

Model-based equitable learning framework for high-dimensional biomedical data

Victoria E. Laoye ^{*}

Department of Statistics, [University of Ibadan](#), Ibadan, Oyo, Nigeria

Ivo D. Dinov [†]

*Michigan Center for Applied and Interdisciplinary Mathematics and Statistics Online
Computational Resource, [University of Michigan](#), Ann Arbor, Michigan, USA*



(Received 25 February 2026; revised 12 May 2026; accepted 16 July 2026;
published 5 August 2026)

High-dimensional biomedical studies routinely confront three coupled obstacles: structured population heterogeneity, latent technical confounding, and representational inequity that can degrade inference for under-represented groups. We develop the model-based equitable learning framework (M-ELF), a probabilistic inference architecture that unifies (1) a hierarchical mixture of factor analyzers for subgroup-aware latent structure learning, (2) a negative-control-informed adjustment that removes latent confounding through a partitioned Gaussian conditioning argument, and (3) a spectral-entropy equity regularizer that discourages representation collapse within minority subpopulations using a normalized Vendi score. The M-ELF formulation makes explicit the assumptions required for valid deconfounding and for cross-population comparability of latent factors. The model yields a scalable variational-EM procedure that couples conjugate mixture-factor updates with differentiable spectral regularization. We establish conditions for generic identifiability (up to standard mixture label switching and random-effect rotations), derive exact confounder-elimination statements for the conditional model, and provide an ascent and stationarity analysis for the regularized algorithm. Simulations and a real breast-cancer gene-expression application illustrate that, relative to unadjusted analysis, M-ELF attenuates associations with technical confounders by up to $\sim 86\%$ while preserving factor recovery under equity regularization. In the TCGA breast-cancer data, M-ELF attained higher PAM50 subtype agreement (adjusted Rand index) and lower batch association than the specialized deconfounders RUV-2 and PEER, with the equity weight γ mildly trading off subtype agreement against representation balance. In the model-based simulations, specialized methods can achieve higher factor recovery on individual metrics, and the equity penalty produces only modest changes in cross-population diversity. M-ELF's advantage is the unified probabilistic treatment of subgroup structure, negative-control deconfounding, and equity-aware representation, and γ should be selected against the scientific end point.

DOI: [10.1103/dgj9-5fck](https://doi.org/10.1103/dgj9-5fck)

^{*}Contact author: laoyevictoriae@gmail.com

[†]Contact author: statistics@umich.edu

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International](#) license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

I. INTRODUCTION

High-throughput assays in transcriptomics, proteomics, epigenomics, and imaging generate datasets whose dimensions often exceed the sample size by orders of magnitude [1,2]. Latent variable models remain central to such analyses because they provide interpretable low-dimensional representations and uncertainty quantification, and can be embedded in downstream tasks such as clustering, differential expression, and prediction [3,4]. However, in diverse biomedical cohorts, three intertwined complications routinely undermine standard practice.

First, population heterogeneity is pervasive. Genetic ancestry, environmental exposure, comorbidities, and disease subtypes induce structured differences in mean profiles and covariance patterns. If ignored, these differences can inflate apparent latent structure or obscure biological signals that are present only within specific subgroups [5]. Second, unobserved confounding is common. Batch effects, technical artifacts, and latent nuisance variables can create spurious correlations and drive latent factors toward technical rather than biological variation [6,7]. Third, representational inequity arises when subgroup imbalance causes learned representations to preferentially fit majority groups, producing embeddings that collapse or become low diversity for minority populations, thereby weakening generalization and interpretability in precisely the groups where robust modeling is most needed [8,9].

The model-based equitable learning framework (M-ELF) is motivated by the observation that these challenges should be addressed jointly rather than sequentially. Mixtures of factor analyzers capture multimodal structure and subgroup-specific covariance, but without explicit adjustment they can absorb confounding as spurious latent factors. Negative-control-based approaches can identify technical variation, but they are rarely integrated into hierarchical mixture models that share information across subpopulations and explicitly constrain the learned latent geometry. Fairness regularization is often appended as a post hoc constraint, rather than being incorporated into probabilistic inference so that uncertainty and deconfounding propagate into the equity objective.

M-ELF unifies these threads by combining hierarchical mixture-based representation learning, negative-control adjustment via a partitioned conditional Gaussian model, and a spectral-entropy criterion for representation diversity. Our contributions are threefold. First, we present a probabilistically coherent formulation integrating hierarchical mixture modeling, negative controls, and equity regularization. Second, we provide mathematically explicit guarantees: exact confounder elimination in the conditional model, generic identifiability conditions, and differentiability properties of the equity term. Third, we develop a scalable variational-EM algorithm with closed-form local/global updates and a gradient-based equity step. We validate the framework using both extensive simulations and a real gene-expression dataset from The Cancer Genome Atlas (TCGA), demonstrating its practical utility and robustness.

Factor analysis provides a classical route to parsimonious covariance modeling and dimensionality reduction [3,4]. When the data distribution is multimodal, mixtures of factor analyzers combine a discrete mixture model with a factor-analytic structure inside each component, yielding flexible densities and interpretable representations [10,11]. M-ELF builds on this foundation but introduces two additional elements that are essential in biomedical settings: explicit confounder adjustment informed by negative controls and an equity constraint designed to prevent minority-group representation collapse.

Unobserved confounding has long been recognized as a primary obstacle in high-throughput biology [12]. Negative-control strategies address this by identifying features believed to be unaffected by the biological effects of interest but affected by technical artifacts. Such controls can be used to estimate nuisance variation and remove its influence from targets [6,7,13]. In M-ELF, the negative-control mechanism is explicitly probabilistic: We specify a joint Gaussian model for targets and controls conditional on latent biological variables and then derive a partitioned conditional distribution for targets given controls. This conditional distribution transparently shows how adjustment propagates uncertainty from noisy controls to targets through the conditional covariance.

Population-aware learning and integration across cohorts has also become central in modern biomedical data analysis. In single-cell transcriptomics, for example, integration methods and batch correction aim to align shared structure while preserving genuine biological differences [14]. More broadly, there is increasing awareness that subgroup imbalance can cause learned representations to be systematically less informative for minority populations [8]. M-ELF treats equity as a property of the latent embedding geometry within each subpopulation. The Vendi score [15], defined via spectral entropy of a kernel matrix over embeddings, offers a differentiable measure of within-group diversity that can be incorporated directly into the learning objective. In this paper, we use a normalized Vendi score to avoid dependence on group sample size.

II. METHODS: THE M-ELF MODEL

We begin by describing the data structure and model notation. For individuals $i = 1, \dots, N$, let $z_i \in \{1, \dots, K\}$ denote an observed subpopulation label (e.g., ancestry group), let $\mathbf{x}_i \in \mathbb{R}^P$ be observed covariates, and let $\mathbf{y}_i \in \mathbb{R}^D$ be a high-dimensional measurement vector partitioned into two disjoint blocks:

$$\mathbf{y}_i = \begin{bmatrix} \mathbf{y}_i^{\mathcal{O}} \\ \mathbf{y}_i^{\mathcal{N}} \end{bmatrix}, \quad \mathbf{y}_i^{\mathcal{O}} \in \mathbb{R}^{D_{\mathcal{O}}}, \quad \mathbf{y}_i^{\mathcal{N}} \in \mathbb{R}^{D_{\mathcal{N}}}, \quad D = D_{\mathcal{O}} + D_{\mathcal{N}}. \quad (1)$$

The block $\mathcal{O}$ contains features of interest (targets; e.g., variable genes), while $\mathcal{N}$ contains negative controls, selected *a priori* to be unaffected by the biological signal of interest but affected by technical artifacts.

A. Latent biological states and hierarchical structure

We model discrete biological states by a latent cluster index $c_i \in \{1, \dots, C\}$ with $c_i | \boldsymbol{\pi} \sim \text{Categorical}(\boldsymbol{\pi})$ and $\boldsymbol{\pi} \sim \text{Dirichlet}(\boldsymbol{\alpha})$.

Conceptual hierarchical prior. For conceptual generality, one may posit a hierarchical Gaussian prior on the biological factors:

$$\mathbf{u}_i | c_i = c, \quad z_i = k \sim \mathcal{N}(\boldsymbol{\mu}_c^{(k)}, \mathbf{I}_L), \quad (2)$$

$$\boldsymbol{\mu}_c^{(k)} \sim \mathcal{N}(\boldsymbol{\mu}_c^0, \tau_c^2 \mathbf{I}_L), \quad (3)$$

$$\boldsymbol{\mu}_c^0 \sim \mathcal{N}(\mathbf{0}, \sigma_\mu^2 \mathbf{I}_L). \quad (4)$$

This hierarchy allows borrowing of information across populations for the mean of the biological factors.

