

BaGGLS: A Bayesian Shrinkage Framework for Interpretable Modeling of Interactions in High-Dimensional Biological Data

Marta S. Lemanczyk^{1*}, Lucas Kock^{2*}, Johanna Schlimme¹,
Nadja Klein³, and Bernhard Y. Renard¹

November 20, 2025

Abstract

Biological data sets are often high-dimensional, noisy, and governed by complex interactions among sparse signals. This poses major challenges for interpretability and reliable feature selection. Tasks such as identifying motif interactions in genomics exemplify these difficulties, as only a small subset of biologically relevant features (e.g., motifs) are typically active, and their effects are often non-linear and context-dependent. While statistical approaches often result in more interpretable models, deep learning models have proven effective in modeling complex interactions and prediction accuracy, yet their black-box nature limits interpretability.

We introduce BaGGLS, a flexible and interpretable probabilistic binary regression model designed for high-dimensional biological inference involving feature interactions. BaGGLS incorporates a Bayesian group global-local shrinkage prior, aligned with the group structure introduced by interaction terms. This prior encourages sparsity while retaining interpretability, helping to isolate meaningful signals and suppress noise. To enable scalable inference, we employ a partially factorized variational approximation that captures posterior skewness and supports efficient learning even in large feature spaces.

In extensive simulations, we compare BaGGLS to frequentist probit regressions (unconstrained and with L1-penalty) as well as a probit model with Markov Chain Monte Carlo (MCMC) sampling under a horseshoe prior. We can show that BaGGLS outperforms the other methods with regard to interaction detection and is many times faster than MCMC sampling under the horseshoe prior. We also demonstrate the usefulness of BaGGLS in the context of interaction discovery from motif scanner outputs (e.g., Find Individual Motif Occurrences (FIMO)) and noisy attribution scores from deep learning models. This shows that BaGGLS is a promising approach for uncovering biologically relevant interaction patterns, with potential applicability across a range of high-dimensional tasks in computational biology.

Keywords: Computational Genomics, Explainability, Global-Local Shrinkage Prior, Interaction Detection, Variational Inference

Acknowledgments: We gratefully acknowledge funding by the German Research Foundation (DFG) via the Research Unit KI-FOR 5363 (grant 459422098).

¹ Hasso Plattner Institute, Digital Engineering Faculty, University of Potsdam, Potsdam, Germany

² Department of Statistics and Data Science, National University of Singapore

³ Scientific Computing Center, Karlsruhe Institute of Technology, Karlsruhe, Germany

* These authors contributed equally; marta.lemanczyk@hpi.de, lucas.kock@nus.edu.sg

1 Introduction

One of the greatest methodological challenges in the biomedical data domain is feature selection in high-dimensional data (Borah et al.; 2024; Yang et al.; 2021). This challenge has historically been central in genomics, but with deep learning (DL) becoming the dominant analytical paradigm and post hoc attribution being routinely used for interpretability, the problem manifests in a different and more intricate form today. DL models excel at learning complex patterns in genomic sequences (Ismail et al.; 2025). However, the resulting models are highly nonlinear and operate in settings where biological features are both strongly interactive (Xie et al.; 2025; Forsberg et al.; 2017) and extremely sparse (Wheeler et al.; 2016). These properties amplify the already difficult task of interpreting which features matter and, crucially, how they interact.

As a result, we continue to struggle to reliably explain feature interactions (Borah et al.; 2024). This difficulty is not merely algorithmic, but rooted in statistical challenges. Among these challenges are the curse of dimensionality, the tendency of interactions to explode combinatorically, sparsity of true effects, and high noise levels (Giraud; 2021). These challenges are in particularly pronounced when working with post hoc attribution maps. Attribution scores are often noisy and spurious (Majdandzic et al.; 2023), and aggregating them to motif-level features produces high-dimensional, sparsely informative predictors whose reliability and error are difficult to determine. Altogether, this creates a setting in which classical feature selection approaches fall short, and even modern statistical approaches do not resolve the core problem reliably.

Consequently, we see a strong need for new methodology that addresses the specific structure of post hoc attribution in genomic deep learning. Here, we propose a structured binary regression model that incorporates a large number of potential main effects and interaction terms into its linear predictor. For interpretability and reliable identification of relevant effects, we impose sparsity on the coefficients. Within the Bayesian framework, sparsity is naturally enforced through shrinkage priors (George and McCulloch; 1997; Ishwaran and Rao; 2005; Liang et al.; 2008; Yanchenko and Bondell; 2025). Many continuous shrinkage priors such as the Bayesian Lasso (Park and Casella; 2008) and the horseshoe (Carvalho et al.; 2010) follow a global-local structure (Polson and Scott; 2011), in which global parameters jointly shrink coefficients toward zero while local parameters allow coefficient-specific deviations. Xu et al. (2017) extend this framework to include group based shrinkage parameters, facilitating structured regularization when coefficients can be meaningfully organized into groups. The inclusion of interaction terms results in many overlapping groups, where each group is formed of all terms involving a specific feature. To accommodate this structure, we extend the group-based shrinkage framework of Xu et al. (2017), adapting it to handle the extremely large number of potential interaction terms present in genomic data. A detailed description of the resulting prior structure is provided in Section 2.2.

Exact Bayesian inference in the resulting high-dimensional Bayesian probit model is challenging and we propose a computationally efficient variational inference (VI; Blei et al.; 2017) algorithm, a commonly applied technique in high-dimensional binary regression models (Zhang et al.; 2019; Ray et al.; 2020). Here, we extend the approach of Fasano et al. (2022) to our novel prior and consider a unified skew-normal (Arrelano-Valle and Azzalini; 2006) approximation for the regression coefficients. This allows more flexibility than the commonly employed mean field approximation (e.g., Durante and Rigon; 2019). The unified skew-normal distribution generalizes the Gaussian distribution to include skewness. Recently, skewness perturbed variational approximations were considered by several authors (e.g., Tan and Chen; 2025; Kock et al.; 2025; Pozza et al.; 2025) theoretically justified by the skewed Bernstein-von Mises theorem (Durante et al.; 2024). The unified skew-normal distribution is also a conjugate prior to the probit model (Durante; 2019; Anceschi et al.; 2023), so that the variational approximation can be efficiently learned using an analytic coordinate ascent updating scheme (e.g., Bishop; 2006; Ray and Szabó; 2022). This enables scalable inference even when the number of interactions is large.

In this paper, we introduce our novel method called *Bayesian Group Global Local Shrinkage* (BaGGLS) in detail and demonstrate through extensive simulations that it outperforms state-of-the-art methods in both computational efficiency and interpretability. Importantly, we also show empirically that BaGGLS fills a methodological gap in post hoc analysis of genomic deep learning models. Many post hoc approaches attempt to understand learned regulatory mechanisms by detecting motifs through attribution scores (Novakovsky et al.; 2023; Van Hilten et al.; 2024; Bartoszewicz et al.; 2021). Attribution methods assign position-wise scores to sequences, and downstream tools such as TF-MoDISco (Shrikumar et al.; 2018) extract motifs from these maps. However, these workflows do not capture interactions between motif patterns, even though such interactions are central to many genomic mechanisms (Xie et al.; 2025).

Motif scanners such as FIMO (Grant et al.; 2011) overcome some limitations by identifying matches to known PWMs from databases like JASPAR (Rauluseviciute et al.; 2024). Yet this produces a high-dimensional feature space containing hundreds of motifs, many of which produce spurious matches. When combined with noisy attribution scores, only a small subset of motifs, and often their interaction, contribute meaningfully to phenotypes. These conditions create exactly the type of high-dimensional, sparse, interaction-rich scenario where classical methods struggle. We show that BaGGLS successfully extracts these interacting signals and provides reliable interpretability.

The remainder of this paper is organized as follows. Section 2 introduces BaGGLS and an efficient VI approach to posterior estimation. Section 3 shows the merits of our approach empirically benchmarking it against state-of-the-art methods while Section 4 illustrates its applicability to high-dimensional biological data along an application in interaction detection for genomic attribution scores. Finally, Section 5 concludes. Code is available at gitlab.com/dacs-hpi/baggl.

2 Bayesian group global-local shrinkage

We formally introduce the model in Section 2.1. Our novel prior is introduced in Section 2.2, and Section 2.3 describes our approach for scalable inference.

