BEATRICE: Bayesian Fine-mapping from Summary Data using Deep Variational Inference

Sayan Ghosal[1], Michael C. Schatz[2], Archana Venkataraman[1,3*]

1 Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, MD, USA
2 Department of Computer Science, Johns Hopkins University, Baltimore, MD, USA
3 Department of Electrical and Computer Engineering, Boston University, Boston, MA, USA

* Correspondences should be directed to archanav@bu.edu

Abstract

We introduce a novel framework BEATRICE to identify putative causal variants from GWAS summary statistics (*https://github.com/sayangsep/Beatrice-Finemapping*). Identifying causal variants is challenging due to their sparsity and to highly correlated variants in the nearby regions. To account for these challenges, our approach relies on a hierarchical Bayesian model that imposes a binary concrete prior on the set of causal variants. We derive a variational algorithm for this fine-mapping problem by minimizing the KL divergence between an approximate density and the posterior probability distribution of the causal configurations. Correspondingly, we use a deep neural network as an inference machine to estimate the parameters of our proposal distribution. Our stochastic optimization procedure allows us to simultaneously sample from the space of causal configurations. We use these samples to compute the posterior inclusion probabilities and determine credible sets for each causal variant. We conduct a detailed simulation study to quantify the performance of our framework across different numbers of causal variants and different noise paradigms, as defined by the relative genetic contributions of causal and non-causal variants. Using this simulated data, we perform a comparative analysis against two state-of-the-art baseline methods for fine-mapping. We demonstrate that BEATRICE achieves uniformly better coverage with comparable power and set sizes, and that the performance gain increases with the number of causal variants. Thus, BEATRICE is a valuable tool to identify causal variants from eQTL and GWAS summary statistics across complex diseases and traits.

Author summary

Fine-mapping provides a way to uncover genetic variants that causally affect some trait of interest. However, correct identification of the causal variants is challenging due to the correlation structure shared across variants. While current fine-mapping approaches take into account this correlation structure, they are often computationally intensive to run and cannot handle infinitesimal effects from non-causal variants. In this paper, we introduce BEATRICE, a novel framework for Bayesian fine-mapping from summary data. Our strategy is to impose a binary concrete prior over the causal configurations that can handle non-zero infinitesimal effects and to infer the posterior probabilities of the causal variant locations using deep variational inference. In a simulation study, we

demonstrate that BEATRICE achieves comparable or better performance to the current fine-mapping methods across increasing numbers of causal variants and increasing noise, as determined by the polygenecity of the trait.

1 Introduction

Genome-Wide Association Studies (GWAS) provide a natural way to quantify the contribution each genetic variant to the observed phenotype [1]. However, the univariate nature of GWAS does not take into account the correlation structure shared between the genetic variants, which arises due to low genomic recombination of nearby DNA regions [2]. Consequently, strong correlations can inflate the effect size of a non-causal genetic variant, thus leading to false positive identifications [3] Fine-mapping [4, 5] addresses this problem by analyzing the correlation structure of the data to identify small subsets of causal genetic variants [5, 6]. These subsets, known as credible sets, capture the uncertainty of finding the true causal variant within a highly correlated region [7]. Unlike p-values, the corresponding posterior inclusion probabilities (PIPs) computed during fine-mapping can be compared across studies of different sample sizes.

Traditional fine-mapping methods can be grouped into two general categories. The first category uses a penalized regression model to predict the output phenotype based on the collection of genetic variants [8, 9]. Popular regularizations like LASSO [10] and Elastic Net [9] simultaneously perform effect size estimation while slowly shrinking the smaller effect sizes to zero. The drawback of penalized regression models is that they optimize phenotypic prediction and, due to the correlation structure, do not always identify the true causal variants. The second category relies on Bayesian modeling. Here, the phenotype is modeled as a linear combination of the genetic variants, with sparsity incorporated into the prior distribution for the model weights. Approximate inference techniques, such as Markov Chain Monte Carlo (MCMC) [11] and variational methods [12] have been used to infer the effect sizes, PIPs, and credible sets. While these approaches represent valuable contributions to the field, they require the raw genotype and phenotype information, which raises privacy and regulatory concerns, particularly in the cases of publicly shared datasets. MCMC sampling also requires a burn-in period, which adds a substantial (100X) runtime overhead.

In response to these concerns, fine-mapping approaches have moved towards using summary statistics, which can be easily shared across sites. For example, the works of [13–15] use a stochastic or exhaustive search to identify the posterior probabilities of the causal configurations. However, exhaustive search based methods are restricted by the number of assumed causal variants, as this leads to an exponential increase in the dimensionality of the approximate posterior distribution. Stochastic search approaches [13] are computationally less expensive, but, by construction, they cannot handle infinitesimal effects from non-causal variants. The most recent contribution to fine-mapping is SuSiE [16, 17], which estimates the variant effect sizes as a sum of "single effects". These "single effect" vectors contain one non-zero element representing a causal variant and are estimated using a Bayesian step-wise selection approach. SuSiE provides a simple framework to robustly estimate PIPs and credible sets; however, there is limited evidence for its performance given the presence of infinitesimal genetic effects. Such scenarios can appear due to polygenicity of the trait, trans-interactions of variants, or varying correlation structure of the genomic region.

In this paper, we introduce BEATRICE, a novel framework for **B**ayesian fin**E**-mapping from summ**A**ry da**T**a using deep va**R**iational **I**nferen**CE**. In contrast to sampling methods, we approximate the posterior distribution of the causal variants given the GWAS summary statistics as a binary concrete distribution [18, 19], whose parameters are estimated using a deep neural network. This unique formulation allows

BEATRICE to use computationally efficient gradient-based optimization to minimize the KL divergence between the proposal binary concrete distribution and the posterior distribution of the causal variants. In addition, our unique optimization strategy samples a representative set of causal configurations in the process of minimizing the empirical KL divergence; these configurations can be used to obtain the PIPs and the credible sets. We compare our model with two state-of-the-art fine-mapping approaches, SuSiE [16] and FINEMAP [13]. We perform an extensive simulation study and quantify the performance of each model across increasing numbers of causal variants and increasing noise, as determined by the degree to which non-causal variants explain the phenotype variance. The runtimes of both SuSiE and BEATRICE are less than one minute, in contrast FINEMAP requires significantly longer to converge. On average BEATRICE achieves 2.2 fold increase in coverage, 0.1 fold increase in AUPRC, and similar power in comparison to SuSiE and FINEMAP.

2 Description of the Method

2.1 Generative Assumptions of Fine-mapping

BEATRICE is based on a generative additive effect model. Formally, let $\mathbf{y} \in \mathbb{R}^{n\times 1}$ denote a vector of (scalar) quantitative traits across n subjects. The corresponding genotype data $\mathbf{X} \in \mathbb{R}^{n\times m}$ is a matrix, where m represents the number of genetic variants in the analysis. Without loss of generality, we assume that the the columns of $\mathbf{X}$ have been normalized to have mean 0 and variance 1, i.e., $\frac{1}{n}\sum_i \mathbf{X}_{ij} = 0$ and $\frac{1}{n}\sum_i \mathbf{X}_{ij}^2 = 1$ for $j = 1, \ldots, m$. The quantitative trait is generated as follows:

$$\mathbf{y} = \mathbf{X}\beta + \eta \qquad \eta \sim N\left(0, \frac{1}{\tau}\mathbf{I}_n\right), \tag{1}$$

where $\beta \in \mathbb{R}^{m\times 1}$ is the effect size, $\eta \in \mathbb{R}^{n\times 1}$ is additive white Gaussian noise with variance $\frac{1}{\tau}$, and $\mathbf{I}_n$ is the $n \times n$ identity matrix.

2.2 Genome Wide Association Studies (GWAS)

GWAS uses a collection of element-wise linear regression models to estimate the effect of each genetic variant. Mathematically, the GWAS effect sizes are computed as $\hat{\beta} = \frac{1}{n}\mathbf{X}^T\mathbf{y}$, with the corresponding vector of normalized z-scores equal to $\mathbf{z} = \sqrt{\frac{\tau}{n}}\mathbf{X}^T\mathbf{y}$ [1, 14]. The main drawback of GWAS is that non-causal genetic variants can have large effect sizes due to polygenicity of the quantitative trait [20], varying degrees of linkage disequilibrium (LD) with causal variants [3], and/or interactions of the variant with enriched genes [20]. One popular strategy to mitigate this drawback is to impose a sparse prior over β given the set of causal variants:

$$\beta \sim N(0, \frac{1}{\tau}\sigma^2\mathbf{\Sigma}_C) \tag{2}$$

$$\mathbf{\Sigma}_C(i,j) = \begin{cases} 0, & i \neq j \\ 1, & i = j \text{ and } i \text{ is causal} \\ \epsilon, & i = j \text{ and } i \text{ is non-causal with non-zero effect} \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

Notice from Eq. (3) that the variance of $\beta(i)$ for a causal variant is $\frac{\sigma^2}{\tau}$ and the variance of $\beta(i)$ for a non-causal variant with non-zero effect is $\epsilon\frac{\sigma^2}{\tau}$, where ϵ is assumed

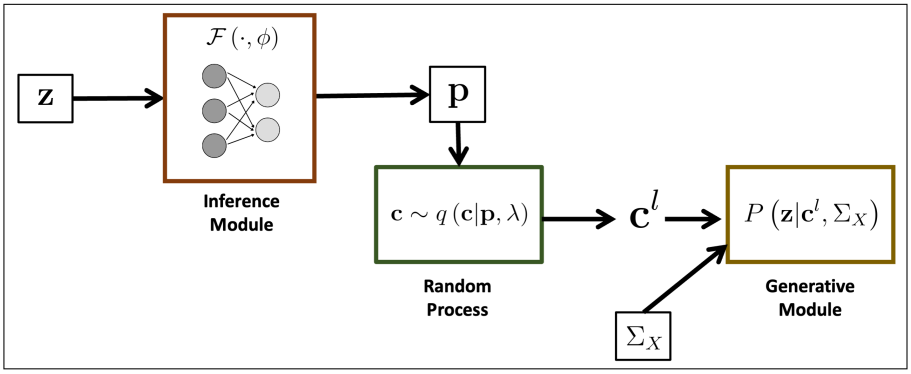


Fig 1. Overview of BEATRICE. The inputs to our framework are the LD matrix $\mathbf{\Sigma}_X$ and the summary statistics $\mathbf{z}$. The inference module uses a neural network to estimate the underlying probability map $\mathbf{p}$. The random process generates random samples $\mathbf{c}^l$ for the Monte Carlo integration in Eq. (12). Finally, the generative module calculates the likelihood of the summary statistics from the sample causal vectors $\mathbf{c}^l$.

to be small. This formulation handles residual influences from the non-causal variants, which are often observed in real-world data. Under this assumed prior, we can show [14, 21] that the normalized GWAS effect sizes $\mathbf{z}$ are distributed as:

$$p(\mathbf{z}|\mathbf{\Sigma}_X, \mathbf{\Sigma}_C) = N\left(\mathbf{z}; 0, \mathbf{\Sigma}_X + \mathbf{\Sigma}_X \left(n\,\sigma^2\,\mathbf{\Sigma}_C\right)\mathbf{\Sigma}_X\right) \tag{4}$$

where $\mathbf{\Sigma}_X = \frac{1}{n}\mathbf{X}^T\mathbf{X}$ is the empirical correlation matrix of the genotype data, also known as the LD matrix. Broadly, the goal of fine-mapping is to identify the diagonal elements of Σ_C that corresponds to 1 given the effect sizes $\mathbf{z}$ and the LD matrix $\mathbf{\Sigma}_X$. The derivation is provided in **Section S1.1** of the Supplement.

