IDENTIFYING BIAS DIRECTIONS FROM POPULATION RESIDUALS IN REFERENCE-SHIFTED DECONVOLUTION

Anonymous authors
Paper under double-blind review

ABSTRACT

We study reference-based linear deconvolution problems, which estimate component fractions from aggregate observations using a potentially shifted component profile. Our study is motivated by gene expression deconvolution. In biomedicine, bulk RNA-seq measures aggregate gene expression in a tissue; to infer cell-type composition, one deconvolutes it by referring to a cell-type gene expression profile, which is yet measured from separate tissues and by a different technology (scRNA-seq). Consequently, the reference profile can significantly differ from the true profile that generated the aggregates. Previous methods either neglect this shift or address it only in preprocessing; they then mostly use least squares to perform deconvolution for each tissue individually. Given that a cohort of similar bulk RNA-seq measurements is usually available, we propose to use post-fit residuals together to correct individual deconvolution bias. Our algorithm is developed from rigorous analysis. We first show that under a shared shift in reference, the population residuals alone cannot identify the fraction biases, motivating us to build a structural bridge that connects markered residuals to the population fraction error. Then, based on the latter, we study the conditions under which individual fraction bias can be safely corrected. The efficacy of our correction algorithm is proven under the PAC framework. Experimentally, we first conduct synthetic studies that demonstrate the residual geometry, bias identifiability, and PAC learning effectiveness. Second, we test an instantiation of our algorithm named PopLS on eight cancer pseudo-bulk datasets and two real-bulk cohorts, which further verify the benefits of population-based deconvolution.

1 INTRODUCTION

Reference-based deconvolution is commonly used in many composition-estimation problems. In the biomedical domain, examples are bulk RNA-seq (RNA sequencing) deconvolution (Dai et al., 2026), tissue-of-origin estimation from cell-free DNA (De Ridder et al., 2024), and microbial source tracking (Huang et al., 2024). Particularly to our interest, bulk RNA-seq measures the *aggregate* gene expression levels within a tissue that consists of numerous cells; and a deconvolution algorithm aims to recover the cell type *fractions* from bulk according to a separate cellular gene expression profile (i.e., a reference). This appears to be a classical reference-based linear inverse problem and is ill-posed because of non-unique solutions. Least squares has been frequently employed as the inverse solver due to its simplicity and efficacy.

A long-standing assumption problem of reference-based deconvolution lies in the inconsistency between target mixtures and their reference. In practice, targets and references are usually generated and measured in different samples, protocols, or times. In bulk RNA-seq deconvolution, bulk tissues and single-cell references can exhibit systematic expression differences even for the same cell type (Newman et al., 2019; Jew et al., 2020; Lu et al., 2025; Zhao et al., 2026). As a result, a deconvolution algorithm is likely to produce biased composition estimates for a target based on a shifted reference. Existing methods address this problem mainly by inserting biological insights (such as signature genes, transcriptome properties) to guide finding more appropriate solutions (Tsoucas et al., 2019; Wang et al., 2019; Lu et al., 2025). However, a major limitation is that they perform deconvolution *independently* for each bulk tissue, while a cohort of bulk RNA-seq is available. We argue that such isolated deconvolutions fail to use the population information, which can help recognize the systematic shift in reference and further correct the individual estimation bias.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

To this end, we complementarily study a group of related deconvolution problems that use a shared and potentially shifted reference. In our context, a cohort of bulk RNA-seq measurements from tumors of the same cancer type is available, such as from TCGA (Hoadley et al., 2018; Zhang et al., 2021), which offers various cancer cohorts with each having hundreds of samples. Say, given m bulk measurements (or targets) $\mathbf{y}_1, \ldots, \mathbf{y}_m \in \mathbb{R}^g$ (g genes' aggregate expression) and a cellular gene expression profile $U \in \mathbb{R}^{g \times p}$ (p cell types), we aim to estimate the cell type fractions $\mathbf{f}_1, \ldots, \mathbf{f}_m \in \mathbb{R}^p$. Existing work fits each $\widehat{\mathbf{f}}_i$ based on U independently. In contrast, we consider the generation of $\mathbf{y}_i = M_i \mathbf{f}_i + \boldsymbol{\xi}_i$, where $M_i = U + \Delta + H_i$ is the true generating profile different from U. Being aware of the shared shift Δ, we first fit each target by least squares using the reference U and then restrict the residuals $\mathbf{r}_i = \mathbf{y}_i - U\widehat{\mathbf{f}}_i$ to attend a stable, component-enriched marker subset $\mathcal{M}$ derived from U. These markered residuals $\mathbf{r}_{i,\mathcal{M}}$ are used together to identify the individual bias direction of $\mathbf{f}_i - \widehat{\mathbf{f}}_i$ and perform safe correction for $\widehat{\mathbf{f}}_i$. Our algorithm is named PopLS, and contributions are summarized below.

• **Residual geometry, identifiability, and markered residuals learning.** Under the shared reference shift $M = U + \Delta$, we show that residuals $\{\mathbf{r}_i\}$ alone cannot identify their fraction errors $\{\widehat{\mathbf{f}}_i - \mathbf{f}_i\}$ without additional structure (Theorem 4.1, Proposition 4.2). To build a mapping from residuals to fractions, we restrict each residual to only attend a marker-subset (i.e., marker genes) $\mathcal{M}$ that is derived from the reference. Then, we show that both the markered population residual $\bar{\mathbf{r}}_{\mathcal{M}}$ and individual residuals $\mathbf{r}_{i,\mathcal{M}}$ achieve tighter error bounds (Theorem 4.3).

• **From markered residuals to safe fraction correction.** We construct a component-enriched marker-response decoder T and estimate the desired population correction $\mathbf{d}^\star$ by $\widehat{\mathbf{d}} = T(\bar{\mathbf{r}}_{\mathcal{M}}) \approx \mathbf{d}^\star$. The errors between $\widehat{\mathbf{d}}$ and $\mathbf{d}^\star$, $\widehat{\mathbf{d}}$ and $\mathbf{d}_i$ (desired individual correction) are analyzed in Theorem 4.4. Finally, we provide sufficient conditions under which the observed $\widehat{\mathbf{d}}$ is guaranteed to reduce individual fraction estimation error (Theorem 4.5).

• **Empirical validation.** We first conduct synthetic studies that demonstrate the residual geometry, bias identifiability, and PAC learning effectiveness. Second, we test fully-pipelined PopLS on eight cancer pseudo-bulk datasets and two real-bulk cohorts. Results show that our population-based deconvolution method achieves the lowest estimation error compared to eight SOTA methods that perform individual deconvolution.

2 Related Work

Residual diagnostics. Residual analysis studies the causes of residuals, which mainly stem from inappropriate modeling or data deviation. Model-related problems include incorrect assumptions, insufficient functional form, missing covariates, etc.; they can be alleviated by fixing assumptions, redefining functions, and adding new features (Li et al., 2023; Escanciano, 2024; Liu et al., 2025). Data-related problems can be anomalous data, group-level shift, and instability under perturbation; they can be addressed by anomaly detection, group correction, adding regularization, or simply increasing the data volume (Cook, 1977; Fisher et al., 2023; Kwon et al., 2025). Our work belongs to the modeling category, for residuals are caused by reference shift and non-uniqueness in solutions. We fix these two issues by explicitly modeling the shift and attending marker modules in residuals.

Multi-task regression. Multi-task learning improves individual task estimation by sharing information across related tasks (Ilbert et al., 2024; Knight & Duan, 2023; Wang & Arora, 2024). (Duan & Wang, 2023) adaptively shrinks task estimates toward a robustly learned prototype; (Kim, 2026) studies matrix-weighted multi-task linear regression and uses second-moment-weighted regularization toward a shared centroid. A major difference of our setting is that every $\mathbf{f}_i$ (coefficients) is estimated against the same reference U (features), whereas in multi-task regression, each task is equipped with own features, thus having more information. Moreover, our reference U is potentially shifted, and we have proven that simply averaging $\widehat{\mathbf{f}}_i$ does not remove the shift effect. We therefore exploit markered residuals and map them to coefficients through a novel bridge T for correction.

Reference-based deconvolution with reference shift. Reference-based bulk RNA-seq deconvolution estimates cell-type fractions for a given bulk, and methods differ mainly in the solvers. Least-squares-based methods include MuSiC (Wang et al., 2019), DWLS (Tsoucas et al., 2019),

ReDeconv (Lu et al., 2025), and Bisque (Jew et al., 2020); they study how to incorporate gene weights, reference variability, or reference-bulk transformations into the fit. Other methods: ν-support vector regression used in CIBERSORTx (Newman et al., 2019); Bayesian/probabilistic solvers used in BayesPrism (Chu et al., 2022) and Statescope (Janssen et al., 2026); neural solvers used in Scaden (Menden et al., 2020), TAPE (Chen et al., 2022), and DECODE (Zhao et al., 2026); non-negative geometric inference unsed in cellGeometry (Lau et al., 2026); and count-space generative bridge used in Count Bridges (Fishman et al., 2026). Despite different solvers, their final composition inferences are done for each bulk individually. Our work is complementary: after shared-reference fits for a group of bulks, we use all their post-fit residuals to construct a shared fraction-correction signal. Besides, most works neglect reference shift except ReDeconv, Bisque, and CIBERSORTx, which adjust the reference in the preprocessing step but not in the modeling.

Extended related work with more comparisons is provided in Appendix F.

3 Preliminaries

We study reference-shifted deconvolution problems that are abstracted from bulk RNA-seq deconvolution. This section details the problem setup, with a compact summary of the core notations defined in Appendix A.

Basic notations. In general, bold lowercase letters are column vectors, and capital letters are matrices. For a matrix X, $\|X\|_{\mathrm{op}}$ is its spectral norm and $X^\dagger$ its Moore-Penrose inverse. Below we denote coefficients, targets, and reference by

$$\mathbf{f}_i \in \mathcal{F} := \{\mathbf{f} \in \mathbb{R}^p : f_k \geq 0,\ \mathbf{1}^\top \mathbf{f} = 1\}, \qquad \mathbf{y}_i \in \mathbb{R}^g, \qquad U \in \mathbb{R}^{g\times p} \text{ w. } \operatorname{rank}(U) = p.$$

In our context, $i \in [m]$ is the bulk tissue index, $\mathbf{f}_i$ the component fractions of p cell types of bulk i (so $\mathcal{F}$ is a $p-1$ dimensional simplex), and $\mathbf{y}_i$ the aggregate gene expression levels in bulk i. U is a separate cell-type gene expression profile, serving as the deconvolution reference.

Aggregate model with shift. We consider the case where aggregates $\{\mathbf{y}_i\}$ can be generated from a true hidden profile M_i different from U. That is,

$$\mathbf{y}_i = M_i \mathbf{f}_i + \boldsymbol{\xi}_i, \qquad M_i = U + \Delta + H_i. \tag{1}$$

Here $M_i \in \mathbb{R}^{g\times p}$ is a latent component-mean operator, representing the underlying true profile to generate $\mathbf{y}_i$. Further, Δ captures a systematic shift shared by the m aggregates, while H_i allows a target-specific deviation. And $\boldsymbol{\xi}_i := \mathbf{y}_i - M_i\mathbf{f}_i$ is a finite-aggregate sampling noise. This general framework unifies three modeling scenarios: matched case ($\Delta = H_i = 0$), shared-shift ($\Delta \neq 0$, $H_i = 0$), and heterogeneous-shift ($\Delta \neq 0$, $H_i \neq 0$). Under the shared-shift scenario, all targets share $M_i = U + \Delta$, and randomness across aggregates comes from $\mathbf{f}_i \sim \mathcal{P}_f$ and $\boldsymbol{\xi}_i$, not from M_i. Note that conditioning on the realized fractions, $\mathcal{P}_f$ does not require a parametric form.

Initial least-squares estimates. The initial reference-based estimates, residuals, and biases are computed as

$$\widehat{\mathbf{f}}_i \in \arg\min_{\mathbf{a}\in\mathbb{R}^p} \|\mathbf{y}_i - U\mathbf{a}\|_2^2 = U^\dagger \mathbf{y}_i, \qquad \mathbf{r}_i := \mathbf{y}_i - U\widehat{\mathbf{f}}_i, \qquad \mathbf{d}_i := \mathbf{f}_i - \widehat{\mathbf{f}}_i. \tag{2}$$

4 Our Results

Our analysis and algorithm are presented in tandem. First, we analyze the geometry of the post-fit residuals and reveal its relation to the shared reference shift (Sec. 4.1). Second, we restrict the residuals to component-enriched marker coordinates and design a reliable PAC learning of the markered population residual $\mathbf{r}^\star_{\mathcal{M}}$ (Sec. 4.2). Third, we build a structural bridge $T(\cdot)$ that links $\mathbf{r}^\star_{\mathcal{M}}$ to the desired population fraction correction $\mathbf{d}^\star$ (Sec. 4.3). $T(\cdot)$ can also be used to identify the direction for target-level correction (Sec. 4.4), with sufficient conditions provided for safe analytic correction and an observable empirical correction (Secs. 4.5). Throughout this section, we mainly study the shared-shift scenario, i.e., $H_i = 0$ and $M_i = U + \Delta$. This is particular to our context as we use Δ to model the discrepancy between RNA sequencing technologies – reference U is generated by single-cell RNA-seq while bulk measurements $\{\mathbf{y}_i\}$ are by tissue RNA-seq. Tissue-level H_i does exist but is negligible compared to Δ, so we include it in the sampling noise $\boldsymbol{\xi}_i$ instead. Nevertheless, we also discuss the extension of $H_i \neq 0$ in Appendix C.4 and highlight the differences.

4.1 RESIDUAL GEOMETRY AND IDENTIFIABILITY

Let $P = UU^\dagger$ be the orthogonal projector onto the column space of the reference U and let $Q = I_g - P$. We decompose the shared reference shift as $\Delta = P\Delta + Q\Delta =: \Delta_P + \Delta_Q$, where Δ_P is the column-space component of Δ relative to U, and Δ_Q is its orthogonal component.

Theorem 4.1 (Geometry of reference shift)**.** *Under (1) with $H_i = 0$ and (2), the initial least-squares coefficient error and reconstruction residual satisfy*

$$\widehat{\mathbf{f}}_i - \mathbf{f}_i = U^\dagger \Delta \mathbf{f}_i + U^\dagger \boldsymbol{\xi}_i = U^\dagger \Delta_P \mathbf{f}_i + U^\dagger \boldsymbol{\xi}_i, \tag{3}$$

$$\mathbf{r}_i = Q\Delta \mathbf{f}_i + Q\boldsymbol{\xi}_i = \Delta_Q \mathbf{f}_i + Q\boldsymbol{\xi}_i. \tag{4}$$

Equations (3)-(4) show the complementary effects of Δ: the column-space component Δ_P is absorbed into the least-squares estimate $\widehat{\mathbf{f}}_i$ and produces the coefficient bias $U^\dagger \Delta_P \mathbf{f}_i$, whereas the orthogonal component Δ_Q produces the systematic residual term $\Delta_Q \mathbf{f}_i$ in $\mathbf{r}_i$. The proof is given in Appendix C.1.

Proposition 4.2 (Residual-only non-identifiability)**.** *Without a structural restriction that links Δ_P and Δ_Q, residuals alone do not identify the coefficient-error direction, even with infinitely many noiseless aggregates:*

$$\Delta_Q \mathbf{f} \not\Rightarrow -U^\dagger \Delta_P \mathbf{f}. \tag{5}$$

The two constructions have complementary statistical meanings. For any arbitrary $B \in \mathbb{R}^{g \times p}$, let $\widetilde{\Delta}_P := PB$ and $\widetilde{\Delta}_Q := QB$ denote the column-space and orthogonal mismatch perturbations. When added to Δ, $\widetilde{\Delta}_P$ may alter $\widehat{\mathbf{f}}_i$ but leaves every noiseless residual unchanged because $Q\widetilde{\Delta}_P = 0$; $\widetilde{\Delta}_Q$ may alter $\mathbf{r}_i$ but leaves $\widehat{\mathbf{f}}_i$ unchanged because $U^\dagger \widetilde{\Delta}_Q = 0$. Thus residual magnitude or cross-sample stability alone cannot determine the fraction-correction direction (complete constructions are given in Appendix C.1). Therefore, this motivates us to build a structural bridge that appropriately links residuals to fractions (Sec. 4.3).

4.2 LEARNING OF THE POPULATION MARKERED RESIDUAL

Now we start presenting our learning algorithm. This sub-section deals with the PAC learning of the population markered residual. Using a subset of marker coordinates, such as marker genes, can substantially improve the learning stability. We restrict the residuals to only attend to a subset of marker coordinates that have two properties: they are enriched for a particular component (cell type) and have relatively low conditional variability within that component. Note that such a selection is totally derived from the reference U, and details can be found in Appendix D. Then, for each component k, let S_k denote its selected marker set and define $\mathcal{M} := \bigcup_{k=1}^{p} S_k$, $q := |\mathcal{M}|$. Let $R_\mathcal{M} \in \{0,1\}^{q \times g}$ be the corresponding row-selection operator and $U_\mathcal{M} := R_\mathcal{M} U$. The initial estimate $\widehat{\mathbf{f}}_i$ remains in (2). Continuing from Theorem 4.1, define

$$\mathbf{r}_{i,\mathcal{M}} := R_\mathcal{M} \mathbf{r}_i = R_\mathcal{M}(\mathbf{y}_i - U\widehat{\mathbf{f}}_i) = R_\mathcal{M}\Delta_Q \mathbf{f}_i + R_\mathcal{M} Q \boldsymbol{\xi}_i. \tag{6}$$

Further define

$$\bar{\mathbf{f}} = \frac{1}{m}\sum_{i=1}^{m} \mathbf{f}_i, \qquad \mathbf{r}^\star_{i,\mathcal{M}} = R_\mathcal{M}\Delta_Q \mathbf{f}_i, \qquad \mathbf{r}^\star_\mathcal{M} = R_\mathcal{M}\Delta_Q \bar{\mathbf{f}}.$$

Let $\bar{\mathbf{r}}_\mathcal{M} = m^{-1}\sum_{i=1}^{m} \mathbf{r}_{i,\mathcal{M}}$ be the observed population markered residual. Then

$$\bar{\mathbf{r}}_\mathcal{M} - \mathbf{r}^\star_\mathcal{M} = \frac{1}{m}\sum_{i=1}^{m} R_\mathcal{M} Q \boldsymbol{\xi}_i. \tag{7}$$

To specify the randomness of $\boldsymbol{\xi}_i$, we consider a conditional Gaussian model. In our context, a bulk tissue i comprises n_i (usually tens of thousands of) cells, whose expression profiles are sampled independently from cell-type-specific distributions with means collected in M_i. Besides, the cell-type composition vector $\mathbf{f}_i$ is independently drawn from $\mathcal{P}_f$; based on $\mathbf{f}_i$, cellular expression profiles are aggregated to form $\mathbf{y}_i = M_i \mathbf{f}_i + \boldsymbol{\xi}_i$. When each represented cell type contains sufficiently many sampled cells, the Central Limit Theorem gives $\mathbb{P}(\boldsymbol{\xi}_i \mid \mathbf{f}_i) \dot{\sim} \mathcal{N}\big(0, n_i^{-1}\Omega(\mathbf{f}_i)\big)$, where $\Omega(\mathbf{f}_i)$ denotes

the covariance. Full derivation of this model from the biological context is given in Appendix B. Since $\mathbf{r}_{i,\mathcal{M}} - \mathbf{r}^{\star}_{i,\mathcal{M}} = R_{\mathcal{M}} Q \boldsymbol{\xi}_i$, conditional independence across the m targets gives

$$\Sigma_{\bar{\mathbf{r}},\mathcal{M}} := \frac{1}{m^2} \sum_{i=1}^{m} \frac{1}{n_i} R_{\mathcal{M}} Q \Omega(\mathbf{f}_i) Q R_{\mathcal{M}}^{\top}. \tag{8}$$

Theorem 4.3 (Learning of the population markered residual signal)**.** *Under the conditional Gaussian approximation above and conditional independence across the m target mixtures, for any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\|\bar{\mathbf{r}}_{\mathcal{M}} - \mathbf{r}^{\star}_{\mathcal{M}}\|_2 \le b_{\mathcal{M}}(\delta) := \sqrt{\operatorname{tr}(\Sigma_{\bar{\mathbf{r}},\mathcal{M}})} + \sqrt{2 \|\Sigma_{\bar{\mathbf{r}},\mathcal{M}}\|_{\mathrm{op}} \log(1/\delta)}. \tag{9}$$

If $b_{\mathcal{M}}(\delta) < \|\mathbf{r}^{\star}_{\mathcal{M}}\|_2$, then $\sin\operatorname{angle}(\bar{\mathbf{r}}_{\mathcal{M}}, \mathbf{r}^{\star}_{\mathcal{M}}) \le \frac{b_{\mathcal{M}}(\delta)}{\|\mathbf{r}^{\star}_{\mathcal{M}}\|_2}$.

The concentration proof, cosine consequence, and equal-n finite-sample specialization are in Appendix C.2. Under the equal-size bounded-covariance assumptions, $b_{\mathcal{M}}$, retains the $O((mn)^{-1/2})$ effective-sample-size law. For the discrepancy at a particular target i, let $\rho_i := \left\|\mathbf{f}_i - \bar{\mathbf{f}}\right\|_2$. Then the noise-free markered-residual displacement satisfies

$$\left\|\mathbf{r}^{\star}_{i,\mathcal{M}} - \mathbf{r}^{\star}_{\mathcal{M}}\right\|_2 \le \rho_i \left\|R_{\mathcal{M}} \Delta_Q\right\|_{\mathrm{op}}. \tag{10}$$

Thus m and n_i reduce the population markered-residual estimation error, while $\rho_i \left\|R_{\mathcal{M}} \Delta_Q\right\|_{\mathrm{op}}$ bounds the population-to-target discrepancy and cannot be reduced by a larger cohort. The noisy discrepancy and sample-level angle bounds are given in Appendix C.2.

4.3 Structural bridge from markered residuals to fraction correction

Define noise-free target-level fraction corrections as $\mathbf{d}^{\star}_i := -U^{\dagger} \Delta_P \mathbf{f}_i$ and population fraction correction as $\mathbf{d}^{\star} := -U^{\dagger} \Delta_P \bar{\mathbf{f}}$. Define the coefficient-bias operator $A := -U^{\dagger} \Delta_P \in \mathbb{R}^{p \times p}$, so that $\Delta_P = -UA$ and $\mathbf{d}^{\star} = A\bar{\mathbf{f}}$. Proposition 4.2 shows that $\mathbf{r}^{\star}_{\mathcal{M}}$ does not generally determine $\mathbf{d}^{\star}$: the residual-visible and coefficient-relevant parts of an unrestricted mismatch can vary independently. Therefore, any residual-based correction must exploit structure linking these two effects of the same mismatch. We first construct a practical decoder from markered-residual space to fraction-correction space and then quantify when such a link is present. We construct the map $T(\cdot)$ as

$$T(\mathbf{r}_{\mathcal{M}}) = CDW\mathbf{r}_{\mathcal{M}},$$

where $W \in \mathbb{R}^{p \times q}$ aggregates marker residuals into component-level scores, $D \in \mathbb{R}^{p \times p}$ calibrates component-wise signs and scales, and $C \in \mathbb{R}^{p \times p}$ removes the common all-component direction. Let $\mathbf{u}_{k,S_k} := (U_{\mathcal{M}})_{S_k,k}$. We use

$$(\mathbf{w}_k)_j = \begin{cases} \dfrac{(\mathbf{u}_{k,S_k})_j}{\|\mathbf{u}_{k,S_k}\|_2^2 + \varepsilon_{\mathrm{reg}}}, & j \in S_k, \\ 0, & j \in S_k^c, \end{cases} \qquad W = [\mathbf{w}_1^{\top}; \ldots; \mathbf{w}_p^{\top}]. \tag{11}$$

To make explicit what information is required for such a map to be meaningful, we relate two effects of the shared shift Δ. The operator $A = -U^{\dagger}\Delta_P$ describes how the coefficient-relevant part of the shift induces systematic fraction error, whereas $R_{\mathcal{M}}\Delta_Q$ describes how the residual-visible part of the same shift appears on the selected marker coordinates.