2.1 Model formulation

We consider the probit model

$$y_i \mid \beta \sim \text{Bern}(\Phi(x_i^\top \beta)) \quad i = 1, \dots, n \quad (1)$$

where $\text{Bern}(\pi)$ denotes the Bernoulli distribution with success probability π , Φ is the cumulative distribution function of the standard normal distribution, $x_i = (x_{i1}, \dots, x_{ip})^\top$ is a vector of pre-defined effects including an intercept and interactions, and $\beta \in \mathbb{R}^p$ is a vector of model parameters to be learned. By introduction of latent variables $z = (z_1, \dots, z_n)^\top$ we can augment (1) as

$$y_i = \mathcal{I}(z_i > 0), \quad z_i \mid \beta \sim \mathcal{N}(x_i^\top \beta, 1),$$

where $\mathcal{N}(\mu, \sigma^2)$ denotes a normal distribution with mean μ and variance σ^2 , and $\mathcal{I}(z_i > 0)$ denotes the indicator function that takes a value of 1 if $z_i > 0$ and 0 otherwise. This representation yields closed-form full-conditionals for β and z , which will become useful for the computational efficient inference algorithm introduced later. The linear predictor x_i is derived from a d -dimensional vector $m_i = (m_{i1}, \dots, m_{id})^\top$ of observed features, $i = 1, \dots, n$, and contains not only linear effects, but also interaction terms of the d features as illustrated in Figure 1B. For example, when considering an intercept, linear effects as well as all possible pairwise multiplicative interactions of the form $m_{il}m_{il'}$, $l \neq l'$, the linear predictor $x_i^\top \beta$ in (1) can be written as $\beta_0 + \sum_{l=1}^d \beta_l m_{il} + \sum_{l=1}^d \sum_{l'>l} \beta_{ll'} m_{il} m_{il'}$. Hence, p is typically much larger than d . This is the structure we consider in Section 4 when we apply BaGGLS to interaction discovery from motif scanner outputs. In this case m_i will be a vector of attribution scores derived from a deep learning architecture. However, our general framework allows us to specify arbitrary effect and interaction terms. In this context, it is important to note that first, even for small and moderate d , the number of potential interactions is large and thus x_i can be high-dimensional. Second, we also explicitly allow for much larger p than the sample size n , that is, $p \gg n$.

2.2 Overlapping group horseshoe prior

If the number of potential effect terms p is large, (1) can be challenging to interpret. Based on common scenarios in biological applications, we make the following assumptions. (i) We assume that β is sparse. That is, most entries are zero and thus only a small number of effect terms, x_{ij} , influence y_i . (ii) We further assume, that only a small subset of features m_{il} significantly influences y and thus for most features all terms involving that feature are jointly zero.

The first assumption can be incorporated by considering an appropriate shrinkage prior. The prior structure also acts as an important regularization to the high-dimensional regression model. Here, we build on the popular horseshoe prior (Carvalho et al.; 2010). The horseshoe prior introduces a global shrinkage parameter τ controlling the overall level of shrinkage jointly for all predictors and local shrinkage parameters λ_j $j = 1, \dots, p$ controlling the shrinkage for the coefficient β_j corresponding to effect x_{ij} . Xu et al. (2017) extend this prior structure to account for group-based shrinkage. The heredity assumption

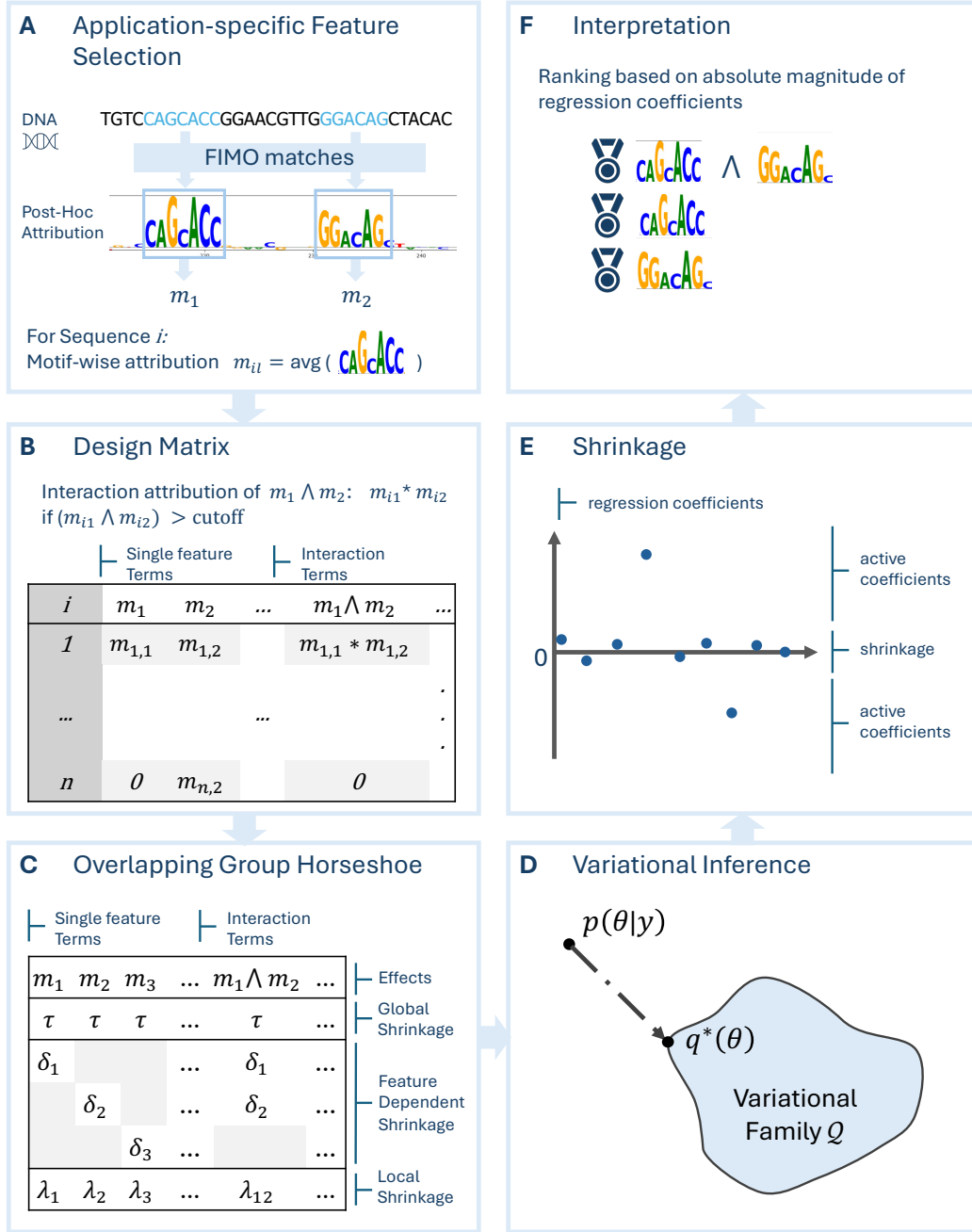


Figure 1: Schematic description of the BaGGLS-workflow. **A)** Feature selection based on the observed data. In our application discussed in Section 4, we use attribution scores from CNNs trained on FIMO matches. **B)** Based on expert knowledge a pre-defined set of potential effects including interpretable interactions is defined. **C)** Our novel overlapping group horseshoe prior matches the structure of the pre-defined effect terms and allows for global, local, and feature dependent shrinkage of the regression coefficients. **D)** Scalable Bayesian inference is carried out by projecting the true posterior onto a tractable family of variational posteriors \mathcal{Q} . **E)** The prior shrinks most regression coefficients towards 0 resulting in a sparse and interpretable regression model. **F)** For interpretation, we propose to rank the effects based on absolute magnitude of their corresponding coefficients in the logistic regression model.

(ii) implies a structure with many overlapping groups. This would necessitate a prohibitively large number of hierarchies in the group horseshoe prior by Xu et al. (2017). We thus propose the following alternation to their grouped global-local shrinkage prior that circumvents the need for many hierarchies

for each predictor.

