A biobank-scale method for learning modulators of gene-environment interaction underlying human complex traits from multiple environmental exposures

Zhengtong Liu, Arush Ramteke, Aakarsh Anand, et al.

Genome Res. published online August 18, 2026
Access the most recent version at doi:10.1101/gr.282105.126

A biobank-scale method for learning modulators of gene-environment interaction underlying human complex traits from multiple environmental exposures

Zhengtong Liu[1], Arush Ramteke[2], Aakarsh Anand[1], Aditya Gorla[3], Moonseong Jeong[1], and Sriram Sankararaman[*1,4,5]

[1] *Department of Computer Science, UCLA, Los Angeles, CA 90095, USA*
[2] *Department of Computer Science, Courant Institute of Mathematical Sciences, NYU, NY 10012, USA.*
[3] *Bioinformatics Interdepartmental Program, UCLA, Los Angeles, CA 90095, USA*
[4] *Department of Human Genetics, David Geffen School of Medicine, UCLA, Los Angeles, CA 90095, USA*
[5] *Department of Computational Medicine, David Geffen School of Medicine, UCLA, Los Angeles, CA 90095, USA*
[*] *Corresponding author: sriram@cs.ucla.edu*

Abstract

It is increasingly recognized that genetic effects on complex traits and diseases are shaped by environmental context. Biobanks that measure diverse environmental exposures alongside genotypes and phenotypes at scale enable systematic study of gene-environment (G×E) interactions. Existing approaches, however, are limited in their ability to accurately model polygenic G×E involving many exposures across genome-wide genetic variants. It is unclear which exposure combinations are relevant for a given trait while distinguishing true interactions from environment-dependent heteroskedastic noise. To address these challenges, we develop *Efficient multi-eNvironmental Gene-environment Interaction iNference Estimator* (ENGINE), a supervised variance-component framework that learns an embedding that combines multiple environmental exposures while jointly estimating additive, G×E, and heteroskedastic noise components. To enable biobank-scale inference, ENGINE makes a single pass over the genotype matrix to cache genotype-dependent summaries, then assembles normal-equation components and gradients at each iteration. In simulations, ENGINE controls type I error rates, achieves high power, and accurately recovers the environmental embedding while remaining efficient at biobank-scale. It is roughly five-fold faster than the state-of-the-art method at biobank scale, making polygenic G×E analysis tractable when both the number of individuals and the number of SNPs reach the millions. Applied to five complex traits paired with lifestyle exposures in $N = 291{,}273$ unrelated White British individuals and $M = 454{,}207$ common SNPs (MAF> 0.01) from the UK Biobank, ENGINE recovered G×E variance that was on average 1.4-fold larger than that captured by a single exposure and 5.5-fold larger than that captured by the first principal component of the exposures.

Introduction

Gene–environment (G×E) interactions—wherein the magnitude and sometimes the direction of a variant's effect depends on environmental context—are increasingly recognized as important contributors to the genetic architecture of complex traits (Boye et al. 2024; Herrera-Luis et al. 2024; Namba et al. 2026; Virolainen et al. 2023; Westerman and Sofer 2024). Growing empirical evidence for context-dependent allelic effects across environments (Dahl et al. 2020; Weine et al. 2025) suggests that explicitly modeling G×E can increase explained variance, help account for components of missing heritability, and improve risk stratification in predictive models (Di Scipio et al. 2023; Li et al. 2019; Motsinger-Reif et al. 2024; Pazokitoroudi et al. 2024; Reay et al. 2020).

In practice, the operative "environment" is rarely a single exposure (Hamra and Buckley 2018; Moore et al. 2019; Patel et al. 2010; Sulc et al. 2020; Wild 2012). Large biobanks now measure a wide array of exposures that collectively constitute the exposome (All of Us Research Program Investigators 2019; Bycroft et al. 2018; Makris et al. 2025; Sudlow et al. 2015; Wild 2005). Within this exposome, lifestyle and behavioral factors, medication use, and sociodemographic context covary and can jointly modulate genetic effects via G×E. For instance, low-density lipoprotein (LDL) levels may reflect interactions between lipid-related loci and age, while statin use, diet, and physical activity further modify those genetic influences (Hartiala et al. 2021; Pazokitoroudi et al. 2024; Sadowski et al. 2024; Shungin et al. 2017). Modeling such multi-environment modulation requires methods that can accommodate many environmental variables, retain interpretability at the level of individual environments, and scale to biobank-sized cohorts.

Polygenic models quantify the genome-wide contribution of G×E by aggregating variant-level interaction effects, thereby offering higher power than single-variant approaches when individual effects are weak. In particular, variance component or mixed models are well-suited to model aggregate genetic signal, arising from additive genetic effects and from interactions such as those due to G×E, within a single framework (Dahl et al. 2020; Ni et al. 2019; Pazokitoroudi et al. 2020; Wu and Sankararaman 2018; Zhong et al. 2023). Early variance component approaches for modeling polygenic G×E (e.g., GCTA-G×E (Yang et al. 2011), MV-GREML (Robinson et al. 2017), MRNM (Ni et al. 2019)) enabled inference for interactions with individual environmental exposures. LEMMA (Kerin and Marchini 2020) models polygenic G×E by learning an environment score—a data-driven combination of exposures–that allows multiple environments to jointly modulate genetic effects. However, because LEMMA does not explicitly model environment-dependent residual variance, estimates of G×E variance can be biased upward under noise heterogeneity, leading to inflated false positive rates (Pazokitoroudi et al. 2024). MonsterLM (Di Scipio et al. 2023) scales to biobank-sized datasets and yields approximately unbiased estimates in many settings; however, its reliance on pruning and partitioning the genome into relatively weakly correlated blocks (and its focus on common variation) can limit power and scope for discovering G×E effects. GxEMM (Dahl et al. 2020) and GENIE (Pazokitoroudi et al. 2024) incorporate context-specific heteroskedasticity within a variance-components framework; GENIE is designed to scale to biobank-sized cohorts, whereas GxEMM becomes computationally infeasible for large samples. Despite these advances, approaches to learn which combinations of environmental exposures contribute to polygenic G×E on a biobank-scale remain limited.

Here we present ENGINE, *Efficient multi-eNvironmental Gene-environment Interaction iNference Estimator*, a method that jointly learns the combination of environments that interact with polygenic effects for a given trait, estimates the contribution of each environment to the interaction signal while producing calibrated variance-component estimates for additive effects, G×E, and noise by explicitly modeling environment-dependent heteroskedasticity. In this study, we establish the statistical foundations of ENGINE and characterize its performance through simulations spanning a range of genetic architectures and patterns of noise heteroskedasticity. We then apply ENGINE to a set of quantitative traits in the UK Biobank to dissect how multiple lifestyle environments jointly modulate polygenic effects.

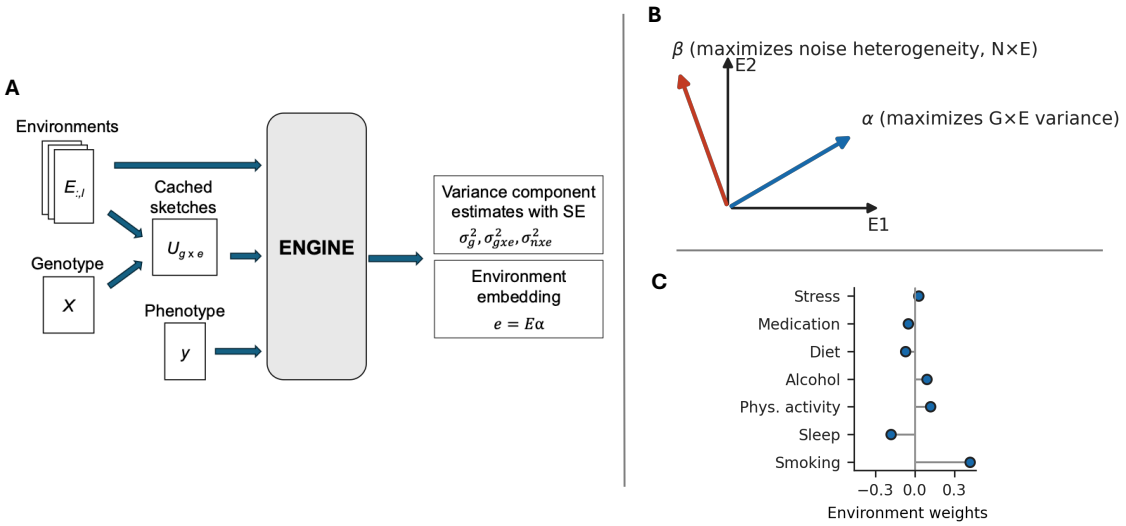

Figure 1: **Overview of ENGINE. A. Pipeline.** Inputs are a genotype matrix, multiple environments, and a phenotype. ENGINE learns an interpretable environmental embedding $\boldsymbol{e}$ and reports point estimates and standard errors for the additive (σ_g^2), interaction ($\sigma_{g\times e}^2$), and noise-by-environment ($\sigma_{n\times e}^2$) variance components. For efficiency, genotype-dependent summaries ($U_{g\times e}$) are cached in a single streaming pass and reused across iterations. **B. Separating heteroskedastic noise from G×E.** With two environments, the combination that maximizes N×E variance can differ from the one that maximizes G×E variance; ENGINE therefore learns weights that maximize G×E while controlling N×E. **C. Interpretable environment weights.** The entries of environmental weights ($\boldsymbol{\alpha}$) quantify the contribution of each environmental exposure to the embedding and to the captured G×E signal.