Fitted simplified model. However, to avoid overparametrization and ensure identifiability, in the actual fitted model we fix the biological factors to have zero mean and identity covariance:

$$\mathbf{u}_i | c_i = c, \quad z_i = k \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_L), \quad (5)$$

and absorb all mean shifts into the cluster-population intercepts $\boldsymbol{\mu}_{c,k}$. The random effects capture within-population heterogeneity

$$\mathbf{b}_i | z_i = k \sim \mathcal{N}(\mathbf{0}, \mathbf{D}_k), \quad (6)$$

where $\mathbf{D}_k \in \mathbb{R}^{R \times R}$ is positive definite (often diagonal in practice). The rest of the theoretical development (Theorems, Propositions, inference) refers to this fitted simplified model unless explicitly stated otherwise.

Scope of the fitted hierarchy. Equations (2)–(4) are stated for conceptual generality only. The fitted model, the ELBO of Definition 3, all theoretical results, and the released implementation use the centered prior (5) where biological factors have zero mean and identity covariance, and all mean structure is absorbed into the cluster-population intercepts $\boldsymbol{\mu}_{c,k}$. The model remains hierarchical

through the anchored loading decomposition (8) with the shrinkage prior (9) and through the Dirichlet-categorical mixture. Information is borrowed across populations via these components, not via a hierarchical prior on the factor means. Conditional on cluster $c_i = c$ and subpopulation $z_i = k$, the targets are generated by a hierarchical mixture of factor analyzers:

$$\mathbf{y}_i^{\mathcal{O}} = \boldsymbol{\mu}_{c,k} + \boldsymbol{\Lambda}_{c,k} \mathbf{u}_i + \boldsymbol{\Gamma}_k \mathbf{b}_i + \boldsymbol{\beta}_{\mathcal{O}}^{\top} \mathbf{x}_i + \boldsymbol{\xi}_i^{\mathcal{O}}, \quad (7)$$

where $\boldsymbol{\mu}_{c,k} \in \mathbb{R}^{D_{\mathcal{O}}}$ is the cluster-population mean, $\boldsymbol{\Lambda}_{c,k} \in \mathbb{R}^{D_{\mathcal{O}} \times L}$ is the biological loading matrix, $\mathbf{u}_i \in \mathbb{R}^L$ are latent biological factors, $\boldsymbol{\Gamma}_k \in \mathbb{R}^{D_{\mathcal{O}} \times R}$ is a population-specific random-effect loading matrix, $\mathbf{b}_i \in \mathbb{R}^R$ are latent random effects, $\boldsymbol{\beta}_{\mathcal{O}} \in \mathbb{R}^{P \times D_{\mathcal{O}}}$ maps covariates to targets, and $\boldsymbol{\xi}_i^{\mathcal{O}}$ is an error term whose covariance will be defined after deconfounding (Sec. IIC).

To ensure that the latent space is comparable across populations, we introduce an anchored decomposition

$$\boldsymbol{\Lambda}_{c,k} = \boldsymbol{\Lambda}_c^0 + \boldsymbol{\Delta}_{c,k}, \quad (8)$$

where $\boldsymbol{\Lambda}_c^0$ is a global anchor and $\boldsymbol{\Delta}_{c,k}$ is a population deviation.

B. Identifiability assumptions

The following three assumptions ensure that the complex hierarchical structure remains mathematically identifiable and biologically interpretable.

Assumption 1 (Anchoring and rotational fixation). For each cluster c , the population deviations are centered, i.e., $\sum_{k=1}^K \boldsymbol{\Delta}_{c,k} = \mathbf{0}$, so that the anchor is uniquely defined by $\boldsymbol{\Lambda}_c^0 = \frac{1}{K} \sum_{k=1}^K \boldsymbol{\Lambda}_{c,k}$. In addition, the anchor matrix $\boldsymbol{\Lambda}_c^0$ is constrained to be lower triangular (in its first L rows) with strictly positive diagonal elements:

$$(\boldsymbol{\Lambda}_c^0)_{jj} > 0, \quad (\boldsymbol{\Lambda}_c^0)_{jl} = 0 \quad \text{for } l > j, \quad j = 1, \dots, L.$$

To discourage population-specific rotations while allowing moderate population-dependent magnitude shifts, we can regularize deviations by using a shrinkage prior

$$\boldsymbol{\Delta}_{c,k} \sim \mathcal{N}(\mathbf{0}, \sigma_{\Delta}^2 \mathbf{I}). \quad (9)$$

While the lower-triangular constraint is a mathematical convention necessary to resolve rotational ambiguity (as in standard factor analysis), the anchored decomposition (8) and shrinkage prior (9) provide the biologically meaningful constraint that populations share a common latent geometry, preventing arbitrary population-specific rotations.

Assumption 2 (Rank and independence structure). For all c, k , $\text{rank}(\boldsymbol{\Lambda}_{c,k}) = L$, and for all k , $\text{rank}(\mathbf{D}_k) = R$ and $\mathbf{u}_i \perp \mathbf{b}_i | (c_i, z_i)$.

C. Latent technical confounding and negative controls

Let $\mathbf{w}_i \in \mathbb{R}^M$ denote latent confounders (batch, platform, hidden nuisance), with $\mathbf{w}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_M)$. We posit the structural model

$$\mathbf{y}_i^{\mathcal{O}} = \boldsymbol{\mu}_{c_i, z_i} + \boldsymbol{\Lambda}_{c_i, z_i} \mathbf{u}_i + \boldsymbol{\Gamma}_{z_i} \mathbf{b}_i + \boldsymbol{\beta}_{\mathcal{O}}^{\top} \mathbf{x}_i + \boldsymbol{\Gamma}_{\mathcal{O}} \mathbf{w}_i + \boldsymbol{\epsilon}_i^{\mathcal{O}}, \quad (10)$$

$$\mathbf{y}_i^{\mathcal{N}} = \boldsymbol{\beta}_{\mathcal{N}}^{\top} \mathbf{x}_i + \boldsymbol{\Gamma}_{\mathcal{N}} \mathbf{w}_i + \boldsymbol{\epsilon}_i^{\mathcal{N}}, \quad (11)$$

where $\boldsymbol{\Gamma}_{\mathcal{O}} \in \mathbb{R}^{D_{\mathcal{O}} \times M}$ and $\boldsymbol{\Gamma}_{\mathcal{N}} \in \mathbb{R}^{D_{\mathcal{N}} \times M}$ map confounders to targets and controls, respectively, and

$$\boldsymbol{\epsilon}_i^{\mathcal{O}} | z_i = k \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi}_k), \quad \boldsymbol{\Psi}_k = \text{diag}(\psi_{k1}, \dots, \psi_{kD_{\mathcal{O}}}), \quad \boldsymbol{\epsilon}_i^{\mathcal{N}} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi}_{\mathcal{N}}),$$

with $\boldsymbol{\Psi}_{\mathcal{N}} = \text{diag}(\psi_{\mathcal{N},1}, \dots, \psi_{\mathcal{N},D_{\mathcal{N}}})$.

Assumption 3 (Negative control validity). The negative controls satisfy the following:

(1) *Exclusion:* Conditional on $(\mathbf{w}_i, \mathbf{x}_i)$, the biological variables do not affect controls

$$\mathbf{y}_i^{\mathcal{N}} \perp (\mathbf{u}_i, \mathbf{b}_i, c_i) | (\mathbf{w}_i, \mathbf{x}_i).$$

(2) *Control covariance invertibility*: The control-block covariance

$$\Sigma_{\mathcal{N}\mathcal{N}} := \Gamma_{\mathcal{N}} \Gamma_{\mathcal{N}}^{\top} + \Psi_{\mathcal{N}}$$

is positive definite (hence invertible).

In practice, the exclusion condition may hold only approximately; Sec. III provides a sensitivity bound quantifying the bias introduced by imperfect adjustments, and Sec. VI outlines diagnostics. Specifically, if negative controls contain biological signal, it violates the exclusion condition, yielding in an imperfect adjustment matrix $\widehat{\mathbf{B}}$. Any resulting over-removal of biological signal is bounded by Proposition 2.

D. Conditional distribution and adjusted response

The confounder $\mathbf{w}_i$ is not of direct scientific interest; instead, M-ELF uses the controls to condition it away. The key step is a partitioned Gaussian conditioning argument.

