *ermB*, *ermC*	0.92 ± 0.09	1	0.07	46.53	1.0	0.14	1.00	1.00 (+0.00)	*ermB*, *ermC*
ERY	*msrA*, *ermA*, *ermB*†, *ermC*†	0.91 ± 0.04	2	0.10	45.05	1.0	0.12	0.50	0.61 (+0.00)	*ermB*, *ermC*
FUS	*fusA**†, *fusC*†	0.89 ± 0.06	3	0.06	29.16	1.0	0.67	0.67	0.70 (+0.30)	*fusA**, *fusC*
TET	*tetK*†, *tetL*, *tetM*†	0.89 ± 0.08	2	0.04	11.85	1.0	0.25	0.50	0.61 (+0.00)	*tetK*, *tetL*
TMP	*dfrS1*†, *dfrG*†	0.88 ± 0.08	3	0.03	23.17	1.0	0.25	0.33	0.47 (+0.23)	*dfrS1*
PEN	*blaZ*†, *mecA*†	0.97 ± 0.03	5	0.04	18.65	0.0	0.25	0.40	0.38 (+0.00)	*blaZ*
MUP	*ileS**†, *mupA*†	0.88 ± 0.12	3	0.08	30.78	0.0	0.14	0.33	0.30 (+0.47)	*ileS**, *mupA*
No recovery within selected panel (Failure Modes)										
MET	*mecA*†	0.98 ± 0.02	17	0.05	26.89	0.0	0.00	0.00	0.00 (+0.19)	–
TEI	*vanZ*	0.90 ± 0.10	1	0.02	11.96	0.0	0.00	0.00	0.00 (+0.00)	–
LIN	*rplC**	0.86 ± 0.18	1	0.02	20.43	0.0	0.00	0.00	0.00 (+0.00)	–

*Indicates point mutation mechanism vs. acquired gene. †Flags known mechanisms recovered by at least one of four bGWAS methods (Plink, Pyseer, FastLMM, treeWAS) reported by Wheeler et al. (2019).
For **nDCG@N**, the main value is per-drug attribution; parentheses show value change for cross-drug attribution (communities with at least one node labelled resistant for any phenotype).
Drug abbreviations: CIP (ciprofloxacin), CLI (clindamycin), ERY (erythromycin), FUS (fusidic acid), GEN (gentamicin), LIN (linezolid), MET (methicillin), MIN (minocycline), MUP (mupirocin), PEN (penicillin), RIF (rifampicin), TEI (teicoplanin), TET (tetracycline), TMP (trimethoprim).

4.4 Panel Selection and Diagnostic Patterns

Each phenotype can be characterised by the diagnostic pair $(N, \text{AUROC@1})$. Low N paired with high AUROC@1 marked compact, mechanistically coherent hotspots (e.g. GEN with $N = 4$, RIF with $N = 1$), where a single locus dominated prediction. Larger panels (e.g. MET with $N = 17$) achieved similarly high AUROC@1 but exhibited distributed attribution, indicating poly-locus architectures and/or residual population structure. Five of 14 phenotypes selected only communities containing resistance-associated unitigs for at least one drug, achieving cross-drug nDCG@N of 1.00.

4.5 Sensitivity Analysis

Locality was the dominant factor (Table 3). Resolution perturbations ($1/2\gamma^*$, $2\gamma^*$) retained moderate localisation (Precision@1 of 0.50 and 0.62), whereas the shuffled ablation $\gamma^\dagger$ collapsed Precision@1 to 0.14 and Recall@N to 0.01, confirming that graph context, not community size, drives fine-mapping performance. At fixed γ^*, swapping the learner produced modest shifts: RF achieved Precision@1 of 0.64 vs. XGBoost (0.50) and L_1-LR (0.43), with all three ranking the same top signals for most phenotypes. This limited model-class dependence suggests that, once features are grouped into coherent genomic units, the GP mapping is largely model-agnostic.

4.6 Qualitative Analysis

PANGENSA performed best on single-mechanism phenotypes: GEN (*aac(6')-Ie*), MIN (*tetM*), and RIF (*rpoB*), the latter two at $\leq$1.43% resistance prevalence. These were also multi-drug resistance regions, where the top community captured signal across multiple phenotypes from distinct mechanisms. For six additional phenotypes (CIP, CLI, ERY, FUS, TET, TMP), known determinants were recovered at rank 1. Notably, PANGENSA identified secondary genes missed by all four base-

Table 3: Sensitivity to locality and function space definitions in the GP mapping formulation.