Let $J \in \mathbb{R}^{p \times d}$ be an indicator matrix where entry J_{jl} indicates if feature m_l contributes to term x_j . We propose the hierarchical prior

$$\begin{aligned}\beta_j \mid \tau, \lambda, \delta &\sim ND \left(0, \tau \lambda_j \prod_{l=1}^d J_{jl} \delta_l \right), & j = 1, \dots, p \\ \tau \mid \nu &\sim \mathcal{IG} \left(\frac{1}{2}, \frac{1}{\nu} \right), \quad \nu \sim \mathcal{IG} \left(\frac{1}{2}, 1 \right), \\ \lambda_j \mid c_j &\sim \mathcal{IG} \left(\frac{1}{2}, \frac{1}{c_j} \right), \quad c_j \sim \mathcal{IG} \left(\frac{1}{2}, 1 \right), & j = 1, \dots, p \\ \delta_l \mid t_l &\sim \mathcal{IG} \left(\frac{1}{2}, \frac{1}{t_l} \right), \quad t_l \sim \mathcal{IG} \left(\frac{1}{2}, 1 \right), & l = 1, \dots, d,\end{aligned}$$

where $\delta = (\delta_1, \dots, \delta_d)^\top$, $\lambda = (\lambda_1, \dots, \lambda_p)^\top$. Here, $\mathcal{IG}(\alpha, \beta)$ denotes an inverse Gamma distribution with shape parameter α and scale parameter β . As for the standard horseshoe prior τ controls global shrinkage, and λ_j controls local shrinkage. In addition, δ_l controls joint group shrinkage for all effects including feature m_l . The structure of the indicator matrix J informing the grouping structure depends on the effects considered in x and needs to be specified upfront. To allow for consistent interpretation of the effect strength as well as consistent shrinkage through the shared shrinkage parameters we assume that all effects x_{ij} , $j = 1, \dots, p$, are standardized to have unit scale across observations, $i = 1, \dots, n$. A schematic description of our prior is given in Figure 1C. Each term has an individual local shrinkage parameter λ_j . The motif dependent group shrinkage parameter δ_l is active exactly for the terms involving the corresponding feature m_l . The parameter τ controls global shrinkage and is shared across all terms. The full vector of the $n + 3p + 2d + 2$ unknown model parameters is $\theta = (z^\top, \beta^\top, \tau, \nu, \lambda^\top, c^\top, \delta^\top, t^\top)^\top$.

2.3 Inference for large data sets

Exact Bayesian inference in the high-dimensional probit model can be computationally challenging. VI emerged as a powerful alternative. The main idea illustrated in Figure 1D is to learn an approximation to the posterior density $p(\theta \mid y)$ using an approximating family of densities \mathcal{Q} . Most commonly, the optimal approximation $q^*(\theta)$ is chosen so that it minimizes the reverse Kullback-Leibler divergence,

$$\mathcal{D}_{\text{KL}} [q(\theta) \parallel p(\theta \mid y)] = \mathbb{E}_q [\log q(\theta) - \log p(\theta \mid y)],$$

where $\mathbb{E}_q[\cdot]$ denotes expectation with respect to $q(\theta)$, among all approximating families in \mathcal{Q} . A popular choice in logistic and probit regression models is the mean field assumption (e.g., Durante and Rigon; 2019), which assumes independence between specific blocks of θ . This choice is computationally efficient, but as recently shown by Fasano et al. (2022) can be too restrictive for large p . Therefore, the authors propose to relax the independence assumption between β and z . This allows for more flexibility in the posterior approximation, as it yields a unified skew-normal (Arrelano-Valle and Azzalini; 2006) approximation for the vector of regression coefficients. Fasano et al. (2022) consider a simple Gaussian prior for β . We extend their approach to the Gaussian scale mixture representation of the BaGGLS prior and consider variational approximations of the form

$$\mathcal{Q} = \left\{ q(\theta) : q(\theta) = q(\beta \mid z) \left(\prod_{i=1}^n q(z_i) \right) q(\tau) q(\nu) \left(\prod_{j=1}^p q(\lambda_j) q(c_j) \right) \left(\prod_{l=1}^d q(\delta_l) q(t_l) \right) \right\}.$$

We describe how $q^*(\theta) = \arg \min_{q(\theta) \in \mathcal{Q}} \mathcal{D}_{\text{KL}} [q(\theta) \parallel p(\theta \mid y)]$ can be derived in a computational attractive manner using a simple coordinate ascent algorithm. The final algorithm updates one variational factor at a time while holding the others fixed, cycling through coordinates until convergence. Under the variational family \mathcal{Q} given above all updates are in closed-form (see Appendix A). It is possible to efficiently sample from $q(\beta) = \int q(\beta \mid z) q(z) dz$ even when $p \gg n$ using the strategies outlined in Bhattacharya et al. (2016) and the posterior mean $\hat{\beta} = \mathbb{E}_q[\beta]$ is given in closed form. Due to the variational approximation, samples from $q^*(\theta)$ will not yield exact uncertainty quantification, and we use the point estimator $\hat{\beta}$ for out of sample prediction and interpretation of the regression model.

3 Simulations

3.1 Illustrative example

Data generating process We generate $m = 500$ data sets with $n = 500$ independent observations. For $m_i = (m_{i1}, \dots, m_{id})$ with $m_{ij} \sim \mathcal{G}(1, 1)$, similar as in our application presented in Section 4 the corresponding vector of predictors x_i consists of an intercept, linear effects m_j , $j = 1, \dots, d$, and all possible pairwise interactions $m_j m_{j'}$ for $j < j'$. We set $d = 10$, so that $p = 56$ in our first set-up. After standardization of the design matrix, we generate observations $y_i \sim \text{Bern}(\Phi(x_i^\top \beta^*))$, where all entries of β^* are zero except for the entries corresponding to m_1 , m_2 , and the interaction $m_1 m_2$. The true regression coefficient β^* is thus extremely sparse and reflects the sparsity assumption described in Section 2.2.

Benchmark methods We consider different regularization techniques for the probit model (1) that result in sparse and interpretable models. We use the same design matrix for all benchmarks, so that the models are directly comparable. In particular, we compare BaGGLS with the following benchmarks:

UC: Unconstrained frequentist probit regression fitted via maximum likelihood,

L1: Frequentist probit regression with L_1 -penalty and default hyperparameters as implemented in `statsmodels` (Seabold and Perktold; 2010),

HS: MCMC sampling for the probit model equipped with the horseshoe prior (Carvalho et al.; 2010) as implemented in `brms` (Bürkner; 2017).

Results Figure 2A shows boxplots of the root mean squared error (RMSE), $\text{RMSE}(\hat{\beta}) = \left(\sum_{j=1}^p (\beta_j^* - \hat{\beta}_j)^2 \right)^{1/2}$, over the 500 independent repetitions. BaGGLS has the smallest average RMSE (0.7022), followed by HS (0.8052), and L1 (1.3827). Figure 2B plots estimates $\hat{\beta}_j$ for all coefficients with $\beta_j^* = 0$ showing effective shrinkage toward zero. As expected, the largest variability occurs for joint effects involving m_1 or m_2 . Non-zero effects ($\beta_j^* \neq 0$) are accurately recovered (Figure 2C). Figure 2D shows a histogram for a representative non-informative coefficient under BaGGLS and under HS from their respective marginal posteriors. Both are centered near zero and on a similar scale. Notably, the marginal posterior $q(\beta_j)$ under BaGGLS is skewed, which is captured due to the unified skew-normal variational family. However, this family cannot capture the characteristic spike around zero. For a representative non-zero coefficient (Figure 2E), BaGGLS concentrates more tightly around the truth, while HS posteriors are more dispersed. Nevertheless, both methods yield similar posterior means. Due to the variational approximation, posterior samples from BaGGLS will not yield exact uncertainty quantification. On this data, BaGGLS takes on average only 0.59 seconds and is therefore more than 100 times faster than MCMC sampling under the horseshoe prior rendering it suitable for large scale applications.

3.2 Scalability

Data generating process To further investigate the performance of BaGGLS we now vary the number of observations n and the number of observed features d in the data generative process described in Section 3.1. Due to the non-linear relationship between d and the number of effects p a slight increase in d leads to a large increase in p and thus to a much more challenging inference task. Here, we consider $n = 500$ and $n = 2,000$ as well as $d = 10, 15, 20$ resulting in a total of 6 scenarios. These values are chosen to reflect common scenarios in real world data and result in a small sample-to-feature ratio n/p . This matches the general structure of the genomic data considered in Section 4.