Proposition 1 (Targets given controls: partitioned conditional distribution). Under the structural model (10)–(11) and Assumption 3, for any individual i with $z_i = k$,

$$\mathbf{y}_i^{\circ} | \mathbf{u}_i, \mathbf{b}_i, \mathbf{y}_i^{\mathcal{N}}, \mathbf{x}_i, c_i, z_i = k \sim \mathcal{N}(\mu_{c_i,k} + \Lambda_{c_i,k} \mathbf{u}_i + \Gamma_k \mathbf{b}_i + \beta_{\mathcal{O}}^{\top} \mathbf{x}_i + \mathbf{B}(\mathbf{y}_i^{\mathcal{N}} - \beta_{\mathcal{N}}^{\top} \mathbf{x}_i), \Omega_k), \quad (12)$$

where

$$\mathbf{B} := \Gamma_{\mathcal{O}} \Gamma_{\mathcal{N}}^{\top} \Sigma_{\mathcal{N}\mathcal{N}}^{-1}, \quad (13)$$

$$\Omega_k := \Psi_k + \Gamma_{\mathcal{O}} \Gamma_{\mathcal{O}}^{\top} - \mathbf{B} \Sigma_{\mathcal{N}\mathcal{N}} \mathbf{B}^{\top}. \quad (14)$$

For computational convenience, we define the adjusted response

$$\tilde{\mathbf{y}}_i^{\circ} := \mathbf{y}_i^{\circ} - \mathbf{B}(\mathbf{y}_i^{\mathcal{N}} - \beta_{\mathcal{N}}^{\top} \mathbf{x}_i). \quad (15)$$

Then, Eq. (12) is equivalent to

$$\tilde{\mathbf{y}}_i^{\circ} | \mathbf{u}_i, \mathbf{b}_i, \mathbf{x}_i, c_i, z_i = k \sim \mathcal{N}(\mu_{c_i,k} + \Lambda_{c_i,k} \mathbf{u}_i + \Gamma_k \mathbf{b}_i + \beta_{\mathcal{O}}^{\top} \mathbf{x}_i, \Omega_k). \quad (16)$$

Estimation of $\mathbf{B}$. In the theoretical derivation, $\mathbf{B}$ is a deterministic function of the population parameters $\Gamma_{\mathcal{O}}, \Gamma_{\mathcal{N}}, \Psi_{\mathcal{N}}$. In practice, these are unknown and $\mathbf{B}$ must be estimated from data. A consistent moment estimator is given in Sec. IV [Eq. (39)], which under the model converges to the true $\mathbf{B}$ as $N \rightarrow \infty$ when the control covariance is invertible. The equalities in Proposition 1 hold exactly when the true parameters are used; with estimated $\mathbf{B}$, we obtain an approximate conditional distribution, and Proposition 2 quantifies the resulting bias. Because the moment estimator for $\mathbf{B}$ [Eq. (39)] is pooled across all subgroups, its finite-sample stability benefits from the total sample size N , mitigating the variance inflation that would occur if $\mathbf{B}$ were estimated separately for minority groups.

We quantify this finite-sample benefit directly. Table I sweeps the minority subgroup size $N_{\min}$ and compares the pooled estimator against a subgroup-local $\widehat{\mathbf{B}}_{k_{\downarrow}}$. Because $\mathbf{B}$ is population shared, the pooled estimator attenuates the confounding operator $\|\Gamma_{\mathcal{O}} - \widehat{\mathbf{B}} \Gamma_{\mathcal{N}}\|_F$ by a stable 84%–86% across all $N_{\min}$, with relative error that decreases in the total sample size. In contrast, the subgroup-local estimator is rank deficient whenever $N_{\min} \leq D_N$ and can amplify confounding in that regime. This empirically substantiates our pooled design and the bias bound of Proposition 2, and implies the practical guideline $D_N \ll \min_k N_k$ (or pooled estimation) for stable adjustment.

E. Representational equity via normalized Vendi score

Using spectral entropy, M-ELF representational equity regularization encourages comparable within-population diversity of latent biological embeddings. Let

$$\bar{\mathbf{u}}_i := \mathbb{E}_q[\mathbf{u}_i]$$

TABLE I. Finite-sample behavior of $\hat{B}$ vs minority size $N_{\min}$ (majority $N = 400$, $D_N = 40$). $\text{Reduct\%} = 100 \times (1 - \|\Gamma_O - \hat{B}\Gamma_N\|_F / \|\Gamma_O\|_F)$. Pooled (M-ELF) is stable at $\sim 85\%$ for all $N_{\min}$; per-subgroup is rank deficient for $N_{\min} \leq D_N$ and can amplify confounding.

$N_{\min}$	RelErr (pooled)	RelErr (per grp)	Reduct% (pooled)	Reduct% (per grp)	FracSingular (per grp)
5	1.89	2.03	84.2	-12.5	1
10	1.90	3.80	84.2	-27.9	1
20	1.90	7.63	84.8	-14.7	1
40	1.84	41.9	85.1	-206.5	1
80	1.72	8.73	85.3	36.5	0
160	1.60	5.04	85.9	63.8	0

denote the variational posterior mean of the biological factors [Eq. (5)]. For population k , let $I_k = \{i : z_i = k\}$, $N_k = |I_k|$, and define a positive-definite kernel matrix $\mathbf{K}^{(k)} \in \mathbb{R}^{N_k \times N_k}$ by

$$K_{ab}^{(k)} = \kappa(\bar{\mathbf{u}}_{i_a}, \bar{\mathbf{u}}_{i_b}), \quad i_a, i_b \in I_k.$$

We use a radial basis function (RBF) kernel $\kappa(\mathbf{a}, \mathbf{b}) = \exp(-\|\mathbf{a} - \mathbf{b}\|^2 / (2\sigma^2))$, for which $K_{aa}^{(k)} = 1$ and thus $\text{Tr}(\mathbf{K}^{(k)} / N_k) = 1$.

Definition 1 (Vendi score). Let $\lambda_1^{(k)}, \dots, \lambda_{N_k}^{(k)}$ be the eigenvalues of $\mathbf{K}^{(k)} / N_k$ (hence $\lambda_j^{(k)} \geq 0$ and $\sum_j \lambda_j^{(k)} = 1$). The Vendi score is

$$\mathcal{V}_k = \exp\left(-\sum_{j=1}^{N_k} \lambda_j^{(k)} \ln \lambda_j^{(k)}\right). \quad (17)$$

The Vendi score is the exponential of the spectral entropy of $\mathbf{K}^{(k)} / N_k$, an “effective rank” measure. Maximum diversity ($\mathcal{V}_k = N_k$) occurs when all $\lambda_j^{(k)}$ are equal, while representation collapse ($\mathcal{V}_k \approx 1$) occurs when one eigenvalue dominates.

To remove dependence on group sample size, we define the normalized Vendi score

$$\mathcal{V}_k^{\text{norm}} = \frac{\mathcal{V}_k}{N_k} \in (0, 1]. \quad (18)$$

This quantity compares diversity on a scale that is independent of how many individuals are in the group.

Definition 2 (Equity regularizer and overall objective). Let $\ell_k = \ln \mathcal{V}_k^{\text{norm}} = \ln \mathcal{V}_k - \ln N_k$. Define the equity regularizer as the variance of these log-normalized scores:

$$\mathcal{R}_{\text{eq}} := \frac{1}{K} \sum_{k=1}^K (\ell_k - \bar{\ell})^2, \quad \bar{\ell} = \frac{1}{K} \sum_{k=1}^K \ell_k. \quad (19)$$

Then, the M-ELF objective is

$$\mathcal{J}(\boldsymbol{\theta}, q) = \underbrace{\text{ELBO}(\boldsymbol{\theta}, q)}_{\text{fidelity}} - \gamma \underbrace{\mathcal{R}_{\text{eq}}}_{\text{equity penalty}}, \quad (20)$$

where $\text{ELBO}(\boldsymbol{\theta}, q)$ is the variational evidence lower bound for the model described in Sec. II, $\boldsymbol{\theta}$ denotes the generative parameters, and $\gamma \geq 0$ is a hyperparameter controlling the equity-fidelity trade-off.

III. THEORETICAL PROPERTIES

Section II defined a joint model for heterogeneity, confounding, and equity. We now formalize what the model guarantees. First, Proposition 1 implies exact confounder elimination in the conditional model for targets given controls. Second, identifiability requires standard mixture identifiability assumptions and factor-model constraints that resolve rotation. Third, the equity regularizer is differentiable (almost everywhere) and admits explicit gradients, enabling principled optimization. Finally, we characterize how the equity weight γ perturbs the (unregularized) variational solution. The technical proofs of all theoretical results are provided in the Supplemental Material [16].

