A TOPOLOGICALLY GUIDED MACHINE LEARNING FRAMEWORK FOR ENHANCED FINE-MAPPING IN WHOLE-GENOME BACTERIAL STUDIES

Tamsin James & Peter Tino
School of Computer Science
University of Birmingham, UK
txj287@student.bham.ac.uk

Nicole Wheeler
Advanced Research and Invention Agency (ARIA)
London, UK

ABSTRACT

This paper proposes a feature selection framework for machine learning–based bacterial genome-wide association studies aimed at uncovering resistance-causing traits. Using a well-characterized *Staphylococcus aureus* pangenome as a ground truth for causal-variant labels, we demonstrate improved control for population structure and enhanced interpretability through the explicit incorporation of genomic context derived from graph-structured data, based on the compacted de Bruijn graph for an assembled pangenome. Our framework successfully uncovers resistance-causing traits for 9 of 14 antibiotics using a significantly reduced feature set, while preserving genomic marker identifiability via unique mappings between the encoded feature space and sequential representations that tag specific genomic loci.

1 INTRODUCTION

Identifying an optimal subset of features from a labeled dataset is an essential property in many biological domains, such as genomics, proteomics, and precision medicine, where treatment plans and/or outcomes are guided by genetic profiles. Causal feature selection (identifying features having a causal relationship with an outcome), or genetic fine-mapping, defines the main objective in bacterial genome-wide association studies (bGWAS), which aim to find statistically significant associations between genotypes and phenotypes through the modeling of observed genotype-phenotype (GP) relationships. For example, antibiotic resistance phenotypes present a critical application where computational prediction of causal genetic variants can inform diagnostic strategies and guide targeted interventions to combat resistance.

Population structure, or ancestry, captures the genetic similarity between isolates due to clonal inheritance and shared lineages, and is the primary confounding factor in resistance bGWAS (Earle et al., 2016). Ancestry confounding manifests as lineage-level differences (stratification of data points), and linkage disequilibrium (LD) across large genomic blocks (multicollinearity) (Mosquera-Rendón et al., 2023). In a machine learning (ML) setting, stratification affects the distribution of samples and can lead to biases in the analysis if not properly accounted for, and multicollinearity affects the model's ability to disentangle the effects of individual predictors. When dealing with the high-dimensional and multicollinear nature of bacterial data, the learning of the GP mapping is inherently ill-posed: the optimization process may yield multiple solutions that equally satisfy the training data, leading to non-unique and unstable mappings that do not generalize well to unseen data.

Accurate bacterial fine mapping requires minimizing the number of candidate variants falsely identified so that experimental studies on resistance remain feasible and focused, rather than overwhelmed by spurious candidates. This necessitates robust control for confounding effects due to ancestry, namely stratification and multicollinearity, however, strict feature selection methods can lead to a reduction in power (Lees et al., 2020). Additionally, fine mapping demands interpretability: genomic sequences should be uniquely identifiable from candidates extracted from the model output. In practice, however, this is typically a one-to-many relationship. Predictive models treating highly correlated variables interchangeably exacerbates this issue, which when combined with the perva-

sive nature of genome-wide LD in bacteria, leads to an inflation in the number of predictions where candidate loci are scattered across the entire genome.

In response to this, we leverage graph-structured genomic data to constrain the function space for the goal of reducing the severity of non-uniqueness in the learned mapping of a statistical model. Specifically, we address identifiability concerns by assigning biological relevance based on genomic structure to features. This then creates a paradigm to address the multicollinearity manifestation of ancestry confounding via harsh feature selection under the assumption that interactions between genetic variants are governed by a distance-based hierarchy.

To summarize, this paper offers the following contributions: **(i)** a general and scalable machine learning bGWAS methodology for phenotype prediction-based bacterial fine mapping, **(ii)** an informed feature selection framework based on graph structured genotype data for integration with existing statistical learning bGWAS pipelines, **(iii)** an experimental ranking framework for selecting subsets of genetic variants as candidates based on model performance and suggesting a level of prediction confidence, and **(iv)** an example instantiation of the framework involving binary resistance classification across 14 antibiotic agents via the modeling of a well-characterized *Staphylococcus Aureus* pangenome.

2 Related Work

Lineage stratification. Numerous regularization techniques exist to explicitly control for population structure manifesting as stratification. For example, Phelan et al. (2016) extract principal components from the genotype matrix, Earle et al. (2016) construct a kinship matrix to regularize a linear mixed model by modeling random effects, and Diaz Caballero et al. (2018) incorporate genetic distances obtained from phylogenetic reconstruction tools. Implicit modeling of population structure is typically achieved using cross-validation (CV) techniques, as demonstrated by Lees et al. (2020) using strain-specific partitions as a leave-one-strain-out methodology for an elastic net model. Inspired by the leave-one-strain-out approach, we also perform cross-validation with strain-specific partitions. However, to reduce computational demands, we partition strains across 5 folds by combining multiple strains to construct validation sets, rather than just validating on a single strain.

Multicollinearity. Feature selection and marker choice provide the dominant approaches to addressing multicollinearity that results from LD and high-dimensionality of bacterial data. For example, Yang & Wu (2023) and Biggs et al. (2021) prioritize features based on variable importances determined by an initial modeling step. Alternatively, Mallawaarachchi et al. (2022) prioritize features based on LD scores, Hyun et al. (2023) on GWAS scores. Saber & Shapiro (2020) and Lees et al. (2020) leverage regularization techniques to shrink coefficients of redundant features. Our approach manages LD effects via an informed feature selection framework, relying on a biological assumption rather than prior causal knowledge or model reasoning.