Performance metrics In addition to the overall RMSE, we consider several other performance metrics. First, we evaluate discrimination via the area under the receiver operating characteristic curve (AUC) and probabilistic accuracy via the Brier score, $n^{-1} \sum_{i=1}^n (\Phi(x_i^\top \hat{\beta}) - y_i)^2$, both evaluated on an additional hold-out test data set with $n = 10,000$ observations. The Brier score is a proper scoring rule (Gneiting and Raftery; 2007). These metrics are useful in comparing the out-of-sample predictive power of the different methods. Although prediction is not the primary objective of BaGGLS, reasonable out-of-sample performance remains desirable. However, when prediction is the main goal rather than interpretability, methods with weaker structural assumptions may be more suitable. To quantify effect

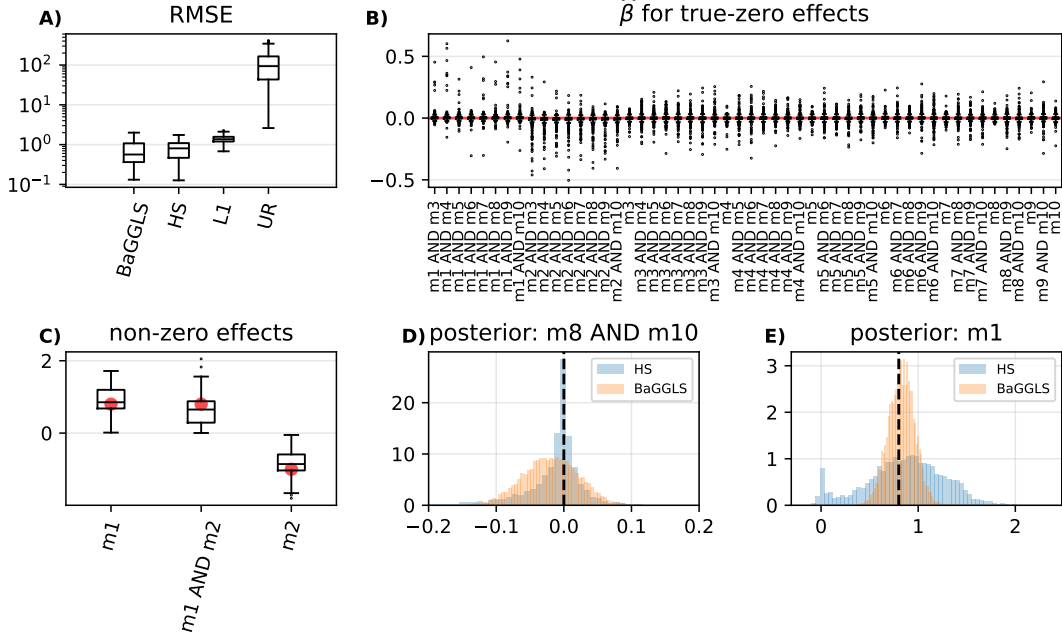


Figure 2: Simulations $n = 500, d = 10$. **A)** RMSE across 500 independent runs on log-scale. Lower values are preferred. Our proposed method outperforms the benchmarks in recovery of the true coefficient vectors **B)** Boxplots for $\hat{\beta}_j$ for all true-zero coefficients across 500 independent runs. For most repetitions, BaGGLS shrinks effects that do not contribute to the response successfully towards zero resulting in an interpretable model. **C)** Boxplots for $\hat{\beta}_j$ for all non-zero coefficients. True values are marked by dots. BaGGLS correctly estimates the effect size for the three active effects in the data generating process. **D)** Samples from the estimated marginal posterior for one true-zero coefficient for HS and BaGGLS. Both posteriors are correctly centered around zero and skewed indicating that the skewed variational approximation used for BaGGLS is helpful in tightly approximating the true posterior. **E)** Samples from the estimated marginal posterior for one non-zero effect for HS and BaGGLS. Again both posteriors are centered around the true value. The variational approximation for BaGGLS is sharper than the posterior under HS.

recovery in high dimensions, we report RMSE computed separately over the active (nonzero) and inactive (zero) entries of β^* . In addition, we report the proportion of runs in which the three truly active effects are ranked among the top 20 and, separately, among the top 3 estimated effects by absolute magnitude. As a measure of overall sparsity in the predictor $\hat{\beta}$, we report the ratio $(\sum_j \hat{\beta}_j^2)^2 / \sum_j \hat{\beta}_j^4$, which serves as a proxy for the effective number of active coefficients. Under the true vector β^* this ratio takes the value 2.8575. Lastly, we report average computation times on a standard laptop for all benchmarks.

Results Values for RMSEs, AUC, the Brier score, and the run times are reported in Appendix B. BaGGLS is compatible with the benchmark methods in terms of out of sample prediction measured by AUC and the Brier score, while more effective at detecting the active terms. Across all simulation scenarios, BaGGLS achieves the lowest RMSE on the active (nonzero) entries of β^* followed by HS. HS imposes stronger shrinkage on zero coefficients as measured by the RMSE on the inactive components, particularly in scenarios with $n = 2,000$, where the sample-to-feature ratio n/p is larger than in scenarios with $n = 500$. In the application considered in Section 4 the sample-to-feature ratio $n/p = 1.33$ is small.

Table 1 reports the proportion of repetitions in which the three truly active terms are ranked among the top 20 and, separately, among the top 3 estimated effects by absolute magnitude. While L1 and UR do not recover the active effects reliably, HS and BaGGLS place all three truly active effects in the top 20 for every run across all scenarios. However, BaGGLS outperforms HS when considering only the top 3 terms. In particular, the interaction effect $m_1 m_2$ is often missed by HS. for $n = 500, d = 20$, the interaction appears in the top 3 in 88 of 100 repetitions for HS versus 93 of 100 for BaGGLS. Similarly, for $n = 500, d = 15$, BaGGLS detects the interaction 93 times, while HS places it outside the top 3 in 16% of runs. These results indicate that BaGGLS is robust at detecting interaction effects, especially when the sample-to-feature ratio is small. In addition, BaGGLS is between 7 and 122 times faster than

	% in top 20 terms			% in top 3 terms			sparsity
	m_1	m_2	m_1m_2	m_1	m_2	m_1m_2	
$n=500, d=10, p=56, n/p=8.93$							
BaGGLS	100%	100%	100%	94%	96%	94%	2.5200 (0.6045)
HS	100%	100%	100%	91%	99%	91%	2.2582 (0.5448)
L1	81%	91%	81%	53%	75%	53%	6.6936 (2.7878)
UR	98%	98%	98%	44%	40%	44%	1.2632 (0.4286)
$n=2000, d=10, p=56, n/p=35.71$							
BaGGLS	100%	100%	100%	100%	100%	100%	2.8241 (0.1598)
HS	100%	100%	100%	100%	100%	100%	2.8046 (0.1561)
L1	94%	100%	94%	85%	99%	85	3.5463 (0.7359)
UR	98%	99%	98%	46%	50%	46%	1.2996 (0.5817)
$n=500, d=15, p=121, n/p=4.13$							
BaGGLS	100%	100%	100%	93%	96%	93%	2.3641 (0.6513)
HS	100%	100%	100%	84%	98%	84%	2.0703 (0.4973)
L1	49%	66%	49%	21%	44%	21%	13.9722 (5.7225)
UR	95%	84%	95%	68%	31%	68%	1.6237 (1.0593)
$n=2000, d=15, p=121, n/p=16.53$							
BaGGLS	100%	100%	100%	100%	100%	100%	2.8329 (0.1255)
HS	100%	100%	100%	100%	100%	100%	2.8049 (0.1258)
L1	88%	99%	88%	61%	92%	61%	5.6560 (1.9977)
UR	92%	94%	92%	48%	49%	48%	1.1880 (0.2497)
$n=500, d=20, p=211, n/p=2.37$							
BaGGLS	100%	100%	100%	93%	94%	93%	2.3405 (0.6667)
HS	100%	100%	100%	88%	99%	88%	1.9953 (0.4971)
L1	13%	19%	13%	8%	13%	8%	31.1999 (10.9530)
UR	—	—	—	—	—	—	—
$n=2000, d=20, p=211, n/p=9.48$							
BaGGLS	100%	100%	100%	100%	100%	100%	2.8352 (0.1220)
HS	100%	100%	100%	100%	100%	100%	2.7953 (0.1115)
L1	71%	88%	71%	41%	72%	41%	9.1345 (3.9088)
UR	75%	81%	75%	47%	49%	47%	1.3885 (1.0950)

Table 1: Simulations. Proportions of runs for which the truly active terms m_1 , m_2 , and m_1m_2 are among the top 20 or top 3 terms respectively for each benchmark method and all simulation scenarios. The last column reports the average value and the standard deviation (in brackets) for the ratio $(\sum_j \hat{\beta}_j^2)^2 / \sum_j \hat{\beta}_j^4$. UR did not reliably converge for the scenario $n = 500, d = 20$. BaGGLS outperforms the benchmarks in detecting relevant effects including the interaction.

HS, depending on the scenario.

4 Application to genomic attribution scores

In this section, we illustrate how BaGGLS is useful as a post hoc processing method for genomic deep learning explanations. To this end, we apply BaGGLS to attribution scores derived from deep learning models trained on synthetic genomic sequences containing real motifs and a known ground truth.