The selected marker subsets make this comparison component-specific. Markers in S_k are enriched in component k and typically have lower reference expression in the other components. Consequently, when a component-related shift leaves a residual-visible effect, the corresponding marker module tends to carry a stronger response than the other modules. This marker structure therefore helps the fraction-relevant and residual-visible effects of the shared shift retain a common component-wise organization. We characterize this structure by

$$V_{\mathcal{M}} := R_{\mathcal{M}} \Delta_Q A^{\dagger}, \qquad E_{\mathcal{M}} := R_{\mathcal{M}} \Delta_Q - V_{\mathcal{M}} A,$$

so that

$$R_{\mathcal{M}} \Delta_Q = V_{\mathcal{M}} A + E_{\mathcal{M}}. \tag{12}$$

Here $V_{\mathcal{M}}, E_{\mathcal{M}} \in \mathbb{R}^{q\times p}$. The map $V_{\mathcal{M}}$ captures the shared component-wise structure between these two effects: in the least-squares sense, it describes how the fraction-error coordinates encoded by A are expressed through the residual-visible marker response $R_{\mathcal{M}}\Delta_Q$. Equivalently, column k of $V_{\mathcal{M}}$ is the marker-space response associated with the kth fraction-error coordinate. An informative marker structure yields stronger intended-module than cross-component responses. The remainder $E_{\mathcal{M}}$ collects the part of the marker-visible response that is not explained by this shared component-wise structure. In particular,

$$E_{\mathcal{M}} = R_{\mathcal{M}}\Delta_Q(I_p - A^{\dagger}A),$$

so the decomposition remains well-defined when A is rank deficient. Marker enrichment supports the component-wise organization represented by $V_{\mathcal{M}}$, while $E_{\mathcal{M}}$ retains residual-visible variation that does not follow it. Define the residual-response coherence error

$$\varepsilon_{\mathrm{rsp}} := \|DWV_{\mathcal{M}} - I_p\|_{\mathrm{op}},$$

and let

$$\tau = \|CDW\|_{\mathrm{op}}, \qquad \varepsilon_{\tan} = \|(I_p - C)\mathbf{d}^{\star}\|_2, \qquad \varepsilon_{\mathrm{rem}} = \|E_{\mathcal{M}}\|_{\mathrm{op}}.$$

Then

$$\|T(\mathbf{r}^{\star}_{\mathcal{M}}) - \mathbf{d}^{\star}\|_2 \le \beta_{\mathcal{M}} := \varepsilon_{\mathrm{rsp}}\|\mathbf{d}^{\star}\|_2 + \varepsilon_{\tan} + \tau\varepsilon_{\mathrm{rem}}. \tag{13}$$

Here $\varepsilon_{\mathrm{rem}}$ measures marker-visible shift not aligned with the fraction-bias coordinates, whereas $\varepsilon_{\mathrm{rsp}}$ measures how accurately DW separates and calibrates the aligned response patterns. Component enrichment supports stronger intended-module than cross-component responses, which Appendix C.3 quantifies through response leakage. Thus marker expression stability controls the stochastic radius $b_{\mathcal{M}}$ in Sec. 4.2, while marker specificity supports structural decoding across components.

4.4 Population fraction-error recovery

Theorem 4.4 (Population recovery of the fraction-error direction)**.** *Let $\widehat{\mathbf{d}} = T(\bar{\mathbf{r}}_{\mathcal{M}})$, and let $V_{\mathcal{M}}$ and $E_{\mathcal{M}}$ be defined by the marker-response decomposition in Eq. (12). Under Theorem 4.3, with probability at least $1-\delta$,*

$$\left\|\widehat{\mathbf{d}} - \mathbf{d}^{\star}\right\|_2 \le b_d(\delta) := \beta_{\mathcal{M}} + \tau b_{\mathcal{M}}(\delta). \tag{14}$$

For target sample i, let $\rho_i = \left\|\mathbf{f}_i - \bar{\mathbf{f}}\right\|_2$ and let $b_{U,i}(\delta_i)$ satisfy $\Pr\{\left\|U^{\dagger}\boldsymbol{\xi}_i\right\|_2 \le b_{U,i}(\delta_i)\} \ge 1-\delta_i$. The target-level radius

$$b_{\mathrm{tot},i} = \beta_{\mathcal{M}} + \tau b_{\mathcal{M}}(\delta) + \rho_i\|A\|_{\mathrm{op}} + b_{U,i}(\delta_i) \tag{15}$$

satisfies

$$\left\|\widehat{\mathbf{d}} - \mathbf{d}_i\right\|_2 \le b_{\mathrm{tot},i}$$

with probability at least $1-\delta-\delta_i$.

Thus $\beta_{\mathcal{M}}$ controls structural marker-bridge error, $\tau b_{\mathcal{M}}(\delta)$ propagates population markered-residual estimation error, $\rho_i\|A\|_{\mathrm{op}}$ controls the population-to-target composition discrepancy, and $b_{U,i}(\delta_i)$ bounds target-specific coefficient noise. Increasing m or n_i reduces the stochastic part of $b_{\mathcal{M}}$, but not the ρ_i-dependent discrepancy terms. The proof is given in Appendix C.4.

4.5 Analytic safety and observable residual empirical step

Define the unit correction direction and analytic margin as

$$\mathbf{v} := \frac{\widehat{\mathbf{d}}}{\left\|\widehat{\mathbf{d}}\right\|_2}, \qquad \gamma_i^{\mathrm{ana}} := \left\|\widehat{\mathbf{d}}\right\|_2 - b_{\mathrm{tot},i}.$$

Theorem 4.5 (Analytic sufficient safe-step condition)**.** *On the event of Theorem 4.4 for which $\left\|\widehat{\mathbf{d}} - \mathbf{d}_i\right\|_2 \le b_{\mathrm{tot},i}$, if $\gamma_i^{\mathrm{ana}} > 0$, then every additive fraction-space step*

$$0 < \lambda_i < 2\gamma_i^{\mathrm{ana}} \tag{16}$$

produces

$$\widetilde{\mathbf{f}}_i = \widehat{\mathbf{f}}_i + \lambda_i\mathbf{v} \quad \textit{with} \quad \left\|\widetilde{\mathbf{f}}_i - \mathbf{f}_i\right\|_2 < \left\|\widehat{\mathbf{f}}_i - \mathbf{f}_i\right\|_2.$$

Because $b_{\mathrm{tot},i}$ contains latent structural quantities such as A, $V_{\mathcal{M}}$, $E_{\mathcal{M}}$, $\mathbf{d}^\star$, and ρ_i, the analytic interval in (16) cannot be evaluated directly from observed data. The recovered direction $\mathbf{v}$ nevertheless determines where the estimate should move; the remaining question is how far to move along that direction. For a candidate magnitude λ, the fraction change $\lambda\mathbf{v}$ is mapped through $U_{\mathcal{M}}$ to the marker response that such an update would explain. A suitable magnitude should therefore produce a forward response whose scale agrees with the observed population markered residual. We compare candidate magnitudes by reference-variance-weighted residual agreement and select the one requiring the least additional rescaling. The full empirical-step construction is given in Appendix C.5.

5 Experiments

We evaluate an algorithm instantiation PopLS (Appendix D) from three complementary perspectives. Section 5.1 first examines the identifiability of population residuals and the accuracy of population-center estimation. Section 5.2 then evaluates whether the estimated residual center provides a reliable correction direction through the proposed bridge mechanism. Section 5.3 finally demonstrates the practical utility of PopLS on pseudo-bulk and real-bulk RNA-seq deconvolution tasks.

Unless otherwise specified, the synthesized experiments conducted in Sections 5.1 and 5.2 use $p = 5$ components, $g = 500$ features, $m = 100$ mixtures, and $n = 500$ observations per mixture. Replicate counts and departures from this default configuration are reported with the corresponding study; curves and summary values are averages over the stated replicates.

5.1 Residual Geometry and Population Markered-Residual Estimation

We first isolate the projection geometry in Theorem 4.1 and Proposition 4.2. We fix the Δ norm at $0.2\|U\|_F$ and vary a parameter $\eta \in \{0, 0.1, \ldots, 1\}$. Increasing η redistributes Δ from the coefficient-relevant component Δ_P to the residual-visible component Δ_Q without changing its total magnitude. Figure 1(a) shows the resulting exchange. Near $\eta = 0$, shift can produce substantial fraction error while leaving almost no reconstruction residual; near $\eta = 1$, it can produce a large residual with almost no fraction error. Thus, the experiment directly verifies the complementary projection effects and shows why residual magnitude alone cannot identify the hidden fraction error (Appendix E.1.1).

Having isolated what is observable, we next test how accurately a cohort size m can estimate the population residual and how closely that center represents an individual target. The concentration study varies m from 2 to 128 and n from 50 to 2,000 observations per mixture, using 20 reference seeds and 20 aggregate repetitions per grid point (Appendix E.1.2). The composition study fixes $n = 2{,}000$, varies $m \in \{4, 8, 16, 32, 64, 128, 256, 512\}$, and scans seven ρ levels from 0 to 0.30, using 20 reference seeds and 10 aggregate repetitions per seed (Appendix E.1.3). Figure 1(b) shows two regimes. Increasing the effective sample size mn reduces stochastic population-estimation error at the $O((mn)^{-1/2})$ rate predicted by Theorem 4.3, whereas nonzero ρ_i leaves a composition-induced discrepancy approaching a nonzero floor bounded by $\rho_i \left\|R_{\mathcal{M}}\Delta_Q\right\|_{\mathrm{op}}$. Extending the cohort-size scan to $m = 512$ further suppresses sampling noise while leaving this deterministic population-to-target discrepancy intact.

5.2 Structural Direction Recovery and Empirical Correction

We test when a well-estimated population residual identifies the fraction-correction direction. The marker experiment varies specificity over $s \in [0, 1]$ and shared-shift strength over $\{0.05, 0.6, 1.0\}$, while fixing the composition radius at 0.12 and unit noise; each configuration uses 40 seeds (Appendix E.2.1). Figure 1(c) shows that a stronger shared signal is insufficient with nonspecific markers. Increasing s reduces cross-component leakage and aligns the recovered and true directions, after which stronger signal improves recovery. Thus, structural identifiability, rather than residual concentration alone, determines direction recovery.

We then perturb marker support W, component correspondence C, and response orientation D in $T = CDW$ (Appendix E.2.3). Together with the specificity and shift scans, these controls expose stable but invalid directions and motivate the observable gate. The gate uses no ground-truth fractions or errors. It combines marker-anchor alignment, population and bootstrap stability, reference-marker and marker-residual enrichment, and residual-derived signal magnitude, retaining the initial estimates

when support is insufficient. Thus concentration alone does not ensure identifiability: a small $b_{\mathcal{M}}$ need not imply a small $\beta_{\mathcal{M}}$; full diagnostics are reported in Appendix E.2.3.

The remaining studies map the target-level radius in Eq. (15) and Theorem 4.4 to finite-sample behavior. Marker specificity and bridge perturbations probe $\beta_{\mathcal{M}}$ (Appendices E.2.1 and E.2.4); scans over m and n characterize $b_{\mathcal{M}}$ and $\tau b_{\mathcal{M}}$ (Appendices E.1.2 and E.1.5); and composition heterogeneity and n_i probe $\rho_i \|A\|_{\mathrm{op}}$ and $b_{U,i}$ (Appendices E.1.3 and E.1.6). A held-out step scan fixes $s = 1$, shared-shift strength 0.8, composition radius 0.12, and unit noise, and varies $\lambda \in \{0, 0.1, \ldots, 2.0\}$ (Appendix E.2.2). Mean MAE is minimized at $\lambda = 0.9$ and Pearson r at $\lambda = 1.0$, after which larger steps overcorrect. Together, these studies separate structural, stochastic, population-to-target, target-specific, and step-size effects (Appendix C.5).

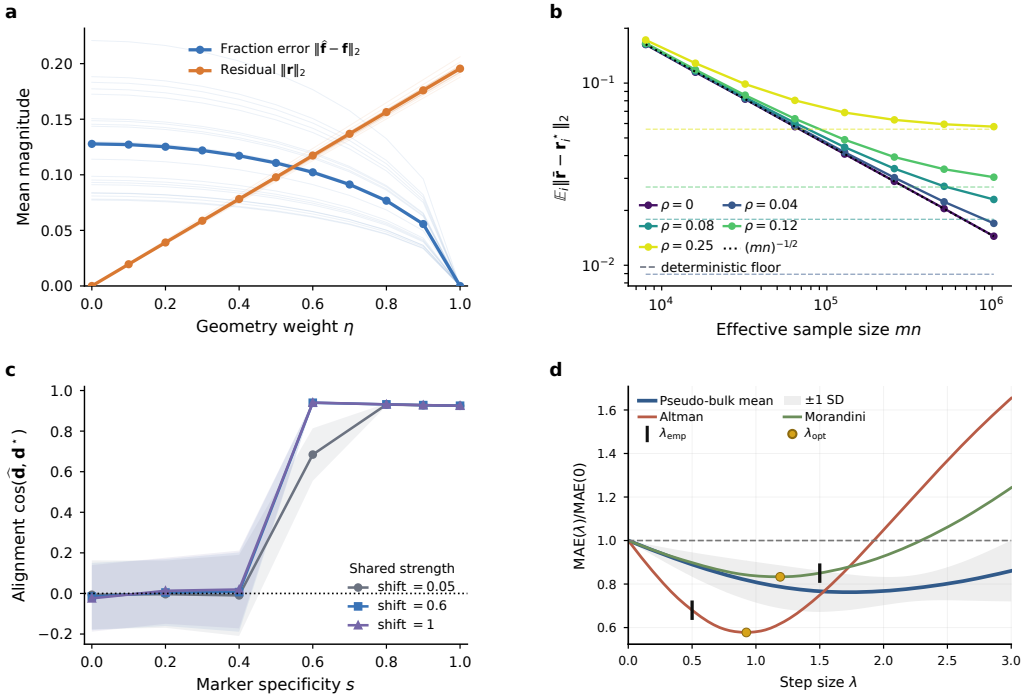

Figure 1: Controlled validation of the theoretical mechanisms and empirical step calibration. (a) Complementary coefficient-relevant and residual-visible effects of reference shift. (b) Increasing effective sample size reduces stochastic population-estimation error, whereas composition heterogeneity leaves a deterministic floor. (c) Component-specific marker structure supports correction-direction recovery. (d) Normalized matrix MAE over $\lambda \in [0, 3]$ for the pseudo-bulk mean and the Altman and Morandini cohorts; markers identify λ_{emp} and the retrospective λ_{opt}.

5.3 Pseudo-Bulk and Real-Bulk RNA-Seq Deconvolution

We next evaluate the full-pipelined PopLS on pseudo-bulk and real-bulk benchmarks. Implementation details and their correspondence to the theory are given in Appendix D. The pseudo-bulk benchmark is derived from eight cancer single-cell RNA-seq cohorts (BRCA, COAD, HNSC, KIRC, LGG, LIHC, LUAD, and STAD). For each cancer, non-overlapping training cells define the reference and held-out cells generate $m = 100$ target mixtures. Mixture fractions are sampled from a weak Dirichlet prior informed by the training-cell type frequencies and are converted into exactly $n_i = 1{,}000$ sampled cells, whose realized fractions serve as the scoring truth. To construct reference shifts resembling those encountered in real deconvolution, we perturb the training-cell expression before constructing U, with gene-wise perturbation magnitudes scaled by empirical coefficient of variation and modulated by an additional 1%–5% direction-preserving fluctuation. PopLS obtains the lowest unweighted mean cancer-level matrix MAE, 0.0199, across the eight evaluated methods and is best on five cancers (Table 1(a); Appendix E.3.1).

Table 1: Dataset-specific matrix MAE on (a) eight cancer pseudo-bulk datasets and (b) two real-bulk cohorts. Small subscripts report $\pm h_{95}$, where h_{95} is the symmetric 95% bootstrap confidence half-width defined in Appendix E.

(a) Eight-cancer pseudo-bulk

Dataset	ReDeconv	Scaden	MuSiC	BayesPrism	TAPE	DECODE	Count Bridges	PopLS (ours)
BRCA	$0.0270_{\pm 0.0022}$	$0.0241_{\pm 0.0014}$	$0.0298_{\pm 0.0028}$	$0.0532_{\pm 0.0036}$	$0.0517_{\pm 0.0035}$	$0.1056_{\pm 0.0043}$	$0.0942_{\pm 0.0059}$	$\mathbf{0.0221}_{\pm 0.0022}$
COAD	$0.0190_{\pm 0.0013}$	$0.0463_{\pm 0.0021}$	$\mathbf{0.0129}_{\pm 0.0013}$	$0.0139_{\pm 0.0008}$	$0.0691_{\pm 0.0022}$	$0.0897_{\pm 0.0015}$	$0.0355_{\pm 0.0028}$	$0.0176_{\pm 0.0014}$
HNSC	$0.0240_{\pm 0.0017}$	$0.0241_{\pm 0.0014}$	$0.0444_{\pm 0.0032}$	$0.0391_{\pm 0.0017}$	$0.0766_{\pm 0.0037}$	$0.1170_{\pm 0.0036}$	$0.0718_{\pm 0.0042}$	$\mathbf{0.0193}_{\pm 0.0015}$
KIRC	$0.0427_{\pm 0.0046}$	$\mathbf{0.0275}_{\pm 0.0018}$	$0.0397_{\pm 0.0045}$	$0.0665_{\pm 0.0027}$	$0.0913_{\pm 0.0074}$	$0.1072_{\pm 0.0060}$	$0.1196_{\pm 0.0072}$	$0.0341_{\pm 0.0036}$
LGG	$0.0281_{\pm 0.0036}$	$0.0200_{\pm 0.0016}$	$0.0229_{\pm 0.0038}$	$0.0432_{\pm 0.0046}$	$0.1703_{\pm 0.0133}$	$0.1852_{\pm 0.0088}$	$0.0625_{\pm 0.0056}$	$\mathbf{0.0193}_{\pm 0.0023}$
LIHC	$0.0208_{\pm 0.0010}$	$0.0245_{\pm 0.0016}$	$0.0357_{\pm 0.0019}$	$0.0769_{\pm 0.0034}$	$0.0642_{\pm 0.0047}$	$0.0913_{\pm 0.0048}$	$0.1362_{\pm 0.0080}$	$\mathbf{0.0150}_{\pm 0.0009}$
LUAD	$0.0214_{\pm 0.0018}$	$0.0255_{\pm 0.0014}$	$\mathbf{0.0102}_{\pm 0.0011}$	$0.0735_{\pm 0.0030}$	$0.0828_{\pm 0.0043}$	$0.1181_{\pm 0.0030}$	$0.0801_{\pm 0.0047}$	$0.0121_{\pm 0.0011}$
STAD	$0.0262_{\pm 0.0021}$	$0.0250_{\pm 0.0015}$	$0.0231_{\pm 0.0014}$	$0.0379_{\pm 0.0020}$	$0.0504_{\pm 0.0026}$	$0.1005_{\pm 0.0029}$	$0.0750_{\pm 0.0052}$	$\mathbf{0.0197}_{\pm 0.0018}$
Mean	0.0262	0.0271	0.0273	0.0505	0.0821	0.1143	0.0844	**0.0199**

(b) Two real-bulk cohorts

Dataset	ReDeconv	Scaden	MuSiC	BayesPrism	TAPE	DECODE	Count Bridges	PopLS (ours)
Altman	$0.1851_{\pm 0.0059}$	$0.3433_{\pm 0.0086}$	$0.2941_{\pm 0.0065}$	$0.2073_{\pm 0.0063}$	$0.3451_{\pm 0.0088}$	$0.2060_{\pm 0.0076}$	$0.2286_{\pm 0.0042}$	$\mathbf{0.0712}_{\pm 0.0049}$
Morandini	$0.0686_{\pm 0.0033}$	$0.0772_{\pm 0.0043}$	$0.0874_{\pm 0.0049}$	$0.1130_{\pm 0.0046}$	$0.1741_{\pm 0.0050}$	$0.0880_{\pm 0.0044}$	$0.0964_{\pm 0.0055}$	$\mathbf{0.0459}_{\pm 0.0037}$
Mean	0.1269	0.2102	0.1907	0.1602	0.2596	0.1470	0.1625	**0.0585**

The real-bulk evaluation provides a complementary end-to-end test in which the reference-to-target discrepancy and the individual theoretical error terms are not experimentally controlled. Altman contains $m = 322$ mixtures over three components and Morandini contains $m = 156$ mixtures over five, both with matched FACS fractions. PopLS obtains matrix MAEs of 0.0712 and 0.0459, respectively, yielding the lowest unweighted two-cohort mean, 0.0585, among the evaluated methods (Table 1(b); Appendix E.3.2).

Lastly, we evaluate empirical step selection over $\lambda \in \{0, 0.05, \ldots, 3.0\}$, with the inputs and recovered direction fixed. PopLS first applies the observable gate described in Section 5.2 and then selects $\widehat{\lambda}_{\mathrm{emp}}$ by matching the forward marker response to the population residual (Appendix C.5). For comparison, the retrospective λ_{opt} minimizes ground-truth MAE over the same grid and is unavailable in practice. Figure 1(d) shows that MAE initially decreases and then increases as larger steps lead to over-correction. The selected steps improve on no correction across all ten benchmarks and retain most of the improvement attained by λ_{opt}; dataset-specific beneficial and 10% near-optimal intervals are reported in Figure 10(d) and Appendices E.3.1–E.3.2.

6 Limitations and Conclusions

Limitations and future work. First, our analytic safe correction algorithm is currently limited to least-squares solvers and conditional Gaussian concentration. Other solvers such as weighted, regularized, probabilistic estimators may induce different coefficient-residual relationships; heavy-tailed or dependent targets also require alternative concentration tools. Second, bias direction recovery requires component-enriched markers and a calibrated response bridge. Third, the optimum correction step λ_{opt} is unobservable in practice as it depends on ground-truth fractions; $\widehat{\lambda}_{\mathrm{emp}}$ therefore serves as an observable proxy without an optimality guarantee. Finally, our empirical algorithm needs robust testing on broader cross-platform and cross-tissue validation.

Conclusions. We study reference-shifted deconvolution that is abstracted from bulk RNA-seq deconvolution. We introduced a post-fit algorithm that maps marker-restricted residuals across related targets to the correction of reference-induced fraction biases. In parallel, we provided analyses of residuals geometry, identifiability, PAC learning effectiveness, and safe correction conditions. In addition, comprehensive synthetic experiments isolatedly studied the effects of different mechanisms on the algorithm and analyses; and a fully-pipelined instantiation PopLS is evaluated on various cancer pseudo-bulk and real-bulk cohorts, verifying the benefits of population-based deconvolution.

AI USE STATEMENT

Generative AI tools were used in this work to assist with research ideation and execution, including refining research questions and methodological considerations, as well as with mathematical reasoning and the development and presentation of proofs. They were also used to support literature retrieval and discovery, including identifying potentially relevant related work, and to assist in drafting, revising, and polishing portions of the manuscript. All AI-assisted mathematical arguments, literature references, methodological decisions, experimental results, and manuscript text were reviewed and verified by the authors. The authors take full responsibility for the correctness, originality, and final content of the work.

ETHICS STATEMENT

This work uses synthetic experiments and previously released transcriptomic datasets to evaluate a computational deconvolution method. We recruited no participants and collected no new biological specimens. The ethics approvals and consent procedures for the biological datasets remain those reported by the source studies, whose publications and accession identifiers are cited in Appendix E. When applied to individual-level biomedical data, transcriptomic deconvolution may reveal sensitive cellular-composition information. Users should therefore comply with applicable access and data-governance requirements, protect participant privacy, and assess whether the reference data adequately represent the target population. PopLS is intended for research use and should not serve as a stand-alone basis for clinical diagnosis or treatment decisions.

