Journal Title Here, 2022, pp. 1–8

doi: DOI HERE
Advance Access Publication Date: Day Month Year
Paper

PAPER

Open Problem: Causal Variant Detection with Machine Learning for Bacterial Genomics

First Author,[1,*] Second Author,[2] Third Author,[3] Fourth Author[3] and Fifth Author[4]

[1]Department, Organization, Street, Postcode, State, Country, [2]Department, Organization, Street, Postcode, State, Country, [3]Department, Organization, Street, Postcode, State, Country and [4]Department, Organization, Street, Postcode, State, Country
[*]Corresponding author. email-id.com
FOR PUBLISHER ONLY Received on Date Month Year; revised on Date Month Year; accepted on Date Month Year

Abstract

How can we identify causal genetic mechanisms that govern bacterial traits? Initial efforts entrusting machine learning models to handle the task of predicting phenotype from genotype return high accuracy scores. However, attempts to extract any meaning from the predictive models are found to be corrupted by falsely identified "causal" features. Relying solely on pattern recognition and correlations is unreliable, significantly so in bacterial genomics settings where high-dimensionality and spurious associations are the norm. Though it is not yet clear whether we can overcome this hurdle, significant efforts are being made towards discovering potential high-risk bacterial genetic variants. In view of this, we set up open problems surrounding phenotype prediction from bacterial whole-genome datasets and extending those to learning causal effects, and discuss challenges that impact the reliability of a machine's decision-making when faced with datasets of this nature.

Key words: keyword1, Keyword2, Keyword3, Keyword4

Introduction

The goal of bacterial genome-wide association studies (bGWAS) is to identify genetic variants that influence a trait or phenotype. These studies traditionally employ statistical methods to perform population genomic analyses to yield a list of candidate genes or genetic markers associated with a phenotype, and have been a significant contributor in uncovering numerous genetic loci that are causally related to a phenotype, e.g., resistance to an antibiotic. Improvements in whole-genome sequencing techniques have led to the generation of increasing amounts of data, creating an impracticality surrounding functional investigations of all loci individually. However, this up-scaling has lead to the prediction of a greater number of significantly associated loci despite efforts to minimize false discovery rate.

Machine learning (ML) algorithms are an obvious successor to bGWAS that may more effectively find signal in genetic noise. To date, existing algorithms have been applied to the data with little to no adaptation. Researchers are finding that these ML models fail to reliably generalize to out-of-distribution examples (Chalka et al. [2023], Hu et al. [2024]), and frequently identify false positive associations (Pearcy et al. [2021]). In addition, they have found that removing all known causal variables from a model does not meaningfully impact model accuracy (Nguyen et al. [2018]). Accordingly, it is crucial to understand the nuances of both genomic data and ML techniques to properly inform analyses, enabling them to mitigate effects that arise from model- and data-specific constraints, for integration with ML pipelines for causal variant detection.

Predictive models do not inherently distinguish between causative and associative relationships. So we raise the question: what course of action should we take to develop a pipeline capable of distinguishing between causal and spurious features in bacterial genomic data? In this paper we present open problems in applying ML to phenotype prediction, with the goal of uncovering causal variants in bacterial datasets. We also discuss challenges specific to the intersection of causal ML and bacterial genomics, and use a dataset of 4,140 *Staphylococcus aureus* isolates to illustrate these challenges with examples.

Genotype-to-phenotype mapping

We first set up the problem of linking genotype to phenotype under the assumption of fixed environmental conditions.

Consider the sets of all biologically feasible genotypes and their corresponding phenotypes collected in the genotype and phenotype spaces $\mathcal{X}$ and $\mathcal{Y}$, respectively. The set $\mathcal{X}$ is finite (although vast due to the combinatorial nature of genetic variations), while $\mathcal{Y}$ can be finite or infinite (even a continuum). The sets represent an idealized, comprehensive view of all possible genetic and phenotypic variations within a bacterial population, reflecting both observed and unobserved

possibilities. In nature, these spaces are constrained by biological and evolutionary limitations, with less fit genotypes being under-represented or absent.

We assume the existence of a ground truth genotype-phenotype (GP) mapping function $\Theta : \mathcal{X} \to \mathcal{Y}$ assigning to each genotype $x \in \mathcal{X}$ the corresponding phenotype $y = \Theta(x) \in \mathcal{Y}$.

At a given time t, the potentially observable genotype and phenotype sets $\mathcal{X}_t \subseteq \mathcal{X}$ and $\mathcal{Y}_t \subseteq \mathcal{Y}$, respectively, are determined by the historical evolution process up to time t. In laboratory settings, the genotype space is further constrained by the experimental design, which can limit or direct the evolutionary process by controlling environmental conditions, selecting specific genotypes, or applying selective pressures.

One may collect an *empirical dataset* $\mathcal{D} = \{(x_i, y_i) | x_i \in \mathcal{X}_t, y_i \in \mathcal{Y}_t, i = 1, 2, ..., n\}$ of n observed genotype-phenotype pairs by sampling from an unknown underlying distribution P_t over $\mathcal{X}_t \times \mathcal{Y}_t$, using a sampling strategy that heavily biases the resultant set of data points. The distribution P_t is dynamic, continuously evolving due to genetic drift, environmental pressures, and other evolutionary processes. Evolutionary processes lead to the formation of distinct genetic clusters, or strains, within the population. Different clusters exhibit significant divergence in genetic and phenotypic traits, especially in isolated environments, leading to P_t with significant multi-modal structures. Note that due to possible "observational noise" the conditional distribution $P_t(y|x)$ over phenotypes $y \in \mathcal{Y}_t$, given a genotype $x \in \mathcal{X}_t$, may not be fully concentrated on the ground truth $\Theta(x)$. Observational noise may refer to variability in the phenotypes caused by measurement errors or epigenetic factors.