Information loss. Genotype matrix representations suffer from information loss due to the omission of genomic context and positional dependencies, as the matrix structure treats each genetic variant as an independent feature and disregards the presence of interactions along the genome. cry (2024) incorporates spatial information within a sequential representation of the genome by prioritizing known causal and statistically significant sites identified in bGWAS and the adjacent $\approx$ 100bps upstream and downstream of these loci. Jaillard et al. (2017) incorporate spatial context by leveraging a graph structure representation of genomic data to reduce feature dimensionality during pre-processing within a traditional statistical testing setting, and during the post-processing stage for analyzing the output of a linear mixed model (Jaillard et al., 2018). We incorporate structural information via a feature selection method that leverages graph structured genomic data within a ML setting. We also consider the full set of features for analysis and therefore avoid limiting the scope of the study by retaining the complete genetic profile.

3 A comprehensive resistance dataset for *Staphylococcus aureus*

To establish a robust ground truth as the foundation for evaluation, we use a dataset in which resistance mechanisms are thoroughly characterized. A total of 4,138 *S. aureus* isolates have been

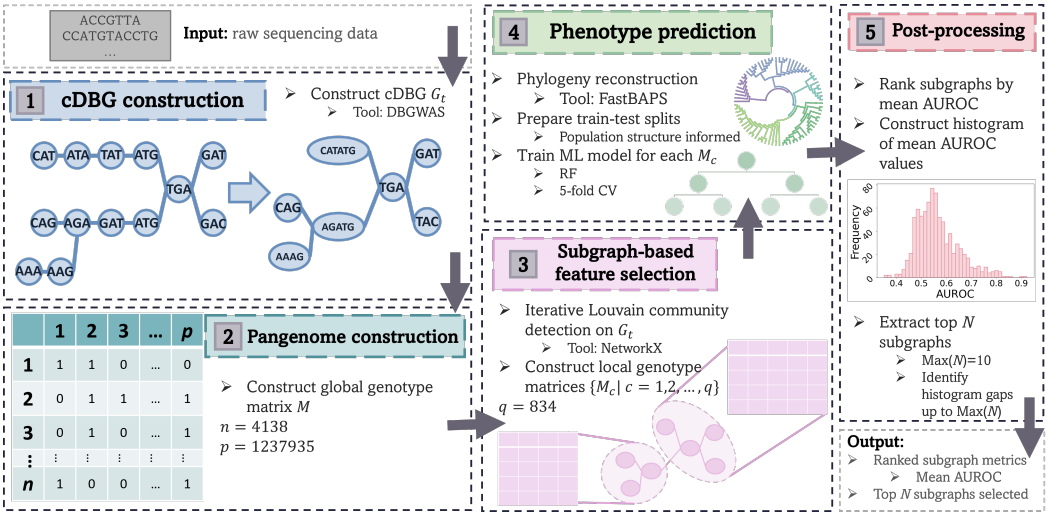

Figure 1: Overview of the proposed workflow for causal cDBG hotspot detection via prediction of binary phenotype measures of resistance to an antibiotic (binned MICs) from local genotype matrices with cDBG subgraph informed feature selection.

previously collected and analyzed by Wheeler et al. (2019), where approximately 98% of phenotypic variation may be explained by the set of known causal resistance mechanisms over 14 rare and common antibiotic agents (rates of resistance ranging from 0.05-97.54%). This collection has previously been used to demonstrate challenges associated with causal analysis in genotype-to-phenotype tasks for bacterial datasets (James et al., 2025). We note that we exclude two agents from our study: VAN (only 2 resistant samples belonging to the same subpopulation), and DAP (no ground truth).

We define the sets of potentially observable genotypes and phenotypes $\mathcal{X}_t$ and $\mathcal{Y}_t$, respectively, determined by the historical evolution processes up to time t. Let $\mathcal{A} = \{a_1, a_2, \ldots, a_{|\mathcal{A}|}\}$ denote the fixed panel of antibiotic agents under study. The 14 agents within our study, and their respective causal mechanisms present within the collection, are listed in Table 1.

The collection of *S. aureus* isolates define an empirical dataset $\mathcal{D} = \{(x_i, \mathbf{y}_i) | x_i \in \mathcal{X}_t, \mathbf{y}_i \in \mathcal{Y}_t, i = 1, 2, ..., n\}, n = 4138$, sampled from an unknown underlying distribution P_t over $\mathcal{X}_t \times \mathcal{Y}_t$. Each genotype $x_i \in \mathcal{X}_t$ is the complete genomic DNA sequence of isolate i, and its associated resistance-profile vector $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{i|\mathcal{A}|}) \in \mathcal{Y}_t$ records the resistance phenotypes for each antibiotic agent $a_\ell \in \mathcal{A}$. Projecting onto a single agent a_ℓ yields $\mathcal{D}^{(\ell)} = \{(x_i, y_{i\ell}) | x_i \in \mathcal{X}_t, y_{i\ell} \in \mathcal{Y}_t^{(\ell)}, i = 1, 2, ..., n\}$ of observed genotype-phenotype pairs. Although assay formats can differ (e.g., binary calls, S/I/R categories, MIC values, etc.), clinical laboratories typically map resistance to every antibiotic to the same categorical scale. Accordingly, we dichotomize each agent into susceptible (0) versus resistant (1), which in principle gives a joint phenotype space $\mathcal{Y}_t = \{0, 1\}^{|\mathcal{A}|}$. In practice, the primary focus is a single-agent setting, where projecting onto a single agent a_ℓ transforms resistance prediction into a binary classification problem, where $\mathcal{Y}_t^{(\ell)} = \{0, 1\}$.