Methods

Estimation of variance components in G×E linear mixed model

Let $\boldsymbol{X} \in \mathbb{R}^{N\times M}$ denote standardized genotypes of M variants on N individuals, $\boldsymbol{y} \in \mathbb{R}^N$ a standardized phenotype, and $\boldsymbol{E} \in \mathbb{R}^{N\times L}$ a matrix of L measured environments (columns are centered and scaled). We compress the L environments into a one–dimensional vector $\boldsymbol{e} = \boldsymbol{E}\boldsymbol{\alpha}$, with interpretable, unit–norm weights $\boldsymbol{\alpha} \in \mathbb{R}^L$ that quantify each environment's contribution to the interaction signal. Intuitively, $\boldsymbol{e}$ collects heterogeneous environmental effects into a single direction along which genetic effects may be modulated.

We model the phenotype $\boldsymbol{y}$ with a linear mixed model comprising four components: an additive genetic effect, a G×E interaction that varies with the environment embedding $\boldsymbol{e}$, a noise-by-environment (N×E) term that captures heteroskedasticity aligned with $\boldsymbol{e}$, and a residual term (Dahl et al. 2020; Jeong et al. 2024; Pazokitoroudi et al. 2020; Wu and Sankararaman 2018):

$$\boldsymbol{y} \;=\; \boldsymbol{X}\boldsymbol{\beta} \;+\; (\boldsymbol{X} \odot \boldsymbol{e})\boldsymbol{\gamma} \;+\; (\boldsymbol{I} \odot \boldsymbol{e})\boldsymbol{\delta} \;+\; \boldsymbol{\epsilon}. \tag{1}$$

Here, $\odot$ denotes the Hadamard (elementwise) product with row-wise broadcasting across SNP columns, so that $\boldsymbol{X} \odot \boldsymbol{e} = \operatorname{diag}(\boldsymbol{e})\,\boldsymbol{X}$ and $\boldsymbol{I} \odot \boldsymbol{e} = \operatorname{diag}(\boldsymbol{e})$; equivalently, we scale each individual genotype and residual noise by their environmental score. The N×E component (context-specific noise) captures heteroskedastic noise across individuals exposed to different environments. Distinguishing this component from the G×E component is crucial for controlling false positives, and such environment-dependent heteroskedasticity is widely observed for complex traits (Pazokitoroudi et al. 2024).

We assume independent, mean-zero random effects with variances that define the components of interest:

$$\boldsymbol{\beta} \sim \mathcal{D}\Big(\mathbf{0}, \tfrac{\sigma_g^2}{M}\boldsymbol{I}_M\Big), \qquad \boldsymbol{\gamma} \sim \mathcal{D}\Big(\mathbf{0}, \tfrac{\sigma_{g\times e}^2}{M}\boldsymbol{I}_M\Big), \qquad \boldsymbol{\delta} \sim \mathcal{D}\big(\mathbf{0}, \sigma_{n\times e}^2\boldsymbol{I}_N\big), \qquad \boldsymbol{\epsilon} \sim \mathcal{D}\big(\mathbf{0}, \sigma_e^2\boldsymbol{I}_N\big),$$

where $\mathcal{D}$ can be any zero-mean distribution with the stated second moments (this moment assumption makes the method-of-moments robust to non-Gaussian effect distributions). The M-vectors $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ represent the

main additive effects and the G×E interaction effects along $\boldsymbol{e}$ respectively, while $\boldsymbol{\delta}$ captures environment-modulated noise (Table S1). Under $\mathbb{E}[\boldsymbol{y}] = \mathbf{0}$, the population covariance decomposes additively into kernels:

$$\mathrm{Cov}[\boldsymbol{y}] \;=\; \sigma_g^2\,\boldsymbol{K}_g \;+\; \sigma_{g\times e}^2\,\boldsymbol{K}_{g\times e}(\boldsymbol{\alpha}) \;+\; \sigma_{n\times e}^2\,\boldsymbol{K}_{n\times e}(\boldsymbol{\alpha}) \;+\; \sigma_e^2\,\boldsymbol{I}_N,$$

where

$$\boldsymbol{K}_g = \tfrac{1}{M}\boldsymbol{X}\boldsymbol{X}^\top, \qquad \boldsymbol{K}_{g\times e}(\boldsymbol{\alpha}) = \tfrac{1}{M}(\boldsymbol{X}\odot\boldsymbol{e})(\boldsymbol{X}\odot\boldsymbol{e})^\top, \qquad \boldsymbol{K}_{n\times e}(\boldsymbol{\alpha}) = \mathsf{diag}\,(\boldsymbol{e})^2\,.$$

Here $\boldsymbol{K}_g$ is the genetic relatedness matrix (GRM), $\boldsymbol{K}_{g\times e}$ is the GRM after scaling each individual's genotypes by e_i, thereby up-weighting pairs of individuals who are both environmentally exposed, and $\boldsymbol{K}_{n\times e}$ is diagonal with entries e_i^2, modeling environment-dependent variance inflation.

Given fixed $\boldsymbol{\alpha}$, we first estimate the variance components $\boldsymbol{\sigma}^2 = \{\sigma_g^2, \sigma_{g\times e}^2, \sigma_{n\times e}^2, \sigma_e^2\}$ using the method-of-moments (MoM), matching the model covariance to the empirical covariance $\boldsymbol{y}\boldsymbol{y}^\top$ via a Frobenius-norm objective:

$$\mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\sigma}^2) = \Big\| \boldsymbol{y}\boldsymbol{y}^\top - \sigma_g^2\boldsymbol{K}_g - \sigma_{g\times e}^2\boldsymbol{K}_{g\times e}(\boldsymbol{\alpha}) - \sigma_{n\times e}^2\boldsymbol{K}_{n\times e}(\boldsymbol{\alpha}) - \sigma_e^2\boldsymbol{I}_N \Big\|_F^2. \tag{2}$$

Because Eq. (2) is quadratic in $\boldsymbol{\sigma}^2$, the minimizer $\widehat{\boldsymbol{\sigma}}^2$ solves the linear system $\boldsymbol{A}\widehat{\boldsymbol{\sigma}}^2 = \boldsymbol{b}$, where $\boldsymbol{A} \in \mathbb{R}^{4\times 4}, \boldsymbol{b} \in \mathbb{R}^4$ are defined:

$$\boldsymbol{A}_{rs} = \mathrm{tr}\,[\boldsymbol{K}_r\boldsymbol{K}_s]\,, \quad \boldsymbol{b}_r = \boldsymbol{y}^\top\boldsymbol{K}_r\boldsymbol{y}, \qquad \text{where } \boldsymbol{K}_r, \boldsymbol{K}_s \in \{\boldsymbol{K}_g, \boldsymbol{K}_{g\times e}(\boldsymbol{\alpha}), \boldsymbol{K}_{n\times e}(\boldsymbol{\alpha}), \boldsymbol{I}_N\}$$

We attach asymptotic standard errors to $\widehat{\boldsymbol{\sigma}}^2 = \boldsymbol{A}^{-1}\boldsymbol{b}$ via the delta method:

$$\mathrm{Cov}\big(\widehat{\boldsymbol{\sigma}}^2\big) \;=\; \boldsymbol{A}^{-1}\,\mathrm{Cov}(\boldsymbol{b})\,\boldsymbol{A}^{-1}, \qquad \mathrm{Cov}(\boldsymbol{b})_{rs} \;=\; 2\,\mathrm{tr}\,[\boldsymbol{K}_r\,\boldsymbol{\Sigma}_y\,\boldsymbol{K}_s\,\boldsymbol{\Sigma}_y]\,,$$

with $\boldsymbol{\Sigma}_y = \mathrm{Cov}[\boldsymbol{y}]$. In practice, we plug in $\widehat{\boldsymbol{\Sigma}}_y = \sum_l \widehat{\sigma}_l^2\boldsymbol{K}_l$ and evaluate the trace contractions efficiently using probe sketches (see "Efficient learning via streaming and pre-computation"). Because $\boldsymbol{A}$ is small and reused, SE computation adds negligible overhead once the necessary contractions are cached.

Optimization of the environment embedding

Next, we alternate to optimize $\boldsymbol{\alpha}$ for fixed variance components $\boldsymbol{\sigma}^2$, such that the learned embedding $\boldsymbol{e} = \boldsymbol{E}\boldsymbol{\alpha}$ concentrates G×E signal while accounting for N×E heteroskedasticity. In particular, we select a single environmental embedding such that the induced kernel $\boldsymbol{K}_{g\times e}(\boldsymbol{\alpha})$ aligns with phenotype covariance while penalizing variance inflation captured by $\boldsymbol{K}_{n\times e}(\boldsymbol{\alpha})$. For notational simplicity, we write $\boldsymbol{K}_{g\times e} = \boldsymbol{K}_{g\times e}(\boldsymbol{\alpha}), \boldsymbol{K}_{n\times e} = \boldsymbol{K}_{n\times e}(\boldsymbol{\alpha})$. Based on the objective in Eq. (2), the terms depending on $\boldsymbol{\alpha}$ simplify to (see Notes S1):

$$\begin{aligned}\mathcal{L}(\boldsymbol{\alpha}) = \; & -2\sigma_{g\times e}^2\,\boldsymbol{y}^\top\boldsymbol{K}_{g\times e}\boldsymbol{y} + (\sigma_{g\times e}^2)^2\,\mathrm{tr}\,[\boldsymbol{K}_{g\times e}^2] + 2\sigma_g^2\sigma_{g\times e}^2\,\mathrm{tr}\,[\boldsymbol{K}_g\boldsymbol{K}_{g\times e}] - 2\sigma_{n\times e}^2\,\boldsymbol{y}^\top\boldsymbol{K}_{n\times e}\boldsymbol{y} \\ & + (\sigma_{n\times e}^2)^2\,\mathrm{tr}\,[\boldsymbol{K}_{n\times e}^2] + 2\sigma_g^2\sigma_{n\times e}^2\,\mathrm{tr}\,[\boldsymbol{K}_g\boldsymbol{K}_{n\times e}] + 2\sigma_{g\times e}^2\sigma_{n\times e}^2\,\mathrm{tr}\,[\boldsymbol{K}_{g\times e}\boldsymbol{K}_{n\times e}]\end{aligned} \tag{3}$$

