Polygenic-by-Environment Adjustment for Binary GWAS with Out-of-Fold Block-PRS and Low-Rank Bilinear Models

Yining Fan [1]

Abstract

Binary-trait genome-wide association studies (GWAS) typically adjust for covariates and additive polygenic background, but environmental variables can also modulate how aggregate genetic liability is expressed. Unmodelled polygenic gene-environment interaction (G×E) can reduce association power and complicate calibration, while existing methods are limited: SNP×exposure scans are expensive and often underpowered, and scalable adjustment methods generally omit exposure-dependent polygenic effects. We propose a cross-fitted, leave-one-chromosome-out pipeline that combines out-of-fold block-level polygenic scores with a low-rank bilinear neural adjustment. The model decomposes the phenotype logit into environmental main effects, additive polygenic effects, and exposure-modulated polygenic interaction terms, which are then included as covariates in chromosome-wise logistic score tests. In simulations with polygenic interaction, the proposed method maintains calibration while achieving up to 4.5% higher genome-wide power and up to 6.9% higher mean causal χ^2 than additive block-PRS adjustment when interaction variance is present. The learned gate also recovers the simulated environmental modulation function, suggesting a practical route to interaction-aware nuisance adjustment for binary GWAS.

1. Introduction

Genome-wide association studies (GWAS) are a central tool for mapping the genetic basis of complex disease, with modern biobank analyses routinely testing millions of variants in large cohorts (Visscher et al., 2017; Loos, 2020). For binary traits, standard pipelines typically combine single-variant association testing with adjustment for non-genetic covariates such as age, sex, and ancestry principal components, often together with an additive predictor of genome-wide polygenic background (Zhou et al., 2018; Mbatchou et al., 2021; Loya et al., 2025; Hof & Speed, 2025). This strategy is scalable and robust, but it usually treats covariates as additive main effects.

In many biomedical settings, this additive view is restrictive. Covariates and exposures such as smoking, body mass index, sex, age, medication use, and recruitment cohort may act as *effect modifiers*: they can change how aggregate genetic liability is translated into disease risk even when individual SNP-level interaction effects are weak and diffuse (Hunter, 2005; Herrera-Luis et al., 2024). This is especially relevant for disease phenotypes, where polygenic signal is broad, environmental measurements are imperfect, and the dominant interaction structure may arise from many small effects rather than a few easily discoverable SNP×exposure pairs. Unfortunately, exhaustive SNP×environment scans incur severe multiple-testing burdens and often have poor power, while additive polygenic predictors do not explicitly model exposure-dependent modulation.

Recent machine learning methods address part of this problem. Methods such as DeepNull learn nonlinear covariate effects and can improve association testing when phenotype-covariate relationships are misspecified by linear models (McCaw et al., 2022). Other mixed-model methods use genome-wide polygenic predictors to account for additive polygenic background (Loh et al., 2018; Mbatchou et al., 2021; Loya et al., 2025). However, neither class of methods is designed to learn a decomposed, out-of-fold (OOF) adjustment for *polygenic-by-environment* structure in binary-trait GWAS. This leaves a gap between additive adjustment, which may leave interaction-driven residual heterogeneity untreated, and explicit SNP×environment discovery, which is statistically and computationally impractical at scale.

We address this gap by targeting a scalable, power-oriented adjustment problem: learning how environmental features modulate aggregate polygenic liability, without attempting to identify individual SNP×environment effects. Rather than testing locus-specific interactions directly, we ask whether one can learn a cross-fitted nuisance adjustment for the component of disease risk explained by interactions

[1]Department of Statistics, University of Oxford. Correspondence to: Yining Fan <yining.fan@st-hughs.ox.ac.uk>.

Accepted at ICML 2026 AI for Science Workshop.

between environmental features and broad polygenic background. We construct OOF block-level polygenic scores using ridge predictors fit within genomic blocks. For each test chromosome, block predictors from that chromosome are excluded, providing leave-one-chromosome-out (LOCO) protection. A neural adjuster then decomposes the phenotype logit into an environmental main-effect term, an additive block-PRS term, and a low-rank bilinear interaction term between block-PRS features and environmental variables. Cross-fitting is used throughout so that the learned adjustment can be incorporated into downstream logistic score testing without reusing the test chromosome in its own adjustment.

We evaluate this design in biobank-inspired simulations of binary disease traits, using realistic allele-frequency and linkage structure and generating phenotypes with exposure-modulated polygenic liability. We compare against naïve GWAS, additive block-PRS adjustment, and a variance-component proxy that approximates interaction kernels using random Fourier features. Across interaction-present regimes, the proposed model improves genome-wide power and increases mean association signal at causal variants while maintaining calibration comparable to additive adjustment. In no-interaction controls, gains disappear or become mildly conservative, supporting the interpretation that the model captures interaction-specific nuisance structure rather than generic overfitting. The learned decomposition also recovers the simulated environmental modulation function, providing an interpretable representation unavailable from the additive and variance-component baselines.

Contributions.

1. We formulate polygenic-by-environment adjustment for binary GWAS as a decomposed prediction problem over OOF block-PRS features and environmental variables, rather than as explicit SNP×environment discovery.
2. We introduce a scalable low-rank bilinear neural architecture that separates environmental main effects, additive polygenic effects, and polygenic-by-environment interaction terms, with LOCO protection for downstream association testing.
3. We provide a simulation benchmark spanning null, weak, and moderate interaction regimes, showing that the proposed adjustment improves discovery when interaction variance is present, preserves calibration on null chromosomes, and recovers interpretable environmental modulation structure.

These results suggest that neural models can complement statistical genetics pipelines by estimating structured nuisance components that improve downstream scientific inference.

2. Related Work

G×E interaction in complex traits. G×E has long been studied as a source of heterogeneity in genetic effects across exposures, developmental stages, and environmental contexts. In complex traits, however, interaction effects are often weak, diffuse, and context-dependent, making them difficult to detect through individual SNP×exposure tests. Recent reviews emphasize that G×E analysis is further complicated by exposure measurement error, population structure, and cohort effects, all of which can challenge robust inference in large observational cohorts (Herrera-Luis et al., 2024). These challenges motivate methods that can account for broad polygenic interaction structure without requiring well-powered discovery of individual interaction loci.