We note that the set of data points are often prone to heavy biases due to environmental conditions (e.g., controlled laboratory settings) and sampling strategies.

4 Encoding the Genotype Data for Statistical Modeling

To facilitate analysis and modeling, appropriate representation spaces of embedded and encoded genotypes $\mathcal{R}$ and X, respectively, must be defined to prepare the genotype data $\{x_i\}$ for input into a specific ML model. In practice, genotype data is captured in the form of raw sequencing reads (e.g., from Illumina sequencing platforms), and undergoes assembly to construct a comprehensive pangenome that captures the full genetic variation of a bacterial population. We may now define a marker space $\mathcal{M}$ that serves as an intermediate representation to capture specific genetic markers

derived from the raw genotype data. These markers represent locatable sequences on the genome (e.g., k-mers, unitigs, genes, or gene clusters).

The space of encoded genotypes X prepared for input may be obtained by employing a hierarchical mapping framework of representation spaces summarized by a composite function ϕ (James et al., 2025),

$$\phi = \rho \circ \tau : \mathcal{X}_t \to \mathcal{M} \xrightarrow{\rho} \mathcal{R} \xrightarrow{\tau} X. \tag{1}$$

A fundamental requirement in designing the hierarchical mapping functions of the composite ϕ in Equation 1 is that they must each ***preserve and encode the core genetic signals underlying phenotype variation***.

When dealing with limited sample sizes, classical ML methods, such as linear regression, decision trees, and support vector machines, are generally preferable to deep learning methods as they require fewer data points to train effectively and are generally considered less susceptible to overfitting (Bashir et al., 2020). A genotype matrix representation strategy facilitates these modeling techniques in bGWAS, allowing for a compact, scalable, and flexible representation of genotype data by encoding each genotype in terms of a set of genetic features (e.g., k-mers, unitigs, genes, and gene clusters). As a result, it offers a robust choice for identifying genetic variants associated with phenotypic traits.

We follow this standard of ML-based bGWAS and adopt a genotype matrix representation for the proposed workflow in Figure 1, thereby ensuring compatibility with existing studies and pipelines.

4.1 cDBG CONSTRUCTION

Many short-read assembly algorithms use de Bruijn graphs (DBGs) to represent the genotype data – a graph structure where each node represents a k-mer, and edges connect nodes that overlap by $k-1$ bases. It is then possible to construct a compacted DBG (cDBG) by collapsing non-branching paths within a DBG, where each node represents a unitig – unique, contiguous sequences formed by aggregating overlapping k-mers. This process results in a more compact and scalable data structure for representing genomes, and a simple example is illustrated in Stage 1 of Figure 1: a DBG (left) of 3-mers, each adjacent pair of nodes overlapping by k-1 nucleotides, undergoes collapsing of non-branching paths into single nodes resulting in a cDBG (right).

The first mapping given in Equation 1 $\mathcal{X}_t \to \mathcal{M}$ allows the information contained within the set of isolates $\{x_i\}$ to be described by a set of p locatable genomic sequences $\mathcal{M} = \{m_j | j = 1, 2, ..., p\}$. This is achieved by converting the raw genomic data into a cDBG G_t, which we construct with DBGWAS (Jaillard et al., 2018), from which a set of p unitigs is derived. The benefit of collapsing overlapping k-mers into non-branching paths (unitigs), is a representation that scales better computationally while retaining many functionally important k-mer patterns in the genome. However, it inevitably discards any explicit record of the complete linear ordering and precise distances among these segments, introducing a controlled form of information loss.

Subsequently, $\mathcal{M} \xrightarrow{\rho} \mathcal{R}$ maps individual isolates x_i onto corresponding paths of finite length $\rho(x_i)$ through G_t, where the collection of observed paths define the representation space $\mathcal{R} = \{\rho(x_i) | i = 1, 2, ..., n\}$ of embedded genotypes. In this transformation, direct information may be lost on factors like unique structural rearrangements that might not be cleanly represented by path-based traversals if they involve highly complex variations. Still, adjacency within each path is preserved, so key local contiguities holding essential signals linking genotype to phenotype should be maintained for a given isolate.

Once the cDBG has been constructed, each unique unitig may be mapped to a gene via exact and near-exact string matching. Gene family sequences in FASTA format were obtained from the publicly available NCBI gene database and searched against (both forward and reverse compliment) the set of unitig sequences. Unitigs that contained a gene family sequence as a substring were labeled accordingly.

4.2 GENOTYPE MATRIX CONSTRUCTION

The final mapping in 1, $\mathcal{R} \xrightarrow{\tau} X$, maps each embedded genotype $\rho(x_i)$ into a binary p-dimensional vector $\phi(x_i) \in \mathbb{R}^p$ within the space $X = \{0,1\}^p$, having components,

$$\phi(x_i)_j = \begin{cases} 1 & \text{if } m_j \in \rho(x_i) \\ 0 & \text{otherwise} \end{cases} \quad \forall j \in \{1, 2, \ldots, p\}.$$

A genotype matrix $M \in \{0,1\}^{n \times p}$ may then be constructed to represent genotypes $\{x_i\}$ within empirical dataset $\mathcal{D}$ as Stage 2 of the workflow in Figure 1.