REPRODUCIBILITY STATEMENT

We provide the necessary materials to facilitate the reproduction of our theoretical and empirical results. For the theoretical analysis, all formal assumptions, theorem statements, auxiliary lemmas, complete proofs, and derivations are provided in Appendices A–C. For the empirical evaluation, we release the PopLS implementation, experimental scripts, configuration files, and simulated datasets in an anonymous repository: https://anonymous.4open.science/r/AnonCode-46FC. For real bulk and single-cell transcriptomic data, the data sources and preprocessing details, together with additional implementation details, baseline settings, ablations, and robustness analyses, are described in Appendix E. The released code supports reproducing PopLS analyses.

REFERENCES

Philip Amortila, Tongyi Cao, and Akshay Krishnamurthy. Mitigating covariate shift in misspecified regression with applications to reinforcement learning. In *Proceedings of the 37th Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pp. 130–160. PMLR, 2024. URL https://proceedings.mlr.press/v247/amortila24a.html.

Donald W. K. Andrews. Tests for parameter instability and structural change with unknown change point. *Econometrica*, 61(4):821–856, 1993. doi: 10.2307/2951764.

Ruiqi Bai, Yijiao Zhang, Hanbo Yang, and Zhongyi Zhu. Transfer learning for high-dimensional quantile regression with distribution shift. *arXiv preprint arXiv:2411.19933*, 2024. URL https://arxiv.org/abs/2411.19933.

R. L. Brown, J. Durbin, and J. M. Evans. Techniques for testing the constancy of regression relationships over time. *Journal of the Royal Statistical Society: Series B (Methodological)*, 37(2): 149–163, 1975. doi: 10.1111/j.2517-6161.1975.tb01532.x.

Bob Chen, Cherie' R. Scurrah, Eliot T. McKinley, Alan J. Simmons, Marisol A. Ramirez-Solano, Xiangzhu Zhu, Nicholas O. Markham, Cody N. Heiser, Paige N. Vega, Andrea Rolong, Hyeyon Kim, Quanhu Sheng, Julia L. Drewes, Yuan Zhou, Austin N. Southard-Smith, Yanwen Xu, James Ro, Angela L. Jones, Frank Revetta, Lynne D. Berry, Hiroaki Niitsu, Mirazul Islam, Karin Pelka, Matan Hofree, Jonathan H. Chen, Siranush Sarkizova, Kimmie Ng, Marios Giannakis, Genevieve M. Boland, Andrew J. Aguirre, Ana C. Anderson, Orit Rozenblatt-Rosen, Aviv Regev, Nir Hacohen, Kenta Kawasaki, Toshiro Sato, Jeremy A. Goettel, William M. Grady, Wei Zheng, M. Kay Washington, Qiuyin Cai, Cynthia L. Sears, James R. Goldenring, Jeffrey L. Franklin, Timothy Su, Won Jae Huh, Simon Vandekar, Joseph T. Roland, Qi Liu, Robert J. Coffey, Martha J. Shrubsole, and Ken S. Lau. Differential pre-malignant programs and microenvironment chart distinct paths to malignancy in human colorectal polyps. *Cell*, 184(26):6262–6280.e26, 2021. doi: 10.1016/j.cell.2021.11.031.

Yanshuo Chen, Yixuan Wang, Yuelong Chen, Yuqi Cheng, Yumeng Wei, Yunxiang Li, Jiuming Wang, Yingying Wei, Ting-Fung Chan, and Yu Li. Deep autoencoder for interpretable tissue-adaptive deconvolution and cell-type-specific gene analysis. *Nature Communications*, 13(1):6735, 2022. doi: 10.1038/s41467-022-34550-9.

Ji-Hye Choi, Bok-Soon Lee, Jeon Yeob Jang, Yun Sang Lee, Hyo Jeong Kim, Jin Roh, Yoo Seob Shin, Hyun Goo Woo, and Chul-Ho Kim. Single-cell transcriptome profiling of the stepwise progression of head and neck cancer. *Nature Communications*, 14(1):1055, 2023. doi: 10.1038/s41467-023-36691-x.

Tinyi Chu, Zhong Wang, Dana Pe'er, and Charles G. Danko. Cell type and gene expression deconvolution with BayesPrism enables Bayesian integrative analysis across bulk and single-cell RNA sequencing in oncology. *Nature Cancer*, 3(4):505–517, 2022. doi: 10.1038/s43018-022-00356-3.

R. Dennis Cook. Detection of influential observation in linear regression. *Technometrics*, 19(1): 15–18, 1977. doi: 10.1080/00401706.1977.10489493.

Yaoyi Dai, Shuai Guo, Yidan Pan, Carla Castignani, Matthew D. Montierth, Peter Van Loo, and Wenyi Wang. A guide to transcriptomic deconvolution in cancer. *Nature Reviews Cancer*, 26(2): 84–103, 2026. doi: 10.1038/s41568-025-00886-9.

Guillaume Davidson, Alexandra Helleux, Yann A. Vano, Véronique Lindner, Antonin Fattori, Marie Cerciat, Reza T. Elaidi, Virginie Verkarre, Cheng-Ming Sun, Christine Chevreau, Mostefa Bennamoun, Hervé Lang, Thibault Tricard, Wolf H. Fridman, Catherine Sautes-Fridman, Xiaoping Su, Damien Plassard, Celine Keime, Christelle Thibault-Carpentier, Philippe Barthelemy, Stéphane M. Oudard, Irwin Davidson, and Gabriel G. Malouf. Mesenchymal-like tumor cells and myofibroblastic cancer-associated fibroblasts are associated with progression and immunotherapy response of clear cell renal cell carcinoma. *Cancer Research*, 83(17):2952–2969, 2023. doi: 10.1158/0008-5472.can-22-3034.

Kobe De Ridder, Huiwen Che, Kaat Leroy, and Bernard Thienpont. Benchmarking of methods for DNA methylome deconvolution. *Nature Communications*, 15(1):4134, 2024. doi: 10.1038/s41467-024-48466-z.

Alexander Dietrich, Lorenzo Merotto, Konstantin Pelz, Bernhard Eder, Constantin Zackl, Katharina Reinisch, Frank Edenhofer, Federico Marini, Gregor Sturm, Markus List, and Francesca Finotello. omnideconv: a unifying framework for using and benchmarking single-cell-informed deconvolution of bulk RNA-seq data. *Genome Biology*, 27(1):6, 2026. doi: 10.1186/s13059-026-03955-w.

Yaqi Duan and Kaizheng Wang. Adaptive and robust multi-task learning. *The Annals of Statistics*, 51 (5):2015–2039, 2023. doi: 10.1214/23-AOS2319.

Juan Carlos Escanciano. A gaussian process approach to model checks. *The Annals of Statistics*, 52 (5):2456–2481, 2024. doi: 10.1214/24-AOS2443.

Jillian Fisher, Lang Liu, Krishna Pillutla, Yejin Choi, and Zaid Harchaoui. Influence diagnostics under self-concordance. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 10028–10076. PMLR, 2023. URL https://proceedings.mlr.press/v206/fisher23a.html.

Nic Fishman, Gokul Gowri, Tanush Kumar, Jiaqi Lu, Valentin de Bortoli, Jonathan S. Gootenberg, and Omar Abudayyeh. Count Bridges enable modeling and deconvolving transcriptomic data. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=4nOZBufbLC.

Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M. Mauck, Shiwei Zheng, Andrew Butler, Maddie J. Lee, Aaron J. Wilk, Charlotte Darby, Michael Zager, Paul Hoffman, Marlon Stoeckius, Efthymia Papalexi, Eleni P. Mimitou, Jaison Jain, Avi Srivastava, Tim Stuart, Lamar M. Fleming, Bertrand Yeung, Angela J. Rogers, Juliana M. McElrath, Catherine A. Blish, Raphael Gottardo, Peter Smibert, and Rahul Satija. Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–3587.e29, 2021. doi: 10.1016/j.cell.2021.04.048.

Katherine A. Hoadley, Christina Yau, Toshinori Hinoue, Denise M. Wolf, Alexander J. Lazar, et al. Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell*, 173(2):291–304.e6, 2018. doi: 10.1016/j.cell.2018.03.022.

Ziyi Huang, Dehan Cai, and Yanni Sun. Towards more accurate microbial source tracking via non-negative matrix factorization (NMF). *Bioinformatics*, 40(Supplement 1):i68–i78, 2024. doi: 10.1093/bioinformatics/btae227.

Romain Ilbert, Malik Tiomoko, Cosme Louart, Ambroise Odonnat, Vasilii Feofanov, Themis Palpanas, and Ievgen Redko. Analysing multi-task regression via random matrix theory with application to time series forecasting. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-3652.

Jurriaan Janssen, Mischa F. B. Steketee, Aryamaan Bose, Saskia van Asten, Paul P. Eijk, Frederike Dijk, Arantza Farina Sarasqueta, Febe van Maldegem, David P. Noske, Idris Bahce, Jan Koster, Juan J. Garcia Vallejo, Richard Schoonhoven, Mark A. van de Wiel, Teodora Radonic, Bauke Ylstra, and Yongsoo Kim. Statescope: an integrative deconvolution framework for discovering cell states in tumors. *Nature Communications*, 17(1):8465, 2026. doi: 10.1038/s41467-026-74997-8.

Brandon Jew, Marcus Alvarez, Elior Rahmani, Zong Miao, Arthur Ko, Kristina M. Garske, Jae Hoon Sul, Kirsi H. Pietiläinen, Päivi Pajukanta, and Eran Halperin. Accurate estimation of cell composition in bulk expression through robust integration of single-cell information. *Nature Communications*, 11(1):1971, 2020. doi: 10.1038/s41467-020-15816-6.

Nayoung Kim, Hong Kwan Kim, Kyungjong Lee, Yourae Hong, Jong Ho Cho, Jung Won Choi, Jung-Il Lee, Yeon-Lim Suh, Bo Mi Ku, Hye Hyeon Eum, Soyean Choi, Yoon-La Choi, Je-Gun Joung, Woong-Yang Park, Hyun Ae Jung, Jong-Mu Sun, Se-Hoon Lee, Jin Seok Ahn, Keunchil Park, Myung-Ju Ahn, and Hae-Ock Lee. Single-cell RNA sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma. *Nature Communications*, 11(1): 2285, 2020. doi: 10.1038/s41467-020-16164-1.

Seok-Jin Kim. Multi-task linear regression without eigenvalue lower bounds: Adaptivity, robustness and safety. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*. PMLR, 2026. URL https://arxiv.org/abs/2605.17126.

Parker Knight and Rui Duan. Multi-task learning with summary statistics. In *Advances in Neural Information Processing Systems*, volume 36, pp. 54020–54031, 2023. doi: 10.52202/075280-2350.

Vikrant Kumar, Kalpana Ramnarayanan, Raghav Sundar, Nisha Padmanabhan, Supriya Srivastava, Mayu Koiwa, Tadahito Yasuda, Vivien Koh, Kie Kyon Huang, Su Ting Tay, Shamaine Wei Ting Ho, Angie Lay Keng Tan, Takatsugu Ishimoto, Guowei Kim, Asim Shabbir, Qingfeng Chen, Biyan Zhang, Shengli Xu, Kong-Peng Lam, Huey Yew Jeffrey Lum, Ming Teh, Wei Peng Yong, Jimmy Bok Yan So, and Patrick Tan. Single-cell atlas of lineage states, tumor microenvironment, and subtype-specific expression programs in gastric cancer. *Cancer Discovery*, 12(3):670–691, 2022. doi: 10.1158/2159-8290.cd-21-0683.

Hyukdong Kwon, Jihnhee Yu, and Mingliang Li. Identifying outlying groups through residual analysis and its application to healthcare expenditure. *Journal of Applied Statistics*, 52(15): 2777–2798, 2025. doi: 10.1080/02664763.2025.2484599.

Rachel Lau, Cankut Çubuk, Athina Spiliopoulou, Pedro Martínez-Paz, Anna E. A. Surace, Liliane Fossati-Jimack, Soumya Raychaudhuri, Costantino Pitzalis, and Myles J. Lewis. cellGeometry: ultra-fast single-cell deconvolution of bulk RNA-Seq using a geometric solution. *Nature Communications*, 17(1):8995, 2026. doi: 10.1038/s41467-026-75762-7.

Lingzhu Li, Xuehu Zhu, and Lixing Zhu. Adaptive-to-model hybrid of tests for regressions. *Journal of the American Statistical Association*, 118(541):514–523, 2023. doi: 10.1080/01621459.2021.1941052.

Dungang Liu, Zewei Lin, and Heping Zhang. A unified framework for residual diagnostics in generalized linear models and beyond. *Journal of the American Statistical Association*, 120(551): 1840–1852, 2025. doi: 10.1080/01621459.2025.2504037.

Songjian Lu, Jiyuan Yang, Lei Yan, Jingjing Liu, Judy Jiaru Wang, Rhea Jain, and Jiyang Yu. Transcriptome size matters for single-cell RNA-seq normalization and bulk deconvolution. *Nature Communications*, 16(1):1246, 2025. doi: 10.1038/s41467-025-56623-1.

Yiming Lu, Aiqing Yang, Cheng Quan, Yingwei Pan, Haoyun Zhang, Yuanfeng Li, Chengming Gao, Hao Lu, Xueting Wang, Pengbo Cao, Hongxia Chen, Shichun Lu, and Gangqiao Zhou. A single-cell atlas of the multicellular ecosystem of primary and metastatic hepatocellular carcinoma. *Nature Communications*, 13(1):4594, 2022. doi: 10.1038/s41467-022-32283-3.

Davide Maran and Csaba Szepesvári. Beyond least squares: Uniform approximation and the hidden cost of misspecification. In *Advances in Neural Information Processing Systems*, volume 38, pp. 168329–168377. Curran Associates, Inc., 2025. doi: 10.52202/085713-5611.

Kevin Menden, Matteo Marouf, Sergio Oller, Anupriya Dalmia, Daniel Sumner Magruder, Karin Kloiber, Peter Heutink, and Stefan Bonn. Deep learning–based cell composition analysis from tissue expression profiles. *Science Advances*, 6(30):eaba2619, 2020. doi: 10.1126/sciadv.aba2619.

Achille O. R. Nazaret, Claudia Shi, and David Blei. On the misspecification of linear assumptions in synthetic controls. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 3790–3798. PMLR, 2024. URL https://proceedings.mlr.press/v238/nazaret24a.html.

Aaron M. Newman, Chloé B. Steen, Chih Long Liu, Andrew J. Gentles, Aadel A. Chaudhuri, Florian Scherer, Michael S. Khodadoust, Mohammad S. Esfahani, Bogdan A. Luca, David Steiner, Maximilian Diehn, and Ash A. Alizadeh. Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nature Biotechnology*, 37(7):773–782, 2019. doi: 10.1038/s41587-019-0114-2.

Parham Rezaei, Filip Kovačević, Francesco Locatello, and Marco Mondelli. High-dimensional analysis of synthetic data selection. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=Y54P2BBPPh.

Takeyuki Sasai and Hironori Fujisawa. Outlier robust and sparse estimation of linear regression coefficients. *Journal of Machine Learning Research*, 26(93):1–79, 2025. URL https://jmlr.org/papers/v26/23-1583.html.

Tabula Sapiens Consortium. The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science*, 376(6594):eabl4896, 2022. doi: 10.1126/science.abl4896.

Daphne Tsoucas, Rui Dong, Haide Chen, Qian Zhu, Guoji Guo, and Guo-Cheng Yuan. Accurate estimation of cell-type composition from gene expression data. *Nature Communications*, 10(1): 2975, 2019. doi: 10.1038/s41467-019-10802-z.

Xuran Wang, Jihwan Park, Katalin Susztak, Nancy R. Zhang, and Mingyao Li. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nature Communications*, 10(1): 380, 2019. doi: 10.1038/s41467-018-08023-x.

Yunjuan Wang and Raman Arora. On the stability and generalization of meta-learning. In *Advances in Neural Information Processing Systems*, volume 37, pp. 83665–83710, 2024. doi: 10.52202/079017-2661.

Sunny Z. Wu, Ghamdan Al-Eryani, Daniel Lee Roden, Simon Junankar, Kate Harvey, Alma Andersson, Aatish Thennavan, Chenfei Wang, James R. Torpy, Nenad Bartonicek, Taopeng Wang, Ludvig Larsson, Dominik Kaczorowski, Neil I. Weisenfeld, Cedric R. Uytingco, Jennifer G. Chew, Zachary W. Bent, Chia-Ling Chan, Vikkitharan Gnanasambandapillai, Charles-Antoine Dutertre, Laurence Gluch, Mun N. Hui, Jane Beith, Andrew Parker, Elizabeth Robbins, Davendra Segara, Caroline Cooper, Cindy Mak, Belinda Chan, Sanjay Warrier, Florent Ginhoux, Ewan Millar, Joseph E. Powell, Stephen R. Williams, X. Shirley Liu, Sandra O'Toole, Elgene Lim, Joakim Lundeberg, Charles M. Perou, and Alexander Swarbrick. A single-cell and spatially resolved atlas of human breast cancers. *Nature Genetics*, 53(9):1334–1347, 2021. doi: 10.1038/s41588-021-00911-1.

Shadi Zahedi, Kent Riemondy, Tian Liu, Andrea M. Griesinger, Andrew M. Donson, April A. Apfelbaum, Rui Fu, Julian Grandvallet Contreras, Michele Crespo, John DeSisto, Madeline M. Groat, Emil Bratbak, Adam Green, Todd C. Hankinson, Michael Handler, Rajeev Vibhakar, Nicholas Willard, Nicholas K. Foreman, Tzu Phang, and Jean Mulcahy Levy. Multi-pronged analysis of pediatric low-grade glioma and ganglioglioma reveals a unique tumor microenvironment associated with BRAF alterations. *Brain Pathology*, 35(6):e70023, 2025. doi: 10.1111/bpa.70023.

Zhenyu Zhang, Kyle Hernandez, Jeremiah Savage, Shenglai Li, Dan Miller, Stuti Agrawal, Francisco Ortuno, Louis M. Staudt, Allison Heath, and Robert L. Grossman. Uniform genomic data analysis in the NCI genomic data commons. *Nature Communications*, 12(1):1226, 2021. doi: 10.1038/s41467-021-21254-9.

Tianyi Zhao, Renjie Liu, Yuzhi Sun, Bingtian Wang, Liyuan Zhang, Qiuhao Chen, Ruibang Luo, Zhiyuan Yuan, Guohua Wang, Liang Cheng, and Yadong Wang. DECODE: deep learning-based common deconvolution framework for various omics data. *Nature Methods*, 23(3):596–608, 2026. doi: 10.1038/s41592-026-03007-y.

Ingvar Ziemann, Stephen Tu, George J. Pappas, and Nikolai Matni. The noise level in linear regression with dependent data. In *Advances in Neural Information Processing Systems*, volume 36, pp. 74903–74920. Curran Associates, Inc., 2023. doi: 10.52202/075280-3273.

A Notation and Assumption Details

Table 2: Core notation used throughout the analysis.

Symbol	Meaning	Symbol	Meaning
Problem quantities			
p, g	Numbers of components and observed features.	$\mathcal{F}, \mathcal{T}$	Composition simplex and its tangent space.
$\mathbf{y}_i$	Aggregate observation; bulk expression vector.	U	Full-column-rank reference operator.
$\mathbf{f}_i$	Unknown true composition vector.	$\widehat{\mathbf{f}}_i$	Initial reference-based least-squares estimate.
m	Number of related target mixtures.	n_i	Elementary units forming aggregate i.
Mismatch and finite-aggregate noise			
M_i	Target-domain component population-mean operator.	Δ	Cohort-shared reference-to-target mismatch.
H_i	Sample-specific mismatch.	$\boldsymbol{\xi}_i$	Finite-aggregate sampling fluctuation.
$\Omega(\mathbf{f}_i)$	Scaled conditional finite-aggregate covariance.	Δ_P, Δ_Q	Column-space and orthogonal parts of Δ.
Projection and residual quantities			
P, Q	$UU^\dagger$ and $I_g - UU^\dagger$.	$\mathbf{r}_i$	Full post-fit residual used for geometry.
$\mathcal{M}, q, R_\mathcal{M}$	Marker union, its size, and row selector.	$U_\mathcal{M}$	Marker-row subset $R_\mathcal{M}U$.
$\mathbf{r}_{i,\mathcal{M}}, \mathbf{r}^\star_{i,\mathcal{M}}$	Observed and noise-free marker residuals.	$\bar{\mathbf{r}}_\mathcal{M}, \mathbf{r}^\star_\mathcal{M}$	Population marker-residual estimate and target.
$\Sigma_{\bar{\mathbf{r}},\mathcal{M}}$	Covariance of $\bar{\mathbf{r}}_\mathcal{M}$; $\sigma^2_\mathcal{M}$ is the uniform marker-residual variance proxy.	$b_\mathcal{M}$	Marker-population concentration radius.
Fraction correction and population discrepancy			
$\mathbf{d}_i, \mathbf{d}^\star_i$	Actual and noise-free sample corrections.	$\mathbf{d}^\star$	Population correction target.
$\bar{\mathbf{f}}$	Mean cohort composition.	ρ_i	Composition distance $\left\|\mathbf{f}_i - \bar{\mathbf{f}}\right\|_2$.
$\widehat{\mathbf{d}}, \mathbf{v}$	Mapped population correction and its unit direction.	$\widetilde{\mathbf{f}}_i$	Corrected composition.
Residual-to-fraction bridge			
$W, \varepsilon_{\mathrm{reg}}$	Marker-residual aggregator and ridge stabilizer.	D	Diagonal sign-and-scale calibration.
A	Coefficient-bias operator $-U^\dagger\Delta_P$.	$V_\mathcal{M}, E_\mathcal{M}$	Marker-space response patterns and remainder.
C	Simplex tangent-space centering matrix.	T	Map $T(\mathbf{r}_\mathcal{M}) = CDW\mathbf{r}_\mathcal{M}$.
$\varepsilon_{\mathrm{rsp}}$	Residual-response coherence error.	$\beta_\mathcal{M}$	Structural marker-bridge error.
τ	$\left\|CDW\right\|_{\mathrm{op}}$.	$\varepsilon_{\tan}, \varepsilon_{\mathrm{rem}}$	Tangent and remainder errors.
Uncertainty and correction			
$b_\mathcal{M}, b_d$	Marker-residual and mapped-direction radii.	$b_{U,i}$	Target coefficient-noise radius.
$b_{\mathrm{tot},i}$	Target-level total error radius.	γ_i^{ana}	Latent analytic directional margin.
λ_i	Additive fraction-space step.	Λ	Finite candidate set for empirical step selection.
$\beta(\lambda)$	Empirical residual-response scale.	G_λ	Marker variance-weighting matrix.