Polygenic adjustment and scalable regression. A standard strategy for improving calibration and power in GWAS is to adjust for additive polygenic background learned from genome-wide data. Mixed-model association methods operationalize this idea at biobank scale by constructing polygenic adjustment terms that summarize the additive effects of many variants. To avoid proximal contamination, these methods commonly use LOCO schemes, in which predictors used for adjustment exclude the chromosome currently being tested (Listgarten et al., 2012; Yang et al., 2014). This motivates our use of cross-fitted, LOCO-compatible polygenic predictors as covariates in downstream score testing.

Learning nonlinear covariate structure for GWAS. Machine learning has recently been used to improve nuisance adjustment in GWAS. DeepNull (McCaw et al., 2022) showed that replacing a strictly linear covariate model with a flexible predictor of nonlinear covariate effects can improve phenotype prediction and association power while maintaining calibration. Our work extends this work beyond nonlinear non-genetic covariate main effects: we model how environmental features modulate aggregate polygenic liability.

Scalable interaction modeling. Explicit SNP×environment modeling is computationally expensive at genome-wide scale and statistically fragile when interaction signal is highly polygenic. We therefore target aggregate polygenic modulation using a low-rank bilinear model over block-level polygenic predictors and environmental features. As a robustness comparator, we also include a kernel-machine-inspired baseline based on random Fourier features (Rahimi & Recht, 2007), which approximates nonlinear environmental feature maps and allows interaction features to be fit by ridge regression

without constructing an $N \times N$ kernel matrix.

3. Problem Setup

We observe data for N individuals. Let $X \in \{0,1,2\}^{N\times M}$ denote genotypes at M SNPs, with chromosome map $\chi(j) \in \{1,\dots,C\}$ for SNP j, and let $y_i \in \{0,1\}$ denote the binary phenotype. Let $e_i \in \mathbb{R}^d$, stacked as $E \in \mathbb{R}^{N\times d}$, denote observed environmental and non-genetic covariates. These variables serve two roles: they are included explicitly as ordinary covariates in the downstream GWAS null model, and they are also used as candidate effect modifiers in the learned nuisance model. We also allow an optional cohort or ascertainment label d_i, used to monitor cohort-specific shortcuts.

Goal. For each SNP j, we test $H_0 : \beta_j = 0$ in a logistic score-test framework, while controlling null calibration and maximizing power. Standard additive adjustment assumes that environmental variables and polygenic background enter the phenotype model as separable main effects. Under polygenic G×E, however, disease risk may depend not only on an additive polygenic component, but also on how environmental features modulate that component. If this interaction-driven nuisance structure is omitted, residual variation can remain systematically structured, reducing discovery power and potentially affecting calibration.

LOCO scheme. To avoid proximal contamination, the adjustment terms used when testing SNPs on chromosome c must not be trained using variants from chromosome c. We therefore require all polygenic features used in the null model for chromosome c to be constructed from chromosomes $\neq c$.

Adjustment target. Our aim is not to test SNP-specific interaction effects. Instead, we seek a cross-fitted nuisance adjustment that captures broad polygenic-by-environment structure and can be included in the null model before single-SNP testing. This distinction is important: the learned interaction term is used to improve calibration and sensitivity of main-effect GWAS, not to claim discovery of individual G×E loci. Throughout, E and d are retained as explicit covariates in the downstream association test. The learned interaction adjustment is added on top of this standard covariate adjustment rather than replacing it.

4. Method

4.1. Overview

For each test chromosome c, our pipeline constructs OOF block-level polygenic predictors using only variants outside chromosome c. These LOCO-protected predictors are then used as nuisance covariates in a second-stage logistic score test for SNPs on chromosome c. We compare four null models: (i) a naïve covariate-only baseline, (ii) an additive block-PRS baseline, (iii) a scalable variance-component proxy for polygenic G×E, and (iv) our low-rank bilinear neural adjustment, which decomposes the phenotype logit into environmental main, additive polygenic, and polygenic-by-environment interaction components.

4.2. Step 1: Out-of-fold block-PRS features

Block construction. We partition SNPs by chromosome and then divide each chromosome into contiguous genotype blocks of size L SNPs. Let $\mathcal{B} = \{1,\dots,B\}$ index all blocks, and let $X_{(b)} \in \mathbb{R}^{N\times m_b}$ denote the genotype submatrix for block b.

Per-block ridge predictors. For each block b, we fit a ridge predictor of the binary phenotype using only individuals in the training fold [1]:

$$\widehat{w}_b = \arg\min_{w} \sum_{i\in\mathcal{T}} \left(y_i - x_{i,(b)}^\top w\right)^2 + \lambda\|w\|_2^2. \quad (1)$$

Although the phenotype is binary, this squared-error ridge objective is used only to construct scalable nuisance predictors, in the spirit of two-stage regression methods such as REGENIE (Mbatchou et al., 2021). Predictions are generated on the held-out fold $\mathcal{V}$, and stacking across folds gives an OOF block predictor matrix

$$Z \in \mathbb{R}^{N\times B}, \qquad Z_{ib} = x_{i,(b)}^\top \widehat{w}_b^{(-k(i))}, \quad (2)$$

where $k(i)$ denotes the fold containing individual i. Thus, Z_{ib} is always predicted by a model not trained on individual i.

LOCO-at-block-level. Let $\mathcal{B} = \{1,\dots,B\}$ denote the set of all genomic blocks, and let $\mathcal{B}(c) \subset \mathcal{B}$ denote the blocks located on chromosome c. For chromosome-wise testing, we retain only non-c blocks:

$$\mathcal{B}^{(-c)} = \mathcal{B} \setminus \mathcal{B}(c), \qquad B_{-c} = |\mathcal{B}^{(-c)}|. \quad (3)$$

All downstream additive and interaction adjustments for chromosome c are trained using $Z^{(-c)} = Z[:,\mathcal{B}^{(-c)}] \in \mathbb{R}^{N\times B_{-c}}$, ensuring that the null model does not contain predictors derived from the chromosome currently being tested. Write $z_i^{(-c)} \in \mathbb{R}^{B_{-c}}$ for row i of $Z^{(-c)}$. The VC proxy and NN adjustment defined below target additional disease-risk dependence on $(z_i^{(-c)}, e_i)$ beyond the explicit covariates $[1, E, d]$.