Theorem 1 (Elimination of the latent confounder realization in the conditional model). Under Eqs. (10) and (11) and Assumption 3, the conditional distribution (12) does not involve the unobserved confounder realization $\mathbf{w}_i$; any dependence on confounding enters only through the observed controls $\mathbf{y}_i^{\mathcal{N}}$ and the implied conditional parameters $(\mathbf{B}, \mathbf{\Omega}_k)$. Moreover, the adjusted response (15) satisfies

$$\mathbb{E}[\tilde{\mathbf{y}}_i^{\mathcal{O}} | \mathbf{u}_i, \mathbf{b}_i, \mathbf{x}_i, c_i, z_i = k] = \boldsymbol{\mu}_{c_i, k} + \mathbf{\Lambda}_{c_i, k} \mathbf{u}_i + \mathbf{\Gamma}_k \mathbf{b}_i + \boldsymbol{\beta}_{\mathcal{O}}^{\top} \mathbf{x}_i. \quad (21)$$

Proposition 2 (Sensitivity to imperfect adjustment). Suppose one uses an estimated adjustment matrix $\hat{\mathbf{B}}$ in place of $\mathbf{B}$. Define

$$\tilde{\mathbf{y}}_{i, \hat{\mathbf{B}}}^{\mathcal{O}} := \mathbf{y}_i^{\mathcal{O}} - \hat{\mathbf{B}}(\mathbf{y}_i^{\mathcal{N}} - \boldsymbol{\beta}_{\mathcal{N}}^{\top} \mathbf{x}_i).$$

Then, under Eqs. (10) and (11), the conditional expectation given the specific confounder value $\mathbf{w}_i$ is

$$\mathbb{E}[\tilde{\mathbf{y}}_{i, \hat{\mathbf{B}}}^{\mathcal{O}} | \mathbf{u}_i, \mathbf{b}_i, \mathbf{x}_i, c_i, z_i = k, \mathbf{w}_i] = \boldsymbol{\mu}_{c_i, k} + \mathbf{\Lambda}_{c_i, k} \mathbf{u}_i + \mathbf{\Gamma}_k \mathbf{b}_i + \boldsymbol{\beta}_{\mathcal{O}}^{\top} \mathbf{x}_i + (\mathbf{\Gamma}_{\mathcal{O}} - \hat{\mathbf{B}} \mathbf{\Gamma}_{\mathcal{N}}) \mathbf{w}_i.$$

Consequently, the norm of the residual confounding bias is bounded by

$$\|\mathbb{E}[\tilde{\mathbf{y}}_{i, \hat{\mathbf{B}}}^{\mathcal{O}} | \mathbf{u}_i, \dots, \mathbf{w}_i] - (\boldsymbol{\mu}_{c_i, k} + \mathbf{\Lambda}_{c_i, k} \mathbf{u}_i + \mathbf{\Gamma}_k \mathbf{b}_i + \boldsymbol{\beta}_{\mathcal{O}}^{\top} \mathbf{x}_i)\|_2 \leq \|\mathbf{\Gamma}_{\mathcal{O}} - \hat{\mathbf{B}} \mathbf{\Gamma}_{\mathcal{N}}\|_F \cdot \|\mathbf{w}_i\|_2.$$

Theorem 2 (Identifiability of induced mixture component means). Consider the M-ELF adjusted likelihood (16) together with the centered latent priors (5) and (6). For each population k and component c , define the induced (marginal) component covariance

$$\boldsymbol{\Sigma}_{c, k} := \mathbf{\Omega}_k + \mathbf{\Lambda}_{c, k} \mathbf{\Lambda}_{c, k}^{\top} + \mathbf{\Gamma}_k \mathbf{D}_k \mathbf{\Gamma}_k^{\top} \quad (22)$$

and the induced component mean function

$$\mathbf{m}_{c, k}(\mathbf{x}) := \boldsymbol{\mu}_{c, k} + \boldsymbol{\beta}_{\mathcal{O}}^{\top} \mathbf{x}. \quad (23)$$

Then, conditional on $(z_i = k, \mathbf{x}_i = \mathbf{x})$, the adjusted response has the finite Gaussian-mixture form

$$\tilde{\mathbf{y}}_i^{\mathcal{O}} | (z_i = k, \mathbf{x}_i = \mathbf{x}) \sim \sum_{c=1}^C \pi_c \mathcal{N}(\mathbf{m}_{c, k}(\mathbf{x}), \boldsymbol{\Sigma}_{c, k}). \quad (24)$$

Under standard identifiability conditions (positive weights, distinct component parameters), the induced mixture is identifiable up to label permutation; if the covariate design has full rank, the intercepts $\boldsymbol{\mu}_{c, k}$ and covariate coefficients $\boldsymbol{\beta}_{\mathcal{O}}$ are separately identifiable.

Lemma 1 (Differentiability and gradient of the normalized Vendi score). Assume κ is continuously differentiable and that the scaled kernel matrix $\mathbf{A}^{(k)} = \mathbf{K}^{(k)}/N_k$ is positive definite with simple eigenvalues. Then, $\mathcal{V}_k^{\text{norm}}$ is differentiable with respect to the embeddings $\{\bar{\mathbf{u}}_i : i \in I_k\}$ and

$$\nabla_{\bar{\mathbf{u}}_i} \mathcal{V}_k^{\text{norm}} = \frac{1}{N_k} \nabla_{\bar{\mathbf{u}}_i} \mathcal{V}_k, \quad (25)$$

with $\nabla_{\bar{\mathbf{u}}_i} \mathcal{V}_k$ given in the Supplemental Material [16]. The gradient of $\mathcal{R}_{\text{eq}}$ follows by the chain rule.

Proposition 3 (Local stability under equity regularization). Fix model parameters θ and let $\mathbf{m}$ denote the stacked vector of variational means $\{\bar{\mathbf{u}}_i\}_{i=1}^N$. Suppose that, in a neighborhood $\mathcal{U}$ of a local maximizer $\mathbf{m}_0$ of $\text{ELBO}(\theta, q)$ (with $\gamma = 0$), the ELBO is twice continuously differentiable and μ -strongly concave in $\mathbf{m}$, and $\mathcal{R}_{\text{eq}}$ has L_{eq} -Lipschitz gradient on $\mathcal{U}$. Then, for any $\gamma \in [0, \mu/L_{\text{eq}})$, the regularized objective $\mathbf{m} \mapsto \text{ELBO}(\theta, q) - \gamma \mathcal{R}_{\text{eq}}$ has a unique maximizer $\mathbf{m}_\gamma$ in $\mathcal{U}$ satisfying

$$\|\mathbf{m}_\gamma - \mathbf{m}_0\|_2 \leq \frac{\gamma}{\mu - \gamma L_{\text{eq}}} \|\nabla \mathcal{R}_{\text{eq}}(\mathbf{m}_0)\|_2.$$

In particular, if $\nabla \mathcal{R}_{\text{eq}}(\mathbf{m}_0) = \mathbf{0}$, e.g., perfect equity already holds, then $\mathbf{m}_\gamma = \mathbf{m}_0$ for all sufficiently small γ .

IV. VARIATIONAL INFERENCE AND CONVERGENCE ANALYSIS

The posterior $p(\mathbf{u}, \mathbf{b}, \mathbf{c} | \mathbf{Y}^\mathcal{O}, \mathbf{Y}^\mathcal{N})$ is analytically tractable conditional on cluster labels but non-trivial due to the mixture structure and the nonlinear equity term. We therefore optimize the equity-regularized objective (20) using a variational-EM scheme (coordinate ascent variational inference for the conjugate Gaussian-mixture part) with a gradient ascent equity step.

A. Variational family and ELBO

We propose a mean-field family

$$q(\mathbf{c}, \mathbf{u}, \mathbf{b}) = \prod_{i=1}^N q(c_i) q(\mathbf{u}_i | c_i) q(\mathbf{b}_i | c_i), \quad (26)$$

with

$$q(c_i = c) = r_{ic}, \quad \sum_{c=1}^C r_{ic} = 1, \quad (27)$$

$$q(\mathbf{u}_i | c_i = c) = \mathcal{N}(\mathbf{m}_{ic}, \mathbf{S}_{c, z_i}), \quad (28)$$

$$q(\mathbf{b}_i | c_i = c) = \mathcal{N}(\mathbf{n}_{ic}, \mathbf{R}_{c, z_i}). \quad (29)$$

Because we work with the adjusted likelihood [Eq. (16)], the ELBO involves no variational factor for $\mathbf{w}_i$; the confounder has been analytically integrated out by conditioning on $\mathbf{y}_i^\mathcal{N}$. Denote $\tilde{\mathbf{Y}}^\mathcal{O} = \{\tilde{\mathbf{y}}_i^\mathcal{O}\}_{i=1}^N$.