B AGGREGATE AND RESIDUAL MODELS FROM IID WITHIN-COMPONENT SAMPLING

We derive the aggregate-level model in Section 3 from conditionally i.i.d. sampling within each component and establish the conditional Gaussian approximation used in Theorem 4.3. Aggregate i contains n_i elementary units, with realized component counts n_{ik} and fractions $f_{ik} = n_{ik}/n_i$. The ℓth unit from component k is $\mathbf{x}_{ik\ell}^{\mathrm{agg}}$, and the pooled observation is $\mathbf{y}_i = n_i^{-1} \sum_k \sum_{\ell=1}^{n_{ik}} \mathbf{x}_{ik\ell}^{\mathrm{agg}}$. Reference column $\mathbf{u}_k$ is likewise the average of n_k^{ref} elementary reference observations. The derivation below gives the induced conditional means and covariances, finite-aggregate fluctuation, reference-side sampling contribution, and population-residual distribution. For aggregate i and component k, assume the elementary vectors $\mathbf{x}_{ik\ell}^{\mathrm{agg}}$ are i.i.d. conditional on the component label, independent across represented component samples, and satisfy

$$\mathbb{E}[\mathbf{x}_{ik\ell}^{\mathrm{agg}}] = \boldsymbol{\mu}_{\mathrm{agg},k}^{(i)}, \qquad \mathrm{Cov}(\mathbf{x}_{ik\ell}^{\mathrm{agg}}) = \Sigma_{\mathrm{agg},ik},$$

with finite second moments and a Lindeberg condition. Let M_i stack these component population means, with column k equal to $\boldsymbol{\mu}_{\mathrm{agg},k}^{(i)}$. Thus M_i is fixed at the target-population level, whereas finite-aggregate randomness arises from sampling elementary observations around these means. Crucially, the elementary observations themselves need not be Gaussian. They may be discrete, skewed, multimodal, or otherwise non-Gaussian; their distribution affects the component means, covariances, and finite-sample approximation error, but not the Gaussian form of the large-aggregate limit. Under the stated independence and Lindeberg/finite-variance conditions, the multivariate CLT applies to the centered sums. Thus the approximate Gaussianity used below is a consequence of averaging sufficiently many elementary units within each represented component, not an assumption on the distribution of a single unit. If the individual distributions have infinite variance or dependence strong enough to invalidate the CLT, the Gaussian theorem does not apply without a different limit or robust concentration argument. Conditional on the realized component counts, the aggregate is

$$\mathbf{y}_i = \frac{1}{n_i} \sum_{k=1}^{p} \sum_{\ell=1}^{n_{ik}} \mathbf{x}_{ik\ell}^{\mathrm{agg}}, \qquad \mathbb{E}[\mathbf{y}_i \mid n_{i1}, \ldots, n_{ip}] = \sum_{k=1}^{p} f_{ik} \boldsymbol{\mu}_{\mathrm{agg},k}^{(i)} = M_i \mathbf{f}_i.$$

Thus the linear combination of component means is the conditional expectation of the aggregate, not an assumption that every sampled unit equals its component mean. Regrouping the centered individual observations gives

$$\mathbf{y}_i - M_i \mathbf{f}_i = \frac{1}{n_i} \sum_{k=1}^{p} \sum_{\ell=1}^{n_{ik}} (\mathbf{x}_{ik\ell}^{\mathrm{agg}} - \boldsymbol{\mu}_{\mathrm{agg},k}^{(i)}) = \sum_{k=1}^{p} f_{ik} \boldsymbol{\varepsilon}_{ik}^{\mathrm{agg}},$$

where

$$\boldsymbol{\varepsilon}_{ik}^{\mathrm{agg}} := \frac{1}{n_{ik}} \sum_{\ell=1}^{n_{ik}} (\mathbf{x}_{ik\ell}^{\mathrm{agg}} - \boldsymbol{\mu}_{\mathrm{agg},k}^{(i)}).$$

For each represented component, assume $n_{ik} \to \infty$. The weak law of large numbers yields $\boldsymbol{\varepsilon}_{ik}^{\mathrm{agg}} \xrightarrow{p} 0$, and hence $\mathbf{y}_i \xrightarrow{p} M_i \mathbf{f}_i$. The Gaussian statement follows from the multivariate CLT. By the Cramér-Wold device,

$$\sqrt{n_{ik}}\, \boldsymbol{\varepsilon}_{ik}^{\mathrm{agg}} \Rightarrow \mathcal{N}(0, \Sigma_{\mathrm{agg},ik}).$$

A fixed weighted sum of these component-wise Gaussian limits is Gaussian. Under conditional independence across component samples,

$$\mathbf{y}_i \mid \mathbf{f}_i \dot{\sim} \mathcal{N}\left(M_i \mathbf{f}_i, \sum_{k=1}^{p} \frac{f_{ik}^2}{n_{ik}} \Sigma_{\mathrm{agg},ik} \right).$$

If $n_{ik} = n_i f_{ik} + o(n_i)$, then

$$\sum_{k=1}^{p} \frac{f_{ik}^2}{n_{ik}} \Sigma_{\mathrm{agg},ik} = \frac{1}{n_i} \sum_{k: f_{ik} > 0} f_{ik} \Sigma_{\mathrm{agg},ik} + o(n_i^{-1}).$$

This gives the $n_i^{-1/2}$ single-aggregate random scale. The requirement that every represented component contain sufficiently many sampled units is important: a large total n_i alone does not guarantee the component-wise CLT when some n_{ik} remain bounded.

The reference side follows the same mechanism. Let

$$\mathbf{u}_k = \frac{1}{n_k^{\mathrm{ref}}} \sum_{\ell=1}^{n_k^{\mathrm{ref}}} \mathbf{x}_{k\ell}^{\mathrm{ref}}, \qquad \mathbb{E}[\mathbf{x}_{k\ell}^{\mathrm{ref}}] = \boldsymbol{\mu}_{\mathrm{ref},k}^0, \qquad \mathrm{Cov}(\mathbf{x}_{k\ell}^{\mathrm{ref}}) = \Sigma_{\mathrm{ref},k}.$$

The LLN gives $\mathbf{u}_k \xrightarrow{p} \boldsymbol{\mu}_{\mathrm{ref},k}^0$ and the CLT gives

$$\mathbf{u}_k = \boldsymbol{\mu}_{\mathrm{ref},k}^0 + \boldsymbol{\varepsilon}_k^{\mathrm{ref}}, \qquad \boldsymbol{\varepsilon}_k^{\mathrm{ref}} \mathrel{\dot\sim} \mathcal{N}\left(0, \frac{\Sigma_{\mathrm{ref},k}}{n_k^{\mathrm{ref}}}\right).$$

For a fixed composition $\mathbf{f}_i$,

$$U\mathbf{f}_i = M_{\mathrm{ref}}^0 \mathbf{f}_i + \sum_{k=1}^{p} f_{ik} \boldsymbol{\varepsilon}_k^{\mathrm{ref}}.$$

Multiplication by the fixed fractions f_{ik} and summation over k are linear operations. Hence $U\mathbf{f}_i$ is Gaussian in the joint CLT limit; under independent reference sampling across components,

$$U\mathbf{f}_i \mid \mathbf{f}_i \mathrel{\dot\sim} \mathcal{N}\left(M_{\mathrm{ref}}^0 \mathbf{f}_i, \sum_{k=1}^{p} \frac{f_{ik}^2}{n_k^{\mathrm{ref}}} \Sigma_{\mathrm{ref},k}\right),$$

We next combine this reference-side limit with aggregate-side fluctuation. Correlated reference component means give the same joint-CLT conclusion after adding the corresponding covariance cross terms.

Finally, write the aggregate and reference fluctuations as

$$\boldsymbol{\varepsilon}_i^{\mathrm{agg}} = \mathbf{y}_i - M_i \mathbf{f}_i, \qquad \boldsymbol{\varepsilon}_i^{\mathrm{ref}} = U\mathbf{f}_i - M_{\mathrm{ref}}^0 \mathbf{f}_i.$$

If the aggregate and reference datasets are sampled independently, these are independent Gaussian vectors in the CLT limit. Therefore

$$\begin{aligned} \mathbf{y}_i - U\mathbf{f}_i &= \{M_i - M_{\mathrm{ref}}^0\}\mathbf{f}_i + \boldsymbol{\varepsilon}_i^{\mathrm{agg}} - \boldsymbol{\varepsilon}_i^{\mathrm{ref}}, \\ \boldsymbol{\varepsilon}_i^{\mathrm{agg}} - \boldsymbol{\varepsilon}_i^{\mathrm{ref}} &\mathrel{\dot\sim} \mathcal{N}\left(0, \sum_{k=1}^{p} \frac{f_{ik}^2}{n_{ik}} \Sigma_{\mathrm{agg},ik} + \sum_{k=1}^{p} \frac{f_{ik}^2}{n_k^{\mathrm{ref}}} \Sigma_{\mathrm{ref},k}\right). \end{aligned}$$

Combining the aggregate- and reference-side fluctuations gives the Gaussian approximation for the aggregate-reference discrepancy:

$$\mathbf{y}_i - U\mathbf{f}_i \mid \mathbf{f}_i \mathrel{\dot\sim} \mathcal{N}\left((M_i - M_{\mathrm{ref}}^0)\mathbf{f}_i, \sum_{k=1}^{p} \frac{f_{ik}^2}{n_{ik}} \Sigma_{\mathrm{agg},ik} + \sum_{k=1}^{p} \frac{f_{ik}^2}{n_k^{\mathrm{ref}}} \Sigma_{\mathrm{ref},k}\right).$$

The subtraction introduces no new approximation: the difference of independent Gaussian vectors is Gaussian, with covariances adding because of independence.

The inverse problem then conditions on the supplied reference matrix U. Once U is observed, $U - M_{\mathrm{ref}}^0$ is fixed and is absorbed into the effective mismatch Δ defined in Section 3. Conditioning on the realized reference therefore yields the finite-aggregate noise model used in the main analysis:

$$\boldsymbol{\xi}_i \mid \mathbf{f}_i \sim \mathcal{N}\left(0, \frac{1}{n_i}\Omega(\mathbf{f}_i)\right), \tag{17}$$

where the exact scaled conditional aggregate covariance is

$$\Omega(\mathbf{f}_i) = n_i \sum_{k:f_{ik}>0} \frac{f_{ik}^2}{n_{ik}} \Sigma_{\mathrm{agg},ik}.$$

Here $\Sigma_{\mathrm{agg},ik}$ is the elementary-observation covariance for component k in target aggregate i; the composition-dependent weights arise from the realized component sampling fractions. Along proportional-sampling sequences $n_{ik} = n_i f_{ik} + o(n_i)$,

$$\Omega(\mathbf{f}_i) = \sum_{k:f_{ik}>0} f_{ik}\Sigma_{\mathrm{agg},ik} + o(1).$$

The finite-n_i definition above is the one that makes (17) equal to the exact conditional covariance before the Gaussian approximation. Conditioning on the observed U absorbs the realized reference-side fluctuation into the effective mismatch Δ, so $\Omega(\mathbf{f}_i)$ in Theorem 4.3 represents only the remaining aggregate-side finite-sampling covariance. Applying the deterministic residual projector Q gives the conditional distribution of the full reconstruction residual:

$$\mathbf{r}_i \mid \mathbf{f}_i \sim \mathcal{N}\left(Q(\Delta + H_i)\mathbf{f}_i, \frac{1}{n_i}Q\Omega(\mathbf{f}_i)Q\right).$$

Conditioning also on the reference-derived marker selector $R_{\mathcal{M}}$ and applying this fixed linear map gives

$$\mathbf{r}_{i,\mathcal{M}} \mid \mathbf{f}_i \sim \mathcal{N}\left(R_{\mathcal{M}}Q(\Delta + H_i)\mathbf{f}_i, \frac{1}{n_i}R_{\mathcal{M}}Q\Omega(\mathbf{f}_i)QR_{\mathcal{M}}^\top\right).$$

Under $H_i = 0$ and conditional independence across target mixtures, averaging these marker noises yields the distribution used in Theorem 4.3:

$$\bar{\mathbf{r}}_{\mathcal{M}} \mid \mathbf{f}_{1:m} \sim \mathcal{N}(\mathbf{r}^\star_{\mathcal{M}}, \Sigma_{\bar{\mathbf{r}},\mathcal{M}}), \qquad \Sigma_{\bar{\mathbf{r}},\mathcal{M}} = \frac{1}{m^2}\sum_{i=1}^{m}\frac{1}{n_i}R_{\mathcal{M}}Q\Omega(\mathbf{f}_i)QR_{\mathcal{M}}^\top.$$

The conditional Gaussian aggregate model used in Theorem 4.3 is supported by the multivariate CLT under conditionally i.i.d. within-component elementary sampling. At the aggregation scales studied here, including exactly 1,000 sampled raw-count cells per pseudo-bulk, the approximation error for the directional and concentration events of interest is small relative to the working δ. It therefore does not materially alter the reported coverage interpretation. Real bulk measurements likewise aggregate many elementary molecular or cellular contributions, without assigning an unobserved cell count. The Gaussian-coverage simulation in Appendix E.1.5 is consistent with this approximation regime.

C Proofs and Auxiliary Results

C.1 Projection geometry and residual-only non-identifiability

Proof of Theorem 4.1. Since $U \in \mathbb{R}^{g\times p}$ has full column rank,

$$U^\dagger = (U^\top U)^{-1}U^\top, \qquad U^\dagger U = I_p.$$

Under the Section 4 model, substituting $\mathbf{y}_i = (U+\Delta)\mathbf{f}_i + \boldsymbol{\xi}_i$ into the least-squares estimator gives

$$\begin{aligned}\widehat{\mathbf{f}}_i &= U^\dagger \mathbf{y}_i \\ &= U^\dagger\{(U+\Delta)\mathbf{f}_i + \boldsymbol{\xi}_i\} \\ &= \mathbf{f}_i + U^\dagger\Delta\mathbf{f}_i + U^\dagger\boldsymbol{\xi}_i.\end{aligned}$$

For the residual,

$$\begin{aligned}\mathbf{r}_i &= \mathbf{y}_i - U\widehat{\mathbf{f}}_i = Q\mathbf{y}_i \\ &= Q\{(U+\Delta)\mathbf{f}_i + \boldsymbol{\xi}_i\} = Q\Delta\mathbf{f}_i + Q\boldsymbol{\xi}_i,\end{aligned}$$

because $QU = U - UU^\dagger U = 0$.

It remains to verify the reference geometry. With full column rank,

$$P^\top = \{U(U^\top U)^{-1}U^\top\}^\top = P, \qquad P^2 = UU^\dagger UU^\dagger = P,$$

so $P = UU^\dagger$ is the orthogonal projector onto the column space of U and $Q = I_g - P$ is the complementary orthogonal projector. Moreover

$$U^\dagger P = U^\dagger UU^\dagger = U^\dagger, \qquad U^\dagger Q = 0.$$

Therefore

$$U^\dagger\Delta = U^\dagger(P+Q)\Delta = U^\dagger\Delta_P, \qquad Q\Delta = \Delta_Q,$$

which proves (3)-(4).

General heterogeneous identities. For the general aggregate model in Section 3, before the Section 4 specialization, the same least-squares algebra gives

$$\widehat{\mathbf{f}}_i - \mathbf{f}_i = U^\dagger(\Delta + H_i)\mathbf{f}_i + U^\dagger \boldsymbol{\xi}_i, \qquad \mathbf{r}_i = Q(\Delta + H_i)\mathbf{f}_i + Q\boldsymbol{\xi}_i.$$

These identities are used in the diagnostic decomposition below and in the heterogeneous-shift extension of Appendix C.4.

Proof of Proposition 4.2. Let $B \in \mathbb{R}^{g\times p}$ be arbitrary and define the column-space mismatch perturbation

$$\widetilde{\Delta}_P := PB.$$

Because $QP = 0$,

$$Q\widetilde{\Delta}_P = QPB = 0.$$

At the same time,

$$U^\dagger \widetilde{\Delta}_P = U^\dagger PB = U^\dagger B,$$

which can be nonzero for a suitable B. Replacing Δ by $\Delta + \widetilde{\Delta}_P$ therefore changes the coefficient term by $U^\dagger \widetilde{\Delta}_P \mathbf{f}_i$ while leaving every noiseless residual unchanged. Even an infinite noiseless residual population cannot distinguish these two shifts.

Conversely, define the orthogonal mismatch perturbation

$$\widetilde{\Delta}_Q := QB.$$

Since $U^\dagger Q = 0$,

$$U^\dagger \widetilde{\Delta}_Q = U^\dagger QB = 0,$$

whereas

$$Q\widetilde{\Delta}_Q = Q^2 B = QB$$

can be nonzero. Replacing Δ by $\Delta + \widetilde{\Delta}_Q$ therefore changes the residual term by $Q\widetilde{\Delta}_Q \mathbf{f}_i$ while leaving the unconstrained least-squares estimate unchanged. The two constructions show that neither residual magnitude nor residual direction identifies the fraction error on the unrestricted shift class.

Proposition 4.2 therefore establishes impossibility on the unrestricted mismatch class. Section 4.3 restores recoverability only after adding a marker-space relation between the actually used residual coordinates $R_{\mathcal{M}}\Delta_Q$ and the coefficient-bias operator $A = -U^\dagger \Delta_P$; selecting specific or stable markers by itself is not sufficient.

C.2 Marker-residual concentration and population-to-target discrepancy

Proof of Theorem 4.3. From Eq. (6),

$$\bar{\mathbf{r}}_{\mathcal{M}} - \mathbf{r}^\star_{\mathcal{M}} = \frac{1}{m}\sum_{i=1}^{m} R_{\mathcal{M}} Q \boldsymbol{\xi}_i.$$

Conditional on $\mathbf{f}_{1:m}$, the observed reference, and its fixed marker selector, each term is Gaussian with zero mean and

$$\mathrm{Cov}(R_{\mathcal{M}} Q \boldsymbol{\xi}_i \mid \mathbf{f}_i) = \frac{1}{n_i} R_{\mathcal{M}} Q \Omega(\mathbf{f}_i) Q R_{\mathcal{M}}^\top.$$

Conditional independence therefore gives

$$\bar{\mathbf{r}}_{\mathcal{M}} - \mathbf{r}^\star_{\mathcal{M}} \sim \mathcal{N}(0, \Sigma_{\bar{\mathbf{r}},\mathcal{M}}).$$

Let $\boldsymbol{\zeta} \sim \mathcal{N}(0, \Sigma_{\bar{\mathbf{r}},\mathcal{M}})$ and write $\boldsymbol{\zeta} = \Sigma^{1/2}_{\bar{\mathbf{r}},\mathcal{M}} \mathbf{g}$ with $\mathbf{g} \sim \mathcal{N}(0, I)$. The map $\mathbf{g} \mapsto \left\| \Sigma^{1/2}_{\bar{\mathbf{r}},\mathcal{M}} \mathbf{g} \right\|_2$ is $\sqrt{\|\Sigma_{\bar{\mathbf{r}},\mathcal{M}}\|_{\mathrm{op}}}$-Lipschitz, and Jensen's inequality gives

$$\mathbb{E}\,\|\boldsymbol{\zeta}\|_2 \le \sqrt{\mathrm{tr}(\Sigma_{\bar{\mathbf{r}},\mathcal{M}})}.$$

Gaussian concentration therefore yields

$$\Pr\left\{\|\zeta\|_2 \geq \sqrt{\operatorname{tr}(\Sigma_{\bar{\mathbf{r}},\mathcal{M}})} + \sqrt{2\,\|\Sigma_{\bar{\mathbf{r}},\mathcal{M}}\|_{\mathrm{op}}\, t}\right\} \leq e^{-t}.$$

Taking $t = \log(1/\delta)$ proves (9).

For the angle, let $\theta = \operatorname{angle}(\bar{\mathbf{r}}_{\mathcal{M}}, \mathbf{r}^\star_{\mathcal{M}})$. If $b_{\mathcal{M}} < \|\mathbf{r}^\star_{\mathcal{M}}\|_2$, then on the PAC event

$$\langle \bar{\mathbf{r}}_{\mathcal{M}}, \mathbf{r}^\star_{\mathcal{M}} \rangle \geq \|\mathbf{r}^\star_{\mathcal{M}}\|_2^2 - \|\bar{\mathbf{r}}_{\mathcal{M}} - \mathbf{r}^\star_{\mathcal{M}}\|_2\, \|\mathbf{r}^\star_{\mathcal{M}}\|_2 > 0.$$

The ray-distance argument then gives

$$\sin\theta \leq \frac{b_{\mathcal{M}}}{\|\mathbf{r}^\star_{\mathcal{M}}\|_2}, \qquad \cos\theta \geq \sqrt{1 - \frac{b_{\mathcal{M}}(\delta)^2}{\|\mathbf{r}^\star_{\mathcal{M}}\|_2^2}}.$$

Equal-size finite-sample specialization and marker stability. Suppose $n_i = n$. Define the marker-residual variance scale

$$\sigma^2_{\mathcal{M}} := \max_i \left\|R_{\mathcal{M}} Q \Omega(\mathbf{f}_i) Q R_{\mathcal{M}}^\top\right\|_{\mathrm{op}} \leq \sigma^2_{\mathcal{M}}.$$

Hence the $q \times q$ positive-semidefinite covariance also satisfies

$$\operatorname{tr}(R_{\mathcal{M}} Q \Omega(\mathbf{f}_i) Q R_{\mathcal{M}}^\top) \leq q\sigma^2_{\mathcal{M}}.$$

Consequently,

$$\|\Sigma_{\bar{\mathbf{r}},\mathcal{M}}\|_{\mathrm{op}} \leq \frac{\sigma^2_{\mathcal{M}}}{mn}, \qquad \operatorname{tr}(\Sigma_{\bar{\mathbf{r}},\mathcal{M}}) \leq \frac{q\sigma^2_{\mathcal{M}}}{mn},$$

and

$$b_{\mathcal{M}}(\delta) \leq \sigma_{\mathcal{M}}\sqrt{\frac{q}{mn}} + \sigma_{\mathcal{M}}\sqrt{\frac{2\log(1/\delta)}{mn}}. \tag{18}$$

This makes the role of marker stability explicit: for fixed q, m, n, a smaller uniform marker-residual variance proxy $\sigma^2_{\mathcal{M}}$ yields a tighter concentration radius. The $O((mn)^{-1/2})$ rate itself is unchanged.

Corollary C.1 (Marker population-to-target discrepancy)**.** *Let $\rho_i := \left\|\mathbf{f}_i - \bar{\mathbf{f}}\right\|_2$. The noise-free discrepancy relation (10) holds. On any event $\|R_{\mathcal{M}} Q \boldsymbol{\xi}_i\|_2 \leq b_{i,\mathcal{M}}$, if*

$$\rho_i \|R_{\mathcal{M}} \Delta_Q\|_{\mathrm{op}} + b_{i,\mathcal{M}} < \|\mathbf{r}^\star_{\mathcal{M}}\|_2,$$

then

$$\sin\operatorname{angle}(\mathbf{r}_{i,\mathcal{M}}, \mathbf{r}^\star_{\mathcal{M}}) \leq \frac{\rho_i \|R_{\mathcal{M}} \Delta_Q\|_{\mathrm{op}} + b_{i,\mathcal{M}}}{\|\mathbf{r}^\star_{\mathcal{M}}\|_2}. \tag{19}$$

Proof. Because

$$\mathbf{r}^\star_{i,\mathcal{M}} - \mathbf{r}^\star_{\mathcal{M}} = R_{\mathcal{M}} \Delta_Q (\mathbf{f}_i - \bar{\mathbf{f}}),$$

submultiplicativity gives (10). Further,

$$\mathbf{r}_{i,\mathcal{M}} - \mathbf{r}^\star_{\mathcal{M}} = R_{\mathcal{M}} \Delta_Q (\mathbf{f}_i - \bar{\mathbf{f}}) + R_{\mathcal{M}} Q \boldsymbol{\xi}_i,$$

so the same acute-angle and ray-distance argument gives (19). Thus statistical concentration and population-to-target discrepancy remain distinct after marker restriction.

C.3 Marker-residual decoder and structural bridge

The decoder in Eq. (11) acts only on the q marker-residual coordinates. The reference matrix is used to choose component-enriched supports and within-module weights. Markers in S_k have strong reference expression in component k and typically weaker expression in other components. This favors marker responses whose intended component module is stronger than their cross-component modules, providing the component-wise organization used by the decoder.

Section 4.3 represents this component-wise relation as

$$R_{\mathcal{M}} \Delta_Q = V_{\mathcal{M}} A + E_{\mathcal{M}}, \qquad V_{\mathcal{M}} = R_{\mathcal{M}} \Delta_Q A^\dagger.$$

Here $V_{\mathcal{M}}$ represents the shared component-wise structure between the fraction-relevant effect encoded by A and the residual-visible marker response. Its columns describe how individual fraction-bias coordinates appear across the selected marker modules after residual projection. With component-specific markers, the intended module is expected to respond more strongly than the other modules, while $E_{\mathcal{M}}$ retains marker-visible variation that is not explained by this shared structure. The quantities below measure how well the intended and cross-component responses are separated.