[1] In implementation, genotypes are standardized using training-fold means and variances, and each block ridge predictor includes an intercept. Eq. 1 omits the intercept for compactness; observed covariates are retained explicitly in the downstream null model and are also provided to the neural nuisance model.

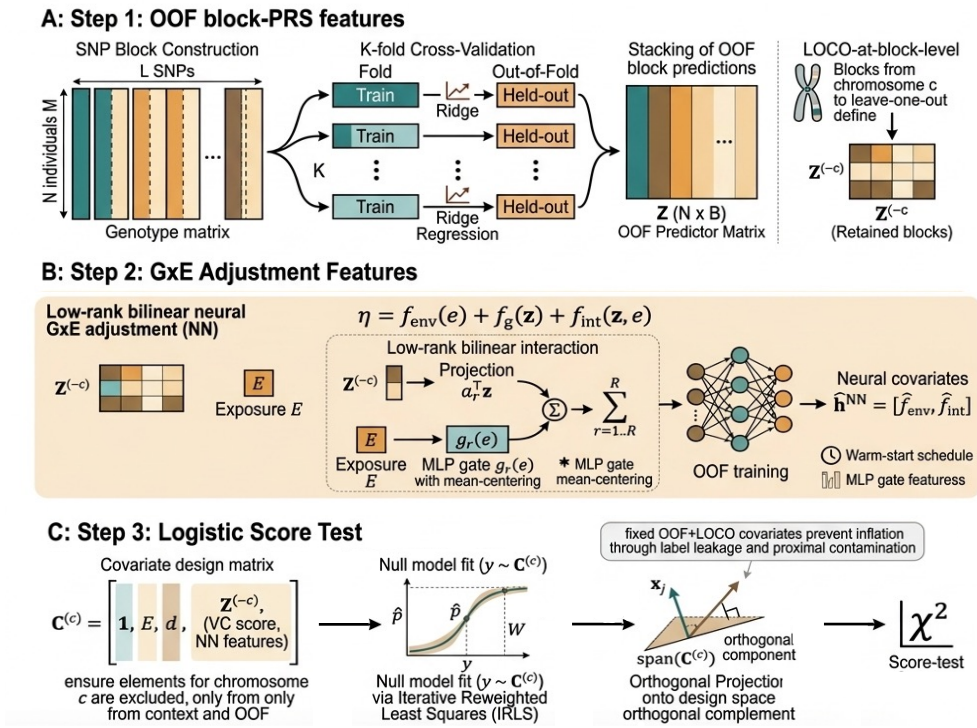

Figure 1. **Overview of the proposed G×E-aware GWAS adjustment pipeline.** Genotypes and exposures are used to generate binary traits with polygenic environmental modulation. We construct OOF block-PRS predictors, apply LOCO protection for each test chromosome, learn additive and interaction nuisance components, and include the resulting adjustment terms in chromosome-wise logistic score tests.

4.3. Additive and kernel-proxy baselines

Additive block-PRS adjustment (BASE). The additive baseline uses the same ordinary covariate adjustment as the downstream GWAS null model and augments it with LOCO block-PRS features included linearly. For chromosome c, the BASE null model is

$$\eta_i = \beta_0 + \beta_E^\top e_i + \beta_D d_i + \beta_Z^\top z_i^{(-c)}, \tag{4}$$

where $\eta_i = \mathrm{logit}\{\Pr(y_i = 1)\}$, e_i denotes the observed environmental and non-genetic covariates, d_i is the optional cohort or ascertainment label, and $z_i^{(-c)} \in \mathbb{R}^{B_{-c}}$ is row i of the LOCO block-PRS matrix $Z^{(-c)}$. Thus, BASE performs standard explicit covariate adjustment through $[1, E, d]$ and additive polygenic adjustment through $Z^{(-c)}$. It does not include nonlinear environmental effects or exposure-dependent modulation of polygenic background, which are the additional nuisance components targeted by the VC proxy and the proposed NN adjustment.

Variance-component proxy (VC). As a scalable proxy for polygenic G×E variance-component adjustment, we construct random Fourier features (RFF) $\phi(e_i) \in \mathbb{R}^r$ approximating an RBF kernel over exposures (Rahimi & Recht, 2007). We then form interaction features

$$u_i = \mathrm{vec}\Big(z_i^{(-c)} \otimes \phi(e_i)\Big) \in \mathbb{R}^{B_{-c}r}. \tag{5}$$

Within each cross-fitting fold, we first fit an additive ridge model on $[E, d, Z^{(-c)}]$, omitting d when no cohort label is available, compute residuals on the training fold, and then fit a second ridge model from u_i to these residuals. This produces an OOF interaction score $\widehat{s}_i^{\mathrm{VC}}$, which is included as an additional covariate in the second-stage null model. This baseline approximates a product interaction kernel, $K_{G\times E} \approx K_G \circ K_E$, without constructing an $N \times N$ kernel matrix.

4.4. Low-rank bilinear neural G×E adjustment

Decomposed phenotype logit. Our main model decomposes the phenotype logit as

$$\eta_i = \alpha_0 + \alpha_E^\top e_i + \alpha_D d_i + f_{\mathrm{env}}(e_i) + f_g(z_i^{(-c)}) + f_{\mathrm{int}}(z_i^{(-c)}, e_i), \tag{6}$$

where e_i and d_i enter as ordinary adjustment variables during neural nuisance learning, while f_{env} captures nonlinear covariate main effects beyond the explicit linear adjustment, f_g captures additive polygenic background, and f_{int} captures exposure-dependent modulation of polygenic effects.

The same ordinary covariates are also retained explicitly in the downstream GWAS null model; the learned NN scores are therefore added on top of standard covariate and additive block-PRS adjustment, rather than replacing them.

Low-rank bilinear interaction. To avoid enumerating SNP×environment terms, we parameterize the interaction using a rank-R bilinear form:

$$f_{\mathrm{g}}(z) = w^{\top} z, \tag{7}$$

$$f_{\mathrm{int}}(z, e) = \sum_{r=1}^{R} \left(a_r^{\top} z\right) g_r(e). \tag{8}$$

Here $a_r^{\top} z$ is a learned projection of the LOCO block-PRS vector, and $g_r(e)$ is an exposure-dependent gate produced by a multilayer perceptron. We apply a $\tanh$ nonlinearity to the gates and center them across individuals, preventing a constant gate from simply rescaling the additive genetic main effect. This low-rank structure scales as $O(NB_{-c}R)$ and targets broad polygenic modulation rather than locus-specific SNP×environment discovery.