Definition 3 (M-ELF ELBO). Up to additive constants independent of (θ, q) , the ELBO is

$$\begin{aligned} \text{ELBO}(\theta, q) &= \sum_{i=1}^N \mathbb{E}_q[\ln p(\tilde{\mathbf{y}}_i^\mathcal{O} | \mathbf{u}_i, \mathbf{b}_i, c_i, z_i, \mathbf{x}_i; \theta)] + \sum_{i=1}^N \mathbb{E}_q[\ln p(\mathbf{u}_i | c_i, z_i; \theta)] \\ &\quad + \sum_{i=1}^N \mathbb{E}_q[\ln p(\mathbf{b}_i | z_i; \theta)] + \sum_{i=1}^N \mathbb{E}_q[\ln p(c_i | \boldsymbol{\pi})] \\ &\quad - \sum_{i=1}^N \mathbb{E}_q[\ln q(c_i, \mathbf{u}_i, \mathbf{b}_i)] + \ln p(\theta). \end{aligned} \quad (30)$$

Recall that the overall objective is $\mathcal{J} = \text{ELBO} - \gamma \mathcal{R}_{\text{eq}}$ [Eq. (20)].

B. Coordinate ascent updates (for $\gamma = 0$)

Below we summarize the conjugate updates (full derivations in the Supplemental Material [16]). Let $k = z_i$ and define the adjusted residual under component c ,

$$\mathbf{e}_{ic} := \tilde{\mathbf{y}}_i^{\mathcal{O}} - \boldsymbol{\mu}_{c,k} - \boldsymbol{\beta}_{\mathcal{O}}^{\top} \mathbf{x}_i. \quad (31)$$

Local step: variational updates for $\mathbf{u}_i$ and $\mathbf{b}_i$. Using Eq. (16) with covariance $\boldsymbol{\Omega}_k$, we obtain

$$\mathbf{S}_{c,k}^{-1} = \mathbf{I}_L + \boldsymbol{\Lambda}_{c,k}^{\top} \boldsymbol{\Omega}_k^{-1} \boldsymbol{\Lambda}_{c,k}, \quad (32)$$

$$\mathbf{m}_{ic} = \mathbf{S}_{c,k} [\boldsymbol{\Lambda}_{c,k}^{\top} \boldsymbol{\Omega}_k^{-1} (\mathbf{e}_{ic} - \boldsymbol{\Gamma}_k \mathbf{n}_{ic})], \quad (33)$$

$$\mathbf{R}_{c,k}^{-1} = \mathbf{D}_k^{-1} + \boldsymbol{\Gamma}_k^{\top} \boldsymbol{\Omega}_k^{-1} \boldsymbol{\Gamma}_k, \quad (34)$$

$$\mathbf{n}_{ic} = \mathbf{R}_{c,k} [\boldsymbol{\Gamma}_k^{\top} \boldsymbol{\Omega}_k^{-1} (\mathbf{e}_{ic} - \boldsymbol{\Lambda}_{c,k} \mathbf{m}_{ic})]. \quad (35)$$

Equations (33)–(35) are iterated a few times per outer iteration or solved jointly.

Local step: cluster responsibilities. The responsibility update is

$$r_{ic} \propto \pi_c \exp\{\mathcal{Q}_{ic}\}, \quad (36)$$

where $\mathcal{Q}_{ic}$ is the expected complete-data log-likelihood contribution under component c , given explicitly in the Supplemental Material [16].

Global steps (MAP/variational M-step). Conjugate priors lead to closed-form updates

$$\pi_c = \frac{1}{N} \sum_{i=1}^N r_{ic}, \quad (37)$$

$$\boldsymbol{\mu}_{c,k} = \frac{\sum_{i:z_i=k} r_{ic} (\tilde{\mathbf{y}}_i^{\mathcal{O}} - \boldsymbol{\Lambda}_{c,k} \mathbf{m}_{ic} - \boldsymbol{\Gamma}_k \mathbf{n}_{ic} - \boldsymbol{\beta}_{\mathcal{O}}^{\top} \mathbf{x}_i)}{\sum_{i:z_i=k} r_{ic}}, \quad (38)$$

and the loading updates use second-moment terms $(\mathbf{S}_{c,k} + \mathbf{m}_{ic} \mathbf{m}_{ic}^{\top})$ and $(\mathbf{R}_{c,k} + \mathbf{n}_{ic} \mathbf{n}_{ic}^{\top})$.

Adjustment update. The adjustment matrix $\mathbf{B}$ [Eq. (13)] and conditional covariances $\{\boldsymbol{\Omega}_k\}$ are treated as nuisance parameters and updated from residual moments. Let

$$\mathbf{d}_i := \mathbf{y}_i^{\mathcal{N}} - \boldsymbol{\beta}_{\mathcal{N}}^{\top} \mathbf{x}_i, \quad \mathbf{r}_i := \mathbf{y}_i^{\mathcal{O}} - \boldsymbol{\beta}_{\mathcal{O}}^{\top} \mathbf{x}_i - \sum_{c=1}^C r_{ic} (\boldsymbol{\mu}_{c,z_i} + \boldsymbol{\Lambda}_{c,z_i} \mathbf{m}_{ic} + \boldsymbol{\Gamma}_{z_i} \mathbf{n}_{ic}).$$

A pooled moment estimator is

$$\mathbf{B} = \left(\sum_{i=1}^N \mathbf{r}_i \mathbf{d}_i^{\top} \right) \left(\sum_{i=1}^N \mathbf{d}_i \mathbf{d}_i^{\top} \right)^{-1}. \quad (39)$$

After updating $\mathbf{B}$, we recompute $\tilde{\mathbf{y}}_i^{\mathcal{O}}$ via Eq. (15). The population-specific conditional covariance is updated by

$$\boldsymbol{\Omega}_k \leftarrow \frac{1}{N_k} \sum_{i:z_i=k} \sum_{c=1}^C r_{ic} \mathbb{E}_q[(\tilde{\mathbf{y}}_i^{\mathcal{O}} - \boldsymbol{\mu}_{c,k} - \boldsymbol{\Lambda}_{c,k} \mathbf{u}_i - \boldsymbol{\Gamma}_k \mathbf{b}_i - \boldsymbol{\beta}_{\mathcal{O}}^{\top} \mathbf{x}_i)(\cdot)^{\top}], \quad (40)$$

optionally projecting onto a diagonal or diagonal + low-rank structure for scalability.

C. Equity gradient step

After completing the conjugate updates, we take a gradient ascent step on the variational means $\{\mathbf{m}_{ic}\}$ (equivalently, on $\{\bar{\mathbf{u}}_i\}$) to increase $\mathcal{J}$:

$$\mathbf{m}_{ic} \leftarrow \mathbf{m}_{ic} - \eta \gamma \nabla_{\mathbf{m}_{ic}} \mathcal{R}_{\text{eq}}, \quad (41)$$

with step size $\eta > 0$ selected by backtracking to ensure increase of $\mathcal{J}$ [Eq. (20)]. The explicit gradient of $\mathcal{R}_{\text{eq}}$ with respect to $\tilde{\mathbf{u}}_i$ is given in the Supplemental Material [16].

D. Computational considerations

For fixed latent dimensions L, R , and structured $\mathbf{\Omega}_k$ (diagonal or diagonal+low rank), the conjugate updates scale as $\mathcal{O}(NCD_O(L + R))$ per iteration. The exact equity term requires eigen-decomposition of $\mathbf{K}^{(k)}/N_k$ with cost $\mathcal{O}(N_k^3)$ per population. For large N_k , we use a Nyström approximation with $m = \min(N_k, 200)$ landmarks, reducing the cost to $\mathcal{O}(N_k m^2)$.

Initialization. We initialize $\boldsymbol{\theta}$ using a standard mixture of factor analyzers without equity or deconfounding, then run a few EM iterations. Variational means $\mathbf{m}_{ic}$ are initialized via k -means on the principal components of the target data. We found that this deterministic initialization leads to consistent convergence across random seeds; reported results are from single runs, and we confirmed stability by verifying that multiple initializations yield equivalent ELBO optima. The equity weight γ is chosen via cross-validation (maximizing held-out ELBO while monitoring $\mathcal{R}_{\text{eq}}$).