Data and motif scanning The dataset consists of synthetic DNA sequences with binary labels indicating the presence of a motif set of interest. We follow a similar simulation protocol as described in Tseng et al. (2024) to evaluate the approach on known ground truth. Depending on the defined motif grammar, the motifs are inserted in randomly generated sequence. Here, we explored the REST motif consisting of two submotifs with a specific order and spacing (Figure 3A). 35,000 sequences are generated for training and 10,000 for validation, with a sequence length of 500 base pairs. To evaluate BaGGLS, we generated additional 45 evaluation datasets with the same grammar each consisting of 2,000 sequences.

FIMO (Grant et al.; 2011) locates motifs by computing significant matches of motifs with position-weight matrices (PWM) from databases in a sequence. Here, we use all latest versions of the human

transcription binding site motifs ($d = 755$) in JASPAR2024 (Rauluseviciute et al.; 2024). Depending on the threshold for the p-value, the results can contain many false positive matches (high threshold) or miss some of the real motif matches (low threshold). We use the default threshold $pthresh = 1e^{-4}$ to investigate the robustness of BaGGLS by exposing it to noisy motif matches by allowing false positives as well as relevant motifs which are not matched by FIMO despite being present.

Attribution scores from deep learning models We train five shallow convolutional neural networks (CNN) on the training data set. We specifically use a simple architecture to not overfit to the synthetic data resulting in an average AUC of 0.91. Details on the architecture and performance can be found in the supplement section C. The trained CNNs are interpreted by the post-hoc attribution method Integrated Gradients (Sundararajan et al.; 2017) to obtain position-wise attribution scores which indicate the contribution of one position to the over-all output of an input sequence. We calculate the scores for all evaluation data sets for each of the five CNNs.

Attribution-based design matrix BaGGLS requires a pre-defined set of features and interactions passed in a design matrix (Figure 1B). Here, we pass all matched motifs by FIMO as possible features to BaGGLS. FIMO returns the start and end positions of the matched motifs within each sequence. With that information, we can aggregate the attribution scores in that subregion to obtain motif-wise contributions. Here, we use the absolute average from the attribution scores at the matched motif positions scaled sequence-wise so that the aggregated motif scores add up to 1. To create interaction features, we calculate the pairwise co-activation based on the product of the average contribution scores. We remove rarely occurring features to avoid inefficiency due to a very large number of features by using a cut-off quantile of 0.95 which can be adjusted by the user. The resulting features still comprise of a large number of motif and interaction features ($p=1,500$), resulting in a ratio of $n/p = 1.33$ similar to the scenarios considered in Section 3. The design matrix consists of 755 single effects and 745 interaction effects. Similarly as in the simulation, we fit BaGGLS for each of the 45 evaluation data sets for each of the 5 CNNs. In total, this results in 225 different scenarios.

Attribution evaluation Due to the sparsity assumption only a small sub-set of regression coefficients $\hat{\beta}$ derived by BaGGLS is meaningfully different from 0. This is the set of effectively active feature and interaction effects detected by BaGGLS. For further interpretation, we calculate the top-20 effects by absolute magnitude of their respective coefficients (see Figure 1F). This can be viewed as the set of the most important 20 effects out of the $p = 1,500$ potential effects initially passed to our method.

For illustration purposes, we consider here data with a known ground truth. This allows us to evaluate whether BaGGLS captures the known main interaction, by checking how often the interaction is included within the top-20 effects. BaGGLS identifies the interaction term in 82.7% of the scenarios whereas the REST submotifs were included within the top-20 in 68.4% and 76.6% of the scenarios, respectively (Figure 3A). It is important to note that when using BaGGLS as post-processing method, its performance is highly dependent on the fit of the base CNN. In particular BaGGLS cannot detect effects that were overlooked by the base model. When analyzing at the rankings of the terms for each of the five CNN models individually (Figure 3B), the interaction term ranks on average higher than the individual submotif effects for each model (Median ranks: interaction term = 4, REST1 = 15, REST2 = 14). This shows that BaGGLS robustly identifies the driving interaction among noisy attribution scores and prioritizes the interaction over individual effects. For this reason, BaGGLS can be used as highly interpretable post-hoc processing method for trained deep learning genomic classifiers. Identifying important interactions is not directly possible from the black-box CNNs, but crucial for understanding the underlying biological process.

5 Discussion

We propose BaGGLS, an interaction detection method for high-dimensional and noisy biological data. To this end, we make the following main contributions. (i) We propose to use a structured probit regression model as a post-hoc analysis tool for deep classifiers. Our model takes a large number of potential effect, including interactions, as input and identifies a small subset of important and truly active effects. (ii) Sparsity is imposed through a novel shrinkage prior that respects the overlapping group structure resulting from the inclusion of many interaction terms in the additive predictor. (iii) We propose a fast algorithm for posterior inference based on a unified skewed-normal approximation. This is crucial

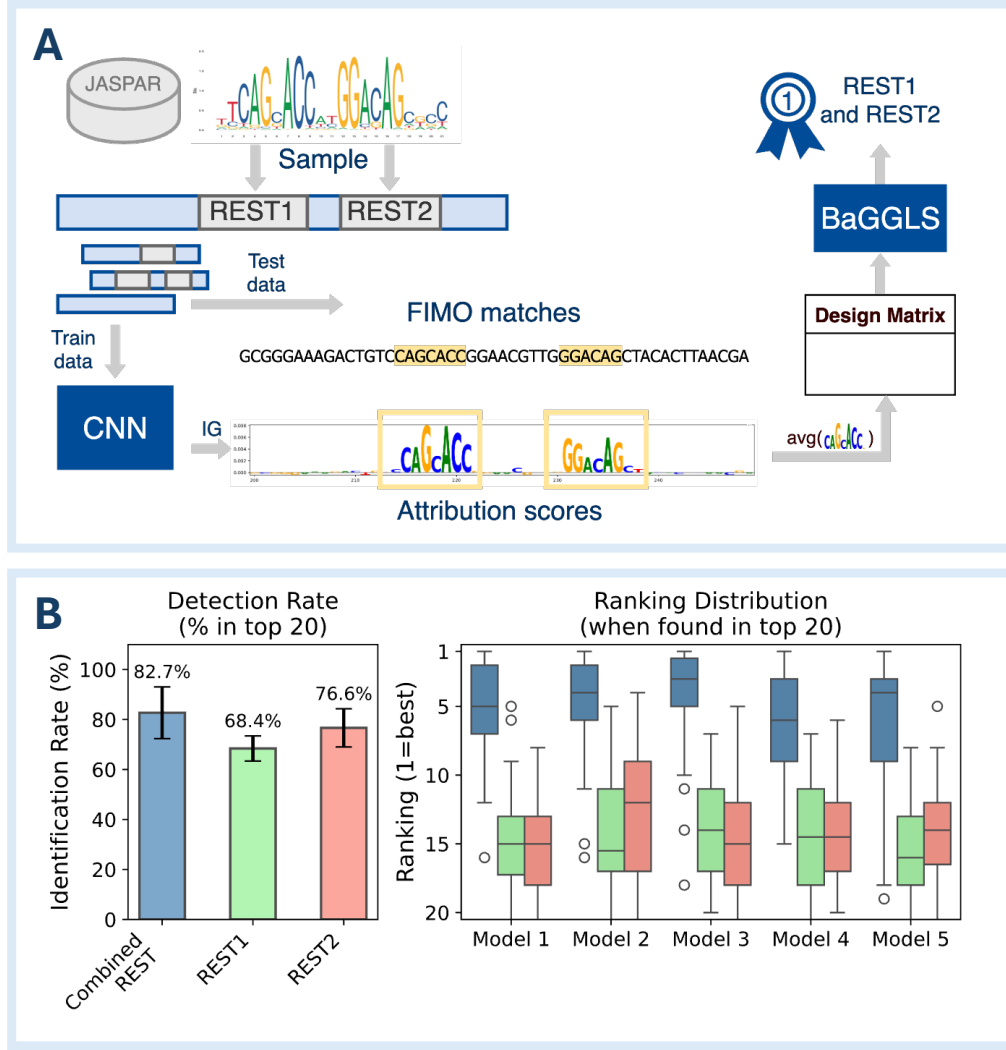


Figure 3: Post-processing of BaGGLS on attribution scores. (A) First, we generated synthetic sequences by inserting the sampled composite REST motif in 500bp long random sequences. We trained five shallow CNN models on that data and interpreted the test data with Integrated Gradients. By applying FIMO on the sequences, we obtain matches which resemble motifs from the JASPAR2024 data base. We average the attribution scores in the matched motif region and compute the design matrix and indicator matrix from those scores for BaGGLS. (B) We applied BaGGLS on the 45 test data sets for each model and measured how often the interaction term was included in the Top 20 terms. BaGGLS ranks the interaction term on average in 82.7% of the test data sets as a top 20 effect while the individual effects in 68.4% and 76.6% for the first and second half of the REST motif respectively (left plot). When looking closer to the exact ranking of the identified terms (right plot), the interaction term receives much higher ranking than the individual effects.

for scalability to scenarios with many potential effects and a small sample-to-feature ratio. (iv) We conduct an extensive simulation study, and show that BaGGLS outperforms state-of-the-art benchmarks in detecting relevant interaction effects. (v) We illustrate the merits of our approach in an application on motif interactions in genomic data which is known to be complex and noisy offering a good use case for comparisons.