The Euclidean gradient $\nabla_{\boldsymbol{\alpha}}\mathcal{L}(\boldsymbol{\alpha})$ is then derived as:

$$\begin{aligned}\nabla_{\boldsymbol{\alpha}}\mathcal{L}(\boldsymbol{\alpha}) = & -4\sigma_{g\times e}^2\boldsymbol{G}_y\boldsymbol{\alpha} + \boldsymbol{g}_{23} - 4\sigma_{n\times e}^2\boldsymbol{E}^\top\big((\boldsymbol{y}\odot\boldsymbol{y})\odot\boldsymbol{e}\big) \\ & + 4\sigma_{n\times e}^4\boldsymbol{E}^\top\big(\boldsymbol{e}^{\odot 3}\big) + 4\sigma_g^2\sigma_{n\times e}^2\boldsymbol{E}^\top\big(\boldsymbol{s}_g\odot\boldsymbol{e}\big) + 8\sigma_{g\times e}^2\sigma_{n\times e}^2\boldsymbol{E}^\top\big(\boldsymbol{s}_g\odot\boldsymbol{e}^{\odot 3}\big),\end{aligned}$$

where

$$\boldsymbol{G}_y = \frac{1}{M}\big(\boldsymbol{X}^\top(\boldsymbol{E}\odot\boldsymbol{y})\big)^\top\big(\boldsymbol{X}^\top(\boldsymbol{E}\odot\boldsymbol{y})\big), \qquad \boldsymbol{G}_E = \frac{1}{M}\big(\boldsymbol{X}^\top\boldsymbol{E}\big)^\top\big(\boldsymbol{X}^\top\boldsymbol{E}\big), \qquad \boldsymbol{s}_g = \mathsf{diag}\,(\boldsymbol{K}_g)\,,$$

and $\boldsymbol{e}^{\odot 3}$ is the element-wise cube of $\boldsymbol{e}$ (Table S1). The vector $\boldsymbol{g}_{23}$ aggregates the trace-gradient contributions from the second and third terms; see "Efficient learning via streaming and pre-computation" for efficient evaluation.

Note that there is a scale non-identifiability between $\boldsymbol{\alpha}$ and $\sigma^2_{g\times e}$: scaling $\boldsymbol{\alpha}$ rescales $\boldsymbol{K}_{g\times e}$ and can be offset by $\sigma^2_{g\times e}$. We therefore fix $\|\boldsymbol{\alpha}\|_2 = 1$ and optimize on the unit sphere $\mathbb{S}^{L-1}$. Defining $\boldsymbol{g} = \frac{1}{N}\nabla_{\boldsymbol{\alpha}}\mathcal{L}(\boldsymbol{\alpha})$, we project to the tangent space at $\boldsymbol{\alpha}$ via $\boldsymbol{g}_\top = \boldsymbol{g} - (\boldsymbol{\alpha}^\top \boldsymbol{g})\,\boldsymbol{\alpha}$ and take a capped geodesic step:

$$\theta \;=\; \min\{\eta\,\|\boldsymbol{g}_\top\|_2,\; \theta_{\max}\}, \qquad \boldsymbol{\alpha}_{\text{new}} \;=\; \boldsymbol{\alpha}\cos\theta \;-\; \frac{\boldsymbol{g}_\top}{\|\boldsymbol{g}_\top\|_2}\sin\theta,$$

where $\theta_{\max}$ is the maximum allowed step size. This exactly preserves unit norm and caps the per-iteration rotation, stabilizing progress when gradients are large. In practice, our proposed scheme of alternating (i) solving the 4×4 MoM system for $\widehat{\boldsymbol{\sigma}}^2$ at the current $\boldsymbol{\alpha}$ and (ii) taking one or a few geodesic steps on $\boldsymbol{\alpha}$ converges quickly in our experiments.

Efficient learning via streaming and pre-computation

The primary computational challenge is that both solving for $\widehat{\boldsymbol{\sigma}}^2$ and $\boldsymbol{\alpha}$-gradient require repeated contractions with genotype-derived kernels. Naïvely rebuilding genotype-based quantities at each iteration is prohibitively expensive at biobank scale. Instead, we perform a single streamed pass over the genotype matrix X to cache probe sketches that suffice for all subsequent evaluations, thereby decoupling per-iteration cost from the number of SNPs M.

We employ Hutchinson's trace estimator (Hutchinson 1989), using $B \ll M$ independent standard-normal probe vectors collected as the rows of $\boldsymbol{W} \in \mathbb{R}^{B\times N}$. Rather than forming the kernel matrices, we multiply each target matrix by the random probes to obtain compact sketches. Traces are approximated by the average Frobenius inner product of the corresponding sketches; accordingly, we compute kernel sketches $\boldsymbol{U}_r = \boldsymbol{K}_r\boldsymbol{W}^\top$ so that the pairwise kernel trace $\operatorname{tr}[\boldsymbol{K}_r\boldsymbol{K}_s]$ follows from $\frac{1}{B}\langle \boldsymbol{U}_r, \boldsymbol{U}_s\rangle_F$. For the additive kernel, $\boldsymbol{U}_g$ is fixed and computed once. Meanwhile, the context-specific kernel $\boldsymbol{K}_{g\times e}$ depends on $\boldsymbol{\alpha}$ only through $\boldsymbol{e} = \boldsymbol{E}\boldsymbol{\alpha}$ and admits a linear decomposition over environments. Let $\boldsymbol{Z}_\ell = \boldsymbol{X} \odot \boldsymbol{E}_{:\ell}$ be the per-environment modulated genotype matrix. During the single streamed pass over SNP blocks, we compute and cache pairwise probe sketches

$$\boldsymbol{U}_{g\times e}[\ell,k] = \begin{cases} \dfrac{1}{M}\big(\boldsymbol{Z}_\ell\boldsymbol{Z}_k^\top + \boldsymbol{Z}_k\boldsymbol{Z}_\ell^\top\big)\boldsymbol{W}^\top, & \ell < k, \\ \dfrac{1}{M}\boldsymbol{Z}_\ell\boldsymbol{Z}_\ell^\top\boldsymbol{W}^\top, & \ell = k, \end{cases}$$

so that for any $\boldsymbol{\alpha}$ we can assemble

$$\boldsymbol{U}_{g\times e}(\boldsymbol{\alpha}) = \sum_{\ell\le k} \alpha_\ell\alpha_k\,\boldsymbol{U}_{g\times e}[\ell,k].$$

Then, both trace and cross-trace terms involving $K_{g\times e}$ may be computed as:

$$\operatorname{tr}\big[\boldsymbol{K}^2_{g\times e}\big] \;\approx\; \tfrac{1}{B}\big\|\boldsymbol{U}_{g\times e}(\boldsymbol{\alpha})\big\|_F^2, \qquad \operatorname{tr}\big[\boldsymbol{K}_g\boldsymbol{K}_{g\times e}\big] \;\approx\; \tfrac{1}{B}\big\langle \boldsymbol{U}_g,\, \boldsymbol{U}_{g\times e}(\boldsymbol{\alpha})\big\rangle_F.$$

To obtain the trace-gradient $\boldsymbol{g}_{23}$ without forming a $L\times L$ matrix, we first compute scalar contractions

$$s^{(1)}_{\ell k} = \tfrac{1}{B}\big\langle \boldsymbol{U}_{g\times e}(\boldsymbol{\alpha}),\, \boldsymbol{U}_{g\times e}[\ell,k]\big\rangle_F, \qquad s^{(2)}_{\ell k} = \tfrac{1}{B}\big\langle \boldsymbol{U}_g,\, \boldsymbol{U}_{g\times e}[\ell,k]\big\rangle_F$$

Next, we define the upper-triangular matrix $\boldsymbol{C} \in \mathbb{R}^{L\times L}$ as:

$$\boldsymbol{C}_{lk} = \begin{cases} 2\,\sigma^4_{g\times e}\,s^{(1)}_{\ell k} \;+\; 2\,\sigma^2_g\sigma^2_{g\times e}\,s^{(2)}_{\ell k} & ,\ \ell \le k \\ 0 & ,\ l > k \end{cases}$$

Then, we derive $\boldsymbol{g}_{23} = (\boldsymbol{C}+\boldsymbol{C}^\top)\boldsymbol{\alpha}$. In practice, we realize $\boldsymbol{g}_{23}$ via two indexed scatter-adds over the $P = L(L+1)/2$ pairs (Duff et al. 2002; Gustavson 1978; Turner et al. 1956), avoiding materializing $\boldsymbol{C}$ while injecting $\boldsymbol{\alpha}$ into all trace terms at a cost that is independent of M.

Finally, to bound memory, we stream the genotype matrix in SNP blocks, updating all sketches and quadratic forms in-place and discarding block-level temporaries. After this single pass, subsequent iterations

reuse the cached quantities: each MoM solves (the 4×4 system) and each $\boldsymbol{\alpha}$-gradient evaluation runs in time independent of M, depending only on the number of individuals N, the number of environments L, and the number of random probes B. Therefore, streaming and pre-computation provide biobank-scale efficiency without sacrificing the statistical structure of the model.

Regularized cross-fitting for G×E

As outlined in Algorithm 1, our procedure alternates between estimating the variance components at the current environment embedding and updating the environment weights to decrease the loss in Eq. (3). While computationally convenient, reusing the same sample for both steps couples learning and evaluation and can introduce target leakage, with the environment weights adapting to noise in the phenotype vector. In the presence of a nonzero genetic main effect σ_g^2 or environment-specific noise $\sigma_{n\times e}^2$, the optimizer may synthesize artificial G×E by borrowing signal from the main effect or from heteroskedasticity. To address overfitting, we introduce variant-level cross-fitting and an ℓ_1-regularized update on the sphere to break this coupling and stabilize the embedding.