Matrix M therefore provides a structured framework for representing the set of n encoded genotypes $\{\phi(x_i)\}$ in terms of the set of p genetic markers $\mathcal{M}$. Within this format, each row $\phi(x_i)$ indicates the presence or absence across all p features – each unitig m_j – for a given isolate i.

The transformation $\tau(\rho(x_i)) = \phi(x_i)$, while drastically simplifying the data structure to facilitate large-scale association studies, results in the loss of the precise ordering and adjacency information inherent in $\rho(x_i)$, only retaining genetic signal with the presence or absence of specific genetic segments (unitigs).

5 BACTERIAL FINE MAPPING

5.1 TASK DEFINITION

Bacterial fine mapping may be defined by the task (James et al., 2025): ***Identify a subset of genomic markers $\mathcal{M}^* \subseteq \mathcal{M}$ that are the true causal variants influencing the phenotype.***

In ML-based bGWAS settings, this is approached through a primary goal of approximating the GP mapping with a predictive model. The true, and unknown, ground truth GP mapping function $\Theta : \mathcal{X} \to \mathcal{Y}$ defines the relationship that maps each genotype $x \in \mathcal{X}$ that may exist for $t \to \infty$, to a corresponding resistance profile $\mathbf{y} = \Theta(x) \in \mathcal{Y}$, with $y^{(\ell)} = \Theta^{(\ell)}(x) \in \mathcal{Y}^{(\ell)}$ for a single-agent setting.

The goal is then to build a predictive model $f^{(\ell)} : \mathcal{X}_t \to \mathcal{Y}_t^{(\ell)}$ that approximates the ground truth $\Theta^{(\ell)}$ using $\mathcal{D}^{(\ell)}$, such that $f^{(\ell)}(x) \approx \Theta^{(\ell)}(x)$ for all $x \in \mathcal{X}$. Fine-mapping then relies on the ability to interpret the model. While it is possible to extract learned parameters from a statistical model, the full function is only transparent for simpler models. Obtaining full transparency is often not possible for bGWAS due to the need for more complex models that can capture the nonlinear relationships that exist within bacterial genomes. Due to this, a more desirable goal is to approximate the GP mapping by focusing on causal markers $\mathcal{M}* \in \mathcal{M}$ alone, assuming a relationship $y^{(\ell)} = \Theta^{(\ell)}(x) \approx f^{(\ell)}(x|_{\mathcal{M}*})$.

5.2 REFINED FINE MAPPING TASK

Within the context of the proposed framework outlined in Figure 1, we consider an amended task: ***Identify a minimal set of subgraphs $\mathcal{S}^* \subseteq \mathcal{S}$ for which their corresponding marker subsets $\mathcal{M}_c$ best explain phenotypic variation.***

The set $\mathcal{S}^*$ of selected subgraphs represent subsets of unitigs that result in models that are the most predictive of a phenotype according to some performance metric (e.g., accuracy, AUROC), from the full set of non-overlapping subgraphs $\mathcal{S}$ derived from the cDBG representation for a bacterial pangenome. Each subgraph within the set $\mathcal{S}^*$ may therefore be considered to be of greater statistical significance, and by extension, contain a subset of more statistically significant features, suggesting the capturing of true causal variants within these neighborhoods. Here, we approximate the GP mapping using the set of features (nodes) $\mathcal{M}_c$ that are present within a single subgraph $g_c \in \mathcal{S}$, given by $y^{(\ell)} = \Theta^{(\ell)}(x) \approx f_c^{(\ell)}(x|_{\mathcal{M}_c})$.

6 CAPTURING SPATIAL CONTEXT WITH TOPOLOGICALLY-INFORMED FEATURE SELECTION

While supervised learning models can learn to capture interactions during the training process, the mappings themselves must prioritize the preservation of biologically relevant interactions. Given that the transformation $\tau(\rho(x_i))$ discards contextual information regarding the interactions and relationships between genetic variants on the genome, supervised models trained on matrix-formatted genotype data will be unable to distinguish between between true epistatic interactions that directly influence phenotypic traits and spurious correlations arising from LD. It is therefore crucial to mitigate information loss in the mapping $\tau(\rho(x_i))$ through explicit incorporation of domain knowledge.

We propose a feature selection framework that selects features corresponding to specific **subgraphs of the cDBG** – which encodes the **topological structure of the pangenome** – to incorporate the graph's embedded structural knowledge into the identification of the genetic features most relevant for phenotype prediction.

Our method is predicated on the assumption that spatial dependencies between variants exist within the genome, manifesting as **causal hotspots** where causal variants are concentrated within a few localized genomic regions defined in physical space, which is visualized by Wheeler et al. (2019) for a *S. aureus* protein. Extending this assumption from physical space to cDBG space, we posit that regions proximal in physical space are also localized within the cDBG. This framework assumes that epistatic effects are primarily confined to within local neighborhoods, enabling the capture of sufficient causal signals within individual subgraphs, while additive effects are distributed over more distant neighborhoods (across multiple subgraphs).

By partitioning the cDBG into subgraphs, the goal is to constrain the learning process by focusing solely on biologically relevant interactions between a reduced set of proximally close variants. This biologically informed paradigm enables extremely harsh feature selection, introducing LD control by removing redundant information and informing the model of higher priority interactions that would otherwise be overlooked due to the assumption of feature independence.