Even though, our general framework allows for arbitrary and more complex effect types to be specified, we have only considered continuous features and interactions based on co-occurrences within our application. The inclusion of more complex effects, for example, by including binary features and boolean interactions in the context of logic regression (Ruczinski et al.; 2003) coupled with regulatory logic in genomics (Buchler et al.; 2003) is a promising direction.

Combining our novel overlapping group shrinkage prior with regression models beyond the probit model, is a further avenue for future research. For example, current deep learning prediction tasks in the genomic domain shift from classification to continuous and categorical responses by considering biological signals like gene expression or read counts directly. Quantitative readouts provide a more fine-grained picture of a biological signal. Extending BaGGLS to continuous responses might be one pathway to account for low-affinity motifs which are sometimes ignored due to thresholds by peak calling methods (Zeitlinger; 2020).

While we demonstrate application to motif detection, other biological domains can also benefit from BaGGLS. In genetics, single-nucleotide polymorphisms are studied in the context of disease. While rare disorders are frequently driven by individual variants with large effects, common-variant contributions to complex traits are highly polygenic and often involve context-dependent effects that are hard to resolve in high-dimensional genotype space (Wray et al.; 2018). Currently, regulatory genomics is shifting from bulk assays to single-cell and related high-resolution assays. These provide finer cell-type specificity but produce sparser, noisier data matrices that change the statistical challenges and opportunities (Bouland et al.; 2023). This shift opens new applications for BaGGLS.

However, there are also limitations visible in the presented use case. In some genomic regions, composite regulatory elements, which are arrangements of multiple nearby binding sites, can alter local sequence preferences and functional readout (Jolma et al.; 2015). Bulk models that assume independent binding sites often miss these dependencies, which leads to only a sparse but important set of interactions being captured. Adjusting the design matrix to include interaction terms based on experimentally validated composite elements can therefore make interaction interpretations more precise. This can further be improved by an iterative approach in which interpretable features from deep models, for example, convolutional filters (Tseng et al.; 2024) or in-silico perturbation readouts (Gjoni and Pollard; 2024), are used to detect candidate composite arrangements that are then incorporated into the design matrix to refine sparsity structure and interaction estimates. The interaction patterns detected by our method could also serve as hypotheses for previously unrecognized composite elements or motif interactions. These hypotheses can be ranked by BaGGLS and subsequently tested in targeted biochemical or reporter assays, providing a systematic way to identify novel regulatory interactions.

References

- Aneschi, N., Fasano, A., Durante, D. and Zanella, G. (2023). Bayesian conjugacy in probit, tobit, multinomial probit and extensions: A review and new results, *Journal of the American Statistical Association* **118**(542): 1451–1469.
- Arrelano-Valle, R. B. and Azzalini, A. (2006). On the unification of families of skew-normal distributions, *Scandinavian Journal of Statistics* **33**(3): 561–574.
- Bartoszewicz, J. M., Seidel, A. and Renard, B. Y. (2021). Interpretable detection of novel human viruses from genome sequencing data, *NAR Genomics and Bioinformatics* **3**(1): lqab004.
- Bhattacharya, A., Chakraborty, A. and Mallick, B. K. (2016). Fast sampling with Gaussian scale mixture priors in high-dimensional regression, *Biometrika* **103**(4): 985–991.
- Bishop, C. M. (2006). *Pattern recognition and machine learning*, Springer.
- Blei, D. M., Kucukelbir, A. and McAuliffe, J. D. (2017). Variational inference: A review for statisticians, *Journal of the American Statistical Association* **112**(518): 859–877.
- Borah, K., Das, H. S., Seth, S., Mallick, K., Rahaman, Z. and Mallik, S. (2024). A review on advancements in feature selection and feature extraction for high-dimensional ngs data analysis, *Functional & Integrative Genomics* **24**(5): 139.
- Bouland, G. A., Mahfouz, A. and Reinders, M. J. (2023). Consequences and opportunities arising due to sparser single-cell rna-seq datasets, *Genome biology* **24**(1): 86.
- Buchler, N. E., Gerland, U. and Hwa, T. (2003). On schemes of combinatorial transcription logic, *Proceedings of the National Academy of Sciences* **100**(9): 5136–5141.
- Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan, *Journal of statistical software* **80**: 1–28.
- Carvalho, C. M., Polson, N. G. and Scott, J. G. (2010). The horseshoe estimator for sparse signals, *Biometrika* **97**(2): 465–480.
- Durante, D. (2019). Conjugate Bayes for probit regression via unified skew-normal distributions, *Biometrika* **106**(4): 765–779.
- Durante, D., Pozza, F. and Szabo, B. (2024). Skewed Bernstein–von Mises theorem and skew-modal approximations, *The Annals of Statistics* **52**(6): 2714–2737.

- Durante, D. and Rigon, T. (2019). Conditionally Conjugate Mean-Field Variational Bayes for Logistic Models, *Statistical Science* **34**(3): 472 – 485.
- Fasano, A., Durante, D. and Zanella, G. (2022). Scalable and accurate variational Bayes for high-dimensional binary regression models, *Biometrika* **109**(4): 901–919.
- Forsberg, S. K., Bloom, J. S., Sadhu, M. J., Kruglyak, L. and Carlborg, Ö. (2017). Accounting for genetic interactions improves modeling of individual quantitative trait phenotypes in yeast, *Nature genetics* **49**(4): 497–503.
- George, E. I. and McCulloch, R. E. (1997). Approaches for Bayesian variable selection, *Statistica sinica* pp. 339–373.
- Giraud, C. (2021). *Introduction to High-Dimensional Statistics*, Chapman and Hall/CRC.
- Gjoni, K. and Pollard, K. S. (2024). Supremo: a computational tool for streamlining in silico perturbation using sequence-based predictive models, *Bioinformatics* **40**(6): btac340.
- Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation, *Journal of the American statistical Association* **102**(477): 359–378.
- Grant, C. E., Bailey, T. L. and Noble, W. S. (2011). Fimo: scanning for occurrences of a given motif, *Bioinformatics* **27**(7): 1017–1018.
- Ishwaran, H. and Rao, J. S. (2005). Spike and slab variable selection: Frequentist and Bayesian strategies, *The Annals of Statistics* **33**(2): 730 – 773.
- Ismail, F. N., Sengupta, A. and Amarasoma, S. (2025). Deep learning for regulatory genomics: A survey of models, challenges, and applications, *Bioinformatics Advances* p. vbaf271.
- Jolma, A., Yin, Y., Nitta, K. R., Dave, K., Popov, A., Taipale, M., Enge, M., Kivioja, T., Morgunova, E. and Taipale, J. (2015). Dna-dependent formation of transcription factor pairs alters their binding specificity, *Nature* **527**(7578): 384–388.
- Kock, L., Tan, L. S., Bansal, P. and Nott, D. J. (2025). Variational inference for hierarchical models with conditional scale and skewness corrections, *arXiv preprint arXiv:2503.18075*.
- Liang, F., Paulo, R., Molina, G., Clyde, M. A. and Berger, J. O. (2008). Mixtures of g priors for Bayesian variable selection, *Journal of the American Statistical Association* **103**(481): 410–423.
- Majdandzic, A., Rajesh, C. and Koo, P. K. (2023). Correcting gradient-based interpretations of deep neural networks for genomics, *Genome Biology* **24**(1): 109.
- Novakovsky, G., Dexter, N., Libbrecht, M. W., Wasserman, W. W. and Mostafavi, S. (2023). Obtaining genetics insights from deep learning via explainable artificial intelligence, *Nature Reviews Genetics* **24**(2): 125–137.
- Park, T. and Casella, G. (2008). The Bayesian lasso, *Journal of the american statistical association* **103**(482): 681–686.
- Polson, N. G. and Scott, J. G. (2011). Shrink globally, act locally: Sparse Bayesian regularization and prediction, *Bayesian Statistics 9*, Oxford University Press.
- Pozza, F., Durante, D. and Szabo, B. (2025). Skew-symmetric approximations of posterior distributions, *Journal of the Royal Statistical Society, Series B: Statistical Methodology*. in press.
- Rauluseviciute, I., Riudavets-Puig, R., Blanc-Mathieu, R., Castro-Mondragon, J. A., Ferenc, K., Kumar, V., Lemma, R. B., Lucas, J., Chèneby, J., Baranasic, D. et al. (2024). Jaspar 2024: 20th anniversary of the open-access database of transcription factor binding profiles, *Nucleic acids research* **52**(D1): D174–D182.
- Ray, K. and Szabó, B. (2022). Variational Bayes for high-dimensional linear regression with sparse priors, *Journal of the American Statistical Association* **117**(539): 1270–1281.
- Ray, K., Szabo, B. and Clara, G. (2020). Spike and slab variational Bayes for high dimensional logistic regression, in H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan and H. Lin (eds), *Advances in Neural Information Processing Systems*, Vol. 33, Curran Associates, Inc., pp. 14423–14434.
- Ruczinski, I., Kooperberg, C. and LeBlanc, M. (2003). Logic regression, *Journal of Computational and graphical Statistics* **12**(3): 475–511.
- Seabold, S. and Perktold, J. (2010). statsmodels: Econometric and statistical modeling with python, *9th Python in Science Conference*.
- Shrikumar, A., Tian, K., Avsec, Ž., Shcherbina, A., Banerjee, A., Sharmin, M., Nair, S. and Kundaje, A. (2018). Technical note on transcription factor motif discovery from importance scores (tf-modisco) version 0.5. 6.5, *arXiv preprint arXiv:1811.00416*.
- Sundararajan, M., Taly, A. and Yan, Q. (2017). Axiomatic attribution for deep networks, *International conference on machine learning*, PMLR, pp. 3319–3328.
- Tan, L. S. and Chen, A. (2025). Variational inference based on a subclass of closed skew normals, *Journal of Computational and Graphical Statistics* **34**(2): 422–436.