To decouple learning and evaluation, we use cross-fitting over variants. We split M SNPs into two disjoint halves (e.g., by chromosome or LD blocks, balancing MAF/LD), obtaining matrices $\boldsymbol{X}_{(1)}, \boldsymbol{X}_{(2)} \in \mathbb{R}^{N\times M/2}$. Using only $\boldsymbol{X}_{(1)}$, we make a single streamed pass to build the required sketches and optimize a unit-norm embedding $\boldsymbol{\alpha}_{(1)}$. Freezing this embedding, we then use only $\boldsymbol{X}_{(2)}$ to assemble $\boldsymbol{A}_{(2)}$ and $\boldsymbol{b}_{(2)}$, solve for $\widehat{\boldsymbol{\sigma}}^2_{(2)}$, and compute standard errors. We then swap roles: learn $\boldsymbol{\alpha}_{(2)}$ on $\boldsymbol{X}_{(2)}$, and evaluate $\widehat{\boldsymbol{\sigma}}^2_{(1)}$ on $\boldsymbol{X}_{(1)}$. This separation prevents $\boldsymbol{\alpha}$ from capturing noise in the same variants used for component estimation. We derive the final embedding as the average of the two held-out solutions followed by renormalization to unit norm, and evaluate variance components on the full dataset to capture variance explained by all SNPs.

Cross-fitting removes the main leakage path, but dense embeddings can still overfit by combining many weak, correlated environments. To constrain model complexity and improve interpretability, we add ℓ_1 regularization on the sphere, solving

$$\min_{\boldsymbol{\alpha}} \ \mathcal{L}(\boldsymbol{\alpha}) \ + \ \lambda\|\boldsymbol{\alpha}\|_1 \quad \text{subject to} \quad \|\boldsymbol{\alpha}\|_2 = 1.$$

The unit-norm constraint fixes the scale ambiguity with $\sigma_{g\times e}^2$, and the ℓ_1 penalty encourages a sparse set of nonzero weights. This reduces effective degrees of freedom, improves robustness under collinearity, and concentrates the embedding on environments with strongest support.

We implement the update in each coordinate via soft-thresholding followed by renormalization,

$$\tilde{\boldsymbol{\alpha}}_i \leftarrow \operatorname{sign}(\boldsymbol{\alpha}_i)\max\big(|\boldsymbol{\alpha}_i| - \tau\lambda,\, 0\big), \qquad \boldsymbol{\alpha} \leftarrow \tilde{\boldsymbol{\alpha}}/\|\tilde{\boldsymbol{\alpha}}\|_2,$$

where τ is the step length along the geodesic move. This "soft-threshold on the sphere" integrates directly into the existing update with negligible overhead. Because a fixed, strong penalty can bias the solution, we apply light annealing: start with large $\lambda = \lambda_0$ to obtain a stable coarse embedding, then decay λ toward a lower bound $\lambda_{\min}$ across outer alternations.

Together, these changes control overfitting. Cross-fitting separates learning $\boldsymbol{\alpha}$ from evaluating $\boldsymbol{\sigma}^2$, blocking gains that arise from relabeling noise as signal. The ℓ_1 term reduces the tendency to inflate $\|\boldsymbol{K}_{g\times e}\|_F$ or mimic $\boldsymbol{K}_g$ without improving fit, and annealing allows the procedure to begin conservatively and refine as evidence accumulates. As a result, the embedding is more stable across resamples and interaction estimates are better calibrated under the null without sacrificing ENGINE's one-pass, streamed MoM efficiency.

Results

Calibration and power

We benchmarked ENGINE against LEMMA, the most directly comparable existing method. Among related variance-component methods, GxEMM does not scale to biobank-scale datasets and GENIE does not have an interpretable embedding. LEMMA, by contrast, is designed for multi-environment G×E inference at biobank scale and learns a one-dimensional environment embedding while also estimating the associated variance

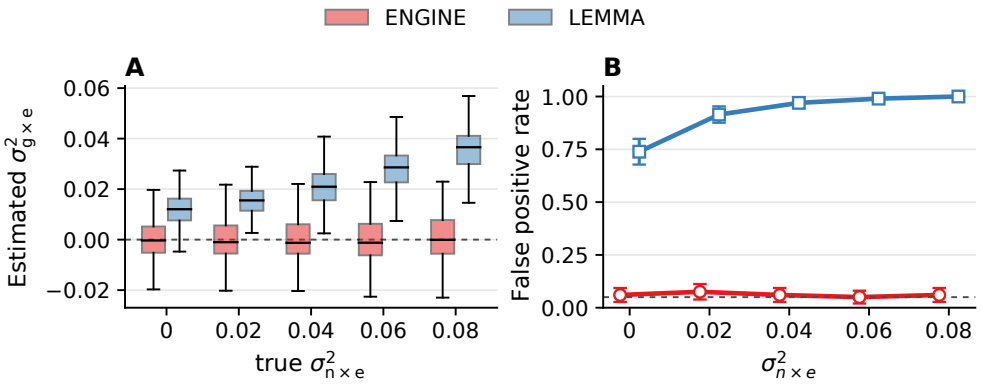

Figure 2: **Calibration under the null across varying levels of heteroskedastic noise.** We simulated phenotypes from $N = 10{,}000$ individuals and $M = 10{,}000$ SNPs with five lifestyle environments under the null of no G×E variance ($\sigma^2_{g\times e} = 0$), varying noise-by-environment variance $\sigma^2_{n\times e} \in \{0, 0.02, 0.04, 0.06, 0.08\}$ across 200 replicates. **A. Estimated $\sigma^2_{g\times e}$.** Boxplots across 200 replicates; ENGINE is marked in red and LEMMA in blue. **B. Type I error.** False-positive rate at significance level 0.05 with 95% binomial confidence intervals.

components. We first assessed calibration under the null of no G×E. We simulated phenotypes for $N = 10{,}000$ unrelated UK Biobank (UKB) participants of White British ancestry using $M = 10{,}000$ SNPs genotyped on the UKB array (see Notes S4). We used five lifestyle environments (Townsend deprivation index, sleep duration, age, smoking status, alcohol frequency), mixing categorical and quantitative measures to assess robustness to different value types. Phenotypes followed Model 1 with $\sigma^2_g = 0.1$ and an environmental score $\boldsymbol{E\alpha}$, where $\boldsymbol{\alpha} \in \mathbb{R}^5$ was simulated from a standard normal distribution. We varied the heteroskedastic noise $\sigma^2_{n\times e} \in \{0, 0.02, 0.04, 0.06, 0.08\}$ while fixing $\sigma^2_{g\times e} = 0$, and ran 200 replicates per setting, comparing ENGINE with LEMMA. When $\sigma^2_{n\times e} = 0$, ENGINE produced $\widehat{\sigma}^2_{g\times e}$ centered at zero, whereas LEMMA exhibited a slight positive bias (mean ≈ 0.010 across replicates; Figure 2A). One plausible source is data reuse: the environmental score is learned and G×E heritability is estimated on the same data, inducing overfitting (also observed for ENGINE without regularization; see "Ablation studies"). As $\sigma^2_{n\times e}$ increased, ENGINE remained unbiased (overall mean across all settings -8.8×10^{-6}), while the bias associated with LEMMA increased with heteroskedasticity. Type-I error was well calibrated for ENGINE: at significance threshold $p < 0.05$, the observed rejection proportions lay within their 95% binomial intervals at all $\sigma^2_{n\times e}$ levels. In contrast, LEMMA showed inflated rejection rates both with and without heteroskedastic noise (Figure 2B).

We next evaluated the statistical power of ENGINE when G×E is present. We fixed $\sigma^2_g = 0.3$, varied $\sigma^2_{g\times e} \in \{0.05, 0.10, 0.15, 0.20, 0.25\}$, and set $\sigma^2_{n\times e} = \sigma^2_{g\times e}$. ENGINE yielded accurate variance-component estimates: $\widehat{\sigma}^2_g$ was centered on the truth across all settings, and $\widehat{\sigma}^2_{g\times e}$ tracked the simulated values closely as $\sigma^2_{g\times e}$ increased (Figure 3A). Power, calculated as the proportion of $p < 0.05$ rejections, approached 1 by $\sigma^2_{g\times e} \approx 0.10$ (Figure 3B). We then assessed recovery of the environment embedding using a sign-invariant metric: the squared correlation r^2 between the predicted and true embeddings, which can be interpreted as the fraction of variation in the true embedding explained by the learned embedding. ENGINE achieved $r^2 > 0.9$ when $\sigma^2_{g\times e} \geq 0.15$ (Fig. 3C). We also confirmed in simulations that the number of random probe vectors B does not affect the accuracy of ENGINE in prediction (Figure S1). We used $B = 100$ by default in all experiments. Together, these results indicate that ENGINE is well calibrated under the null, powerful when interactions are present, and accurately recovers both variance components and the environment embedding.