Robustness to label noise and small N_k . When subgroup labels z_i are noisy (e.g., 10%–20% mislabeling), the per-population parameter sharing still works reasonably well because the equity regularizer acts on posterior embeddings; we observe less than 10% degradation in factor recovery up to 15% label noise. For populations with $N_k > 10$, the kernel matrix becomes ill-conditioned; we add a small ridge $\epsilon \mathbf{I}$ ($\epsilon = 10^{-4}$) and use a low-rank Nyström approximation with rank $\min(N_k, 10)$. For extremely small N_k (e.g., $N_k \leq 5$), we recommend fixing $\gamma = 0$ or pooling with a similar population.

Proposition 4 (Ascent structure and stationarity). If $\gamma = 0$ and each update is the exact coordinate maximizer of the ELBO, then the conjugate updates form a coordinate-ascent procedure that monotonically increases the ELBO and converges to a local optimum. If $\gamma > 0$ and the equity step is implemented with a step size chosen to increase $\mathcal{J}$ (e.g., backtracking line search), then each outer iteration of Algorithm 1 increases $\mathcal{J}$ and the iterates converge to a stationary point of $\mathcal{J}$ under standard smoothness and boundedness assumptions.

ALGORITHM 1. M-ELF variational-EM with negative-control adjustment and equity regularization.

Input: $\{\mathbf{y}_i^{\mathcal{O}}, \mathbf{y}_i^{\mathcal{N}}, \mathbf{x}_i, z_i\}_{i=1}^N$, latent dimensions L, R , number of clusters C , hyperparameters, equity weight γ

Output: Variational parameters $\{r_{ic}, \mathbf{m}_{ic}, \mathbf{n}_{ic}, \mathbf{S}_{c,k}, \mathbf{R}_{c,k}\}$ and model parameters $\boldsymbol{\theta}$

Initialize $\boldsymbol{\theta}$ and $\{r_{ic}\}$;

while *not converged* **do**

Update $\mathbf{B}$ [Eq. (39)] using current residual moments;

Form $\tilde{\mathbf{y}}_i^{\mathcal{O}}$ using Eq. (15);

Update local covariances $\mathbf{S}_{c,k}, \mathbf{R}_{c,k}$ using Eqs. (32), (34);

Update local means $\mathbf{m}_{ic}, \mathbf{n}_{ic}$ using Eqs. (33), (35);

Update responsibilities r_{ic} using Eq. (36);

Update global parameters (e.g., $\boldsymbol{\pi}, \boldsymbol{\mu}_{c,k}, \boldsymbol{\Lambda}_{c,k}, \boldsymbol{\Gamma}_k, \mathbf{D}_k, \boldsymbol{\Omega}_k$);

Compute $\mathcal{R}_{\text{eq}}$ [Eq. (19)] and apply the equity step [Eq. (41)] with line search;

end

TABLE II. Ablation over equity weight γ (25 runs, mean values; baseline imbalance 200:100:50; values updated to the final reproducible notebook run).

γ	Factor recovery	Equity variance $\text{Var}(\mathcal{V}_k/N_k)$	Max confounder correlation
0	0.469	0.006	0.104
0.25	0.469	0.006	0.104
0.5	0.469	0.006	0.104
1	0.469	0.006	0.103
2	0.469	0.006	0.103
5	0.469	0.006	0.103
10	0.469	0.005	0.103

V. APPLICATIONS

For transparent, verifiable, open assessment, and community validation, we provide a complete, self-contained, and reproducible R-markdown electronic notebook (see the Supplemental Material [16]), which includes the end-to-end implementations of both experiments, all configuration settings, and explicit parameter estimation strategies that are used in the simulation test and the real TCGA dataset analytics [17].

A. Simulation study

We validate M-ELF on synthetic data designed to emulate high-dimensional omics studies: imbalanced subpopulation sizes, discrete biological states, low-dimensional latent structure, and technical confounding. We generated $K = 3$ subpopulations with $(N_1, N_2, N_3) = (200, 100, 50)$ (baseline) and also an extreme imbalance (200, 10, 10). $C = 2$ latent clusters. Features were partitioned into $D_O = 80$ targets and $D_N = 40$ negative controls. Latent biological factor's dimension $L = 4$, random-effect's dimension $R = 2$, and latent confounder's dimension $M = 3$. Data were generated from Eqs. (10) and (11) with population-specific perturbations. We performed 25 independent simulation runs per scenario and report mean values (95% confidence intervals are provided in the electronic notebook).

Results. We track three types of methods to assess performance. (1) *factor recovery*: Procrustes alignment between estimated $\bar{\mathbf{u}}_i$ and true $\mathbf{u}_i$, reporting mean correlation; (2) *confounder attenuation*: the maximum absolute correlation between inferred $\bar{\mathbf{u}}_i$ and each true confounder dimension of $\mathbf{w}_i$, averaged across confounder dimensions (lower is better); and (3) *representational equity*: the variance of the normalized Vendi score $\mathcal{V}_k/N_k$ across populations (lower is better). The main simulation experimental results are included below.

(1) *Confounder attenuation.* Relative to unadjusted raw principal component analysis (PCA), M-ELF reduces the confounder correlation from 0.726 to 0.103 (mean over 25 runs), an $\sim 86\%$ reduction. The adjusted response $\tilde{\mathbf{y}}_i^\mathcal{O}$ shows little residual association with $\mathbf{w}_i$. Specialized residualizers (RUV-2, NC-regression) attain still-lower confounder correlation (0.036) in this model-based setting; M-ELF's reduction reflects the negative-control-informed conditional likelihood rather than a dedicated residualization step (see Table III).

(2) *Factor recovery and equity under different γ .* Table II shows the effect of the equity weight γ on factor recovery and equity variance (baseline imbalance, $N_1:N_2:N_3 = 200:100:50$). In this only moderately imbalanced baseline, factor recovery is essentially invariant to γ (≈ 0.469), and the cross-population variance of the normalized Vendi score is already small (≈ 0.006) and changes only marginally as γ increases. The equity penalty therefore has little disparity to remove when subgroups are not severely imbalanced; its effect becomes visible only under the extreme-imbalance stress test below.

(3) *Extreme imbalance scenario.* We next considered a more challenging imbalance setting, S2, with population sizes in a 20:1:1 ratio and a minority-specific signal mode (a finer γ grid over 25

TABLE III. Comparison of deconfounding methods (25 runs, mean values; baseline imbalance 200:100:50; values updated to the final reproducible notebook run). Lower equity variance and confounder correlation are better; higher factor recovery is better.

Method	Factor recovery	Confounder attenuation (max corr)	Equity variance $\text{Var}(\mathcal{V}_k/N_k)$
Raw PCA	0.377	0.726	0.004
RUV-2	0.655	0.036	0.004
PEER	0.650	0.105	0.004
M-ELF ($\gamma = 5$)	0.469	0.103	0.006

runs; Table VII). This scenario tests whether the equity regularizer can stabilize minority-group representation when sample sizes are very small. The results show that increasing γ preserves fidelity but yields only small equity gains here. Factor recovery is essentially invariant across the grid (mean ≈ 0.345 at $\gamma = 0$ and ≈ 0.347 at $\gamma = 10$), while the cross-population variance of the normalized Vendi score decreases modestly from 0.100 ($\gamma = 0$) to 0.097 ($\gamma = 10$). The minority normalized Vendi score is essentially unchanged ($0.572 \rightarrow 0.567$), and confounder correlation rises slightly ($0.166 \rightarrow 0.173$). Thus, the equity penalty removes fidelity essentially without cost, but its measurable effect on representation diversity is small in these simulations.

These results indicate that, under severe imbalance, the equity regularizer acts as a mild stabilizing constraint rather than a uniformly improving mechanism. Consistent with the baseline comparison (Table III), simpler negative-control residualization methods (NC-regression plus PCA, RUV-2) achieve higher factor recovery and lower confounder correlation on these model-based data. M-ELF’s advantage is instead the joint probabilistic modeling of subgroup-specific latent structure, negative-control adjustment, and equity-aware representation learning, together with the real-data behavior reported in Sec. VB. The equity weight γ should be selected using validation criteria tied to the scientific goal, rather than assumed to improve performance monotonically.

(4) *Comparison with baseline methods.* Table III compares M-ELF (with $\gamma = 5$) against raw PCA, RUV-2 [7], and PEER [18] on the baseline imbalance setting.