Classifier	**Resolution** γ	**Precision**@1	**Precision**@N	**Recall**@N
Resolution comparison / locality sensitivity (RF; perturb γ).				
RF	$1/2\gamma^*$	0.50 ± 0.52	0.50 ± 0.39	0.36 ± 0.35
RF	γ^*	0.64 ± 0.50	0.45 ± 0.36	0.29 ± 0.28
RF	$2\gamma^*$	0.62 ± 0.51	0.36 ± 0.35	0.25 ± 0.30
Locality ablation (size-matched shuffling at γ^*; removes graph context).				
RF	γ^*	0.64 ± 0.50	0.45 ± 0.36	0.29 ± 0.28
RF	$\gamma^\dagger$	0.14 ± 0.36	0.13 ± 0.30	0.01 ± 0.02
Model-class (function space) sensitivity at fixed partition γ^*.				
XGBoost	γ^*	0.50 ± 0.52	0.40 ± 0.39	0.26 ± 0.30
L1-LR	γ^*	0.43 ± 0.51	0.28 ± 0.35	0.21 ± 0.28
RF	γ^*	0.64 ± 0.50	0.45 ± 0.36	0.29 ± 0.28

All entries report mean $\pm$ s.d. over 14 antibiotic drugs.
Resolution / locality sensitivity. RF recovery under γ perturbations ($1/2\gamma^*$, γ^*, $2\gamma^*$).
Locality ablation. $\gamma^\dagger$ denotes size-matched shuffled communities at γ^* (preserves size distribution; removes cDBG topology).
Model-class sensitivity. Learner choice (RF, XGBoost, L1-penalised logistic regression) at fixed γ^*.

line bGWAS methods of Wheeler et al. (Wheeler et al., 2019): *ermB* (ERY/CLI), *parE* (CIP), and *tetL* (TET), each present in only 3–7 genomes, demonstrating that community-level modelling amplifies low-prevalence resistance-associated signals.

Three failure modes account for non-recovery. First, sub-unitig variation (point mutations, regulatory changes) cannot be captured by presence/absence features, affecting (*gyrB*; CIP), (*rplC*; LIN), and (*ileS*; MUP). Second, near-ubiquitous markers such as *mecA* (4,138/4,139 isolates) are perfectly confounded with clonal background and unrewarded by LOSO. Third, mechanisms concentrated within one or two lineages are attenuated under collapsed-LOSO when held-out folds contain too few positives, affecting (*tetM*; TET) and (*dfrG*; TMP). We note that AUROC-based community ranking is a prioritisation tool for experimental follow-up, not a formal causal estimator; selected panels should be interpreted as candidate regions enriched for resistance-associated signal.

5 Conclusion

In this work, we introduced PANGENSA, a graph-constrained machine learning framework that mitigates the high multicollinearity and underdetermined regimes in bacterial GWAS by encoding genomic locality as a structural prior. By partitioning the pangenome cDBG into distinct communities and reframing the GP mapping as independent subproblems, PANGENSA ensures that candidate variants inherently map to localised genomic regions.

Our empirical results on 14 *S. aureus* phenotypes show strong predictive performance, with top-ranked communities achieving AUROC@1 values ranging from 0.86 to 1.00. PANGENSA localised resistance-associated signal for 11 of 14 antibiotics without explicit causal interventions or global feature pooling. Notably, the community-level constraint successfully amplified rare, low-prevalence resistance mechanisms that were missed by standard global baselines.

Despite these successes, PANGENSA has notable limitations. The framework relies on presence/absence encodings of unitigs, rendering it unable to properly represent sub-unitig variations such as point mutations. Furthermore, near-ubiquitous markers perfectly linked to clonal backgrounds (e.g., *mecA*) remain heavily confounded due to the strictness of the leave-one-strain-out protocol. Ultimately, while community-ranked predictive scores are not formal causal estimators, PANGENSA serves as a powerful and highly interpretable prioritisation tool to guide wet-lab experimental validation of novel AMR determinants.

REFERENCES

Francis Bach, Rodolphe Jenatton, Julien Mairal, and Guillaume Obozinski. Structured sparsity through convex optimization. 2012.

RV Broekema, OB Bakker, and IH Jonkers. A practical view of fine-mapping and gene prioritization in the post-genome-wide association era. *Open biology*, 10(1), 2020.

