# PSL-GWAS: A Microbial GWAS Method Using Statistical Relational Learning

Alex Miller[1], Eriq Augustine[1], Elijah Pandolfo[1], Lise Getoor[1]

**1** Computer Science and Engineering, University of California, Santa Cruz, Santa Cruz, California, United States of America

## 1 Abstract

Microbial genome-wide association studies (mGWAS) are a new, quickly growing area of research that aims to identify genetic variants that are associated with phenotypes of interest in microbes. We introduce PSL-GWAS, a flexible mGWAS method that makes use of the statistical relational learning framework Probabilistic Soft Logic (PSL.) Our method can be readily adapted to incorporate domain knowledge or the output of other mGWAS methods. We show that PSL-GWAS performs comparably to well-established mGWAS methods on a dataset of 355 E. Coli samples and antibiotic resistance phenotypes.

## 2 Introduction

Microbial genome-wide association studies (mGWAS) look at identifying what genetic variants (differences from an organism's typical DNA sequence) are associated with some phenotype (observable trait), such as drug resistance. Many existing methods construct a pan-genome (set of all genetic variants) from a mGWAS dataset and identify which are likely to cause a phenotype. Microbial GWAS methods have successfully identified genomic markers associated with microbial virulence and drug resistance, and this is particularly important given the emergence of multi-drug resistant bacteria. The understanding gained from mGWAS results and analysis can inform drug design as well as treatment decisions for a particular disease.

Straightforward association tests are generally insufficient due to the highly-structured nature of most mGWAS datasets. Strong selective pressure in bacteria for traits like antibiotic resistance, the frequency of horizontal gene transfer (the swapping of genetic material between bacteria), and the asexual reproduction of bacteria often create strong structure within a mGWAS dataset. Within this structure, a large number of non-causal (or "hitchhiker") variants in the pan-genome will appear to be highly correlated with a phenotype. The factors that lead to population structure also result in a large and diverse pan-genome which may add computational costs, because having a larger pan-genome forces a mGWAS method to consider more variants.

A near-universal feature of mGWAS methods is that they take steps to account for population structure. Some common ways that existing methods deal with population structure are based on applying principal component analysis to the sample-sample distance matrix (like in the mGWAS package SEER (1)) or working directly with the structure described in an evolutionary history like Scoary (2) does.

We introduce PSL-GWAS, a mGWAS framework that uses the statistical relational learning framework Probabilistic Soft Logic (Bach, et al. 3). By leveraging relationships

in the dataset captured in interpretable first-order logical rules, our method assigns a confidence score to each genetic variant for each phenotype in the dataset. Our method addresses the issue of population structure through directly using the distance matrix and can jointly infer genetic variant-phenotype association scores over multiple phenotypes. Like many other existing GWAS approaches, our approach uses DNA words of length $k$ ($k$-mers) to capture many different forms of genetic variation including single-nucleotide polymorphisms (SNPs) and base pair insertion/deletion (indels.) Our method performs comparably with current, well-known mGWAS methods on the task of assigning high confidence scores to $k$-mers that have been confirmed to cause antibiotic resistance in E. Coli.

# 3 Related work

San, et al. (4) provide a thorough review of mGWAS literature. We focus here on two popular methods: PySEER and Scoary. These methods both apply a population structure-naive filter to the genetic variants in the dataset, and then perform a structure-aware analysis of gene-phenotype association. Both methods return a set $k$-mers identified as significantly associated with a particular phenotype.

PySEER (5) is a Python package that contains ready-to-go implementations of several mGWAS methods, including the fixed effects model SEER (1) that it is named after. Depending on the model, either a chi-squared or simple correlation test are applied as an initial filter. To address population structure in SEER, metric multi-dimensional scaling is applied to the sample-sample distance matrix and the resulting eigenvectors are included as covariates in either a linear or logistic regression model, depending on whether the phenotype is binary or continuous. MDS and related principal component analysis-based methods for structure correction are commonly used in other linear model-based GWAS methods such as BugWAS (6) and HAWK (7). Other methods implicitly account for population structure. For example, in PySEER's linear mixed model, the full similarity matrix is included as random effects.

Scoary (2) is a lightweight package that implements a phylogeny-aware pairwise comparison algorithm that counts non-intersecting sample pairs where samples A and B differ in both gene presence and displayed trait. A binomial test is then used to compute p-values. The authors note that care must be taken to account for the many possible pairings of samples and propose some methods for addressing this issue and validating the significant genes it returns.