Residual-response coherence. Write

$$V_{\mathcal{M}} = [\mathbf{v}_{\mathcal{M},1}, \dots, \mathbf{v}_{\mathcal{M},p}], \qquad W = [\mathbf{w}_1^\top; \dots; \mathbf{w}_p^\top], \qquad D = \mathrm{diag}(d_1, \dots, d_p).$$

Then

$$(DWV_{\mathcal{M}})_{\ell k} = d_\ell \, \mathbf{w}_\ell^\top \mathbf{v}_{\mathcal{M},k}.$$

Define the diagonal calibration error and cross-component response leakage

$$\varepsilon_{\mathrm{cal}} := \max_k |d_k \mathbf{w}_k^\top \mathbf{v}_{\mathcal{M},k} - 1|, \qquad \varepsilon_{\mathrm{cross}} := \max_{\ell \neq k} |d_\ell \mathbf{w}_\ell^\top \mathbf{v}_{\mathcal{M},k}|.$$

Therefore

$$\|DWV_{\mathcal{M}} - I_p\|_F^2 \leq p\varepsilon_{\mathrm{cal}}^2 + p(p-1)\varepsilon_{\mathrm{cross}}^2,$$

and hence

$$\varepsilon_{\mathrm{rsp}} = \|DWV_{\mathcal{M}} - I_p\|_{\mathrm{op}} \leq \sqrt{p\varepsilon_{\mathrm{cal}}^2 + p(p-1)\varepsilon_{\mathrm{cross}}^2}. \tag{20}$$

Under ideal diagonal calibration $d_k = (\mathbf{w}_k^\top \mathbf{v}_{\mathcal{M},k})^{-1}$, define

$$s_{\min} := \min_k |\mathbf{w}_k^\top \mathbf{v}_{\mathcal{M},k}|, \qquad \mu_{\mathrm{cross}} := \max_{\ell \neq k} |\mathbf{w}_\ell^\top \mathbf{v}_{\mathcal{M},k}|.$$

If $s_{\min} > 0$, then $\varepsilon_{\mathrm{cal}} = 0$ and

$$\varepsilon_{\mathrm{rsp}} \leq \sqrt{p(p-1)}\, \frac{\mu_{\mathrm{cross}}}{s_{\min}}. \tag{21}$$

Thus strong intended response and weak cross-component response make the marker-residual bridge nearly coordinate preserving after calibration. Reference expression specificity helps construct such component-enriched modules, but Eq. (21) makes explicit that the decisive quantity is their response to $V_{\mathcal{M}}$, not their ability to invert an arbitrary reference mixture.

Proof of the marker-response bridge bound. From Eq. (12) and $\mathbf{d}^\star = A\bar{\mathbf{f}}$,

$$\mathbf{r}_{\mathcal{M}}^\star = R_{\mathcal{M}}\Delta_Q \bar{\mathbf{f}} = V_{\mathcal{M}}\mathbf{d}^\star + E_{\mathcal{M}}\bar{\mathbf{f}}.$$

Therefore

$$\begin{aligned} T(\mathbf{r}_{\mathcal{M}}^\star) - \mathbf{d}^\star &= CDW(V_{\mathcal{M}}\mathbf{d}^\star + E_{\mathcal{M}}\bar{\mathbf{f}}) - \mathbf{d}^\star \\ &= C(DWV_{\mathcal{M}} - I_p)\mathbf{d}^\star - (I_p - C)\mathbf{d}^\star + CDWE_{\mathcal{M}}\bar{\mathbf{f}}. \end{aligned}$$

Since C is an orthogonal projector, $\|C\|_{\mathrm{op}} = 1$, and $\left\|\bar{\mathbf{f}}\right\|_2 \leq 1$. Hence

$$\begin{aligned} \|T(\mathbf{r}_{\mathcal{M}}^\star) - \mathbf{d}^\star\|_2 &\leq \varepsilon_{\mathrm{rsp}} \|\mathbf{d}^\star\|_2 + \|(I_p - C)\mathbf{d}^\star\|_2 + \|CDW\|_{\mathrm{op}} \|E_{\mathcal{M}}\|_{\mathrm{op}} \\ &= \varepsilon_{\mathrm{rsp}} \|\mathbf{d}^\star\|_2 + \varepsilon_{\tan} + \tau\varepsilon_{\mathrm{rem}} = \beta_{\mathcal{M}}, \end{aligned} \tag{22}$$

which proves (13). The bound separates residual-response leakage, tangent incompatibility, and unstructured marker-visible mismatch. Expression stability is absent from this structural term because it enters separately through $b_{\mathcal{M}}$.

C.4 Population fraction-error recovery

Proof of Theorem 4.4. Add and subtract $T(\mathbf{r}_{\mathcal{M}}^\star)$:

$$\begin{aligned} \left\|\widehat{\mathbf{d}} - \mathbf{d}^\star\right\|_2 &= \|T(\bar{\mathbf{r}}_{\mathcal{M}}) - \mathbf{d}^\star\|_2 \\ &\leq \|T(\bar{\mathbf{r}}_{\mathcal{M}} - \mathbf{r}_{\mathcal{M}}^\star)\|_2 + \|T(\mathbf{r}_{\mathcal{M}}^\star) - \mathbf{d}^\star\|_2 \\ &\leq \tau \|\bar{\mathbf{r}}_{\mathcal{M}} - \mathbf{r}_{\mathcal{M}}^\star\|_2 + \beta_{\mathcal{M}} \\ &\leq \tau b_{\mathcal{M}}(\delta) + \beta_{\mathcal{M}} = b_d(\delta). \end{aligned} \tag{23}$$

Thus $\tau b_{\mathcal{M}}$ is the population marker-residual statistical contribution, whereas $\beta_{\mathcal{M}}$ is the structural marker-response contribution.

For a target sample, since $-U^\dagger \Delta_P = A$,

$$\mathbf{d}_i^\star - \mathbf{d}^\star = A(\mathbf{f}_i - \bar{\mathbf{f}}), \qquad \|\mathbf{d}_i^\star - \mathbf{d}^\star\|_2 \leq \rho_i \|A\|_{\mathrm{op}}.$$

The actual correction satisfies

$$\mathbf{d}_i = \mathbf{f}_i - \widehat{\mathbf{f}}_i = \mathbf{d}_i^\star - U^\dagger \boldsymbol{\xi}_i.$$

Therefore, on the event $\left\|U^\dagger \boldsymbol{\xi}_i\right\|_2 \le b_{U,i}(\delta_i)$,

$$\begin{aligned}\left\|\widehat{\mathbf{d}} - \mathbf{d}_i\right\|_2 &\le \left\|\widehat{\mathbf{d}} - \mathbf{d}^\star\right\|_2 + \|\mathbf{d}^\star - \mathbf{d}_i^\star\|_2 + \left\|U^\dagger \boldsymbol{\xi}_i\right\|_2 \\ &\le b_d(\delta) + \rho_i \|A\|_{\mathrm{op}} + b_{U,i}(\delta_i),\end{aligned}$$

which is (15). Increasing the number of training aggregates only decreases the population marker-residual stochastic term; it does not remove composition mismatch or target-specific coefficient noise.

Heterogeneous-shift extension. The marker-space bridge is imposed on the cohort-common mismatch Δ; no marker-alignment condition is imposed on the remaining sample-specific deviations H_i. For $H_j \neq 0$,

$$\begin{aligned}\mathbf{r}_{j,\mathcal{M}} &= R_{\mathcal{M}} Q(\Delta + H_j)\mathbf{f}_j + R_{\mathcal{M}} Q \boldsymbol{\xi}_j, \\ \bar{\mathbf{r}}_{\mathcal{M}} &= \mathbf{r}_{\mathcal{M}}^\star + \frac{1}{m}\sum_{j=1}^m R_{\mathcal{M}} Q H_j \mathbf{f}_j + \frac{1}{m}\sum_{j=1}^m R_{\mathcal{M}} Q \boldsymbol{\xi}_j.\end{aligned}$$

Define

$$\mathbf{h}_{Q,\mathcal{M}} := \frac{1}{m}\sum_{j=1}^m R_{\mathcal{M}} Q H_j \mathbf{f}_j.$$

On the same sampling event as the main theorem,

$$\|\bar{\mathbf{r}}_{\mathcal{M}} - \mathbf{r}_{\mathcal{M}}^\star\|_2 \le \|\mathbf{h}_{Q,\mathcal{M}}\|_2 + b_{\mathcal{M}}(\delta),$$

so

$$\left\|\widehat{\mathbf{d}} - \mathbf{d}^\star\right\|_2 \le \beta_{\mathcal{M}} + \tau b_{\mathcal{M}}(\delta) + \tau \|\mathbf{h}_{Q,\mathcal{M}}\|_2 .$$

For target i, let $\mathbf{d}_i^{\mathrm{het},\star} := -U^\dagger(\Delta + H_i)\mathbf{f}_i$. Then

$$\left\|\mathbf{d}_i^{\mathrm{het},\star} - \mathbf{d}^\star\right\|_2 \le \rho_i \|A\|_{\mathrm{op}} + \left\|U^\dagger H_i \mathbf{f}_i\right\|_2 .$$

Finally, $\mathbf{d}_i = \mathbf{d}_i^{\mathrm{het},\star} - U^\dagger \boldsymbol{\xi}_i$, and the heterogeneous target radius is

$$b_{\mathrm{tot},i}^{(\mathrm{het})} = \beta_{\mathcal{M}} + \tau b_{\mathcal{M}}(\delta) + \tau \|\mathbf{h}_{Q,\mathcal{M}}\|_2 + \rho_i \|A\|_{\mathrm{op}} + \left\|U^\dagger H_i \mathbf{f}_i\right\|_2 + b_{U,i}(\delta_i). \tag{24}$$

Thus $\left\|\widehat{\mathbf{d}} - \mathbf{d}_i\right\|_2 \le b_{\mathrm{tot},i}^{(\mathrm{het})}$ with probability at least $1 - \delta - \delta_i$. Setting every $H_j = 0$ recovers the main model.

Corollary C.2 (Directional consequence of population fraction-error recovery)**.** *Under the event of Theorem 4.4, if $b_d(\delta) < \|\mathbf{d}^\star\|_2$, then*

$$\sin \operatorname{angle}(\widehat{\mathbf{d}}, \mathbf{d}^\star) \le \frac{b_d(\delta)}{\|\mathbf{d}^\star\|_2}.$$

Proof. The bound (14) places $\widehat{\mathbf{d}}$ inside a ball of radius b_d around $\mathbf{d}^\star$. When $b_d < \|\mathbf{d}^\star\|_2$, the same acute-half-space and ray-distance argument used in Appendix C.2 yields the stated angular bound.

C.5 Analytic safety and observable residual empirical magnitude

Proof of Theorem 4.5. Let $\mathbf{e}_i = \widehat{\mathbf{f}}_i - \mathbf{f}_i = -\mathbf{d}_i$. For any unit vector $\mathbf{v}$,

$$\begin{aligned}\|\mathbf{e}_i + \lambda \mathbf{v}\|_2^2 - \|\mathbf{e}_i\|_2^2 &= \lambda^2 + 2\lambda\langle \mathbf{e}_i, \mathbf{v}\rangle \\ &= \lambda^2 - 2\lambda\langle \mathbf{d}_i, \mathbf{v}\rangle.\end{aligned} \tag{25}$$

We now apply (25) with $\mathbf{v} = \widehat{\mathbf{d}} / \left\|\widehat{\mathbf{d}}\right\|_2$. The target-level error bound of Theorem 4.4 implies

$$\begin{aligned}\langle \mathbf{d}_i, \mathbf{v} \rangle &= \langle \widehat{\mathbf{d}}, \mathbf{v} \rangle + \langle \mathbf{d}_i - \widehat{\mathbf{d}}, \mathbf{v} \rangle \\ &\geq \left\|\widehat{\mathbf{d}}\right\|_2 - \left\|\mathbf{d}_i - \widehat{\mathbf{d}}\right\|_2 \|\mathbf{v}\|_2 \\ &\geq \left\|\widehat{\mathbf{d}}\right\|_2 - b_{\mathrm{tot},i} = \gamma_i^{\mathrm{ana}}.\end{aligned}$$

Hence

$$\|\mathbf{e}_i + \lambda \mathbf{v}\|_2^2 - \|\mathbf{e}_i\|_2^2 \leq \lambda^2 - 2\lambda\gamma_i^{\mathrm{ana}} = \lambda(\lambda - 2\gamma_i^{\mathrm{ana}}),$$

which is strictly negative whenever $\gamma_i^{\mathrm{ana}} > 0$ and $0 < \lambda < 2\gamma_i^{\mathrm{ana}}$. This proves Theorem 4.5.

Observable residual empirical step. The empirical construction uses the same direction $\mathbf{v}$ as Section 4.5. For a candidate magnitude λ in a finite set Λ, define

$$\mathbf{f}_i(\lambda) := \widehat{\mathbf{f}}_i + \lambda \mathbf{v}.$$

The observed population marker residual is

$$\bar{\mathbf{r}}_{\mathcal{M}} := \frac{1}{m} \sum_{i=1}^{m} \left(\mathbf{y}_{i,\mathcal{M}} - U_{\mathcal{M}} \widehat{\mathbf{f}}_i \right),$$

and the marker-space forward response induced by candidate λ is

$$\bar{\mathbf{h}}(\lambda) := \frac{1}{m} \sum_{i=1}^{m} U_{\mathcal{M}} \{ \mathbf{f}_i(\lambda) - \widehat{\mathbf{f}}_i \}. \tag{26}$$

Thus $\bar{\mathbf{h}}(\lambda)$ is the marker-expression change predicted by moving the fitted compositions a distance λ along $\mathbf{v}$. For marker gene g, let σ_{gk}^2 be its reference variance in component k and define

$$\bar{v}_g(\lambda) := \frac{1}{m^2} \sum_{i=1}^{m} \sum_{k=1}^{p} f_{ik}(\lambda) \sigma_{gk}^2, \qquad G_\lambda := \left\{ \mathrm{diag}\left(\bar{v}_g(\lambda) + \varepsilon \right) \right\}^{-1}.$$

The diagonal weights in G_λ down-weight marker coordinates whose reference-based forward response is more variable. The variance-weighted aligned scale is

$$\beta(\lambda) := \frac{\bar{\mathbf{h}}(\lambda)^\top G_\lambda \bar{\mathbf{r}}_{\mathcal{M}}}{\bar{\mathbf{h}}(\lambda)^\top G_\lambda \bar{\mathbf{h}}(\lambda)}. \tag{27}$$

Equivalently, $\beta(\lambda)$ is the variance-weighted least-squares scale needed to rescale the candidate forward response to match the observed marker residual. Hence $\beta(\lambda) = 1$ means that the candidate magnitude already matches the observed residual scale without further rescaling. Among candidates with a nondegenerate denominator, define

$$\widehat{\lambda} \in \arg\min_{\lambda \in \Lambda} \{ \beta(\lambda) - 1 \}^2. \tag{28}$$

Thus $\widehat{\lambda}$ is the candidate magnitude whose marker-space forward response best matches the observed population marker residual in variance-weighted scale. Setting $\lambda_i = \widehat{\lambda}$ gives

$$\widehat{\mathbf{f}}_i^{\mathrm{new}} = \widehat{\mathbf{f}}_i + \lambda_i \mathbf{v}.$$

The criterion uses fitted fractions, marker reference means and variances, and observed target marker residuals, without estimating the full shift Δ.

D Practical Instantiation of the Population-Residual Principle

Population-Guided Least Squares (PopLS) instantiates the population-residual principle for transcriptomic deconvolution. The initial estimate is fitted separately to each bulk using a relatively broad aligned feature set and the fixed single-cell reference, without pooling information across target bulks.

Reference means supply the design matrix and reference variances define feature weights for this per-bulk fit. This stage plays the role of the broad U and $\widehat{\mathbf{f}}_i$ in Section 4.

The correction stage selects a narrower marker subset $\mathcal{M}$ from the rows already present in U using the stable-signature-gene procedure of ReDeconv (Lu et al., 2025). Pairwise Welch t-tests together with reference mean fold changes first identify component-enriched candidates; these candidates are then ranked by their within-component coefficient of variation, $\mathrm{CV}_{gk} = \sigma_{gk}/\mu_{gk}$, and the lowest-CV genes are retained for each component. The resulting sets S_k are united in original gene order and aligned with the target features and reference records, so $U_{\mathcal{M}} = R_{\mathcal{M}}U$ remains a row subset of the broad fit reference.

For each bulk, PopLS first forms the post-fit reconstruction residual from the broad initial fit and then keeps only the selected marker coordinates,

$$\mathbf{r}_{i,\mathcal{M}} = R_{\mathcal{M}}(\mathbf{y}_i - U\widehat{\mathbf{f}}_i).$$

It does not recompute $\widehat{\mathbf{f}}_i$ from $U_{\mathcal{M}}$. Marker residual features are standardized by reference-based predicted variability. Within each component-enriched marker module, PopLS summarizes the standardized residuals for every bulk and then aggregates them across the cohort, weighting the component score by its initial abundance. This cohort summary is the practical counterpart of learning $\bar{\mathbf{r}}_{\mathcal{M}}$. The component modules and their reference magnitudes provide the residual-aggregation role of W; orientation and scale calibration followed by removal of the common all-components offset correspond to D and C.

The observable gate is constructed from the controlled bridge-validity experiment summarized in Section 5.2 and detailed in Appendix E.2.3. That experiment perturbs marker support W, component correspondence C, and response orientation D to expose stable but structurally invalid directions. The implementation evaluates structural support without using true fractions or true errors. Disjoint marker folds provide an agreement check for the proposed component direction, while marker-to-background residual strength checks that the shared signal is component enriched rather than diffuse. These observable diagnostics determine whether the recovered direction is sufficiently supported to be used; they provide observable evidence for marker-response coherence and direction reliability before the empirical step is selected.

For a supported direction, PopLS uses the observable residual empirical step of Appendix C.5. Candidate magnitudes $\lambda \in \Lambda$ move the initial estimates along the recovered unit direction $\mathbf{v}$ and are mapped forward through $U_{\mathcal{M}}$. Equation (28) selects the magnitude whose variance-weighted forward response best matches the observed cohort marker residual. The selected magnitude is applied as $\widehat{\mathbf{f}}_i + \lambda_i \mathbf{v}$; if the structural direction gate fails, the initial estimates are retained. Thus the implemented order is broad per-bulk fitting, marker restriction, marker-population aggregation, structural direction validation, observable residual-based magnitude selection, and the final shared update.

E Detailed Experimental Protocols

The controlled evidence below is grouped by the theoretical quantity it isolates. Transcriptomic benchmarks are then presented with their data construction, fixed evaluation procedure, comparison protocol, and source information.

All fraction-error metrics use exactly aligned prediction and truth matrices. After flattening the evaluation arrays, denote the J paired entries by $\{(\widehat{\pi}_j, \pi_j)\}_{j=1}^{J}$, where $\widehat{\pi}_j$ and π_j denote the predicted and true proportions, respectively. We report

$$\mathrm{MAE} = \frac{1}{J}\sum_{j=1}^{J}|\widehat{\pi}_j - \pi_j|, \qquad \mathrm{RMSE} = \left\{\frac{1}{J}\sum_{j=1}^{J}(\widehat{\pi}_j - \pi_j)^2\right\}^{1/2},$$

$$r = \frac{\sum_{j=1}^{J}(\widehat{\pi}_j - \overline{\widehat{\pi}})(\pi_j - \bar{\pi})}{\left\{\sum_{j=1}^{J}(\widehat{\pi}_j - \overline{\widehat{\pi}})^2 \sum_{j=1}^{J}(\pi_j - \bar{\pi})^2\right\}^{1/2}}, \qquad \mathrm{CCC} = \frac{2s_{\widehat{\pi},\pi}}{s_{\widehat{\pi}}^2 + s_{\pi}^2 + (\overline{\widehat{\pi}} - \bar{\pi})^2},$$

where $s_{\widehat{\pi},\pi} = J^{-1}\sum_j(\widehat{\pi}_j - \overline{\widehat{\pi}})(\pi_j - \bar{\pi})$ and $s_{\widehat{\pi}}^2, s_{\pi}^2$ use the corresponding J^{-1} normalization. MAE measures typical absolute fraction error and RMSE places greater weight on large deviations; lower

is better for both. Pearson r measures pattern correlation, whereas Lin's concordance correlation coefficient (CCC) additionally measures calibration-sensitive agreement; higher is better for both.

Uncertainty is reported as an estimate with confidence half-width h_{95}, displayed as a small $\pm h_{95}$ subscript. Benchmark intervals use 10,000 bootstrap resamples of complete target mixtures or bulk samples, retaining all component fractions within a resampled sample; Table 7 instead bootstraps the 20 replicate-level MAEs within each cancer.

E.1 Residual Geometry and Population Estimation

E.1.1 Column-Space and Residual-Space Effects of Reference Mismatch

This experiment tests the projection geometry of Theorem 4.1 and the residual-only non-identifiability in Proposition 4.2: the coefficient effect $\widehat{\mathbf{f}}_i - \mathbf{f}_i = U^\dagger \Delta \mathbf{f}_i$ depends on the reference column-space component, whereas the observable residual $\mathbf{r}_i = Q\Delta \mathbf{f}_i$ depends on its orthogonal component. We generated full-column-rank $U \in \mathbb{R}^{500\times 5}$ with condition number 3, drew $m = 100$ compositions from $\mathrm{Dirichlet}(10\mathbf{1})$, and used no observation noise, finite-cell noise, regularization, marker selection, gating, or simplex projection. For an independent Gaussian seed matrix G, define

$$G_P = \frac{PG}{\|PG\|_F}, \qquad G_Q = \frac{QG}{\|QG\|_F}, \qquad \Delta(\eta) = 0.2\|U\|_F\{\sqrt{1-\eta^2}\, G_P - \eta G_Q\},$$

with $\eta \in \{0, 0.1, \ldots, 1.0\}$. Here G_P has unit Frobenius norm in the reference column space: it changes the unconstrained least-squares coefficients but is absorbed by the reference fit and therefore creates no Q-residual. Conversely, G_Q has unit Frobenius norm in the orthogonal complement: it is visible in the reconstruction residual but does not change the unconstrained least-squares coefficients. Thus $\eta = 0$ is a purely column-space shift, $\eta = 1$ is a purely orthogonal shift, and intermediate values change only the geometry of the fixed-strength mismatch, not its Frobenius magnitude.

Writing $s_0 = 0.2\|U\|_F$, the orthogonality of the two normalized components gives $\|\Delta(\eta)\|_F = s_0$, $\|P\Delta(\eta)\|_F/s_0 = \sqrt{1-\eta^2}$, and $\|Q\Delta(\eta)\|_F/s_0 = \eta$. The generator therefore changes the allocation between the two subspaces while holding U, P, Q, and total mismatch strength fixed.

Across twenty seeds, the maximum absolute errors in $\widehat{\mathbf{f}} - \mathbf{f} = U^\dagger \Delta \mathbf{f}$ and $\mathbf{r} = Q\Delta\mathbf{f}$ were 1.922×10^{-15} and 5.740×10^{-15}. At $\eta = 0$, mean coefficient error was 0.1278 while mean residual magnitude was 8.04×10^{-16}; at $\eta = 1$, the corresponding values were 3.67×10^{-16} and 0.1955. Main Figure 1(a) places both quantities on the same axis against η. Thin translucent curves retain the original twenty seed trajectories, and the heavier curves show the across-seed means.

This means that the column-space and orthogonal mismatch components encode complementary information. An observable residual cannot by itself recover the hidden column-space bias, exactly as Theorem 4.1 and Proposition 4.2 require.

E.1.2 Population Residuals Concentrate at the Effective-Sample-Size Rate

This experiment tests the $O((mn)^{-1/2})$ concentration law underlying Theorem 4.3. It uses $R_{\mathcal{M}} = I_g$, the all-feature special case of the fixed-row-selection theorem, to isolate the effective-sample-size scaling without marker selection. It isolates finite-aggregate sampling by fixing one shared orthogonal shift and a uniform composition, setting composition variation and sample-specific shift to zero, and drawing exact Gaussian component-wise aggregate means. The grid used $m \in \{2, 4, 8, 16, 32, 64, 128\}$ and $n \in \{50, 100, 200, 500, 1000, 2000\}$, with twenty reference seeds and twenty aggregate repetitions per grid point.

The residual identity held with maximum relative error 8.126×10^{-15}. Single-aggregate noise scaled as $n^{-1/2}$, covariance trace as $(mn)^{-1}$, and the population error $e_j = \|\bar{\mathbf{r}} - \mathbf{r}^\star\|_2$ in this $R_{\mathcal{M}} = I_g$ special case as $(mn)^{-1/2}$. To estimate the scaling exponent, we fit

$$\log e_j = \alpha + \beta \log(m_j n_j) + \varepsilon_j$$

by ordinary least squares. Theory predicts $\beta = -1/2$; the fitted exponent was -0.5003 with a two-sided 95% regression confidence interval $[-0.5006, -0.5000]$. The interval tightly contains $-1/2$, supporting the square-root law.