Our goal is to build a *predictive model* $f : \mathcal{X} \to \mathcal{Y}$ approximating Θ using the empirical dataset $\mathcal{D}_t \subset \mathcal{X}_t \times \mathcal{Y}_t$, such that $f(x) \approx \Theta(x)$ for all $x \in \mathcal{X}$. The challenge is that we have at our disposal only a *limited sample* from the subset $\mathcal{X}_t \times \mathcal{Y}_t$ of the full space $\mathcal{X} \times \mathcal{Y}$ over which we would like to build the predictive model. An additional challenge is posed by the need to represent the input genotypes x and output phenotypes y in a form suitable for the machine learning technique employed. Assume that our parametrized machine learning predictive model (with parameters $w \in W$) takes as inputs and outputs elements of the input and output spaces X and Y, respectively. The genotypes $x \in \mathcal{X}$ are represented through their feature representations $\phi(x) \in X$, $\phi : \mathcal{X} \to X$. Analogously, the phenotypes $y \in \mathcal{Y}$ are represented through an invertible map $\psi : \mathcal{Y} \to Y$. The genotype feature representations $\phi(x)$ may not contain all the relevant information for the phenotype prediction. Due to these challenges, an additional set of potentially informative covariate variables $z \in Z_t$ is introduced to form extended inputs (x, z) to the parametrized machine learning model $F : X \times Z \times W \to Y$. **Examples of covariates - fixed environment.** Hence, the phenotype prediction for a genotype $x \in \mathcal{X}$ with additional covariates $z \in Z_t$ is realized through the parametrized machine learning model (with parameter values $w \in W$) as

$$\hat{y} = \psi^{-1}(F(\phi(x), z; w)).$$

Representation spaces

To facilitate analysis and modeling, the actual genotype space $\mathcal{X}_t$ must first be mapped into a suitable representation space $\mathcal{R}$. This chosen representation space $\mathcal{R}$ can be any mathematical structure appropriate for representing genotype data, such as graphs, tensors, matrices or sequences. A marker space $\mathcal{M}$ serves as an intermediate representation that captures specific genetic markers derived from the raw genotype data that are locatable sequences on the genome (e.g., SNPs, k-mers, unitigs, or genes etc) through the processing of raw genetic data (i.e., sequencing or genotyping, alignment, variant calling, annotation and marker extraction).

We define a general embedding function $\rho : \mathcal{M} \to \mathcal{R}$ which maps the marker space $\mathcal{M}$ into a suitable representation space $\mathcal{R}$ of **embedded features** that facilitates analysis and modeling, allowing each input genotype $x \in \mathcal{X}_t$ to be represented as $\rho(x)$ in the new space $\mathcal{R}$, $\rho : \mathcal{X}_t \to \mathcal{R}$. The embedding function ρ should preserve relevant genotypic properties.

Finally, the model representation space X defines the space of **encoded features** prepared for input into the particular ML model. This is represented by a map $\tau : \mathcal{R} \to X$, giving us the genotype representation $\phi = \rho \circ \tau : \mathcal{X}_t \to X$. As outlined above, the genotype features may be extended with additional informative co-variates $z \in Z$ which may capture nonlinear relationships and interactions that are not explicitly included in representations $\phi(x) \in X$.

We may summarize the composite function ϕ by a hierarchical mapping framework of representation spaces,

$$\mathcal{X}_t \to \mathcal{M} \xrightarrow{\rho} \mathcal{R} \xrightarrow{\tau} X.$$

Task 1

Given an extended training set $\tilde{\mathcal{D}} = \{(x_i, z_i, y_i) |\ x_i \in \mathcal{X}_t, z_i \in Z_t, y_i \in \mathcal{Y}_t, i = 1, 2, ..., n\}$, the training process (adaptation of the parameters w) can be posed as an optimization problem. The task may be defined as ***optimize the parameters*** w ***by solving optimization problem***:

$$w_{\text{opt}} = \arg\min_{w} \sum_{i=1}^{n} \ell(\psi^{-1}(F(\phi(x_i), z_i; w)), y_i), \qquad (1)$$

determined by an appropriate loss function $\ell : \mathcal{Y}_t \times \mathcal{Y}_t \to \mathbb{R}_{\geq 0}$.

Genetic fine mapping

In bacterial fine mapping, the goal is to identify **causal genetic variants** – specific DNA sequence alterations within a population that directly influence phenotypic traits. This involves distinguishing true causal variants from those merely in linkage disequilibrium with causal ones and spuriously associated with the phenotype.

Task 2

Identify a subset of genomic markers $\mathcal{M}^* \subseteq \mathcal{M}$ ***that are the true causal variants influencing the phenotype.***

Given that the learning of the GP mapping is well-posed, the goal is to optimally disentangle causal from non-causal features such that the true causal variants may be identified. Since the GP task on its own may not be specific enough to pinpoint the causal features, additional measures to constrain the feature space may need to be deployed. For example, Aliee et al. [2023] suggest a deep neural network with architecture and loss functions specifically informed by causality analysis.

Besides direct task-driven architectural and loss function manipulation, additional regularization pressure R_{causal} may be introduced (e.g. in an additive form) to help zoom on the causal features. For example, denoting the genotype encoding mapping using marker set $\mathcal{N} \subseteq \mathcal{M}$ by $\phi(x; \mathcal{N})$, the size of the marker set by $|\mathcal{N}|$, and a class of ML models input-compatible

with this encoding by $\mathcal{F}_{\phi(\cdot;\mathcal{N})}$, we can express the overall optimization problem as:

$$\mathcal{M}^* = \arg \min_{|\mathcal{N}|;\mathcal{N}\subseteq\mathcal{M}} \Big[\min_{F\in\mathcal{F}_{\phi(\cdot;\mathcal{N})}} \sum_{i=1}^{n} \ell(\psi^{-1}(F(\phi(x_i;\mathcal{N}), z_i)), y_i) + \lambda R_{causal}(F) \Big]. \quad (2)$$

Open problems

Given the optimization problem 1, the goal is to find the parameter set w_{opt} that minimizes the total loss over the training data. The function $F(\phi(x_i), z_i; w))$ produces some intermediate output based on the transformed input $\phi(x_i)$, covariates z_i, and the parameters w, in the phenotype representation space Y. The inverse mapping ψ^{-1} is then applied to this output to obtain a prediction $\hat{y}_i$ before computing the loss with respect to the true target y_i. In some cases the map ψ can be simply be the identity map.