2.3 The Deep Bayesian Variational Model

BEATRICE uses a variational inference framework for fine-mapping. For convenience, we represent the diagonal elements of $\mathbf{\Sigma}_C$ by the vector $\mathbf{c} \in \mathbb{R}^{m\times 1}$, and by construction, $\mathbf{c}$ encodes the causal variant locations. Figure 1 provides an overview of BEATRICE. Our framework consists of three main components: an inference module, a random sampler, and a generative module. The inputs to BEATRICE are the summary statistics $\mathbf{z}$ and the LD matrix $\mathbf{\Sigma}_X$. The inference module estimates the parameters $\mathbf{p}$ of our proposal distribution $q(\cdot; \mathbf{p}, \lambda)$ using a neural network. The random process sampler uses the parameters $\mathbf{p}$ to randomly sample potential causal vectors $\mathbf{c}$ according to the given proposal distribution. Finally, the generative module calculates the likelihood of the observed summary statistics $\mathbf{z}$ according to Eq. (4).

2.3.1 Proposal Distribution

The goal of fine-mapping is to infer the posterior distribution $p(\mathbf{c}|\{\mathbf{z}, \mathbf{\Sigma_X}\})$, where $\mathbf{c}$ corresponds to the diagonal elements of $\mathbf{\Sigma}_C$. Due to the prior formulation in Eqs. (2-3), solving for the true posterior distribution is computationally intractable, as it requires a combinatorial search over the possible causal configurations. Thus, we approximate the posterior distribution $p(\mathbf{c}|\{\mathbf{z}, \mathbf{\Sigma_X}\})$ with a binary concrete distribution $q(\mathbf{c}; \mathbf{p}, \lambda)$ [18], where the parameters $\mathbf{p}$ of the distribution are functions of the inputs $\{\mathbf{z}, \mathbf{\Sigma_X}\}$. Samples $\mathbf{c}$ generated under a binary concrete distribution can be viewed as continuous relaxations of independent Bernoulli random variables. This reparametrization [19] allows us to learn $\mathbf{p}$ from the data using standard gradient descent.

Formally, let $\mathbf{c}_i$ and $\mathbf{p}_i$ denote the ith element of the vectors $\mathbf{c}$ and $\mathbf{p}$, respectively. Each entry of $\mathbf{c}$ is independent and is distributed as follows:

$$q(\mathbf{c}_i; \mathbf{p}_i, \lambda) = \frac{\lambda \mathbf{p}_i \mathbf{c}_i^{-\lambda-1}(1-\mathbf{p}_i)(1-\mathbf{c}_i)^{-\lambda-1}}{\left(\mathbf{p}_i \mathbf{c}_i^{-\lambda} + (1-\mathbf{p}_i)(1-\mathbf{c})^{-\lambda}\right)^2}, \tag{5}$$

where the parameter λ controls the extent of relaxation from a Bernoulli distribution.

We can easily sample from the binary concrete distribution in Eq. (5) via

$$\mathbf{c}_i = \xi\left(\frac{\log\left(\frac{U}{1-U}\right) + \log\left(\frac{\mathbf{p}_i}{1-\mathbf{p}_i}\right)}{\lambda}\right), \tag{6}$$

where $\xi(\cdot)$ is the sigmoid function, and the random variable U is sampled from a uniform distribution over the interval $[0, 1]$. As seen, $\mathbf{p}_i$ specifies the underlying probability map and U provides stochasticity for the sampling procedure in Eq. (6). We note that the gradient of Eq. (6) with respect to $\mathbf{p}_i$ tends to have a low variance in practice, which helps to stabilize the optimization.

Intuitively, every element of the binary concrete random vector $\mathbf{c}$ can be regarded as a continuous relaxation from a Bernoulli random variable. Specifically, the parameter $\mathbf{p}$ captures the underlying probability map, analogous to the selection probability of a Bernoulli distribution. The parameter λ controls the extent of relaxation from the 0/1 Bernoulli distribution, such that increasing λ results in a smoother transition between the extremal values $\{0, 1\}$. This continuous representation allows us to model the infinitesimal effects of the non-causal variants. Additionally, the underlying probability map $\mathbf{p}$ captures the relative importance of a variant containing a causal signal. The two unique properties of the probability maps are $P(\mathbf{c}_i > \frac{1}{2}) = \mathbf{p}_i$ and $\lim_{\lambda\to 0} P(\mathbf{c}_i = 1) = \mathbf{p}_i$. The first property indicates that $\mathbf{p}_i$ controls the degree to which $\mathbf{c}_i$ assumes low values close to 0 and high values close to 1. This property also give BEATRICE flexibility to handle genetic variants with different levels of association, thus aligning with our generative process that assumes some non-causal variants may have small non-zero effects. The second property implies that a high probability $\mathbf{p}_i$ at location i is highly indicative of a causal variant. Taken together, the binary concrete distribution has an easily-optimized parameterization with desirable properties.

2.3.2 Variational Inference

We select the variational parameters $\{\mathbf{p}, \lambda\}$ to minimize the Kullback–Leibler (KL) divergence between the proposal distribution and the posterior distribution of the causal vector $\mathbf{c}$ given the input data $\{\mathbf{z}, \mathbf{\Sigma_X}\}$, that is

$$\{\mathbf{p}^*, \lambda^*\} = \arg\min_{\{\mathbf{p},\lambda\}} KL\left(q(\mathbf{c}; \mathbf{p}, \lambda) \,||\, p(\mathbf{c}|\{\mathbf{z}, \mathbf{\Sigma_X}\})\right) \tag{7}$$

Using Bayes' Rule, we can show that the optimization in Eq. (7) can be rewritten

$$\{\mathbf{p}^*, \lambda^*\} = \arg\min_{\{\mathbf{p},\lambda\}} KL\left(q(\mathbf{c}; \mathbf{p}, \lambda) \,||\, p(\mathbf{c}; \mathbf{p}_0, \lambda_0)\right)) - E_{q(\cdot;\mathbf{p},\lambda)}\left[\log\left(p(\mathbf{z}|\mathbf{\Sigma_X}, \mathbf{c})\right)\right], \tag{8}$$

where we have assumed an element-wise binary concrete prior $p(\mathbf{c}; \mathbf{p}_0, \lambda_0)$ over the vector $\mathbf{c}$. We fix the relaxation parameter to be small ($\lambda = 0.01$) and the probability map to be uniform $\mathbf{p}_0 = \left[\frac{1}{m}, \ldots, \frac{1}{m}\right]^{\mathrm{T}}$. Thus, the first term of Eq. (8) can be viewed as a regularizer that encourages sparsity in causal vectors $\mathbf{c}$. The second term of Eq. (8) can be interpreted as the likelihood of the observed test statistics. The works of [22, 23] have demonstrated that under certain assumptions, the likelihood term of the summary statistics is the same as the original data likelihood $p(\mathbf{y}|\mathbf{X}, \mathbf{c})$ derived from Eq. (1).

During optimization, the relaxation parameter λ is annealed [18, 19] to a small non-zero value (0.01) with fixed constant rate, and the underlying probability map $\mathbf{p}$ is optimized using gradient descent. Specifically, we use a neural network to generate the vector $\mathbf{p} = \mathcal{F}(\mathbf{z}; \phi)$. The details of the neural network architecture are provided in **Section S1.4** of the Supplement. Practically speaking, the neural network ties the input data $\{\mathbf{z}, \Sigma_X\}$ to the parameter space of the proposal distribution in a data-driven fashion. Empirically, we find that generating $\mathbf{p}$ as a function of the input data regularizes the model and leads to a stable optimization.

Optimizing $\mathbf{p}^*$ now amounts to learning the parameters of the neural network ϕ. Given a fixed value of λ, the neural network loss function follows from Eq. (8) as

$$\mathcal{L}(\phi) = KL\left(q(\mathbf{c}; \mathbf{p}(\phi), \lambda) \,||\, p(\mathbf{c}; \mathbf{p}_0, \lambda_0)\right)) - E_{q(\cdot;\mathbf{p}(\phi),\lambda)}\left[\log\left(p(\mathbf{z}|\mathbf{\Sigma_X}, \mathbf{c})\right)\right], \tag{9}$$

where we have defined $\mathbf{p}(\phi) \triangleq \mathcal{F}(\mathbf{z}; \phi)$ for notational convenience.

2.3.3 Optimization Strategy

The expectations in Eq. (9) do not have closed-form expressions. Therefore, we use Monte Carlo integration to accurately approximate $\mathcal{L}(\phi)$ in the regime of small λ, i.e., when the binary concrete distribution behaves similar to a Bernoulli distribution.

Let $\mathbf{c}^1(\phi), \ldots, \mathbf{c}^L(\phi)$ be a collection of causal vectors sampled independently from $q(\cdot|\mathbf{p}(\phi), \lambda)$ according to Eq. (6). The likelihood term of Eq. (9) is computed as

$$E_{q(\cdot;\mathbf{p}(\phi),\lambda)}\left[\log\left(p(\mathbf{z}|\mathbf{\Sigma_X}, \mathbf{c})\right)\right] = \frac{1}{L}\sum_{l=1}^{L} \log\left(p\left(\mathbf{z}|\mathbf{\Sigma_X}, \mathbf{c}^l(\phi)\right)\right), \tag{10}$$

where the right-hand side probability is computed according to Eq. (4) by substituting $\mathbf{c}^l(\phi)$ for the diagonal entries of $\mathbf{\Sigma}_C$ in each term of the summation. Once again, the continuous relaxation used to generate $\mathbf{c}^l(\phi)$ in Eq. (6) allows us to directly optimize ϕ.

We approximate the first term of Eq. (9) under the assumption of small $\{\lambda, \lambda_0\}$ on the order of 0.01. In this case, the binary concrete distribution behaves like a $\{0, 1\}$ Bernoulli distribution. Under these conditions, we can write the first term of Eq. (9) as

$$\begin{aligned} &KL\left(q(\mathbf{c}; \mathbf{p}(\phi), \lambda) \,||\, p(\mathbf{c}; \mathbf{p}_0, \lambda_0)\right)) \\ &\quad \approx \sum_{i=1}^{m}\left[\mathbf{p}_i(\phi)\log\left(\frac{\mathbf{p}_i(\phi)}{p_0}\right) + (1-\mathbf{p}_i(\phi))\log\left(\frac{1-\mathbf{p}_i(\phi)}{1-p_0}\right)\right], \end{aligned} \tag{11}$$

where p_0 is a fixed scalar parameter used to construct the (constant) prior vector $\mathbf{p}_0$. We note that the criteria $\{\lambda \to 0.01, \lambda_0 = 0.01\}$ is satisfied in practice, as λ is annealed during the optimization to progressively smaller values and λ_0 is fixed *a priori*.