- Tseng, A. M., Eraslan, G., Diamant, N. L., Biancalani, T. and Scalia, G. (2024). A mechanistically interpretable neural-network architecture for discovery of regulatory genomics, *ICLR 2024 Workshop on Machine Learning for Genomics Explorations*.
- Van Hilten, A., Katz, S., Saccenti, E., Niessen, W. J. and Roshchupkin, G. V. (2024). Designing interpretable deep learning applications for functional genomics: a quantitative analysis, *Briefings in Bioinformatics* **25**(5).
- Wheeler, H. E., Shah, K. P., Brenner, J., Garcia, T., Aquino-Michaels, K., Consortium, G., Cox, N. J., Nicolae, D. L. and Im, H. K. (2016). Survey of the heritability and sparse architecture of gene expression traits across human tissues, *PLoS genetics* **12**(11): e1006423.
- Wray, N. R., Wijmenga, C., Sullivan, P. F., Yang, J. and Visscher, P. M. (2018). Common disease is more complex than implied by the core gene omnigenic model, *Cell* **173**(7): 1573–1580.
- Xie, Z., Sokolov, I., Osmala, M., Yue, X., Bower, G., Pett, J. P., Chen, Y., Wang, K., Cavga, A. D., Popov, A. et al. (2025). Dna-guided transcription factor interactions extend human gene regulatory code, *Nature* pp. 1–10.
- Xu, Z., Schmidt, D. F., Makalic, E., Qian, G. and Hopper, J. L. (2017). Bayesian sparse global-local shrinkage regression for selection of grouped variables, *arXiv preprint arXiv:1709.04333*.
- Yanchenko, E. and Bondell, Howard D. and Reich, B. J. (2025). The R2D2 prior for generalized linear mixed models, *The American Statistician* **79**(1): 40–49.
- Yang, P., Huang, H. and Liu, C. (2021). Feature selection revisited in the single-cell era, *Genome Biology* **22**(1): 321.
- Zeitlinger, J. (2020). Seven myths of how transcription factors read the cis-regulatory code, *Current opinion in systems biology* **23**: 22–31.
- Zhang, C.-X., Xu, S. and Zhang, J.-S. (2019). A novel variational Bayesian method for variable selection in logistic regression models, *Computational Statistics & Data Analysis* **133**: 1–19.

A Variational Inference

The full vector of the $n + 3p + 2d + 2$ unknown model parameters is $\theta = (z^\top, \beta^\top, \tau, \nu, \lambda^\top, c^\top, \delta^\top, t^\top)$. We consider the following variational approximation

$$q(\theta) = q(\beta | z) \left(\prod_{i=1}^n q(z_i) \right) q(\tau) q(\nu) \left(\prod_{j=1}^p q(\lambda_j) q(c_j) \right) \left(\prod_{l=1}^d q(\delta_l) q(t_l) \right). \quad (2)$$

Note that (2) is not the fully factorized mean-field approximation as we do neither assume independence between β and z nor between the individual entries of β . Instead we consider the partially factorized approximation introduced by Fasano et al. (2022). The structure (2) implies that $q^*(\beta | z)$ is a p -dimensional Gaussian distribution, $q^*(z_i)$ is a truncated normal distribution and all remaining factors of $q^*(\theta)$ are inverse gamma distributions. We can thus write

$$\begin{aligned} q^*(\beta | z) &= \phi_p(\beta; B(\beta)z, \Sigma(\beta)); \\ q^*(z_i) &= 1_{(2y_i-1)z_i > 0} \frac{\phi(z_i; \mu(z_i), \sigma^2(z_i))}{\Phi((2y_i-1)\mu(z_i)(\sigma^2(z_i))^{-\frac{1}{2}})} & i = 1, \dots, n; \\ q^*(\tau) &= p_{\text{IG}}(\tau; a(\tau), b(\tau)); \\ q^*(\nu) &= p_{\text{IG}}(\nu; a(\nu), b(\nu)); \\ q^*(\lambda_j) &= p_{\text{IG}}(\lambda_j; a(\lambda_j), b(\lambda_j)) & j = 1, \dots, p; \\ q^*(c_j) &= p_{\text{IG}}(c_j; a(c_j), b(c_j)) & j = 1, \dots, p; \\ q^*(\delta_l) &= p_{\text{IG}}(\delta_l; a(\delta_l), b(\delta_l)) & l = 1, \dots, d; \\ q^*(t_l) &= p_{\text{IG}}(t_l; a(t_l), b(t_l)) & l = 1, \dots, d. \end{aligned}$$

The optimal variational parameters are given as

$$\begin{aligned}
\Sigma(\beta) &= \left(X^\top X + \text{diag} \left(\frac{a(\tau)a(\lambda_1)}{b(\tau)b(\lambda_1)} \prod_{J_{1l}=1} \frac{a(\delta_l)}{b(\delta_l)}, \dots, \frac{a(\tau)a(\lambda_p)}{b(\tau)b(\lambda_p)} \prod_{J_{pl}=1} \frac{a(\delta_l)}{b(\delta_l)} \right) \right)^{-1} \\
B(\beta) &= \Sigma(\beta) X^\top \\
\mu(z_i) &= \sigma^2(z_i) x_i^\top \Sigma(\beta) X_{-i}^\top (\mathbb{E}_q[z_1], \dots, \mathbb{E}_q[z_{i-1}], \mathbb{E}_q[z_{i+1}], \dots, \mathbb{E}_q[z_n])^\top \\
\sigma^2(z_i) &= (1 - x_i^\top \Sigma(\beta) x_i)^{-1} \\
a(\tau) &= \frac{p+1}{2} \\
b(\tau) &= \frac{1}{2} \sum_{j=1}^p \left(\mathbb{E}_q[\beta_j^2] \frac{a(\lambda_j)}{b(\lambda_j)} \prod_{J_{jl}=1} \frac{a(\delta_l)}{b(\delta_l)} \right) + \frac{a(\nu)}{b(\nu)} \\
a(\nu) &= 1 \\
b(\nu) &= \frac{a(\tau)}{b(\tau)} + 1 \\
a(\lambda_j) &= 1 \\
b(\lambda_j) &= \frac{1}{2} \mathbb{E}_q[\beta_j^2] \frac{a(\tau)}{b(\tau)} \prod_{J_{jl}=1} \frac{a(\delta_l)}{b(\delta_l)} + \frac{a(c_j)}{b(c_j)} \\
a(c_j) &= 1 \\
b(c_j) &= \frac{a(\lambda_j)}{b(\lambda_j)} + 1 \\
a(\delta_l) &= \frac{\sum_{j=1}^p J_{jl} + 1}{2} \\
b(\delta_l) &= \frac{a(\tau)}{b(\tau)} \sum_{j=1}^p J_{jl} \frac{a(\lambda_l)}{b(\lambda_l)} \mathbb{E}_q[\beta_j^2] + \frac{a(t_l)}{b(t_l)} \\
a(t_l) &= 1 \\
b(t_l) &= 1 + \frac{a(\delta_l)}{b(\delta_l)},
\end{aligned}$$