The core issue arises because the mapping from the parameter space to the space of predictive models F may not be injective. This means that this mapping mapping is many-to-one and the desired parameter setting is not uniquely identifiable. However, even if the mapping from the parameter space to the space of predictive models was injective, because our training sample is finite and sparse, several parameter settings w can lead to exactly (or almost exactly) the same phenotype predictions on the sample genotypes. Hence, multiple parameter settings $w \in W$ can produce the same predictions $\hat{y}_i \in \mathcal{Y}_t$ on the training data. Consequently, **the inverse mapping from predicted phenotypes back to parameters is one-to-many**: different parameter settings w can lead to the same predicted phenotype $\hat{y}_i$ for given inputs $\phi(x_i)$ and z_i.

As a result, the model cannot uniquely determine the parameters w that map the input genotypes $\phi(x_i)$ to the observed phenotypes y_i. This non-uniqueness manifests as multiple solutions to the optimization problem, where different parameter sets minimize the loss function equally well, rendering the problem ill-posed. This can have serious consequences for approaches that detect "important" input features (e.g. causal variants) based on the learnt parameter vector w (e.g. Automatic Relevance Determination or Matrix Relevance Learning (Wipf and Nagarajan [2007], Schneider et al. [2010])).

Open problem 1

OP1.A - *Is it possible to reformulate the genotype-to-phenotype mapping task to be well-posed?*

Learning of the genotype-to-phenotype mapping is inherently ill-posed (when unconstrained), so in order to render the task as well-posed, we need to address the violations of Hadamard's criteria – existence, uniqueness, and stability – by introducing additional constraints and modifications to the problem structure.

Given that there exists a "ground truth" formulation of the GP mapping that is unknown to us, we further ask:

OP1.B - *What are the necessary conditions and representations required to achieve a well-posed task?*

To make learning of the GP mapping well-posed, additional constraints must be employed in the formulation of data representations (mappings ϕ and ψ) and of the predictive model $F(\cdot; w)$ itself. These aspects can involve e.g. feature selection, dimensionality reduction techniques, and incorporation of domain knowledge. One may attempt to enforce the constraints, for example, through additive regularization

$$\min_{F\in\mathcal{F}_{\text{constrained}}} \sum_{i=1}^{n} \ell(\psi^{-1}(F(\phi(x_i), z_i; w)), y_i) + \lambda R(F), \quad (3)$$

of the core optimization problem (1). Here $\mathcal{F}_{\text{constrained}}$ is a restricted function space incorporating prior knowledge and potentially of reduced complexity compared to $\mathcal{F}$, $\lambda \geq 0$ is a regularization parameter that controls the trade-off between fitting the data and satisfying the regularization term, and $R(\cdot)$ is a regularization functional that imposes additional constraints on F.

Open problem 2

Does there exist a well-posed genotype-to-phenotype mapping that also satisfies requirements for bacterial fine mapping? And if so, what are the necessary conditions and representations to achieve this?

Challenges in genotype-to-phenotype prediction

Assessing the well-posedness of predicting bacterial phenotypes from genotypes involves examining whether the optimization problem defined in Equation 1 meets Hadamard's criteria. According to Hadamard's criteria, a problem is well-posed if a solution **exists**, is **unique**, and depends continuously on the input data (is **stable**)(Hadamard [2014]).

In the context of bacterial genotype-to-phenotype mapping, these criteria are often unmet, rendering the problem **ill-posed**. A primary reason for this ill-posedness is the nature of the inverse mapping from phenotype predictions $\hat{y}$ to model parameters w, which is inherently one-to-many. The one-to-many relationship arises due to several factors: limited sampling in the dataset, information loss in feature representations, unmeasured confounders and observational noise, and model complexity coupled with function space limitations. Each of these factors contributes to the ambiguity and uncertainty in accurately mapping genotypes to phenotypes using machine learning approaches.

To elucidate these challenges, we examine an example real dataset: the *Staphylococcus aureus* pangenome, representing the empirical dataset $\mathcal{D}_t$. This dataset encompasses genetic variation across the entire genome $\mathcal{X}_t$ and associated phenotypes $\mathcal{Y}_t$, specifically binary measures of antibiotic resistance across 16 different drugs. While our analysis focuses on this collection of isolates, the identified challenges are broadly applicable to various bacterial species, albeit to differing extents. Additional analyses have been previously detailed (Wheeler et al. [2019]), providing a foundation for our discussion.

Limited sampling in the dataset

The empirical dataset $\mathcal{D}_t$, from which we construct the extended training set $\tilde{\mathcal{D}}$, is a limited and potentially biased sample from the underlying distribution P_t over $\mathcal{X}_t \times \mathcal{Y}_t$, and may not contain sufficient information to uniquely determine the optimal parameters w_{opt}. Due to evolutionary processes like genetic drift and selection, P_t can be multi-modal and may not capture the full variability of genotype-phenotype relationships.

The dataset may lack diversity, missing critical genotype variations that influence the phenotype, leading to an underdetermined system where multiple parameter settings fit the available data.

Without adjusting for confounders, a solution w_{opt} may exist since there are parameters w that minimize the loss function ℓ over the training set $\tilde{\mathcal{D}}$, but it may not represent the true underlying relationship. Additionally, the existence of a reliable solution may also be prevented due to insufficient data and small sample size, due to the dataset $\mathcal{D}_t$ only representing a *limited sample* from the subset $\mathcal{X}_t \times \mathcal{Y}_t$. Data scarcity is a prevalent issue in bacterial genomics, with whole genome sequencing dataset sizes typically on the order of 10^3 isolates, with single-source datasets struggling to reach over 5,000 samples for most species (exceptions being *Salmonella enterica* and *Mycobacterium tuberculosis*). The *Staphylococcus aureus* dataset contains $n = 4140$ samples.