The above approximations allow us to rewrite the neural network loss as

$$\begin{aligned} \mathcal{L}(\phi) \approx &-\frac{1}{L}\sum_{l=1}^{L}\log N\left(\mathbf{z}; 0, \mathbf{\Sigma}_X + \mathbf{\Sigma}_X\left(n\,\sigma^2\,\mathbf{\Sigma}_C^l(\phi)\right)\mathbf{\Sigma}_X\right) \\ &+ \sum_{i=1}^{m}\mathbf{p}_i(\phi)\log\left(\frac{\mathbf{p}_i(\phi)}{p_0}\right) + (1-\mathbf{p}_i(\phi))\log\left(\frac{1-\mathbf{p}_i(\phi)}{1-p_0}\right), \end{aligned} \tag{12}$$

where $\mathbf{\Sigma}_C^l(\phi)$ corresponds to the diagonal matrix using the vector $\mathbf{c}^l(\phi)$ as the diagonal entries. We use a stochastic gradient descent optimizer [24] to minimize the loss $\mathcal{L}(\phi)$ with respect to the neural network weights ϕ. This process is detailed in Algorithm 1.

2.3.4 Computational Complexity

Each iteration of stochastic gradient descent requires us to compute the data log-likelihood term $\left[\log N\left(\mathbf{z}; 0, \mathbf{\Sigma}_X + \mathbf{\Sigma}_X\left(n\,\sigma^2\,\mathbf{\Sigma}_C^l(\phi)\right)\mathbf{\Sigma}_X\right)\right]$. This computation is expensive due to the covariance matrix inversion, whose run-time is on the order of $O(m^3)$, where m is the total number of variants. To mitigate this issue, the works of [25] show that if $\mathbf{\Sigma}_C^l(\phi)$ is sparse, then the matrix inversion can be done with order $O(k^3) + O(mk^2)$ run-time, where k is the number of non-zero diagonal elements of $\mathbf{\Sigma}_C^l(\phi)$. We leverage this result in the optimization by thresholding the elements of $\mathbf{c}^l(\phi)$ to set small values exactly to zero. In every iteration, we sparsify $\mathbf{c}_t^l$ by considering the top 50 non-zero locations of $\mathbf{c}_t^l$ with values $\mathbf{c}_t^l(i) > 0.01$. This strategy provides a way to optimize the parameters of our models in $O(50^3) + O(m50^2)$ run-time for all scenarios. We also regularize $\mathbf{\Sigma}_X$ with a small diagonal load to ensure invertibility of the covarance matrix at each iteration. Finally, we run stochastic gradient descent with a batch size of one to further speed up BEATRICE. Effectively, this means that we sample a single $\mathbf{c}^l(\phi)$ at each epoch rather than perform a true Monte Carlo integration. The authors of [26] have previously shown that a single random sample ($L = 1$) is sufficient to guarantee convergence to a local minimum of Eq. (12). Algorithm 1 provides a detailed description of these optimization steps.

3 Verification and Comparison

3.1 Causal Configurations and Posterior Inclusion Probabilities

The desired outputs of each fine-mapping method are Posterior Inclusion Probabilities (PIPs) and credible sets. PIPs estimate how likely each variant is causal as a measure of its importance. Credible sets identify the subset of variants that are likely to contain a causal variant, which captures the uncertainty of finding the true variant.

The main challenge to estimating the posterior probability of a given causal configuration (i.e., set of causal variant locations) is the exponentially large search space. Let $\mathbf{b}$ denote a binary vector with a value of 1 at causal locations and a value of 0 at non-causal locations. At a high level, $\mathbf{b}$ can be viewed as a binarized version of the causal vector $\mathbf{c}$ in the previous sections. Using Bayes' Rule, the posterior probability of $\mathbf{b}$ given the input data $\{\mathbf{z}, \mathbf{\Sigma_X}\}$ can be written as follows:

$$p\left(\mathbf{b}|\mathbf{z}, \mathbf{\Sigma}_X\right) = \frac{p\left(\mathbf{z}|\mathbf{\Sigma}_X, \mathbf{b}\right) p\left(\mathbf{b}\right)}{\sum_{\mathbf{b}' \in \mathcal{B}} p\left(\mathbf{z}|\mathbf{\Sigma}_X, \mathbf{b}'\right) p\left(\mathbf{b}'\right)} \tag{13}$$

Algorithm 1 Optimization scheme to minimize Eq. (12)

$\mathcal{B}^R = \{\}$
Initialize ϕ_0
for $t = [1 \ldots T]$ **do**
  Generate $\mathbf{p}(\phi_t) = \mathcal{F}(\mathbf{z}; \phi_t)$
  Randomly sample $\mathbf{c}_t^l$ according to Eq. (6)
  Binarize $\mathbf{c}_t^l$ to $\mathbf{b}_t^l$ and add to $\mathcal{B}^R$
  $\mathcal{S}_t^l = \{i\} \;\; s.t. \;\; \mathbf{c}_t^l(i) > 0.01$
  Prune set $\mathcal{S}_t^l$ such that it consists of 50 indices.
  $\mathbf{c}_t^l(i) = 0$ if $i \notin \mathcal{S}_t^l$
  Generate $\mathcal{L}(\phi_t)$ according to Eq. (12)
  $\phi_{t+1} = \phi_t - StepSize \nabla \mathcal{L}(\phi^t)$
end for

where $\mathcal{B}$ is the set of all 2^m possible causal configurations. Once again, $\mathbf{z}$ captures the summary statistics and $\mathbf{\Sigma}_X$ is the LD matrix. Even though $\mathcal{B}$ is exponentially large, it has been argued [27] that the majority of these configurations have negligible probability and do not contribute to the denominator of Eq. (13).

Our stochastic optimization provides a natural means to track causal configurations with non-negligible probability to compute $p\left(\mathbf{b}|\mathbf{z}, \mathbf{\Sigma}_X\right)$. Namely, at each iteration of stochastic gradient descent, we randomly generate a sample causal vector $\mathbf{c}^l$ to minimize Eq. (12). In parallel, we binarize the vector $\mathbf{c}^l$ via

$$\mathbf{b}_i^l = \begin{cases} 1, & \mathbf{c}_i^l > \gamma, \\ 0, & otherwise \end{cases}$$

and add the resulting vector $\mathbf{b}^l$ to a reduced set of causal configurations $\mathcal{B}^R$. The variational objective ensures that our proposal distribution converges to the true posterior distribution of the causal vectors. Thus, the samples $\mathbf{c}^l$ lie near modes of the posterior distribution which is the neighborhood of non-negligible probability.

In this work, we use a threshold $\gamma = 0.1$ to binarize the vectors $\mathbf{c}^l$. Thresholding at $\gamma = 0.1$ only considers variants whose estimated effect size variance is $> 0.1\sigma^2$. This operation prunes out spurious non-causal configurations generated by the non-causal variants. A higher threshold is beneficial in the presence of high interaction effects from non-causal variants and a lower threshold could be useful when the causal variants are weakly associated with the outcome. Empirically, we find that the threshold value of 0.1 preserves the main interactions between variants. However, the user of BEATRICE can adjust this threshold as needed.

After obtaining the sampled vectors, we replace the exhaustive set $\mathbf{B}$ in Eq. (13) with the reduced set $\mathcal{B}^R$ for tractable computation of $p\left(\mathbf{b}|\mathbf{z}, \mathbf{\Sigma}_X\right)$. We then compute the posterior inclusion probability (PIP) of each variant by summing the probabilities over the subset of $\mathcal{B}^R$ with a value of 1 at that variant location. Mathematically,

$$P(\mathbf{b}_i = 1|\mathbf{z}, \mathbf{\Sigma}_X) \approx \sum_{\mathbf{b} \in \mathcal{S}} p\left(\mathbf{b}|\mathbf{z}, \mathbf{\Sigma}_X\right) \tag{14}$$

$$\text{s.t. } \mathcal{S} \subset \mathcal{B}^R \text{ and } \mathcal{S} = \{\mathbf{b} \,|\, \mathbf{b}_i = 1\} \tag{15}$$

where $\mathcal{S}$ is a subset of $\mathcal{B}^R$ that contains binary configurations with 1 at location i.

Finally, we identify the credible sets in two steps. First, in a conditional step-wise fashion, we identify the variants with the highest conditional probability given the previously selected variants. This strategy identifies the set of "key" variants with a high probability of being causal. Second, we determine the credible set for each key variant, by computing the conditional inclusion probabilities of each variant given the key variants and adding variants to the credible set. A detailed description of this process can be found in the Supplementary Methods document (**Section S1.3 in S1 text**) .

3.2 Baselines

We compare our approach with the state-of-the-art methods, FINEMAP and SuSiE.

FINEMAP: This approach uses a stochastic shotgun search to identify causal configurations with non-negligible posterior probability. FINEMAP defines the neighborhood of a configuration at every step by deleting, changing or adding a causal variant from the current configuration. The next iteration samples from this neighborhood, thus reducing the exponential search space to a smaller high-probability region. Finally, the identified causal configurations are used to determine the posterior

inclusion probabilities for each variant. The computationally efficient shotgun approach makes FINEMAP a viable tool for finemapping from multiple GWAS summary data in [28, 29]. We prune the credible sets of FINEMAP [13] via the approach used in [16] for this task. Details of this procedure are provided in **Section S1.5** of the Supplement.

SuSiE: The recent works of [16, 17] introduced an iterative Bayesian selection approach for fine-mapping that represents the variant effect sizes as a sum of "single-effect" vectors. Each vector contains only one non-zero element, which represents the causal signal. In addition to finding causal variants, SuSiE provides a way to quantify the uncertainty of the causal variants locations via credible sets. SuSiE has also been used widely to find putative causal variants GWAS summary statistics [30, 31].

3.3 Evaluation Strategy

We evaluate several metrics of performance in our simulation study.

Evaluating PIPs: We have compared the quality of the PIPs via the AUPRC metric. AUPRC (area under the precision-recall curve) is computed by sweeping a threshold on the PIPs and computing precision and recall against the true configuration of causal and non-causal variants. High precision indicates a low false positive rate in the estimated causal variants. High recall indicates that the model correctly identifies more of the causal variants. Thus, the AUPRC, can be viewed as a holistic measure of performance across both classes. AUPRC is also robust to severe class imbalance [32], which is the case in fine-mapping, as the number of causal variants is small. Additionally, we visualize power vs. FDR for different thresholds of the PIPs. Following standard nomenclature, both power and recall measures the probability of detection ($\frac{TP}{P}$) and FDR (1-precision) measures the type-I error ($\frac{FP}{PP}$). Therefore, the power vs. FDR curve provides a visual comparison, while AUPRC gives the numeric quantification of the performances.