Garth R Brown, Vichet Hem, Kenneth S Katz, Michael Ovetsky, Craig Wallin, Olga Ermolaeva, Igor Tolstoy, Tatiana Tatusova, Kim D Pruitt, Donna R Maglott, et al. Gene: a gene-centered information resource at ncbi. *Nucleic acids research*, 43(D1):D36–D42, 2015.

Peter E Chen and B Jesse Shapiro. The advent of genome-wide association studies for bacteria. *Current opinion in microbiology*, 25:17–24, 2015.

Fan Chung and Linyuan Lu. The average distances in random graphs with given expected degrees. *Proceedings of the National Academy of Sciences*, 99(25):15879–15882, 2002.

Caitlin Collins and Xavier Didelot. A phylogenetic method to perform genome-wide association studies in microbes that accounts for population structure and recombination. *PLoS computational biology*, 14(2):e1005958, 2018.

G Csardi and T Nepusz. The igraph software package for complex network research. interjournal complex systems 1695. *Available at igraph. org/. Accessed November*, 30:2015, 2006.

Sarah G Earle, Chieh-Hsi Wu, Jane Charlesworth, Nicole Stoesser, N Claire Gordon, Timothy M Walker, Chris CA Spencer, Zamin Iqbal, David A Clifton, Katie L Hopkins, et al. Identifying lineage effects when controlling for population structure improves power in bacterial association studies. *Nature microbiology*, 1(5):1–8, 2016.

Dario Fasino, Arianna Tonetto, and Francesco Tudisco. Generating large scale-free networks with the chung–lu random graph model. *Networks*, 78(2):174–187, 2021.

Jacques Hadamard. *Lectures on Cauchy's problem in linear partial differential equations*. Courier Corporation, 2014.

Aric Hagberg, Pieter Swart, and Daniel S Chult. Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.

Hussein Hazimeh, Rahul Mazumder, and Peter Radchenko. Grouped variable selection with discrete optimization: Computational and statistical perspectives. *The Annals of Statistics*, 51(1):1–32, 2023.

Jason C Hyun, Erol S Kavvas, Jonathan M Monk, and Bernhard O Palsson. Machine learning with random subspace ensembles identifies antimicrobial resistance determinants from pan-genomes of three pathogens. *PLoS computational biology*, 16(3):e1007608, 2020.

Adam D Irwin, Lachlan JM Coin, Patrick NA Harris, Menino Osbert Cotta, Michelle J Bauer, Cameron Buckley, Ross Balch, Peter Kruger, Jason Meyer, Kiran Shekar, et al. Optimising treatment outcomes for children and adults through rapid genome sequencing of sepsis pathogens. a study protocol for a prospective, multi-centre trial (direct). *Frontiers in Cellular and Infection Microbiology*, 11:667680, 2021.

Magali Jaillard, Leandro Lima, Maud Tournoud, Pierre Mahé, Alex Van Belkum, Vincent Lacroix, and Laurent Jacob. A fast and agnostic method for bacterial genome-wide association studies: Bridging the gap between k-mers and genetic events. *PLoS genetics*, 14(11):e1007758, 2018.

T James, B Williamson, P Tino, and N Wheeler. Whole-genome phenotype prediction with machine learning: Open problems in bacterial genomics. *arXiv preprint arXiv:2502.07749*, 2025.

Erol S Kavvas, Edward Catoiu, Nathan Mih, James T Yurkovich, Yara Seif, Nicholas Dillon, David Heckmann, Amitesh Anand, Laurence Yang, Victor Nizet, et al. Machine learning and structural analysis of mycobacterium tuberculosis pan-genome identifies genetic signatures of antibiotic resistance. *Nature communications*, 9(1):4306, 2018.

Jee In Kim, Finlay Maguire, Kara K Tsang, Theodore Gouliouris, Sharon J Peacock, Tim A McAllister, Andrew G McArthur, and Robert G Beiko. Machine learning for antimicrobial resistance prediction: current practice, limitations, and clinical perspective. *Clinical microbiology reviews*, 35(3):e00179–21, 2022.

John A Lees, Minna Vehkala, Niko Välimäki, Simon R Harris, Claire Chewapreecha, Nicholas J Croucher, Pekka Marttinen, Mark R Davies, Andrew C Steer, Steven YC Tong, et al. Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes. *Nature communications*, 7(1):12797, 2016.