Information loss in representations

Information loss in the feature representations $\phi : \mathcal{X} \to X$ for genotypes and $\psi : \mathcal{Y} \to Y$ for phenotypes leads to non-injectivity in the mapping F in optimization 1 by failing to preserve all relevant details necessary for accurate prediction.

The mapping $\phi : \mathcal{X} \to X$ may fail to preserve all relevant genetic information (i.e., due to dimensionality reduction or feature selection) necessary for accurate prediction.

Similarly, $\psi : \mathcal{Y} \to Y$ may simplify phenotypic information and obscure differences between phenotypes that are important for learning the mapping. Phenotype space compression – converting continuous or multiclass traits into binary categories – focuses the learning task by reducing the complexity of the output space, but at the expense of information loss.

Following the hierarchical framework in Section 2.1, the **forward mappings** ρ and τ must prevent the loss of variant information and preserve unique genetic information relevant to causality. Ideally, they should maintain a one-to-one mapping for features between spaces to ensure full identifiability for fine-mapping tasks, or should be informed with relevant biological properties to enhance interpretability when extracting causal insights.

Practical examples of information loss arise when applying dimensionality reduction and feature selection techniques to manage high-dimensional genetic data. While these methods are essential for computational efficiency and mitigating overfitting, they can inadvertently eliminate relevant genetic factors crucial for accurately modeling the true mapping function F in the optimization task. Excluding such factors risks producing suboptimal models that fail to capture the underlying biological relationships.

Feature selection methods reduce dimensionality of the feature space, and by extension constrain the function space $\mathcal{F}$, by selecting a subset of features before modeling (or as part of the model fitting) to focus the model on the most informative genetic variants. Given the dataset $\mathcal{D}_t$ and representation function $\phi : \mathcal{X}_t \to X = \mathbb{R}^D$, the feature selection objective is to define a selection function $\sigma : X \to X_S$, where $X_S = \mathbb{R}^d \subset X$ is the reduced representation space with $d < D$ selected features $S \subseteq \{1, 2, \ldots, D\}$. When using a genotype matrix M, σ maps each genotype representation $\phi(x_i) \in \mathbb{R}^D$ to a reduced $(1 \times d)$ feature vector $\sigma(\phi(x_i)) \in \mathbb{R}^d$. Commonly used techniques include performing statistical tests (using univariate tests to select genetic features, e.g., SNPs, unitigs etc, significantly associated with the phenotype), recursive feature elimination (iteratively removing features based on their importance to the model), and embedded methods (algorithms that perform feature selection as part of the model training process).

Dimensionality reduction is the process of transforming data from a high-dimensional representation space X to a lower-dimensional space $X' \subset X$ while preserving essential properties or structures of the original data, useful for the phenotype prediction task at hand.

Genetic markers

The choice of marker space $\mathcal{M}$ critically shapes the representation genotype space X, influencing the well-posedness of the GP mapping by enhancing biological relevance. For example, **linkage disequilibrium (LD)**-induced redundancy may be mitigated through marker selection, with unitigs and genes supporting more robust GP mapping and precise fine mapping. LD refers to the non-random association of alleles at different loci in the genome.

A single nucleotide polymorphism, or **SNP**, is a single base-pair change at a specific genomic location between isolates, hence they capture single base-pair variations but miss complex genetic changes.

k-mers represent contiguous nucleotide sequences of fixed length k extracted from raw genomic data, offering broader genomic variation coverage but resulting in high-dimensional, redundant feature spaces, and facilitate the assembly of genomes via construction of a **de Bruijn Graph (DBG)** – a graph structure where each node represents a k-mer, and edges connect nodes that overlap by $k - 1$ bases.

Unitigs and **compacted DBG (cDBG)** structures aggregate overlapping k-mers (non-branching paths within a DBG) into unique, contiguous sequences, providing more efficient and less redundant representations.

Genes consolidate multiple variants, lowering dimensionality and improving interpretability, though they sacrifice resolution and exclude non-coding regions.

Uniqueness in the GP mapping is not guaranteed due to the **high dimensionality** and redundancy in genomic data (high **multicollinearity**); multiple parameter sets may yield the same minimal loss. The genome $x \in \mathcal{X}$ is typically measured in base pairs, with size falling anywhere within an approximate range of 100 kbp to 10 Mbp for bacteria. The *Staphylococcus aureus* genome is typically ∼2.8Mbp, but the dataset grows with the capture of genetic variants. The genome representations $\phi(x) \in X_t$ are typically high-dimensional. For example, the *Staphylococcus aureus* dataset is noted to follow a unitig representation that results in $p \approx 1.2$m parameters, notably reduced in comparison to alternative variant representations (i.e., k-mers). Genome-wide linkage disequilibrium (LD) is the phenomenon resulting in non-random association of alleles at different loci, which in bacterial populations extends throughout the entire genome due to clonal replication and leads to strong correlation structures across genome representations.

Extending the task of constructing a well-posed GP problem to bacterial fine mapping (Section 3.1) – inferring causal relationships from observational data – poses additional significant challenges, especially given the lack of controlled interventions that are typically available in experimental settings. Namely, the **marker space** $\mathcal{M}$ must include all potential causal variants with precise, ideally unique, genomic locations to ensure **identifiability** (uniquely associating

genetic variants $\mathcal{M}^* \subseteq \mathcal{M}$ with phenotypic variations) and **interpretability** (biologically contextualizing relationships between genetic variants and phenotypic traits).