Cross-fitted training. For each chromosome c, the neural model is trained in K folds using $Z^{(-c)}$, E, and the optional cohort label d when available. Each held-out fold receives OOF predictions for the total logit and its components. To stabilize the decomposition, we first train the environmental and additive genetic main effects with the interaction term switched off, then gradually increase the interaction contribution during training. When a cohort label is available, we optionally add a domain-adversarial loss on the environmental representation using a gradient reversal layer to discourage cohort-specific shortcuts.

Neural adjustment covariates. The OOF neural covariates included in the second-stage null model are

$$\widehat{h}_i^{\mathrm{NN}} = \left[\widehat{f}_{\mathrm{env}}(e_i), \widehat{f}_{\mathrm{int}}(z_i^{(-c)}, e_i)\right]. \tag{9}$$

These learned scores are included in addition to the explicit baseline covariates $[1, E, d, Z^{(-c)}]$. Thus, ordinary covariates enter both the neural nuisance-learning stage and the final GWAS association test: the neural model uses them to learn nonlinear and interaction-dependent nuisance structure, while the score test still performs standard covariate adjustment explicitly. We do not include $\widehat{f}_g$ separately in the second-stage covariates because the LOCO block-PRS matrix $Z^{(-c)}$ is already included linearly and spans this additive component.

4.5. Step 2: Chromosome-wise logistic score test

For a given chromosome c, define the null design matrix

$$C^{(c)} = \left[\mathbf{1},\, E,\, d,\, Z^{(-c)},\, \text{learned adjustment scores}\right]. \tag{10}$$

The learned adjustment-score block is omitted for the naïve and BASE null models, is s_i^{VC} for the VC proxy, and is $\widehat{h}_i^{\mathrm{NN}}$ for the proposed NN adjustment. We fit the logistic null model $y \sim C^{(c)}$ using iteratively reweighted least squares with a small ridge term for numerical stability; see Appendix A for details. This yields fitted probabilities $\widehat{p}_i$ and diagonal weight matrix $W = \mathrm{diag}\{\widehat{p}_i(1-\widehat{p}_i)\}_{i=1}^{N}$. For each SNP j on chromosome c, with standardized genotype vector x_j, we compute the one-degree-of-freedom logistic score statistic after projecting x_j onto the W-weighted orthogonal complement of the columns of $C^{(c)}$.

Because learned adjustment terms are generated OOF and exclude variants on the tested chromosome, they reduce individual-level label leakage and proximal contamination. We therefore treat them as nuisance covariates in the association test and evaluate calibration empirically on chromosomes with no simulated causal variants.

4.6. Computational considerations

The block-PRS stage is parallelizable over blocks and cross-fitting folds. The low-rank neural interaction scales as $O(NB_{-c}R)$, where B_{-c} is the number of retained LOCO blocks for chromosome c and R is the interaction rank. The VC proxy scales as $O(NB_{-c}r)$ for RFF dimension r. Both approaches avoid explicit SNP×environment enumeration and avoid $N \times N$ kernel construction, making them compatible with chromosome-wise GWAS workflows. Implementation details are provided in Appendix C.

5. Results

5.1. Simulation grid and evaluation metrics

We evaluate the proposed adjustment in controlled simulations where the ground-truth additive genetic, environmental-modulation, and polygenic interaction components are known. We simulate genotypes for $N = 10{,}000$ unrelated individuals using `msprime` (Baumdicker et al., 2022), under a GBR demographic model inferred from the 1000 Genomes Project (Spence & Song, 2019). To keep experiments computationally tractable while retaining linkage and realistic allele-frequency structure, we simulate 8 chromosomes, each of length 2×10^5 bp, and remove monomorphic variants.

Binary phenotypes are generated from a liability model (Lee et al., 2011) with additive genetic variance h^2 and a global polygenic G×E component with variance $\sigma^2_{G\times E}$. Briefly,

causal SNPs are restricted to a subset of chromosomes, leaving the others as null chromosomes for calibration assessment. The latent liability is generated as

$$\ell_i = g_i + u_i + \epsilon_i, \qquad u_i \propto f(e_i) g_i,$$

where g_i is a standardized additive polygenic component and $f(e_i)$ is a centered, standardized nonlinear environmental modulation function; full details are given in Appendix B. We sweep $h^2 \in \{0.2, 0.4, 0.6\}$, $\sigma^2_{G\times E} \in \{0, 0.1, 0.2\}$, where $\sigma^2_{G\times E} = 0$ is a true no-interaction control. Each grid cell is repeated for 30 independent replicates. We compare four approaches: naïve GWAS, additive block-PRS adjustment (BASE), the random-feature variance-component proxy (VC), and the proposed low-rank bilinear neural adjustment (NN).

We evaluate three aspects of performance: calibration, discovery, and mechanistic recovery. Calibration is measured using Type I error at $p < 0.05$ on null chromosomes. Discovery is measured by Bonferroni-threshold power, $p < \alpha/M$ with $\alpha = 0.05$, and by mean χ^2 at causal SNPs. Because our Bonferroni-threshold false-positive metric is computed as the average false-positive rate across null SNPs, we report it as FPR at α/M rather than family-wise error rate.

5.2. Calibration: interaction adjustment remains well controlled

Figure 2 shows Type I error on chromosomes with no simulated causal variants. The naïve GWAS baseline is consistently inflated, indicating that observed covariates alone do not absorb the polygenic structure induced by the simulated genomes. BASE substantially improves calibration, bringing the null Type I rate close to the nominal 0.05 level.

Adding interaction-aware adjustment does not reintroduce inflation. Both the VC proxy and the NN adjustment remain well controlled across interaction regimes, with the NN often slightly conservative at larger h^2. This is important because the NN introduces a flexible learned nuisance component; the results show that cross-fitting and LOCO protection empirically preserve calibration in the simulated setting.