Efficiency and scaling

We benchmarked ENGINE against LEMMA and the exact-per-update ENGINE baseline, which recomputes the exact traces and rebuilds the normal equations at every embedding update. Recall that the efficient

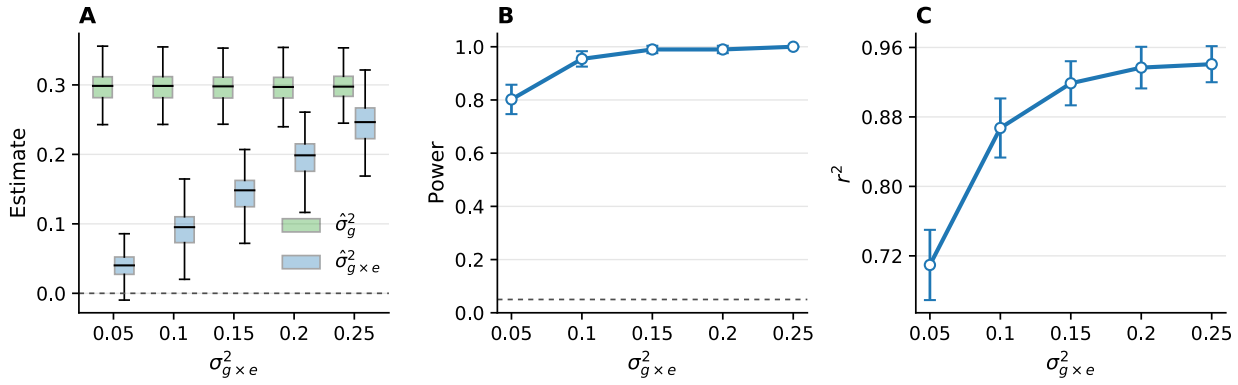

Figure 3: **Power and estimation accuracy as a function of interaction strength.** We simulated phenotypes from $N = 10{,}000$ individuals and $M = 10{,}000$ SNPs with five lifestyle environments, setting $\sigma^2_g = 0.3$, $\sigma^2_{g\times e} \in \{0.05, 0.10, 0.15, 0.20, 0.25\}$, and $\sigma^2_{n\times e} = \sigma^2_{g\times e}$ (200 replicates per setting). **A. Variance estimates.** Boxplots of $\widehat{\sigma}^2_g$ are in green and $\widehat{\sigma}^2_{g\times e}$ in blue across replicates; the horizontal dashed line marks zero. **B. Power.** Rejection rate at significance level 0.05 with 95% binomial confidence intervals. **C. Embedding accuracy.** Squared correlation between the learned environmental embedding and the ground truth.

implementation of ENGINE makes a single streaming $\mathcal{O}(NM)$ pass over the genotype matrix to cache sketches of GRMs; subsequent embedding updates reuse these caches without scanning the raw genotypes. For the experiments below, we used five lifestyle environments by default. With SNPs fixed at $M = 10{,}000$, number of environmental exposures at $L = 5$ and individuals varied $N \in \{1, 3, 5, 10, 30, 50, 100\}$ k, ENGINE is highly scalable: while LEMMA is slightly faster for smaller sample sizes when the number of individuals is less than $10,000$, its runtime grows more steeply so that for sample sizes of $N = 100,000$ its runtime is $\sim 5\times$ that of ENGINE (Figure 4A). The exact-per-update ENGINE baseline is computationally prohibitive, taking ~ 900 s at $N = 1,000$ ($\sim 45\times$ slower than ENGINE). In a second experiment, with individuals fixed at $N = 10{,}000$ and SNPs varied $M \in \{1, 3, 5, 10, 30, 50, 100\}$ k, ENGINE is marginally slower at the smallest M but becomes faster once the number of SNPs reaches $10,000$; the exact-per-update ENGINE baseline remains infeasible as its per-iteration cost grows with M (Figure 4B). In an additional experiment with both individuals and SNPs fixed at $N = M = 10,000$, ENGINE exhibits similar runtime scaling in the number of environments as LEMMA (Figure S2). We further report that ENGINE fits the full UK Biobank array dataset ($M = 454{,}207$ variants; $N = 291{,}273$ individuals; $L = 5$ environments) in approximately seven hours on a single CPU core.

Ablation studies

We evaluated how cross-fitting (A/B holdout) and ℓ_1 penalization of the environment embedding affect false-positive control under the no G×E calibration scenarios, varying heterogeneous noise $\sigma^2_{n\times e}$ and running 100 replicates per setting at significance level 0.05; unless noted, the regularization parameter was $\lambda = 5 \times 10^{-3}$. Without either safeguard, the procedure was severely inflated. At $\sigma^2_{n\times e} = 0$, the false positive rate (FPR) was 0.48. As noise heterogeneity increased, inflation lessened but remained substantial (FPR = 0.22 at $\sigma^2_{n\times e} = 0.08$). Each safeguard alone reduced inflation: cross-fitting was most effective when $\sigma^2_{n\times e}$ was larger, with FPR decreasing from 0.11 to 0.06 as $\sigma^2_{n\times e}$ rose from 0.02 to 0.08, whereas ℓ_1 regularization helped most when noise heterogeneity was absent or small (FPR ≈ 0.09 at $\sigma^2_{n\times e} = 0.02$) and was slightly less protective as $\sigma^2_{n\times e}$ grew (rising to ≈ 0.11 at 0.08). Used together, the two mechanisms restored near-nominal behavior, with FPR ≈ 0.05 at $\sigma^2_{n\times e} = 0$ and mean FPR across settings 0.06 (95% binomial CI [0.01, 0.11]) (Figure 5). These results indicate that cross-fitting and ℓ_1 regularization together enable calibrated false positive rates across the range of tested noise heteroskedasticity levels.

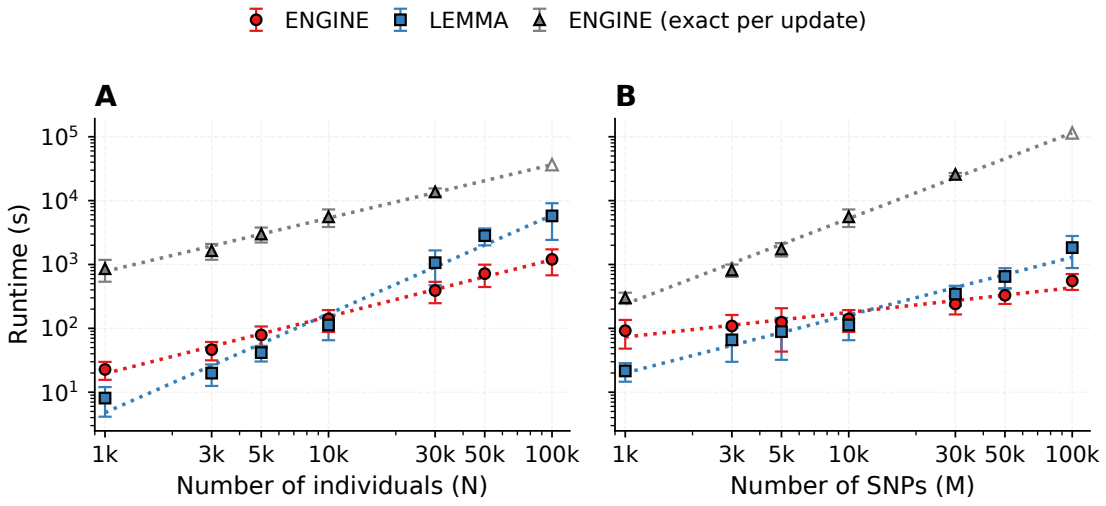

Figure 4: **Runtime benchmarks against alternative approaches.** Points show mean wall-clock runtime over 200 replicates; error bars denote one standard deviation. Dotted curves are power-law fits over the observed range, and open markers at 100 k indicate values obtained by extrapolating these fits. Both axes are on a logarithmic scale. **A. Scaling with sample size.** The number of SNPs is fixed at $M = 10{,}000$ and we vary the number of individuals $N \in \{1, 3, 5, 10, 30, 50, 100\}$ k. The runtime for the exact-per-update ENGINE baseline at $N = 100$ k is extrapolated because the method did not complete within the 8-hour wall-clock limit under 30 GB of memory. **B. Scaling with SNP count.** The number of individuals is fixed at $N = 10{,}000$ and we vary the number of SNPs $M \in \{1, 3, 5, 10, 30, 50, 100\}$ k. The runtime for the exact-per-update ENGINE baseline at $M = 100$ k is likewise extrapolated. All benchmarks were run on a single core of an Intel Xeon Gold 6140 CPU (36 cores, 2.30 GHz) with a maximum of 30 GB memory and an 8-hour wall-clock limit per replicate.

G×E associated with lifestyle factors in the UK Biobank

Lifestyle context can modulate the genetic architecture of complex traits, with prior studies reporting interactions across socioeconomic status, physical activity, sleep, smoking, alcohol intake, and diet for body mass index (BMI) (Ding et al. 2018; Justice et al. 2017; Kilpeläinen et al. 2011; Qi et al. 2012; Robinson et al. 2015; Tyrrell et al. 2017; Wang et al. 2018; Watson et al. 2012). From 42 derived environments including dietary and non-dietary measures and their interactions with sex and age (Notes S3; Table S2), we pre-screened and retained ten that each showed significant $\sigma^2_{g\times e}$ with BMI in UK Biobank, after adjusting for sex, age, and the top 20 genetic principal components. Applying ENGINE to these ten environments uncovered a substantial, interpretable lifestyle G×E component (Table S3). Using common array-genotyped SNPs with MAF > 0.01 ($M = 454{,}207$), we trained the environmental embedding on $N = 50{,}000$ individuals and then estimated variance components on the full set of $N = 291{,}273$ unrelated White British participants. The learned embedding yielded $\sigma^2_{g\times e} = 2.34 \times 10^{-2}$ (SE: 4.2×10^{-3}), exceeding the interaction variance attributable to any single exposure (Figure 6A).