Coverage, Power and Size of the Credible Sets: We follow the strategy of [16, 17] to define a credible set. A credible set is defined as a collection of variants that contain a single causal variant with a probability equal to the coverage. Given that the number of causal variants can be arbitrary, we use two metrics to assess the quality of the credible sets. Specifically, coverage is the percentage of credible sets that contain a causal variant, and power is the percentage of causal variants identified by *all* the credible sets. Higher coverage indicates that the method is confident about its prediction of *each* causal variant, whereas higher power indicates the method can accurately identify all the causal variants.

One caveat is that a method can generally achieve both higher coverage and higher power simply by adding variants to the credible sets. To counter this trend, we report the average size of the credible sets identified by each method. Ideally, we would like the credible sets to be as small as possible while retaining high coverage and high power.

4 Applications

4.1 Setup for Simulation Experiments

Genotype Simulations: We use the method of [33] to simulate genotypes $\mathbf{X}$ based on data from the 1000 Genomes Project. We select an arbitrary sub-region ($39.9Mb - 40.9Mb$) from Chromosome 2 as the base. After filtering for rare variants (MAF < 0.02), the remaining 3.5K variants are used to simulate pairs of haplotypes to

generate 10,000 unrelated individuals. We chose a MAF threshold of 0.02, as it lies in the middle of the range $0.01 - 0.05$ commonly used in GWAS studies [34]. In each experiment below, we randomly select $m = 1000$ variants and $n = 5000$ individuals to generate the phenotype data.

Phenotype Generation: We generate the phenotype $\mathbf{y}$ from a standard mixed linear model [22], where the influences of the causal variants are modeled as fixed effects, and the influences of other non-causal variants are modeled as random effects. In this case, the genetic risk for a trait is spread over the entire dataset, with each variant having small individual effects, as per the polygenicity assumption of a complex trait. We randomly select the causal variants in our simulations. Thus, some simulations will have causal variants in LD, while others will select causal variants with low correlation.

Given a set of d causal variants C, let $\mathbf{X}_C \in \mathbf{R}^{n \times d}$ denote the corresponding subset of the genotype data and $\mathbf{X}_{NC} \in \mathbf{R}^{n \times m-d}$ denote the remaining non-causal variants. From here, we generate the phenotype data $\mathbf{y}$ as follows:

$$\begin{aligned}
\mathbf{y} &= \mathbf{X}_C\, \beta + \mathbf{g}_{NC} + \epsilon \triangleq \mathbf{g}_C + \mathbf{g}_{NC} + \epsilon \\
\mathbf{g}_{NC} &\sim N\left(0, \frac{1}{m-d}\, \mathbf{X}_{NC}\, \mathbf{X}_{NC}^T\right) \\
\beta &\sim N\left(0, \mathbf{I}_d\right) \\
\epsilon &\sim N\left(0, \alpha^2 \mathbf{I}_n\right)
\end{aligned}$$

where β is the d-dimensional effect sizes sampled from a Gaussian, and ϵ is an zero-mean Gaussian noise with variance α^2. The random variable $\mathbf{g}_{NC}$ models the effect of the non-causal variants as a multivariate Gaussian vector with mean 0 and covariance $\frac{1}{m-d}\mathbf{X}_{NC}\mathbf{X}_{NC}^T$. Likewise, $\mathbf{g}_C = \mathbf{X}_C\, \beta$ captures the effect of the causal variants.

In our experiments, we define ω^2 as the total phenotypic variance attributed to the genotype (e.g., both $\mathbf{g}_C$ and $\mathbf{g}_{NC}$) and p as the proportion of this variance associated with the causal variants in $\mathbf{g}_C$. Using the strategy described in [35], we enforce these conditions by normalizing the phenotype $\mathbf{y}$ as follows:

$$\begin{aligned}
\tilde{\mathbf{y}} &= \sqrt{\frac{p\,\omega^2}{var(\mathbf{g}_C)}}\, \mathbf{g}_C + \sqrt{\frac{(1-p)\,\omega^2}{var(\mathbf{g}_{NC})}}\, \mathbf{g}_{NC} + \tilde{\epsilon} \qquad (16) \\
\tilde{\epsilon} &\sim N(0, (1-\omega^2)\mathbf{1}_n)
\end{aligned}$$

where $var(\mathbf{g}_C)$ and $var(\mathbf{g}_{NC})$ are the empirical variances of $\mathbf{g}_C$ and $\mathbf{g}_{NC}$, respectively.

In order to replicate real-world scenarios, we use a GWAS framework to estimate the effect size $\hat{\beta}_i$ of each variant i based on the phenotype $\tilde{\mathbf{y}}$ and the un-normalized genotype data $\tilde{\mathbf{X}}$. Specifically, we use a simple linear model and ordinary least squares to estimate the effect sizes. From here, we convert the estimated effect sizes to z-scores via $\mathbf{z}_i = \frac{\hat{\beta}_i}{se\left(\hat{\beta}_i\right)}$, where $se(\cdot)$ denotes the standard error. The LD matrix is computed as $\mathbf{\Sigma}_X = \frac{1}{n}\mathbf{X}^T\mathbf{X}$, where $\mathbf{X}$ is the normalized genotype data. The z-scores and LD matrix are input to each of the fine-mapping methods above.

Noise Configurations: We evaluate the performance of each method while varying the number of causal variants d, the total genotype variance ω^2, and the proportion of this variance associated with the causal variants p. Formally, we sweep over one order of magnitude for $d = [1, 4, 8, 12]$, $\omega^2 = [0.1, 0.2, 0.4, 0.5, 0.7, 0.8]$, and $p = [0.1, 0.3, 0.5, 0.7, 0.9]$. For each noise setting, we randomly generate 20 datasets by

independently re-sampling the causal variant locations, the effect sizes $\{\beta_i\}$, the non-causal component $\mathbf{g}_{NC}$, and the noise $\tilde{\epsilon}$. We run all three fine-mapping methods over a total of $4 \times 6 \times 5 \times 20 = 2400$ configurations for a comphrehensive evaluation.

4.2 Application to Real-World SNP Data

We compare the performance of each finemapping method on a GWAS study of Alzheimer's Disease. Alzheimer's Disease is a polygenic disorder, making it an ideal test bed to evaluate each model. We use the publicly available GWAS summary statistics released by [36] to obtain the z-scores and the 1000 Genome data to generate the LD matrices. We compare the finemapping results of SuSiE and BEATRICE. Notice that we cannot run FINEMAP because the reported GWAS statistics only contain the z-scores, whereas FINEMAP requires the effect sizes $\hat{\beta}_i$ and the corresponding standard errors to perform finemapping. This real-world study also highlights a major drawback of FINEMAP, should the original effect sizes and standard errors be unavailable.

We run SuSiE in their default setting. However, for BEATRICE we set $\gamma_{key} = 0.5$ to identify sparse and highly selective credible sets. The other parameters are set at their default values.

Data Acquisition: The recent GWAS analysis performed by [36] identified multiple statistically significant index SNPs in Alzheimer's Disease. We have restricted our analysis to the 5 index SNPs with the smallest p-values and their corresponding clumped variants. We only use the SNPs from GWAS statistics that overlap with the publicly available 1000 Genome Phase-3 data.

Data Preprocessing: We first filter out the variants not present in the 1000 Genome database. We further removed the ambiguous SNPs (i.e. those with complementary alleles, either C/G or A/T SNPs) due to the lack of strand information. After filtering, we ensured the counted alleles were identical in the GWAS summary statistics and the 1000 Genome data. We identify the clumps using 1000 Genome data and the GWAS statistics, where each clump contains SNPs less than 250KB away with r2 > 0.1 and p-value < 0.001. This choice of clumping setting is widely used in other studies [29]. Finally, we merge overlapping clumped regions into a single region. The information about the clumped variants is reported in Supplementary Table 1. The resulting variants within each clump are then used to generate LD matrices from the 1000 Genome data.

4.3 Results

4.3.1 Simulated Experiments

Varying the Number of Causal Variants: Figure 2 illustrates the performance of each method (BEATRICE, FINEMAP, and SuSiE) while increasing the number of causal variants from $d = 1$ to $d = 12$. The points denote the mean performance across all noise configurations (ω^2, p) for fixed d, and the error bars represent the 95% confidence interval across these configurations. We note that BEATRICE achieves a uniformly higher AUPRC than both baseline methods, which suggests that BEATRICE can better estimate the PIPs than FINEMAP or SuSiE. BEATRICE also provides a $0.9 - 1.4$-fold increase in coverage than the baselines while maintaining a similar power, which indicates that the credible sets generated by BEATRICE are more likely to contain a causal variant as compared to SuSiE and FINEMAP. Finally, we note that BEATRICE , identifies the same or smaller credible set sizes than FINEMAP, and SuSiE. Taken together, as the number of causal variants increases, BEATRICE gives us

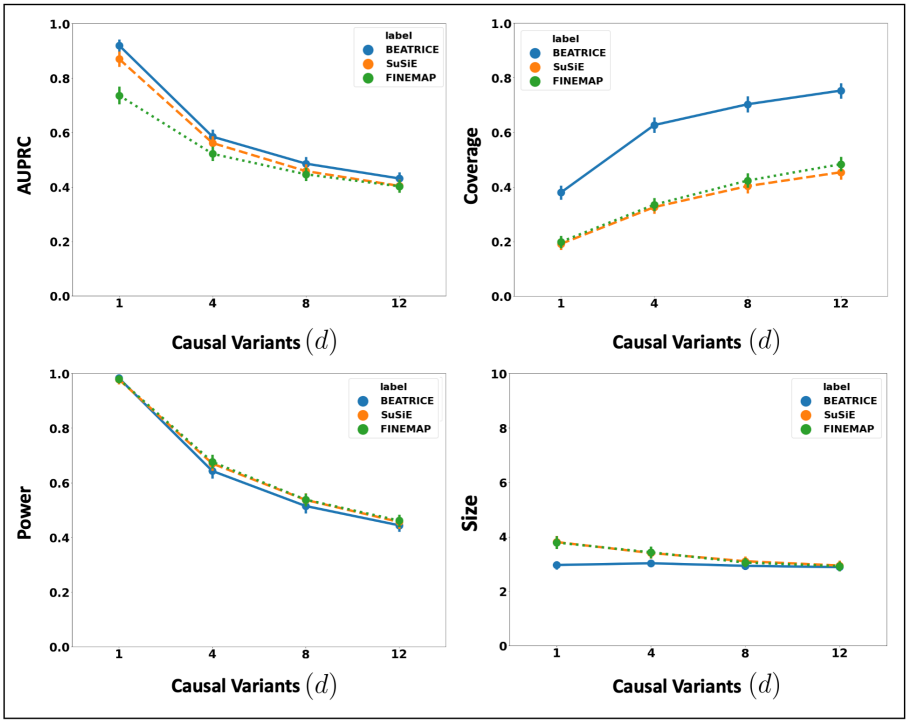


Fig 2. The performance metrics for the three methods across varying numbers of causal variants. Along the x-axis, we plot the number of causal variants, and across the y-axis, we plot the mean and confidence interval (95%) of each metric. We calculate the mean by fixing d to a specific value $d = d^*$ and sweep over all the noise settings where $d = d^*$.

a better estimate of the PIPs and arguably better credible sets. Unlike the baseline methods, BEATRICE does not impose any prior assumptions over the total number of causal variants, which may lead to its improved performance.