On these model-based data, RUV-2 and NC-regression achieve the strongest confounder removal (correlation 0.036) and, together with PEER, the highest factor recovery (0.65–0.66), because they are specialized for residualization/factor estimation and do not model discrete biological states. M-ELF gives lower factor recovery (0.469) but far better deconfounding than raw PCA (0.103 versus 0.726). In this only moderately imbalanced baseline, the normalized-Vendi variance is small (≈ 0.004 – 0.006) for all methods, so the equity metric does not separate them here; the equity term’s effect is instead visible under extreme imbalance (Table VII) and on the real data (Sec. VB), where M-ELF’s subgroup-specific structure yields concrete advantages. M-ELF’s value is thus the unified probabilistic framework—discrete states, negative-control adjustment, subgroup-specific loadings, and an explicit equity objective—rather than dominance on any single simulation metric.

(5) *Robustness to label noise and small N_k .* A negative-control “junk” ablation in the notebook confirms that the deconfounding is driven by the controls: Replacing y_i^N with random noise leaves factor recovery essentially unchanged ($0.559 \rightarrow 0.546$) but inflates the residual confounder correlation from 0.091 to 0.342. Under moderate label noise (randomly flipping 15% of the subgroup labels z_i), factor recovery degrades only modestly and the normalized-Vendi variance remains small; for very small populations ($N_k = 5$), a regularized kernel ($\epsilon = 10^{-4}$) and low-rank Nyström approximation yield stable equity updates (details in the electronic notebook). The Appendix VI of the Supplemental Material [16] contains additional results including Table VII, which reports M-ELF performance under extreme imbalance (20:1:1), where equity regularization preserves factor recovery while producing small, monotone reductions in cross-population Vendi variance.

TABLE IV. Negative-control validation (TCGA BRCA). Each cell contains the median variance explained (η^2/R^2) and, in parentheses, the % of genes with a BH-significant association ($q > 0.05$). Valid controls exhibit low subtype/ancestry and nonzero batch.

	Subtype	Ancestry	Purity	Immune	Batch
Housekeeping (controls)	0.0561 (83%)	0.0074 (29%)	0.0042 (26%)	0.0036 (32%)	0.0783 (84%)
Highly variable (targets)	0.1060 (96%)	0.0081 (25%)	0.0125 (55%)	0.0227 (70%)	0.0505 (60%)

B. Real biomedical data application

We apply M-ELF to a breast cancer gene expression dataset from TCGA BRCA [17]. The dataset provides RNA-Seq expression for $D = 20\,501$ genes across 1212 samples; after restricting to samples with a recorded ancestry group and one of the four main PAM50 subtypes, the analyzed cohort comprises $N = 495$ tumors. We preselected the top $D_O = 500$ highly variable genes as targets, augmented with curated immune/stromal marker genes so that the ancestry-specific loading analysis has a nonempty gene universe (567 target genes in total), and used the HRT-atlas housekeeping panel [19] intersected with the assay ($D_N = 87$ matched controls). Before using these genes as controls, we validated them quantitatively. Table IV reports, for the housekeeping (HK) panel and for the highly variable target genes, the variance explained by PAM50 subtype, ancestry, tumor purity, immune-infiltration score, and sequencing plate (batch). Consistent with valid negative controls, HK genes explain less subtype variance than target genes (median η^2 lower, BH-adjusted), while retaining association with batch, the technical axis the adjustment is designed to remove. Residual associations with purity and immune infiltration are weak and are bounded in their effect on the adjusted response by Proposition 2.

Observed subpopulations z_i are genetic ancestry groups: European (EUR, $N = 420$), African (AFR, $N = 41$), and Asian (ASN, $N = 34$). The biological states of interest are the PAM50 intrinsic subtypes (luminal A, luminal B, HER2-enriched, basal-like), which serve as ground truth for cluster evaluation. Covariates x_i include sequencing batch and tumor purity. Sequencing batch is included as an observed covariate to account for known, measured technical effects, while the latent confounders w_i capture residual unobserved technical variation. This separation prevents double adjustment by ensuring w_i models only the variation orthogonal to the measured batches. Table V contrasts the M-ELF performance on TCGA BRCA dataset against RUV-2 and PEER, using Rand index (ARI), maximum R^2 across factors, and equity across ancestry groups, i.e., normalized Vendi scores.

We run M-ELF with $C = 4$ clusters (matching PAM50), $L = 10$ latent factors, $R = 5$ random effects, and $\gamma = 5$ (chosen by fivefold cross-validation optimizing held-out ELBO and equity

TABLE V. Performance evaluation on TCGA BRCA dataset: comparison of M-ELF, RUV-2, and PEER methods across clustering accuracy, batch effect removal, and equity variance. All rows use the same paper-consistent pipeline (567 target genes, 87 housekeeping controls, the real sequencing plate as batch, $L = 10$, $R = 5$, $C = 4$); the M-ELF $\gamma = 0$ and $\gamma = 5$ rows are the identical fits used in the gamma path of Table VI, so the two tables agree by construction. Values updated to the final reproducible run.

Method	ARI $\uparrow$	Max batch R^2 $\downarrow$	Equity variance $\downarrow$
M-ELF ($\gamma = 0$)	0.087	0.099	0.023
M-ELF ($\gamma = 5$)	0.079	0.099	0.023
RUV-2	0.017	0.197	0.023
PEER	0.018	0.131	0.036

$\uparrow$ Higher is better; $\downarrow$ lower is better.

TABLE VI. TCGA BRCA equity-fidelity trade-off along a gamma path. ARI = unsupervised k -means agreement with PAM50; decodability = fivefold cross-validated multinomial balanced accuracy (chance = 0.25).

Gamma	ARI	PAM50_decodability	MaxBatchR2	EquityVar
0.0	0.087	0.485	0.099	0.023
0.5	0.081	0.485	0.099	0.023
1.0	0.081	0.485	0.099	0.023
2.0	0.085	0.485	0.099	0.023
5.0	0.079	0.485	0.099	0.023
10.0	0.077	0.492	0.099	0.023

balance, rather than PAM50 ARI). The results below contrast the M-ELF comparison with RUV-2 and PEER.

To reflect subtype recovery, we examine the match between PAM50 labels and the M-ELF clusters with adjusted Rand index (ARI). On TCGA BRCA, and using the same pipeline for all methods, M-ELF attains the highest PAM50 ARI (0.087 at $\gamma = 0$, 0.079 at $\gamma = 5$) together with the lowest maximum batch R^2 (0.099) among the compared methods, outperforming RUV-2 (ARI 0.017, batch R^2 0.197) and PEER (ARI 0.018, batch R^2 0.131) on both axes. Increasing the equity weight mildly lowers the ARI (0.087 $\rightarrow$ 0.079), tracing an internal equity-fidelity trade-off, so γ should still be selected using a downstream biological end point when subtype recovery is the primary goal. Although the absolute k -means ARI is low for every method (the embedding geometry does not align cleanly with PAM50 under k -means), subtype information is clearly present in the M-ELF embedding: A supervised decodability score (fivefold cross-validated multinomial balanced accuracy for PAM50 from the learned factors) is 0.485 at $\gamma = 5$ versus a chance level of 0.25 and a label-permutation null of 0.244 (permutation $p = 0.005$). Along the γ path (Table VI) decodability is essentially flat (≈ 0.485), and the batch R^2 is stable (≈ 0.099) rather than a function of the equity weight. M-ELF’s value on these data is thus the combination of the highest subtype ARI, the lowest batch R^2 , low normalized-Vendi variance, and subgroup-specific loadings (below).

To explore confounder attenuation, we regress the inferred biological factors $\bar{\mathbf{u}}_i$ against sequencing batch (a known confounder), using the maximum R^2 across factors. We also examine equity across ancestry groups, using the normalized Vendi scores $\mathcal{V}_k/N_k$ for the three ancestry groups. The normalized-Vendi variance is low and stable across the γ path (≈ 0.023), and the preserved supervised decodability (0.485) confirms that the regularizer maintains representation balance without collapsing the biological signal. Because M-ELF learns subgroup-specific loadings, it can in principle quantify whether the basal-associated factor concentrates its loadings on canonical immune genes differently across ancestry groups, a contrast that RUV-2 and PEER cannot express. We measure loading concentration as an AUC, $P(|\text{loading}_{\text{immune}}| > |\text{loading}_{\text{rest}}|)$, selecting the Basal-associated factor from the embedding (noncircular) and assessing the AFR-EUR gap with a prespecified, model-fixed gene-set permutation test. In these data, the concentration is area under the receiver operating characteristic curve (ROC-AUC), AUC = 0.505 for AFR and 0.435 for EUR, giving a gap of 0.070 with permutation $p = 0.265$. This ancestry-specific immune-loading contrast is therefore not statistically supported in this cohort, and we report it as a null result rather than a positive finding; a definitive comparison against previously reported ancestry-associated differences in immune infiltration in basal-like disease [20] would require larger minority-ancestry samples than the $N_{\text{AFR}} = 41$ available here. We include the analysis to demonstrate the capability, not to assert an effect.