5.3. Power gains are specific to interaction-present regimes

Figure 3 summarizes the relative power gain of NN adjustment over BASE across the full simulation grid. The pattern is consistent with an interaction-specific effect. When the simulated interaction variance is zero, NN adjustment provides no systematic power benefit and is slightly conservative in two of the three heritability settings. This behavior is expected: in the absence of polygenic G×E, the additional interaction module has no true interaction signal to exploit. When interaction variance is nonzero, the NN adjustment yields positive gains over the additive baseline in most settings. At $\sigma^2_{G\times E} = 0.2$, NN improves genome-wide power by approximately 1.8%, 1.7%, and 4.5% at $h^2 = 0.2, 0.4, 0.6$, respectively. These results suggest that the method does not act as a generic power-inflating transformation; instead, its benefit is concentrated in regimes where exposure-dependent polygenic modulation is present.

5.4. Interaction-aware adjustment increases causal signal

At the moderate sample size used here, Bonferroni-corrected threshold discoveries are discrete and therefore power differences are modest. We therefore report both genome-wide power and a continuous signal-strength metric: the mean χ^2 statistic at causal SNPs. This metric directly measures whether an adjustment sharpens association signal at true causal variants, even when few variants cross the genome-wide threshold.

Figure 4 shows the relative gain in mean causal χ^2 for NN compared with naïve GWAS, BASE, and VC. When $\sigma^2_{G\times E} = 0$, gains over BASE and VC are essentially zero, as expected when there is no interaction component to exploit. As interaction variance increases, the NN yields clear gains in causal signal: for $\sigma^2_{G\times E} = 0.2$, mean causal χ^2 increases by approximately 0.8%, 2.4%, and 6.9% over BASE at $h^2 = 0.2, 0.4, 0.6$, respectively, with comparable gains over the VC proxy. Relative to naïve GWAS, the gain is larger, reaching roughly 21% at $h^2 = 0.6$, reflecting both improved calibration and better adjustment for polygenic structure. These results provide a threshold-free view of association improvement and further support the intended behavior of the method: the learned interaction term increases causal signal specifically when exposure-modulated polygenic structure is present.

Additional replicate-level diagnostics are shown in Appendix D.1. The power-calibration trade-off confirms that NN adjustment improves discovery without moving into the inflated Type I error region occupied by naïve GWAS. In contrast, the NN points remain close to the calibrated polygenic-adjustment methods while achieving comparable or higher power in interaction-present regimes. This replicate-level view supports the main-text summary that the proposed low-rank bilinear adjustment improves association testing without introducing uncontrolled inflation.

5.5. Mechanistic validation: the NN recovers the simulated G×E structure

Because the simulations provide ground-truth components, we can test whether the learned model recovers the intended mechanism rather than merely improving prediction. Figure 5 reports component-level recovery in the interaction

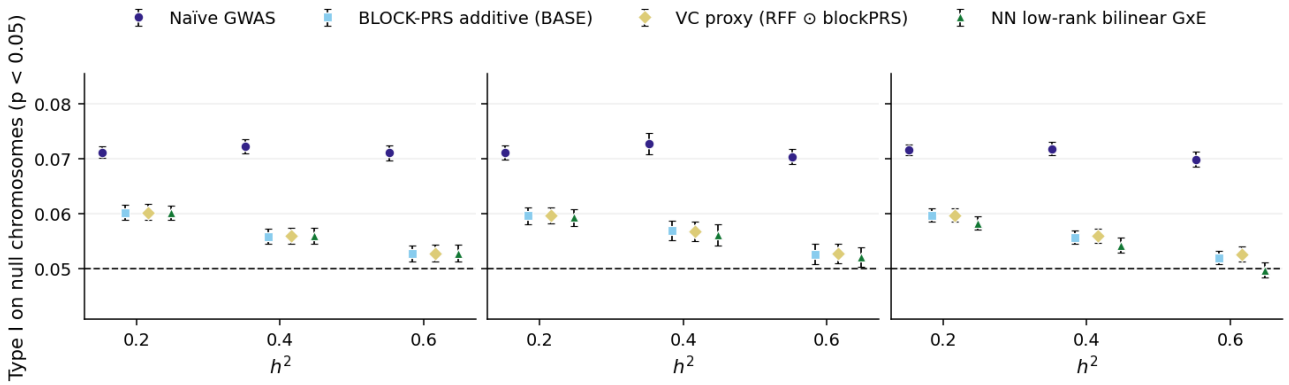

Figure 2. **Calibration on null chromosomes.** Type I error at $p < 0.05$ on chromosomes containing no simulated causal variants, stratified by interaction variance: **left**: $\sigma^2_{G\times E} = 0$; **middle**: $\sigma^2_{G\times E} = 0.1$; **right**: $\sigma^2_{G\times E} = 0.2$. Points and error bars show means and 95% bootstrap confidence intervals (CIs) across 30 replicates; the dashed line denotes the nominal 0.05 level.

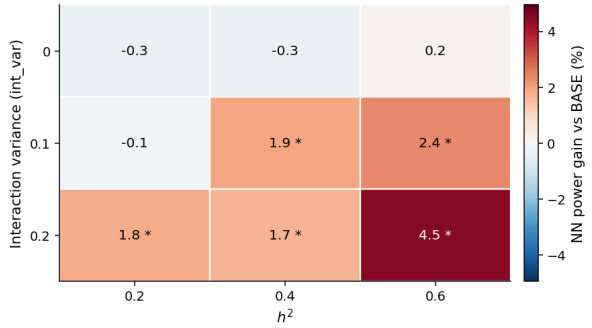

Figure 3. **Interaction-specific power gains.** Mean relative gain in genome-wide power of NN adjustment compared with BASE across the simulation grid. Asterisks indicate bootstrap 95% CIs excluding zero. Gains are near zero under the no-interaction control and become positive when interaction variance is present.

regime $\sigma^2_{G\times E} = 0.2$. We measure recovery by the replicate-level R^2 between inferred components and their corresponding simulated components.

The NN interaction score shows modest but consistent alignment with the simulated interaction component and improves over the VC proxy. The learned environmental gate recovers the modulation function more strongly, indicating that the model captures exposure-dependent modulation of aggregate genetic liability rather than simply absorbing residual variation through additional flexibility. This component-level recovery is a key advantage over the baselines: naïve and additive adjustment provide no explicit interaction decomposition, while the VC proxy produces only an interaction score without a learned environmental modulation function. Thus, the proposed model provides both a calibrated adjustment for association testing and an interpretable representation of the simulated G×E mechanism.