We then compared the learned embedding to unsupervised baselines constructed as the top principal component (PC) from (i) the selected ten exposures and (ii) the full set of 42 derived environments. ENGINE produced both a larger $\sigma^2_{g\times e}$ and stronger statistical evidence than either PC baseline: the estimate from the learned embedding was around 6× that of the top PC from the ten exposures ($\sigma^2_{g\times e} = 3.88 \times 10^{-3}$) and substantially larger than the top PC from all 42 environments ($\sigma^2_{g\times e} = 1.35 \times 10^{-3}$; Figure 7A). A paired comparison over leave-one-out jackknife blocks (100 blocks) confirmed that blockwise estimates of $\sigma^2_{g\times e}$ were consistently higher for the learned embedding than for the top PC of the ten exposures (one-sided $p = 6.62 \times 10^{-15}$; see Notes S2). Targeting the interaction structure directly thus yields measurable gains over unsupervised projections that need not align with the G×E signal. Finally, the learned embedding also exceeded the strongest single-environment signal, with $\sigma^2_{g\times e}$ significantly larger than the maximum across

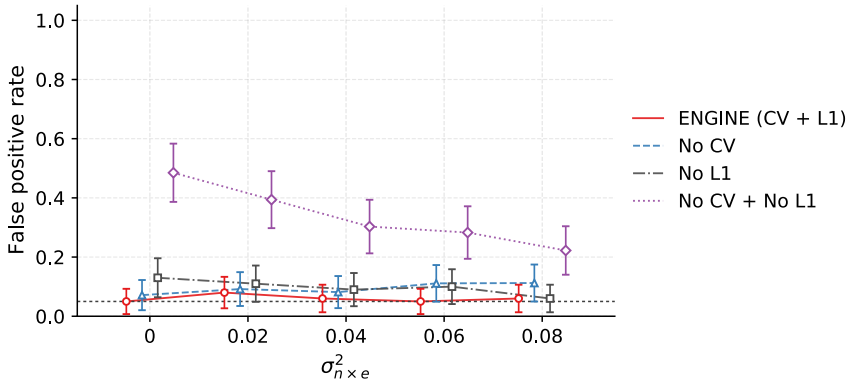

Figure 5: **Ablation of cross-fitting (CV) and ℓ_1 regularization on false-positive control.** False positive rate under no true G×E as a function of heterogeneous noise $\sigma^2_{n\times e}$. Curves compare ENGINE with both safeguards (CV+ℓ_1), without CV, without ℓ_1, and with neither. The false positive rate is calculated as P(rejection at $p < 0.05$) across 100 replicates; error bars show 95% binomial CIs; the dashed line marks the significance level 0.05 (ℓ_1 penalty $\lambda = 5 \times 10^{-3}$).

individual environments (jackknife $p = 2.67 \times 10^{-6}$).

The learned weights $\widehat{\boldsymbol{\alpha}}$ provide additional interpretability. Smoking-related measures (and their interactions) received the largest magnitudes, while nontrivial weight was also assigned to other features (e.g., time spent watching TV and TDI-related measures), supporting a diffuse architecture in which many exposures contribute modestly to the overall G×E component (Figure 6B). This pattern aligns with prior evidence that lifestyle factors collectively shape BMI's genetic effects (Kerin and Marchini 2020; Moore et al. 2019).

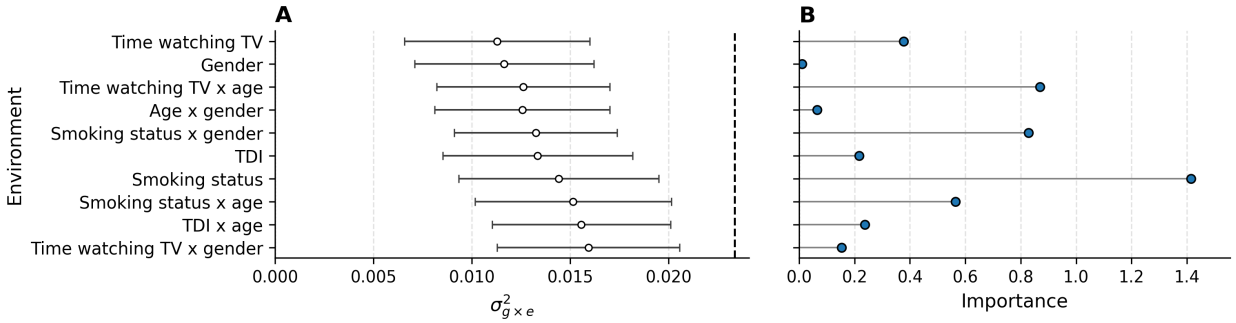

Figure 6: **Lifestyle G×E for BMI: single environments vs. learned embedding. A. Single-exposure estimates.** Estimates of $\sigma^2_{g\times e}$ for each selected environmental exposure, with error bars indicating the estimated standard errors; the dashed vertical line marks the learned embedding estimate ($\sigma^2_{g\times e} = 2.34 \times 10^{-2}$). **B. Learned embedding weights.** The magnitude of learned embedding weights for the same features.

In addition to BMI, we extended our lifestyle-exposure analysis to four traits—basal metabolic rate, standing height, LDL cholesterol (LDL-C) direct, and gamma glutamyltransferase (GGT). For all five traits, after residualizing each trait on fixed-effect covariates (basal metabolic rate additionally adjusted for BMI), the learned embedding yielded a significantly larger estimate of $\sigma^2_{g\times e}$ than the top PC of the same environments (jackknife $p < 0.05/(42 \times 5)$, Bonferroni-corrected across exposures and traits). To quantify the magnitude of G×E relative to additive effects, we computed $r_{g\times e} = \sigma^2_{g\times e}/\sigma^2_g$ on the full dataset; across traits, this ratio ranged from 2.47% for height to 10.25% for LDL-C levels (Figure 7B). Within the context of lifestyle G×E, these results indicate that height has a largely additive genetic architecture, whereas lifestyle exposures substantially modulate the genetic architecture of LDL-C levels (Jelenkovic et al. 2016; Laville et al. 2022; Pazokitoroudi et al. 2024; Robinson et al. 2017; Yengo et al. 2022). Comparisons with the single

environment that attains the largest $\sigma^2_{g\times e}$ estimate show that the learned embedding captures additional G×E signal for height and GGT, while marginally for LDL-C direct and basal metabolic rate (Figure 7A; Figure S3). These trait-specific differences indicate that not all traits exhibit substantial additional G×E beyond the contribution of the single most interacting environment.

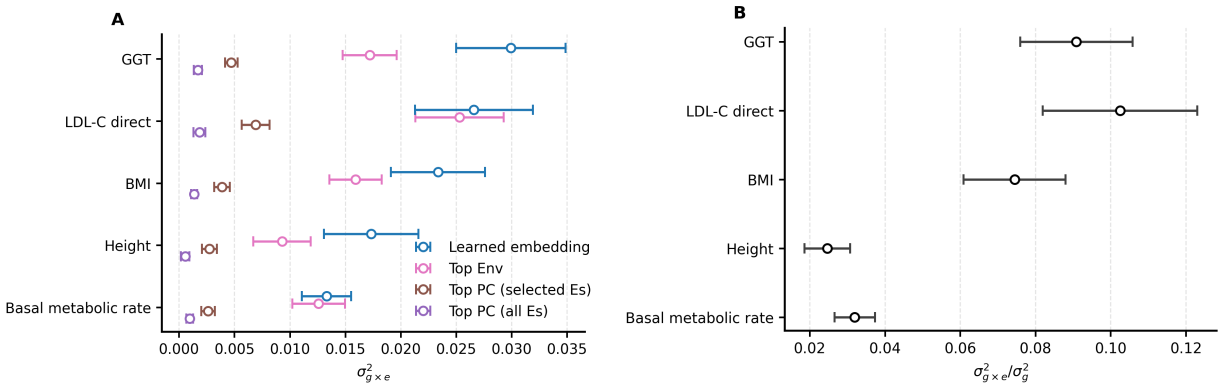

Figure 7: **Lifestyle G×E across traits. A. Learned embedding vs. baselines.** Estimates of $\sigma^2_{g\times e}$ across traits (basal metabolic rate, height, BMI, LDL-C direct, GGT) for four approaches: the learned environmental embedding (blue), the best single environment ("Top Env", pink), the top principal component computed from the selected environments (brown), and the top principal component from all environments (purple). Points show estimates; horizontal bars denote the standard errors. **B. G×E proportion of additive variance.** Estimates of the ratio $r_{g\times e} = \sigma^2_{g\times e}/\sigma^2_g$ for the learned embedding across traits, with error bars indicating the estimated standard errors.

Discussion

We present ENGINE, a scalable multi-environment G×E estimator that learns a one-dimensional environmental embedding to concentrate interaction signal while explicitly modeling context-dependent noise. The central computational idea is to precompute genotype-dependent summaries and cache them, so each iteration can form gradients and update variance components without repeated passes over the full genotype matrix. This design makes training practical at biobank scale. Methodologically, cross-fitting and ℓ_1 regularization reduce leakage from additive effects and noise-by-environment components, improving calibration while retaining strong power under heterogeneous noise.

The separation of gene-environment interaction from context-specific noise rests on several assumptions. The effects associated with the variance component are mutually independent. We also require that the covariance kernels corresponding to the GxE and NxE components must be sufficiently distinguishable, which usually holds when there is variation in genetic relationship across individuals.

Interpreting the learned embedding requires care. Because the interaction term is bilinear, optimization over the environment weights is nonconvex and may converge to different local optima depending on initialization. In addition, two edge cases warrant caution. First, although our simulations show that ENGINE remains robust to highly correlated environments (Figure S5), extreme near-collinearity among environments can make the embedding weights difficult to identify and remains an important direction for future work. Second, when the underlying genetic effects are small and close to zero, the signal driving the learned embedding is weak, and the embedding can fit noise and may fail to converge. Nevertheless, we found that ENGINE yielded generally consistent embeddings and variance-component estimates across random seeds (Figure S4). We also confirmed that the uncertainty in the learned embedding is negligible relative to variance component sampling error (Notes S5; Figure S6). In practice, stability can be improved by imposing mild structure—for example, constraining weights to be nonnegative after preprocessing and using multiple random starts to report a consensus embedding when feasible. We also note that the embedding is trained in the European-ancestry cohort; transfer to other populations depends on the degree to which causal vari-

ants and G×E architectures are shared across ancestries. Finally, we estimate variance components via the method of moments. While potentially less statistically efficient than restricted maximum likelihood (REML) under particular assumptions (Patterson and Thompson 1971), the method-of-moments approach is substantially faster and integrates naturally with cached summaries (Yue et al. 2021; Zhou 2017).