where

$$\begin{aligned}
\mathbf{E}_q[\beta_j^2] &= (\Sigma(\beta) + \Sigma(\beta) X^\top \text{diag} (\sigma^2(z_1) - (\mathbf{E}[z_1] - \mu(z_1))\mathbf{E}[z_1], \dots, \sigma^2(z_n) - (\mathbf{E}[z_n] - \mu(z_n))\mathbf{E}[z_n]) X \Sigma(\beta))_{jj} \\
&\quad + (B(\beta) (\mathbf{E}[z_1], \dots, \mathbf{E}[z_n])^\top)_j^2; \\
\mathbf{E}_q[z_i] &= \mu(z_i) + (2y_i - 1) \sqrt{\sigma^2(z_i)} \frac{\phi \left(\mu(z_i) \sigma^2(z_i)^{-\frac{1}{2}} \right)}{\Phi \left((2y_i - 1) \mu(z_i) \sigma^2(z_i)^{-\frac{1}{2}} \right)}.
\end{aligned}$$

We refer to Fasano et al. (2022) for technical details in the derivation of $q^*(\beta, z)$. Using this set of equations $q^*(\theta)$ can be updated via coordinate ascent (Bishop; 2006). Note that each step involves inverting the $p \times p$ dimensional matrix $\Sigma(\beta)$. If $p > n$, the Woodbury matrix identity is used to invert a lower dimensional $n \times n$ matrix instead.

B Additional simulation results

Table 2 summarizes the overall RMSE, the RMSE computed separately on the active (nonzero) and inactive (zero) entries of β^* , the area under the receiver operating characteristic curve (AUC), and the Brier score, $n^{-1} \sum_{i=1}^n (\Phi(x_i^\top \hat{\beta}) - y_i)^2$, evaluated on an additional hold-out test data set with $n = 10,000$ observations. We report average values as well as standard deviations (in brackets) over 100 repetitions. The table also reports the average run-time on a standard laptop for each of the six simulation scenarios considered.

	RMSE (\downarrow)	RMSE non zero (\downarrow)	RMSE true-zero (\downarrow)	AUC (\uparrow)	Brier score (\downarrow)	time (in s)
<i>n=500, d=10, p=56, n/p=8.93</i>						
BaGGLS	0.7022 (0.4180)	0.5627 (0.3889)	0.3881 (0.2220)	0.8648 (0.0040)	0.1483 (0.0026)	0.5866
HS	0.8052 (0.3671)	0.6876 (0.3388)	0.3964 (0.1960)	0.8660 (0.0031)	0.1478 (0.0020)	71.5908
L1	1.3827 (0.2706)	0.6892 (0.2880)	1.1727 (0.2278)	0.8457 (0.0063)	0.1616 (0.0042)	0.3983
UR	114.2356 (85.8997)	14.4545 (10.4597)	113.2603 (85.3364)	0.7569 (0.0845)	0.2601 (0.0839)	0.0238
<i>n=2000, d=10, p=56, n/p=35.71</i>						
BaGGLS	0.2356 (0.1305)	0.1775 (0.1159)	0.1381 (0.0924)	0.8674 (0.0013)	0.1472 (0.0007)	10.2991
HS	0.2186 (0.1247)	0.1737 (0.1196)	0.1191 (0.0683)	0.8676 (0.0010)	0.1471 (0.0006)	250.9252
L1	0.7699 (0.1804)	0.4266 (0.2162)	0.6206 (0.1074)	0.8611 (0.0019)	0.1510 (0.0012)	0.4775
UR	59.8707 (44.3895)	7.3706 (5.4846)	59.4063 (44.0615)	0.8398 (0.0245)	0.1736 (0.0245)	4.3269
<i>n=500, d=15, p=121, n/p=4.13</i>						
BaGGLS	0.8325 (0.4839)	0.6967 (0.4367)	0.4219 (0.2704)	0.8533 (0.0058)	0.1558 (0.0036)	7.8429
HS	0.9311 (0.3800)	0.8220 (0.3459)	0.4177 (0.2040)	0.8558 (0.0035)	0.1545 (0.0023)	90.8542
L1	2.3274 (0.3461)	0.8842 (0.3459)	2.1265 (0.3364)	0.8056 (0.0100)	0.1956 (0.0073)	6.7732
UR	86.4223 (67.9976)	11.6303 (7.5754)	85.0680 (68.2881)	0.7597 (0.0383)	0.2590 (0.0384)	5.9180
<i>n=2000, d=15, p=121, n/p=16.53</i>						
BaGGLS	0.2267 (0.1152)	0.1673 (0.1031)	0.1410 (0.0787)	0.8614 (0.0012)	0.1504 (0.0007)	26.7289
HS	0.1994 (0.1115)	0.1653 (0.1063)	0.1029 (0.0546)	0.8621 (0.0007)	0.1500 (0.0004)	298.2052
L1	1.2030 (0.2149)	0.5456 (0.2493)	1.0519 (0.1639)	0.8459 (0.0029)	0.1603 (0.0018)	5.7420
UR	53.0565 (33.0476)	4.9558 (2.7377)	52.8008 (32.9721)	0.8287 (0.0162)	0.1788 (0.0174)	6.8981
<i>n=500, d=20, p=211, n/p=2.37</i>						
BaGGLS	0.8575 (0.4544)	0.7135 (0.4310)	0.4288 (0.2511)	0.8476 (0.0076)	0.1593 (0.0044)	15.6204
HS	0.9166 (0.3401)	0.8283 (0.3253)	0.3707 (0.1627)	0.8516 (0.0035)	0.1571 (0.0021)	107.5822
L1	3.9889 (0.5015)	1.3117 (0.3719)	3.7540 (0.4599)	0.7648 (0.0114)	0.2447 (0.0091)	71.3119
UR	—	—	—	—	—	—
<i>n=2000, d=20, p=211, n/p=9.48</i>						
BaGGLS	0.2388 (0.1003)	0.1709 (0.0942)	0.1528 (0.0754)	0.8565 (0.0016)	0.1536 (0.0009)	48.2972
HS	0.2050 (0.0984)	0.1710 (0.0945)	0.1045 (0.0513)	0.8576 (0.0007)	0.1529 (0.0005)	371.5898
L1	1.7003 (0.2597)	0.7289 (0.2766)	1.5154 (0.2332)	0.8294 (0.0038)	0.1724 (0.0027)	34.3788
UR	37.2856 (24.8385)	3.4586 (2.0772)	37.0293 (24.8942)	0.8184 (0.0096)	0.1826 (0.0080)	12.4084

Table 2: Simulations. Average values and standard deviations (in brackets) across all independent repetitions for the six simulation scenarios for the overall RMSE, the RMSE computed on the active (nonzero) and inactive (zero) entries of β^* , the area under the receiver operating characteristic curve (AUC), and the Brier score, as well as the average computation times. Bold values indicate the best performing method in terms of the best mean value as well as methods potentially tied according to a one-sided t-test with level $\alpha = 0.05$. UR did not reliably converge for the scenario $n = 500, d = 20$.

C Application to genomic attribution scores

This section gives additional details on the deep learning models considered within our application presented in Section 4 of the main paper. The architecture first employs a one-dimensional convolutional layer with 8 filters (kernel size = 8), designed to scan the input sequences for motif patterns. The output of this layer is processed by a ReLU activation function and batch normalization. This is followed by two fully connected layers, each with 16 units and ReLU activations. To mitigate overfitting, a dropout layer with a rate of 0.3 is applied after each fully connected layer. The architecture concludes with a dense output layer followed by a sigmoid activation function, which outputs a probability score representing the likelihood of the input sequence containing the target motif set. We trained 5 models and evaluated the predictive performance on a separate test data set consisting of 2000 samples (see Table 3).

Model:	Model 1	Model 2	Model 3	Model 4	Model 5
AUC	0.9157 \pm 0.0074	0.9137 \pm 0.0060	0.9037 \pm 0.0069	0.9088 \pm 0.0062	0.9099 \pm 0.0067

Table 3: Performance of the CNNs on the 45 evaluation sequence datasets with each 2000 sequences. AUC are calculated on each dataset separately, since they were used separately for fitting BaGGLS. On average, an AUC of 0.91 could be achieved.