6.1 GENERATING LOCALIZED GENOTYPE MATRICES VIA CDBG-BASED PARTITIONING

Let G_t be the compacted de Bruijn graph (cDBG) built from the *S. aureus* pangenome. We partition its nodes (unitigs) into q non-overlapping communities $\mathcal{S} = \{g_1, \ldots, g_q\}$ such that each community g_c induces a disjoint marker subset $\mathcal{M}_c \subset \mathcal{M}$, and $\bigcup_{c=1}^{q} \mathcal{M}_c = \mathcal{M}$. Restricting the global genotype matrix M to the columns in $\mathcal{M}_c$ yields a *local* matrix M_c that is analyzed independently in downstream models (Stage 3, Figure 1).

We apply the Louvain community detection algorithm to a NetworkX (Hagberg et al., 2008) graph from the python-louvain PyPi package, using default resolution=1.0, to iteratively split and merge communities for partitioning G_t (Alg. 1, Appendix A) until every subgraph satisfies $1300 \leq |\mathcal{M}_c| \leq 1600$ *or* a maximum number of iterations (max_itr) is reached, producing $q = 834$ subgraphs. The interval was chosen to: **(i)** include enough unitigs (nodes) such that an entire gene may be comfortably captured within a single subgraph, **(ii)** maintain a dimensionality of $p_c < n$, and **(iii)** limit CPU/RAM demands on commodity hardware. The selected thresholds comfortably span the vast majority ($\approx 99.9\%$) of *S. aureus* genes, which average around 1kb in length (Méric et al., 2015).

Because many unitigs share identical presence/absence patterns, a genome-wide model can only assign importance to an arbitrary representative, obscuring biological interpretation. Constraining the feature space to a single cDBG neighborhood eliminates patterns that span distant loci, re-establishing a one-to-one correspondence between feature *subsets* and physical genomic context. Although the size constraint is agnostic to gene boundaries and may split some genes, testing under these stringent, quasi-arbitrary limits lets us evaluate whether cDBG-guided selection can still concentrate causal signal despite LD and collinearity. In effect, each subgraph acts as an LD-aware window that preserves positional information while providing a tractable, biologically informed reduction of the high-dimensional genotype matrix.

Antibiotic	Known mechanisms present in collection	# resistant subgraphs	Top N selected	# resistant in top N	Max. $\overline{\text{A}}_{c,\ell}$
GEN	*aaacA-aphD*	3	8	0	0.98
PEN	***blaZ***, *mecA*	13	3	1	0.92
MET	*mecA*	4	10	0	0.96
FUS	***fusA****, *fusB*†, ***fusC***	3	4	2	0.85
TEI	*vanA*†, *vanZ*	11	10	0	0.86
ERY	*msrA, ermA,* ***ermB***, ***ermC***	9	5	2	0.92
CLI	*linA, ermA, ermB,* ***ermC***	8	2	1	0.83
MUP	*ileS**, ***ileS-2***	8	3	2	0.88
LIN	*rplC, 23S*†	3	2	0	0.69
CIP	***grlA****, ***grlB****, ***gyrA***, ***gyrB***	4	6	2	0.87
RIF	***rpoB****	1	3	1	0.98
MIN	***tetM***	4	7	1	1.0
TET	***tetK***, *tetL, tetM*	8	10	1	0.84
TMP	*dfrA**, *dfrG, dfrB*†	6	4	0	0.83

Table 1: Predicting antibiotic resistant sites in the *S. aureus* cDBG. To distinguish between causal mechanisms that are due to an acquired gene vs. individual unitig mutations, $*$ indicates causal mutations. Genes that are not labeled within the dataset (for providing the ground truth) are indicated with †. Bold genes indicate those that have corresponding nodes present within the top N subgraphs.

7 Empirical analysis

For each of the 14 antibiotic agents $a_\ell \in \mathcal{A}$, we fit multiple independent random-forest (RF) classifiers using scikit-learn (Pedregosa et al., 2011) to predict binary resistance phenotypes $\{y_i^{(\ell)}\}$ from a local genotype M_c . Due to incomplete resistance measurement coverage for 7 of the 14 documented antibiotics, sample sizes vary from 908 to 4138 isolates. Specifically, for each drug a_ℓ, we defined 5-fold cross-validation splits and selected a hyperparameter configuration h_ℓ via "GridSearchCV". A RF $f^{(\ell)}(x|_{\mathcal{M}_c})$ was then trained on each local genotype matrix M_c independently to predict $y^{(\ell)}$, fixing hyperparameters to h_ℓ. Further details on training procedures are outlined in Appendix B) for Stage 4 of Figure 1.

We note that although we train RF classifiers on phenotype prediction, the methodology in Figure 1 is not limited to classification tasks or RF models, and may be amended for training alternative models instead depending on the desired goals of the study.

We evaluate the performance of each model $f^{(\ell)}(x|_{\mathcal{M}_c})$ on the classification of resistance to each antibiotic agent a_ℓ, corresponding to Stage 5 in Figure 1. Each model $f^{(\ell)}(x|_{\mathcal{M}_c})$ has a corresponding score $\overline{\text{A}}_{c,\ell}$, determined by taking the mean of the AUROC values over each of the 5 folds, which we use for evaluation. As there are a total of q subgraphs, a total set of q scores is produced for each antibiotic a_ℓ, where each score $\overline{\text{A}}_{c,\ell}$ indicates the predictive performance for the model trained on the specific subset of unitigs p_c contained within M_c, representing the information within a localized genomic region across genotypes $\{x_i\}$.