6. Discussion and Future Work

We presented a scalable approach for polygenic G×E adjustment in binary-trait GWAS. The method combines OOF block-level polygenic predictors with LOCO protection and a decomposed neural architecture that separates environmental main effects, additive polygenic effects, and low-rank polygenic-by-environment interactions. In simulations with realistic linkage structure and polygenic architecture, the proposed adjustment improves causal-signal enrichment and yields interaction-specific power gains while maintaining calibration on null chromosomes.

Limitations. Our experiments are intentionally simulation-based because the key quantities of interest, the environmental modulation function and the polygenic interaction component, are not directly observable in real data. This controlled setting lets us evaluate two properties that are otherwise difficult to measure jointly: calibration on chromosomes known to be null and component recovery against known G×E ground truth. As a falsification check, we include true no-interaction regimes; in these settings, NN gains disappear or become mildly conservative rather than producing generic power inflation. The simulations nevertheless simplify real biobank analyses, where exposures may be noisy, correlated with ascertainment or ancestry, and only partially capture the relevant environment. In addition, our method targets diffuse polygenic modulation rather than locus-specific SNP×environment discovery.

Extension to continuous traits. Although we focus on binary disease phenotypes, the same OOF/LOCO adjustment can be used for quantitative traits by replacing the logistic loss with a Gaussian loss and replacing the logistic score test with a residualized linear association test. We leave empirical evaluation of continuous traits to future work.

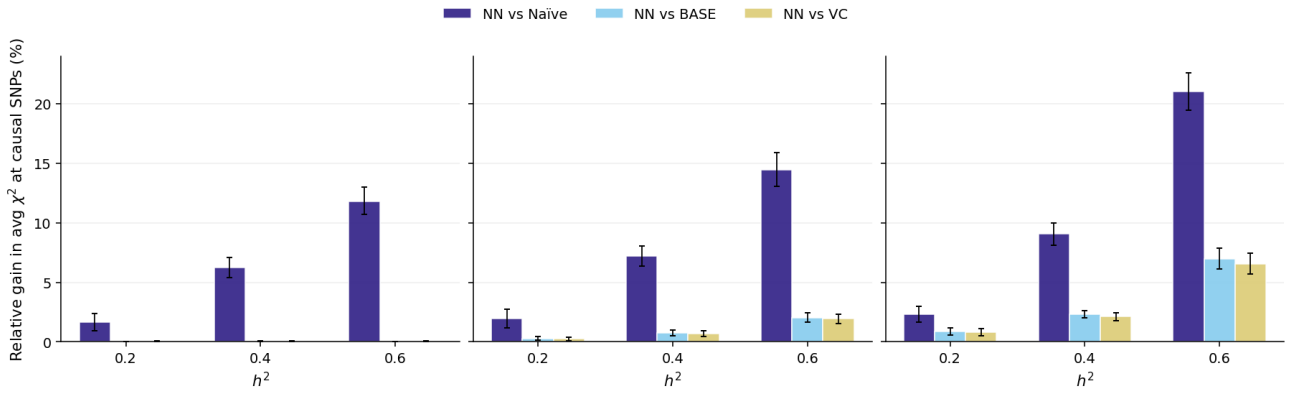

Figure 4. **Causal association signal improves under NN adjustment.** Relative gain in the mean χ^2 statistic at causal SNPs for NN adjustment compared with naïve GWAS, BASE, and the VC proxy. **Left**: $\sigma^2_{G\times E} = 0$; **middle**: $\sigma^2_{G\times E} = 0.1$; **right**: $\sigma^2_{G\times E} = 0.2$. Bars show means across 30 replicates and error bars show 95% bootstrap CIs.

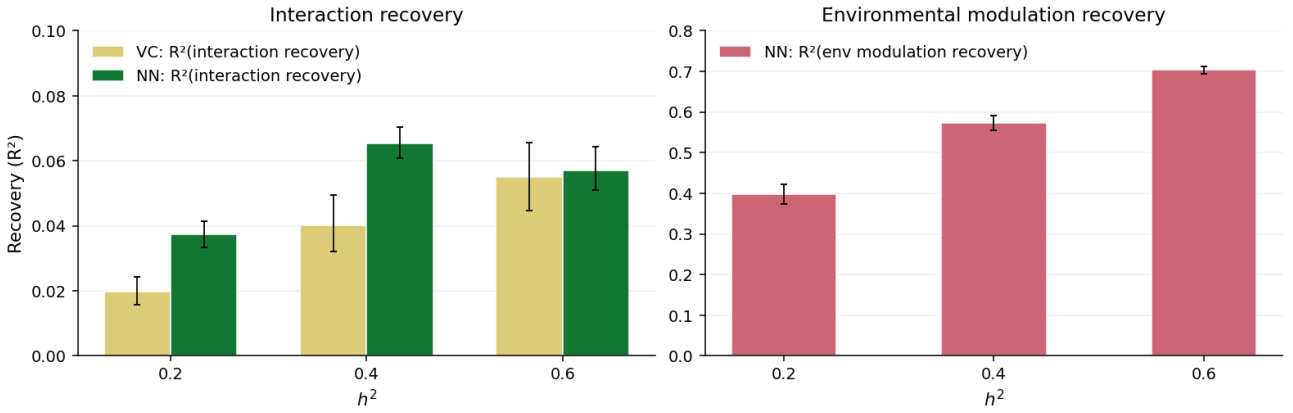

Figure 5. **Mechanistic recovery under polygenic G×E.** Results are shown for the main interaction regime with $\sigma^2_{G\times E} = 0.2$. **Left**: recovery of the simulated polygenic interaction component, measured by R^2 between the inferred interaction score and the ground-truth interaction component. **Right**: recovery of the environmental modulation function, measured by R^2 between the learned NN gating component and the ground-truth modulation function. Bars show means and 95% bootstrap CIs across 30 replicates.

Future work. Future work should prioritize real-cohort validation through calibration diagnostics, known-locus enrichment, cohort replication, and sensitivity to exposure definitions. Methodologically, the interaction module could be extended to multiple exposure groups or time-varying exposures, and additional regularizers could improve identifiability by encouraging orthogonality between additive and interaction components. Multi-ancestry settings will also require ancestry-aware block construction and domain-robust training, since both block predictors and environmental distributions may shift across populations.