John A Lees, T Tien Mai, Marco Galardini, Nicole E Wheeler, Samuel T Horsfield, Julian Parkhill, and Jukka Corander. Improved prediction of bacterial genotype-phenotype associations using interpretable pangenome-spanning regressions. *MBio*, 11(4):10–1128, 2020.

Jeanneth Mosquera-Rendón, Claudia Ximena Moreno-Herrera, Jaime Robledo, and Uriel Hurtado-Páez. Genome-wide association studies (gwas) approaches for the detection of genetic variants associated with antibiotic resistance: a systematic review. *Microorganisms*, 11(12):2866, 2023.

Interagency Coordination Group on Antimicrobial Resistance. No time to wait: securing the future from drug-resistant infections. *Report to the Secretary General of the United Nations*, 2019.

William R Pearson and David J Lipman. Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences*, 85(8):2444–2448, 1988.

Niklas Pfister, Evan G. Williams, Jonas Peters, Ruedi Aebersold, and Peter Bühlmann. Stabilizing variable selection and regression, 2021. URL `https://arxiv.org/abs/1911.01850`.

Morteza M Saber and B Jesse Shapiro. Benchmarking bacterial genome-wide association study methods using simulated genomes and phenotypes. *Microbial genomics*, 6(3):e000337, 2020.

James Emmanuel San, Shakuntala Baichoo, Aquillah Kanzi, Yumna Moosa, Richard Lessells, Vagner Fonseca, John Mogaka, Robert Power, and Tulio de Oliveira. Current affairs of microbial genome-wide association studies: approaches, bottlenecks and analytical pitfalls. *Frontiers in microbiology*, 10:3119, 2020.

Daniel J Schaid, Wenan Chen, and Nicholas B Larson. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nature Reviews Genetics*, 19(8):491–504, 2018.

R Greg Stacey, Michael A Skinnider, and Leonard J Foster. On the robustness of graph-based clustering to random network alterations. *Molecular & Cellular Proteomics*, 20, 2021.

Gerry Tonkin-Hill, John A Lees, Stephen D Bentley, Simon DW Frost, and Jukka Corander. Fast hierarchical bayesian analysis of population structure. *Nucleic acids research*, 47(11):5539–5549, 2019.

Vincent Traag, Fabio Zanini, Ryan Gibson, Oren Ben-Kiki, Tom Kelly, Britny Farahdel, and Dafne van Kuppevelt. vtraag/leidenalg: 0.10.0, 2023.

Vincent A Traag, Ludo Waltman, and Nees Jan Van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific reports*, 9(1):1–12, 2019.

Qingbo S Wang and Hailiang Huang. Methods for statistical fine-mapping and their applications to auto-immune diseases. In *Seminars in immunopathology*, volume 44, pp. 101–113. Springer, 2022.

Nicole E Wheeler, Sandra Reuter, Claire Chewapreecha, John A Lees, Beth Blane, Carolyne Horner, David Enoch, Nicholas M Brown, M Estée Török, David M Aanensen, et al. Contrasting approaches to genome-wide association studies impact the detection of resistance mechanisms in staphylococcus aureus. *bioRxiv*, pp. 758144, 2019.

Yang Wu, Zhili Zheng, Loic Thibaut, Tian Lin, Qian Feng, Hao Cheng, Loic Yengo, Michael E Goddard, Naomi R Wray, Peter M Visscher, et al. Genome-wide fine-mapping improves identification of causal variants. *medRxiv*, pp. 2024–07, 2025.

A Code availability

The codebase, including the custom grouped-LOSO splitting algorithm, cross-validation scripts, and graph partitioning implementations, is available at `https://github.com/tejames42/pangensa`.

B Dataset details

B.1 Data availability

Collections of isolates were selected and aggregated into a unified dataset by Wheeler et al. (at bioRxiv, doi:10.1101/758144, 2019).

Corresponding raw Illumina sequencing reads are available in the European Nucleotide Archive (ENA, `https://www.ebi.ac.uk/ena`) under project accessions ERP001012 (Reuter et al., *Genome Research*, doi:10.1101/gr.196709.115, 2016), PRJEB3174 (Coll et al., *Science Translational Medicine*, doi:10.1126/scitranslmed.aak9745, 2017), and PRJEB2755, PRJEB2756, and PRJEB2944 (Donker et al., *Microbial Genomics*, doi:10.1099/mgen.0.000113, 2017).