7.1 Ranking cDBG subgraphs for causal hotspot detection

To compare across subgraph models for a given phenotype and determine which subgraphs may be predicted as high risk, we rank subgraphs according to their respective $\overline{\text{A}}_{c,\ell}$. As we expect the causal signal to be concentrated within just a few subgraphs – confirmed by our ground truths ranging from 1-13 subgraphs for each phenotype – we also expect this to be reflected within $\overline{\text{A}}_{c,\ell}$ values as gaps in the distribution, where we predict resistance mechanisms with causal signal concentrated to fewer subgraphs results in a more distinct separation in $\overline{\text{A}}_{c,\ell}$ scores between resistant and non-resistant sites. Based on this, we select N top subgraphs for a given phenotype based on $\overline{\text{A}}_{c,\ell}$ scores, with an upper limit of $N = 10$. This choice corresponds to a maximal selection of $\approx 1.2\%$ of the full set of unitigs, depending on node distribution over the selected subgraphs.

To determine the number of top subgraphs N selected for downstream analysis, we implemented a function that constructs a histogram over all $\overline{\text{A}}_{c,\ell}$. Bins with zero observations indicate a gap in the distribution, and the function identifies the rank of the subgraph that marks this discontinuity. Specifically, if a bin is empty, the algorithm locates the subgraph with its $\overline{\text{A}}_{c,\ell}$ closest to the upper boundary of that gap and considers its index as a potential cutoff only if it is within the top 10 subgraphs. The maximum of these is used to set N. The default of $N = 10$ is set for instances where no discontinuity in the distribution exists.

7.2 BASELINES

To provide a baseline for comparison and demonstrate biological relevance of candidates, we train a RF classifier on a subset of features. This dataset was constructed from the set of all M_c containing unitigs labeled as causal according to our ground truth across all 14 antibiotics, ensuring complete coverage of causal signal for each drug. The resultant dataset contained 95630 unitigs (7.7% of the nodes in the cDBG G_t) over 64 subgraphs.

Additionally, Wheeler et al. (2019) provide a baseline for comparison with traditional bGWAS methods and highlight the causal mechanisms identified with them.

8 RESULTS

(i) cDBG-guided feature selection reveals genomic subgraphs linked to resistance with machine learning.

The predictive performance across subgraphs were ranked (with highest $\overline{\text{A}}_{c,\ell}$ assigned rank 1) for each antibiotic a_ℓ, by assigning a rank to each subgraph $g_c \in \mathcal{S}$ based on corresponding $\overline{\text{A}}_{c,\ell}$. Extracting the top N performing subgraphs for each drug reveals that selecting features based on the cDBG structure returns high predictive performances for significantly reduced feature sets.

The top $\overline{\text{A}}_{c,\ell}$ values ($N = 1$) for each drug are given in Table 1, where the top ranking subgraph for 13 out of the 14 total drugs returned a value of $\overline{\text{A}}_{c,\ell} \geq 0.83$ when trained on features sets containing between 1435 to 1671 unitigs.

The total number of unitigs selected within the top N subgraphs for each of the 14 drugs resulted in 2860 to 15858 candidate nodes being selected, corresponding to a unitig selection rate between 0.23% and 1.28%. Within the total set of top N subgraphs selected within each of the 14 antibiotic-specific studies, mechanisms associated with a total of 9 drugs were uncovered. Within these, previously undiscovered (with traditional bGWAS) resistance mechanisms to 3 drugs were captured within the subgraph feature selection framework (*ermB* for ERY, *ermC* for CLI, and *gyrB, grlB* for CIP). It was also stated that performance was worst for predicting ERY and FUS resistance with an elastic net, where multiple gene variants contribute to resistance. This was not reflected within our subgraph-based approach, where resistant sites were uncovered for *fusA, fusC* and *ermB, ermC*, suggesting that a subgraph-based feature selection approach can aid in identifying resistance mechanisms that are the product of multiple interacting variants where traditional methods struggle. Additionally, it may be used in combination with traditional methods to increase confidence in predictions.

(ii) Focused cDBG-subgraph feature selection minimizes spurious associations while retaining causal variants.

By comparing the performance along with the sets of selected candidate features between those selected via a subgraph-based approach and those selected within the baseline model, we can demonstrate the ability to focus on the true causal variants by successfully disentangling causal from spurious features across multiple resistance studies. The baseline model contained a reduced feature set (7.7% of the unitigs), while the subgraph-based approach considered complete genetic variation. Despite the reduced feature set, the baseline returned sets of candidate nodes (features having non-zero importance) that made up between 4.1% and 24.6% of the limited set, with representation in all 64 of the subgraphs that were selected to construct the dataset across all 14 antibiotic analyses, compared to the 1-10 subgraphs selected out of the 834 total in the subgraph approach.

Assessing the number of true discoveries for each method (causal variants correctly selected as candidates), the baseline returned a higher value of true discoveries for 9 of the 14 antibiotics, and the remaining 5 had a greater number of true discoveries with the subgraph approach, however it should be noted that the number of false discoveries is greatly inflated for the baseline. Notable differences were seen for resistance predictions in CIP, where the subgraph-based method uncovered 152 true causal variants, compared to 43 in the baseline, and RIF, where the subgraph method correctly discovered all 5 of the known causal variants compared to just 1 in the baseline. Other than resistance to the drug LIN, these were the only other two antibiotics that are solely the result of causal mutations, rather than acquired genes.