Representing the genotype space

In practice, one considers a finite set of markers $\mathcal{M} = \{m_1, m_2, ..., m_p\}$, for example the set of unitigs – nodes of a cDBG G_t. The representation space $\mathcal{R}$ is then the space of all finite length paths through the graph G_t. Hence, each genotype $x \in \mathcal{X}_t$ is represented by the corresponding path $\rho(x)$ in G_t. For our machine learning model, the paths in G_t (elements of $\mathcal{R}$) can be encoded e.g. by marking out the unitigs visited by the path $\rho(x)$. In other words, the representations space is the p-dimensional binary hypercube, $X = \{0,1\}^p$, where the j-th coordinate of $\phi(x) = \tau(\rho(x)) \in X$, $\phi(x)_j$, is 1 if and only if the unitig m_j is visited by the path $\rho(x)$ (i.e. if the unitig is contained in the genotype x), and 0 otherwise.

The genotypes of the data set $\mathcal{D} = \{(x_i, y_i) | x_i \in \mathcal{X}_t, y_i \in \mathcal{Y}_t, i = 1, 2, ..., n\}$ can then be compactly represented via a **genotype matrix** $M \in \{0,1\}^{n \times p}$,

$$M = \begin{pmatrix} \phi(x_1)_1 & \phi(x_1)_2 & \cdots & \phi(x_1)_p \\ \phi(x_2)_1 & \phi(x_2)_2 & \cdots & \phi(x_2)_p \\ \vdots & \vdots & \ddots & \vdots \\ \phi(x_n)_1 & \phi(x_n)_2 & \cdots & \phi(x_n)_p \end{pmatrix}$$

Each row $\phi(x) = (\phi(x_1)_1, \phi(x_1)_2, \cdots, \phi(x_1)_p)$ corresponds to the genotype of sample i across all p genetic features.

Of course, other types of representations are possible and in general entries $\phi(x_i)_j$ can be binary values (presence or absence of a feature), categorical variables, or numerical values, depending on the nature of the genetic data representation. If a suitable dimensionality reduction technique is applied, the final genotype representations may be of dimensionality $d < p$.

While the genotype matrix is widely used (Ma et al. [2020], Macesic et al. [2020], Burgaya et al. [2023]), it is important to consider whether alternative representations might better capture the biological complexity of the data for certain modeling approaches, such as graph-based (Yang et al. [2021]) or sequence-based representations (Wiatrak et al. [2024]).

Unmeasured confounders and observational noise

Population structure

Population structure arises from the evolutionary history and genetic relatedness within the sample population. It affects both genotype and phenotype distributions in P_t, introducing correlations that can confound the genotype-phenotype relationship (spurious associations between genotypes and phenotypes). **It is a hard requirement that confounders must be accounted for either implicitly and/or explicitly for a GP mapping problem to be well-posed.**

Explicit modeling involves directly incorporating known confounders into the ML model through additional features or adjustments. Extending to fine mapping, reliable causal inference demands **causal sufficiency** – no unmeasured confounders influencing both the cause and effect variables – which necessitates explicit modeling of confounders. Explicit modeling approaches rely on accurately approximating the unmeasured confounders through various techniques. Population structure may be handled explicitly via additional covariates z or separately via regularization, typically through extracting principal components (PCs) from genotype data using methods such as Principal Component Analysis (PCA), or regularization with a similarity matrix K.

Derived from the genotype data, K captures how genetically similar each pair of individuals is within a population, which can be calculated using measures such as genetic distances, phylogenetic distances from reconstructed trees, or other appropriate similarity metrics. The matrix K influences the model through regularization, penalizing large differences in predicted phenotypes $\hat{y}$ for individuals that are genetically similar. An example (Yurtseven et al. [2023]) of enforcing this additional constraint is given by:

$$\min_{F \in \mathcal{F}_{\text{constrained}}} \sum_{i=1}^{n} \ell(\psi^{-1}(F(\phi(x_i), z_i; w)), y_i) + \lambda \bar{F}^\intercal K \bar{F}. \quad (4)$$

Here, $F : X \times Z \times W \to \mathbb{R}$ and

$$\bar{F} = (F(\phi(x_1), z_1; w), F(\phi(x_2), z_2; w), ..., F(\phi(x_n), z_n; w))^\intercal$$

is the vector of model predictions across the data sample.

Implicit modeling relies on the ML model's inherent ability to learn and adjust for confounders through complex pattern recognition without directly incorporating specific confounder variables into the model. Complex ML models, such as deep neural networks, inherently possess the capacity to learn and adjust for confounders through enhanced nonlinearities and richness in the functions, provided sufficiently large data sample is available. However, bacterial datasets typically exhibit limited sample sizes, constraining the applicability of these complex models that demand extensive amounts of data to achieve robust generalization.

Consequently, the prevalent method for implicitly accounting for population structure in bacterial studies is the use of of cross-validation techniques, which are well-suited for simpler models. Strain-specific cross-validation is a tailored validation strategy designed to indirectly account for population structure by ensuring that related or similar strains are appropriately represented in both training and validation sets. These methods are focused on model evaluation and generalizability, via assessing how well a model handles confounders. Traditional methods using random partitioning (Pincus et al. [2020]) can inadvertently place related strains in both sets, thus inflating performance metrics due to the model effectively memorizing strain-specific features rather than learning generalizable patterns. Strain-specific cross-validation partitions the data based on genetic clusters (Lees et al. [2020]). In this approach, entire strains are held out during each fold of training and validation, which tests the model's ability to generalize across different genetic backgrounds and reduce the risk of overfitting to specific strains.

Observational noise

Observational noise refers to random variability in the data that cannot be attributed to the variables being studied, making it challenging for models to learn accurate mappings.

Observational noise may be attributed to measurement errors (e.g., sequencing errors and phenotype measurements), stochastic gene expression (e.g., random fluctuations in antibiotic resistance levels unrelated to genetic variability), processing errors (e.g., DBG assembly), or technical/detection limitations (e.g., low expression genes falling below detection thresholds).