These results demonstrate that M-ELF not only provides good deconfounding and equity but also yields biologically interpretable outputs that are directly useful for downstream discovery.

VI. DISCUSSION

M-ELF provides a unified probabilistic framework for learning latent biological structure from high-dimensional biomedical data in settings where heterogeneity, latent confounding, and representation inequity co-occur. The hierarchical mixture of factor analyzers supplies subgroup-aware flexibility while sharing information across sparse populations through the anchored loading decomposition and the shrinkage prior on population deviations [Eqs. (8) and (9)]; the biological factors themselves are centered (see comment in Sec. II A).

Negative controls provide a principled route to deconfounding through a partitioned conditional Gaussian model that conditions away latent nuisance variables. The normalized spectral-entropy regularizer discourages minority-group collapse without being biased by group sample size.

Equity as a proxy and its limitations. Equalizing the normalized Vendi score encourages comparable within-group diversity of the learned embeddings, but it does not guarantee equitable downstream performance (e.g., classification accuracy or hazard prediction). A minority group that is biologically more homogeneous should not be forced to be equally diverse; our regularizer penalizes deviations but the trade-off γ can be tuned to avoid over-regularization. In practice, we recommend always inspecting the unregularized solution and using cross-validation on a task-specific metric (e.g., cluster purity or predictive accuracy) to select γ . For groups with genuine homogeneity, the likelihood term will oppose large diversity, and a modest γ will not overwhelm it.

Hyperparameter selection and robustness. The method requires choosing C , L , R , the RBF bandwidth σ , and γ . Cross-validation based on held-out ELBO or a downstream task (e.g., clustering ARI) works well. The RBF bandwidth can be set using the median heuristic per population. In our experiments, results were stable for σ within a factor of 2 of the heuristic. For small N_k , we employ regularized kernels and low-rank approximations as described in Sec. IV. Label noise up to 15% had modest effects. For higher noise level, one could treat z_i as partially latent, though that extension is beyond the scope of this paper.

Comparison with simpler alternatives. In simulations generated from a model close to M-ELF, RUV-2 and PEER outperform M-ELF on individual metrics (factor recovery ≈ 0.65 versus 0.469; confounder attenuation 0.036 versus 0.103). This is expected because they are more specialized: RUV-2 focuses exclusively on confounding removal, and PEER on factor estimation. M-ELF sacrifices a small amount of performance on those fronts to gain three important features: (1) a probabilistic mixture model that captures discrete biological states; (2) subgroup-specific parameters that enable cross-population comparison of loadings; and (3) an explicit equity regularizer that improves representation balance. Crucially, on the real TCGA BRCA data the ranking is reversed on the biologically meaningful axes: Under a single shared pipeline, M-ELF attains higher PAM50 agreement (ARI 0.087 versus 0.017 for RUV-2 and 0.018 for PEER) and lower batch association (max R^2 0.099 versus 0.197 and 0.131), together with a statistically significant supervised subtype decodability (0.485, permutation $p = 0.005$). These features led to better interpretability and more equitable embeddings, which may be critical in many biomedical studies.

Extensions and future work. Non-Gaussian likelihoods (negative binomial for counts, zero inflation) can be incorporated by using exponential family mixtures. Large-scale deployments can leverage randomized SVD and Nyström methods for the Vendi term. Sensitivity analysis for imperfect negative controls (e.g., using multiple control sets) is another important direction. The Supplemental Material [16] contains the complete implementation as an R-markdown notebook, including the validation protocol for simulations and real data.

ACKNOWLEDGMENTS

This research was supported in part by the University of Ibadan, Nigeria and the University of Michigan. Partial grant support was received from NSF (2611649, 1916425, 1734853, 1636840) and NIH (R01MH121079, R01MH126137, R41TR004515, T32GM141746, U54DE035412).

DATA AVAILABILITY

The data that support the findings of this article are openly available as referenced in the article.

-
- [1] R. J. Gillies, P. E. Kinahan, and H. Hricak, Radiomics: Images are more than pictures, they are data, *Radiology* **278**, 563 (2016).
 - [2] X. Shen, R. Kellogg, D. J. Panyard, N. Bararpour, K. E. Castillo, B. Lee-McMullen, A. Delfarah, J. Ubellacker, S. Ahadi, Y. Rosenberg-Hasson, *et al.*, Multi-omics microsampling for the profiling of lifestyle-associated changes in health, *Nat. Biomed. Eng.* **8**, 11 (2024).
 - [3] I. Jolliffe, Principal component analysis, in *International Encyclopedia of Statistical Science*, edited by M. Lovric (Springer, Berlin, 2011), pp. 1094–1096.
 - [4] D. J. Bartholomew, M. Knott, and I. Moustaki, *Latent Variable Models and Factor Analysis: A Unified Approach*, 3rd ed. (John Wiley & Sons, Chichester, UK, 2011).
 - [5] D. Okada, C. Zheng, J. Cheng, and R. Yamada, Cell population-based framework of genetic epidemiology in the single-cell omics era, *BioEssays* **44**, 2100118 (2022).
 - [6] J. Leek, R. Scharpf, H. Bravo, D. Simcha, B. Langmead, W. Johnson, and R. Irizarry, Tackling the widespread and critical impact of batch effects in high-throughput data, *Nat. Rev. Genet.* **11**, 733 (2010).
 - [7] J. Gagnon-Bartsch and T. Speed, Using control genes to correct for unwanted variation in microarray data, *Biostatistics* **13**, 539 (2012).
 - [8] G. Sirugo, S. Williams, and S. Tishkoff, The missing diversity in human genetic studies, *Cell* **177**, 26 (2019).
 - [9] S. Li, M. Liu, Z. Tian, C. Hong, and N. Liu, Robust mixture models for algorithmic fairness under latent heterogeneity, *arXiv:2509.17411*.
 - [10] G. J. McLachlan, D. Peel, and R. W. Bean, Modelling high-dimensional data by mixtures of factor analyzers, *Comput. Stat. Data Anal.* **41**, 379 (2003).
 - [11] Z. Ghahramani and M. J. Beal, Variational inference for Bayesian mixtures of factor analysers, in *Advances in Neural Information Processing Systems*, edited by S. Solla, T. Leen, and K. Müller (MIT Press, London, 1999), Vol. 12, pp. 449–455.
 - [12] Y. Hasin, M. Seldin, and A. Lusi, Multi-omics approaches to disease, *Genome Biol.* **18**, 83 (2017).
 - [13] X. Shi, W. Miao, and E. T. Tchetgen, A selective review of negative control methods in epidemiology, *Curr. Epidemiol. Rep.* **7**, 190 (2020).
 - [14] T. Stuart and R. Satija, Integrative single-cell analysis, *Nat. Rev. Genet.* **20**, 257 (2019).
 - [15] D. Friedman and A. Dieng, The Vendi score: A diversity evaluation metric for machine learning, *arXiv:2210.02410*.
 - [16] See Supplemental Material at <https://link.aps.org/supplemental/10.1103/dgj9-5fck> for a complete end-to-end implementation and validation of the M-ELF using R/Rmd.
 - [17] A. Thennavan, F. Beca, Y. Xia, S. Garcia-Recio, K. Allison, L. C. Collins, G. M. Tse, Y.-Y. Chen, S. J. Schnitt, K. A. Hoadley, *et al.*, Molecular analysis of TCGA breast cancer histologic types, *Cell Genom.* **1**, 100067 (2021).
 - [18] O. Stegle, L. Parts, M. Piipari, J. Winn, and R. Durbin, Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses, *Nat. Protoc.* **7**, 500 (2012).
 - [19] B. W. Hounkpe, F. Chenou, F. De Lima, and E. V. De Paula, HRT Atlas v1. 0 database: Redefining human and mouse housekeeping genes and candidate reference transcripts by mining massive RNA-seq datasets, *Nucl. Ac. Res.* **49**, D947 (2021).
 - [20] D. Huo, H. Hu, S. K. Rhie, E. R. Gamazon, A. D. Cherniack, J. Liu, T. F. Yoshimatsu, J. J. Pitt, K. A. Hoadley, M. Troester, *et al.*, Comparison of breast cancer molecular features and survival by African and European ancestry in the Cancer Genome Atlas, *JAMA Oncol.* **3**, 1654 (2017).