Impact Statement

This work aims to improve the reliability and statistical efficiency of genome-wide association studies by learning structured polygenic-by-environment adjustment terms for binary traits. If validated in real cohorts, such methods could help researchers make better use of biobank-scale genetic and environmental data, improve discovery power for complex diseases, and reduce false or unstable associations caused by inadequately modelled background risk. More broadly, the approach illustrates how machine learning can be used not to replace statistical genetics workflows, but to estimate structured nuisance components that improve downstream scientific inference.

Potential risks arise because genetic, health, and environmental data are sensitive, and exposures may correlate with ancestry, socioeconomic status, geography, or healthcare access. The method should therefore be used with appropriate data governance, ancestry- and cohort-aware validation, calibration checks, and transparent exposure definitions. The present study is simulation-based and makes no claims about specific diseases, populations, or clinical deployment.

Acknowledgements and Funding Disclosure

The author is supported by the Clarendon Fund at the University of Oxford. The author thanks the ICML 2026 AI for Science Workshop reviewers and organizers for their feedback and for providing the venue for presenting this work.

References

Baumdicker, F., Bisschop, G., Goldstein, D., Gower, G., Ragsdale, A. P., Tsambos, G., Zhu, S., Eldon, B., Ellerman, E. C., Galloway, J. G., Gladstein, A. L., Gorjanc, G., Guo, B., Jeffery, B., Kretzschumar, W. W., Lohse, K., Matschiner, M., Nelson, D., Pope, N. S., Quinto-Cortés, C. D., Rodrigues, M. F., Saunack, K., Sellinger, T., Thornton, K., van Kemenade, H., Wohns, A. W., Wong, Y., Gravel, S., Kern, A. D., Koskela, J., Ralph, P. L., and Kelleher, J. Efficient ancestry and mutation simulation with msprime 1.0. *Genetics*, 220:iyab229, 2022. doi: 10.1093/genetics/iyab229.

Herrera-Luis, E., Benke, K., Volk, H., et al. Gene–environment interactions in human health. *Nature Reviews Genetics*, 25:768–784, 2024. doi: 10.1038/s41576-024-00731-z.

Hof, J. P. and Speed, D. LDAK-KVIK performs fast and powerful mixed-model association analysis of quantitative and binary phenotypes. *Nature Genetics*, 57:2116–2123, 2025. doi: 10.1038/s41588-025-02286-z.

Hunter, D. J. Gene–environment interactions in human diseases. *Nature Reviews Genetics*, 6:287–298, 2005. doi: 10.1038/nrg1578.

Lee, S. H., Wray, N. R., Goddard, M. E., and Visscher, P. M. Estimating missing heritability for disease from genome-wide association studies. *American Journal of Human Genetics*, 88:294–305, 2011. doi: 10.1016/j.ajhg.2011.02.002.

Listgarten, J., Lippert, C., Kadie, C., et al. Improved linear mixed models for genome-wide association studies. *Nature Methods*, 9:525–526, 2012. doi: 10.1038/nmeth.2037.

Loh, P.-R., Kichaev, G., Gazal, S., Schoech, A. P., and Price, A. L. Mixed-model association for biobank-scale datasets. *Nature Genetics*, 50:906–908, 2018. doi: 10.1038/s41588-018-0144-6.

Loos, R. J. 15 years of genome-wide association studies and no signs of slowing down. *Nature Communications*, 11(1):5900, 2020. doi: 10.1038/s41467-020-19653-5.

Loshchilov, I. and Hutter, F. Decoupled Weight Decay Regularization. *arXiv e-prints*, art. arXiv:1711.05101, November 2017. doi: 10.48550/arXiv.1711.05101.

Loya, H., Kalantzis, G., Cooper, F., et al. A scalable variational inference approach for increased mixed-model association power. *Nature Genetics*, 57:461–468, 2025. doi: 10.1038/s41588-024-02044-7.

Mbatchou, J., Barnard, L., Backman, J., et al. Computationally efficient whole-genome regression for quantitative and binary traits. *Nature Genetics*, 53:1097–1103, 2021. doi: 10.1038/s41588-021-00870-7.

McCaw, Z. R., Colthurst, T., Yun, T., et al. DeepNull models non-linear covariate effects to improve phenotypic prediction and association power. *Nature Communications*, 13:241, 2022. doi: 10.1038/s41467-021-27930-0.

Rahimi, A. and Recht, B. Random features for large-scale kernel machines. In *Proceedings of the 21st International Conference on Neural Information Processing Systems*, NIPS'07, pp. 1177–1184, Red Hook, NY, USA, 2007. Curran Associates Inc. ISBN 9781605603520.

Spence, J. P. and Song, Y. S. Inference and analysis of population-specific fine-scale recombination maps across 26 diverse human populations. *Science Advances*, 5: eaaw9206, 2019. doi: 10.1126/sciadv.aaw9206.

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., and Yang, J. 10 years of GWAS discovery: Biology, function, and translation. *The American Journal of Human Genetics*, 101(1):5–22, 2017. doi: 10.1016/j.ajhg.2017.06.005.

Yang, J., Zaitlen, N., Goddard, M., et al. Advantages and pitfalls in the application of mixed-model association methods. *Nature Genetics*, 46:100–106, 2014. doi: 10.1038/ng.2876.

Zhou, W., Nielsen, J. B., Fritsche, L. G., et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nature Genetics*, 50:1335–1341, 2018. doi: 10.1038/s41588-018-0184-y.

A. Logistic score test

We fit the logistic null model $y \sim C^{(c)}$ using iteratively reweighted least squares with a small ridge term for numerical stability. This yields fitted probabilities $\widehat{p}_i$ and diagonal weight matrix

$$W = \mathrm{diag}\{\widehat{p}_i(1-\widehat{p}_i)\}_{i=1}^N.$$

For each SNP j on chromosome c, let x_j denote its standardized genotype vector. We first compute the weighted residual of x_j after projection onto the null-design columns:

$$\widetilde{x}_j = x_j - C^{(c)} \left\{ (C^{(c)})^\top W C^{(c)} + \rho P_C \right\}^{-1} (C^{(c)})^\top W x_j,$$

where ρ is a small ridge parameter and P_C is the penalty matrix, with zeros on any unpenalized columns such as the intercept. The score, variance estimate, and one-degree-of-freedom statistic are

$$U_j = \widetilde{x}_j^\top (y - \widehat{p}), \qquad V_j = \widetilde{x}_j^\top W \widetilde{x}_j, \qquad T_j = \frac{U_j^2}{V_j}.$$

We report p-values from the χ^2_1 reference distribution.