Model complexity and function space limitations

The complexity of the predictive model F and and limitations of the function space it operates within can cause non-injectivity in the mapping from parameters w to predicted phenotypes $\hat{y}$. An overparameterized model, with a high-dimensional parameter space W, may contain redundancies or non-identifiable parameters. Conversely, an underparameterized model may lack the capacity to approximate the true genotype-to-phenotype mapping Θ.

Limited sample sizes inherent to bacterial datasets impose significant constraints on the choice of predictive models, often being forced to rely on simpler models, which may risk being underparameterized. To mitigate the lack of inherent ability in capturing complex biological relationships, it is essential to enhance these simpler models with additional constraints or incorporate prior biological knowledge.

ML view of datasets

Before adding additional constraints to address ill-posedness of the GP mapping task, it is essential to understand how ML models inherently "see" datasets and how their assumptions might align (or conflict) with knowledge of bacterial population behavior.

ML models typically assume that data points are **independent and identically distributed (iid)**, which is often violated in bacterial datasets due to evolutionary processes that introduce dependencies among genotypes – genotypes present within the actual distribution over $\mathcal{X}_t \times \mathcal{Y}_t$, from which $\mathcal{D}_t$ is sampled, cannot be assumed to be independently generated.

The **stationarity** of the data distribution is another critical assumption, yet bacterial populations may experience shifts over time and geographical location driven by selective pressures and founder effects, leading to non-stationary data generating processes.

Furthermore, ML models assume that the dataset contains a **complete and representative set of features** relevant to the target outcome. However, the complexity of bacterial genomes and technological limitations can result in incomplete feature sets, where not all relevant genetic variants are observed. **Measurement accuracy** is also presumed, but in practice, sequencing errors and phenotype measurement inaccuracies introduce noise into the data.

Domain knowledge is not inherently reflected within many ML models due to **uniform treatment of features** (all features being treated equally). The biological relevance of certain genetic markers may therefore not be captured within the model, and will limit the ability of a model to make reliable predictions by exclusion of these parameters.

Many ML models also assume **smoothness** and **stability** in the map on representations from genotypes to phenotypes $F : X \to Y$. Smoothness here refers to F being continuous between metric spaces (X, d_X), and (Y, d_Y), where d_X and d_Y correspond to distance metrics that measure distances between elements $\phi(x_i) \in X$ and $\psi(y_i) \in Y$, respectively. F is continuous if $\forall \epsilon > 0$, $\exists \delta > 0$ such that $d_X(\phi(x_1), \phi(x_2)) < \delta \implies d_Y(\psi(y_1), \psi(y_2)) < \epsilon$, where ϵ and δ are positive real numbers representing *small* distances in the representation spaces X and Y, respectively. However, in bacterial GP mappings, the ground truth Θ is often non-smooth due to causal sparsity, where minor genetic differences can lead to significant phenotypic changes. Predictive models assuming smoothness in F may struggle to generalize when attempting to approximate Θ with a smooth mapping f.

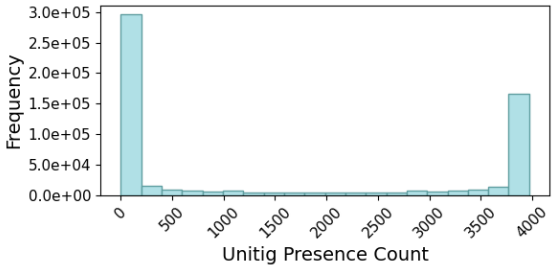

Fig. 1. Histogram of unitig presence count across 4140 *Staphylococcus aureus* isolates.

For *Staphylococcus aureus*, just 61 acquired genes and 82 mutations form the set of known causal mechanisms across 16 phenotypes (Wheeler et al. [2019]). There is also additional sparsity in the feature set, where the majority of unitigs are only present in a handful of samples. Figure 1 displays the occurrence rate for each of the 1,238,055 unitigs across all samples, with their frequency ranging from 1 to 4,139 out of 4,140 samples. A large portion of unitigs are only present in 83 (2%) or fewer samples. This level of sparsity in the data exacerbates the issue of overfitting, making it much harder to generalize to new data.

Choice of predictive model

This section focuses on models facilitated by a genotype matrix representation M – the predominant data representation for GP prediction tasks – and does not cover models that inherently require alternative data representations

Linear models. Linear models assume a linear relationship between genetic features and the phenotype, wherein the genetic features influence the phenotype independently and additively. They provide a straightforward approach and interpretable results, with the addition of penalty terms better managing high-dimensionality (Burgaya et al. [2023], Saber and Shapiro [2020]). This limits the function space $\mathcal{F}$ to linear functions with parameters $w^\intercal = (w_\phi^\intercal, w_z^\intercal)$ – where in the optimization 1, $F(\phi(x_i), z_i; w)) = w_\phi^\intercal \phi(x_i) + w_z^\intercal z_i$ represents a linear function of the transformed genotype $\phi(x_i)$ – which are unable to capture the complex nonlinear interactions often present in bacterial populations.

Nonlinear models. The relationship between features and the phenotype may be nonlinear and involve interactions. Expanding the function space $\mathcal{F}$ to include nonlinear functions increases flexibility but also the risk of overfitting if not properly regularized. Among the non-linear models, **tree-based models** have enjoyed particular attention (Karanth et al. [2022], Allen et al. [2021], Batisti Biffignandi et al. [2024]). Decision trees and ensemble methods (random forests, gradient boosting) can naturally model interactions and nonlinearities. The phenotype can be predicted based on hierarchical decision rules derived from the genetic features. The function space $\mathcal{F}$ is constrained by the structure of the trees and hyperparameters like depth and the number of trees. Interpretability can sometimes be challenging due to the complexity of the ensemble models.

Parameter fitting can be performed in various settings, but if F is formulated as a probabilistic model (conditional on an extended input $(\phi(x), z)$ a full distribution over phenotypes

is provided), a **Bayesian framework** can be adopted. Bayesian approaches incorporate prior distributions over model parameters, allowing for natural integration of prior knowledge and uncertainty quantification. The posterior distribution over the model parameters w is given by $p(w|\mathcal{D}) \propto p(\mathcal{D}|w)p(w)$, where $p(w)$ is the prior and the likelihood $p(\mathcal{D}|w)$ is provided by our probabilistic model.