(iii) AUROC distributions across local cDBG models reflect confidence in predicted variants.

In bGWAS, the reliability of selected candidates may be judged according to the signal distribution over genetic loci positions in the genome, where a more condensed signal peak is considered a stronger indication that candidates correspond to true causal variants (Wheeler et al., 2019). This however relies upon identifying the corresponding location of each genetic feature on the genome prior to establishing a level of confidence in results. By selecting features based on cDBG structure, this signal peak analysis is inherently captured within the individual $\overline{\mathrm{A}}_{c,\ell}$ scores due to features being derived from subgraphs in the cDBG, where a smaller value of N combined with higher $\overline{\mathrm{A}}_{c,\ell}$ values indicate a stronger and more localized causal signal, which in most cases would suggest more trustworthy candidates. Phenotypes indicating a stronger and more localized signal include PEN and RIF, both of which involved discovery of a corresponding resistant site. Phenotype studies where 10 subgraphs were selected, indicating no discernible gap in the histogram, include MET, TEI, and TET, for which only a single resistance site was uncovered for TET alone.

9 Conclusion

This paper presents an approach to bacterial fine mapping that derives structural knowledge from graph structured genetic data to address multicollinearity and high-dimensionality in genetic cohorts. We discover that selecting features based on localized subgraphs within a cDBG provided additional control for confounding effects due to ancestry, specifically those manifesting as genome-wide LD, combined with assigning biological relevance to features as an explicit method for distinguishing between correlated variables for uniquely identifying relevant genetic marker sequences. Constraining the function space through the injection of structural information within the feature selection stage enabled the successful identification of multiple resistance causing loci in *S. aureus* with low counts of false discoveries. We also observed that resistance mechanisms that are the product of multiple interacting variants are identifiable using our proposed approach. Overall, this paper highlights the potential of using structural information to control for population structure and also enhance interpretability for more focused bacterial fine mapping for some resistance profiles in a limited setting. A complete description of scenarios that enable the uncovering of underlying causal mechanisms remain unclear, requiring a more thorough exploration of the behavior of causal traits for resistance phenotypes where deconfounding efforts demonstrate a weakened ability in disentangling causal and spurious features.

References

Quantitative measurement of antibiotic resistance in mycobacterium tuberculosis reveals genetic determinants of resistance and susceptibility in a target gene approach. *Nature communications*, 15(1):488, 2024.

Daniel Bashir, George D Montañez, Sonia Sehra, Pedro Sandoval Segura, and Julius Lauw. An information-theoretic perspective on overfitting and underfitting. In *AI 2020: Advances in Artificial Intelligence: 33rd Australasian Joint Conference, AI 2020, Canberra, ACT, Australia, November 29–30, 2020, Proceedings 33*, pp. 347–358. Springer, 2020.

Matthew B Biggs, Kelly Craig, Esther Gachango, David Ingham, and Mathias Twizeyimana. Genomics-and machine learning-accelerated discovery of biocontrol bacteria. *Phytobiomes Journal*, 5(4):452–463, 2021.

Julio Diaz Caballero, Shawn T Clark, Pauline W Wang, Sylva L Donaldson, Bryan Coburn, D Elizabeth Tullis, Yvonne CW Yau, Valerie J Waters, David M Hwang, and David S Guttman. A genome-wide association analysis reveals a potential role for recombination in the evolution of antimicrobial resistance in burkholderia multivorans. *PLoS pathogens*, 14(12):e1007453, 2018.

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.

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.

Jason C Hyun, Jonathan M Monk, Richard Szubin, Ying Hefner, and Bernhard O Palsson. Global pathogenomic analysis identifies known and candidate genetic antimicrobial resistance determinants in twelve species. *Nature Communications*, 14(1):7690, 2023.

Magali Jaillard, Maud Tournoud, Leandro Lima, Vincent Lacroix, Jean-Baptiste Veyrieras, and Laurent Jacob. Representing genetic determinants in bacterial gwas with compacted de bruijn graphs. *bioRxiv*, pp. 113563, 2017.

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.

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.

Sudaraka Mallawaarachchi, Gerry Tonkin-Hill, Nicholas J Croucher, Paul Turner, Doug Speed, Jukka Corander, and David Balding. Genome-wide association, prediction and heritability in bacteria with application to streptococcus pneumoniae. *NAR genomics and bioinformatics*, 4(1): lqac011, 2022.

Guillaume Méric, Maria Miragaia, Mark de Been, Koji Yahara, Ben Pascoe, Leonardos Mageiros, Jane Mikhail, Llinos G Harris, Thomas S Wilkinson, Joana Rolo, et al. Ecological overlap and horizontal gene transfer in staphylococcus aureus and staphylococcus epidermidis. *Genome biology and evolution*, 7(5):1313–1328, 2015.

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.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

Jody Phelan, Francesc Coll, Ruth McNerney, David B Ascher, Douglas EV Pires, Nick Furnham, Nele Coeck, Grant A Hill-Cawthorne, Mridul B Nair, Kim Mallard, et al. Mycobacterium tuberculosis whole genome sequencing and protein structure modelling provides insights into anti-tuberculosis drug resistance. *BMC medicine*, 14:1–13, 2016.

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

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.

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.