Main Figure 1(b) uses the matched-composition curve to show the same concentration law within the population-to-target error. Figure 2(a) gives the full m-by-n grid: the horizontal axis is effective sample size mn on a logarithmic scale, the vertical axis is the all-feature special-case residual error $\|\bar{\mathbf{r}} - \mathbf{r}^\star\|_2$ on a logarithmic scale, and cohort size is encoded by color. The fitted line is compared with the theoretical $-1/2$ slope.

This means that m and n enter through the same effective sample size in this controlled regime: either more independent mixtures or more cells per mixture reduces stochastic population-residual error at the predicted square-root rate.

E.1.3 Composition Heterogeneity Creates an Irreducible Population-to-Target Floor

This experiment tests the population-to-target discrepancy relation in the all-feature special case $R_{\mathcal{M}} = I_g$ of Eq. (10),

$$\mathbf{r}_i^\star - \mathbf{r}^\star = Q\Delta(\mathbf{f}_i - \bar{\mathbf{f}}), \qquad \|\mathbf{r}_i^\star - \mathbf{r}^\star\|_2 \le \rho_i \|Q\Delta\|_{\mathrm{op}}.$$

thereby separating stochastic population estimation from the deterministic composition-induced floor. It fixes $n = 2000$, varies $m \in \{4, 8, 16, 32, 64, 128, 256, 512\}$ and $\rho = \max_i \|\mathbf{f}_i - \bar{\mathbf{f}}\|_2$ over seven levels from 0 to 0.30, and uses 20 reference seeds with 10 aggregate repetitions per seed (200 repetitions per finite-sampling condition). The deterministic study contains 256,000 sample-level observations.

For a Gaussian matrix $G \in \mathbb{R}^{g\times p}$, the base observable shift is

$$\Delta_{Q,\mathrm{iso}} = QG\{G^\top QG\}^{-1/2}, \qquad \Delta_{Q,\mathrm{iso}}^\top \Delta_{Q,\mathrm{iso}} = I_p.$$

The four geometries then take the Section 4 form

$$\Delta_Q = Q\Delta = c\Delta_{Q,\mathrm{iso}} R\,\mathrm{diag}(\sqrt{\kappa_1}, \ldots, \sqrt{\kappa_p})R^\top, \qquad \|Q\Delta\|_F = 0.1\|U\|_F.$$

The isotropic geometry uses $\boldsymbol{\kappa} = (1, 1, 1, 1, 1)$ and therefore equal component-direction scales. The well-conditioned geometry uses $(1, 1.35, 1.70, 2.20, 3.0)$ to introduce moderate directional variation, and the anisotropic geometry uses $(1, 2, 3.5, 6, 10)$ to create a wider spectral spread. For these three cases, R is orthogonal and has first column $\mathbf{f}_0/\|\mathbf{f}_0\|_2$, aligning one spectral axis with the baseline composition. The near-low-rank geometry uses $R = I_p$ and $(1, 0.20, 0.05, 0.015, 0.005)$, concentrating most observable energy in a small number of directions. The scalar c normalizes every construction to the same Frobenius strength, so the comparison changes geometry rather than total mismatch magnitude.

The discrepancy identity

$$Q\Delta\mathbf{f}_i - Q\Delta\bar{\mathbf{f}} = Q\Delta(\mathbf{f}_i - \bar{\mathbf{f}})$$

held to 1.56×10^{-17}, with no operator-norm-bound violations. It isolates the deterministic population-to-target term in Eq. (10). In the finite-sampling model $\mathbb{E}\|\bar{\mathbf{r}} - Q\Delta\mathbf{f}_i\|_2^2 = a + b/(mn)$, the fitted intercept agreed with the deterministic floor with slope 1.0003, intercept -3.37×10^{-6}, and $R^2 = 0.99997$. Main Figure 1(b) plots this single response field against mn for representative ρ values, including the added $m = 256$ and $m = 512$ cohorts. Figure 2(b) shows all seven composition distances against cohort size; solid curves are empirical errors and horizontal dashed lines are the corresponding deterministic floors. Panel (c) plots the directly computed deterministic discrepancy magnitude $\|Q\Delta(\mathbf{f}_i - \bar{\mathbf{f}})\|_2$, averaged over the prespecified seeds and perturbation sets, against ρ for all four observable-shift geometries.

This means that increasing m suppresses sampling error but cannot eliminate deterministic composition-induced discrepancy. The four constructions verify the same decomposition under isotropic, moderately conditioned, anisotropic, and near-low-rank observable shifts.

E.1.4 Recoverability under Shared and Sample-Specific Shift

This experiment tests the competition among the shared observable signal, sample-specific mismatch, and sampling noise in

$$\mathbf{r}_i = Q\Delta\mathbf{f}_i + QH_i\mathbf{f}_i + Q\boldsymbol{\xi}_i,$$

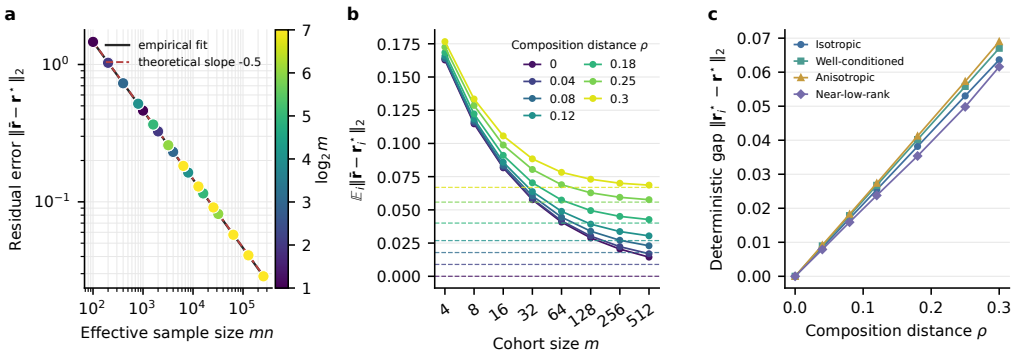

Figure 2: Population residual estimation and target discrepancy. (a) Residual error versus effective sample size mn. (b) Population-to-target residual gap versus cohort size for composition distances ρ; dashed lines are deterministic floors. (c) Deterministic composition-induced floor versus ρ for isotropic, well-conditioned, anisotropic, and near-low-rank observable shifts.

using the all-feature special case of the marker-response geometry to connect composition discrepancy with the sample-specific heterogeneity term in the target-level error budget of Theorem 4.4. The effective SNR below is an empirical summary statistic, not a theorem quantity. The experiment varies the shared observable mismatch, sample-specific mismatch, cohort size, and composition heterogeneity while fixing $p = 5$ mixture components, $g = 500$ observed features, and $n = 200$ elementary cells per aggregate. The three terms are respectively the shared observable population signal, sample-specific mismatch, and finite-aggregate sampling noise. The main grid fixes $m = 64$, varies $\|Q\Delta\|_F/\|U\|_F$ from 0 to 0.4, varies the sample-specific shift relative to the shared shift from 0 to 4, and uses composition radii 0.02 and 0.10, with 200 Monte Carlo populations per condition.

The effective population signal-to-noise ratio is

$$\mathrm{SNR}_{\mathrm{pop}} = \frac{\|Q\Delta\bar{\mathbf{f}}\|_2}{\left[\|m^{-1}\sum_{i=1}^m QH_i\mathbf{f}_i\|_2^2 + m^{-2}\sum_{i=1}^m n^{-1}\,\mathrm{tr}\{Q\Omega(\mathbf{f}_i)Q\}\right]^{1/2}}.$$

Larger values mean that the common residual signal dominates sample-specific perturbation and sampling noise. A population is labeled recovered when its shared-direction cosine exceeds 0.90. For a threshold t on $\mathrm{SNR}_{\mathrm{pop}}$, the ROC true-positive rate is the fraction of recovered populations with SNR above t, and the false-positive rate is the corresponding fraction among unrecovered populations; the AUC is the area traced by these rates over all thresholds. The ratio almost perfectly separates the two regimes (AUC $= 0.998$). A logistic regression of recovery on $\log \mathrm{SNR}_{\mathrm{pop}}$ gives McFadden pseudo-$R^2 = 0.887$, summarizing how much of the recovery transition is captured by this one-dimensional predictor. Under the fitted model, $\mathrm{SNR}_{\mathrm{pop}} \approx 2.16$ corresponds to a 90% estimated probability of shared-direction recovery.

Figure 3(a) places shared shift strength $\|Q\Delta\|_F/\|U\|_F$ on the horizontal axis and relative sample-specific shift on the vertical axis under low and moderate composition heterogeneity; color is the probability that shared-direction cosine exceeds 0.90. The two-subgroup control in panel (b) uses subgroup shift-separation angle (0°-90°) on the horizontal axis and mean population-to-sample direction cosine on the vertical axis, with separate curves for $m \in \{4, 8, 16, 32, 64, 128\}$. Larger cohorts can stabilize the population center while population-to-subgroup alignment deteriorates as the directions separate.

This means that a recoverable cohort-level direction need not be valid for each subgroup: the sample-specific H_i discrepancy term does not vanish merely by increasing m.

E.1.5 Finite-Sample Calibration of Directional Recovery

This calibration combines the residual concentration of Theorem 4.3, the residual-to-fraction bridge, and the directional consequence of Theorem 4.4 through the decoder-only specialization $\Phi(\bar{\mathbf{r}}_{\mathcal{M}}) = W\bar{\mathbf{r}}_{\mathcal{M}}$. It evaluates four properties: finite-sample coverage under matched Gaussian noise and

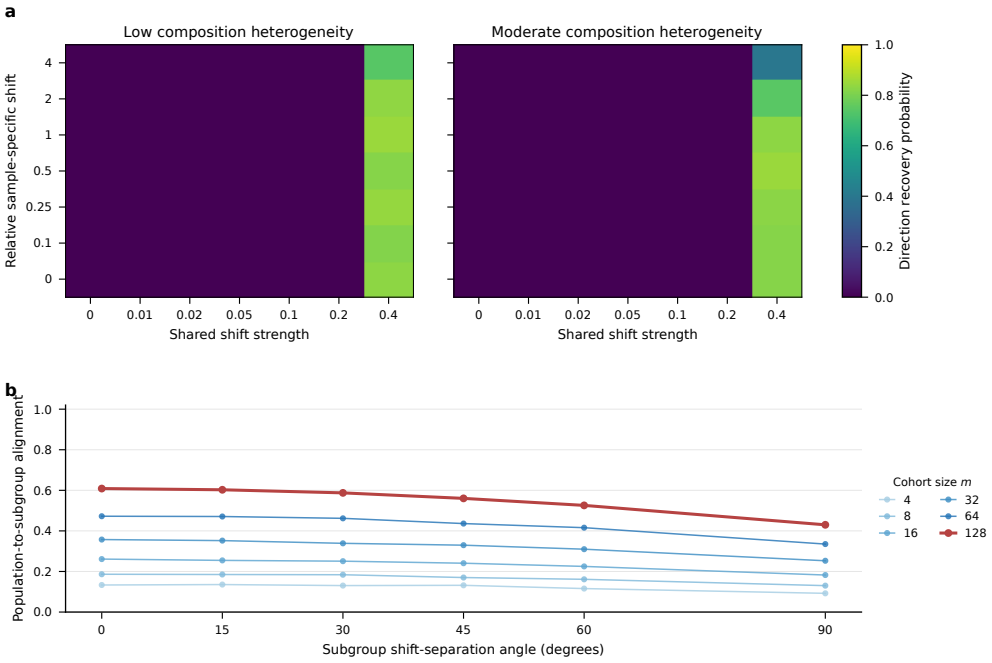

Figure 3: Shared-direction recovery and subgroup alignment. (a) Recovery probability over shared and sample-specific shift strengths. (b) Population-to-subgroup direction cosine versus subgroup separation for six cohort sizes.

informative markers; the dependence of required cohort size on m and n; abstention when markers contain no component information; and behavior under deliberate violation of Gaussian residual concentration. "Matched Gaussian" means that aggregate fluctuations follow the conditional Gaussian approximation in Theorem 4.3, with covariance matched to the generating model used for the theoretical radius. "Informative markers" means that the simulated reference modules are component enriched and that the residual-visible mismatch uses the corresponding component-specific marker modules, producing strong intended marker response and low cross-component response leakage in the sense of Section 4.3. Reference enrichment determines marker identity and support; the residual-response condition supplies the structural coherence required for direction recovery. The model uses $p = 5$, $g = 500$, 50 markers per component, $m \in \{10, 20, 50, 100, 200, 500\}$, $n \in \{100, 250, 500, 1000, 2500, 5000\}$, $\delta \in \{0.10, 0.05, 0.01\}$, and 5,000 Monte Carlo repetitions per condition.

This calibration isolates the decoder stage W before sign/scale calibration D and tangent-space centering C; the tested map is therefore $\Phi(\bar{\mathbf{r}}_{\mathcal{M}}) = W\bar{\mathbf{r}}_{\mathcal{M}}$, not a second unrelated bridge. The full theoretical map remains $T(\mathbf{r}_{\mathcal{M}}) = CDW\mathbf{r}_{\mathcal{M}}$. The tested certificate is

$$\|\Phi(\bar{\mathbf{r}}_{\mathcal{M}}) - \alpha\mathbf{d}^\star\|_2 \le \|W\|_{\mathrm{op}} b_{\mathcal{M}}(\delta) + b_{\mathrm{id}} = b_{\mathrm{total}}, \qquad c_{\mathrm{lower}} = \frac{\alpha\|\mathbf{d}^\star\|_2 - b_{\mathrm{total}}}{\alpha\|\mathbf{d}^\star\|_2 + b_{\mathrm{total}}},$$

when $b_{\mathrm{total}} < \alpha\|\mathbf{d}^\star\|_2$; otherwise it abstains. Empirical PAC coverage is the fraction of Monte Carlo repetitions in which the realized residual or directional error satisfies its corresponding high-probability bound. The quantity $q_{0.05}$ is the empirical fifth percentile of direction cosine and summarizes lower-tail directional recovery.

Matched-Gaussian residual and fraction-direction coverage were both 1.0000, and the worst $q_{0.05}$ exceeded its certified lower bound by 0.0259. Required cohort size followed the predicted n^{-1} and $\|\mathbf{d}^\star\|_2^{-2}$ laws. With non-informative markers, the bridge was declared non-identifiable and the certificate abstained. Student-t_3 denotes Student's t distribution with three degrees of freedom. If $Z \sim t_3$, the stress condition uses $10Z/\sqrt{3}$ before applying the target covariance square root. Thus the latent t_3 is first variance-normalized and then deliberately inflated tenfold; the resulting

condition is not variance-matched to the Gaussian condition. Its much heavier tails make finite-sample Gaussian concentration and the corresponding central-limit approximation unreliable at the tested scale, deliberately violating the premise of Theorem 4.3. Residual-bound coverage falls to zero even though downstream empirical direction recovery can remain strong; this does not validate the Gaussian theorem outside its premise. In bulk RNA-seq deconvolution, approximate Gaussian residuals arise naturally by averaging many finite-variance cell-level contributions through the multivariate central limit theorem. In other inverse problems, the corresponding concentration model must instead be derived or empirically validated for the relevant elementary observations.

In Figure 4, panel (a) places cells per aggregate n on the horizontal axis, cohort size m on the vertical axis, and c_{lower} in color. Panel (b), cited from the main text for certificate calibration, plots $q_{0.05}$ minus the certified lower bound against mn, with positive values indicating conservative coverage. Panel (c) is an assumption-outcome matrix: rows are matched Gaussian, Student-t_3, and non-informative-marker regimes; columns are residual and direction coverage, with "N/A" marking abstention. Panel (d) plots required theoretical cohort size against n and against $\|\mathbf{d}^\star\|_2$ to expose the two sample-complexity laws.

This means that the finite-sample certificate is conservative in its intended regime, fails its residual premise under deliberate heavy tails, and abstains when bridge information is absent.

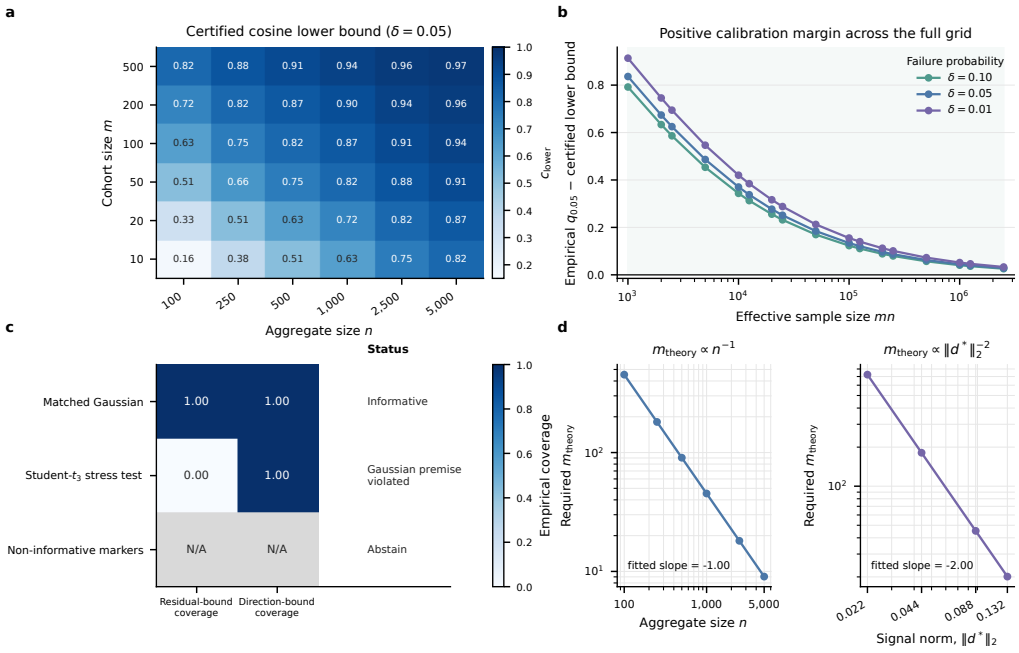

Figure 4: Finite-sample calibration. (a) Certified cosine lower bound over m and n. (b) Empirical $q_{0.05}$ minus the bound. (c) Coverage and abstention by regime. (d) Required cohort size versus n and $\|\mathbf{d}^\star\|_2$.

E.1.6 TARGET-SPECIFIC COEFFICIENT-NOISE CALIBRATION

This experiment isolates the target-specific coefficient-noise term $b_{U,i}(\delta_i)$ entering Theorem 4.4, without residual-bridge estimation, marker identification, gating, or an empirical correction step. Under the matched-Gaussian finite-aggregate model,

$$\boldsymbol{\xi}_i \mid \mathbf{f}_i \sim \mathcal{N}\big(0, n_i^{-1}\Omega(\mathbf{f}_i)\big)\,, \qquad Z_i = \big\|U^\dagger \boldsymbol{\xi}_i\big\|_2\,,$$

the analytic radius was computed from $\Sigma_{U,i} = n_i^{-1}U^\dagger\Omega(\mathbf{f}_i)(U^\dagger)^\top$ as $b_{U,i}(\delta_i) = \sqrt{\operatorname{tr}(\Sigma_{U,i})} + \sqrt{2\,\|\Sigma_{U,i}\|_{\text{op}}\log(1/\delta_i)}$. We fixed a full-column-rank $U \in \mathbb{R}^{500\times 5}$ with condition number 3, the uniform composition $\mathbf{f}_i = (0.2,\ldots,0.2)^\top$, and the component-level diagonal covariance structure,

Table 3: Directional-certificate diagnostics from 5,000 Monte Carlo repetitions per condition.

Regime	Status	Fraction PAC	Residual PAC	Diagnostic
Matched Gaussian	Informative	1.0000	1.0000	$q_{0.05}$ margin $+0.0259$
Non-informative markers	Abstain	–	–	Median cosine -0.0000
Student-t_3	Premise violated	1.0000†	0.0000	Residual bound fails
Sample-complexity grid	–	–	–	Slopes -1.0000, -2.0000

†The empirical fraction-direction result under Student-t_3 noise does not validate the Gaussian theorem because its residual-concentration premise has zero empirical coverage.

while varying $n_i \in \{50, 100, 200, 500, 1000, 2000\}$ and $\delta_i \in \{0.10, 0.05, 0.01\}$. Each n_i used 10,000 Monte Carlo draws from five independent seeds.

Figure 5(a) shows that the analytic radius and empirical 95th percentile follow the predicted square-root scaling, with log–log slopes -0.500 and -0.501, respectively. Panel (b) reports empirical coverage minus nominal coverage. Across all 18 (n_i, δ_i) conditions, coverage ranges from 0.9990 to 1.0000 and never falls below nominal. The radius is valid but conservative: $b_{U,i}/q_{1-\delta_i}$ ranges from 1.56 to 1.79. As an implementation check, Monte Carlo covariance closely matches its analytic counterpart, with maximum trace-relative error 0.0085.

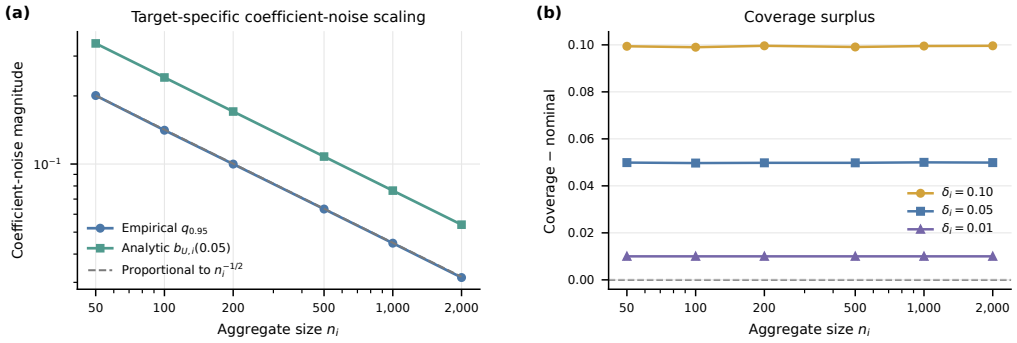

Figure 5: Target-specific coefficient-noise calibration. (a) The empirical 95th percentile of $\left\|U^\dagger \boldsymbol{\xi}_i\right\|_2$ and the analytic radius $b_{U,i}(0.05)$ follow the predicted $n_i^{-1/2}$ scaling. (b) Empirical coverage minus nominal coverage for three confidence levels; all tested conditions remain above zero.

E.2 Residual-to-Fraction Recovery and Safe Rejection

The real-cell-derived geometry uses single-cell cohorts from BRCA, COAD, HNSC, KIRC, LGG, LIHC, LUAD, and STAD; Table 4 lists their sources and dimensions.

E.2.1 Marker Specificity Supports the Shared-Latent Bridge

This experiment probes observable consequences of the marker-supported shared-latent bridge in Section 4.3. Marker specificity changes how selectively W responds to component-enriched patterns, whereas shared-signal strength changes the structured population signal relative to unstructured mismatch and sampling noise. The sample-level target is $\mathbf{d}_i = \mathbf{f}_i - \widehat{\mathbf{f}}_i$; $\mathbf{d}^\star$ remains reserved for the population correction. The reported sample-wise cosines are empirical diagnostics rather than a separate sample-level identifiability theorem. The joint scan used 40 seeds and crossed $s \in \{0, 0.2, 0.4, 0.6, 0.8, 0.9, 1\}$ with shared-shift strengths $\{0.05, 0.6, 1.0\}$, while holding $m = 100$,

Table 4: Original sources and dimensions of the eight cancer single-cell datasets.