# 4 PSL Overview

PSL-GWAS uses the statistical relational learning frame Probabilistic Soft Logic (PSL) to predict genetic variants that are likely associated with phenotypes of interest. PSL induces a hinge-loss Markov Random Field (HL-MRF) that encodes a knowledge base of variables and relations among them. The HL-MRF is associated with the following conditional probability distribution of unobserved variables $\mathbf{Y}$ and evidence $\mathbf{X}$:

$$P(\mathbf{Y}|\mathbf{X}) \propto exp(-\sum_{j=1}^{m} w_j \phi_j)$$

$w$ is a vector of weights and $\phi$ is a vector of convex potential functions that correspond to user-defined first-order logical rules that serve as templates for potentials. PSL relaxes template rules with Łukasiewicz logic, an intermediate logic that defines the

following fundamental operations:

$$p \wedge q = max(0, p + q - 1)$$
$$p \vee q = min(1, p + q)$$
$$\neg p = 1 - p$$

These rules relate *predicates*, which for our purposes can be considered as templates for individual variables in the dataset. Predicates are defined by a unique name and some fixed number of arguments $c$ that are associated with it. Predicates are instantiated with constant identifiers from the data to create *ground atoms* which correspond to particular variables in the dataset. An instantiation of a template relates ground atoms and is called a *ground rule*. Ground rules each have a corresponding potential in the HL-MRF. Each template rule is associated with a single weight $w_i$ that is tied to every potential generated from it. In the $j$-th ground rule, let $I_j^-$ be the indices of negated variables in the disjunctive normal form of the rule and $I_j^+$ be the non-negated variable indices. The potential $\phi_j$ is defined as:

$$\phi_j = max\left(0, 1 - \sum_{i \in I_j^+} y_i - \sum_{i \in I_j^-} (1 - y_i)\right)^p$$
$$p \in \{1, 2\}$$

The choice of $p$ is left to the user.

Since the potentials are convex, minimizing their sum to perform maximum a posteriori (MAP) inference is a convex optimization problem that can be solved efficiently. In our experiments, we use tandem inference (8), an out-of-core optimization algorithm that stores potentials on disk and uses stochastic gradient descent to minimize the MAP inference objective. In our experiments, removing the requirement that a model must fit in RAM was necessary to have PSL-GWAS work with our full dataset.

# 5 PSL-GWAS

PSL-GWAS uses Probabilistic Soft Logic (PSL), the statistical relational learning framework described in Section 4 to infer a confidence score for each $k$-mer, phenotype pair. PSL-GWAS takes in genome information (represented as a collection of continuous sub-sequences of DNA) and observed phenotype information for some set of $n$ isolates. The data construction process specifies the values of observed variables and enumerates the unobserved, or *target*, variables whose values are inferred. In our experiments, the inference targets are confidence scores for every $k$-mer, phenotype pair in the dataset. All other variables are observed.

## 5.1 K-mer counting and filtering

PSL-GWAS relies external tools for $k$-mer counting—in our experiments, the genomic data is in FASTA format and we use tools that can work directly with these files. PSL-GWAS uses the tool's output to create a map from $k$-mers to the sets of samples that contain them. We then consider all possible $k$-mer, phenotype pairs and apply two steps of filtering. First, any $k$-mer that is present in fewer than $n$ samples is filtered out immediately. Second, to reduce computational costs and the number of false positives, we filter out $k$-mer, phenotype pairs where the correlation between the $k$-mer's presence and display of the phenotype is below some threshold. This is analogous to the use of

Fisher's exact test in Scoary or the $\chi^2$ test in SEER. We consider the simple correlation between a $k$-mer K and phenotype P:

$$C(K, P) = \frac{\text{\# of samples with K that display P}}{\text{\# of samples with K}}$$

## 5.2 PSL data

We define the following predicates to use in our model:

1. Contains(S, K) — 1 if sample S has $k$-mer K, 0 otherwise.
2. SamplePheno(S, P) — 1 if sample S displays phenotype P, 0 otherwise.
3. SimilarPheno(P1, P2) — The degree of similarity between phenotypes P1 and P2. For example, a user might set SimilarPheno(Ceftazidime, Ceftriaxone) to 1 because they are both cephalosporin antibiotics.
4. DissimilarSample(S1, S2) — The degree of dissimilarity between samples S1 and S2. By default, PSL-GWAS uses a distance matrix output by Mash to set these values.
5. SimilarSample(S1, S2) — The degree of similarity between samples S1 and S2. By default, PSL-GWAS defines similarity as the complement of the dissimilarity.
6. KmerPheno(K, P) — The PSL-GWAS confidence score that $k$-mer K causes phenotype P. Unlike previous predicates, this is used to represent the unknown confidence scores that we infer.