Feature attribution in fine mapping

Interpretability is crucial for fine mapping to accurately distinguish causal variants from confounded associations, making feature attribution methods a popular choice for extracting causal insights. Ensemble methods such as random forests and gradient boosting machines, along with ability to handle high-dimensional data, provide insights into feature importances (Buckley and Harvey [2021], Rahman et al. [2018], Wassan et al. [2018]).

Feature importances provide a quantitative measure of how much each genetic feature $\phi(x)_j$ contributes to the model's prediction $\hat{y}$, derived from X. A high feature importance score suggests that a particular feature has a strong association with the phenotype of interest.

SHAP values estimate the contribution of each $\phi(x_i)_j$ to the model's prediction $\hat{y}_i$ for a given sample i, so can highlight specific genetic variants that consistently influence the phenotype across different strains or isolates. However, these methods capture *associations* rather than *direct causal relationships*, and are prone to highlighting regions of the genome that are spuriously associated with the phenotype rather than causally related due to the high rates of LD in bacterial populations.

Incorporating domain knowledge

To address the challenges posed by the one-to-many inverse mapping from phenotypes to model parameters, integrating domain-specific knowledge into the genotype-to-phenotype mapping can be used to constrain the function space and improve model interpretability.

Spatial dependencies

Spatial dependencies refer to the relationships and interactions between genetic variants based on their physical locations on the genome, and possibly their positions within representational structures such as cDBGs by extension. Spatial dependencies can be integrated into the feature representation $\phi(x)$ by encoding the physical or representational positions of genetic variants. This integration allows the predictive model F to account for the localized interactions and collective effects of adjacent genetic variants. For instance, variants that are physically proximal on the chromosome may be more likely to interact epistatically, influencing the phenotype in a manner that is not apparent when considering each variant independently. Spatial dependencies facilitate the incorporation of genomic domain knowledge, such as the tendency of causal variants to cluster within specific genomic regions or regulatory elements, and can constrain the model by prioritizing interactions among nearby variants

Figure 2A illustrates a Manhattan plot for bGWAS results using Pyseer (Lees et al. [2018]), highlighting the close proximity of causal variants in the genome. Figure 2B depicts a 3-dimensional protein structure of the main chromosomal locus for fusidic acid resistance in *Staphylococcus aureus*. It is evident that known causal mutations, along with compensatory mutations (mutations which co-occur with causal mutations to mitigate any negative impacts they have on protein function) and novel bGWAS candidates, exist within spatial "hotspots" – novel candidates obtained via bGWAS (yellow) are close to known causative mutations (red). Most classical ML models accept data in a tabular format, where features are treated as independent variables and lack any representation of spatial dependencies (standard representation of genomic data discussed in Section 6.2).

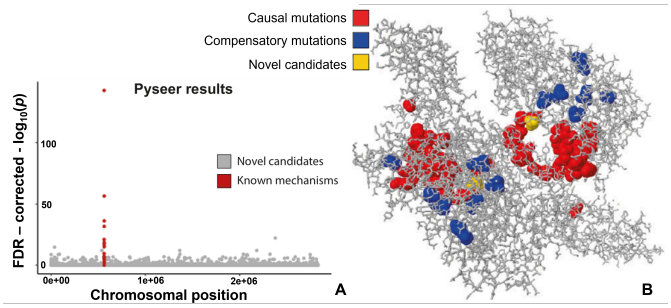

Fig. 2. A: Manhattan plot for bGWAS single nucleotide polymorphism Pyseer results. B: 3-dimensional structure obtained experimentally of a single *Staphylococcus aureus* protein. Figure adapted from Wheeler et al. [2019].

Prior causal knowledge

In the formal framework, prior causal knowledge can be incorporated by selectively augmenting the feature representation $\phi(x)$ with information about known causal variants or by structuring the predictive function F to prioritize these variants during the learning process.

Conclusion

Several open problems remain in bacterial causal variant detection. The first is determining whether the genotype-to-phenotype mapping task can be reformulated to be well-posed, along with identifying the necessary conditions and representations required for this transformation. Addressing this foundational issue is crucial before considering causal variant detection, as it establishes the basis for any subsequent fine mapping efforts. If a well-posed mapping is achievable, the next challenge is to determine whether it can also satisfy the specific requirements for bacterial fine mapping, including the necessary conditions and representations to achieve this.

We have pinpointed major areas that need to be addressed to overcome these challenges. Developing robust representation spaces is essential, as they must preserve biological properties and causal information while minimizing spurious associations. The construction of these representation spaces depends on factors such as model choice and the effective incorporation of domain knowledge at appropriate stages of the pipeline. Additionally, there is a balance to be struck between model complexity and interpretability for bacterial fine mapping. While more complex models may better capture the governing biological mechanisms, they often reduce the ease with which results can be interpreted, while also demanding larger sample sizes than those typically available. Therefore, interpretability

must be ensured through complementary approaches, such as post hoc analysis or specialized interpretability techniques, to maintain biological relevance between spaces.

Competing interests

No competing interest is declared.

Author contributions statement

Acknowledgments

References

H. Aliee, F. Kapl, S. Hediyeh-Zadeh, and F. J. Theis. Conditionally invariant representation learning for disentangling cellular heterogeneity. *arXiv preprint arXiv:2307.00558*, 2023.

J. P. Allen, E. Snitkin, N. B. Pincus, and A. R. Hauser. Forest and trees: exploring bacterial virulence with genome-wide association studies and machine learning. *Trends in microbiology*, 29(7):621–633, 2021.