Increasing the Genotype Contribution: Figure 3 shows the performance of each method while increasing the genetically-explained variance from $\omega^2 = 0.1$ to $\omega^2 = 0.8$. Similar to the above experiment, the points in the figure denote the mean performance across all other noise configurations (d, p) for fixed ω^2, and the error bars represent the 95% confidence intervals across these configurations. We note that BEATRICE achieves a significantly higher AUPRC than FINEMAP and a slightly higher AUPRC than SuSiE. When evaluating the credible sets, we observe similar trends in coverage (BEATRICE is $0.25 - 2.34$ folds higher) and power (similar performance across methods). All three methods identify credible sets of similar size. We submit that BEATRICE achieves the best trade-off across the four performance metrics.

Varying the Contributions of Causal and Non-Causal Variants: Figure 4 illustrates the performance of each method while increasing the contribution of the causal variants from $p = 0.1$ to $p = 0.9$. Once again, the points denote the mean performance across all other noise configurations (d, ω^2) for fixed p, and the error bars represent the 95% confidence intervals across these configurations. From an application standpoint, the presence of non-causal variants with small non-zero effects makes it difficult to detect the true causal variants. Accordingly, we observe a performance boost across all methods when p is larger. Similar to our previous experiments, BEATRICE provides the best AUPRC, with converging performance as $p \to 1$. In addition,

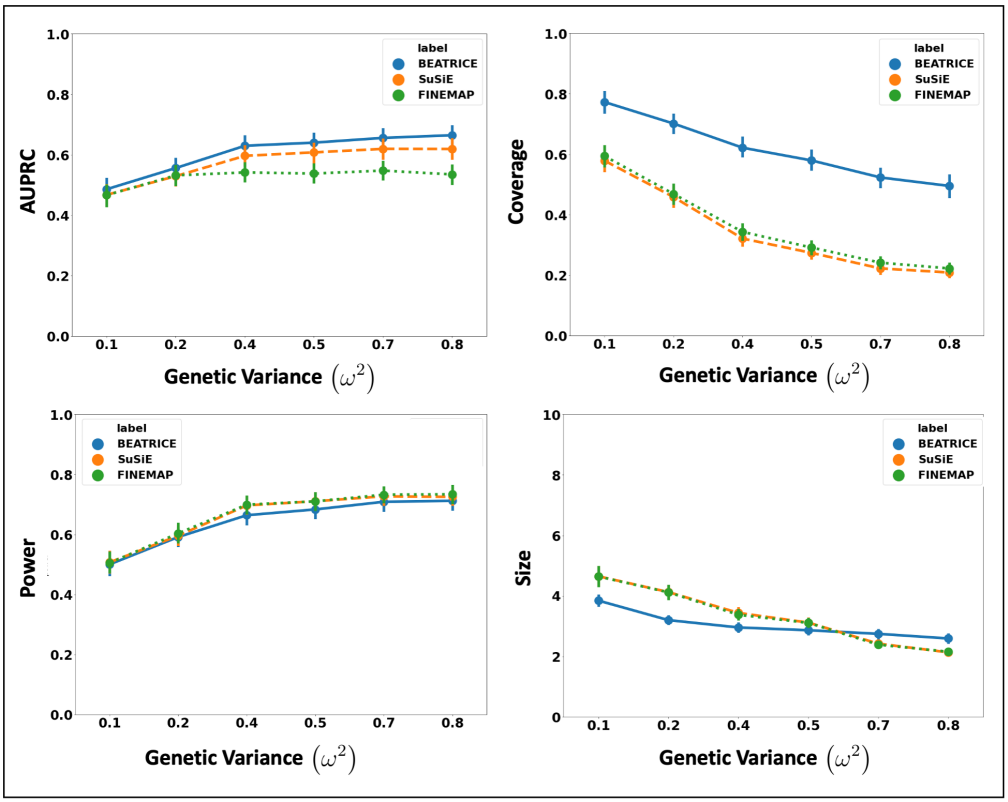


Fig 3. The performance metric for increasing phenotype variance explained by genetics. Along the x-axis, we plot the variance explained by genetics (ω^2), and across the y-axis, we plot each metric's mean and confidence interval (95%). We calculate the mean by fixing ω^2 to a specific value $\omega = \omega^*$ and sweep over all the noise settings where $\omega = \omega^*$.

BEATRICE identifies smaller credible sets with significantly higher coverage while maintaining power. Thus, we conclude that BEATRICE is the most robust of the three methods to the presence of noise from non-causal variants. This performance gain may arise from our binary concrete proposal distribution for the causal vector $\mathbf{c}$, which provides flexibility to accommodate varying degrees of association.

Power vs. FDR Performance Comparison: Figure 5 shows the mean performance of all the methods across 20 simulation settings with fixed SNP heritability $\omega^2 = 0.2$ and multiple causal variants $d = \{1, 4, 8, 12\}$ and multiple infinitesimal effects from non-causal SNPs $p = \{0.1, 0., 3, 0.5, 0.9\}$. We use linear interpolation to project the results across simulation runs onto the same x-y axes. When p is small, i.e., the scenario in which most of the phenotype variance can be explained by the infinitesimal effects from non-causal variants, we notice that BEATRICE gives the best performance in terms of power and FDR. This result shows the BEATRICE generates PIPs that are robust in the presence of high infinitesimal effects from nearby SNPs. Additional comparisons for other SNP heritability values are provided in Section S1.6 of the Supplementary Text.

4.3.2 Finemapping Results on Real-World Data

Figure 6 shows the absolute z-scores of the SNPs in each of the 5 clumps. The GWAS statistics published by [36] reported multiple SNPs in the first clump with z-scores of

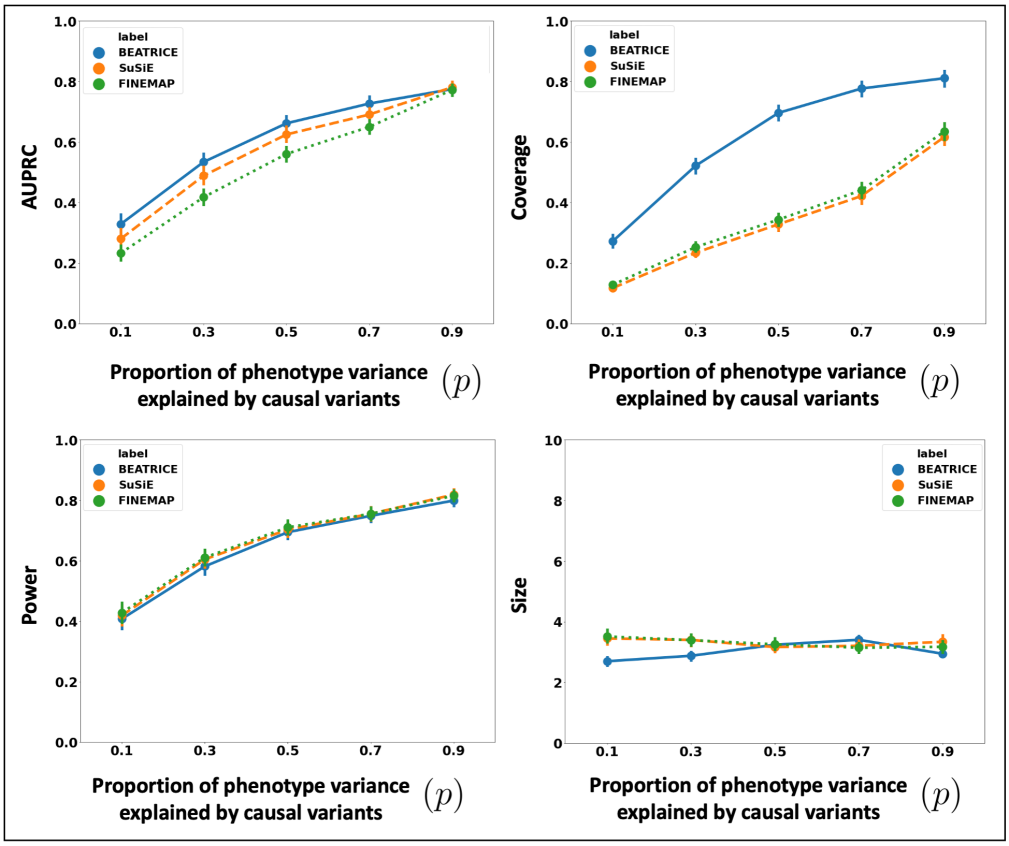


Fig 4. The performance metric for multiple levels of noise introduced by non-causal variants. The noise level (p) is explained by the variance ratio of non-causal variants vs. causal variants. Along the x-axis, we plot the noise level (p); across the y-axis, we plot each metric's mean and confidence interval (95%). We calculate the mean by fixing p to a specific value $p = p^*$ and sweep over all the noise settings where $p = p^*$.

inf. For numerical stability, we clip these z-scores to 200. In the other clumps, we notice multiple z-scores with high values. The inflation in the GWAS statistics could result from the infinitesimal effects of multiple SNPs within each clump. This scenario is similar to the simulation settings when p is small.

Figure 7 shows the PIPs identified by each finemapping method, colored by locus. The x-axis corresponds to the SNP index, and the y-axis reports the corresponding PIP values. In Figure 8, we provide the mean size of the credible sets identified by each method and the number of overlapping SNPs between credible sets of SuSiE and BEATRICE.

We further explored SNPs within the credible set of the first clump because that region overlies TOMM40, APOC1, and APOE genes, which are widely identified as potential disease-causing loci for Alzheimer's Disease [37–40]. In Supplementary Table 2 and Supplementary Table 3, we report the SNPs present in the credible sets identified byBEATRICE and SuSiE, respectively. We also provide the p-values from the GWAS statistics and the PIPs as identified by each method. Overall, we observe a two-fold increase in the number of SNPs with very small p-values (≈ 0) that BEATRICE finds, as compared to SuSiE.