## 5.3 PSL model

A PSL model consists of weighted logical rules and constraints that capture known relationships within the dataset. The relationships leveraged by rules in PSL-GWAS belong to one of four general classes: simple correlation between $k$-mer and phenotype, similarity between phenotypes, population structure-based rules, and a negative prior. The rules are given below:

1. Simple correlation
   (a) 1.0: Contains(S, K) ∧SamplePheno(S, P) →KmerPheno(K, P) ^2
   If a sample contains a $k$-mer and displays a phenotype, there is evidence that $k$-mer causes that phenotype.
   (b) 1.0: Contains(S, K) ∧¬SamplePheno(S, P) →¬KmerPheno(K, P) ^2
   If a sample contains a $k$-mer and does not display a phenotype, there is evidence that $k$-mer does not cause that phenotype.
2. Phenotype similarity
   (a) 1.0: SimilarPheno(P1, P2) ∧KmerPheno(K, P1) →KmerPheno(K, P2) ^2
   The same $k$-mer can cause two very similar phenotypes.
   (b) 1.0: ¬SimilarPheno(P1, P2) ∧KmerPheno(K, P1) →¬KmerPheno(K, P2) ^2
   The same $k$-mer likely does not cause two very dissimilar phenotypes.
3. Population structure

(a) 1.0: SimilarSample(S1, S2) ∧SamplePheno(S1, P) ∧Contains(S1, K) ∧¬SamplePheno(S2, P) ∧¬Contains(S2, K) →KmerPheno(K, P) ^2

If two samples are similar, and one contains a $k$-mer and displays a phenotype and the other does neither, there is evidence that $k$-mer causes that phenotype.

(b) 1.0: DissimilarSample(S1, S2) ∧SamplePheno(S1, P) ∧Contains(S1, K) ∧SamplePheno(S2, P) ∧Contains(S2, K) →KmerPheno(K, P) ^2

If two samples are dissimilar, and both display a phenotype and both contain the same $k$-mer, there is evidence that $k$-mer causes the phenotype.

(c) 1.0: ¬DissimilarSample(S1, S2) ∧SamplePheno(S1, P) ∧Contains(S1, K) ∧SamplePheno(S2, P) ∧Contains(S2, K) →¬KmerPheno(K, P) ^2

If two samples are similar, and both display a phenotype and both contain the same $k$-mer, there is not evidence that $k$-mer causes the phenotype. This rule limits the number of $k$-mers with very high confidence scores.

4. Negative prior

(a) 1.0: ¬KmerPheno(K, P) ^2

Most $k$-mers do not cause a particular phenotype.

PSL-GWAS assigns a uniform weight to every rule, though we plan to develop a method for selecting better weights in the future. PSL's flexibility allows PSL-GWAS models to be easily adapted to incorporate domain-specific information. Logical rules can be added or discarded easily, and modelers have free choice in setting the observed values for any predicate they use. This also allows users to incorporate the output of other microbial GWAS methods in PSL-GWAS. For example, users could add the rule:

$$\textsc{Baseline}(\texttt{K}, \texttt{P}) \rightarrow \textsc{KmerPheno}(\texttt{K}, \texttt{P})$$

Finally, to make our model computationally feasible, we use a strategy known as *blocking*, in which we only consider the top $p$% most similar sample pairs in the population structure-based rules. We refer to the hyperparameter $p$ as the blocking threshold. Although it's omitted from the rules above for clarity, we add an additional *blocking predicate* to the left-hand side of the population structure rules to ensure those rules are only grounded with the selected most-similar pairs.

# 6 Experiments

## 6.1 Dataset

We evaluated our method on a dataset of 355 E. Coli samples collected from two hospitals in the United States. The dataset contains each sample's assembled genome and associated binary phenotype data, where the phenotypes of interest are resistance to one of these seven antibiotics: ceftazidime, ceftriaxone, cefepime, tobramycin, gentamicin, cefazolin, and ampicillin. Missing phenotype observations are marked with "NA." For truth data, we used 11 genes that are present in the dataset and are known to cause resistance to some subset of the seven antibiotics for which we have data.

### 6.2 PSL-GWAS Hyperparameters

We use $k$-mers to capture genetic variation in our experiments. $k$-mers are flexible enough to capture variation caused by several major types of mutation, including single-nucleotide polymorphisms (SNPs), base pair insertions/deletions (indels), and copy number variations. We count $k$-mers with fsm-lite and estimate the full sample-sample distance matrix using Mash. We use fsm-lite's default $k$-mer frequency cutoffs of 1% at the minimum and 99% at the maximum, and a fixed $k$-mer length of 31. We then apply the filtering steps described in Section 5.1 with a frequency threshold of 5 and a correlation threshold of 0.9.