G. Batisti Biffignandi, L. Chindelevitch, M. Corbella, E. J. Feil, D. Sassera, and J. A. Lees. Optimising machine learning prediction of minimum inhibitory concentrations in klebsiella pneumoniae. *Microbial Genomics*, 10(3):001222, 2024.

S. J. Buckley and R. J. Harvey. Lessons learnt from using the machine learning random forest algorithm to predict virulence in streptococcus pyogenes. *Frontiers in Cellular and Infection Microbiology*, 11:809560, 2021.

J. Burgaya, J. Marin, G. Royer, B. Condamine, B. Gachet, O. Clermont, F. Jaureguy, C. Burdet, A. Lefort, V. de Lastours, et al. The bacterial genetic determinants of escherichia coli capacity to cause bloodstream infections in humans. *PLoS Genetics*, 19(8):e1010842, 2023.

A. Chalka, T. J. Dallman, P. Vohra, M. P. Stevens, and D. L. Gally. The advantage of intergenic regions as genomic features for machine-learning-based host attribution of salmonella typhimurium from the usa. *Microbial genomics*, 9(10):001116, 2023.

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

K. Hu, F. Meyer, Z.-L. Deng, E. Asgari, T.-H. Kuo, P. C. Münch, and A. C. McHardy. Assessing computational predictions of antimicrobial resistance phenotypes from microbial genomes. *bioRxiv*, pages 2024–01, 2024.

S. Karanth, C. K. Tanui, J. Meng, and A. K. Pradhan. Exploring the predictive capability of advanced machine learning in identifying severe disease phenotype in salmonella enterica. *Food Research International*, 151:110817, 2022.

J. A. Lees, M. Galardini, S. D. Bentley, J. N. Weiser, and J. Corander. Pyseer: a comprehensive tool for microbial pangenome-wide association studies. *Bioinformatics*, 34(24): 4310–4312, 2018.

J. A. Lees, T. T. Mai, M. Galardini, N. E. Wheeler, S. T. Horsfield, J. Parkhill, and J. Corander. Improved prediction of bacterial genotype-phenotype associations using interpretable pangenome-spanning regressions. *MBio*, 11(4): 10–1128, 2020.

K. C. Ma, T. D. Mortimer, M. A. Duckett, A. L. Hicks, N. E. Wheeler, L. Sánchez-Busó, and Y. H. Grad. Increased power from conditional bacterial genome-wide association identifies macrolide resistance mutations in neisseria gonorrhoeae. *Nature communications*, 11(1):5374, 2020.

N. Macesic, O. J. Bear Don't Walk IV, I. Pe'er, N. P. Tatonetti, A. Y. Peleg, and A.-C. Uhlemann. Predicting phenotypic polymyxin resistance in klebsiella pneumoniae through machine learning analysis of genomic data. *Msystems*, 5(3): 10–1128, 2020.

M. Nguyen, T. Brettin, S. W. Long, J. M. Musser, R. J. Olsen, R. Olson, M. Shukla, R. L. Stevens, F. Xia, H. Yoo, et al. Developing an in silico minimum inhibitory concentration panel test for klebsiella pneumoniae. *Scientific reports*, 8 (1):421, 2018.

N. Pearcy, Y. Hu, M. Baker, A. Maciel-Guerra, N. Xue, W. Wang, J. Kaler, Z. Peng, F. Li, and T. Dottorini. Genome-scale metabolic models and machine learning reveal genetic determinants of antibiotic resistance in escherichia coli and unravel the underlying metabolic adaptation mechanisms. *Msystems*, 6(4):e00913–20, 2021.

N. B. Pincus, E. A. Ozer, J. P. Allen, M. Nguyen, J. J. Davis, D. R. Winter, C.-H. Chuang, C.-H. Chiu, L. Zamorano, A. Oliver, et al. A genome-based model to predict the virulence of pseudomonas aeruginosa isolates. *MBio*, 11(4): 10–1128, 2020.

S. F. Rahman, M. R. Olm, M. J. Morowitz, and J. F. Banfield. Machine learning leveraging genomes from metagenomes identifies influential antibiotic resistance genes in the infant gut microbiome. *MSystems*, 3(1):10–1128, 2018.

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

P. Schneider, K. Bunte, H. Stiekema, B. Hammer, T. Villmann, and M. Biehl. Regularization in matrix relevance learning. *IEEE Transactions on Neural Networks*, 21(5):831–840, 2010. doi: 10.1109/TNN.2010.2042729.

J. T. Wassan, H. Wang, F. Browne, and H. Zheng. Paam-ml: a novel phylogeny and abundance aware machine learning modelling approach for microbiome classification. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 44–49. IEEE, 2018.

N. E. Wheeler, S. Reuter, C. Chewapreecha, J. A. Lees, B. Blane, C. Horner, D. Enoch, N. M. Brown, M. Estée Török, D. M. Aanensen, et al. Contrasting approaches to genome-wide association studies impact the detection of resistance mechanisms in staphylococcus aureus. *bioRxiv*, page 758144, 2019.

M. Wiatrak, A. Weimann, A. M. Dinan, M. Brbić, and R. A. Floto. Sequence-based modelling of bacterial genomes enables accurate antibiotic resistance prediction. *bioRxiv*, pages 2024–01, 2024.

D. P. Wipf and S. S. Nagarajan. A new view of automatic relevance determination. In *Neural Information Processing Systems*, 2007. URL `https://api.semanticscholar.org/CorpusID:9583791`.

Y. Yang, T. M. Walker, S. Kouchaki, C. Wang, T. E. Peto, D. W. Crook, C. Consortium, and D. A. Clifton. An end-to-end heterogeneous graph attention network for mycobacterium tuberculosis drug-resistance prediction. *Briefings in bioinformatics*, 22(6):bbab299, 2021.

A. Yurtseven, S. Buyanova, A. A. Agrawal, O. O. Bochkareva, and O. V. Kalinina. Machine learning and phylogenetic analysis allow for predicting antibiotic resistance in m. tuberculosis. *BMC microbiology*, 23(1):404, 2023.