Several directions could broaden applicability and robustness. First, cache size currently scales as $O(L^2)$ in the number of exposures L, which may be memory-limiting for hundreds to thousands of environments. A practical remedy is to compress storage using anchored sketches—for example, maintaining a single "anchor" environment sketch and storing per-exposure projections or scaling factors relative to this anchor—thereby preserving fast iterations while reducing memory. Second, correlated exposures remain challenging: we currently standardize and whiten inputs for stability, but learning directly on correlated features would improve usability and interpretability. Covariance-aware parameterizations or structured regularization that explicitly models correlation could replace preprocessing while preserving identifiability. Third, ENGINE currently learns linear combinations of exposures; incorporating nonlinearity could capture higher-order environmental effects without sacrificing scalability. A polynomial kernel over the environments, for instance, would let the learned embedding weights index the importance of polynomial features. Learning an arbitrary non-linear function directly is less compatible with cached sketches and remains to be explored. Also, our framework learns a one-dimensional environment embedding, which is well suited to settings where the dominant GxE signal lies along a single direction in environment space, such as a cluster of lifestyle exposures. When interaction effects instead span multiple orthogonal axes of exposure, a one-dimensional projection may attenuate signal, and extending the embedding to K dimensions is a natural direction for future work. Beyond variance decomposition, the learned embedding can be fixed and reused in downstream analyses to estimate stabilized SNP-by-embedding interaction effects, facilitating locus prioritization and biological interpretation. An additional theoretical direction is to establish convergence guarantees for sketch-based updates and error bounds that quantify approximation accuracy as a function of the number of random probe vectors.

Overall, ENGINE provides an efficient and well-calibrated framework for discovering environmental contexts that modulate genetic effects on complex traits, bridging statistical interpretability with biobank-scale computation.

Code availability

The ENGINE software is publicly available at https://github.com/sriramlab/ENGINE. A frozen snapshot of the version used in this study is archived at Zenodo https://doi.org/10.5281/zenodo.20618212.

Acknowledgements

This research was conducted using the UK Biobank Resource under application 33127. We thank the participants of UK Biobank for making this work possible. This work was funded, in part, by NIH grants R35GM153406 (Z.L., M.J., A.A., and S.S.), 1R21HG013393 (A.R.) and NSF grant CAREER-1943497 (S.S.).

Author contributions: S.S. conceived and supervised the project. Z.L. and A.R. developed the method. Z.L. and S.S. wrote the manuscript. Z.L, A.R., and A.A. wrote the software code and performed the analyses. Z.L., A.G., and M.J. interpreted the findings. All authors read, reviewed, and approved the final manuscript.

Algorithm 1 Alternating optimization for learning the environment embedding

Require: $\boldsymbol{X}$ (streamed in SNP blocks), $\boldsymbol{E}\in\mathbb{R}^{N\times L}$, $\boldsymbol{y}\in\mathbb{R}^{N}$, step size η, cap $\theta_{\max}$, tolerances $\varepsilon_{\text{ang}}, \varepsilon_{\text{obj}}$
Ensure: $\widehat{\boldsymbol{\alpha}}\in\mathbb{S}^{L-1}$, $\widehat{\boldsymbol{\sigma}}^2 = (\widehat{\sigma}_g^2, \widehat{\sigma}_{g\times e}^2, \widehat{\sigma}_{n\times e}^2, \widehat{\sigma}_n^2)$

Precompute
Sample rows of $\boldsymbol{W} \in \mathbb{R}^{B\times N}$ from $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$
Cache $\boldsymbol{U}_g = \boldsymbol{K}_g\boldsymbol{W}^\top$
Cache

$$\boldsymbol{U}_{g\times e}[\ell,k] \;=\; \begin{cases} \dfrac{1}{M}\,\boldsymbol{Z}_\ell\boldsymbol{Z}_\ell^\top\boldsymbol{W}^\top, & \ell = k, \\ \dfrac{1}{M}\left(\boldsymbol{Z}_\ell\boldsymbol{Z}_k^\top + \boldsymbol{Z}_k\boldsymbol{Z}_\ell^\top\right)\boldsymbol{W}^\top, & \ell < k, \end{cases} \qquad \ell \le k,$$

where $\boldsymbol{Z}_\ell = \boldsymbol{X}\odot\boldsymbol{E}_{:,\ell}$.
Cache $\boldsymbol{G}_y$, $\boldsymbol{G}_E$, $\boldsymbol{s}_g$
Pick $\boldsymbol{\alpha}^{(0)} \in \mathbb{S}^{L-1}$.
Initialize $\mathcal{L}^{(0)} \leftarrow +\infty$.
for $t = 0, 1, 2, \ldots$ **do**
$\boldsymbol{e}^{(t)} \leftarrow \boldsymbol{E}\boldsymbol{\alpha}^{(t)}$

$$\boldsymbol{U}_{g\times e}(\boldsymbol{\alpha}^{(t)}) \leftarrow \sum_{\ell\le k}\alpha_\ell^{(t)}\alpha_k^{(t)}\,\boldsymbol{U}_{g\times e}[\ell,k].$$

Update $\boldsymbol{\sigma}^2$ (MoM): build $\boldsymbol{A}(\boldsymbol{\alpha}^{(t)})$, $\boldsymbol{b}(\boldsymbol{\alpha}^{(t)})$, and solve

$$\boldsymbol{A}(\boldsymbol{\alpha}^{(t)})\,\widehat{\boldsymbol{\sigma}}^{2\,(t)} = \boldsymbol{b}(\boldsymbol{\alpha}^{(t)}).$$

Compute gradient wrt $\boldsymbol{\alpha}$:
Assemble $\boldsymbol{g}(\boldsymbol{\alpha}^{(t)})$ (including trace terms via scatter-adds)
Set $\boldsymbol{g}_\top \leftarrow \boldsymbol{g} - (\boldsymbol{\alpha}^{(t)\top}\boldsymbol{g})\,\boldsymbol{\alpha}^{(t)}$
if $\|\boldsymbol{g}_\top\|_2 = 0$ **then**
break
end if
Update $\boldsymbol{\alpha}$:
$\theta \leftarrow \min\{\eta\|\boldsymbol{g}_\top\|_2, \theta_{\max}\}$

$$\boldsymbol{\alpha}^{(t+1)} \leftarrow \boldsymbol{\alpha}^{(t)}\cos\theta - \frac{\boldsymbol{g}_\top}{\|\boldsymbol{g}_\top\|_2}\sin\theta$$

Recompute $\boldsymbol{\sigma}^2$ at the new iterate:
$\boldsymbol{e}^{(t+1)} \leftarrow \boldsymbol{E}\boldsymbol{\alpha}^{(t+1)}$

$$\boldsymbol{U}_{g\times e}(\boldsymbol{\alpha}^{(t+1)}) \leftarrow \sum_{\ell<k}\alpha_\ell^{(t+1)}\alpha_k^{(t+1)}\,\boldsymbol{U}_{g\times e}[\ell,k] + \sum_{\ell}(\alpha_\ell^{(t+1)})^2\boldsymbol{U}_{g\times e}[\ell,\ell]$$

Build $\boldsymbol{A}(\boldsymbol{\alpha}^{(t+1)})$, $\boldsymbol{b}(\boldsymbol{\alpha}^{(t+1)})$, and solve

$$\boldsymbol{A}(\boldsymbol{\alpha}^{(t+1)})\,\widehat{\boldsymbol{\sigma}}^{2\,(t+1)} = \boldsymbol{b}(\boldsymbol{\alpha}^{(t+1)}).$$

Compute

$$\mathcal{L}^{(t+1)} \leftarrow \frac{1}{N}\,\mathcal{L}\Big(\boldsymbol{\alpha}^{(t+1)}, \widehat{\boldsymbol{\sigma}}^{2\,(t+1)}\Big).$$

Stop if

$$1 - \langle\boldsymbol{\alpha}^{(t+1)}, \boldsymbol{\alpha}^{(t)}\rangle < \varepsilon_{\text{ang}} \quad \text{and} \quad \frac{|\mathcal{L}^{(t)} - \mathcal{L}^{(t+1)}|}{\max\{1, |\mathcal{L}^{(t)}|\}} < \varepsilon_{\text{obj}}$$

for several consecutive iterations.
end for
return $\widehat{\boldsymbol{\alpha}} \leftarrow \boldsymbol{\alpha}^{(t+1)}$, $\widehat{\boldsymbol{\sigma}}^2 \leftarrow \widehat{\boldsymbol{\sigma}}^{2\,(t+1)}$

References

All of Us Research Program Investigators. 2019. The "all of us" research program. *New England Journal of Medicine* **381**: 668–676.

Boye C, Nirmalan S, Ranjbaran A, and Luca F. 2024. Genotype× environment interactions in gene regulation and complex traits. *Nature Genetics* **56**: 1057–1068.

Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, Motyer A, Vukcevic D, Delaneau O, O'Connell J, et al.. 2018. The uk biobank resource with deep phenotyping and genomic data. *Nature* **562**: 203–209.

Dahl A, Nguyen K, Cai N, Gandal MJ, Flint J, and Zaitlen N. 2020. A robust method uncovers significant context-specific heritability in diverse complex traits. *The American Journal of Human Genetics* **106**: 71–91.

Di Scipio M, Khan M, Mao S, Chong M, Judge C, Pathan N, Perrot N, Nelson W, Lali R, Di S, et al.. 2023. A versatile, fast and unbiased method for estimation of gene-by-environment interaction effects on biobank-scale datasets. *Nature Communications* **14**: 5196.

Ding M, Ellervik C, Huang T, Jensen MK, Curhan GC, Pasquale LR, Kang JH, Wiggs JL, Hunter DJ, Willett WC, et al.. 2018. Diet quality and genetic association with body mass index: results from 3 observational studies. *The American Journal of Clinical Nutrition* **108**: 1291–1300.

Duff IS, Heroux MA, and Pozo R. 2002. An overview of the sparse basic linear algebra subprograms: The new standard from the blas technical forum. *ACM Transactions on Mathematical Software (TOMS)* **28**: 239–267.