In an exploratory analysis, we investigate the biological consequences of the SNPs present in the credible sets of the first clump as identified by each method. We extracted

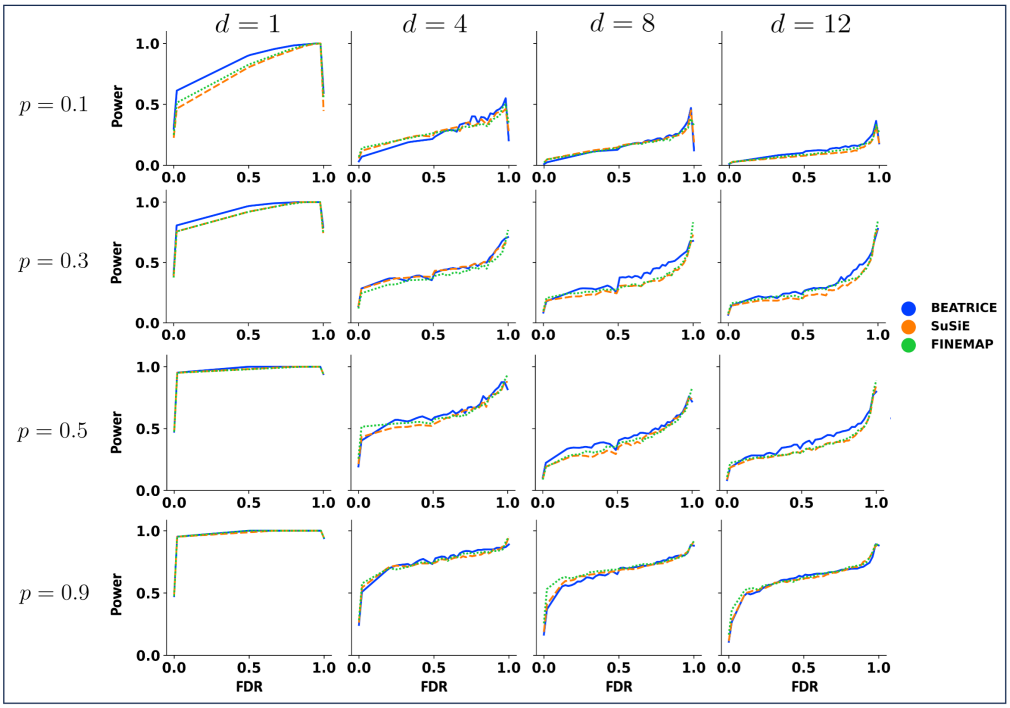


Fig 5. Power vs. FDR curve for three models across multiple causal variants $d = [1, 4, 8, 12]$, and multiple proportion of phenotype variance explained by causal variants $p = [0.1, 0.3, 0.5, 0.9]$, while fixing SNP heritability at $\omega^2 = 0.2$. Each row and column corresponds to a specific value of p and d, respectively. In each plot, the y-axis captures power, and the x-axis represents FDR.

all genes tagged by BEATRICE and SuSiE from the Ensemble VEP annotation, which expands the GENCODE boundaries by 5kb to account for upstream/downstream flanking regulatory regions. Supplementary Table 4 and Supplementary Table 5 show the tagged genes, and the biological consequences of the SNPs identified by BEATRICE and SuSiE, respectively. Both approaches tag genes that involve APOE, TOMM40 [38], APOC1 [39], and PVRL2 [40], all of which have been previously associated with Alzheimer's Disease. In addition, BEATRICE identify genes like MARK4 [41] (not found by SuSiE) associated with tau pathology [37] in Alzheimer's Disease.

5 Discussion

BEATRICE is a novel, robust, and general purpose tool for fine-mapping that can be used across a variety of studies. One key contribution of BEATRICE over methods like FINEMAP and SuSiE is its ability to discern infinitesimal effects from non-causal variants, including non-causal variants in high LD with true causal variants. Our simulated experiments capture this improved performance by sweeping the proportion of the observed variance attributed to causal (fixed effects) and non-causal (random effects) genetic variants. This parameter $p \in [0, 1]$ is swept over its natural domain, such that $p = 1$ implies that the only link between the genotype and phenotype comes from the causal variants. At this extreme, Figure 4 shows that all methods achieve comparable performance. However, as p decreases, meaning that the effects of non-causal variants increase, BEATRICE outperforms both baselines.

We further probe this behavior by illustrating the element-wise PIPs and the

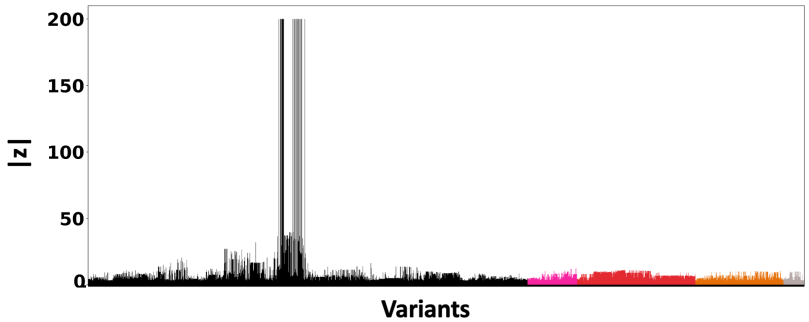


Fig 6. Z-scores of the variants present each of the 5 clumps. The scores are obtained from a GWAS study of Alzheimer's Disease [36]. Each locus is colored differently. The x-axis corresponds to the SNP index present in each locus. The y-axis reports the absolute z-scores.

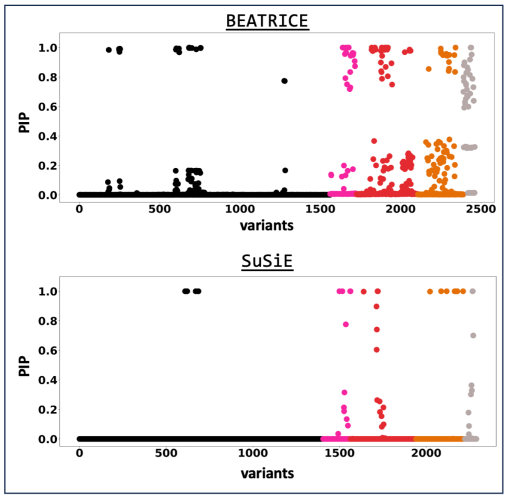


Fig 7. The posterior inclusion probabilities obtained by running the finemapping models on top-10 significant loci of Alzheimer's GWAS. Each locus here is colored differently. On the y-axiz, we report the PIP values; on the x-axis, we report the SNPs present in each locus.

credible sets identified by all three methods under two simulation settings: $\{d = 1, \omega^2 = 0.2, p = 0.9\}$ (Figure 9) and $\{d = 1, \omega^2 = 0.2, p = 0.1\}$ (Figure 10). As seen in Figure 9, the variance explained by the non-causal variants is small, so the causal variant is easy to distinguish and has been correctly identified by all three approaches. In contrast, we see in Figure 10 that when the non-causal variants play a larger role, the causal variant no longer has the maximum GWAS z-score. The high z-score observed in Figure 10 can be largely attributed to the LD structure present between SNPs. Following the generative assumptions of finemapping, we can show that the estimated GWAS effect sizes can be written as $\hat{\beta}_i = \sum_j r_{ij}\beta_j$, where $\hat{\beta}_i$ is the estimated effect size, r_{ij} is the correlation between the i-th and j-th SNPs, and β_j is the true effect size. This expression shows that in the presence of infinitesimal effect ($\beta_j \neq 0$) the LD structure can inflate the estimated size of a variant, leading to a high z-score. In this case, only BEATRICE correctly identifies the causal variant and assigns it the highest PIP. Both FINEMAP and SuSiE give uncertain predictions, as captured by the large credible sets and multiple high PIPs. We conjecture that BEATRICE takes advantage of the binary concrete distribution to model non-causal variants with non-zero effects, while using the

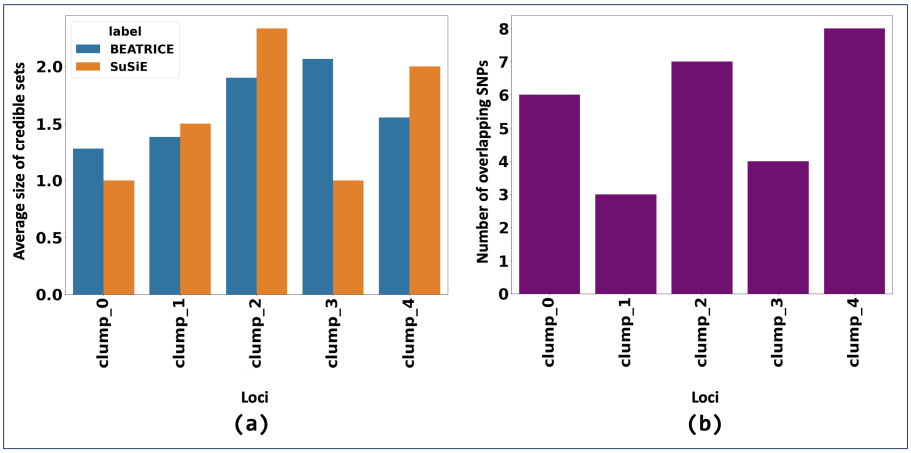


Fig 8. (a) The average size of the credible sets. (b) The number of overlapping SNPs between the credible sets identified by SuSiE and BEATRICE.

sparsity term of $\mathcal{L}(\cdot)$ to prioritize potentially causal variants.

A second contribution of BEATRICE is our strategic integration of neural networks within a larger statistical framework. Specifically, we use the neural network in Figure 1 as an inference engine to estimate the parameters $\mathbf{p}$ of our proposal distribution. In this case, the standard over-parameterization in the neural network helps BEATRICE to manage the complexity of the data while providing a buffer against overfitting. BEATRICE leverage the continuous representation of the causal vectors $\mathbf{c}^l$ to backpropagate the gradients through the random sampler and train the network. Additionally, the continuous representation of $\mathbf{c}^l$ results in low-variance gradients with respect to the underlying probability map, thus leading to a stable optimization.

Related to the above point, a third contribution of BEATRICE is its ability to efficiently build and evaluate a representative set of causal configurations during the optimization process. This set identifies key regions of the exponential search space to compute the PIPs and credible sets. In particular, we keep track of the sampled vectors at every iteration of the optimization, as described in Section 3.1. By minimizing the KL divergence between the proposal distribution and the true posterior distribution, we ensure that the randomly sampled causal vectors slowly converge to the causal configurations that have non-negligible posterior probability. Our strategy lies in stark contrast with traditional mean-field approaches, where independence assumptions between elements of the proposal distribution do not allow for joint inference of the causal configurations. Furthermore, this strategy allows us to efficiently estimate the PIPs in finite run-time. However, we note that the current implementation of BEATRICE does not estimate the posterior variant effect sizes. Instead, BEATRICE uses the binary concrete vectors to model the variance of the effect sizes. This property allows our model to adjust for infinitesimal effects from the non-causal variants. Figure 11 compares the average run-time of each method across all parameter settings. We observe that the run-time of BEATRICE and SuSiE are less than one minute. In contrast, FINEMAP requires significantly more time to converge.