Annotated assemblies were produced with the Sanger bacterial pipeline (`https://github.com/sanger-pathogens/vrcodebase`) and unitigs were extracted using DBGWAS (`https://gitlab.com/leoisl/dbgwas`). No new raw sequencing data were generated for this work. The processed sequencing data are available from the corresponding author upon reasonable request.

Phenotype labels and class balance

Phenotypes were encoded as binary S/R labels; genotype quality filtering and phenotype determination methods are described by Wheeler et al. (Wheeler et al., 2019). Figure 2 summarises label availability and class balance per drug. Label counts ranged from $n = 908$ (MUP) to $n = 4{,}139$ (MET, PEN), with 8 of 14 antibiotics tested in all isolates. Resistance prevalence varied from 0.14% (LIN; 3 of 45 lineages) to 97.54% (PEN; 43 of 45 lineages).

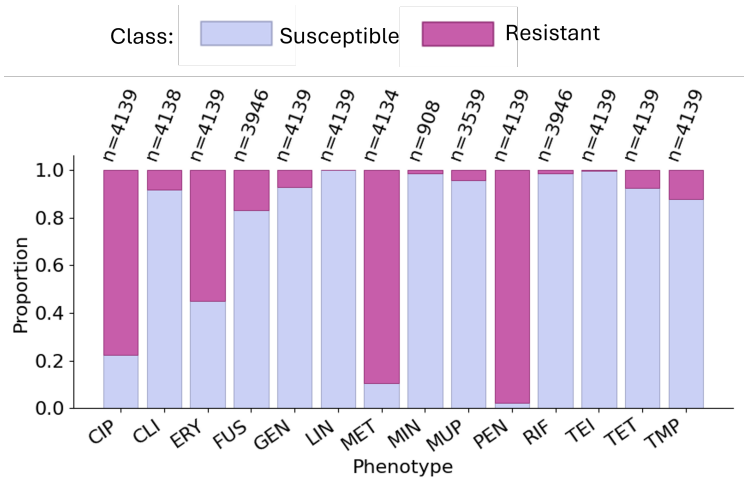

Figure 2: **Phenotype label coverage across antibiotics.** Number of genomes with an available S/R label and corresponding class balance per drug.

Population structure

Population structure was inferred using FastBAPS (Tonkin-Hill et al., 2019), yielding both a coarse (level-1) and fine (level-2) clustering. The fine clustering, used throughout this study, divides the cohort into 33 clonal complexes (CCs) and 12 orphan sequence types (singleton clusters), totalling

45 distinct labels. The five dominant CCs are CC22/EMRSA-15 ($n = 2{,}802$), CC30/EMRSA-16 ($n = 436$), CC5 ($n = 168$), CC8 ($n = 147$), and CC1 ($n = 145$). Figure 3 shows the full composition.