Gustavson FG. 1978. Two fast algorithms for sparse matrices: Multiplication and permuted transposition. *ACM Transactions on Mathematical Software (TOMS)* **4**: 250–269.

Hamra GB and Buckley JP. 2018. Environmental exposure mixtures: questions and methods to address them. *Current epidemiology reports* **5**: 160–165.

Hartiala JA, Hilser JR, Biswas S, Lusis AJ, and Allayee H. 2021. Gene-environment interactions for cardiovascular disease. *Current atherosclerosis reports* **23**: 75.

Herrera-Luis E, Benke K, Volk H, Ladd-Acosta C, and Wojcik GL. 2024. Gene–environment interactions in human health. *Nature Reviews Genetics* **25**: 768–784.

Hutchinson MF. 1989. A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines. *Communications in Statistics-Simulation and Computation* **18**: 1059–1076.

Jelenkovic A, Hur YM, Sund R, Yokoyama Y, Siribaddana SH, Hotopf M, Sumathipala A, Rijsdijk F, Tan Q, Zhang D, et al.. 2016. Genetic and environmental influences on adult human height across birth cohorts from 1886 to 1994. *Elife* **5**: e20320.

Jeong M, Pazokitoroudi A, Liu Z, and Sankararaman S. 2024. Scalable summary-statistics-based heritability estimation method with individual genotype level accuracy. *Genome Research* **34**: 1286–1293.

Justice AE, Winkler TW, Feitosa MF, Graff M, Fisher VA, Young K, Barata L, Deng X, Czajkowski J, Hadley D, et al.. 2017. Genome-wide meta-analysis of 241,258 adults accounting for smoking behaviour identifies novel loci for obesity traits. *Nature communications* **8**: 14977.

Kerin M and Marchini J. 2020. Inferring gene-by-environment interactions with a bayesian whole-genome regression model. *The American Journal of Human Genetics* **107**: 698–713.

Kilpeläinen TO, Qi L, Brage S, Sharp SJ, Sonestedt E, Demerath E, Ahmad T, Mora S, Kaakinen M, Sandholt CH, et al.. 2011. Physical activity attenuates the influence of fto variants on obesity risk: a meta-analysis of 218,166 adults and 19,268 children. *PLoS medicine* **8**: e1001116.

Laville V, Majarian T, Sung YJ, Schwander K, Feitosa MF, Chasman DI, Bentley AR, Rotimi CN, Cupples LA, de Vries PS, et al.. 2022. Gene-lifestyle interactions in the genomics of human complex traits. *European Journal of Human Genetics* **30**: 730–739.

Li J, Li X, Zhang S, and Snyder M. 2019. Gene-environment interaction in the era of precision medicine. *Cell* **177**: 38–44.

Makris KC, Baccarelli A, Silverman EK, and Wright RO. 2025. How exposomic tools complement and enrich genomic research. *Cell Genomics* **5**.

Moore R, Casale FP, Jan Bonder M, Horta D, Franke L, Barroso I, and Stegle O. 2019. A linear mixed-model approach to study multivariate gene–environment interactions. *Nature genetics* **51**: 180–186.

Motsinger-Reif AA, Reif DM, Akhtari FS, House JS, Campbell CR, Messier KP, Fargo DC, Bowen TA, Nadadur SS, Schmitt CP, et al.. 2024. Gene-environment interactions within a precision environmental health framework. *Cell Genomics* **4**.

Namba S, Sonehara K, Koyanagi YN, Kikuchi T, Ojima T, Edahiro R, Sato G, Yamaji T, Tomofuji Y, Ueda H, et al.. 2026. A cross-population compendium of gene–environment interactions. *Nature* pp. 1–10.

Ni G, Van Der Werf J, Zhou X, Hyppönen E, Wray NR, and Lee SH. 2019. Genotype–covariate correlation and interaction disentangled by a whole-genome multivariate reaction norm model. *Nature communications* **10**: 2239.

Patel CJ, Bhattacharya J, and Butte AJ. 2010. An environment-wide association study (ewas) on type 2 diabetes mellitus. *PloS one* **5**: e10746.

Patterson HD and Thompson R. 1971. Recovery of inter-block information when block sizes are unequal. *Biometrika* **58**: 545–554.

Pazokitoroudi A, Liu Z, Dahl A, Zaitlen N, Rosset S, and Sankararaman S. 2024. A scalable and robust variance components method reveals insights into the architecture of gene-environment interactions underlying complex traits. *The American Journal of Human Genetics* **111**: 1462–1480.

Pazokitoroudi A, Wu Y, Burch KS, Hou K, Zhou A, Pasaniuc B, and Sankararaman S. 2020. Efficient variance components analysis across millions of genomes. *Nature communications* **11**: 4020.

Qi Q, Chu AY, Kang JH, Jensen MK, Curhan GC, Pasquale LR, Ridker PM, Hunter DJ, Willett WC, Rimm EB, et al.. 2012. Sugar-sweetened beverages and genetic risk of obesity. *New England Journal of Medicine* **367**: 1387–1396.

Reay WR, Atkins JR, Carr VJ, Green MJ, and Cairns MJ. 2020. Pharmacological enrichment of polygenic risk for precision medicine in complex disorders. *Scientific reports* **10**: 879.

Robinson MR, English G, Moser G, Lloyd-Jones LR, Triplett MA, Zhu Z, Nolte IM, van Vliet-Ostaptchouk JV, Snieder H, Study LC, et al.. 2017. Genotype–covariate interaction effects and the heritability of adult body mass index. *Nature genetics* **49**: 1174–1181.

Robinson MR, Hemani G, Medina-Gomez C, Mezzavilla M, Esko T, Shakhbazov K, Powell JE, Vinkhuyzen A, Berndt SI, Gustafsson S, et al.. 2015. Population genetic differentiation of height and body mass index across europe. *Nature genetics* **47**: 1357–1362.

Sadowski M, Thompson M, Mefford J, Haldar T, Oni-Orisan A, Border R, Pazokitoroudi A, Cai N, Ayroles JF, Sankararaman S, et al.. 2024. Characterizing the genetic architecture of drug response using gene-context interaction methods. *Cell Genomics* **4**.

Shungin D, Deng WQ, Varga TV, Luan J, Mihailov E, Metspalu A, Consortium G, Morris AP, Forouhi NG, Lindgren C, et al.. 2017. Ranking and characterization of established bmi and lipid associated loci as candidates for gene-environment interactions. *PLoS genetics* **13**: e1006812.

Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, Downey P, Elliott P, Green J, Landray M, et al.. 2015. Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine* **12**: e1001779.

Sulc J, Mounier N, Günther F, Winkler T, Wood AR, Frayling TM, Heid IM, Robinson MR, and Kutalik Z. 2020. Quantification of the overall contribution of gene-environment interaction for obesity-related traits. *Nature communications* **11**: 1385.

Turner MJ, Clough RW, Martin HC, and Topp LJ. 1956. Stiffness and deflection analysis of complex structures. *Journal of the Aeronautical Sciences* **23**: 805–823.

Tyrrell J, Wood AR, Ames RM, Yaghootkar H, Beaumont RN, Jones SE, Tuke MA, Ruth KS, Freathy RM, Davey Smith G, et al.. 2017. Gene–obesogenic environment interactions in the uk biobank study. *International journal of epidemiology* **46**: 559–575.

Virolainen SJ, VonHandorf A, Viel KCMF, Weirauch MT, and Kottyan LC. 2023. Gene–environment interactions and their impact on human health. *Genes & Immunity* **24**: 1–11.

Wang T, Heianza Y, Sun D, Huang T, Ma W, Rimm EB, Manson JE, Hu FB, Willett WC, and Qi L. 2018. Improving adherence to healthy dietary patterns, genetic risk, and long term weight gain: gene-diet interaction analysis in two prospective cohort studies. *BMJ* **360**.

Watson NF, Harden KP, Buchwald D, Vitiello MV, Pack AI, Weigle DS, and Goldberg J. 2012. Sleep duration and body mass index in twins: a gene-environment interaction. *Sleep* **35**: 597–603.

Weine E, Smith SP, Knowlton RK, and Harpak A. 2025. Trade-offs in modeling context dependency in complex trait genetics. *Elife* **13**: RP99210.

Westerman KE and Sofer T. 2024. Many roads to a gene-environment interaction. *The American Journal of Human Genetics* **111**: 626–635.

Wild CP. 2005. Complementing the genome with an "exposome": the outstanding challenge of environmental exposure measurement in molecular epidemiology. *Cancer Epidemiology Biomarkers & Prevention* **14**: 1847–1850.

Wild CP. 2012. The exposome: from concept to utility. *International journal of epidemiology* **41**: 24–32.

Wu Y and Sankararaman S. 2018. A scalable estimator of snp heritability for biobank-scale data. *Bioinformatics* **34**: i187–i194.

Yang J, Lee SH, Goddard ME, and Visscher PM. 2011. Gcta: a tool for genome-wide complex trait analysis. *The American journal of human genetics* **88**: 76–82.

Yengo L, Vedantam S, Marouli E, Sidorenko J, Bartell E, Sakaue S, Graff M, Eliasen AU, Jiang Y, Raghavan S, et al.. 2022. A saturated map of common genetic variants associated with human height. *Nature* **610**: 704–712.

Yue K, Ma J, Thornton T, and Shojaie A. 2021. Rehe: Fast variance components estimation for linear mixed models. *Genetic epidemiology* **45**: 891–905.

Zhong W, Chhibber A, Luo L, Mehrotra DV, and Shen J. 2023. A fast and powerful linear mixed model approach for genotype-environment interaction tests in large-scale gwas. *Briefings in Bioinformatics* **24**: bbac547.

Zhou X. 2017. A unified framework for variance component estimation with summary statistics in genome-wide association studies. *The annals of applied statistics* **11**: 2027.