The final contribution of BEATRICE is its simple and flexible design. Importantly, BEATRICE can easily incorporate priors based on the functional annotations of the variants. Formally, in the current setup, the prior over $\mathbf{c}$ is effectively constant, as captured by $p_0 = \frac{1}{m}$. We can integrate prior knowledge simply by modifying the distribution of p_0 across the variants. Thus, BEATRICE is a general-purpose tool for fine-mapping. Going one step further, a recent direction in fine-mapping is to aggregate data across multiple studies to identify causal variants [25]. Here, different LD matrices

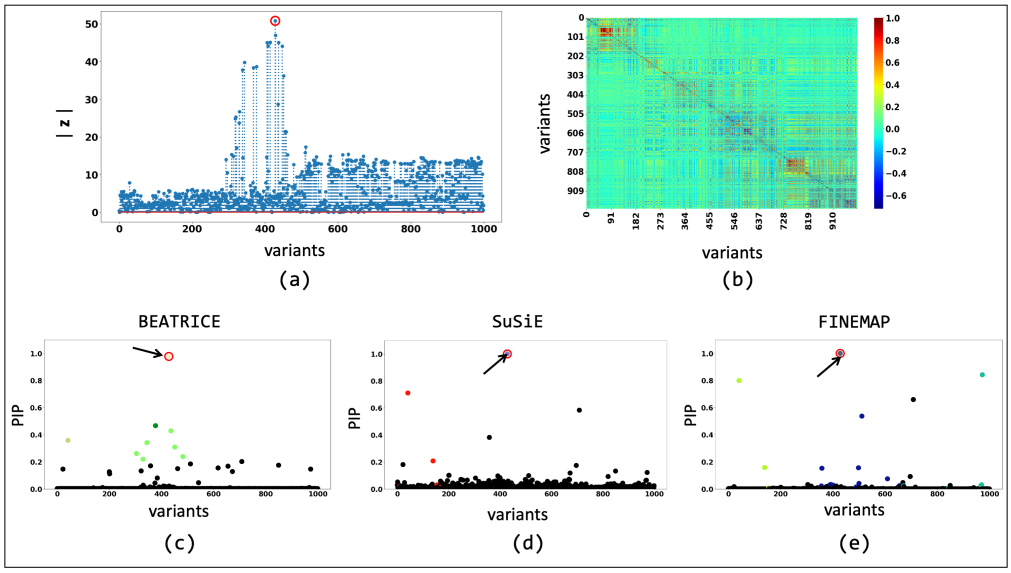


Fig 9. The fine-mapping performance of BEATRICE , SuSiE, and FINEMAP at a noise setting of $\{d = 1, \omega^2 = 0.2, p = 0.9\}$. (a) The absolute z-score of each variant as obtained from GWAS. (b) Pairwise correlation between the variants. (c), (d), and (e) are the posterior inclusion probabilities of each variant as identified by BEATRICE , SuSiE, and FINEMAP, respectively. The red circle marked by an arrow shows the location of the causal variant. We have further color-coded the variants based on their assignment to credible sets. The non-black markers represent the variants assigned to a credible set. Additionally, the variants in a credible set are marked by the same color.

across studies helps to refine the fine-mapping results. BEATRICE can be applied in this context as well simply by modifying Eq. (12) as follows:

$$\mathcal{L}(\phi) = -\frac{1}{S\,L}\sum_{s=1}^{S}\sum_{l=1}^{L}\log\left(N\left(\mathbf{z}_s; 0, \mathbf{\Sigma}_{X_s} + \mathbf{\Sigma}_{X_s}\left(n\sigma^2\mathbf{\Sigma}_C^l(\phi)\right)\mathbf{\Sigma}_{X_s}\right)\right) + \sum_i \mathbf{p}_i \log\left(\frac{\mathbf{p}_i}{p_0}\right) + (1-\mathbf{p}_i)\log\left(\frac{1-\mathbf{p}_i}{1-p_0}\right) \tag{17}$$

where s denotes each separate study, S is the total number of studies in the analysis, and $\mathbf{z}_s, \mathbf{\Sigma}_{X_s}$ are the summary statistics for each study.

In this work, we have shown that BEATRICE is highly efficient in handling the complexity that arises due to mutations with infinitesimal effects [20, 42]. Thus, we believe that the advantages of BEATRICE will be more evident when considering polygenic traits and diseases. Additionally, the high coverage and small size of credible sets reported in Figure 2–4 show that BEATRICE can successfully prioritize variants in the presence of LD. This property is in stark contrast with the baseline finemapping approaches that generate a large number of credible sets that do not contain a causal variant. Taken together, we believe BEATRICE could be useful in eQTL studies, where multiple variants within a locus can show strong association due to the complex LD structure present in the human genome [43]. Additionally, there may be multiple causal variants within a locus, which adds to the complexity of the finemapping problem [44].

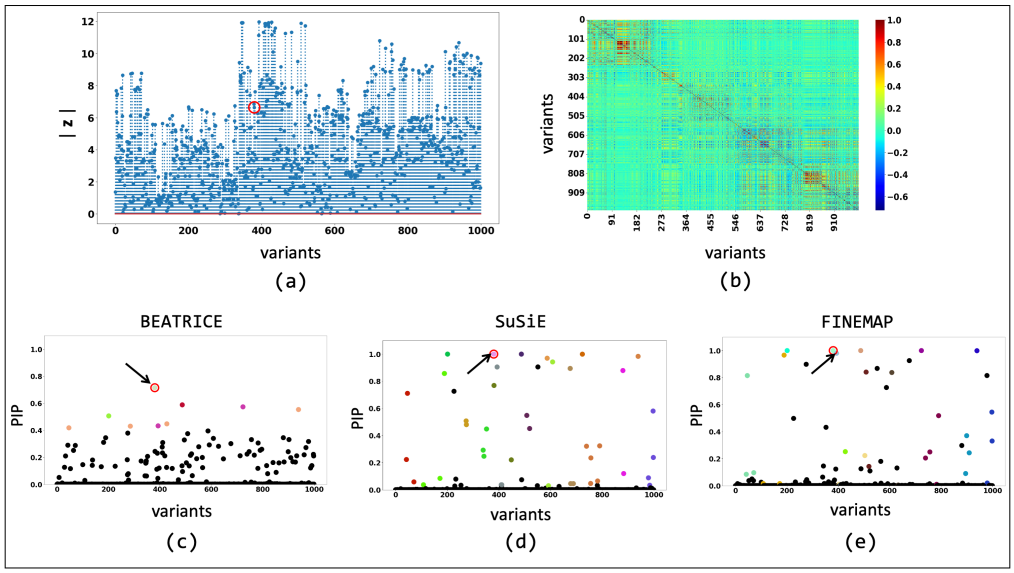


Fig 10. The fine-mapping performance of BEATRICE , SuSiE, and FINEMAP at a noise setting of $\{d = 1, \omega^2 = 0.2, p = 0.1\}$. (a) The absolute z-score of each variant as obtained from GWAS. (b) Pairwise correlation between the variants. (c), (d), and (e) are the posterior inclusion probabilities of each variant as identified by BEATRICE , SuSiE, and FINEMAP, respectively. The red circle marked by an arrow shows the location of the causal variant. We have further color-coded the variants based on their assignment to credible sets. The non-black markers represent the variants assigned to a credible set. Additionally, the variants in a credible set are marked by the same color.

5.1 Code Availability

We have compiled the code for BEATRICE and its dependencies into a docker image, which can be found at *https://github.com/sayangsep/Beatrice-Finemapping*. We have also provided installation instructions and a detailed description of the usage. The compact packaging will allow any user to directly download and run BEATRICE on their data. Namely, all the user must specify are a directory path to the summary statistics (i.e., z-scores), the LD matrix, and the number of subjects. Figure 12 shows the outputs generated by BEATRICE. The results are output in (1) a PDF document that displays the PIPs and corresponding credible sets, (2) a table with PIPs, (3) a text file with credible sets, and (4) a text file with the conditional inclusion probability of the variants within the credible sets. The user can also generate the neural network losses describe in Eq. (12) by adding a flag to the run command.

5.2 Summary

We present BEATRICE, a novel Bayesian framework for fine-mapping that identifies potentially causal variants within GWAS risk loci through the shared LD structure. Using a variational approach, we approximate the posterior probability of the causal location(s) via a binary concrete distribution. We leverage the unique properties of binary concrete random variables to build an optimization algorithm that can successfully model variants with differing levels of association. Moreover, we introduce a new strategy to build a reduced set of causal configurations within the exponential search space that can be neatly folded into our optimization routine. This reduced set is used to approximate the PIPs and identify credible sets. In a detailed simulation study,

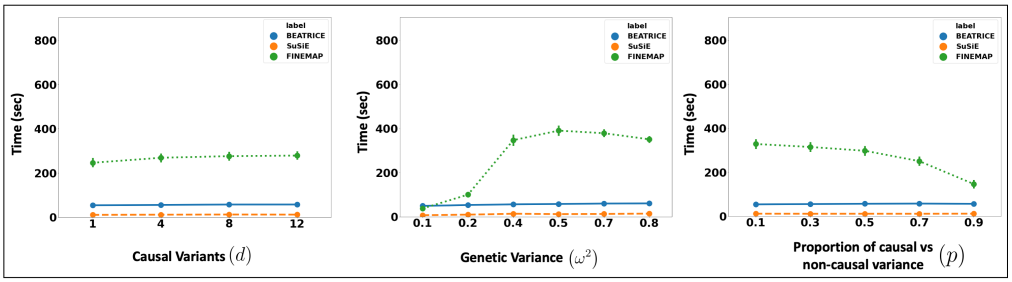
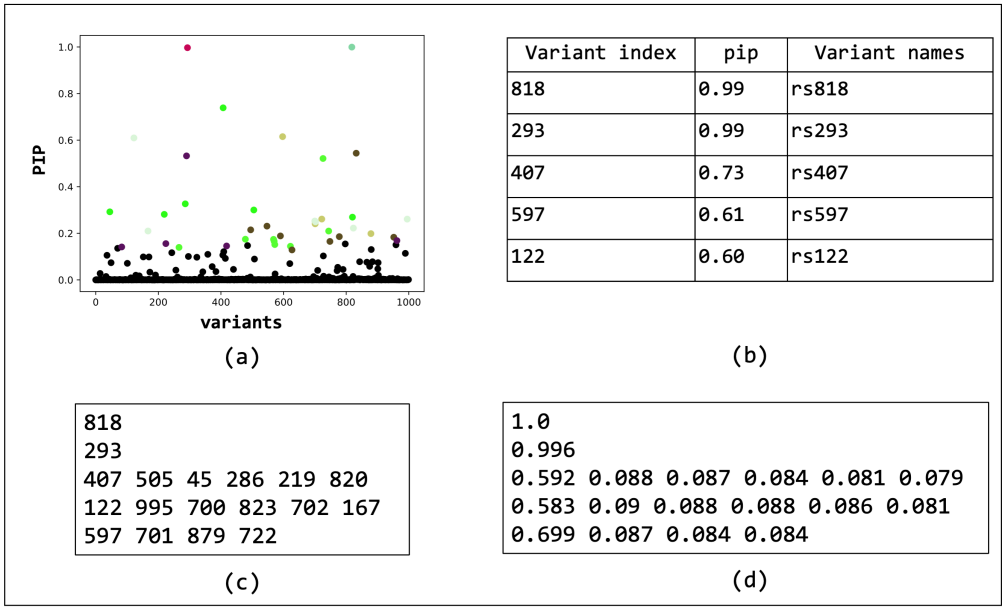


Fig 11. The runtime comparison of BEATRICE , SuSiE, and FINEMAP across all the simulation settings.