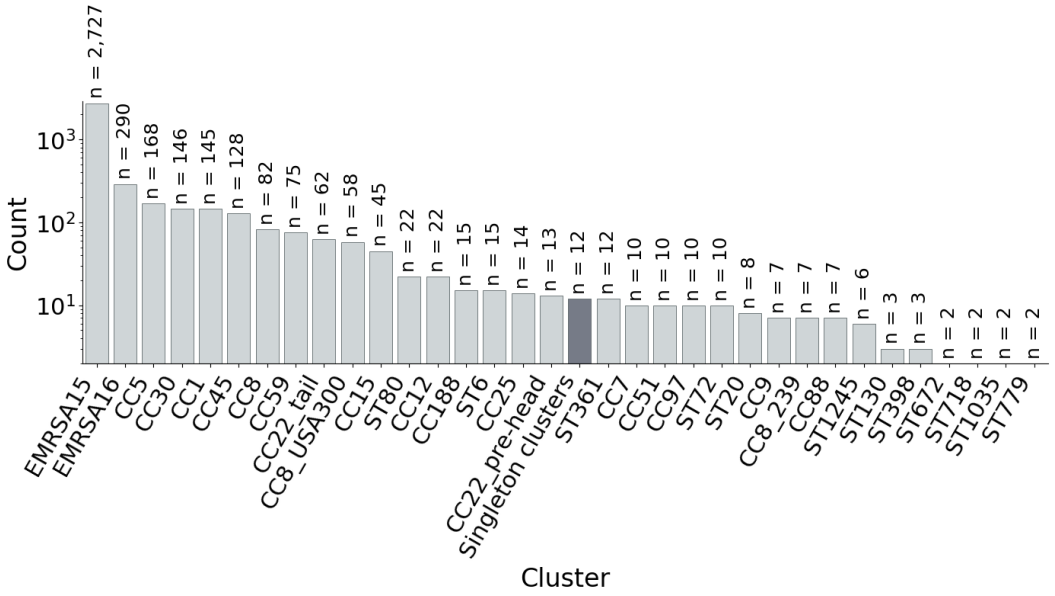

Figure 3: **Population structure of the *S. aureus* cohort.** Composition of the fine (level-2) FastBAPS clustering. Clusters are grouped into clonal complexes (CCs) and singleton sequence types, totalling 45 labels.

PANGENOME GRAPH

The cDBG was constructed from all short-read assemblies using DBGWAS (Jaillard et al., 2018). The resulting graph contained $|V_t| = 1{,}238{,}044$ unitig nodes. A single giant connected component accounted for 1,237,935 nodes; the remaining nodes were distributed across five minor components (sizes 2–84), which were excluded from downstream analysis. Genotypes were encoded as binary unitig presence/absence vectors. Collapsing identical presence/absence patterns reduced the feature count from 1,238,044 to 733,415 unique equivalence classes.

GROUND TRUTH ANNOTATION

Unitigs were annotated with known resistance gene families by searching each unitig sequence (both strands) for matches to a curated library of resistance gene sequences in FASTA format (Pearson & Lipman, 1988) from NCBI's gene database (Brown et al., 2015). Unitigs containing a gene sequence as a substring were labelled with that gene family. Four gene–drug pairs lacked matching unitigs and were excluded from ground-truth evaluation: (*fusB*; FUS), (*23S rRNA*; LIN), (*dfrB*; TMP), and (*vanA*; TEI). The full set of recoverable known mechanisms is listed in Table 2.

C DATA-DRIVEN PANGENOME GRAPH PARTITIONING

This section is devoted to the assessment of the quality of the partition $\mathcal{C}$ on G_t (Fig. 1, Stage 1) assembled from observed data $\mathcal{D}_t$, sampled at time t under the assumption of fixed environmental conditions.

Specifically, we perform data-driven partitioning of the cDBG as part of Stage 1 of the workflow (Fig. 1) to hard-code a locality prior in the GP mapping function, and assess it against a set of design goals: (i) high internal connectivity, (ii) compactness, (iii) low inter-community leakage, (iv) biological coherence, and (v) maximised loss-contrast between causal and non-causal genomic regions. We assess design goals (i)–(iv) in this section.

Using a Leiden partitioning framework (Algorithm 1) we scanned a broad range of resolution parameter γ and found a a single sharp optimum at $\gamma^* \approx 28.02$, which was stable across different search criteria (e.g. detecting an empty gap in the $\log_{10}$ community size distribution, and varying the bracketing tolerance for refining γ). This invariance suggests a true, robust mesoscale structure in the pangenome rather than a tuning artifact. At γ^*, densely co-occurring genomic modules (clusters of neighbouring unitigs) emerge naturally, separated from background connectivity. In other words, γ^* marks the point where an over-aggregated giant community splits into multiple internally dense, well-connected communities. Because the Leiden method guarantees connectivity within each detected module (Traag et al., 2019), the partition at γ^* yields feature sets supporting biological coherence and represents the coarsest resolution at which coherent genomic neighbourhoods are resolved across the genome. This partitioning aids compactness (fewer, larger features) while limiting the fragmentation of causal signal.

Our coarse resolution sweep spanned four orders of magnitude in γ (from 0.1 to 100) to avoid local optima, with a fixed random seed to ensure reproducibility (Stacey et al., 2021). We implemented the Leiden algorithm via the `leidenalg` package (Traag et al., 2023) via Python-igraph (Csardi & Nepusz, 2006) on the cDBG initialised in NetworkX(Hagberg et al., 2008). The initial Leiden partition (at γ^*) contained 3,947 communities, which we further merged by a greedy, size-constrained procedure, where each community was iteratively merged into the neighbouring community with which it shared the highest total edge weight, provided the merged size did not exceed the largest original community (2,352 nodes). This merging step reduced 3,947 Leiden communities to a final set of 705 merged communities. These 705 merged communities constitute the local genomic modules used in subsequent PANGENSA analysis (Fig. 1b).