For sample-sample dissimilarity observations, we use the genetic distance estimates from Mash. Distances are scaled into $[0, 1]$ by subtracting the minimum distance and dividing by the maximum. We then filter out all but the top $p = 10\%$ most similar sample pairs. The complements of the remaining dissimilarity observations are then used for similarity. We define phenotype similarity as the correlation between two phenotypes in the original dataset.

We perform MAP inference using tandem inference (8), an out-of-core optimization algorithm that leverages stochastic gradient descent. We found that using this algorithm was necessary to apply PSL-GWAS to the full dataset. Because we found that the whole model wouldn't even fit in 250 gigabytes of RAM, all non-out-of-core optimization algorithms available in PSL, such as the alternating direction method of multipliers (ADMM) (9), couldn't be used in our experiments.

### 6.3 Baseline

We compare PSL-GWAS to the well-known, well-documented, and flexible microbial GWAS package PySEER. Specifically, we use PySEER's linear mixed model (LMM) and whole genome elastic net (WG) with default hyperparameters to identify significant $k$-mers. We use fsm-lite to count $k$-mers and use Mash (10) for genetic distance estimation.

### 6.4 Evaluation metrics

Many microbial GWAS programs, including our baseline PySEER, return a set of $k$-mers that pass some $p$-value threshold for significance. In contrast, PSL-GWAS returns a ranked list of $k$-mers. For our evaluation, we sort the $k$-mers returned by PySEER by $p$-value to rank them.

If a $k$-mer is part of one of the 11 causal genes in the dataset (i.e. is a substring of the gene sequence) then we consider it a hit. We evaluate our method on the mean reciprocal rank of the hits and hits/precision at K. In the results that follow, metrics are first computed with respect to individual phenotypes, and are then used to compute a mean metric value across all phenotypes.

## 7 Experiment results

Table 1 gives the mean values for mean reciprocal rank and precision at K, and the standard deviations, for each method. PySEER-LMM refers to PySEER's implementation of a linear mixed model, and PySEER-WG refers to its whole genome model. Within each column, bolding a model's mean metric value indicates that 1) its difference with non-bolded values is statistically significant and 2) its difference with other bolded values is not statistically significant. We use a paired $t$ test with a $p$ value cutoff of $\alpha = 0.05$ to determine whether the difference in mean value of a metric between any two methods is statistically significant.

All three methods are capable of identifying significant k-mers in a meaningful capacity and show statistically insignificant differences in their mean reciprocal ranks (MRRs.) PySEER-WG and PSL-GWAS perform comparably well at ranking known causal $k$-mers close to the top of their results, as shown in their mean hits/precision at $K = 100$. PySEER-LMM does not find any known causal $k$-mers at $K = 100$. However, all three methods achieve non-trivial precisions at $K = 500$ and the differences between each method are statistically insignificant. For hits/precision at $K > 500$, both PySEER models perform comparably to each other, and both outperform PSL-GWAS.

| **Method** | **MRR** | **Prec. @100** | **Prec. @500** | **Prec. @1000** |
|---|---|---|---|---|
| PSL-GWAS | **0.0026 ± 0.0028** | **0.3614 ± 0.2698** | **0.2594 ± 0.1321** | 0.1577 ± 0.0763 |
| PySEER-LMM | **0.0007 ± 0.0006** | 0.0 ± 0.0 | **0.5106 ± 0.3199** | **0.5686 ± 0.1508** |
| PySEER-WG | **0.0020 ± 0.0028** | **0.1771 ± 0.3515** | **0.3357 ± 0.3684** | **0.4569 ± 0.2394** |

| **Method** | **Prec. @5000** | **Prec. @10000** | **Prec. @50000** | **Prec. @100000** |
|---|---|---|---|---|
| PSL-GWAS | 0.0398 ± 0.0256 | 0.0205 ± 0.0129 | 0.0044 ± 0.0026 | 0.0027 ± 0.0013 |
| PySEER-LMM | **0.17 ± 0.045** | **0.0853 ± 0.0225** | **0.0182 ± 0.0032** | **0.0102 ± 0.0030** |
| PySEER-WG | **0.2138 ± 0.0638** | **0.1147 ± 0.0383** | **0.0304 ± 0.0137** | **0.0152 ± 0.0068** |

**Table 1.** PSL-GWAS vs. PySEER. Mean values of mean reciprocal rank and precision at $K$ for $K \in \{100, 500, 1000, 5000, 10000, 50000, 100000\}$.