Cancer	Donors	Cells	Types	Genes	Single-cell source
COAD	89	69,504	16	28,951	HTAN VUMC (Chen et al., 2021)
BRCA	25	99,290	9	17,929	GSE176078 (Wu et al., 2021)
LUAD	25	81,834	10	18,001	GSE131907 (Kim et al., 2020)
LIHC	6	21,504	6	17,157	GSE149614 (Lu et al., 2022)
LGG	14	20,995	6	17,882	GSE232316 (Zahedi et al., 2025)
STAD	26	112,718	11	18,169	GSE183904 (Kumar et al., 2022)
HNSC	18	21,164	8	15,265	GSE181919 (Choi et al., 2023)
KIRC	7	17,188	7	15,421	GSE210042 (Davidson et al., 2023)

$n = 500$, composition radius 0.12, and noise scale 1. For specificity s, the reference construction is

$$U(s) = U_{\text{bg}} + (1 - s)M_{\text{base}} + sM_{\text{spec}},$$

followed by the common conditioning and Frobenius normalization. Within each 50-feature component module, M_{base} adds the same marker contribution to all $p = 5$ components, whereas M_{spec} assigns that contribution only to the intended component. Thus $s = 0$ gives nondiscriminative marker modules and $s = 1$ gives fully component-assigned modules. Candidate directions use only the reference matrix, marker modules, and residuals, with true fractions held out for cosine evaluation. The proposed decoder uses the intended modules with nonnegative contrast weights. The all-gene control replaces component modules by all $g = 500$ features with signed contrasts; the random-module control permutes genes into disjoint modules of the same size; and the wrong-marker control cyclically assigns each component the next component's module. These controls respectively remove enrichment, destroy component membership, or preserve enrichment while assigning the wrong identity.

At shared-shift strength 0.6, mean proposed cosine increased from -0.015 at zero marker specificity to 0.925 at full specificity, with Spearman correlation 0.643. In the joint grid, the transition occurred near $s = 0.6$: mean cosine was 0.684 at shared strength 0.05 and 0.940 at strengths 0.6 and 1.0. Below $s = 0.4$, all three shared-strength curves remained near zero; above $s = 0.8$, they converged near 0.93. In a representative high-specificity condition, population and mean sample-wise cosines were both 0.927. The matched marker direction reached 0.932, whereas all-gene residuals, randomized modules, and wrong marker assignments gave -0.902, -0.377, and -0.342.

Increasing marker specificity reduces cross-component marker-response leakage and improves the component-specific response of the marker decoder, while stronger shared signal improves separation of the structured residual-visible component from unstructured and sampling perturbations. Main Figure 1(c) shows their joint effect. Figure 6(a) places cell type on the horizontal axis and the normalized direction component on the vertical axis; dark bars are the true sample correction $\mathbf{d}_i$ and blue bars are the residual-derived direction. This means that component specificity and sufficient shared signal, not residual magnitude alone, support direction recovery.

E.2.2 Residual-Guided Updates Reduce Fraction Error

This experiment characterizes sensitivity to the correction magnitude while holding the residual-derived direction fixed. Step selection used 40 independent synthetic populations with specificity $s = 1$, shared-shift strength 0.8, $m = 100$, $n = 500$, composition radius 0.12, and unit noise scale. Each population was split into non-overlapping validation and evaluation halves. We evaluated $\lambda \in \{0, 0.1, \ldots, 2.0\}$ from the same initial estimate, with $\lambda = 0$ denoting no correction. Mean validation MAE selected $\lambda = 0.9$, with the smaller step used to break a tie. Correction effects are reported on the held-out halves, while separate shift, specificity, cohort-size, and composition-radius scans test generalization beyond the selection condition.

In the high-specificity, high-shift condition, correction reduced MAE by 0.0408 and RMSE by 0.0476 with no harmful corrections. Mean gated gain was zero for weak shifts and 0.0100 for strong shifts; the summarized gate rejected all low-specificity conditions and retained a mean MAE reduction of 0.0195 in high-specificity conditions.

Table 5: Simulation step-size sensitivity. Values are mean ± standard deviation across 40 seeds.

λ	MAE	SD	Pearson r	SD
0.0	0.0671	0.0002	0.5945	0.0235
0.1	0.0618	0.0003	0.6292	0.0222
0.2	0.0567	0.0006	0.6656	0.0210
0.3	0.0518	0.0010	0.7033	0.0200
0.4	0.0468	0.0013	0.7415	0.0190
0.5	0.0418	0.0016	0.7789	0.0178
0.6	0.0369	0.0019	0.8138	0.0162
0.7	0.0319	0.0022	0.8441	0.0143
0.8	0.0280	0.0022	0.8673	0.0134
0.9	**0.0265**	0.0027	0.8814	0.0161
1.0	0.0271	0.0035	**0.8851**	0.0230
1.1	0.0288	0.0043	0.8780	0.0321
1.2	0.0309	0.0051	0.8610	0.0415
1.3	0.0332	0.0058	0.8357	0.0501
1.4	0.0357	0.0063	0.8042	0.0570
1.5	0.0384	0.0069	0.7686	0.0622
1.6	0.0419	0.0074	0.7309	0.0658
1.7	0.0464	0.0077	0.6925	0.0679
1.8	0.0511	0.0078	0.6546	0.0690
1.9	0.0556	0.0076	0.6179	0.0692
2.0	0.0599	0.0074	0.5830	0.0685

Table 5 shows a continuous sensitivity landscape rather than a monotone trend. Relative to the no-correction baseline $\lambda = 0$, increasing λ first reduces under-correction, reaches the lowest mean MAE at $\lambda = 0.9$ and the highest Pearson correlation at $\lambda = 1.0$, and then loses performance through over-correction by $\lambda = 2.0$. Figure 6(b) gives the corresponding representative fraction refinement. The practical magnitude is selected from the observable marker-residual response in Appendix C.5.

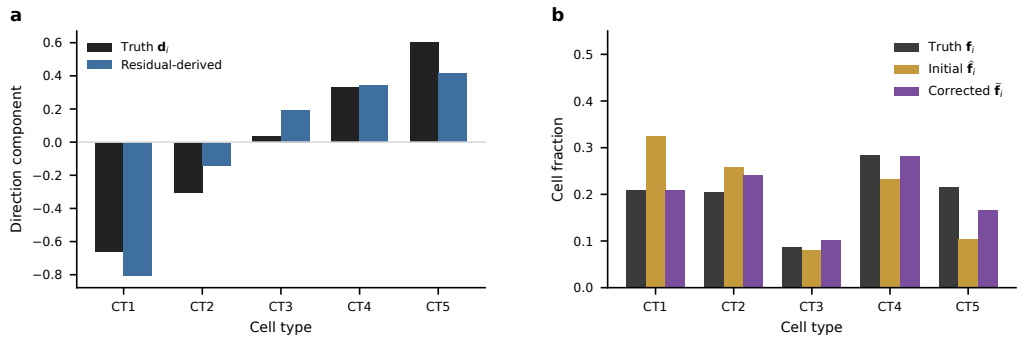

Figure 6: Direction recovery and fraction correction. (a) True and residual-derived unit directions. (b) Truth, initial estimate, and residual-guided refinement.

E.2.3 Stable Directions Can Still Be Harmful

This experiment tests Proposition 4.2 and the marker bridge $T(\mathbf{r}_{\mathcal{M}}) = CDW\mathbf{r}_{\mathcal{M}}$: high cross-sample stability cannot identify the coefficient-error direction unless decoder support, component assignment, and orientation calibration are structurally valid. The five candidate constructions are designed to isolate different failures of $T = CDW$: support-related perturbations affect the component-enriched structure encoded by W, identity perturbations disrupt the component correspondence represented by C, and sign perturbations reverse the calibrated orientation represented by D. The proposed marker candidate uses the intended non-overlapping 50-feature modules and their nonnegative component contrasts. The all-gene candidate replaces each enriched module by all $g = 500$ standardized residual

features and uses the corresponding signed all-feature contrast, creating an enrichment-support failure rather than an orientation flip. For every seed, the randomized-module candidate applies one random permutation to the 500 feature indices and takes five disjoint, size-matched blocks of 50 features; sampling is without replacement, so module size and non-overlap are preserved while component structure is removed, disrupting component-specific marker-response coherence. The sign-flipped candidate is the global negative of the proposed marker output: its marker support and W are unchanged, but its calibrated orientation is reversed, corresponding to $D_{\text{flip}} = -D$ and hence $T_{\text{flip}}(\mathbf{r}_{\mathcal{M}}) = -T(\mathbf{r}_{\mathcal{M}})$. The wrong-marker candidate cyclically maps coordinate k to module and contrast $k+1$ (with the final coordinate mapped to the first), preserving enrichment but permuting the intended component assignment, so the corresponding marker-response map $DWV_{\mathcal{M}}$ becomes cross-component rather than identity-like. These are diagnostic perturbations of marker support, orientation, and component assignment; they are not separate additive terms in the structural bound (13).

The gate combined marker-anchor alignment, population and bootstrap stability, reference marker enrichment, marker-residual enrichment, and residual-derived signal magnitude. Ground-truth fractions, true errors, and the simulated shift were excluded from gate inputs.

In the reference high-specificity, high-shift condition, the proposed marker direction had mean cosine 0.925 and population stability 1.000, and the gate accepted all 40 replicates. Stability alone did not distinguish correctness: the all-gene, sign-flipped, and wrong-marker candidates also had stability at least 0.9999, but their mean cosines were -0.924, -0.925, and -0.362. Random modules had stability 0.985 and mean cosine -0.479; 82.5% of these replicates were stable but wrong, and every such replicate was rejected. Across the pooled adverse cases classified as stable but wrong ($n = 3{,}222$), rejection was 1.000 for all four control families. The gate substantially reduced harmful accepted updates across the controls; randomized modules retained a small gated harmful rate of 0.049, so the empirical gate should be interpreted as selective rather than as an absolute safety certificate. Acceptance of the proposed direction rose from zero in low-specificity or weak-shift settings to a mean of 0.958 in high-specificity or strong-shift settings.

These controls illustrate the same non-identifiability principle as Proposition 4.2: residual stability alone does not determine alignment with the coefficient-error direction. Figure 7(b) isolates this failure. Candidates that disrupt marker support W, component correspondence C, or response orientation D can remain highly stable while pointing away from the true correction, whereas the direction-sensitive observable gate rejects these unsupported mappings. Stability is therefore insufficient without structural support for the residual-to-fraction map.

Figure 7 shows the full evidence chain. Panel (a) plots candidate-wise cosine to the true correction direction, panel (b) plots correction-direction alignment against population stability and distinguishes accepted from rejected candidates, panel (c) reports the harmful-update rate before and after gating, and panel (d) plots gate acceptance against marker specificity or shared-shift strength. Table 6 gives the reference-condition values.

Table 6: Reference-condition gate diagnostics over 40 seeds per candidate. Small subscripts report symmetric 95% confidence half-widths over seeds.

Candidate	Direction cosine	Population stability	Gate acceptance	Harmful rate, raw $\rightarrow$ gated
Proposed marker	$0.925_{\pm 0.005}$	1.000	1.000	$0.000 \rightarrow 0.000$
All-gene residual	$-0.924_{\pm 0.015}$	1.000	0.000	$1.000 \rightarrow 0.000$
Random modules	$-0.479_{\pm 0.143}$	0.985	0.125	$0.912 \rightarrow 0.049$
Sign-flipped marker	$-0.925_{\pm 0.005}$	1.000	0.000	$1.000 \rightarrow 0.000$
Wrong markers	$-0.362_{\pm 0.014}$	1.000	0.000	$0.729 \rightarrow 0.000$

E.2.4 Robustness to Bridge Perturbations

This experiment stress-tests the effective structural error in Eq. (13), $\|T(\mathbf{r}^{\star}_{\mathcal{M}}) - \mathbf{d}^{\star}\|_2 \leq \beta_{\mathcal{M}}$, which supplies the structural contribution in Theorem 4.4. The random and bias-direction perturbations form a bounded stress test around an ideal bridge rather than an application benchmark. For each of

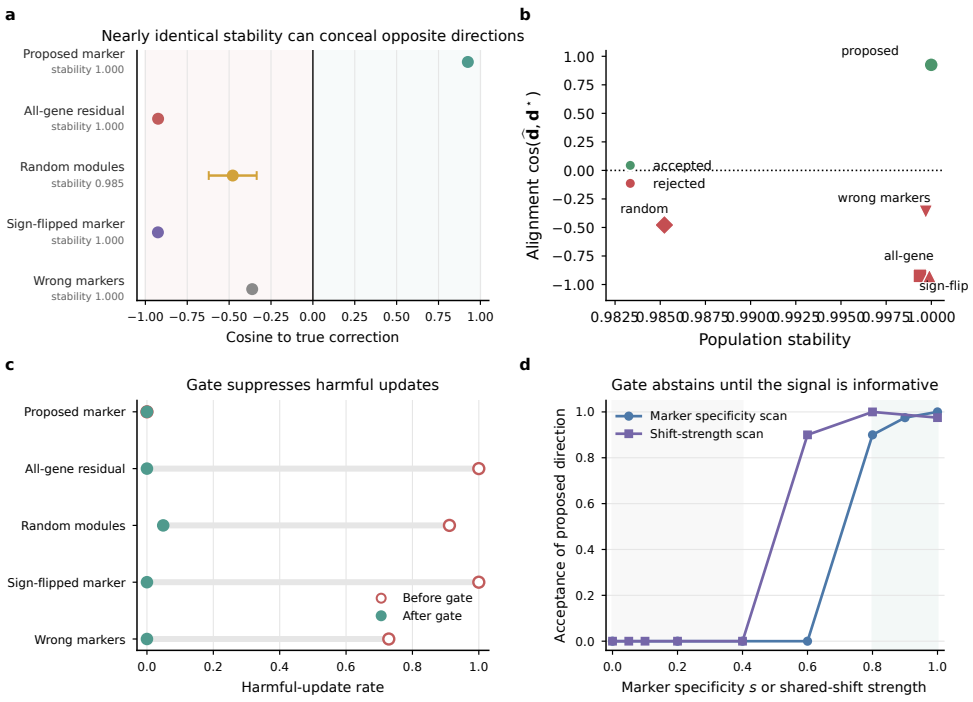

Figure 7: Gate diagnostics. (a) Direction cosine with symmetric 95% confidence half-width. (b) Correction-direction alignment versus population stability, with accepted and rejected candidates distinguished by color. (c) Harmful-update rate before and after gating. (d) Acceptance over specificity and shared-shift scans.

eight cancer-specific reference geometries, a controlled tangent mismatch imposes an OLS fraction bias with mean absolute magnitude 0.03. The residual decoder is constructed so that its unperturbed output aligns with the true fraction-error direction. The initial OLS estimate is compared with two projected bridge updates: Bridge 0.5 applies half of the recovered correction, whereas Bridge 1.0 applies the full correction. Thus the unperturbed setting is a direct sanity check: the half update should halve the imposed error and the full update should recover the target fractions to numerical precision.

The bridge information is then stressed by two distinct perturbations. Random noise of level $\sigma \in \{0, 0.03, 0.05, 0.10\}$ adds independent Gaussian feature noise with gene-wise scale σ times the across-mixture standard deviation. Bias-direction amplification of level $\alpha \in \{0, 0.03, 0.05, 0.10\}$ applies

$$\mathbf{Y}^{(\sigma,\alpha)} = \mathbf{Y} + \mathbf{\Xi}_\sigma - \alpha\, \mathbf{1}_m \mathbf{b}^\top U^\top,$$

where $\mathbf{b} \in \mathbb{R}^p$ is the original OLS bias direction shared across mixtures, and $\mathbf{1}_m \mathbf{b}^\top U^\top$ repeats its induced gene-space perturbation across the m rows of $\mathbf{Y}$. The second term is therefore a systematic column-space perturbation that increases the estimate along the original bias-related direction; it is not an angular rotation. The full 4×4 grid holds the reference, decoder, realized proportions, and baseline bias direction fixed and uses 20 random replicates per cancer and condition.

Figure 8 shows all 16 perturbation conditions as annotated heatmaps. With no perturbation, OLS MAE is 0.03000, Bridge 0.5 gives 0.01500, and Bridge 1.0 is numerically zero. Random noise gradually increases full-update error, while increasing α produces the expected systematic displacement in all three estimates. At the strongest tested combination, $\sigma = \alpha = 0.10$, the unweighted eight-cancer mean MAEs are 0.03302 for OLS, 0.01805 for Bridge 0.5, and 0.00331 for Bridge 1.0. Both update magnitudes therefore retain substantially lower fraction error than OLS throughout the tested range, although the full update progressively departs from exact recovery as bridge information is perturbed. This result supports bounded robustness of the residual-to-fraction bridge and does not imply effectiveness under arbitrary noise.

We next test the complementary zero-shift regime using the original raw-count single-cell pool for each cancer. If $\mathcal{X}_k = \{\mathbf{x}_{k1}, \ldots, \mathbf{x}_{kN_k}\}$ is the fixed empirical pool for component k, its reference column is $U_{:,k} = N_k^{-1} \sum_j \mathbf{x}_{kj}$. Every pseudo-bulk independently samples cells with replacement from these same component pools and averages their raw counts. We draw compositions with $\alpha_k = 2N_k / \max_\ell N_\ell$, convert each draw to exactly 1,000 cells by flooring and assigning the remainder to the largest component, and use the realized integer proportions as truth. Conditional on those proportions and the empirical pools, $\mathbb{E}[\mathbf{y}_i \mid \mathbf{f}_i, \{\mathcal{X}_k\}] = U\mathbf{f}_i$. Each cancer has 20 independently resampled cohorts of $m = 100$ bulks, giving 16,000 bulk realizations overall.

Each cohort uses unconstrained OLS, $\widehat{\mathbf{f}}_i = U^\dagger \mathbf{y}_i$, followed by the same simplex projection for all reported methods. Dominant-marker modules with reference fold change above 1.5 define a fixed reference-only decoder W; its marker least-squares weights supply the sign and scale calibration, so $D = I$ in this control. The tested population correction is $\widehat{\mathbf{d}} = CW\bar{\mathbf{r}}_{\mathcal{M}}$, and Bridge 0.5 and Bridge 1.0 multiply this recovered vector rather than normalizing it to a fixed length. The observable gate in Appendix E.2.3 uses disjoint marker support, population stability, marker residual enrichment, and direction magnitude without access to the true fractions or the zero-shift label. In this matched-reference control, an accepted cohort uses the prespecified $0.9\widehat{\mathbf{d}}$ update and rejection returns the projected initial estimate; this magnitude is distinct from the observable residual empirical step in Appendix C.5.

In this construction, $\Delta = H_i = 0$, $\widehat{\mathbf{f}}_i - \mathbf{f}_i = U^\dagger \boldsymbol{\xi}_i$, and $\mathbf{r}_i = Q\boldsymbol{\xi}_i$, while $\mathbf{r}^\star_{\mathcal{M}} = \mathbf{d}^\star = 0$. The deterministic check $Y = FU^\top$ recovered the raw OLS fractions to at most 5.21×10^{-15} and left residual norm at most 1.12×10^{-12} across cancers. Direction cosine to $\mathbf{d}^\star = 0$ is therefore undefined and is recorded as N/A, not as zero or one.

Table 7(b) shows that the forced corrections had no consistent effect. The unweighted mean MAE changed from 0.014843 initially to 0.014838 for Bridge 0.5 and 0.014850 for Bridge 1.0. Paired gains changed sign across replicates and cancers; for example, the small positive LIHC gains coexisted with small negative BRCA gains. The gate accepted one of 160 cancer-replicate cohorts, giving a gated mean MAE of 0.014844. A leave-one-bulk version, which excludes target i when constructing its direction, had mean MAE 0.014894 and is reported as an additional diagnostic and omitted from the table. Thus finite-cell noise can create tiny realization-specific changes, but the matched empirical means do not support a stable nonzero correction.

A supplementary disjoint condition splits cells within component and, for COAD, within the available donor identifier. Reference and target pools do not overlap and no artificial training-side perturbation is added. Its mean MAE changed from 0.016302 initially to 0.016198 and 0.016186 for the half and full updates. Finite reference estimation still leaves a fixed empirical reference–target discrepancy in this setting, so its results are reported separately rather than combined with the matched-reference control in Table 7.

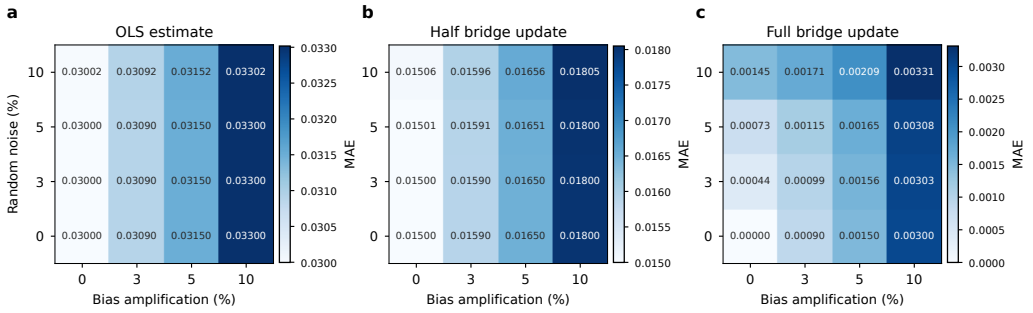

Figure 8: Mean MAE over random-noise and bias-amplification levels for (a) initial OLS, (b) half bridge update, and (c) full bridge update.

Table 7: Fraction-estimation error under bridge perturbation and matched references. Panel (a) uses the strongest bridge perturbation. Panel (b) uses matched empirical reference and sampling pools. Small subscripts report symmetric 95% bootstrap half-widths over 20 complete replicate cohorts within each cancer; Mean is the unweighted cancer average without an interval.

(a) Perturbed bridge

Cancer	Initial	Bridge 0.5	Bridge 1.0
BRCA	$0.03298_{\pm 0.00004}$	$0.01798_{\pm 0.00003}$	$0.00326_{\pm 0.00002}$
COAD	$0.03300_{\pm 0.00004}$	$0.01803_{\pm 0.00002}$	$0.00355_{\pm 0.00002}$
HNSC	$0.03301_{\pm 0.00003}$	$0.01802_{\pm 0.00002}$	$0.00317_{\pm 0.00002}$
KIRC	$0.03304_{\pm 0.00003}$	$0.01815_{\pm 0.00002}$	$0.00357_{\pm 0.00002}$
LGG	$0.03307_{\pm 0.00004}$	$0.01817_{\pm 0.00003}$	$0.00341_{\pm 0.00003}$
LIHC	$0.03300_{\pm 0.00004}$	$0.01799_{\pm 0.00004}$	$0.00325_{\pm 0.00003}$
LUAD	$0.03299_{\pm 0.00003}$	$0.01800_{\pm 0.00002}$	$0.00313_{\pm 0.00002}$
STAD	$0.03303_{\pm 0.00002}$	$0.01805_{\pm 0.00001}$	$0.00316_{\pm 0.00001}$
Mean	0.03302	0.01805	0.00331

(b) Matched reference

Cancer	Initial	Bridge 0.5	Bridge 1.0
BRCA	$0.019092_{\pm 0.000600}$	$0.019097_{\pm 0.000584}$	$0.019150_{\pm 0.000577}$
COAD	$0.006347_{\pm 0.000102}$	$0.006347_{\pm 0.000101}$	$0.006356_{\pm 0.000100}$
HNSC	$0.012313_{\pm 0.000277}$	$0.012305_{\pm 0.000280}$	$0.012309_{\pm 0.000278}$
KIRC	$0.029181_{\pm 0.000901}$	$0.029181_{\pm 0.000900}$	$0.029186_{\pm 0.000906}$
LGG	$0.008130_{\pm 0.000203}$	$0.008129_{\pm 0.000214}$	$0.008148_{\pm 0.000232}$
LIHC	$0.014505_{\pm 0.000233}$	$0.014487_{\pm 0.000225}$	$0.014482_{\pm 0.000227}$
LUAD	$0.017584_{\pm 0.000496}$	$0.017579_{\pm 0.000488}$	$0.017600_{\pm 0.000500}$
STAD	$0.011593_{\pm 0.000231}$	$0.011578_{\pm 0.000227}$	$0.011572_{\pm 0.000233}$
Mean	0.014843	0.014838	0.014850

E.3 Transcriptomic Deconvolution Benchmarks

The metric and uncertainty definitions appear at the start of Appendix E. Main Table 1 reports cancer- and cohort-specific MAE with symmetric 95% bootstrap confidence half-widths. Figures 9-10 provide the complementary RMSE, Pearson, and CCC comparisons.