We assessed the quality of this partition at γ^* against the design goals. (This evaluation considered only communities from the main giant component of the cDBG, excluding the five tiny disconnected components.)

Figure 4a plots the degree-corrected density enrichment R_{dens} = (observed density)/(expected density under a Chung-Lu random graph) for each of the 700 merged communities. The Chung-Lu model does not introduce correlations among node degrees, and represents the fundamental null model for finding community structures(Fasino et al., 2021). Across all community sizes, we observe $R_{\text{dens}} \gg 1$, often orders of magnitude above the null expectation, indicating strong within-community edge concentration (dense internal connectivity) beyond random chance(Chung & Lu, 2002). Notably, global transitivity in the graph is uniformly low, meaning the enrichment is not driven by many triangles (which would inflate clustering coefficients), but rather by higher-order connectivity within modules.

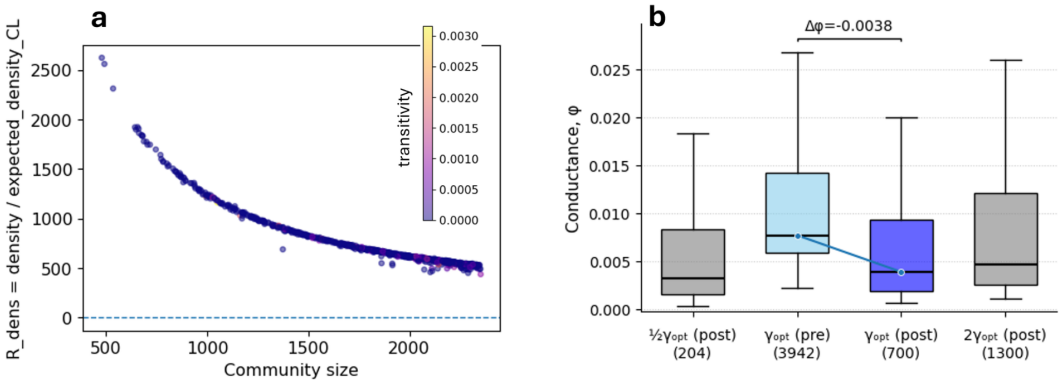

Figure 4: **Quality of cDBG partitioning at the data-driven resolution. (a)** The degree-corrected density enrichment R_{dens} for the merged cDBG communities at resolution $\gamma = \gamma^*$, quantifying within-community edge concentration relative to a Chung-Lu null. **(b)** Conductance (φ) distributions comparing pre-merged vs merged communities at γ^*, up to a maximum community size constraint, and neighbouring resolutions $(1/2\gamma^*, 2\gamma^*)$.

Conductance analyses (Fig. 4b) confirmed that our γ^* partition yields well-separated, internally cohesive modules. At γ^*, the median conductance improved by $\Delta\varphi = -0.0038$ after merging

(from the initial Leiden partition with 3,942 communities to the merged set of 700), indicating fewer boundary edges per unit internal degree despite the communities becoming larger on average. Importantly, partitions obtained at $1/2\gamma^*$ (204 communities) and $2\gamma^*$ (1,300 communities) also showed comparably low conductance distributions. This demonstrates that the mesoscale partitioning is robust to moderate changes in resolution. By design, conductance φ measures the fraction of a community's edge volume that lies across its boundary, so a downward shift signifies a direct improvement in community quality, not an artifact of size effects. Since all partitions are produced by Leiden (yielding connected communities), we can be confident that the observed conductance gains reflect truly well-knit genomic modules, rather than the trivial effect of splitting disconnected pieces.

In summary, our graph partition achieved low inter-community leakage and high internal coherence, where the merged communities at γ^* are internally dense yet boundary-sparse, and remain well-connected by construction. These properties ensure that PANGENSA's features (unitig communities) correspond to interpretable genomic units (local neighbourhoods in the genome) while being compact enough for downstream modelling (community size $< n$).