Ming-Ren Yang and Yu-Wei Wu. A cross-validated feature selection (cvfs) approach for extracting the most parsimonious feature sets and discovering potential antimicrobial resistance (amr) biomarkers. *Computational and Structural Biotechnology Journal*, 21:769–779, 2023.

A SUBGRAPH GENERATION

Algorithm 1: Generate Non-Overlapping Subgraphs

Input: A graph G_t, minimum subgraph size threshold *min_size*, maximum subgraph size threshold *max_size*, maximum iterations *max_itr*

Output: A set of non-overlapping subgraphs $\mathcal{S}$

Procedure `generate_subgraphs` (G_t, min_size, max_size, max_itr):
- **init**
  - $\mathcal{S} \leftarrow G_t$
  - $itr = 0$
- **while** $itr < max\_itr$ **do**
  - **while** *any(*size$(c) > max_size$ for all c in $\mathcal{S}$*)* **do**
    - Update $\mathcal{S} \leftarrow$ `split_large_communities` ($\mathcal{S}$)
    - Update $\mathcal{S} \leftarrow$ `merge_small_communities` ($\mathcal{S}$)
    - Update $itr \leftarrow itr + 1$
- **return** $\mathcal{S}$

Function `split_large_communities` ($\mathcal{S}$, max_size):
- **init**
  - $\mathcal{S}_{\text{split}} \leftarrow \{c \in \mathcal{S} \mid \text{size}(c) < max\_size\}$
  - $\mathcal{S}_{\text{large}} \leftarrow \{c \in \mathcal{S} \mid \text{size}(c) > max\_size\}$
- **while** $\mathcal{S}_{large} \neq \emptyset$ **do**
  - **foreach** c in $\mathcal{S}_{\text{large}}$ **do**
    - Pop c from $\mathcal{S}_{\text{large}}$
    - $partitions \leftarrow$ Partition c using the Louvain method
    - **foreach** p in $partitions$ **do**
      - **if** size$(p) < max_size$ **then**
        - Add p to $\mathcal{S}_{\text{split}}$
      - **else**
        - Add p to $\mathcal{S}_{\text{large}}$
- **return** $\mathcal{S}_{\text{split}}$

Function `merge_small_communities` ($\mathcal{S}$, min_size):
- **init**
  - $\mathcal{S}_{\text{small}} \leftarrow \{c \in \mathcal{S} \mid \text{size}(c) < min\_size\}$
- **while** $\mathcal{S}_{\text{small}} \neq \emptyset$ **do**
  - **foreach** s in $\mathcal{S}_{\text{small}}$ **do**
    - Pop s from $\mathcal{S}, \mathcal{S}_{\text{small}}$
    - $neighbors \leftarrow \{c \in \mathcal{S} \mid$ edge exists between c and $s\}$
    - $n \leftarrow \arg\min_c(size(c) \mid c \in neighbors)$
    - Update $n \in \mathcal{S} \leftarrow \text{merge}(n, s)$
  - Update $\mathcal{S}_{\text{small}} \leftarrow \{c \in \mathcal{S} \mid \text{size}(c) < min\_size\}$

return $\mathcal{S}$

B RESISTANCE CLASSIFICATION

Each local genotype matrix M_c was used to train a corresponding RF model $f^{(\ell)}(x|_{\mathcal{M}_c})$ configured with the same set of hyperparameters h_ℓ. Hyperparameters were initially selected using 5-fold CV for a subset of 6 randomly sampled matrices M_c, confirming robustness across models to variations in h_ℓ. The final set of hyperparameters were fixed according to the modal parameters across

the set of representative models as: 'bootstrap': True, 'max_depth': None, 'min_samples_leaf': 4, 'min_samples_split': 2, 'n_estimators': 200. We also used the 'balanced' mode to compute sample weights and adjust them according to class frequencies in the input data.

For model evaluation, we adopted a 5-fold CV strategy with validation sets determined by subpopulation assignments. Our method is a more limited version of a leave-one-strain-out method (Lees et al., 2020), in favor of reduced computation over a more comprehensive evaluation. Cluster membership for each sample was derived from a phylogenetic reconstruction performed with the software package FastBAPS (Tonkin-Hill et al., 2019). Validation sets were then constructed using entire clusters, each cluster being assigned to the validation set of one of the 5 folds, resulting in variable sizes. Assignments were managed to ensure a fair and systematic distribution of samples across each fold regarding cluster size and resistance class representation across both training and test sets, and were uniquely determined for each drug.

C LIMITATIONS

Louvain partitioning. We partition according to to Louvain community detection algorithm with subgraph size constraints, however alternative sizes or community detection frameworks may be more suitable depending on the underlying causal structure for resistance. For example, better management for structures with high causal spread. We also don't consider distances between subgraphs, treating them each as independent entities meaning relationships spanning across borders are excluded from the model's function space.

Strain representation. Strains are treated independently, however their exist relationships on the strain level as well as on the sample level. As a result, certain strains may be more predictive of others, which we don't account for when dividing strains between validation sets. This may reflect in variable and biased mean AUROC values ($\overline{\mathrm{A}}_{c,\ell}$). As we consider the mean across folds, we do not account for biases due to genetic relatedness between strains.

Feature invariance. Statistical models and strain-based CV assume feature invariance, where the effect of causal features on the phenotype is assumed to remain consistent across all environments (strains). Biological mechanisms don't always conform to this rule, so ignoring the possibility of environmental dependencies along with certain types of interactions such as suppressor epistasis, where the presence of a non-causal variant may mask the effect of a causal variant on the phenotype.