E.3.1 Cross-Cancer Pseudo-Bulk Benchmark

The pseudo-bulk benchmark contains BRCA, COAD, HNSC, KIRC, LGG, LIHC, LUAD, and STAD, whose original single-cell sources are listed in Table 4. Each cancer contributes 100 mixtures. Let c_k be the training-set count of cell type k and set $\alpha_k = 2c_k / \max_\ell c_\ell$. A composition is drawn from $\mathrm{Dirichlet}(\boldsymbol{\alpha})$ and converted to exactly 1,000 integer cell counts. This concentration choice preserves the reference cohort's broad type-frequency pattern while avoiding extremely concentrated mixtures, creating mild composition similarity without forcing identical fractions. Raw held-out cells are then sampled with replacement within type and their counts are summed. The realized integer count fractions, not the continuous Dirichlet draw, are the scoring truth.

Reference-side mismatch is introduced before the reference matrix U is constructed. For each gene, its empirical coefficient of variation (CV) is estimated from the raw training single cells and used to scale the magnitude of the training-side perturbation, so genes with larger empirical cell-to-cell variability receive larger expression fluctuations. An additional random fluctuation between 1% and 5% is applied in the same direction as the underlying perturbation, changing its magnitude without reversing its sign. This perturbation acts directly on the unprocessed training-cell expression profiles before downstream feature selection and reference construction; the held-out cells used to generate pseudo-bulk targets are not subjected to it. The resulting reference U is therefore altered indirectly through the perturbed training cells rather than by a post-hoc transformation of U itself.

This construction addresses a limitation of same-source pseudo-bulk evaluation. When both the reference and target pools are drawn from the same single-cell cohort, an unperturbed split produces an unusually well-matched reference–target problem whose remaining variability is largely finite-cell sampling noise. The CV-guided training-side perturbation introduces controlled, gene-heterogeneous reference mismatch while retaining the empirical single-cell structure, providing a closer stress test for the reference–bulk discrepancies encountered in real deconvolution. The complementary matched-reference control in Appendix E.2.4 omits this artificial perturbation while retaining finite-cell resampling; forced corrections then have essentially no systematic mean effect and the observable gate accepts only one of 160 cancer–replicate cohorts. The deterministic check $Y = FU^\top$ reported there additionally removes sampling fluctuation and recovers the matched model to numerical precision.

The comparison includes ReDeconv, MuSiC, BayesPrism, Scaden, TAPE, DECODE, and Count Bridges (Lu et al., 2025; Wang et al., 2019; Chu et al., 2022; Menden et al., 2020; Chen et al., 2022; Zhao et al., 2026; Fishman et al., 2026). Every method is evaluated on the same 100 target mixtures and aligned component labels within a cancer. Because both pools originate from the same single-cell dataset, the benchmark does not contain a genuine single-cell-versus-bulk sequencing-platform effect; the CV-guided training-side perturbation serves instead as a controlled surrogate for reference-to-target mismatch. All methods receive the same benchmark-level reference and pseudo-bulk target inputs and start from the same feature universe, without additional method-specific cross-platform normalization. Preprocessing intrinsic to an individual implementation remains part of that method.

Across the complete eight-method comparison, PopLS has the best unweighted cross-cancer mean on all three complementary metrics: RMSE 0.03454, Pearson 0.97594, and CCC 0.97437. In Figure 9, boxes summarize the eight cancer-level values and overlaid points show individual cancers. PopLS has the lowest RMSE, highest Pearson, and highest CCC in HNSC, LIHC, and STAD. Scaden leads all three metrics in BRCA, KIRC, and LGG; BayesPrism has the lowest RMSE in COAD, whereas MuSiC has the highest COAD Pearson and CCC and leads all three metrics in LUAD. The mean advantage is therefore not a universal per-cancer win.

This means that the population-residual method improves the overall accuracy-correlation-calibration tradeoff under fixed reference mismatch, but the cancer-level exceptions remain scientifically important. The conclusion is about cross-cancer performance, not dominance on every dataset.

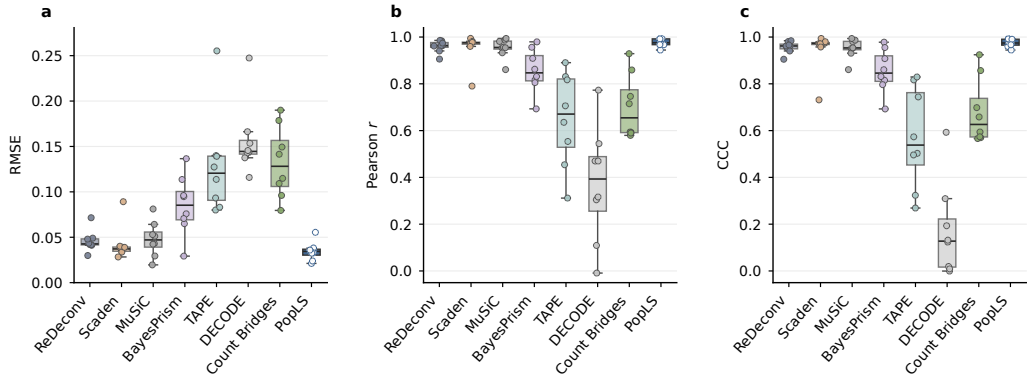

Figure 9: Pseudo-bulk performance across eight cancers and eight methods. (a) Matrix RMSE. (b) Flattened Pearson r. (c) Lin's CCC. Points are cancers; boxes summarize their distributions.

Main Table 1(a) gives the corresponding MAE view. PopLS has the lowest unweighted mean MAE, 0.0199, and the lowest point estimate in BRCA, HNSC, LGG, LIHC, and STAD. MuSiC leads COAD and LUAD, while Scaden leads KIRC.

Step-size sensitivity. Keeping the reference, markers, gate, and recovered direction fixed, we compared the observable selector with a retrospective MAE scan over 601 equally spaced values in $[0, 3]$. We define λ_{opt} post hoc as the value minimizing ground-truth MAE over this scan. Across the eight pseudo-bulk datasets, the selected step retained a median 98.5% (range 69.8–100%) of the

improvement attained by the scanned optimum and lay in the prespecified 10% near-optimal interval for five datasets. The scanned optimum reached the upper boundary for COAD and STAD, so those optima should be interpreted as boundary-limited (Figure 10(d)).

E.3.2 Evaluation on Real-Bulk Cohorts

The real-bulk evaluation uses two cohorts distributed through deconvData (Dietrich et al., 2026). Altman combines 322 blood bulk RNA-seq profiles (SRX4218303-SRX4218503) with FACS proportions and a Tabula Sapiens blood/PBMC reference (Tabula Sapiens Consortium, 2022). Its evaluated components are Lymphocytes, Monocytes, and Neutrophils. Morandini combines immune-cell bulk profiles (GSM5774456-GSM5774614) with FACS proportions and the Hao PBMC CITE-seq reference (GSE164378) (Hao et al., 2021). Of the 159 bulk profiles, 156 have FACS truth and are included after exact identifier matching. Its evaluated components are B cells, Monocytes, NK cells, T cells CD4, and T cells CD8. Table 9 separates the bulk/truth source from its paired single-cell reference. Table 8 summarizes the method inputs and configurations used for the real-bulk comparison.

Table 8: Reference inputs, target inputs, and configurations for the real-bulk comparison.

Method	Reference input	Target input	Configuration
ReDeconv	CLTS-normalized reference	Bulk TPM	Default
MuSiC	Raw counts with subject labels	Bulk TPM	Default
BayesPrism	Raw counts	Rounded bulk TPM	Default
Scaden	Raw counts	Bulk TPM	5,000 simulated mixtures; 128 epochs; batch 128; seed 42
TAPE	Raw counts	Bulk TPM	5,000 simulated mixtures; 128 epochs; batch 128; seed 42; adaptive prediction
DECODE	Raw counts	Bulk TPM	1,000/300 training/validation mixtures; 40+80 epochs; seed 2021
Count Bridges	Raw counts	Bulk TPM	Up to 256 genes; 1,000 training steps; 500 generated cells
PopLS (ours)	Raw counts	Bulk TPM	Default

Table 9: Real-bulk cohorts, truth sources, accessions, and paired single-cell references.

Cohort	Bulk/truth source	Accession	Paired single-cell reference	Samples	Components
Altman	deconvData; FACS	SRX4218303–SRX4218503	Tabula Sapiens blood/PBMC, 10x	322	3
Morandini	deconvData; FACS	GSM5774456–GSM5774614	Hao PBMC CITE-seq, GSE164378	156	5

Each method receives its method-appropriate reference and target representation. PopLS applies the population-residual refinement principle described in Appendix D. Only cell types that map to the same biological component in the reference and orthogonal truth are retained. Cell types and cells outside this intersection are excluded when constructing the benchmark reference and therefore do not enter deconvolution or scoring. For each cohort, predictions and orthogonal truth are restricted to the same retained component set $\mathcal{K}$ and renormalized over that set before scoring:

$$f_{ik}^{\mathrm{eval}} = \frac{f_{ik}}{\sum_{\ell \in \mathcal{K}} f_{i\ell}}, \qquad \widehat{f}_{ik}^{\mathrm{eval}} = \frac{\widehat{f}_{ik}}{\sum_{\ell \in \mathcal{K}} \widehat{f}_{i\ell}}.$$

Thus predictions and truth use the same retained-component denominator definition. Samples are matched by exact identifier. The reported mean is the unweighted average of the two cohort-level values, not a pooled sample-level score.

PopLS obtains matrix RMSE 0.09151 on Altman and 0.06392 on Morandini. It also has the highest Pearson (0.89975 and 0.81908) and CCC (0.89168 and 0.81028) in both cohorts. Its two-cohort means are RMSE 0.07772, Pearson 0.85942, and CCC 0.85098. The remaining methods show larger and more dataset-dependent errors. Figure 10(a–c) shows the two cohort-level values for every method; its boxes summarize those two values and do not estimate sampling uncertainty.

This means that the final population-residual procedure is consistently strongest on absolute error, pattern recovery, and calibration for these two retained cohorts. Because only two cohorts are available, the result supports cross-cohort consistency for Altman and Morandini rather than a broad population-level uncertainty claim.

The MAE analysis in main Table 1(b) is consistent with the RMSE comparison: PopLS has the lowest point estimate on Altman, Morandini, and their unweighted mean, with mean MAE 0.0585.

Step-size sensitivity. The same fixed-input scan on $[0, 3]$ selected observable steps 0.50 and 1.50 for Altman and Morandini, compared with retrospective optima 0.925 and 1.19. Their empirical-versus-optimal MAEs were 0.0712 versus 0.0605 and 0.0459 versus 0.0450, retaining 75.8% and 89.8% of the attainable improvement. Neither selected step fell inside its 10% near-optimal interval (Figure 10(d)).

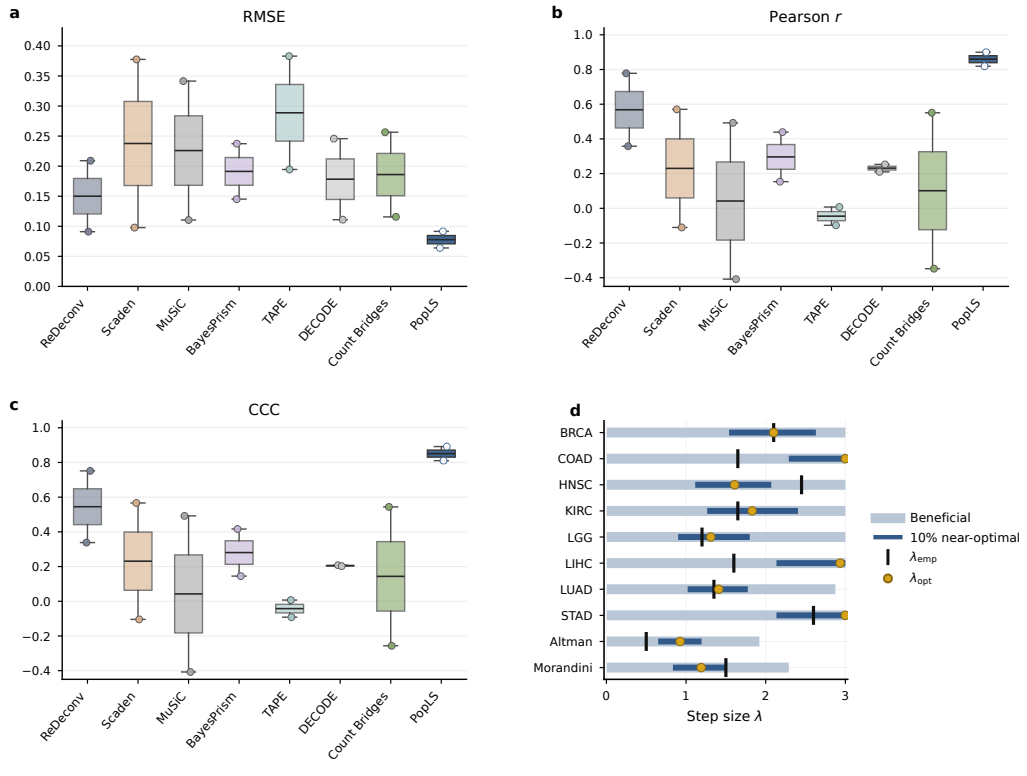

Figure 10: Real-bulk performance and step-size calibration. (a) Matrix RMSE. (b) Flattened Pearson r. (c) Lin's CCC. In (a–c), points are cohorts and boxes summarize the two values. (d) Dataset-specific beneficial and 10% near-optimal intervals, λ_{emp}, and the retrospective λ_{opt} over $\lambda \in [0, 3]$.

F Extended Related Work

Residual analysis in regression and reference-shifted inverse problems. Residual-based methods examine the errors left after fitting to determine whether the model has missed a systematic pattern and whether particular observations or groups are responsible for the poor fit. Classical influence diagnostics identify observations whose removal would substantially change the fitted regression,

using both their fitting errors and their positions in the design space (Cook, 1977). More recent perturbation diagnostics measure how fitted results change under controlled perturbations of the observed data and provide finite-sample guarantees for this sensitivity analysis (Fisher et al., 2023). Another line checks whether residuals still contain reproducible patterns after fitting. Moment and conditional-moment tests compare selected residual averages or relationships with what the fitted model predicts (Li et al., 2023); related conditional-moment procedures test residual behavior across observed covariates (Escanciano, 2024); and functional residual methods provide flexible residual representations when the observations have functional structure (Liu et al., 2025). These works use post-fit errors primarily for diagnosis.

PopLS retains this diagnostic role and additionally uses a stable marker-restricted cohort residual signal to recover information about hidden fraction error under a shared reference shift. In the regime $H_i = 0$, Theorem 4.1 gives

$$\mathbf{r}_i = Q\Delta\mathbf{f}_i + Q\boldsymbol{\xi}_i, \qquad \widehat{\mathbf{f}}_i - \mathbf{f}_i = U^\dagger\Delta\mathbf{f}_i + U^\dagger\boldsymbol{\xi}_i.$$

Thus the same reference shift affects two quantities differently: $Q\Delta$ determines the part that remains visible in the reconstruction residual, whereas $U^\dagger\Delta$ determines the systematic error in the estimated fractions. Equivalently, $\Delta_Q = Q\Delta$ is residual-visible, while the component represented by $\Delta_P = P\Delta$ is relevant to the fraction estimate. Proposition 4.2 shows that residual magnitude or stability alone cannot determine the hidden fraction error: a stable residual may occur without a fraction error, and a fraction error may occur without a residual. A structural residual-to-fraction bridge is therefore required before the residual signal can support correction.

Reference-shifted estimation and broader distribution shift. Distribution-shift methods study how estimation or prediction changes when the target data follow a different relation or distribution from the one used to construct the estimator, and how the resulting loss of alignment can be reduced. Least-squares analyses quantify how a changed linear relation affects estimation error, prediction risk, and sample requirements (Maran & Szepesvári, 2025). Covariate-shift methods study changes in the input distribution between source and target data (Amortila et al., 2024). Related work studies mismatch in synthetic-control construction (Nazaret et al., 2024), transfer under parameter, covariate, and residual shifts (Bai et al., 2024), and covariance changes in synthetic-data selection (Rezaei et al., 2026). Robust-regression methods address complementary failures such as sparse contamination, outliers, or dependent noise (Sasai & Fujisawa, 2025; Ziemann et al., 2023).

PopLS considers the more structured reference shift $M = U + \Delta$ after the common-reference fit has already been obtained. It separates the effect of the shift that is absorbed into the estimated fractions from the effect that remains in the observable residual, and uses an explicit structural bridge to connect these two effects. The problem is therefore not only to make an estimator robust to a changed distribution, but to learn a correction-relevant signal from the post-fit behavior of many targets that share the same shifted reference.

Population residuals and multi-sample structure. Multi-sample residual methods combine fitting errors across observations, time points, or groups so that weak but repeated departures from a fitted model are easier to detect than they would be from a single noisy residual. CUSUM procedures accumulate residual evidence along an ordered sequence to detect changes in the fitted relation (Brown et al., 1975). Related change-point tests determine whether model parameters change somewhere in the sample even when the change location is unknown (Andrews, 1993). Grouped-residual procedures combine evidence within predefined groups to identify groups that behave abnormally relative to the fitted model (Kwon et al., 2025). These methods use aggregated residuals mainly to strengthen detection of instability or outlying groups.

PopLS retains this diagnostic role but additionally treats the aggregated marker residual as an estimable population object that can support fraction correction. Under $H_i = 0$,

$$\bar{\mathbf{r}}_{\mathcal{M}} = R_{\mathcal{M}}Q\Delta\bar{\mathbf{f}} + \frac{1}{m}\sum_{i=1}^{m} R_{\mathcal{M}}Q\boldsymbol{\xi}_i = \mathbf{r}^\star_{\mathcal{M}} + \frac{1}{m}\sum_{i=1}^{m} R_{\mathcal{M}}Q\boldsymbol{\xi}_i, \qquad \mathbf{d}^\star = -U^\dagger\Delta_P\bar{\mathbf{f}}.$$

Increasing the cohort size or aggregate size concentrates $\bar{\mathbf{r}}_{\mathcal{M}}$ around the marker-visible population signal $\mathbf{r}^\star_{\mathcal{M}}$. Lower conditional marker variability tightens this concentration, but concentration alone does not identify $\mathbf{d}^\star$; the marker-response bridge supplies the structural information needed to

connect residual space to fraction space. The discrepancy between the population signal and target i additionally depends on the deterministic composition displacement $\|\mathbf{f}_i - \bar{\mathbf{f}}\|$, so accurate population estimation and a bounded population-to-target discrepancy are distinct requirements.

Multi-task regression and shared population structure. Multi-task methods improve an individual estimation problem by borrowing information from other related tasks. The shared information can describe which tasks are similar, where their parameters are centered, which model components are common, or what summary information is available across tasks. Task-specific unknowns correspond to $\{\mathbf{f}_i\}$, while a shared parameter center corresponds to $\bar{\mathbf{f}}$. Some methods allow closely related tasks to share more information while preventing clearly different tasks from being forced toward the same solution (Duan & Wang, 2023). Other methods estimate task-specific parameters around a common center and use each task's second-moment information to decide how strongly it should rely on that center (Kim, 2026). Shared-component methods allow common structure together with task-specific deviations and explicitly guard against harmful transfer between dissimilar tasks (Ilbert et al., 2024). Related tasks can also be estimated jointly when only task-level summary statistics, rather than the original observations, are available (Knight & Duan, 2023). Meta-learning approaches use a population of previous tasks to learn information that can improve estimation on a new task (Wang & Arora, 2024).

The key structural difference is that every PopLS target $\mathbf{y}_i$ is inverted through the same fixed reference U. Consequently, all initial fits share the same $U^\dagger$ and the same projectors

$$P = UU^\dagger, \qquad Q = I_g - P.$$

Conventional multi-task problems may share features or parameter structure, but their tasks generally retain task-specific observations, design information, covariance, or loss geometry. Because PopLS targets share the same U, their post-fit residuals are directly comparable. More importantly, directly averaging the individual composition estimates does not remove the systematic bias induced by the shared reference shift. Under $H_i = 0$,

$$\frac{1}{m}\sum_{i=1}^{m}(\widehat{\mathbf{f}}_i - \mathbf{f}_i) = U^\dagger \Delta\bar{\mathbf{f}} + \frac{1}{m}\sum_{i=1}^{m} U^\dagger \boldsymbol{\xi}_i.$$

As the cohort grows, the zero-mean sampling term can concentrate toward zero while the shared reference-shift term $U^\dagger\Delta\bar{\mathbf{f}}$ remains. A population center formed directly from the target-wise composition estimates can therefore stay systematically displaced from the true composition center even in a large cohort. PopLS instead estimates the observable population object in marker-residual space,

$$\bar{\mathbf{r}}_{\mathcal{M}} \longrightarrow \mathbf{r}^\star_{\mathcal{M}} = R_{\mathcal{M}} Q \Delta\bar{\mathbf{f}},$$

and then uses separately justified marker-response structure to map $\mathbf{r}^\star_{\mathcal{M}}$ toward $\mathbf{d}^\star = -U^\dagger\Delta_P\bar{\mathbf{f}}$. This separates estimation of the shared population signal, structural identification of the hidden correction, and the discrepancy between the population center and an individual target.

Deconvolution solvers and the PopLS distinction. Least-squares deconvolution estimates cell fractions by choosing a non-negative combination of reference profiles that best reconstructs the observed bulk expression. Methods in this family mainly differ in how genes are weighted or how the bulk and reference are prepared before solving the fit. MuSiC solves weighted NNLS and uses variability estimated from the single-cell reference to construct its gene weights (Wang et al., 2019). DWLS uses weighted least squares and updates the contribution of genes according to the current mixture estimate, leading to its dampened weighting scheme (Tsoucas et al., 2019). ReDeconv uses variance-weighted NNLS together with reference-side normalization and stable-gene selection (Lu et al., 2025). Bisque first transforms bulk expression to better match the single-cell reference and then estimates the fractions with NNLS (Jew et al., 2020).

Other reference-based methods replace least squares with different inference mechanisms. CIBERSORTx uses ν-support vector regression (Newman et al., 2019). Bayesian or probabilistic solvers include BayesPrism (Chu et al., 2022) and Statescope (Janssen et al., 2026). Neural approaches include Scaden (Menden et al., 2020), TAPE (Chen et al., 2022), and DECODE (Zhao et al., 2026). cellGeometry uses non-negative geometric inference (Lau et al., 2026), whereas Count Bridges uses a count-space generative bridge (Fishman et al., 2026).

Across the solver families above, most reference-based methods treat each target bulk primarily as its own inverse problem. PopLS introduces cross-target information after the initial target-wise least-squares fits: all targets are first fitted with the common broad reference U, their post-fit residuals are restricted to the selected marker rows, and these marker residuals are aggregated to estimate

$$\mathbf{r}^{\star}_{\mathcal{M}} = R_{\mathcal{M}} Q \Delta \bar{\mathbf{f}}.$$

A separately justified marker-response bridge maps this signal toward the hidden population correction

$$\mathbf{d}^{\star} = -U^{\dagger} \Delta_P \bar{\mathbf{f}}.$$

The marker-restricted population residual is used both to assess whether an individual target is compatible with the cohort-level residual structure and, when the bridge is informative, to guide a gated fraction correction. PopLS therefore does not replace the initial least-squares solver; it adds a population-residual stage after the target-wise fits. The present theory focuses on this least-squares setting because the common U, P, and Q make the relation between the observable residual shift and the hidden fraction error explicit.

Broader reference-based mixture inference. Reference-based mixture estimation also appears outside transcriptomics. Examples include cell-free-DNA tissue-of-origin inference, microbial source tracking, and related source-separation problems (De Ridder et al., 2024; Huang et al., 2024). In each setting, a target mixture is interpreted using component references measured separately from that target. A systematic shift between those references and the true target components can therefore bias the inferred mixture weights, creating the same distinction between observable reconstruction error and fraction-relevant error studied here.