Variant index	pip	Variant names
818	0.99	rs818
293	0.99	rs293
407	0.73	rs407
597	0.61	rs597
122	0.60	rs122

Fig 12. Overview of the outputs generated by BEATRICE. (a) The PIPs are displayed and color coded by their assignment to credible sets. (b) A table with the PIPs and the corresponding name of the variants. (c) A text file with the credible sets. Here each row represent a credible set and the entries are indices of the variants present in the credible set. The first column of each row represents the key index. (d) The conditional inclusion probability of each of the credible variants given all the key variants. The calculations can be found in **Section S1.3** of the Supplements.

we compared BEATRICE with two state-of-the-art baselines and demonstrated the advantages of BEATRICE under different noise settings. Finally, our model does not have any prior on the causal variants and is agnostic to the original GWAS study. Hence, BEATRICE is a powerful tool to refine the results of a GWAS or eQTL analysis. It is also flexible enough to accommodate a variety of experimental settings.

Acknowledgments

This work was supported by the National Science Foundation CAREER Award 1845430 (PI: Venkataraman), the National Institutes of Health Awards R01-HD108790 (PI: Venkataraman), U24-HG010263 (PI: Schatz) and U41-HG006620 (PI: Schatz).

References

1. Uffelmann E, Huang QQ, Munung NS, de Vries J, Okada Y, Martin AR, et al. Genome-wide association studies. Nature Reviews Methods Primers 2021 1:1. 2021;1(1):1–21.

2. Visscher PM, Brown MA, McCarthy MI, Yang J. Five Years of GWAS Discovery. The American Journal of Human Genetics. 2012;90(1):7–24.

3. Brzyski D, Peterson CB, Sobczyk P, Candès EJ, Bogdan M, Sabatti C. Controlling the Rate of GWAS False Discoveries. Genetics. 2017;205:61–75. doi:10.1534/GENETICS.116.193987.

4. Maller JB, McVean G, Byrnes J, Vukcevic D, Palin K, Su Z, et al. Bayesian refinement of association signals for 14 loci in 3 common diseases. Nature Genetics 2012 44:12. 2012;44:1294–1301. doi:10.1038/ng.2435.

5. Spain SL, Barrett JC. Strategies for fine-mapping complex traits. Human Molecular Genetics. 2015;24(R1):R111–R119.

6. Schaid DJ, et al. From genome-wide associations to candidate causal variants by statistical fine-mapping. Nature Reviews Genetics 2018 19:8. 2018;19(8).

7. Hutchinson A, Watson H, Wallace C. Improving the coverage of credible sets in Bayesian genetic fine-mapping. PLOS Computational Biology. 2020;16:e1007829. doi:10.1371/JOURNAL.PCBI.1007829.

8. Cho S, et al. Elastic-net regularization approaches for genome-wide association studies of rheumatoid arthritis. BMC proceedings. 2009;3(Suppl 7).

9. Sabourin J, et al. Fine-Mapping Additive and Dominant SNP Effects Using Group-LASSO and Fractional Resample Model Averaging. Genetic Epidemiology. 2015;39(2).

10. Tibshirani R. Regression Shrinkage and Selection Via the Lasso. Journal of the Royal Statistical Society: Series B (Methodological). 1996;58:267–288. doi:10.1111/J.2517-6161.1996.TB02080.X.

11. Guan Y, et al. BAYESIAN VARIABLE SELECTION REGRESSION FOR GENOME-WIDE ASSOCIATION STUDIES AND OTHER LARGE-SCALE PROBLEMS. The Annals of Applied Statistics. 2011;5(3).

12. Carbonetto P, Stephens M. Scalable Variational Inference for Bayesian Variable Selection in Regression, and Its Accuracy in Genetic Association Studies. Bayesian Analysis. 2012;7(1):73–108.

13. Benner C, Spencer CCA, Havulinna AS, Salomaa V, Ripatti S, Pirinen M. FINEMAP: efficient variable selection using summary data from genome-wide association studies. Bioinformatics. 2016;32(10):1493–1501.

14. Chen W, Larrabee BR, Ovsyannikova IG, Kennedy RB, Haralambieva IH, Poland GA, et al. Fine mapping causal variants with an approximate bayesian method using marginal test statistics. Genetics. 2015;200(3):719–736.

15. Hormozdiari F, Kostem E, Kang EY, Pasaniuc B, Eskin E. Identifying causal variants at loci with multiple signals of association. Genetics. 2014;198(2):497–508.

16. Zou Y, Carbonetto P, Wang G, Stephens M. Fine-mapping from summary data with the "Sum of Single Effects" model. PLOS Genetics. 2022;18:e1010299. doi:10.1371/JOURNAL.PGEN.1010299.

17. Wang G, Sarkar A, Carbonetto P, Stephens M. A simple new approach to variable selection in regression, with application to genetic fine mapping. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2020;82(5):1273–1300.

18. Maddison CJ, et al. The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables. 5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings. 2016;.

19. Jang E, Gu S, Poole B. Categorical Reparameterization with Gumbel-Softmax. 5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings. 2016;.

20. Cheng W, Ramachandran S, Crawford L. Estimation of non-null SNP effect size distributions enables the detection of enriched genes underlying complex traits. PLOS Genetics. 2020;16:e1008855. doi:10.1371/JOURNAL.PGEN.1008855.

21. Servin B, Stephens M. Imputation-Based Analysis of Association Studies: Candidate Regions and Quantitative Traits. PLOS Genetics. 2007;3:e114. doi:10.1371/JOURNAL.PGEN.0030114.

22. Pirinen M, Donnelly P, Spencer CCA. EFFICIENT COMPUTATION WITH A LINEAR MIXED MODEL ON LARGE-SCALE DATA SETS WITH APPLICATIONS TO GENETIC STUDIES. The Annals of Applied Statistics. 2013;7(1):369–390.

23. Wakefield J. Bayes factors for Genome-wide association studies: Comparison with P-values. Genetic Epidemiology. 2009;33(1):79–86.

24. Kingma DP, Ba JL. Adam: A Method for Stochastic Optimization. 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings. 2014;.

25. LaPierre N, Taraszka K, Huang H, He R, Hormozdiari F, Eskin E. Identifying causal variants by fine mapping across multiple studies. PLOS Genetics. 2021;17(9):e1009733.

26. Kingma DP, Welling M. Auto-Encoding Variational Bayes; 2013. Available from: `https://arxiv.org/abs/1312.6114`.

27. Hans C, Dobra A, West M. Shotgun Stochastic Search for "Large p" Regression. https://doiorg/101198/016214507000000121. 2012;102:507–516. doi:10.1198/016214507000000121.

28. Wang Z, et al. An autoimmune pleiotropic SNP modulates IRF5 alternative promoter usage through ZBTB3-mediated chromatin looping. Nature Communications 2023 14:1. 2023;14(1).

29. Trubetskoy V, et al. Mapping genomic loci implicates genes and synaptic biology in schizophrenia. Nature 2022 604:7906. 2022;604(7906).

30. Albiñana C, Zhu Z, Borbye-Lorenzen N, Boelt SG, Cohen AS, Skogstrand K, et al. Genetic correlates of vitamin D-binding protein and 25-hydroxyvitamin D in neonatal dried blood spots. Nature Communications 2023 14:1. 2023;14(1):1–16. doi:10.1038/s41467-023-36392-5.

31. Li Y, Dang X, Chen R, Wang J, Li S, Mitchell BL, et al. Cross-ancestry genome-wide association study and systems-level integrative analyses implicate new risk genes and therapeutic targets for depression. medRxiv. 2023; p. 2023.02.24.23286411. doi:10.1101/2023.02.24.23286411.

32. Davis J, Goadrich M. The relationship between precision-recall and ROC curves. ACM International Conference Proceeding Series. 2006;148:233–240. doi:10.1145/1143844.1143874.

33. Dimitromanolakis A, Xu J, Krol A, Briollais L. sim1000G: A user-friendly genetic variant simulator in R for unrelated individuals and family-based designs. BMC Bioinformatics. 2019;20:1–9. doi:10.1186/S12859-019-2611-1/TABLES/4.

34. Belmont JW, et al. The International HapMap Project. Nature 2004 426:6968. 2003;426(6968).

35. Meyer HV, Birney E. PhenotypeSimulator: A comprehensive framework for simulating multi-trait, multi-locus genotype to phenotype relationships. Bioinformatics. 2018;34:2951–2956. doi:10.1093/BIOINFORMATICS/BTY197.

36. Wightman DP, et al. A genome-wide association study with 1,126,563 individuals identifies new risk loci for Alzheimer's disease. Nature Genetics. 2021;53(9):1276–1282. doi:10.1038/s41588-021-00921-z.

37. Lund H, Gustafsson E, Svensson A, Nilsson M, Berg M, Sunnemark D, et al. MARK4 and MARK3 associate with early tau phosphorylation in Alzheimer's disease granulovacuolar degeneration bodies. Acta Neuropathologica Communications. 2014;2(1):22. doi:10.1186/2051-5960-2-22.

38. Chen YC, et al. TOMM40 Genetic Variants Cause Neuroinflammation in Alzheimer's Disease. International Journal of Molecular Sciences. 2023;24(4):4085. doi:10.3390/IJMS24044085/S1.

39. Zhou Q, et al. Association between APOC1 Polymorphism and Alzheimer's Disease: A Case-Control Study and Meta-Analysis. PLOS ONE. 2014;9(1):e87017. doi:10.1371/JOURNAL.PONE.0087017.

40. Cooper YA, et al. Functional regulatory variants implicate distinct transcriptional networks in dementia. Science. 2022;377(6608).

41. Shamsi A, et al. MARK4 Inhibited by AChE Inhibitors, Donepezil and Rivastigmine Tartrate: Insights into Alzheimer's Disease Therapy. Biomolecules 2020, Vol 10, Page 789. 2020;10(5):789.

42. Zhang Y, et al. Estimation of complex effect-size distributions using summary-level statistics from genome-wide association studies across 32 complex traits. Nature Genetics 2018 50:9. 2018;50(9):1318–1326.

43. Zou J, et al. Leveraging allelic imbalance to refine fine-mapping for eQTL studies. PLOS Genetics. 2019;15(12).

44. Abell NS, et al. Multiple causal variants underlie genetic associations in humans. Science. 2022;375(6586):1247–1254.

45. Cui R, et al. Improving fine-mapping by modeling infinitesimal effects. bioRxiv. 2022;doi:10.1101/2022.10.21.513123.

Supporting Information

S1 Text. Detailed Methods and Additional Applications. Further details of the method, including the distribution of GWAS effect sizes, properties of binary concrete vectors, identification of credible sets, neural network architecture for the inference module, and baseline hyperparameter settings. Document also contains a comparison with SuSiE-inf [45] and a detailed comparison across simulation settings.

Supplementary Table 1. The list of SNPs present within each of the 5 clumps.

Supplementary Table 2. The SNPs present in the credible set of the first clump, as identified by BEATRICE . We provide the PIPs and the p-values as reported in [36].

Supplementary Table 3. The SNPs present in the credible set of the first clump, as identified by SuSiE. We provide the PIPs and the p-values as reported in [36].

Supplementary Table 4. The biological annotation and consequences of the variant present in the credible set of the first clump, as identified by BEATRICE.

Supplementary Table 5. The biological annotation and consequences of the variant present in the credible set of the first clump, as identified by SuSiE.