B. Phenotype data-generating process

Binary phenotypes are generated from a liability-threshold model with additive polygenic liability and exposure-modulated polygenic interaction. Causal SNPs are sampled uniformly from a prespecified subset of chromosomes, chosen here as the even-indexed chromosomes, leaving the remaining chromosomes as null chromosomes for calibration assessment. For individual i, we first construct an additive genetic component

$$g_i = \sqrt{h^2}\, \frac{x_i^\top \beta - \overline{x^\top \beta}}{\mathrm{sd}(x^\top \beta)},$$

where β is nonzero only at causal SNPs. We then define a mean-zero, unit-variance environmental modulation function using age, sex, BMI, and smoking:

$$f(E_i) = 0.55 \sin(2.5\,\mathrm{BMI}_{i,\mathrm{std}}) + 0.25(\mathrm{smoking}_i - \overline{\mathrm{smoking}}) + 0.20\,\mathrm{age}_{i,\mathrm{std}}(\mathrm{sex}_i - \overline{\mathrm{sex}}),$$

followed by centering and scaling to unit variance. The polygenic interaction component is

$$u_i = \sqrt{\sigma^2_{G\times E}}\, \frac{f(E_i) g_i - \overline{f(E) g}}{\mathrm{sd}(f(E) g)}.$$

All components are centered and scaled empirically within each replicate; because exposures are generated independently of genotypes and $f(E)$ is centered, the additive and interaction components are approximately uncorrelated in the simulated population. The latent liability is

$$\ell_i = g_i + u_i + \epsilon_i, \qquad \epsilon_i \sim N\big(0, 1 - h^2 - \sigma^2_{G\times E}\big).$$

Binary outcomes are obtained by thresholding liability at the target prevalence $\pi = 0.15$:

$$y_i = \mathbf{1}\{\ell_i \geq q_{1-\pi}(\ell)\}.$$

We sweep $h^2 \in \{0.2, 0.4, 0.6\}$ and $\sigma^2_{G\times E} \in \{0, 0.1, 0.2\}$, with $\sigma^2_{G\times E} = 0$ serving as a true no-interaction control, and repeat each setting for 30 independent replicates.

C. Implementation Details

Unless otherwise stated, we use $K = 5$ cross-fitting folds and genomic blocks of $L = 500$ SNPs. Block ridge predictors use penalty $\lambda = 50$, with genotypes standardized using training-fold means and variances. The VC proxy uses $r = 32$ random Fourier features for the exposure kernel. The NN adjustment uses interaction rank $R = 8$, an MLP gate with 2 hidden layers of widths 64 and 32, and is trained with `AdamW` (Loshchilov & Hutter, 2017) with learning rate 5×10^{-3}, weight decay 10^{-2}, batch size 1024, and 100 epochs. Hyperparameters are fixed across simulation replicates and grid settings.

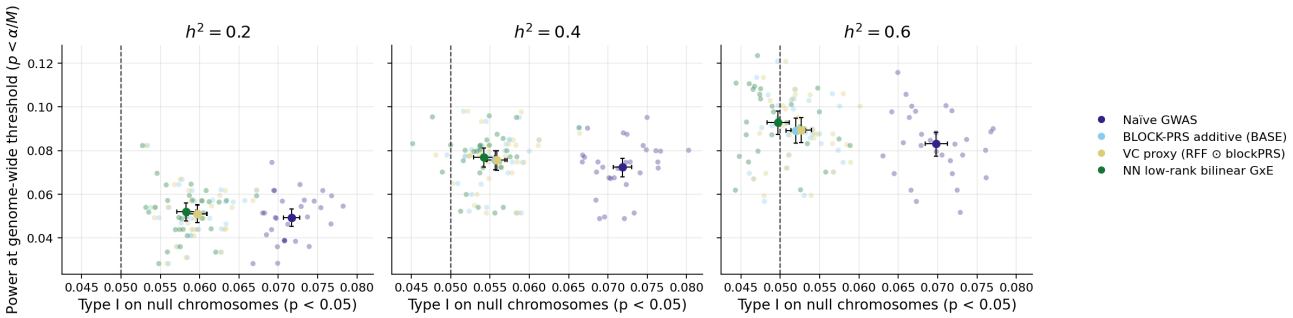

Figure 6. **Power–calibration trade-off under polygenic G×E.** Each point represents one simulation replicate in the main interaction regime, $\sigma^2_{G\times E} = 0.2$. The x-axis shows Type I error on null chromosomes at $p < 0.05$, and the y-axis shows genome-wide power at the Bonferroni-corrected threshold $p < \alpha/M$. Large markers with error bars show mean and bootstrap 95% confidence intervals across 30 replicates. The dashed vertical line marks nominal Type I error 0.05.

D. Additional Simulation Results

D.1. Power-calibration trade-off across replicates

Figure 6 provides a replicate-level view of the trade-off between discovery power and null calibration in the main interaction regime, $\sigma^2_{G\times E} = 0.2$. Each point corresponds to one simulation replicate, with the x-axis showing Type I error on null chromosomes at $p < 0.05$ and the y-axis shows genome-wide power at the Bonferroni-corrected threshold $p < \alpha/M$, with $\alpha = 0.05$. Large markers with error bars denote the mean and bootstrap 95% confidence intervals across replicates.

The naïve GWAS baseline is consistently shifted to the right, indicating inflated Type I error on null chromosomes. In contrast, all polygenic-adjusted methods substantially improve calibration. Among the adjusted methods, the NN low-rank bilinear G×E model lies in a favorable region of the trade-off: it remains close to the nominal calibration target while matching or improving discovery power relative to BASE and the VC proxy. This replicate-level visualization supports the summary results in the main text and shows that the reported gains are not driven by a small number of outlier runs.