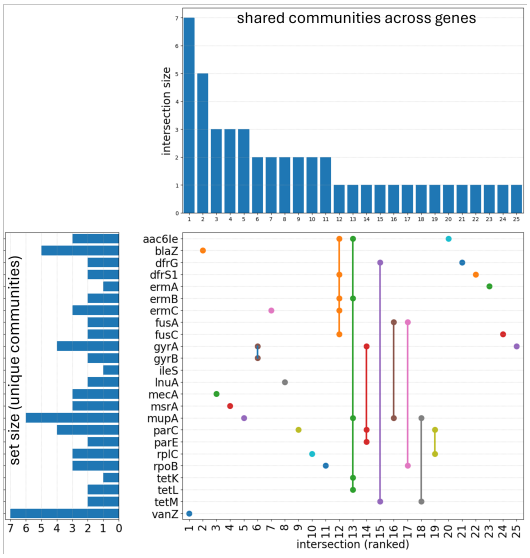

Figure 5: **Shared genomic contexts across resistance genes.** UpSet plot showing the overlap among communities containing unitigs annotated to 24 known ABR genes. The top bar graph indicates the size of each intersection (how often a specific combination of genes co-occurs in the same community). The left bar graph shows the number of unique communities containing each gene (set size), distinguishing highly localised signals (genes found in few contexts) from widely spread signals (genes in many contexts). Each filled circle in the matrix indicates membership of a gene in a given community intersection. Unconnected single dots denote gene-specific contexts, while vertically connected dots mark genes co-occurring in the same community.

The spread of known resistance gene signals across these communities is visualized in Fig. 5. Each of 24 curated resistance genes is mapped to the set of communities in which its annotated unitigs occur, and the overlaps are shown in an UpSet plot. Vertical connected dots indicate groups of genes co-occurring in the same community (i.e., multi-gene contexts), while disconnected single dots indicate mechanisms that occupy distinct communities. The top bar plot ranks multi-gene intersections by size (number of communities sharing that specific gene combination), and the left bar plot shows the number of distinct communities per gene.

This reveals which resistance mechanisms are highly localised (a gene appears in only one or few communities) versus widely distributed and poly-locus (a gene's unitigs span many different communities). In our data, the most localised signals, each confined to a single community, include (*ermA*; CLI-ERY), (*ileS*; MUP), and (*tetK*; TET), whereas the most widely distributed is (*vanZ*; TEI) appearing in 7 communities. Multi-mechanism loci manifest as vertical stacks of dots, indicat-

ing multiple known genes co-residing in the same community (e.g., certain plasmids or transposons carrying several resistance genes).

C.1 PANGENSA PARTITIONING APPROACH

Algorithm 1 cDBG Partitioning and Refinement (PANGENSA)

Require: cDBG $G_t = (V_t, E_t)$, Resolution range $\Gamma = [0.1, 100]$
Ensure: Final partition $\mathcal{C}_{final}$
Phase 1: Structure Discovery
for $\gamma \in \Gamma$ **do**
  $\mathcal{C}_\gamma \leftarrow$ Leiden(G_t, γ)
  $score_\gamma \leftarrow$ StabilityGap$(\mathcal{C}_\gamma)$ {Detect gap in size distribution}
end for
$\gamma^* \leftarrow \arg\max_\gamma score_\gamma$ {Found $\gamma^* \approx 28.02$}
$\mathcal{C}_{init} \leftarrow$ Leiden(G_t, γ^*)
Phase 2: Size-Constrained Merging
$S_{max} \leftarrow \max_{c \in \mathcal{C}_{init}} |c|$ {Set size constraint (e.g., 2,352)}
repeat
  $merged \leftarrow$ False
  for $c_i \in \mathcal{C}_{init}$ **do**
    $c_{best} \leftarrow \arg\max_{c_j \in \text{Neighbors}(c_i)}$ Weight(c_i, c_j)
    if $|c_i| + |c_{best}| \leq S_{max}$ **then**
      $\mathcal{C}_{init} \leftarrow (\mathcal{C}_{init} \setminus \{c_i, c_{best}\}) \cup \{c_i \cup c_{best}\}$
      $merged \leftarrow$ True
    end if
  end for
until not $merged$
return $\mathcal{C}_{init}$

D IMPLEMENTATION DETAILS

Models were implemented via `scikit-learn` and `xgboost` with fixed hyperparameters. Random Forest: `n_estimators`=256, `max_depth`=30, `min_samples_split`=5, `max_features`='sqrt', `class_weight`='balanced'. XGBoost: `n_estimators`=400, `learning_rate`=0.05, `max_depth`=3, `colsample_bytree`=0.8, `min_child_weight`=1. L_1-LR: `penalty`='l1', `solver`='liblinear', `C`=10. Grouped-LOSO cross-validation partitioned the 45 FastBAPS lineages into $K \leq 5$ folds, strictly preventing lineage leakage between training and test sets.