Investigating a potentially-causal genetic variant can be a lengthy or expensive process, which may drive researchers to only consider the most highly-ranked $k$-kmers returned by a microbial GWAS method. These results suggest that PSL-GWAS may have comparable utility to a linear mixed model or whole-genome elastic net in this kind of setting.

# 8 Conclusion and future work

Microbial genome-wide association studies (mGWAS) are a recent research area that involves identifying genetic variants that are associated with particular phenotypes. Given the rise of multi-drug resistant strains of bacteria, mGWAS results concerning traits such as virulence and antibiotic resistance are particularly important from a public health perspective. We introduce PSL-GWAS, a mGWAS method that is robust in its ability to handle binary or continuous phenotype measurements, richly model heavily structured data, and incorporate other GWAS methods in its predictions.

PSL-GWAS uses the statistical relational learning framework Probabilistic Soft Logic (PSL) to rank variants in the pan-genome of samples in a mGWAS dataset. Our method combines a variety of information sources and infers $k$-mer, phenotype association confidence scores over all phenotypes simultaneously. We then compare PSL-GWAS against two baselines: PySEER's implementations of a linear mixed model and a whole genome elastic net.

On an E. Coli antibiotics-resistance dataset, we find that PSL-GWAS outperforms a linear mixed model and performs comparably to a whole genome elastic net. All three methods perform similarly at placing known causal $k$-mers in their top 500 results, with PSL-GWAS and the whole genome model outperforming the linear mixed model when only considering the top 100. All three methods achieve comparable mean reciprocal ranks with respect to known causal $k$-mers, indicating the PSL-GWAS may be a similarly effective mGWAS method to either of these baselines.

Our assessment of PSL-GWAS's performance is limited by several factors. First, we only have 2 baselines to compare against and only evaluate PSL-GWAS on a single dataset. It is also limited by our choice of uniform weights for all of the rules—future

work in this area can give a better sense of PSL-GWAS's effectiveness. Restricting ourselves to a single set of hyperparameters is a related limitation.

It is also worth noting that only the phenotype similarity rules capture relationships among more than one unobserved variable. It is possible that our current model design overlooks PSL's unique ability to collectively reason about the relations between the association scores of separate $k$-mers.

Future work may address the limitations of this study and better describe PSL-GWAS's efficacy as a mGWAS method. Evaluating on multiple datasets, addressing the choice of weights, and potentially performing an ablation study are all natural directions for further work.

# References


1. Lees JA, Vehkala M, Välimäki N, Harris SR, Chewapreecha C, Croucher NJ, et al. Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes. Nature Communications. 2016;7(1):12797. doi:10.1038/ncomms12797.

2. Brynildsrud O, Bohlin J, Scheffer L, Eldholm V. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. Genome Biology. 2016;17(1):238. doi:10.1186/s13059-016-1108-8.

3. Bach SH, Broecheler M, Huang B, Getoor L. Hinge-Loss Markov Random Fields and Probabilistic Soft Logic;.

4. San JE, Baichoo S, Kanzi A, Moosa Y, Lessells R, Fonseca V, et al. Current Affairs of Microbial Genome-Wide Association Studies: Approaches, Bottlenecks and Analytical Pitfalls. Frontiers in Microbiology. 2020;10.

5. Lees JA, Galardini M, Bentley SD, Weiser JN, Corander J. pyseer: a comprehensive tool for microbial pangenome-wide association studies. Bioinformatics. 2018;34(24):4310–4312. doi:10.1093/bioinformatics/bty539.

6. Earle SG, Wu CH, Charlesworth J, Stoesser N, Gordon NC, Walker TM, et al. Identifying lineage effects when controlling for population structure improves power in bacterial association studies. Nature Microbiology. 2016;1(5). doi:10.1038/nmicrobiol.2016.41.

7. Rahman A, Hallgrímsdóttir I, Eisen M, Pachter L. Association mapping from sequencing reads using $k$-mers. eLife. 2018;7:e32920. doi:10.7554/eLife.32920.

8. Srinivasan S, Augustine E, Getoor L. Tandem Inference: An Out-of-Core Streaming Algorithm for Very Large-Scale Relational Inference. Proceedings of the AAAI Conference on Artificial Intelligence. 2020;34(06):10259–10266. doi:10.1609/aaai.v34i06.6588.

9. Boyd S, Parikh N, Chu E, Peleato B, Eckstein J; 2011.

10. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, et al. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biology. 2016;17(1). doi:10.1186/s13059-016-0997-x.